TECHNOSUP

Les FILIÈRES TECHNOLOGIQUES des ENSEIGNEMENTS SUPÉRIEURS

STATISTIQUES

Statistique mathématique

Applications commentées

Jean-Pierre BOULAY

ellipses

Pour plus de livres visitez notre site web :

biblio-scientifique.com

TECHNOSUP

Les FILIÈRES TECHNOLOGIQUES des ENSEIGNEMENTS SUPÉRIEURS

STATISTIQUES

Statistique mathématique

Applications commentées

Jean-Pierre BOULAY

ellipses

Du même auteur, dans la même collection

- Calcul des probabilités 224 p. (B) J-P. BOULAY

Dans la même collection

- Probabilités pour modéliser et décider 256 p. (A) N. SAVY
- Modélisation probabiliste pour l'ingénieur 312 p. (C) A. SMOLARZ
- Assimiler et utiliser les statistiques 288 p. (A) L. PIBOULEAU
- Statistiques et expérimentation en biologie 192 p. (A) J-Cl. LABERCHE

ISBN 978-2-7298-5602-1

32, rue Bargue 75740 Paris cedex 15

www.editions-ellipses.fr

AVANT-PROPOS

Ayant pour objet d'exploiter et d'interpréter l'information contenue dans des données le plus souvent entachées de variations et d'incertitudes, les méthodes statistiques recouvrent :

- la statistique descriptive dont le but est de traduire l'information sous forme synthétique et efficace et de dégager les caractéristiques majeures du phénomène étudié ;
- la statistique inférentielle (dite mathématique) qui, à partir des données contenues dans un échantillon, vise à formuler prévisions et décisions étendues à toute une population, ce qui requiert modélisation probabiliste et évaluation des risques d'erreurs.

C'est ce dernier aspect qui est développé dans cet ouvrage, les applications quasi universelles des outils présentés touchant les domaines les plus divers comme la politique (sondages), les sciences sociales (prévisions économiques), la médecine (diagnostics et expérimentation des traitements), l'industrie (contrôle de qualité), l'agriculture (rendements et procédés), la biologie (évolution des espèces),...

***

Historiquement, les premières opérations statistiques remontent, à travers recensements et gestions diverses, à plus de 2000 ans avant notre ère (Egypte, Chine, Incas...). Toutefois, il faut attendre le XVIII^{ème} siècle pour voir apparaître représentations graphiques et constructions de tableaux qui forment les bases de la statistique descriptive, la statistique inférentielle ne trouvant, quant à elle, son essor qu'au XX^{ème} siècle après l'émergence, au cours du XIX^{ème} siècle, de l'étude des lois fondamentales pour la modélisation probabiliste à commencer la loi normale.

Ainsi Ronald Aylmer FISHER (1890-1962) est-il présenté comme le père de l'estimation statistique selon le principe du « maximum de vraisemblance » et contribue-t-il au développement de la théorie des tests d'hypothèse, William Sealy GOSSET alias STUDENT (1876-1937) apportant ici, de par son expérience professionnelle en production industrielle, une contribution déterminante au plan des applications.

Egon PEARSON (1895-1980) et Jezzy NEYMAN (1894-1981) posent quant à eux les bases de la théorie des tests, Jezzy NEYMAN étant également, suite à ses travaux sur l'estimation par intervalle de confiance et sur l'échantillonnage par stratification, un fondateur des techniques modernes de sondage.

Enfin, Abraham WALD (1902-1950) invente le concept des tests statistiques séquentiels ouvrant ainsi une évolution importante des applications statistiques dans le domaine du contrôle industriel de qualité.

Comme le lecteur pourra le constater dans la table des matières à la lumière de la diversité marquante des tests aujourd'hui disponibles, la présente liste susmentionnée des pionniers de la statistique mathématique, ne saurait être exhaustive.

Entre autres, on peut citer ainsi, William Edwards DEMING (1900-1993), Wallodi WEIBULL (1887-1979), Charles SPEARMAN (1863-1945), précurseurs respectivement, du management par la qualité, de la théorie de la fiabilité, de l'analyse factorielle.

***

S'appuyant sur le calcul des probabilités dont les techniques usuelles sont supposées acquises et maîtrisées, cet ouvrage a pour objet de présenter les principales méthodes utilisées en statistique mathématique, leur illustration par des problèmes concrets étant, dans ce cadre, une préoccupation majeure.

Il s'adresse donc à un large public qui est celui des étudiants des écoles d'ingénieurs, des I.U.T, mais aussi des écoles de commerce et des facultés dans les nombreux domaines qui ont été mentionnés précédemment. Il concerne également ceux qui, dans un cadre professionnel, sont confrontés à des problèmes d'estimation ou de décision statistique.

Le plan est classique. Dans le premier chapitre, les lois de probabilité rencontrées usuellement en statistique sont rappelées et les distributions d'échantillonnage les plus courantes sont caractérisées et étudiées. Par ailleurs, une introduction à la pratique des sondages y est également présentée.

Le second chapitre traite le délicat problème de l'estimation ponctuelle et par intervalle de confiance, les propriétés des estimateurs et leurs modes de construction étant largement développées et illustrées à travers divers modèles classiques et plusieurs techniques particulières d'estimation.

Le troisième chapitre rassemble quant à lui, une présentation des principaux tests paramétriques et non paramétriques qui peuvent être mis en œuvre pour répondre à la question du choix entre deux hypothèses (décision statistique), notamment à des fins de conformité, de comparaison, d'ajustement, d'indépendance. Les exemples d'application y tiennent là encore une place majeure.

Enfin, le dernier chapitre porte sur une initiation aux modèles de régression linéaire simple et multiple, dûment illustrée.

Quant à la présentation, elle comprend pour chacun des chapitres ci-dessus, un rappel de cours restreint au strict nécessaire, puis un ensemble consistant d'applications commentées regroupées par thèmes, une série d'exercices corrigés complétant ces développements aux fins d'entraînement.

J'adresse enfin, tous mes remerciements les plus chaleureux au Professeur Claude CHEZE, directeur de la collection, pour toute la confiance qu'il m'a accordée et ses encouragements à concrétiser le difficile challenge d'un tel projet, à la suite de mon précédent ouvrage paru en 2008 et portant sur le calcul des probabilités.

J'associe également à ces remerciements Philippe MONVOISIN, Professeur et responsable du département informatique à l'Ecole Spéciale des Travaux Publics, et mon fils William, pour toute l'aide technique apportée dans le montage de cet ouvrage.

TABLE DES MATIERES

Chapitre I : Echantillonnage

Chapitre II : Estimation

Chapitre IV : Régression

CHAPITRE I

ECHANTILLONNAGE

A - Rappels de cours

1. Lois de probabilités

1.1 Définitions et caractérisations

Les principales lois de probabilités, leurs conditions de validité, et leurs paramètres représentatifs sont rappelées dans le tableau ci-dessous:

Loi	Nature	Définition	Caractérisation	E(X)	Var(X)
BERNOULLI	Discrète	Variable indicatrice d'un caractère au cours de n épreuves de BERNOULLI (*)	Valeurs : $\{0,1\}$ $\Pr ob(X=0)=q$ $\Pr ob(X=1)=p$	p	q
Binomiale **B(n,p)**	Discrète	Occurrence d'un caractère au cours de n épreuves de BERNOULLI indépendantes	Valeurs : $\{0,1,2....,n\}$ $\Pr ob(X=x)=C_n^x p^x q^{n-x}$	$n.p$	$n.p.q$
Hypergéométrique	Discrète	Occurrence d'un caractère au cours de n épreuves de BERNOULLI dépendantes (à savoir le tirage sans remise d'un échantillon de taille n dans une population de taille N)	Valeurs : $\{0,1,2....,n\}$ $\Pr ob(X=x)=\frac{C_{N.p}^x . C_{N.q}^{n-x}}{C_N^n}$	$n.p$	$\frac{N-n}{N-1}.npq$
POISSON **P(a)**	Discrète	Occurrence des événements relativement rares	Valeurs : N $\Pr ob(X=x)=e^{-a}.\frac{a^x}{x!}$	a	a

(*) Pour rappel, l'épreuve de BERNOULLI est une épreuve dans laquelle, seuls sont possibles, les résultats C (avec la probabilité p) et $\overline{C}$ (avec la probabilité complémentaire $q=1-p$).

Loi	Nature	Définition	Caractérisation	E(X)	Var(X)
Géométrique	Discrète	Nombre de tentatives nécessaires jusqu'à l'obtention du caractère C à travers des épreuves de BERNOULLI indépendantes	Valeurs : N^* $\Pr ob(X=x)=q^{x-1}.p$	$\frac{1}{p}$	$\frac{q}{p^2}$
Binomiale négative	Discrète	Nombre de tentatives jusqu'à l'obtention r fois d'un caractère C à travers des épreuves de BERNOULLI indépendantes	Valeurs : $[r,+\infty[$ $\Pr ob(X=x)=C_{x-1}^{r-1}p^r.q^{x-r}$	$\frac{r}{p}$	$\frac{r.q}{p^2}$
Uniforme $U_{[a,b]}$	Continue	Probabilité uniforme sur $[a,b]$	Valeurs : $[a,b]$ $f(x)=\frac{1}{b-a}.1_{[a,b]}(x)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Exponentielle	Continue	Caractéristique des durées de vie des équipements qui ne vieillissent pas (loi « sans mémoire »)	Valeurs : R^+ $f(x)=\lambda.e^{-\lambda.x}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Gamma n	Continue	Loi de la somme de n variables aléatoires exponentielles indépendantes	Valeurs : R^+ $f(x)=\frac{\lambda^n.e^{-\lambda.x}.x^{n-1}}{(n-1)!}$	$\frac{n}{\lambda}$	$\frac{n}{\lambda^2}$
Normale $N(m,\sigma)$	Continue	Loi « universelle » vers laquelle convergent une large part des autres lois	Valeurs : R $f(x)=\frac{1}{\sigma.\sqrt{2\pi}}.e^{\frac{(x-m)^2}{2.\sigma^2}}$ (tables de valeurs en annexes)	m	σ^2

Loi	Nature	Définition	Caractérisation	E(X)	Var(X)
Chi-deux $\chi^2(n)$	Continue	Loi de la somme $\sum_{i=1}^{i=n} X_i^2$ où les X_i sont des variables normales, centrées, réduites, et indépendantes	Valeurs : R^+ $f(x)=\frac{x^{\frac{n}{2}-1}.e^{-\frac{x}{2}}}{2^{\frac{n}{2}}.\Gamma(\frac{n}{2})}$ avec $\Gamma(n)=\int_0^{+\infty} t^{n-1}.e^{-t}.dt$ (tables de valeurs en annexes)	n	$2n$
STUDENT **T(n)**	Continue	Loi de $T=\frac{X}{\sqrt{\frac{Y}{n}}}$ où X est normale centrée réduite et où Y suit la loi du chi -deux $\chi^2(n)$	Valeurs : R^+ $f(x)=\frac{\Gamma(\frac{n+1}{2}).(1+\frac{x^2}{n})^{-\frac{(n+1)}{2}}}{\sqrt{n.\pi}.\Gamma(\frac{n}{2})}$ (tables de valeurs en annexes)	0 n≻1 (indéterminée pour n=1)	$\frac{n}{n-2}$ n≻2 (infinie pour n≤2)
FISHER SNEDECOR **F(n,p)**	Continue	Loi de $F=\frac{X/n}{Y/p}$ où X et Y suivent respectivement les lois $\chi^2(n)$ et $\chi^2(p)$	Valeurs : R^+ $f(x)=\frac{n^{\frac{n}{2}}.p^{\frac{p}{2}}.\Gamma(\frac{n+p}{2}).x^{\frac{n}{2}-1}}{\Gamma(\frac{n}{2}).\Gamma(\frac{p}{2}).(n.x+p)^{\frac{n+p}{2}}}$ (tables de valeurs en annexes)	$\frac{p}{p-2}$ p≻2	Voir renvoi (*) ci-dessous.

(*) La variance de la loi de FISHER SNEDECOR est égale à $(\frac{p}{p-2})^2.\frac{2.(n+p-2)}{n.(p-4)}$ pour $p \succ 4$.

1.2 Propriétés de convergence

• Le **théorème central limite** tient une place fondamentale dans la justification desdites convergences. Pour rappel, son énoncé est le suivant :

Soit $X_n, n \in N$, une suite de variables aléatoires indépendantes de même loi d'espérance m et de variance σ^2 finies. Alors, la somme $Z=\sum_{i=1}^{i=n} X_i$ converge pour n assez grand (en pratique à partir de n=30) vers la loi normale de moyenne $n.m$ et d'écart -type $\sigma.\sqrt{n}$.

• *Sur un plan plus général,* les lois de probabilités mentionnées dans le paragraphe précédent satisfont à un **ensemble de convergences**, essentielles pour les applications en statistique, et qui s'énoncent comme suit :

- La **loi hypergéométrique** *converge*, pour N grand, vers la **loi binomiale** $B(n, p)$ (condition le plus souvent satisfaite dès lors qu'on est amené à pratiquer un sondage).

Pratiquement, cette convergence est satisfaite pour $\frac{N}{n} \geq 10$.

- La **loi binomiale** $B(n, p)$ *converge*, pour n assez grand et p ni trop voisin de 1 ni de 0 vers la **loi normale** $N(m = n.p, \sigma^2 = n.p.q)$.

C'est le **théorème de MOIVRE- LAPLACE** qui résulte de l'application du théorème central limite au cas particulier de la somme de n variables aléatoires de BERNOULLI indépendantes.

Au plan pratique, plusieurs conditions de validité de cette convergence sont applicables. On peut retenir entre autres, $n \geq 30$ **et** $n.p \succ 5$ **et** $n.q \succ 5$, ou, $n \geq 30$ **et** $n.p \geq 15$ **et** $n.p.q \succ 5$.

- La **loi binomiale** $B(n, p)$ *converge*, pour n assez grand, et p faible (ou voisin de 1) vers la **loi de POISSON** de paramètre $a = n.p$.

Au plan pratique, on peut citer, entre autres, la condition $n \geq 30$ **et** $p \leq 0{,}1$ **et** $n.p \prec 15$.

- La **loi de POISSON** de paramètre a *converge*, pour n assez grand, vers la **loi normale** $N(m = a, \sigma^2 = a)$.

Au plan pratique, la convergence en question devient satisfaisante dès que $a \succ 15$.

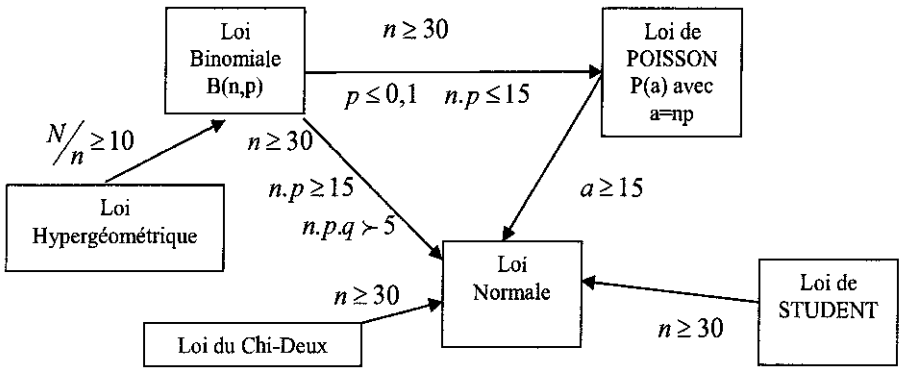
- La **loi de STUDENT**, $T(n)$, *converge*, pour n assez grand, vers la **loi normale** centrée réduite $N(0,1)$.

Au plan pratique, cette approximation devient satisfaisante dès que $n \geq 30$.

- La **loi dn chi-deux**, $\chi^2(n)$, *converge*, pour n assez grand, vers la **loi normale** $N(m = n, \sigma^2 = 2n)$.

Ici encore, cette approximation est vérifiée à partir de $n = 30$.

Le schéma ci-dessous résume les propriétés de convergence susmentionnées :



2. Statistiques et distributions d'échantillonnage

2.1 Le principe de l'inférence statistique

L'objectif est d'évaluer la valeur inconnue d'un paramètre caractéristique déterminé au sein d'une *population*, à travers le *prélèvement d'un échantillon* et une expression du paramètre en question en fonction des observations faites (**principe de l'inférence statistique**). Il faut distinguer dans ce processus :

a) les *techniques de prélèvement de l'échantillon* $(X_1, X_2, ..., X_n)$ dont la forme la plus simple est celle d'un tirage aléatoire avec remise (échantillons dits de « BERNOULLI » non exhaustifs) ;

b) les *données* $(x_1, x_2, ..., x_n)$ fournies par un échantillon particulier et *l'estimation* qui en résulte pour le paramètre θ inconnu, soit $\hat{\theta} = T_n(x_1, x_2, ..., x_n)$;

c) l'étude des variations aléatoires de l'estimation $T_n(x_1, x_2, ..., x_n)$ en fonction des divers échantillons $(x_1, x_2, ..., x_n)$ que l'on peut extraire de la population, c'est-à-dire la caractérisation de la loi de la *statistique* $T_n(X_1, X_2, ..., X_n)$ (dite encore « *estimateur* »), loi formant la *distribution d'échantillonnage*.

Il est précisé qu'on appelle « statistique » toute fonction des observations faites.

2.2 Le cas d'une moyenne

• Considérant une variable aléatoire X (de moyenne m inconnue et de variance σ^2 connue ou non) et un échantillon $(X_1, X_2, ..., X_n)$ de n valeurs indépendantes prises par X (échantillons de type « BERNOULLI »), la transposition de l'expression probabiliste de $E(X)$ conduit, pour ce qui est de la moyenne, à la **statistique** $\overline{X} = \dfrac{\sum_{i=1}^{i=n} X_i}{n}$ dont il découle immédiatement $E(\overline{X}) = m$, $Var(\overline{X}) = \dfrac{\sigma^2}{n}$.

La linéarité de l'espérance mathématique entraîne immédiatement $E(\overline{X}) = \dfrac{\sum_{i=1}^{i=n} E(X_i)}{n}$, soit $E(\overline{X}) = \dfrac{n.m}{n} = m$. Par ailleurs, l'indépendance des X_i entraîne $Var(\overline{X}) = \dfrac{\sum_{i=1}^{i=n} Var(X_i)}{n^2}$, étant entendu, par ailleurs que $Var(a.X) = a^2.Var(X)$. Finalement, on obtient bien $Var(\overline{X}) = \dfrac{n.\sigma^2}{n^2} = \dfrac{\sigma^2}{n}$.

• Pour ce qui est de la **distribntion d'échantillonnage**, le *théorème central limite* entraîne, pour $n \geq 30$, la convergence de $\overline{X}$ vers la **loi normale** $N(m, \dfrac{\sigma}{\sqrt{n}})$. En d'autres termes, la variable $\dfrac{\overline{X} - m}{\dfrac{\sigma}{\sqrt{n}}}$ suit la loi normale centrée réduite $N(0,1)$.

Plus encore, désignant par $\widehat{S^2}$ l'estimateur ponctuel de la variance σ^2 lorsque cette dernière est inconnue ($\widehat{S^2} = \frac{1}{n-1}.\sum_{i=1}^{i=n}(X_i - \overline{X})^2$ suivant résultats présentés dans le chapitre II), la variable $\frac{\overline{X} - m}{\frac{\hat{S}}{\sqrt{n}}}$ suit la **loi de STUDENT**, $T(n-1)$, à $\nu = n-1$ degrés de libertés.

La démonstration de ce résultat est présentée dans l'application 1.1 du présent chapitre.

2.3 Le cas d'une proportion

• Considérant la fréquence inconnue p d'un caractère C dans une population et la variable aléatoire X qui décrit l'occurrence de C dans des échantillons de taille n aléatoires, indépendants (prélèvements avec remise), la transposition de l'expression probabiliste de $E(X)$ conduit, pour ce qui est de la fréquence inconnue p, à la **statistique** $F_n = \frac{X}{n}$ dont il est évident que $E(F_n) = p$ et $Var(F_n) = \frac{p.q}{n}$.

En effet, X suit la loi binomiale $B(n, p)$ de moyenne $n.p$ et de variance $n.p.q$. La linéarité de l'espérance entraîne $E(F_n) = \frac{E(X)}{n} = \frac{n.p}{n} = p$. Par ailleurs, $Var(F_n) = \frac{1}{n^2}.Var(X)$, soit $Var(F_n) = \frac{p.q}{n}$.

On remarquera que p représente aussi l'espérance de la loi de BERNOULLI associée à chaque élément prélevé de l'échantillon. Dès lors et par application des résultats du paragraphe 4 susmentionné pour ce qui concerne les moyennes, la statistique représentative de p est fournie par $\overline{X} = \frac{\sum_{i=1}^{i=n} X_i}{n}$ où les X_i forment une suite de n variables aléatoires de BERNOULLI indépendantes.

La somme $\sum_{i=1}^{i=n} X_i$ constituant la variable X de loi binomiale $B(n, p)$, on on retrouve ainsi l'expression $\frac{X}{n}$ qui caractérise F_n.

Cette analogie d'une proportion avec une moyenne sera couramment utilisée par la suite.

• Pour ce qui est de la **distribution d'échantillonnage**, le *théorème de* MOIVRE LAPLACE justifie, pour $n \geq 30$ et p ni trop faible, ni trop voisin de 1 (critères pratiques rappelés précédemment), la possibilité d'approcher la loi de F_n par la **loi normale** de moyenne p et de variance $\frac{p.q}{n}$, soit la loi $N(p, \sqrt{\frac{p.q}{n}})$.

Il est précisé que dans l'hypothèse contraire où p est faible, voire n petit, on pourra mener des calculs directs à partir des lois binomiales et de POISSON et déterminer ainsi la distribution d'échantillonnage de $F_n = \frac{X}{n}$.

2.4 Le cas d'une variance

• Soient X une variable aléatoire (de moyenne m connue ou non et de variance σ^2 inconnue) et $(X_1, X_2, ..., X_n)$ un échantillon de n valeurs indépendantes prises par X (échantillon de type « Bernoullien »). La transposition de l'expression probabiliste de Var(X) conduit, pour ce qui est de la variance, à la **statistique** $S^2 = \frac{1}{n}.\sum_{i=1}^{i=n}(X_i - m)^2$ (resp. la statistique $S'^2 = \frac{1}{n}.\sum_{i=1}^{i=n}(X_i - \overline{X})^2$ lorsque m est inconnue).

Le lecteur se méfiera néanmoins que, dans l'hypothèse où m est inconnue, c'est l'estimateur « non biaisé », $\widehat{S^2} = \frac{1}{n-1}.\sum_{i=1}^{i=n}(X_i - \overline{X})^2$ qu'il faudra retenir (et non S'^2) -> se reporter pour cela au chapitre II.

On montre dans l'application 1.2 proposée ci-après, que $E(S^2) = \sigma^2$ (resp. $E(\widehat{S^2}) = \frac{n-1}{n}.\sigma^2$ lorsque m est inconnue). Par ailleurs, il est montré également dans la même application que $Var(S^2) = \frac{\mu_4}{n} - \frac{\sigma^4}{n}$ et $Var(\widehat{S^2}) = \frac{\mu_4}{n} - \frac{n-3}{n.(n-1)}.\sigma^4$ (μ_4 désignant le moment d'ordre 4 de la variable centrée $X - E(X)$, soit $\mu_4 = E\left[(X - E(X))^4\right]$.

• Pour ce qui est de la **distribution d'échantillonnage,** et sous l'hypothèse de la *normalité* de la loi de X (échantillons dits « gaussiens »), la variable $\frac{n.S^2}{\sigma^2} = \frac{\sum_{i=1}^{i=n}(X_i - m)^2}{\sigma^2}$ suit la **loi du chi- deux** à n degrés de libertés, soit $\chi^2(n)$.

De même, la variable $\frac{(n-1).\widehat{S^2}}{\sigma^2} = \frac{n.S'^2}{\sigma^2} = \frac{\sum_{i=1}^{i=n}(X_i - \overline{X})^2}{\sigma^2}$ suit la **loi du chi- deux** à $n-1$ degrés de libertés, soit $\chi^2(n-1)$.

Le premier de ces résultats est immédiat puisque la loi du chi- deux, $\chi^2(n)$, caractérise la somme des carrés de n variables aléatoires, normales, centrées, réduites, indépendantes, ce qui est le cas pour les variables $\frac{X_i - m}{\sigma}$. Quant au second résultat, sa démonstration est proposée dans l'application 1.1 ci-après.

• Pour $n \geq 30$, on pourra *approcher* la **loi du chi- deux**, soit $\chi^2(n)$, par la **loi normale** de moyenne n et de variance $2n$, soit $N(n, \sqrt{2n})$, et ceci conformément au *théorème central limite.*

Cette convergence est assez simple à établir. On rappelle tout d'abord que si $(U_1, U_2, ..., U_n)$ forment une suite de n variables aléatoires indépendantes, il en est de même de la suite $(U_1^2, U_2^2, ..., U_n^2)$. En effet, partant d'un n- uplet $(U_1, U_2, ..., U_n)$ de densité de probabilité $f(u_1, u_2, ..., u_n)$, il est évident que l'indépendance des U_i entraîne, pour cette densité sur R^n une expression égale au produit $\prod_{i=1}^{i=n}\varphi(u_i)$ des densités $\varphi(u_i)$ de chacune des variables U_i.

Dès lors, le changement de variables ($Y_1 = U_1^2, Y_2 = U_2^2, ..., Y_n = U_n^2$) conduit, pour le n- uplet $(U_1^2, U_2^2, ..., U_n^2)$ à la densité de probabilité élémentaire :

$$f(\sqrt{y_1}, \sqrt{y_2}, ..., \sqrt{y_n}).|J|.dy_1.dy_2....dy_n$$

où, le jacobien J est égal au déterminant :

$$J = \begin{vmatrix} \frac{1}{2.\sqrt{y_1}} & 0 & ... & 0 \\ 0 & \frac{1}{2.\sqrt{y_2}} & ... & 0 \\ . & . & ... & . \\ 0 & 0 & ... & \frac{1}{2.\sqrt{y_n}} \end{vmatrix} = \frac{1}{2^n.\sqrt{y_1}.\sqrt{y_2}....\sqrt{y_n}}$$

Or, la décomposition de $f(y_1, y_2, ..., y_n)$ en fonction des produits des densités $\varphi(y_i)$ conduit, pour la densité du n- uplet $(Y_1, Y_2, ..., Y_n)$ au produit ci-dessous :

$$\frac{\varphi(\sqrt{y_1}).dy_1}{2.\sqrt{y_1}}.\frac{\varphi(\sqrt{y_2}).dy_2}{2.\sqrt{y_2}}....\frac{\varphi(\sqrt{y_n}).dy_n}{2.\sqrt{y_n}}$$

qui est le produit des densités de probabilités de chacune des variables $U_1^2, U_2^2, ..., U_n^2$. Ainsi l'indépendance des U_i^2 est-elle établie.

Si on considère désormais la suite des variables normales, centrées, réduites, et indépendantes, soient $U_i = \frac{X_i - m}{\sigma}$ (loi $N(0,1)$ de densité de probabilité $\varphi(u) = \frac{1}{\sqrt{2\pi}}.\exp(-\frac{u^2}{2})$), on remarque que $E(U_i^2) = Var(U_i) + \left[E(U_i)\right]^2 = 1$ (puisque $E(U_i) = 0$).

D'autre part, $Var(U_i^2) = E(U_i^4) - \left[E(U_i^2)\right]^2$ avec $E(U_i^4) = \frac{1}{\sqrt{2\pi}}.\int_{-\infty}^{+\infty} t^4.\exp(-\frac{t^2}{2}).dt$,

soit $E(U_i^4) = \left[-\frac{1}{\sqrt{2\pi}}.t^3.\exp(-\frac{t^2}{2})\right]_{-\infty}^{+\infty} + 3.\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} t^2.\exp(-\frac{t^2}{2}).dt$ (suivant intégration par parties). Suite à la nullité du premier des deux termes ci-dessus, il reste $E(U_i^4) = 3.\frac{1}{\sqrt{2\pi}}.\int_{-\infty}^{+\infty} t^2.\exp(-\frac{t^2}{2}).dt = 3.E(U_i^2) = 3.$

Ainsi obtient-on, le résultat, $Var(U_i^2) = 3 - 1^2 = 2$.

En résumé, le théorème central limite appliqué aux n variables aléatoires indépendantes U_i^2 de moyenne égale à 1 et de variance égale à 2, entraîne la convergence de la somme $\sum_{i=1}^{i=n} U_i^2$ vers la loi normale de moyenne n et de variance $2n$, ce qui forme le résultat annoncé.

2.5 Récapitulatif concernant espérance, proportion, et variance

- Tous les résultats précédents qui, rappelons le, correspondent au cas d'un échantillonnage aléatoire élémentaire avec replacement (tirages non exhaustifs), sont résumés dans le tableau présenté ci-après.

Objet	moyenne (m)	proportion (p)	variance (σ^2)
Statistique associée	$\overline{X} = \frac{\sum_{i=1}^{i=n} X_i}{n}$	$F_n = \frac{X}{n}$ (où $X = \sum_{i=1}^{i=n} X_i$ suit la loi binomiale $B(n,p)$ et où les X_i sont des variables de BERNOULLI)	$S^2 = \frac{1}{n}.\sum_{i=1}^{i=n}(X_i - m)^2$ (lorsque m est connue) $\widehat{S^2} = \frac{1}{n-1}.\sum_{i=1}^{i=n}(X_i - \overline{X})^2$ (lorsque m est inconnue)
Paramètres représentatifs	$E(\overline{X}) = m$ $Var(\overline{X}) = \frac{\sigma^2}{n}$	$E(F_n) = p$ $Var(F_n) = \frac{p.q}{n}$	$E(S^2) = \sigma^2$ $E(\widehat{S^2}) = \sigma^2$ $Var(S^2) = \frac{\mu_4}{n} - \frac{\sigma^4}{n}$ $Var(\widehat{S^2}) = \frac{\mu_4}{n} - \frac{(n-3)}{n.(n-1)}.\sigma^4$ (où μ_4 désigne le moment d'ordre 4 de la variable centrée $X - E(X)$, soit $\mu_4 = E\left[(X - E(X))^4\right]$
Distributions d'échantillonnage	$\frac{\overline{X} - m}{\frac{\sigma}{\sqrt{n}}}$ converge vers la loi normale $N(0,1)$ (pour $n \geq 30$) Lorsqu'il s'agit d'échantillons gaussiens, $\frac{\overline{X} - m}{\frac{\sigma}{\sqrt{n}}}$ suit la loi $N(0,1)$ et par ailleurs, $\frac{\overline{X} - m}{\frac{\hat{S}}{\sqrt{n}}}$ suit la loi $T(n-1)$ de STUDENT à $\nu = n-1$ degrés de libertés.	$\frac{F_n - p}{\sqrt{\frac{p.q}{n}}}$ converge vers la loi normale $N(0,1)$ (pour $n \geq 30$) (du moins, pour p ni trop faible, ni trop voisin de 1). Sinon, on fera un calcul direct et on utilisera la loi de POISSON lorsque n est assez grand et p faible).	Supposant l'échantillon gaussien, la variable $\frac{n.S^2}{\sigma^2}$ suit la loi du chi- deux $\chi^2(n)$ à n degrés de libertés. La variable $\frac{(n-1).\widehat{S^2}}{\sigma^2}$ suit la loi du chi- deux $\chi^2(n-1)$ à $n-1$ degrés de libertés. Dans le cas d'échantillons non gaussiens, on pourra effectuer des calculs directs, voire utiliser le théorème central limite pour $n \geq 30$.

• **D'autres estimateurs** sont également rencontrés en statistique et ils donnent lieu, pour certains d'entre eux, à des développements dans ce chapitre et les suivants. On peut citer ainsi, des statistiques portant sur :

- les paramètres représentatifs d'une variable aléatoire comme *la médiane, les quartiles, les déciles, l'étendue...* ;

- *les différences de moyennes, proportions, variances, entre populations, échantillons...* ;

- *les coefficients de la droite de régression, le coefficient de corrélation...*

3. La pratique de l'échantillonnage

Les résultats de statistique mathématique sont usuellement développés dans l'hypothèse **d'échantillons élémentaires** de taille n, prélevés aléatoirement (suivant la loi uniforme) dans une population de taille N, et ceci avec remise (**tirages non exhaustifs**), ce qui assure la propriété d'indépendance entre les composantes de l'échantillon. Mais, en pratique, les méthodes de sondage font appel fréquemment à des processus plus complexes en fonction de la nature des problèmes étudiés (contrôle industriel, analyse de mesures, enquêtes sociologiques...).

• Il y a tout d'abord, et toujours dans le cadre d'un **échantillonnage aléatoire simple**, le cas de prélèvements sans remise dans une population de taille N (**tirages exhaustifs**), méthode dont il est montré dans les applications ci-après qu'elle conduit aux mêmes résultats que ceux du tirage non exhaustif pour ce qui est des statistiques $\overline{X}$ et F_n associées respectivement aux moyennes et proportions, les variances desdites statistiques étant cependant à corriger par le facteur $\frac{N-n}{N-1}$.

• Plus généralement, il faut faire une *distinction* entre les **méthodes de prélèvement empirique** et les **méthodes aléatoires** dans lesquelles les éléments sondés résultent d'un tirage aléatoire au sein de la population (base de sondage). Pour chacune de ces méthodes, des techniques plus ou moins évoluées comme les **quotas**, le **prélèvement par grappes**, la **stratification**, et les **plans à plusieurs degrés**, permettent une amélioration notable de l'efficacité. Certains de ces aspects, sont abordés dans les applications ci-après.

Plus précisément, le *choix raisonné* est le plus classique parmi les *méthodes empiriques* d'échantillonnage (c'est par exemple, sonder une personne sur dix, sonder les personnes dont les noms commencent par A...).

La *méthode des quotas* est très usitée dans le cadre des cette approche empirique. Elle consiste à partitionner la population suivant un certain nombre de critères (sexe, classes d'âge, catégories professionnelles...), l'échantillon étant construit au prorata des effectifs suivant un taux réducteur dit « *taux de sondage* ». Les quotas sont imposés aux enquêteurs et le choix des éléments qui composent l'échantillon est laissé à leur initiative.

Pour ce qui est des *sondages aléatoires*, le tirage avec remise (indépendance des éléments de l'échantillon), voire sans remise, constitue le procédé le plus élémentaire utilisé, l'usage d'une simulation de la loi uniforme pouvant faciliter le choix des éléments de l'échantillon.

Plusieurs techniques permettent d'alléger la base de sondage, c'est à dire d'éviter de travailler sur la population de référence dans sa totalité. Les *sondages à plusieurs degrés* sont les plus courants ici. Par exemple, travaillant sur les médecins, on pourra en premier lieu, tirer au sort un certain nombre de villes (premier degré de sondage) puis au sein de chaque ville, dresser la liste des médecins en activité et former un échantillon à partir de ceux-ci (deuxième degré de sondage).

Citons également une méthode proche qui est le *prélèvement par grappes* (par exemple, un ménage est constitué d'une grappe de personnes, une ville forme une grappe de ménages…).

Enfin, la *stratification* (qui est très proche de la méthode des quotas) permet une amélioration notable de la précision des estimations effectuées à partir de l'échantillon, son principe étant de s'assurer que ce dernier est bien représentatif des diverses configurations rencontrées au sein de la population. L'idée est de découper cette dernière en groupes homogènes (strates) par rapport à un critère donné (exercice reconductible à des niveaux successifs de plus en plus précis), et de constituer un échantillon par prélèvement dans chacune des strates (au prorata des effectifs ou suivant d'autres méthodes -> cf. application 2.2 ci- après).

B - Applications

1. Distributions d'échantillonnage et propriétés

Dans cette partie, il est proposé d'étudier les distributions d'échantillonnage les plus courantes et de justifier certaines de leurs propriétés.

1.1 Moyenne et variance dans le cas d'échantillons gaussiens (théorème de FISHER)

Enoncé : Considérant un échantillon $(X_1, X_2, ..., X_n)$ de taille n pour une variable aléatoire X de loi normale $N(m,\sigma)$ il est proposé de montrer que $\dfrac{\overline{X}-m}{\dfrac{\hat{S}}{\sqrt{n}}}$ suit la loi de STUDENT à $\nu = n-1$ degrés de libertés et que la statistique $\dfrac{(n-1).\widehat{S^2}}{\sigma^2}$ suit la loi du chi- deux à $\nu = n-1$ degrés de libertés (loi $\chi^2(n-1)$).

PARTIE I

1°) V étant une variable aléatoire suivant la loi du chi- deux à n degrés de libertés (loi notée $\chi^2(n)$), exprimer la fonction caractéristique $\Phi_V(t)$ de la variable V.

2°) V_1 et V_2 étant deux variables de chi- deux supposées indépendantes et respectivement à n_1 et n_2 degrés de libertés, exprimer la fonction caractéristique de la somme $V_1 + V_2$, soit $\Phi_{V_1+V_2}(t)$, et en déduire la loi suivie par cette somme.

3°) Déduire du résultat précédent que si V_1 et V_2 sont deux variables aléatoires indépendantes, si V_1 suit la loi du chi- deux à n_1 degrés de libertés, soit $\chi^2(n_1)$, et qu'enfin si $V = V_1 + V_2$ suit la loi $\chi^2(n)$ (avec $n \succ n_1$), alors la variable aléatoire V_2 suit la loi du chi- deux à $\nu_2 = n - n_1$ degrés de libertés.

4°) Montrer que si X suit la loi normale, centrée, réduite, soit $N(0,1)$, la variable aléatoire $Y = X^2$ suit la loi du chi- deux à un degré de liberté.

PARTIE II

$(X_1, X_2, ..., X_n)$ étant l'échantillon susmentionné en introduction, on admettra le résultat suivant lequel l'estimateur de la variance σ^2 est fourni par $\widehat{S^2} = \dfrac{1}{(n-1)} \cdot \sum_{i=1}^{i=n} (X_i - \overline{X})^2$ (cf. chapitre II).

1°) Montrer que $\frac{(n-1).\widehat{S^2}}{\sigma^2} = V - V_1$ où $V = \sum_{i=1}^{i=n} (\frac{X_i - m}{\sigma})^2$ et $V_1 = n.(\frac{\overline{X} - m}{\sigma})^2$.

2°) On rappelle que, s'agissant de variables aléatoires de loi normale, la condition nécessaire et suffisante d'indépendance se ramène à la nullité de la covariance, ce qui est loin d'être le cas de façon générale.

2-a) Montrer que $\overline{X}$ et les variables aléatoires $X_i - \overline{X}$ sont indépendantes, $\forall i / 1 \le i \le n$.

2-b) En déduire l'indépendance de $\overline{X}$ avec les variables aléatoires $(X_i - \overline{X})^2, \forall i / 1 \le i \le n$.

2-c) Etablir que $\overline{X}$ et $\frac{(n-1).\widehat{S^2}}{\sigma^2}$ sont indépendantes et en conclure l'indépendance de V_1 avec $\frac{(n-1).\widehat{S^2}}{\sigma^2}$.

3°) Reconnaissant les lois de V et de V_1, en déduire la loi de $\frac{(n-1).\widehat{S^2}}{\sigma^2}$ (distribution d'échantillonnage de la variance).

4°) Etablir que $T = \frac{\overline{X} - m}{\hat{S}/\sqrt{n}}$ suit la loi de STUDENT à $\nu = n-1$ degrés de libertés.

5°) Montrer que les statistiques $\overline{X}$ et $\frac{(n-1).\widehat{S^2}}{\sigma^2}$ sont indépendantes.

Solution : I-1°) Lorsque X suit la loi $\chi^2(n)$, sa densité de probabilité à pour expression $f(x) = \frac{x^{\frac{n}{2}-1}.e^{-\frac{x}{2}}}{2^{\frac{n}{2}}.\Gamma(\frac{n}{2})}$.

Il s'ensuit $\Phi_X(t) = E\left[e^{i.t.X}\right] = \int_0^{+\infty} \frac{x^{\frac{n}{2}-1}.e^{(i.t-\frac{1}{2}).x}}{2^{\frac{n}{2}}.\Gamma(\frac{n}{2})}.dx = \frac{1}{2^{\frac{n}{2}}.\Gamma(\frac{n}{2})}.\int_0^{+\infty} x^{\frac{n-2}{2}}.e^{-(1-2.i.t).\frac{x}{2}}.dx$.

Posant $(1-2.i.t).x = u$, il vient $x = \frac{u}{1-2.i.t}$ et $\Phi_X(t) = \frac{(1-2.i.t)^{-\frac{n}{2}}}{2^{\frac{n}{2}}.\Gamma(\frac{n}{2})}.\int_0^{+\infty} u^{\frac{n}{2}-1}.e^{-\frac{u}{2}}.du$ (après développement et simplifications). Or et par définition de la densité de probabilité $f(x)$ de la loi $\chi^2(n)$, $\int_0^{+\infty} \frac{u^{\frac{n}{2}-1}.e^{-\frac{u}{2}}}{2^{\frac{n}{2}}.\Gamma(\frac{n}{2})}.du = 1$ En conclusion, on obtient $\Phi_X(t) = (1-2.i.t)^{-\frac{n}{2}}$.

I-2°) La *fonction caractéristique de la somme* $V_1 + V_2$, soit $\Phi_{V_1+V_2}(t)$ est égale au *produit des fonctions caractéristiques* puisqu'il s'agit de deux variables aléatoires indépendantes. On a donc $\Phi_{V_1+V_2}(t) = (1-2.i.t)^{-\frac{n_1}{2}}.(1-2.i.t)^{-\frac{n_2}{2}} = (1-2.i.t)^{-(\frac{n_1+n_2}{2})}$. On reconnaît là, la fonction caractéristique de la loi $\chi^2(n_1+n_2)$.

I-3°) V et V_1 ont pour fonctions caractéristiques respectives, $\Phi_V(t)=(1-2.i.t)^{-\frac{n}{2}}$ et $\Phi_{V_1}(t)=(1-2.i.t)^{-\frac{n_1}{2}}$. Considérant la fonction caractéristique $\Phi_{V_2}(t)$ de la variable aléatoire V_2, l'indépendance de V_1 et de V_2 entraîne $\Phi_{V_1+V_2}(t)=\Phi_{V_1}(t).\Phi_{V_2}(t)$. Ainsi, obtient-on $\Phi_{V_2}(t)=\dfrac{(1-2.i.t)^{-\frac{n}{2}}}{(1-2.i.t)^{-\frac{n_1}{2}}}=(1-2.i.t)^{-\frac{(n-n_1)}{2}}$, ce qui établit le résultat cherché, à savoir $V-V_1$ suit la *loi du chi- deux* à $n-n_1$ degrés de libertés, soit $\chi^2(n-n_1)$.

I-4°) Lorsque $n=1$, la variable $\chi^2(1)$ a pour densité de probabilité $g(y)=\dfrac{e^{-\frac{y}{2}}}{\Gamma(\frac{1}{2}).\sqrt{2y}}$.

On notera que la fonction $\Gamma(\frac{1}{2})=\int_0^{+\infty} t^{\frac{1}{2}-1}.e^{-t}.dt$ est égale, suivant le changement de variable $t=u^2$, à l'intégrale $\int_0^{+\infty}\dfrac{e^{-u^2}}{u}.2u.du$, soit $2.\int_0^{+\infty} e^{-u^2}.du=\sqrt{\pi}$. Ainsi, $\chi^2(1)$ a-t-elle pour densité de probabilité $g(y)=\dfrac{e^{-\frac{y}{2}}}{\sqrt{2.\pi.y}}$.

Ceci dit, considérant la variable aléatoire X, *normale, centrée, réduite* de loi $N(0,1)$, c'est à dire de densité de probabilité $f(x)=\dfrac{e^{-\frac{x^2}{2}}}{\sqrt{2.\pi}}$, le changement de variable $Y=X^2$ conduit au calcul ci-dessous, le caractère non injectif de la transformation appelant ici des précautions puisque $\Pr ob(0\le Y\le y)=\Pr ob(-\sqrt{y}\le X\le\sqrt{y})$ et non pas $\Pr ob(0\le X\le\sqrt{y})$!!.

Dès lors en recourant à la fonction de répartition, $F(x)=\Pr ob(X\le x)$, on obtient l'expression $\Pr ob(0\le Y\le y)=F(\sqrt{y})-F(-\sqrt{y})$. Désignant par $g(y)$ la densité de probabilité de Y, il en résulte $g(y)=\dfrac{d\Pr ob(0\le Y\le y)}{dy}=\dfrac{F'(\sqrt{y})}{2.\sqrt{y}}-(-\dfrac{F'(\sqrt{y})}{2.\sqrt{y}})$.

Mais, pour tout $u, F'(u)=f(u)$, avec pour le cas présent, $f(u)=\dfrac{e^{-\frac{u^2}{2}}}{\sqrt{2.\pi}}$ (puisqu'il s'agit de la loi $N(0,1)$). Finalement, $g(y)=\dfrac{f(\sqrt{y})}{2.\sqrt{y}}+\dfrac{f(\sqrt{y})}{2.\sqrt{y}}=\dfrac{e^{-\frac{y}{2}}}{\sqrt{2.\pi.y}}$. Il s'agit bien de *la loi du χ^2 à un degré de liberté.*

II-1°) On peut écrire $(n-1).\widehat{S^2}=\sum_{i=1}^{i=n}(X_i-\overline{X})^2=\sum_{i=1}^{i=n}\left[(X_i-m)-(\overline{X}-m)\right]^2$, ce qui conduit au développement $(n-1).\widehat{S^2}=\sum_{i=1}^{i=n}(X_i-m)^2+\sum_{i=1}^{i=n}(\overline{X}-m)^2-2.\sum_{i=1}^{i=n}(X_i-m).(\overline{X}-m)$.

Or, d'une part $\sum_{i=1}^{i=n}(\overline{X}-m)^2 = n.(\overline{X}-m)^2$, et d'autre part, $\sum_{i=1}^{i=n}(\overline{X}-m).(X_i-m)$ est égal à $(\overline{X}-m).(\sum_{i=1}^{i=n}X_i - n.m) = n.(\overline{X}-m)^2$ puisque $\sum_{i=1}^{i=n}X_i = n.\overline{X}$. En définitive, on obtient l'expression $(n-1).\frac{\widehat{S^2}}{\sigma^2} = \sum_{i=1}^{i=n}(\frac{X_i-m}{\sigma})^2 - n.\frac{(\overline{X}-m)^2}{\sigma^2}$ qui est le résultat annoncé.

II-2°-a) S'agissant de *variables aléatoires normales*, *l'indépendance* de $\overline{X}$ avec les variables $X_i-\overline{X}$ revient à montrer la *nullité de la covariance* (puisque cela devient alors une *condition nécessaire et suffisante*).

Cette *covariance* est égale à $E\left[\overline{X}.(X_i-\overline{X})\right]-E(\overline{X}).E(X_i-\overline{X})$, expression dans laquelle le second terme est nul puisque, par linéarité, $E(X_i-\overline{X}) = E(X_i)-E(\overline{X}) = m-m = 0$. Quant au premier terme, son développement conduit à $E\left[\overline{X}.(X_i-\overline{X})\right] = E\left[\sum_{k=1}^{k=n}\frac{X_k.X_i}{n} - \overline{X}^2\right] = \frac{1}{n}.\sum_{k=1}^{k=n}E(X_k.X_i) - E(\overline{X}^2)$, compte tenu de la linéarité de l'espérance mathématique.

Mais, $E(\overline{X}^2) = Var(\overline{X}) + E(\overline{X})^2 = \frac{\sigma^2}{n} + m^2$. Par ailleurs, l'indépendance deux à deux des variables X_i et X_j entraîne $\text{cov}(X_i,X_j) = E(X_i.X_j) - E(X_i).E(X_j) = 0$, ou encore $E(X_i.X_j) = E(X_i).E(X_j) = m^2$. Enfin, $E(X_i^2) = Var(X_i) + E(X_i)^2 = \sigma^2 + m^2$.

Il s'ensuit $\frac{1}{n}.\sum_{k=1}^{k=n}E(X_i.X_k) = \frac{1}{n}.E(X_i^2) + \frac{1}{n}.\sum_{\substack{k=1\\k\neq i}}^{k=n}E(X_i.X_k) = \frac{1}{n}.\left[(\sigma^2+m^2)+(n-1).m^2\right]$,

ceci compte tenu des expressions obtenues ci-dessus. En conséquence, $E\left[\overline{X}.(X_i-\overline{X})\right]$ est égale à $\frac{1}{n}.(\sigma^2+n.m^2) - \frac{\sigma^2}{n} - m^2 = 0$. Ceci établit la nullité de la covariance entre $\overline{X}$ et $X_i-\overline{X}$, résultat vérifié pour tout $i / 1 \le i \le n$.

Se référant à *l'approche géométrique des variables aléatoires* et du *problème de la corrélation linéaire*, on remarquera que *l'indépendance* ci-dessus, se traduit au regard du produit scalaire classique $\prec X,Y \succ = E(X.Y)$, par *l'orthogonalité* de $\overline{X}$ avec $X_i-\overline{X}$, $(\overline{X} \perp X_i-\overline{X})$, et même par l'orthogonalité de $\overline{X}$ avec le sous-espace vectoriel engendré par les $X_i-\overline{X}$, soit $H = Vect(X_1-\overline{X}, X_2-\overline{X}, ..., X_n-\overline{X})$.

II-2°-b) Dans la mesure où l'indépendance de X et de Y entraîne pour toutes fonctions φ et ψ et ceci $\forall(X,Y)$, l'indépendance du couple $(\varphi(X),\psi(Y))$ (résultat justifié dans les rappels de cours précédents -> cf. paragraphe 6), il en résulte l'indépendance (ou encore, l'orthogonalité) de $\overline{X}$ avec chacune des variables $(X_i-\overline{X})^2$ et à fortiori avec le sous-espace vectoriel qu'elles engendrent, soit :

$$H = Vect((X_1-\overline{X})^2, (X_2-\overline{X})^2, ..., (X_n-\overline{X})^2) = \left\{Z / Z = \sum_{i=1}^{i=n}\alpha_i.(X_i-\overline{X})^2\right\}.$$

Ce résultat est vérifié entre autres, pour la somme $\sum_{i=1}^{i=n}(X_i-\overline{X})^2 \in H$ et même pour la pour la variable aléatoire $\frac{(n-1).\widehat{S^2}}{\sigma^2}$.

II-2°-c) De l'indépendance de $\overline{X}$ avec $\frac{(n-1).\widehat{S^2}}{\sigma^2}$ et toujours pour les raisons développées en 2-b) quant au lien entre l'indépendance de X et de Y et celle de $\varphi(X)$ et $\psi(Y)$, découle l'indépendance de $V_1 = n.(\frac{\overline{X}-m}{\sigma})^2$ avec $\frac{(n-1).\widehat{S^2}}{\sigma^2}$.

II-3°) $V=\sum_{i=1}^{i=n}(\frac{X_i-m}{\sigma})^2$ suit, en sa qualité de somme de n variables aléatoires, indépendantes, de loi normale, centrée, réduite, $N(0,1)$, la *loi du chi- deux à n degrés de libertés*, soit $\chi^2(n)$. Ce résultat mentionné dans les rappels de cours est une conséquence immédiate des propriétés établies dans la partie I, puisque chaque variable $(\frac{X_i-m}{\sigma})^2$ est de type $\chi^2(1)$ et que la somme deux variables $\chi^2(1)$ suit une loi $\chi^2(2)$ et ainsi de suite...

- D'autre part, $V_1 = U^2$, où $U=\frac{\overline{X}-m}{\sigma/\sqrt{n}}$ suit la loi normale $N(0,1)$ puisque $E(\overline{X})=m$ et que $Var(\overline{X})=\frac{\sigma}{\sqrt{n}}$. La loi de V_1 est donc, d'après la question I-4°), la loi du chi- deux à un degré de liberté, soit $\chi^2(1)$.
- Par ailleurs, on a montré dans les questions II-2°) précédentes que les variables V_1 et $\frac{(n-1).\widehat{S^2}}{\sigma^2}$ étaient indépendantes.
- Enfin, il a été établi dans la question II-1°) que $\frac{(n-1).\widehat{S^2}}{\sigma^2}=V-V_1$.

De tout cela, et compte tenu du résultat établi à la question I-3°), on en conclut que $\frac{(n-1).\widehat{S^2}}{\sigma^2}$ suit la *loi du chi- deux à $n-1$ degrés de libertés*, soit $\chi^2(n-1)$.

II-4°) Il est rappelé que si X et Y suivent respectivement la *loi normale centrée réduite*, $N(0,1)$, et la *loi du chi- deux à n degrés de libertés*, $\chi^2(n)$, la variable $T=\frac{X}{\sqrt{Y/n}}$ suit la *loi de STUDENT à n degrés de libertés*, soit $T(n)$ (résultat mentionné dans les rappels de cours « paragraphe 1 » et dont la démonstration est faisable à partir des techniques usuelles de calcul des probabilités). Or, $\frac{\overline{X}-m}{\hat{S}/\sqrt{n}}=\frac{\frac{\overline{X}-m}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1).\widehat{S^2}/\sigma^2}{(n-1)}}}$.

Par identification de l'expression précédente avec T, on reconnaît, au numérateur, la variable normale centrée réduite (de loi $N(0,1)$), soit X, et au dénominateur, la variable $\sqrt{Y/n-1}$ où, Y, qui est égale à $(n-1).\widehat{S^2}/\sigma^2$ suit la loi du chi- deux à $\nu = n-1$ degrés de libertés, soit $\chi^2(n-1)$, d'après la question précédente II-3°).

Dans ces conditions, $T = \dfrac{X}{\sqrt{Y/n-1}}$ suit la *loi de STUDENT à* $\nu = n-1$ *degrés de libertés* compte tenu du rappel précédent.

Les résultats qui viennent d'être obtenus ici relativement aux lois de $\dfrac{\overline{X}-m}{\sigma/\sqrt{n}}, \dfrac{\overline{X}-m}{\hat{S}/\sqrt{n}}$, et $\widehat{S^2}$, pour des échantillons gaussiens, constituent le **théorème de FISHER**.

1.2 Paramètres représentatifs des statistiques décrivant la variance

Enoncé : On considère un échantillon de taille n, soit $(X_1, X_2, ..., X_n)$, de n valeurs indépendantes d'une variable aléatoire X de moyenne m et de variance σ^2. Soient $S^2 = \frac{1}{n}.\sum_{i=1}^{i=n}(X_i - m)^2$ et $\hat{S}^2 = \frac{1}{n-1}.\sum_{i=1}^{i=n}(X_i - \overline{X})^2$ les statistiques associées à la variance (respectivement lorsque m est connue ou non).

1°) Calculer $E(S^2)$ et $E(\hat{S}^2)$.

2-a) Calculer $Var(S^2)$.

2-b) Calculer $Var(\hat{S}^2)$.

2-c) Etudier le cas particulier où $(X_1, X_2, ..., X_n)$ est un échantillon gaussien de n valeurs indépendantes d'une variable aléatoire X de loi $N(m, \sigma)$.

Solution : 1°) La *linéarité de l'espérance mathématique* permet d'écrire immédiatement, le développement $E(S^2) = \frac{1}{n}.E\left[\sum_{i=1}^{i=n}(X_i - m)^2\right] = \frac{1}{n}.\sum_{i=1}^{i=n}E\left[(X_i - m)^2\right] = \frac{n.\sigma^2}{n} = \sigma^2$.

D'autre part, $(n-1).E(\hat{S}^2) = E\left[\sum_{i=1}^{i=n}(X_i - \overline{X})^2\right] = E\left[\sum_{i=1}^{i=n}(X_i^2 - 2.X_i.\overline{X} + \overline{X}^2)\right]$, soit par décomposition, $E\left[\sum_{i=1}^{i=n}X_i^2 - 2.\overline{X}.\sum_{i=1}^{i=n}X_i + \sum_{i=1}^{i=n}\overline{X}^2\right]$. Mais, $\sum_{i=1}^{i=n}X_i = n.\overline{X}$ et $\sum_{i=1}^{i=n}\overline{X}^2 = n.\overline{X}^2$.

Ainsi, $(n-1).E(\hat{S}^2) = E\left[\sum_{i=1}^{i=n}X_i^2 - 2.n.\overline{X}^2 + n.\overline{X}^2\right]$, soit par *linéarité de l'espérance mathématique*, $(n-1).E(\hat{S}^2) = \sum_{i=1}^{i=n}E(X_i^2) - n.E(\overline{X}^2)$.

Or, $Var(X_i) = E(X_i^2) - E(X_i)^2 \Rightarrow E(X_i^2) = \sigma^2 + m^2$ et $Var(\overline{X}) = E(\overline{X}^2) - E(\overline{X})^2 \Rightarrow E(\overline{X}^2) = \frac{\sigma^2}{n} + m^2$. Il en résulte $(n-1).E(\hat{S}^2) = n.(\sigma^2 + m^2) - n.(\frac{\sigma^2}{n} + m^2) = (n-1).\sigma^2$.

Finalement $E(\hat{S}^2)=\sigma^2$. Lorsque la moyenne m est inconnue, on remarquera ici que la statistique $S'^2=\frac{1}{n}.\sum_{i=1}^{i=n}(X_i-\overline{X})^2$ conduirait à l'espérance $E(S'^2)=\frac{n-1}{n}.\sigma^2$. Ainsi la moyenne des estimateurs de σ^2 faite à partir d'une suite d'échantillons de taille n et de la statistique $S'^2=\frac{1}{n}.\sum_{i=1}^{i=n}(X_i-\overline{X})^2$ conduit-elle à une valeur moyenne limite « **biaisée** » par rapport à la valeur inconnue σ^2 à estimer. D'où, le recours à la statistique $\hat{S}^2=\frac{n}{n-1}.S'^2$ pour compenser cette « *erreur de parallaxe* » (se reporter au chapitre II).

2-a) $Var(S^2)=\frac{1}{n^2}.Var(\sum_{i=1}^{i=n}(X_i-m)^2)$. Or l'indépendance deux à deux des variables aléatoires X_i, voire X_i-m, entraîne l'indépendance deux à deux, des variables $(X_i-m)^2$ (cf. calcul mené en rappel de cours du présent chapitre, paragraphe 6).

Ainsi $Var(\sum_{i=1}^{i=n}(X_i-m)^2)=\sum_{i=1}^{i=n}Var\left[(X_i-m)^2\right]=E\left[(X_i-m)^4\right]-\left[E\left[(X_i-m)^2\right]\right]^2$, soit $Var(\sum_{i=1}^{i=n}(X_i-m)^2)=\mu_4-\sigma^4$ où μ_4 désigne le *moment d'ordre 4 des variables centrées* X_i-m. En conclusion, $Var(S^2)=\frac{\mu_4}{n}-\frac{\sigma^4}{n}$.

2-b) $Var(\hat{S}^2)=\frac{1}{(n-1)^2}.Var\left[\sum_{i=1}^{i=n}(X_i-\overline{X})^2\right]$. Pour simplifier les calculs qui, au demeurant, restent lourds, il est proposé de travailler sur les *variables centrées* $U_i=X_i-m$ dont l'indépendance est induite par celle des X_i.

- On a $\overline{X}=\sum_{i=1}^{i=n}\frac{(m+U_i)}{n}=m+\overline{U}$, avec $\overline{U}=\frac{\sum_{i=1}^{i=n}U_i}{n}$. Ainsi $X_i-\overline{X}=(U_i+m-\overline{U}-m)$, soit $X_i-\overline{X}=U_i-\overline{U}$ et à fortiori :

$$Var(\hat{S}^2)=\frac{1}{(n-1)^2}.Var\left[\sum_{i=1}^{i=n}(X_i-\overline{X})^2\right]=\frac{1}{(n-1)^2}.Var\left[\sum_{i=1}^{i=n}(U_i-\overline{U})^2\right].$$

Par contre, il est facile de vérifier que les variables $X_i-\overline{X}$ et à fortiori, les variables $U_i-\overline{U}$ ne sont plus indépendantes deux à deux, malgré l'indépendance des X_i (resp. des U_i).

- On peut écrire $\hat{S}^2=\frac{1}{n-1}.(\sum_{i=1}^{i=n}U_i^2-n.(\frac{U_1+U_2+....+U_n}{n})^2)$. Par ailleurs, on a par définition, $Var(\hat{S}^2)=E\left[(\hat{S}^2)^2\right]-\left[E(\hat{S}^2)\right]^2$, le dernier de ces deux termes étant égal à σ^4 d'après le résultat de la 1ère question. Avant de passer au développement du carré de $\hat{S}^2$, on peut écrire $\hat{S}^2=\frac{1}{n-1}.\left[U_1^2+U_2^2+....+U_n^2-\frac{n}{n^2}.(U_1^2+U_2^2+....+U_n^2+\sum_{\substack{i=1\\i\neq j}}^{i=n}\sum_{j=1}^{j=n}U_i.U_j)\right]$.

Ainsi, $\hat{S}^2 = \frac{1}{n-1}.\left[\sum_{i=1}^{i=n}(1-\frac{1}{n}).U_i^2 - \frac{1}{n}.\sum_{\substack{i=1\\ i\neq j}}^{i=n}\sum_{j=1}^{j=n}U_i.U_j\right] = \frac{1}{n}.\sum_{i=1}^{i=n}U_i^2 - \frac{1}{n.(n-1)}.\sum_{\substack{i=1\\ i\neq j}}^{i=n}\sum_{j=1}^{j=n}U_i.U_j$.

Dans ces conditions, $E\left[(\hat{S}^2)^2\right] = E\left[(\frac{1}{n}.\sum_{i=1}^{i=n}U_i^2 - \frac{1}{n.(n-1)}.\sum_{\substack{i=1\\ i\neq j}}^{i=n}\sum_{j=1}^{j=n}U_i.U_j)^2\right]$. Décomposant cette différence en posant $A = \frac{1}{n}.\sum_{i=1}^{i=n}U_i^2$ et $B = \frac{1}{n.(n-1)}.\sum_{\substack{i=1\\ i\neq j}}^{i=n}\sum_{j=1}^{j=n}U_i.U_j$, on obtient successivement $A^2 = \frac{1}{n^2}.\sum_{i=1}^{i=n}U_i^4 + \frac{1}{n^2}.\sum_{\substack{i=1\\ i\neq j}}^{i=n}\sum_{j=1}^{j=n}U_i^2.U_j^2$, $A.B = \frac{1}{n^2.(n-1)}.\sum_{\substack{i=1\\ i\neq j}}^{i=n}\sum_{j=1}^{j=n}\sum_{k=1}^{k=n}U_i.U_j.U_k^2$, et enfin $B^2 = \frac{1}{n^2.(n-1)^2}.\left[2.\sum_{\substack{i=1\\ i\neq j}}^{i=n}\sum_{j=1}^{j=n}U_i^2.U_j^2 + \sum_{\substack{i=1\\ i\neq j\neq k}}^{i=n}\sum_{j=1}^{j=n}\sum_{k=1}^{k=n}U_i^2.U_j.U_k\right]$. S'agissant de cette dernière expression de B^2 et plus particulièrement du coefficient 2 portant sur le premier de ses deux termes, il suffit de faire un développement pour le cas particulier $n=3$, par exemple, pour s'apercevoir que les doubles produits $(U_i.U_j).(U_j.U_i)$ génèrent une somme supplémentaire de $U_i^2.U_j^2$, ce qui motive le facteur multiplicatif en question.

- L'indépendance deux à deux des U_i et des U_j entraîne $E(U_i.U_j) = E(U_i).E(U_j) = 0$ puisque, par ailleurs, les variables U_i sont centrées. Plus largement, les variables $\varphi(U_i)$ et $\psi(U_j)$ sont indépendantes pour toutes fonctions φ et ψ, ce qui dans le cas de la fonction $z \to z^2$, implique l'indépendance deux à deux des U_i^2 et des U_j^2. On a donc, par nullité de la covariance, $E(U_i^2.U_j^2) = E(U_i^2).E(U_j^2) = \sigma^4$ (puisque $E(U_i^2) = Var(U_i) + \left[E(U_i)\right]^2$, avec $E(U_i) = 0$). De même, $E(U_i^2.U_j.U_k) = E(U_i^2).E(U_j).E(U_k) = 0$ (toujours en raison du caractère centré des variables U_i). Enfin, $E(U_i^3.U_j) = E(U_i^3).E(U_j) = 0$.

- En définitive, $E\left[(\hat{S}^2)^2\right] = E(A^2 - 2.A.B + B^2)$, c'est-à-dire :

$E\left[(\hat{S}^2)^2\right] = \frac{1}{n^2}.\sum_{i=1}^{i=n}E(U_i^4) + \frac{1}{n^2}.\sum_{\substack{i=1\\ i\neq j}}^{j=n}E(U_i^2.U_j^2) + \frac{2}{n^2.(n-1)^2}.\sum_{\substack{i=1\\ i\neq j}}^{j=n}E(U_i^2.U_j^2)$, tous les autres termes étant nuls pour les raisons précisées ci-dessus.

- Finalement l'expression de $Var(\hat{S}^2)$, qui est égale à $E\left[(\hat{S}^2)^2\right] - \left[E(\hat{S}^2)\right]^2$, s'écrit compte tenu de tous les résultats ci-dessus :

$$Var(\hat{S}^2) = \frac{n.\mu_4}{n^2} + \frac{n.(n-1)}{n^2}.\sigma^4 + \frac{2.n.(n-1)}{n^2.(n-1)^2}.\sigma^4 - \sigma^4 = \frac{\mu_4}{n} - \frac{(n-3)}{n.(n-1)}.\sigma^4$$

(μ_4 désignant, pour rappel, le *moment d'ordre 4 de la variable centrée* U_i, soit $\mu_4 = E(U_i^4) = E\left[(X_i - m)^4\right]$.

2-c) Dans le cas *d'échantillons gaussiens*, $\mu_4 = E\left[(X_i - m)^4\right] = \sigma^4 . E\left[(\frac{X_i - m}{\sigma})^4\right]$, où la variable aléatoire $\xi = \frac{X_i - m}{\sigma}$ désigne la *variable normale, centrée, réduite*, de loi $N(0,1)$. Ainsi, $\mu_4 = \sigma^4 . \int_{-\infty}^{+\infty} \frac{t^4 . \exp(-\frac{t^2}{2})}{\sqrt{2\pi}} . dt$. Posant $U = t^3, dV = t.\exp(-\frac{t^2}{2})$, soit $dU = 3.t^2.dt$ et $V = -\exp(-\frac{t^2}{2})$, il vient immédiatement, suivant intégration par parties :

$$\mu_4 = \frac{1}{\sqrt{2\pi}} \left[-t^3 . \exp(-\frac{t^2}{2}) \right]_{-\infty}^{+\infty} + \frac{3.\sigma^4}{\sqrt{2\pi}} . \int_{-\infty}^{+\infty} t^2 . \exp(-\frac{t^2}{2}) . dt$$

Le premier de ces deux termes est nul. Quant au second, on reconnaît, dans l'intégrale $\frac{1}{\sqrt{2\pi}} . \int_{-\infty}^{+\infty} t^2 . \exp(-\frac{t^2}{2}) . dt$, le moment d'ordre deux $E(\xi^2)$ dont il est immédiat qu'il est égal à 1, puisque $Var(\xi) = E(\xi^2) - \left[E(\xi)\right]^2 = 1$ et que $E(\xi) = 0$. En bref, $\mu_4 = 3.\sigma^4$.

Il s'ensuit, $Var(S^2) = \frac{2.\sigma^4}{n}$ et $Var(\hat{S}^2) = \frac{2.\sigma^4}{n-1}$.

1.3 Distributions d'échantillonnage des rapports de variances (loi de FISHER – SNEDECOR)

Enoncé : 1°) On considère deux variables aléatoires indépendantes, soient X et Y, suivant respectivement les lois du chi- deux à n et p degrés de libertés.

1-a) Exprimer le loi de du rapport $Z = X/Y$.

1-b) En déduire que la variable $F = \frac{X/n}{Y/p}$ suit la loi de FISHER-SNEDECOR à n et p degrés de libertés.

2°) Soient $(X_1, X_2, ..., X_n)$ et $(Y_1, Y_2, ..., Y_p)$ deux échantillons indépendants de tailles n et p extraits de deux populations normales de moyennes respectives m_X et m_y et de variances respectives σ_X^2 et σ_Y^2 (échantillons gaussiens).

2-a) Considérant les statistiques $S_X^2 = \frac{1}{n} . \sum_{i=1}^{i=n} (X_i - m_X)^2$ et $S_Y^2 = \frac{1}{p} . \sum_{i=1}^{i=p} (Y_i - m_Y)^2$, montrer que la statistique $F = \frac{S_X^2 / \sigma_X^2}{S_Y^2 / \sigma_Y^2}$ suit la loi $F(n, p)$ de FISHER- SNEDECOR à n et p degrés de libertés.

2-b) On forme les statistiques $S'^2_X = \frac{1}{n} . \sum_{i=1}^{i=n} (X_i - \overline{X})^2$ et $S'^2_Y = \frac{1}{p} . \sum_{i=1}^{i=p} (Y_i - \overline{Y})^2$.

Montrer que la statistique $F=\dfrac{n.S'^2_X/(n-1).\sigma_X^2}{p.S'^2_Y/(p-1).\sigma_Y^2}$ suit la $F(n-1,p-1)$ de FISHER-SNEDECOR à $n-1$ et $p-1$ degrés de libertés.

3°) On considère deux échantillons de tailles respectives 20 et 30 extraits de deux populations distribuées normalement avec des variances respectivement égales à 25 et 16. Quelle est la probabilité pour que la variance du premier échantillon soit supérieure au double de celle du second de ces échantillons ?

Solution : 1-a) Par définition (cf. rappels de cours), X et Y qui suivent les lois $\chi^2(n)$ et $\chi^2(p)$ ont donc pour densités de probabilité respectives $\dfrac{x^{\frac{n}{2}-1}.e^{-\frac{x}{2}}}{2^{\frac{n}{2}}.\Gamma(\frac{n}{2})}$ et $\dfrac{y^{\frac{p}{2}-1}.e^{-\frac{y}{2}}}{2^{\frac{p}{2}}.\Gamma(\frac{p}{2})}$, ce qui entraîne, pour le couple (X,Y), la densité de probabilité élémentaire :

$$\frac{x^{\frac{n}{2}-1}.e^{-\frac{x}{2}}.y^{\frac{p}{2}-1}.e^{-\frac{y}{2}}}{2^{\frac{n}{2}}.2^{\frac{p}{2}}.\Gamma(\frac{n}{2}).\Gamma(\frac{p}{2})}.dx.dy$$

Pour déterminer la loi de $Z=X/Y$, on va utiliser le changement de variables $(X=Y.Z, Y=Y)$, transformation dont le *jacobien* est $J=\begin{vmatrix} Y & 0 \\ Z & 1 \end{vmatrix}=Y$. Ainsi (Y,Z) a-t-il pour densité de probabilité élémentaire $\dfrac{z^{\frac{n}{2}-1}.y^{\frac{n}{2}-1}.e^{-\frac{z.y}{2}}.y^{\frac{p}{2}-1}e^{-\frac{y}{2}}.y}{2^{\frac{n}{2}}.\Gamma(\frac{n}{2}).2^{\frac{p}{2}}.\Gamma(\frac{p}{2})}.dy.dz$.

La *loi des probabilités marginales* conduit à la densité $\varphi(z)$ de la variable aléatoire Z, à savoir $\varphi(z)=\dfrac{z^{\frac{n}{2}-1}.\int_0^{+\infty} y^{(\frac{n+p}{2}-1)}.e^{-\frac{y.(1+z)}{2}}.dy}{2^{\frac{n}{2}}.2^{\frac{p}{2}}.\Gamma(\frac{n}{2}).\Gamma(\frac{p}{2})}$. Posant $\dfrac{y.(1+z)}{2}=u \Rightarrow y=\dfrac{2.u}{1+z}$ et $dy=\dfrac{2.du}{1+z}$, on obtient $\varphi(z)=\dfrac{z^{\frac{n}{2}-1}.2^{\frac{n}{2}}.2^{\frac{p}{2}}.\int_0^{+\infty} u^{\frac{n+p}{2}-1}.e^{-u}.du}{2.2^{\frac{n}{2}}.\Gamma(\frac{n}{2}).2^{\frac{p}{2}}.\Gamma(\frac{p}{2}).(1+z)^{\frac{n+p}{2}}}$, soit après simplifications et rappel de l'expression de *l'intégrale d'*EULER, $\Gamma(\alpha)=\int_0^{+\infty} t^{\alpha-1}.e^{-t}.dt$, $\varphi(z)=\dfrac{\Gamma(\frac{n+p}{2})}{\Gamma(\frac{n}{2}).\Gamma(\frac{p}{2})}.\dfrac{z^{\frac{n}{2}-1}}{(1+z)^{\frac{n+p}{2}}}$.

2-b) Le nouveau changement de variable $F=p.\dfrac{Z}{n}$ conduit, pour la variable aléatoire F à la densité de probabilité élémentaire $\dfrac{\Gamma(\frac{n+p}{2})}{\Gamma(\frac{n}{2}).\Gamma(\frac{p}{2})}.\dfrac{n^{\frac{n}{2}-1}.f^{\frac{n}{2}-1}.p^{\frac{n}{2}}.p^{\frac{p}{2}}.n.f}{(p+n.f)^{\frac{n+p}{2}}.p^{\frac{n}{2}-1}.p}.df$.

Après simplifications, il reste pour ce qui est de la densité de probabilité $\psi(f)$ de la variable aléatoire F, l'expression donnée en page suivante.

$\psi(f)=\dfrac{\Gamma(\frac{n+p}{2})}{\Gamma(\frac{n}{2}).\Gamma(\frac{p}{2})}.\dfrac{n^{\frac{n}{2}}.p^{\frac{p}{2}}.f^{\frac{n}{2}-1}}{(p+n.f)^{\frac{n+p}{2}}}$. Il s'agit bien de la *loi de* FISHER-SNEDECOR, dont la densité de probabilité est rappelée au début de ce chapitre.

2°) Les statistiques $n.S_X^2/\sigma_X^2$ et $p.S_Y^2/\sigma_Y^2$ suivent les lois du chi- deux, respectivement à n et p degrés de libertés (cf. rappels de cours). De la question précédente, il résulte que la statistique $F=\dfrac{n.S_X^2/n.\sigma_X^2}{p.S_Y^2/p.\sigma_Y^2}=\dfrac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2}$ suit la *loi de* FISHER-SNEDECOR, $F(n,p)$ à n et p degrés de libertés.

• De même, les statistiques $n.S'^2_X/\sigma_X^2$ et $p.S'^2_Y/\sigma_Y^2$ suivent les *lois de chi- deux*, respectivement à $n-1$ et $p-1$ degrés de libertés (cf. rappels de cours et application 1.1 du présent chapitre). Il s'ensuit que la statistique $F=\dfrac{n.S'^2_X/(n-1).\sigma_X^2}{p.S'^2_Y/(p-1).\sigma_Y^2}$ suit *la loi de* FISHER-SNEDECOR, $F(n-1,p-1)$ à $n-1$ et $p-1$ degrés de libertés.

3°) Soient $S'^2_X=\dfrac{1}{n_X}.\sum_{i=1}^{i=n_X}(X_i-\overline{X})^2$ et $S'^2_Y=\dfrac{1}{n_Y}.\sum_{i=1}^{i=n_Y}(Y_i-\overline{Y})^2$ les statistiques qui décrivent les variances calculées pour chacun des échantillons prélevés respectivement dans les deux populations considérées ($n_X=20, n_Y=30, \sigma_X^2=25, \sigma_Y^2=16$). Précisons cependant qu'on a pris ici les *variances calculées sur l'échantillon et non pas les variances corrigées* $\widehat{S_X}^2$ et $\widehat{S_Y}^2$.

Il s'agit ici de calculer la probabilité $Prob(S'^2_X \succ 2.S'^2_Y)$, soit $Prob(S'^2_X/S'^2_Y \succ 2)$. Or, la variable $F=\dfrac{n_X.S'^2_X/(n_X-1).\sigma_X^2}{n_Y.S'^2_Y/(n_Y-1).\sigma_Y^2}$ suit la loi de FISHER-SNEDECOR, $F(n_X-1,n_Y-1)$.

Ainsi, $Prob(S'^2_X/S'^2_Y \succ 2) \Rightarrow Prob(F \succ 2.\dfrac{n_X.(n_Y-1).\sigma_Y^2}{n_Y.(n_X-1).\sigma_X^2})$, soit $Prob(F \succ 1,30)$ où F est la loi de FISHER-SNEDECOR, $F(19,29)$.

Des tables plus élaborées que celles annexées au présent ouvrage sont nécessaires ici pour obtenir un résultat suffisamment précis. En effet, ces tables indiquent, pour $\nu_X=20, \nu_Y=29$, les résultats $Prob(F \succ 1,94)=0,05$ et $Prob(F \succ 2,21)=0,025$. La seule conclusion qu'on peut en tirer est que $Prob(F \succ 1,30)$ est nettement supérieure à 0,05 voire 0,10.

En fait, l'appel à un calculateur trouvé sur internet fournit plus précisément, pour la loi $F(19,29)$, et par approximations successives les évaluations ci-après.

L'appel à l'un des calculateurs accessibles par internet fournit plus précisément, par approximations successives et pour la loi $F(19,29)$, les évaluations :

$Prob(F \succ 1,684) = 0,10$ - $Prob(F \succ 1,404) = 0,20$ - $Prob(F \succ 1,230) = 0,30$
$Prob(F \succ 1,31) = 0,25$ - $Prob(F \succ 1,29) = 0,26$

La réponse cherchée est donc environ, 0,255.

1.4 Distributions d'échantillonnage des différences de moyennes

Enoncé : Considérant deux échantillons indépendants de tailles n_1 et n_2, soient $(X_1, X_2, ..., X_{n_1})$ et $(Y_1, Y_2, ..., Y_{n_2})$ extraits de deux populations P_1 et P_2 dont les moyennes et variances sont respectivement (m_1, m_2) et (σ_1^2, σ_2^2), il est proposé de déterminer la distribution d'échantillonnage de la statistique $\overline{X} - \overline{Y}$ (avec $\overline{X} = \frac{\sum_{i=1}^{i=n_1} X_i}{n_1}$ et $\overline{Y} = \frac{\sum_{i=1}^{i=n_2} Y_i}{n_2}$).

PARTIE I

On suppose, dans cette partie, que les variances σ_1^2 et σ_2^2 sont connues.

I-1°) Caractériser la loi limite suivie par la statistique $\overline{X} - \overline{Y}$ dans le cas de grands échantillons ($n_1 \geq 30, n_2 \geq 30$).

I-2°) Quelles conditions faut-il imposer pour conduire les calculs à terme dans le cas de petits échantillons ($n_1 \prec 30, n_2 \prec 30$) ?

I-3°) Les lampes électriques fabriquées par un industriel A ont une durée de vie moyenne de $2000h$ avec un écart-type de $300h$. Celles fabriquées par un industriel B, ont une durée de vie moyenne de $1500h$ avec un écart-type de $200h$.

Testant des échantillons aléatoires de tailles 100 et 130, respectivement pour chacune des fabrications en question, déterminer la probabilité pour que les lampes prélevées dans la fabrication émanant de A aient une durée de vie moyenne au moins supérieure de $600h$ à celles prélevées dans la fabrication issue de B.

PARTIE II

On suppose, dans cette partie, que les variances σ_1^2 et σ_2^2 sont inconnues.

II-1°) Dans l'hypothèse de grands échantillons ($n_1 \geq 30, n_2 \geq 30$), par quelle loi limite peut-on approcher la loi de $\overline{X} - \overline{Y}$?.

II-2°) Etant cette fois, dans l'hypothèse de petits échantillons ($n_1 \prec 30, n_2 \prec 30$) et supposant $\sigma_1 = \sigma_2 = \sigma$, caractériser la loi de $\overline{X} - \overline{Y}$.

II-3°) Soient deux populations de chevaux de courses, à savoir les bons sauteurs et les mauvais sauteurs. On étudie la hauteur du garrot que l'on suppose être distribuée normalement (c'est-à-dire suivant une loi normale), dans les deux populations. Pour cela, on prélève un échantillon dans chacune de ces deux populations, ce qui donne les résultats présentés dans le tableau présenté en page suivante.

Population	Taille de l'échantillon	Moyenne	Ecart- type corrigé (*)
Bons sauteurs	$n_1 = 50$	$\overline{x_1} = 164$	$\widehat{s_1} = 4,7$
Mauvais sauteurs	$n_2 = 40$	$\overline{x_2} = 161,5$	$\widehat{s_2} = 5,2$

(*) On rappelle que $\widehat{S}^2 = \frac{1}{n-1}.\sum_{i=1}^{i=n}(X_i - \overline{X})^2$.

II-3°-a) Dans l'hypothèse où la hauteur du garrot est la même (en moyenne) pour chacune des populations « bons sauteurs » et « mauvais sauteurs », dans quel intervalle se situe la statistique $d = \left|\overline{X_1} - \overline{X_2}\right|$ dans 95% des cas ?

II-3°-b) Les résultats observés ici, pour $\overline{X_1}$ et $\overline{X_2}$ sont-ils conformes à l'hypothèse susmentionnée suivant laquelle il n'y a pas de différence significative entre les moyennes au garrot des bons et des mauvais sauteurs ?

II-3°-c) Reprendre la question précédente dans le cas de deux échantillons d'effectifs $n_1 = 15$ et $n_2 = 12$.

Solution : I-1°) Pour $n_1 \geq 30$ et $n_2 \geq 30$, le **théorème central limite** justifie la *convergence* de chacune des statistiques $\overline{X} = \frac{\sum_{i=1}^{i=n_1} X_i}{n_1}$ et $\overline{Y} = \frac{\sum_{i=1}^{i=n_2} Y_i}{n_2}$ vers les *lois normales* respectives $N(m_1, \frac{\sigma_1^2}{n_1})$ et $N(m_2, \frac{\sigma_2^2}{n_2})$. Il en résulte que la différence $\overline{X} - \overline{Y}$ *converge* également vers une *loi normale* d'espérance $E(\overline{X}) - E(\overline{Y})$ suivant la linéarité de l'espérance mathématique, et de variance égale à $Var(\overline{X}) + Var(\overline{Y})$ et non pas $Var(\overline{X}) - Var(\overline{Y})$, car il ne faut pas oublier que $Var(a.X) = a^2.Var(X), \forall (a, X)$.

Finalement, la statistique $\overline{X} - \overline{Y}$ *converge* vers la *loi normale* $N(m_1 - m_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$, ce qui s'écrit aussi $\frac{\overline{X} - \overline{Y} - (m_1 - m_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$ *converge* vers la *loi normale, centrée, réduite*, $N(0,1)$.

I-2°) Dans le cas de **petits échantillons**, cette convergence n'est plus vérifiée et la connaissance des lois de X et de Y est nécessaire pour mener à terme un calcul exact.

En particulier, **dans le cas d'échantillons gaussiens**, $\overline{X}$ et $\overline{Y}$ sont des *variables aléatoires normales* et on aura donc $\frac{\overline{X} - \overline{Y} - (m_1 - m_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$ qui *suit* la loi $N(0,1)$.

I-3°-a) Avec les notations précédentes, on a pour l'exemple proposé, $(m_1 = 2000, n_1 = 100, \sigma_1 = 300)$ et $(m_2 = 1500, n_2 = 130, \sigma_2 = 200)$.

On cherche à évaluer $Prob(\overline{X}-\overline{Y}\geq 600)$. Les conditions de convergence vers la loi normale étant satisfaites ici, on a donc $\dfrac{\overline{X}-\overline{Y}-(2000-1500)}{\sqrt{\dfrac{300^2}{100}+\dfrac{200^2}{130}}}$ dont après calculs, l'expression s'écrit $\dfrac{\overline{X}-\overline{Y}-500}{34,75}$, qui suit, approximativement, la loi normale $N(0,1)$.

Ainsi, $Prob(\overline{X}-\overline{Y}\geq 600)$ s'écrit, en notant par ξ la variable normale, centrée, réduite, $N(0,1)$, $Prob(\xi \geq \dfrac{600-500}{34,75}=2,87)=0,0021$.

• On notera que c'est très faible, mais il est bien évident que les fluctuations de $\overline{X}$ et de $\overline{Y}$ diminuent très sensiblement quand on augmente la taille de l'échantillon (plus précisément, cette variation est inversement proportionnelle à $\sqrt{n}$).

II-1°) Dans le cas où σ_1 **et** σ_2 **sont inconnues**, le **théorème central limite** autorise ici encore, lorsqu'il s'agit de **grands échantillous**, la *convergence* de la statistique $\dfrac{\overline{X}-\overline{Y}-(m_1-m_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1}+\dfrac{\sigma_2^2}{n_2}}}$ vers la *loi normale* $N(0,1)$.

On pourra alors finaliser les calculs en approchant $\sqrt{\dfrac{\sigma_1^2}{n_1}+\dfrac{\sigma_2^2}{n_2}}$ par $\sqrt{\dfrac{\widehat{S}_1^{\,2}}{n_1}+\dfrac{\widehat{S}_2^{\,2}}{n_2}}$ avec $\widehat{S}_1^{\,2}=\dfrac{1}{n_1-1}.\sum_{i=1}^{i=n_1}(X_i-\overline{X})^2$ et $\widehat{S}_2^{\,2}=\dfrac{1}{n_2-1}.\sum_{i=1}^{i=n_2}(Y_i-\overline{Y})^2$.

II-2°) Dans le cas de **petits échantillons**, l'hypothèse $\sigma_1=\sigma_2=\sigma$ est quasiment *incontournable* pour mener à bien les calculs (car dans le cas contraire, il faut recourir à des tables très complexes en fonction du rapport des variances).

• On notera cependant qu'en dépit de son caractère restrictif, cette hypothèse d'égalité de la variance entre les deux populations, est souvent vérifiée. En effet, ce qu'on mesure principalement, ce sont les traitements ou les fabrications d'un même objet ce qui généralement ne modifie par la variance, du moins à un instant donné. Cette dernière est plus fonction de l'usure d'une phénomène au cours du temps, que des variations instantanées de réglages.

• Regroupant les deux échantillons de tailles n_1 et n_2 pour former un estimateur commun de σ_1^2 et σ_2^2, soit $\widehat{S}^2$, on remarque que $(n_1-1).\dfrac{\widehat{S}_1^{\,2}}{\sigma_1^2}$ et $(n_2-1).\dfrac{\widehat{S}_2^{\,2}}{\sigma_2^2}$ suivent respectivement les *lois du chi- deux*, $\chi^2(n_1-1)$ et $\chi^2(n_2-1)$ (se reporter à l'application 1.1 du présent chapitre). Il s'ensuit que $(n_1-1).\dfrac{\widehat{S}_1^{\,2}}{\sigma_1^2}+(n_2-1).\dfrac{\widehat{S}_2^{\,2}}{\sigma_2^2}$ suit la *loi du chi- deux* de type $\chi^2(n_1-1+n_2-1)$, soit $\chi^2(n_1+n_2-2)$ (toujours d'après le même exercice 1.1 mentionné ci-dessus).

• Remplaçant $\frac{(n_1-1).\widehat{S_1}^2+(n_2-1).\widehat{S_2}^2}{\sigma^2}$ par $\frac{(n_1-1+n_2-1).\widehat{S}^2}{\sigma^2}$, on conclut ainsi que la statistique $(n_1+n_2-2).\frac{\widehat{S}^2}{\sigma^2}$ suit la loi $\chi^2(n_1+n_2-2)$.

• Dans ces conditions, $\frac{\overline{X}-\overline{Y}-(m_1-m_2)}{\hat{S}.\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}=\frac{\frac{\overline{X}-\overline{Y}-(m_1-m_2)}{\sigma.\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}}{\sqrt{\frac{(n_1+n_2-2)}{(n_1+n_2-2)}.\frac{\hat{S}^2}{\sigma^2}}}$, s'identifie à $\frac{U}{\sqrt{\frac{V}{n_1+n_2-2}}}$,

où U suit *la loi normale, centrée, réduite*, $N(0,1)$, et où V suit la *loi du chi- deux* à $\nu=n_1+n_2-2$ degrés de libertés, soit $\chi^2(n_1+n_2-2)$. Conformément aux résultats montrés dans l'application 1.1 du présent chapitre, il s'agit de *la loi de* STUDENT à $\nu=n_1+n_2-2$ degrés de libertés.

II-3°-a) Dans l'exemple proposé ici, on se trouve dans le cas de *grands échantillons* avec des variances *inconnues et non nécessairement égales*. D'après les résultats de la question II-1°) précédente, la statistique $\frac{\overline{X}-\overline{Y}-(m_1-m_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$ suit donc la loi normale $N(0,1)$ (« *suit* » et non « *converge vers* » car la distribution de référence (celle du caractère étudié qui est en l'occurrence la hauteur du garrot, est supposée être une loi normale).

Désignant par t_α le nombre vérifiant $Prob(-t_\alpha \le \xi \le t_\alpha)=0,95$, où ξ est la variable normale, centrée, réduite, de loi $N(0,1)$, il vient par lecture dans la table des valeurs de ξ (cf. annexes), $t_\alpha=1,96$. Il en résulte immédiatement et sous l'hypothèse $m_1=m_2$, l'encadrement $-1,96.\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}} \le \overline{X_1}-\overline{X_2} \le 1,96.\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$.

Remplaçant σ_1 et σ_2 par leurs estimations $\widehat{s_1}=4,7$ et $\widehat{s_2}=5,2$, il vient numériquement, l'encadrement cherché $-2,07 \le \overline{X_1}-\overline{X_2} \le +2,07$. Ce résultat, signifie que, sous l'hypothèse d'une hauteur moyenne de garrot égale pour les bons et les mauvais sauteurs, la différence des moyennes $\overline{X_1}$ et $\overline{X_2}$ associées aux échantillons de tailles 50 et 40, respectivement prélevés dans les populations « *bons sauteurs* » et «*mauvais sauteurs* », se trouve comprise entre -2,07 et +2,07 dans 95% des cas (soit, à l'extérieur de cet **intervalle dit « de confiance** », dans 5% des cas).

• Or, concrètement, et pour l'exemple choisi, on a $\overline{x_1}-\overline{x_2}=164-161,5=2,5$. Il s'agit d'une valeur à l'extérieur de l'intervalle susmentionné. Il est donc prudent ici de **rejeter** l'affirmation « *il n'y pas de différence de hauteur de garrot entre les bons et les mauvais sauteurs* », le **risque** pris à travers cette décision étant celui d'avoir $\overline{X_1}-\overline{X_2} \notin [-2,07;+2,07]$ alors qu'on a $m_1=m_2$, soit 5%.

Cet exercice préfigure la **décision statistique** (*théorie des tests*), développée au chapitre III.

• La théorie exacte qui consisterait à supposer $\sigma_1 = \sigma_2 = \sigma$ et à utiliser la statistique $\dfrac{\overline{X} - \overline{Y} - (m_1 - m_2)}{\hat{S}.\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$ où $(n_1 + n_2 - 2).\hat{S}^2 = (n_1 - 1).\hat{S}_1^{\,2} + (n_2 - 1).\hat{S}_2^{\,2}$ conduit, numériquement, à l'estimation $\hat{s} = 4,9278$. Considérant, par ailleurs, la loi de STUDENT à $50 + 40 - 2 = 88$ degrés de libertés, soit T, et le nombre t_α vérifiant $\Pr ob(-t_\alpha \le T \le t_\alpha) = 0,95$, il s'ensuit $t_\alpha = 1,99$ et l'encadrement :

$$-1,99 \times 4,9278 \times \sqrt{\frac{1}{50} + \frac{1}{40}} \le \overline{X_1} - \overline{X_2} \le +1,99 \times 4,9278 \times \sqrt{\frac{1}{50} + \frac{1}{40}}$$

Après calculs, on obtient le résultat $-2,08 \le \overline{X_1} - \overline{X_2} \le 2,08$ qui est très proche du résultat antérieur, ce qui valide l'approximation faite précédemment en remplaçant les σ_i par leurs estimations $\hat{s}_i$.

II-3°-b) Par contre, lorsqu'on a des *petits échantillons* la méthode exacte qui consiste à supposer $\sigma_1 = \sigma_2 = \sigma$ et à passer par la *loi de* STUDENT est incontournable. Par exemple, pour $n_1 = 15, n_2 = 12$, on trouve successivement $\hat{s} = 4,9262, t_\alpha = 2,06$, et l'encadrement $-3,93 \le \overline{X_1} - \overline{X_2} \le +3,93$.

Cette fois, la même *différence observée* $\overline{x_1} - \overline{x_2} = 164 - 161,5 = 2,5$ se situe dans la zone d'acceptation de l'hypothèse "$m_1 = m_2$". En fait, et comme on le verra dans le chapitre III, l'augmentation de la taille des échantillons et à fortiori des informations disponibles, affine la précision des conclusions et la pertinence de la décision.

• Enfm, dans cette application comme la suivante, on considère *extraire des échantillons de deux populations différentes.* Mais, il est bien évident que les résultats obtenus ici *demeurent valides* lorsqu'il s'agit de *comparer deux échantillons au sein d'une même population.*

1.5 Distributions d'échantillonnage des différences de proportions

Enoncé : On considère deux populations P_1 et P_2 au sein desquelles un caractère C donné est rencontré avec les probabilités respectives p_1 et p_2.

1°) On forme deux échantillons indépendants et avec remise extraits de chacune des deux populations en question et de tailles respectives n_1 et n_2. On considère les statistiques F_1 et F_2 définies par les fréquences observées de C dans chacun desdits échantillons (fréquences empiriques). Supposant n_1 et n_2 assez grands (en pratique, $n_1 \ge 30, n_2 \ge 30$), préciser vers quelle loi converge $F_1 - F_2$?

2°) Une étude révèle que dans une population donnée, 35% des personnes ont les yeux bleus. Quelle est la probabilité pour que les fréquences observées sur deux échantillons distincts de taille 200 extraits de cette population, soient distantes d'au moins 5% ?

Solution : 1°) Se référant à la variable X_1 égale au nombre de personnes ayant les yeux bleus parmi les n_1 personnes constituant l'échantillon extrait de la population P_1, il est immédiat que X_1 suit la loi *binomiale* $B(n_1, p_1)$ dont la *convergence vers la loi normale* est assurée pour $n_1 \geq 30$ et p ni trop faible, ni trop voisin de 1. Plus précisément, il s'agit de la *loi normale* $N(n_1.p_1, \sigma^2 = n_1.p_1.q_1)$.

Quant à $F_1 = \frac{X_1}{n_1}$, c'est aussi, à la limite, une *loi normale* d'espérance $E(F_1) = \frac{1}{n_1}.E(X_1) = p_1$ et de variance $Var(F_1) = \frac{1}{n_1^2}.n_1.p_1.q_1 = \frac{p_1.q_1}{n_1}$. Bien entendu, il en est de même pour F_2 dont la loi limite est la *loi normale* $N(p_2, \sigma^2 = \frac{p_2.q_2}{n_2})$.

• Dans ces conditions et compte tenu de l'indépendance entre les échantillons, la statistique $F_1 - F_2$ *converge vers la loi normale* de moyenne $E(F_1 - F_2) = E(F_1) - E(F_2) = p_1 - p_2$ et de variance $Var(F_1 - F_2) = Var(F_1) + Var(F_2)$, soit $Var(F_1 - F_2) = \frac{p_1.q_1}{n_1} + \frac{p_2.q_2}{n_2}$. Une autre façon d'écrire ce résultat est la convergence de la variable, centrée, réduite, $\frac{F_1 - F_2 - (p_1 - p_2)}{\sqrt{\frac{p_1.q_1}{n_1} + \frac{p_2.q_2}{n_2}}}$ vers la *variable normale* de loi $N(0,1)$.

2°) Comme il a été remarqué en fin d'application I-4°) du présent chapitre, le raisonnement susmentionné reste valable lorsqu'il s'agit de comparer deux échantillons issus d'une même population. Ainsi, transcrivant les conditions de l'application numérique proposée, on a $p_1 = p_2 = p = 0,35$ et $n_1 = n_2 = 200$, la question posée étant d'évaluer $\Pr ob(|F_1 - F_2| \geq 0,05)$.

ξ étant la variable normale, centrée, réduite, égale à $\frac{F_1 - F_2}{\sqrt{p.q}.\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ (puisque $p_1 = p_2 = p$), la probabilité cherchée s'écrit en définitive :

$$\Pr ob(|\xi| \geq \frac{0,05}{\sqrt{0,35 \times 0,65}.\sqrt{\frac{1}{200} + \frac{1}{200}}} = 1,048)$$

Utilisant la table annexée qui donne les valeurs de la fonction de répartition $\Pi(t)$ de la variable aléatoire normale, centrée, réduite, ξ, on a immédiatement $\Pr ob(|\xi| \geq 1,048) = 2.\Pr ob(\xi \geq 1,048) = 2.[1 - \Pi(1,048)] = 2.(1 - 0,853) = 0,293$.

1.6 La différence entre estimation et estimateur

Cet exercice illustre les propriétés mises en évidence dans les applications précédentes concernant moyenne et variance et permet de bien comprendre la différence entre la donnée particulière d'un échantillon et la distribution d'échantillonnage, voire plus largement le principe de l'inférence statistique.

Enoncé : Une population comprend les valeurs suivantes 3,5,7,9,12 relativement à un caractère donné.

1°) Evaluer la moyenne et l'écart- type de ces valeurs au sein de la population.

2°) On se propose d'estimer ces paramètres à partir d'échantillons de taille $n=2$ extraits avec remise de cette population et ceci de façon uniforme.

2-a) Enumérer tous les échantillons qui peuvent être ainsi extraits et pour chacun, calculer la valeur de la statistique $\overline{X}=\frac{\sum_{i=1}^{i=n} X_i}{n}$. Interpréter la série de valeurs ainsi obtenue.

2-b) On forme la moyenne des valeurs ci-dessus. Quels résultats obtient-on et quelle interprétation peut-on en tirer ?

2-c) On forme la variance de la série des valeurs exprimées à la question 2-a). Se rapprochant de la question 1°), retrouver ainsi le résultat classique $Var(\overline{X})=\frac{\sigma^2}{n}$.

2-d) Pour chacun des échantillons énumérés en question 2-a), calculer les variances associées $S'^2=\frac{1}{n}.\sum_{i=1}^{i=n}(X_i-\overline{X})^2$.

2-e) On forme la moyenne des variances susmentionnées en 2-d). Interpréter ici encore le résultat obtenu et retrouver l'expression classique $E(\hat{S}^2)=\sigma^2$.

Solution : 1°) Les cinq valeurs qui sont proposées ici au sein de la population considérée de taille $N=5$, ont immédiatement pour moyenne m et pour variance σ^2 :

$$m=\frac{3+5+7+9+12}{5}=7,2 \text{ et } \sigma^2=\frac{3^2+5^2+7^2+9^2+12^2}{5}-(7,2)^2=9,76, \text{ soit } \sigma=3,124.$$

2-a) Les échantillons de taille $n=2$ que l'on peut extraire avec remise de la population en question de taille $N=5$, sont en nombre égal à $N^n=5^2=25$. Plus précisément, leur énumération conduit à la série d'échantillons ci-dessous :

(3,3)	(3,5)	(3,7)	(3,9)	(3,12)
(5,3)	(5,5)	(5,7)	(5,9)	(5,12)
(7,3)	(7,5)	(7,7)	(7,9)	(7,12)
(9,3)	(9,5)	(9,7)	(9,9)	(9,12)
(12,3)	(12,5)	(12,7)	(12,9)	(12,12)

Pour chacun des ces 25 échantillons possibles, la valeur de la statistique $\overline{X}=\frac{\sum_{i=1}^{i=2} X_i}{2}$ est calculée ci- après :

3	4	5	6	7,5	4	5
6	7	8,5	5	6	7	8
9,5	6	7	8	9	10,5	7,5
8,5	9,5	10,5	12			

Ces valeurs sont toutes équiprobables (hypothèse de l'uniformité des prélèvements aléatoires) et ceci suivant la *probabilité uniforme* $\frac{1}{N^n}=\frac{1}{25}$ puisqu'il s'agit, qui plus est, de tirages avec remise.

La série des 25 *valeurs* possibles de $\overline{X}$, munie de la *probabilité uniforme* représente la loi de probabilité de $\overline{X}$. C'est **la distribution d'échantillonnage** !

2-b) Le calcul de $E(\overline{X})$ à partir des 25 valeurs u_i ci-dessus et de leurs probabilités $p_i = \frac{1}{25}(1 \le i \le 25)$, conduit à $E(\overline{X}) = \sum_{i=1}^{i=25} u_i.p_i = \frac{1}{25}.(3+4.....+12) = 7,2$.

Or cette valeur de 7,2, c'est aussi la valeur de la moyenne calculée sur l'ensemble de la population dans la 1ère question. On retrouve donc ici le résultat $E(\overline{X}) = m$.

• En faisant la moyenne de l'ensemble des moyennes $\overline{X}$ calculées à partir des N^n échantillons que l'on peut extraire de la population de référence de taille N, on obtient la valeur de la moyenne sur l'ensemble de la population.

2-c) De même, le calcul de $Var(\overline{X})$ conduit à former, à partir des 25 valeurs précédentes de $\overline{X}$, soient $u_i (1 \le i \le 25)$, l'expression $Var(\overline{X}) = \frac{1}{25}.\sum_{i=1}^{i=25}(u_i - 7,2)^2 = \frac{1}{25}.\sum_{i=1}^{i=25} u_i^2 - 7,2^2$.

Il s'ensuit, numériquement, $Var(\overline{X}) = \frac{1418}{25} - 7,2^2 = 4,88$.

• Comparant avec la variance calculée sur l'ensemble de la population, soit $\sigma^2 = 9,76$. calculée sur l'ensemble de la population (cf. 1ère question), on retrouve ici le résultat classique $Var(\overline{X}) = \frac{\sigma^2}{n}$ (avec $n = 2$ puisque les échantillons extraits ici, sont de taille 2).

2-d) Reprenant le raisonnement précédent pour ce qui est des variances, on obtient pour chacun des 25 échantillons et comme expressions de S'^2, les résultats $\frac{1}{2}.\left[(3-3)^2 + (2-3)^2\right] = 0$, $\frac{1}{2}.\left[(3-4)^2 + (5-4)^2\right] = 1$,.... Soit la série des 25 valeurs :

0	1	4	9	20,25	1	0
1	4	12,25	4	1	0	1
6,25	9	4	1	0	2,25	20,25
12,25	6,25	2,25	0			

2-e) Ici encore, on a affaire pour ce qui est de S'^2, à une variable aléatoire à 25 valeurs v_i équidistribuées de probabilités associées $p_i = \frac{1}{N} = \frac{1}{25}(1 \le i \le 25)$, variable dont le calcul de l'espérance mathématique conduit à $E(S'^2) = \sum_{i=1}^{i=25} v_i.p_i = \frac{1}{25}.\sum_{i=1}^{i=25} v_i = 4,88$.

• Or se référant à $\sigma^2 = 9,76$ qui représente la variance sur la totalité de la population, on constate ici que $E(S'^2)$ vérifie la relation $E(S'^2) = \frac{n}{n-1}.\sigma^2$ (avec $n = 2$). En considérant pour chaque échantillon, la statistique $\hat{S}^2 = \frac{n}{n-1}.S'^2$ (soit $\hat{S}^2 = 2.S'^2$), on retrouve le résultat $E(\hat{S}^2) = \sigma^2$.

• Les mécanismes développés ci-dessus montrent, *comment à partir des échantillons et des distributions des statistiques d'échantillonnages*, on est conduit à *tirer des conclusions* quant à *l'ensemble de la population considérée.*

C'est la propriété de **l'inférence statistique** qui sert de *fil conducteur à l'ensemble des techniques de statistique mathématique* présentées dans les chapitres suivants.

2. Exemples de méthodes d'échantillonnage

Même si les tirages aléatoires équiprobables et avec remise, restent l'hypothèse la plus couramment admise dans les développements de statistique mathématique, d'autres méthodes plus perfectionnées sont susceptibles d'augmenter considérablement l'efficacité de l'échantillonnage. Plusieurs exemples en sont présentés ci-après.

2.1 Les sondages aléatoires (équiprobables) sans replacement (sondages dits « exhaustifs »)

Enoncé : Dans cette application, on constitue des échantillons de taille n par prélèvements aléatoires sans répétition (sans remise) au sein d'une population de taille N dans laquelle m et σ^2 désignent la moyenne et la variance du caractère aléatoire étudié.

Pour chaque élément "i" de cette population, la valeur du caractère X en question est notée x_i.

1-a) On associe à chaque élément "i" susmentionné, la variable indicatrice ε_i égale à 1 si l'élément fait partie de l'échantillon de taille n considéré et à 0, dans le cas contraire.

Montrer que $\overline{X} = \dfrac{\sum_{i=1}^{i=n} X_i}{n} = \dfrac{1}{n}.\sum_{i=1}^{i=N} X_i.\varepsilon_i$.

1-b) Caractériser, pour i fixé, la loi de $X_i.\varepsilon_i$ et en déduire l'expression de $E(\overline{X})$.

1-c) Pour tout couple $(i,j)/1 \le i \le N, 1 \le j \le N$, calculer $E(\varepsilon_i.\varepsilon_j)$.

1-d) Formant $\left[\sum_{i=1}^{i=N}(X_i - m)\right]^2$, montrer que $\sum_{\substack{i=1 \\ i \neq j}}^{i=N}\sum_{j=1}^{j=N}(X_i - m).(X_j - m) = -N.\sigma^2$. En déduire la valeur de $Var(\overline{X})$.

2°) Montrer que la statistique $S'^2 = \dfrac{1}{n}.\sum_{i=1}^{i=n}(X_i - \overline{X})^2$ a pour moyenne $E(S'^2) = (\dfrac{N}{N-1}).(\dfrac{n-1}{n}).\sigma^2$.

3°) En reprenant la population de taille $N = 5$ de l'application 1.6 précédente et ses valeurs (3, 5, 7, 9, 12), retrouver les résultats ci-dessus.

Solution : 1-a) Le résultat est en fait une évidence. La revue de tous les éléments $i/1 \le i \le N$ de la population et le filtrage induit par $\varepsilon_i = 1$ *si l'élément appartient à l'échantillon considéré* et $\varepsilon_i = 0$ *sinon*, conduit immédiatement à la relation cherchée :

$$\overline{X} = \frac{1}{n}.\sum_{i=1}^{i=n} X_i = \frac{1}{n}.\sum_{i=1}^{i=N} X_i.\varepsilon_i$$

Il est important de noter ici la différence de signification de l'indice de sommation i entre les deux sommes susmentionnées comme cela est expliqué ci-après.

Dans $\frac{1}{n}.\sum_{i=1}^{i=n} X_i$, X_i désigne la valeur de X pour l'élément de rang i dans l'échantillon formé et constitue donc une variable aléatoire. Dans $\frac{1}{n}.\sum_{i=1}^{i=N} X_i.\varepsilon_i$, X_i est la valeur associée sans ambiguïté à l'élément numéroté i dans la population considérée et est donc une valeur déterminée lorsque i est fixé (autrement dit, pour i fixé, X_i est constant).

1-b) La variable $X_i.\varepsilon_i$ a pour valeurs X_i et 0 avec les probabilités respectives $\Pr ob(\varepsilon_i = 1) = \frac{n}{N}$ et $\Pr ob(\varepsilon_i = 0) = 1 - \frac{n}{N}$.

Il s'ensuit par linéarité de l'espérance mathématique, $E(\overline{X}) = \frac{1}{n}.\sum_{i=1}^{i=N} E(X_i.\varepsilon_i)$ avec $E(X_i.\varepsilon_i) = n.\frac{X_i}{N} + (1 - \frac{n}{N}).0 = \frac{n}{N}.X_i$. En résumé, $E(\overline{X}) = \frac{1}{n}.\frac{n}{N}.\sum_{i=1}^{i=N} X_i = m$.

1-c) $E(\varepsilon_i^2) = 1^2.\Pr ob(\varepsilon_i = 1) + 0^2.\Pr ob(\varepsilon_i = 0) = \Pr ob(\varepsilon_i = 1) = \frac{n}{N}$. D'autre part, pour tout couple $(i, j)/i \neq j$, la loi du couple $(\varepsilon_i, \varepsilon_j)$ peut être représentée par le tableau de contingences ci-dessous :

$\varepsilon_i \setminus \varepsilon_j$	0	1
0	$\frac{(N-n)}{N}.\frac{(N-n-1)}{(N-1)}$	$\frac{(N-n)}{N}.\frac{n}{(N-1)}$
1	$\frac{n}{N}.\frac{(N-n)}{(N-1)}$	$\frac{n}{N}.\frac{(n-1)}{(N-1)}$

Ainsi a-t-on $E(\varepsilon_i.\varepsilon_j) = \frac{n}{N}.\frac{(n-1)}{(N-1)}$.

1-d) $\left[\sum_{i=1}^{i=N}(X_i - m)\right]^2 = \sum_{i=1}^{i=N}(X_i - m)^2 + \sum_{\substack{i=1 \\ i \neq j}}^{i=N}\sum_{j=1}^{j=N}(X_i - m).(X_j - m)$. Mais, il est bien évident que $\left[\sum_{i=1}^{i=N}(X_i - m)\right]^2 = \left[\sum_{i=1}^{i=N} X_i - N.m\right]^2 = 0$ puisque sur l'ensemble de la population on a $m = \frac{\sum_{i=1}^{i=N} X_i}{N}$. Il en découle la relation $\sum_{\substack{i=1 \\ i \neq j}}^{i=N}\sum_{j=1}^{j=N}(X_i - m).(X_j - m) = -\sum_{i=1}^{i=N}(X_i - m)^2 = -N.\sigma^2$, parce que sur l'ensemble de la population on a $\sigma^2 = \frac{1}{N}.\sum_{i=1}^{i=N}(X_i - m)^2$.

- Revenant à la variance $Var(\overline{X}) = E\left[(\overline{X} - m)^2\right]$ dont on cherche l'expression, on a d'une part $\overline{X} - m = \frac{1}{n}.\sum_{i=1}^{i=n}(X_i - m) = \frac{1}{n}.\sum_{i=1}^{i=N}(X_i - m).\varepsilon_i$, et d'autre part, suivant élévation au carré, $(\overline{X} - m)^2 = \frac{1}{n^2}.\sum_{i=1}^{i=N}(X_i - m)^2.\varepsilon_i^2 + \frac{1}{n^2}.\sum_{\substack{i=1 \\ i \neq j}}^{i=N}\sum_{j=1}^{j=N}(X_i - m).(X_j - m).\varepsilon_i.\varepsilon_j$.

Dès lors, par passage à l'espérance mathématique et remarquant, ici encore, que les valeurs $X_i - m$ et $X_j - m$ sont constantes pour i et j fixés, il vient compte tenu des expressions de $E(\varepsilon_i^2)$ et de $E(\varepsilon_i.\varepsilon_j)$ établies en 1-c), le résultat :

$$E(\overline{X}-m)^2 = \frac{1}{n^2}.\frac{n}{N}.\sum_{i=1}^{i=N}(X_i-m)^2 + \frac{1}{n^2}.\frac{n.(n-1)}{N.(N-1)}.\sum_{\substack{i=1\\ i\neq j}}^{i=N}\sum_{j=1}^{j=N}(X_i-m).(X_j-m)$$

Or, $\sum_{i=1}^{i=N}(X_i-m)^2 = N.\sigma^2$ et $\sum_{\substack{i=1\\ i\neq j}}^{i=N}\sum_{j=1}^{j=N}(X_i-m).(X_j-m) = -N.\sigma^2$. Par substitution dans l'expression précédente, on obtient $E(\overline{X}-m)^2 = \frac{1}{n^2}.\frac{n}{N}.N.\sigma^2 - \frac{n.(n-1)}{n^2.N.(N-1)}.N.\sigma^2$, soit en définitive, $E(\overline{X}-m)^2 = \frac{\sigma^2}{n}.(1-\frac{n-1}{N-1}) = \frac{N-n}{N-1}.\frac{\sigma^2}{n}$.

• Ainsi, lorsqu'il s'agit d'échantillons *exhaustifs*, la variance est-elle multipliée par le facteur $\frac{N-n}{N-1}$, expression qui nous est déjà familière à travers la *distinction entre la loi binomiale et la loi hypergéométrique.*

2°) On va s'inspirer ici de l'application 1.2 du présent chapitre. On a $n.S'^2 = \sum_{i=1}^{i=n}(X_i-\overline{X})^2 = \sum_{i=1}^{i=n}X_i^2 - n.\overline{X}^2$ et donc, par *linéarité de l'espérance mathématique*, $n.E(S'^2) = \sum_{i=1}^{i=n}E(X_i^2) - n.E(\overline{X}^2)$. Mais, $E(X_i^2) = Var(X_i) + E(X_i)^2 = \sigma^2 + m^2$ et de même, $E(\overline{X}^2) = Var(\overline{X}) + E(\overline{X})^2 = \frac{N-n}{N-1}.\frac{\sigma^2}{n} + m^2$ (d'après les résultats des questions antérieures).

En développant, $n.E(S'^2) = n.(\sigma^2+m^2) - n.\frac{\sigma^2}{n}.\frac{N-n}{N-1} - n.m^2 = n.\sigma^2.(1-\frac{N-n}{n.(N-1)})$, soit en conclusion :

$$E(S'^2) = \sigma^2.\left[\frac{n.N-n-N+n}{n.(N-1)}\right] = \frac{N}{N-1}.\frac{n-1}{n}.\sigma^2$$

3°) Reprenant l'application 1.6 précédente et la population de valeurs (3, 5, 7, 9, 12) de laquelle on extrait, cette fois, des échantillons de taille $n = 2$, sans remise, ces derniers se trouvent être en nombre C_5^2, soit dix possibilités représentées par les couples ci-dessous :

(3,5)	(3,7)	(3,9)	(3,12)	(5,7)
(5,9)	(5,12)	(7,9)	(7,12)	(9,12)

(on notera que les sélections (3,7) et (7,3) induisent le même échantillon et ainsi de suite...).

• La *série des moyennes* $\overline{x}$ conduit immédiatement aux valeurs ci-dessous :

4	5	6	7,5	6	7	8,5	8	9,5	10

• Son *espérance mathématique*, qui est égale à la moyenne précédente calculée sur la base de l'ensemble des échantillons possibles soit dix, est donc $\frac{1}{10}.(4+5+....+10,5)=7,2$. On retrouve le résultat $E(\overline{X})=m$.

• Quant à la *variance* des 10 moyennes en question, elle est égale à la somme :

$$\frac{1}{10}.\left[(4-7,2)^2+....(10,5-7,2)^2\right]=3,66.$$

Rappelant que $\sigma^2=9,76$, on retrouve la relation $Var(\overline{X})=\frac{N-n}{N-1}.\frac{\sigma^2}{n}$, puisqu'on a numériquement, $Var(\overline{X})=\frac{9,76}{2}.\frac{5-2}{5-1}=3,66$.

• De la même manière, le calcul de S'^2 pour chacun des dix échantillons possibles conduit à la série :

1	4	9	20,25	1	4	12,25	1	6,25	2,25

L'espérance mathématique est égale à $E(S'^2)=\frac{1}{10}.\left[1+4+....+2,25\right]=6,1$. Or, c'est aussi $\frac{N}{N-1}.\frac{n-1}{n}.\sigma^2$ avec $\sigma^2=9,76, N=5, n=2$. Ainsi la relation $E(S'^2)=\frac{N}{N-1}.\frac{n-1}{n}.\sigma^2$ établie à la 2ème question est-elle vérifiée pour l'exemple proposé.

2.2 Les sondages par stratification

La méthode vise à découper la population de référence de taille N en groupes homogènes au regard de certains critères, la constitution de l'échantillon de taille n à partir de ces groupes ayaut pour effet une meilleure représentativité et notamment un amortissement des fluctuations inhérentes à un prélèvement aléatoire élémentaire tels ceux décrits jusqu'à présent.

Enoncé : La population de taille N étant découpée en k strates de tailles $N_h (1\le h\le k)$ et les échantillons prélevés dans chacune des strates étant de tailles $n_h (1\le h\le k$ et $n_1+n_2...+n_k=n)$, on désignera par :

- m et σ^2 la moyenne et la variance du caractère aléatoire X étudié, sur l'ensemble de la population ;

- m_h et σ_h^2 la moyenne et la variance de X au sein de chacune des strates $(1\le h\le k)$;

- $(X_{h,1}, X_{h,2},...., X_{h,n_h})$ l'échantillon de taille n_h extrait de la sous- population "h" et $\overline{X_h}, \widehat{S_h}^2$ les moyennes et variances associées à savoir, pour rappel, les statistiques définies par les relations $\overline{X_h}=\frac{1}{n_h}.\sum_{i=1}^{i=n_h} X_{h,i}, \widehat{S_h}^2=\frac{1}{n_h-1}.\sum_{i=1}^{i=n_h}(X_{h,i}-\overline{X_h})^2$.

1°) On forme la statistique $\overline{X}=\sum_{h=1}^{h=k}\frac{N_h}{N}.\overline{X_h}$. Calculer $E(\overline{X})$ et $Var(\overline{X})$ dans chacune des hypothèses d'un sondage non exhaustif et d'un sondage exhaustif.

2°) On se propose dans cette question de présenter plusieurs méthodes de répartition des tailles des échantillons entre les strates, l'hypothèse de tirages exhaustifs étant supposée vérifiée ici.

2-a) A quelle condition $\overline{X} = \frac{\sum_{h=1}^{h=k}\sum_{i=1}^{i=n_h} X_{h,i}}{n}$?. En déduire une procédure de construction de l'échantillon (*méthode des prélèvements suivant taux de sondage uniforme*).

2-b) On se propose dans cette question de déterminer les n_h de sorte que $Var(\overline{X})$ soit minimale. Montrer que cela conduit aux conditions optimales $\frac{n_h}{N_h.\sigma_h}.\sqrt{\frac{N_h-1}{N_h}} = \text{constante}$ $(\forall h / 1 \le h \le k)$ et $\frac{n_h}{N_h.\sigma_h} = \text{constante}$ $(\forall h / 1 \le h \le k)$, dans le cas d'échantillonnages respectivement exhaustifs et non exhaustifs. En déduire une procédure de construction de l'échantillon (*méthode de l'échantillon optimum dite de* NEYMAN).

2-c) On suppose cette fois que le coût de l'enquête est fixé à l'avance, soit C, et on note par c_h les coûts unitaires de sondage pour chacune des strates h. Montrer que cela conduit aux conditions optimales $\sqrt{\frac{N_h-1}{N_h}}.\frac{n_h}{\frac{N_h.\sigma_h}{\sqrt{c_h}}} = \text{constante}$ $(\forall h / 1 \le h \le k)$ et $\frac{n_h}{\frac{N_h.\sigma_h}{\sqrt{c_h}}} = \text{constante}$, $(\forall h / 1 \le h \le k)$ dans le cas d'échantillonnages respectivement exhaustifs et non exhaustifs.

3°) On se propose ici de comparer les méthodes précédentes au plan de la précision de l'estimateur $\overline{X}$, précision caractérisée par $Var(\overline{X})$. On considère ainsi une population de 300 entreprises réparties en 4 strates dont les tailles N_h et les écart-types σ_h sont décrits ci-dessous, le paramètre étudié étant le chiffre d'affaires annuel (en millions d'euros).

	Strate 1	Strate 2	Strate 3	Strate 4
N_h	105	40	75	80
σ_h	6,35	5,52	8,75	34,53

On souhaite effectuer un échantillonnage de taille $n = 60$. Calculer $Var(\overline{X})$ pour chacune des hypothèses :

3-a) échantillonnages aléatoires stratifiés représentatifs (taux de sondage uniforme) avec et sans remise ;

3-b) échantillonnages aléatoires optimaux de NEYMAN avec et sans remise.

4°) Que peut-on conclure des résultats de la 3ème question ?

Solution : 1°) Par *linéarité de l'espérance mathématique*, on peut écrire les relations $E(\overline{X}) = \sum_{h=1}^{h=k} \frac{N_h}{N} E(\overline{X_h}) = \sum_{h=1}^{h=k} \frac{N_h m_h}{N} = m$. On notera, qui plus est, que ce résultat est valable quel que soit le mode de prélèvement, exhaustif ou non.

D'autre part, l'indépendance entre les strates, entraîne par *pseudo linéarité de la variance*, les relations $Var(\overline{X}) = \sum_{h=1}^{h=k} Var(\frac{N_h}{N}.\overline{X_h}) = \sum_{h=1}^{h=k} \frac{N_h^2}{N^2}.Var(\overline{X_h})$.

• Pour des **prélèvements non exhaustifs**, on a $Var(\overline{X_h}) = \frac{\sigma_h^2}{n_h} \Rightarrow Var(\overline{X}) = \sum_{h=1}^{h=k} \frac{N_h^2}{N^2}.\frac{\sigma_h^2}{n_h}$.

• Pour des **prélèvements exhaustifs**, on a suivant les résultats de l'application 2.1 précédente, $Var(\overline{X_h}) = \frac{N_h - n_h}{N_h - 1}.\frac{\sigma_h^2}{n_h} \Rightarrow Var(\overline{X}) = \sum_{h=1}^{h=k} \frac{N_h^2}{N^2}.\frac{N_h - n_h}{N_h - 1}.\frac{\sigma_h^2}{n_h}$.

2-a) $\overline{X} = \sum_{h=1}^{h=k} \frac{N_h}{N}.\sum_{i=1}^{i=n_h} \frac{X_{h,i}}{n_h} = \sum_{h=1}^{h=k}\sum_{i=1}^{i=n_h} \frac{N_h}{N.n_h}.X_{h,i}$ s'écrit sous la forme $\frac{1}{n}.\sum_{h=1}^{h=k}\sum_{i=1}^{i=n_h} X_{h,i}$ si et seulement si $\frac{N_h}{N.n_h} = \frac{1}{n} (\forall h / 1 \le h \le k)$. Il s'ensuit la condition $\frac{n_h}{N_h} = \frac{n}{N} (\forall h / 1 \le h \le k)$.

Ce rapport constant $f = \frac{n_1}{N_1} = \frac{n_2}{N_2} .... = \frac{n_k}{N_k} = \frac{n}{N}$ constitue le **taux de sondage**, les échantillons prélevés dans les strates l'étant ainsi au *prorata des effectifs*. On parle en l'occurrence **d'échantillonnage stratifié représentatif**.

2-b) S'agissant **d'échantillonnage exhaustif**, il s'agit de trouver les valeurs des effectifs $n_h (1 \le h \le k)$ qui rendent minimale la fonction :

$$U(n_1, n_2, ..., n_k) = \sum_{h=1}^{h=k} \frac{N_h^2}{N^2}.\frac{N_h - n_h}{N_h - 1}.\frac{\sigma_h^2}{n_h},$$

les variables $n_1, n_2, ..., n_k$ étant liées par la relation $n_1 + n_2 .... + n_k = n$.

On notera que $U(n_1, n_2, ..., n_k) = Var(\overline{X}) = E\left[(X - m)^2\right]$ représente la précision de l'estimation puisque c'est aussi la norme $\|X - m\|^2$, c'est-à-dire le carré de la distance entre l'estimateur de la moyenne et cette dernière sur l'ensemble de la population (ce sujet est développé dans le chapitre suivant).

• Selon la **méthode des multiplicateurs de LAGRANGE**, la fonction auxiliaire définie par $V(n_1, n_2, ..., n_k) = U(n_1, n_2, ..., n_k) + \lambda.\left[\sum_{h=1}^{h=k} n_h - n\right]$ a pour extremums les solutions des équations $\frac{\partial V}{\partial n_h} = 0 (1 \le h \le k)$. On obtient donc, pour h fixé, la relation :

$$\frac{N_h^2}{N^2.(N_h - 1)}.\left[-\frac{\sigma_h^2}{n_h} - \frac{(N_h - n_h).\sigma_h^2}{n_h^2}\right] + \lambda = 0$$

soit, après simplifications et pour tout $h / 1 \le h \le k$, l'expression :

$$n_h^2 = \frac{1}{\lambda}.\frac{N_h^2}{N^2}.\sigma_h^2.\frac{N_h}{N_h - 1} \quad (E)$$

Ecrivant que $\sum_{h=1}^{h=k} n_h = n$, il vient $\frac{1}{\sqrt{\lambda}}.\sum_{h=1}^{h=k} \frac{N_h.\sigma_h}{N}.\sqrt{\frac{N_h}{N_h - 1}} = n$, relation de laquelle on peut extraire la valeur de λ qui est une constante.

En fait, la relation précédente (E), se décline pour tout h, sous la forme :

$$\sqrt{\lambda} = \text{constante} = \frac{N_h}{N}.\frac{\sigma_h}{n_h}.\sqrt{\frac{N_h}{N_h - 1}}$$

relation qu'on peut écrire également $\frac{n_h}{N_h.\sigma_h}.\sqrt{\frac{N_h - 1}{N_h}}$ = constante, puisque N est une donnée déterminée. Ainsi est donc établie la condition d'optimalité cherchée, à savoir :

$$\frac{n_h}{N_h.\sigma_h}.\sqrt{\frac{N_h - 1}{N_h}} = \text{constante } (\forall h / 1 \le h \le k).$$

(L'écriture de la relation $n_1 + n_2 + ... + n_k = n$ permettra de déterminer la valeur la constante en question et à fortiori les valeurs des n_h comme cela est illustré dans la question suivante).

- Lorsqu'il s'agit d'un **échantillonnage non exhaustif**, on constatera aisément en réitérant le calcul précédent à partir de l'expression $Var(\overline{X}) = \sum_{h=1}^{h=k} \frac{N_h^2}{N^2}.\frac{\sigma_h^2}{n_h}$, que l'on obtient la condition d'optimalité :

$$\frac{n_h}{N_h.\sigma_h} = \text{constante } (\forall h / 1 \le h \le k).$$

C'est d'ailleurs, le résultat qu'on obtiendrait encore plus simplement, en faisant tendre N_h vers l'infini dans la relation obtenue pour un échantillonnage exhaustif. En effet, on sait que la loi hypergéométrique (tirages exhaustifs) converge vers la loi binomiale (tirages non exhaustifs) lorsque la taille de la population de référence est grande (mathématiquement $N \to +\infty$).

2-c) Toujours suivant la *méthode des multiplicateurs de* LAGRANGE, et d'abord dans l'hypothèse d'un **échantillonnage exhaustif**, l'optimisation de la fonction $Var(\overline{X}) = \sum_{h=1}^{h=k} \frac{N_h^2}{N^2}.\frac{N_h - n_h}{N_h - 1}.\frac{\sigma_h^2}{n_h}$ sous la contrainte $\sum_{h=1}^{h=k} c_h.n_h = C$, conduit, pour tout h, aux relations $\frac{N_h^2}{N^2.(N_h - 1)}.\left[-\frac{\sigma_h^2}{n_h} - \frac{(N_h - n_h).\sigma_h^2}{n_h^2}\right] + \lambda.c_h = 0 (\forall h / 1 \le h \le k)$. En simplifiant, on obtient $\frac{N_h^2.\sigma_h^2.N_h}{N^2.(N_h - 1).n_h^2} = \lambda.c_h$, soit la relation :

$$\sqrt{\frac{N_h - 1}{N_h}}.\frac{n_h}{N_h.\sigma_h / \sqrt{c_h}} = \text{constante } (\forall h / 1 \le h \le k).$$

- Pour des **échantillons non exhaustifs**, la relation ci-dessus s'écrit immédiatement :

$$\frac{n_h}{N_h.\sigma_h / \sqrt{c_h}} = \text{constante } (\forall h / 1 \le h \le k).$$

3°) Les conditions exprimées ci-dessus supposent les σ_h connus. Dans le cas contraire, on se contentera de leurs estimations fournies par $\widehat{S_h}$.

La déclinaison des calculs et résultats précédents menés dans la 2ème question, conduit au tableau ci-dessous, relativement aux données numériques qui sont proposées ici.

Objet	Variable	Strate 1	Strate 2	Strate 3	Strate 4	Σ
Données liées aux strates	N_h σ_h	105 6,35	40 5,52	75 8,75	80 34,53	300
Echantillons stratifiés représentatifs	$n_h = \frac{n}{N}.N_h$	21	8	15	16	60
	$(\frac{N_h}{N})^2.\frac{\sigma_h^2}{n_h}$	0,235	0,068	0,319	5,299	5,921 **(1)**
	$(\frac{N_h}{N})^2.\frac{N_h-n_h}{N_h-1}.\frac{\sigma_h^2}{n_h}$	0,190	0,055	0,258	4,293	4,797 **(2)**
Echantillons optimaux de NEYMAN non exhaustifs	$N_h.\sigma_h$	666,75	220,80	656,25	2762,4	4306,2
	$\frac{N_h.\sigma_h}{\sum_{h=1}^{h=4} N_h.\sigma_h}.\sum_{h=1}^{h=4} n_h$	9,29	3,07	9,15	38,48	
	n_h (après arrondi à l'entier le plus proche)	9	3	9	39	60
	$(\frac{N_h}{N})^2.\frac{\sigma_h^2}{n_h}$	0,549	0,181	0,532	2,17	3,435 **(3)**
Echantillons de NEYMAN exhaustifs	$N_h.\sigma_h.\sqrt{\frac{N_h}{N_h-1}}$	669,94	223,61	660,67	2779,83	4334,06
	n_h (après arrondi à l'entier le plus proche)	9	3	9	39	
	$(\frac{N_h}{N})^2.\frac{N_h-n_h}{N_h-1}.\frac{\sigma_h^2}{n_h}$	0,507	0,171	0,474	1,128	2,280 **(4)**

En résumé, la précision de l'estimation, qui est caractérisée par $Var(\overline{X})$ prend les valeurs (1), (2), (3), (4) susmentionnées et rassemblées ci-dessous :

Mode d'échantillonnage	$Var(\overline{X})$
Echantillons stratifiés représentatifs et non exhaustifs	5,921
Echantillons stratifiés représentatifs et exhaustifs	4,797
Echantillons stratifiés optimaux de NEYMAN et non exhaustifs	3,435
Echantillons stratifiés optimaux de NEYMAN et exhaustifs	2,280

Les résultats obtenus confirment la *légère supériorité des tirages exhaustifs sur ceux qui sont non exhaustifs*. Par ailleurs, ils montrent que *l'échantillonnage optimal* selon NEYMAN est bien *plus précis que l'échantillonnage stratifié suivant taux de sondage uniforme*. Bien évidemment, *la stratification demeure elle-même plus précise* (le plus souvent) *que l'échantillonnage élémentaire classique.*

On remarquera également que les *tailles* n_h des échantillons sont logiquement *croissantes avec les tailles* N_h des strates. *Mais l'écart- type est aussi un facteur important*. Plus les données sont dispersées, plus il faut prélever des informations. Ainsi la taille de l'échantillon de la strate n°4 est-elle très importante en comparaison avec celles des échantillons extraits des autres strates.

2.3 Les sondages à probabilités inégales (estimation d'nn total)

On se propose de sélectionner ici les unités entrant dans l'échantillon en fonction de probabilités d'inclusion proportionnelles à l'intérêt qu'elles présentent en termes d'information sur le caractère étudié.

Enoncé : On considère une population U de taille N au sein de laquelle un certain caractère aléatoire X prend la valeur X_k pour chacun des éléments k de ladite population. On note par t_X la somme des valeurs de X sur l'ensemble des éléments de la population, soit $t_X = \sum_{k=1}^{k=n} X_k$. On note par π_k la probabilité d'inclusion de l'élément $k / 1 \le k \le N$, dans l'échantillon formé que l'on notera S. Plus largement, on notera par π_{kl} la probabilité d'inclusion de deux éléments k et l de la population U considérée, dans l'échantillon S formé.

1°) Associant à chaque élément k, une variable indicatrice ε_k égale à 1 si l'élément est choisi dans l'échantillon et à 0 sinon (variable introduite par CORNFIELD « 1944 »), caractériser la loi de probabilité de la variable aléatoire ε_k puis montrer que $\text{cov}(\varepsilon_k, \varepsilon_l) = \pi_{kl} - \pi_k.\pi_l, \forall (i,j) / 1 \le i \le N, 1 \le j \le N$.

2°) On considère l'estimateur $\widehat{t_X}$ de t_X de HORVITZ et THOMPSON (1952) caractérisé par $\widehat{t_X} = \sum_{\epsilon \in S} \frac{X_k}{\pi_k}$ (k désignant ici les indices des éléments retenus dans l'échantillon S).

2-a) Supposant que toutes les probabilités d'inclusion soient non nulles, montrer que $E\left[\widehat{t_X}\right] = t_X$.

2-b) Toujours dans l'hypothèse $\pi_k \succ 0, \forall k \in U$, établir que :

$$Var(\widehat{t_X}) = \sum_{k \in U} \sum_{l \in U} \frac{X_k}{\pi_k} . \frac{X_l}{\pi_l} . \text{cov}(\varepsilon_k, \varepsilon_l)$$

3°) On considère une population de quatre entreprises A,B,C,D comptant respectivement 800, 200, 50, et 30 salariés. Dans cette population on veut estimer le nombre de salariés à partir d'échantillons de taille 2.

3-a) On effectue tout d'abord un sondage aléatoire élémentaire sans remise. Après avoir dénombré le nombre d'échantillons possibles (on pourra se reporter à l'application 1.6 du présent chapitre), évaluer la moyenne et la variance des résultats possibles quant à l'estimation du nombre total des salariés sur la population en question.

3-b) On forme un plan de sondage à probabilités inégales, π_k, définies de sorte que leur somme soit égale à la taille de l'échantillon, soit $n=2$, et qu'elles soient représentatives du degré d'inclusion de chacune des quatre entreprises A, B, C, D, dans l'échantillon formé. Ainsi, le choix $\pi_k=1$ signifie que l'on souhaite nécessairement la présence de A dans tout échantillon de taille 2 formé à partir de la population des quatre entreprises considérées. Les données sont les suivantes :

Entreprise k	Effectif des salariés	Probabilité d'inclusion π_k
A	800	1
B	200	0,5
C	50	0,3
D	30	0,2

Pour le plan considéré, toujours exhaustif, évaluer la moyenne et la variance des résultats possibles quant à l'estimation du nombre total de salariés sur la population en question. Qu'en conclure ?

Solution : 1°) Pour chaque élément $k \in U$, la *variable indicatrice,* soit ε_k, est caractérisée par les valeurs 1 et 0 de probabilités associées π_k et $1-\pi_k$. Soit $(\varepsilon_k,\varepsilon_l)$ un couple de variables indicatrices, on a $\text{cov}(\varepsilon_k,\varepsilon_l)=E[\varepsilon_k.\varepsilon_l]-E(\varepsilon_k).E(\varepsilon_l)$, expression dans laquelle

$E[\varepsilon_k.\varepsilon_l]=\sum_{k\in U}\sum_{l\in U}u_k.v_l.\text{Prob}(\varepsilon_k=u_k,\varepsilon_l=v_l)$.

Dans la mesure où, parmi les quatre valeurs possibles du couple $(\varepsilon_k,\varepsilon_l)$, trois d'entre elles contiennent au moins un zéro, la double somme ci-dessus, se réduit au seul terme $E[\varepsilon_k.\varepsilon_l]=\text{Prob}(\varepsilon_k=1,\varepsilon_l=1)=\pi_{kl}$.

On a par ailleurs, $E(\varepsilon_k)=1.\pi_k+0.(1-\pi_k)=\pi_k, \forall k\in U$. En conclusion, on a donc $\text{cov}(\varepsilon_k,\varepsilon_l)=\pi_{kl}-\pi_k.\pi_l$ qui est le résultat cherché.

2-a) $E\left[\widehat{t_X}\right]=E\left[\sum_{k\in S}\frac{X_k}{\pi_k}\right]=E\left[\sum_{k\in U}\frac{X_k.\varepsilon_k}{\pi_k}\right]$. Par *linéarité de l'espérance mathématique* et remarquant que pour k fixé appartenant à U, X_k est déterminé (et donc *non aléatoire*), il en résulte $E\left[\widehat{t_X}\right]=\sum_{k\in U}\frac{X_k}{\pi_k}.E(\varepsilon_k)=\sum_{k\in U}\frac{X_k}{\pi_k}.\pi_k=\sum_{k\in U}X_k=t_X$ .Anticipant le chapitre II, l'estimateur $\widehat{t_X}$ dont , en moyenne, la valeur est celle du total t_X sur l'ensemble de la population U est dit « **sans biais** ».

2-b) $Var(\widehat{t_X})=Var(\sum_{k\in S}\frac{X_k}{\pi_k})=Var(\sum_{k\in U}\frac{X_k.\varepsilon_k}{\pi_k})$. De façon générale, on sait que pour toutes variables aléatoires U_i, $Var(\sum_{i=1}^{i=n}U_i)=\sum_{i=1}^{i=n}Var(U_i)+\sum_{\substack{i=1\\i\neq j}}^{i=n}\sum_{j=1}^{j=n}\text{cov}(U_i,U_j)$.

S'agissant de $Var(\widehat{t_x})$, on a donc :

$$Var(\widehat{t_X})=\sum_{k\in U}Var(\frac{X_k.\varepsilon_k}{\pi_k})+\sum_{\substack{k\in U\\k\neq l}}\sum_{l\in U}\text{cov}(\frac{X_k.\varepsilon_k}{\pi_k},\frac{X_l.\varepsilon_l}{\pi_l}).$$

• Or pour tout $k \in U$, $Var(\frac{X_k.\varepsilon_k}{\pi_k}) = (\frac{X_k}{\pi_k})^2.Var(\varepsilon_k) = (\frac{X_k}{\pi_k})^2.\pi_k.(1-\pi_k)$. En effet, $Var(\varepsilon_k) = E\left[\varepsilon_k^2\right] - E(\varepsilon_k)^2$ avec $E(\varepsilon_k^2) = 1^2\pi_k + 0^2.(1-\pi_k) = \pi_k$ et $E(\varepsilon_k) = \pi_k$. Ainsi a-t-on, $Var(\varepsilon_k) = \pi_k - \pi_k^2 = \pi_k.(1-\pi_k)$.

• Par ailleurs, $\forall\ (a,b)$ scalaires réels, $\text{cov}(a.X, b.Y) = E\left[(a.X).(b.Y)\right] - E(a.X).E(b.Y)$, soit par *linéarité de l'espérance mathématique* et après factorisation, $\text{cov}(a.X, b.Y) = a.b.\left[E(X.Y) - E(X).E(Y)\right] = a.b.\text{cov}(X,Y)$.

On peut donc écrire, $\text{cov}(\frac{X_k.\varepsilon_k}{\pi_k}, \frac{X_l.\varepsilon_l}{\pi_l}) = \frac{X_k}{\pi_k}.\frac{X_l}{\pi_l}.\text{cov}(\varepsilon_k, \varepsilon_l)$.

• En résumé, $Var(\widehat{t_X}) = \sum_{k\in U}(\frac{X_k}{\pi_k})^2.\pi_k.(1-\pi_k) + \sum_{\substack{k\in U \\ k\neq l}}\sum_{l\in U}\frac{X_k}{\pi_k}.\frac{X_l}{\pi_l}.\text{cov}(\varepsilon_k, \varepsilon_l)$. Remarquant que $\text{cov}(\varepsilon_k, \varepsilon_k) = Var(\varepsilon_k)$, on peut écrire $Var(\widehat{t_X})$ sous la forme synthétique ci-dessous :

$$Var(\widehat{t_X}) = \sum_{k\in U}\sum_{l\in U}(\frac{X_k}{\pi_k}.\frac{X_l}{\pi_l}).\text{cov}(\varepsilon_k, \varepsilon_l)$$

(avec, pour rappel de la 1ère question, $\text{cov}(\varepsilon_k, \varepsilon_l) = \pi_{kl} - \pi_k.\pi_l$. On notera en outre, que $Var(\varepsilon_k) = \text{cov}(\varepsilon_k, \varepsilon_k) = \pi_{kk} - \pi_k^2 = \pi_k - \pi_k^2 = \pi_k.(1-\pi_k)$).

3-a) Le nombre d'échantillons de taille 2 qu'on peut prélever sans remise de la population considérée de taille 4 est immédiatement $C_4^2 = 6$. Ces échantillons et l'estimateur du nombre des salariés qu'ils génèrent sont les suivants :

Echantillon	Probabilité de tirage	Estimation du nombre total de salariés pour la population considérée (*)
(A, B)	1/6	2000
(A, C)	1/6	1700
(A, D)	1/6	1660
(B, C)	1/6	500
(B, D)	1/6	460
(C, D)	1/6	160

(*) On estime ici le nombre total de salariés dans toute la population par *effet multiplicateur* entre le nombre total d'entreprises dans la population et le nombre total d'entreprises considérées dans l'échantillon. Plus précisément, ce coefficient multiplicateur qui est égal à $\frac{N}{n}$ a pour valeur 2 dans le cas présent. Ainsi, pour l'exemple du premier échantillon (A, B) dont l'effectif est $800 + 200 = 1000$ salariés, l'estimation qui en résulte, pour ce qui est des quatre entreprises qui forment la population considérée, est égale à 2000 salariés, et ainsi de suite....

- La série précédente a pour *espérance mathématique*, la valeur :

$$\frac{1}{6}.(2000+1700+....+160)=1080.$$

Cette valeur est aussi le nombre total de salariés sur l'ensemble de la population. On retrouve donc ici la propriété des **estimateurs sans biais** (cf. chapitre II), à savoir la relation $E(\widehat{t_X})=t_X$.

Mais cet estimateur est *très dispersé* puisque sa variance est égale à :

$$\frac{1}{6}.\left[(2000-1080)^2+(1700-1080)^2+....+(160-1080)^2\right]=522400.$$

3-b) Dans l'hypothèse d'un plan à probabilités inégales et tenant compte de la nécessaire inclusion de A dans l'échantillon formé ($\pi_1=1$), le nombre d'échantillons possibles n'est plus que de trois et génère suivant l'expression $\widehat{t_X}=\sum_{k\in S}\frac{X_k}{\pi_k}$ de HORVITZ et THOMPSON. les estimations correspondantes ci-dessous :

Echantillon	Estimation	Probabilité d'obtention de l'échantillon (*)
(A, B)	$\frac{800}{1}+\frac{200}{0,5}=1200$	0,5
(A, C)	$\frac{800}{1}+\frac{50}{0,3}=966,666$	0,3
(A, D)	$\frac{800}{1}+\frac{30}{0,2}=950$	0,2

(*) En effet, A étant d'ores et déjà choisi dans l'échantillon, les probabilités d'obtention des trois échantillons possibles sont celles associées aux entreprises B, C, et D, soient respectivement 0,5, 0,3, et 0,2.

- On constate que l'estimateur de HORVITZ et THOMPSON est lui aussi *sans biais* puisqu'on a $E\left[\widehat{t_X}\right]=1200\times0,5+966,666\times0,3+950\times0,2=1080$. Quant à la variance, elle est égale à $0,5\times(1200-1080)^2+0,3\times(966,666-1080)^2+0,2\times(950-1080)^2=14433,33$

C'est *nettement plus faible* que la variance de l'estimateur par prélèvements élémentaires sans remise calculé précédemment et dont il est rappelé que la valeur était égale à 522400.

- On pourra montrer que le résultat obtenu ci-dessus, est aussi celui qu'on obtient par la formule générale $Var(\widehat{t_X})=\sum_{k\in U}\sum_{l\in U}\frac{X_k.X_l}{\pi_k.\pi_l}.\left[\pi_{kl}-\pi_k.\pi_l\right]$. Les calculs correspondants sont développés ci-après, les valeurs de π_{kl} et de $\frac{\pi_{kl}-\pi_k.\pi_l}{\pi_k.\pi_l}$ étant préalablement portées dans le tableau de la page suivante.

$\pi_{kl} \setminus \frac{\pi_{kl}-\pi_k.\pi_l}{\pi_k.\pi_l}$	A	B	C	D
A	1 \ 0	0,5 \ 0	0,3 \ 0	0,2 \ 0
B	0,5 \ 0	0,5 \ 1	0 \ -1	0 \ -1
C	0,3 \ 0	0 \ -1	0,3 \2,33	0 \ -1
D	0,2 \ 0	0 \ -1	0 \ -1	0,2 \ 4

Il s'ensuit $Var(\widehat{t_X}) = (200)^2 \times 1 - (200\times 50) - (200\times 30) - (200\times 50) + 2,333\times(50)^2$ $-(50\times 30) - (200\times 30) - (50\times 30) + 4\times(30)^2 = 49433,33 - 35000 = 14433,33$. On retrouve ainsi le même résultat.

• Comme on a pu le constater, le *plan de sondage à probabilités inégales* a conduit ici à une *précision nettement supérieure* à celle d'un *sondage simple.* Il faut cependant remarquer que le même exercice avec des probabilités de tirage respectivement égales pour A, B, C, D à 0,2, 0,3, 0,5, et 1 conduirait, par contre, à des résultats très mauvais et nettement supérieurs à 522400 (précision du sondage élémentaire).

En fait, il est primordial de lier la probabilité d'inclusion π_k aux valeurs prises quant au caractère étudié X et on peut montrer d'ailleurs que $Var(\widehat{t_X})$ est minimale lorsqu'on choisit les π_k proportionnelles à la valeur du caractère étudié, pour chacune des strates de la population.

C - Exercices complémentaires

1. Les durées de vie moyenne de tubes de télévision ont pour valeur 3000 h avec un écart-type de 70 h, lesdites durées de vie étant supposées indépendantes et de loi normale. Si on prend au hasard dix de ces tubes, trouver la probabilité que l'écart-type de l'échantillon obtenu soit compris entre 60 et 80 h.

Solution : Soient X, la durée de vie aléatoire d'un tube de télévision, et (m,σ) les paramètres représentatifs associés ($m = 3000, \sigma = 70$). Notant par X_i les durées de vie des tubes formant l'échantillon de taille n, l'écart- type de l'échantillon ou plutôt sa variance, est exprimée par la statistique $S^2 = \frac{1}{n}.\sum_{i=1}^{i=n}(X_i - m)^2$.

Or on sait que $\xi^2 = \frac{n.S^2}{\sigma^2}$ suit la **loi du chi- deux** à $\nu = n$ degrés de libertés. Dès lors, on peut écrire la probabilité proposée, à savoir, $\Pr ob(60 \le S \le 80)$, sous la forme équivalente $\Pr ob(\frac{10\times 60^2}{70^2} \le \xi^2 \le \frac{10\times 80^2}{70^2}) = \Pr ob(7,34 \le \xi^2 \le 13,06)$.

L'appel à un *calculateur en ligne disponible sur internet,* conduit, pour la variable du chi- deux à 10 degrés de libertés, aux évaluations $\Pr ob(\xi^2 \ge 7,34) = 0,693$ et $\Pr ob(\xi^2 \ge 13,06) = 0,22$. Il en résulte $\Pr ob(7,34 \le \xi^2 \le 13,06) = \Pr ob(\xi^2 \ge 7,34) - \Pr ob(\xi^2 \succ 13,06) = 0,473$.

2. Le responsable qualité d'une entreprise contrôle 20 objets dans chaque lot de 1000 objets, avant de laisser expédier aux clients, les lots en question. Il accepte seulement les lots pour lesquels il ne trouve aucun objet non conforme dans l'échantillon prélevé. Dans le cas contraire, le lot est trié, unité par unité.

Quelle est la probabilité pour qu'un lot contenant une proportion $p = 0,05$ d'objets non conformes soit acceptée ?

Solution : Désignant par X la variable aléatoire décrivant le nombre d'objets non conformes au sein d'un échantillon de taille $n = 20$ extrait sans remise d'un lot de taille $N = 1000$ objets, la question posée est d'évaluer la probabilité conditionnelle $\Pr ob(X = 0 / p = 0,05)$.

Or X suit la **loi hypergéométrique** caractérisée en l'occurrence par l'équation :

$$\Pr ob(X = x) = \frac{C_{N.p}^{x}.C_{N.q}^{n-x}}{C_N^n} \text{ (avec } p = 0,05, N = 1000, q = 0,95, n = 20.$$

Il en résulte numériquement, $\Pr ob(X = 0 / p = 0,05) = \frac{C_{950}^{20}}{C_n^{1000}} = 0,355$, suivant calcul effectué sur tableur et après développements et simplifications des C_n^{1000}.

3. Une entreprise fabrique des sacs en plastique pour les enseignes de distribution. Elle s'intéresse au poids maximal que ces sacs peuvent supporter sans se déchirer. On suppose que le poids maximal en question est une variable aléatoire suivant la loi normale de moyenne 5 kg et d'écart- type 1 kg.

1°) Sur 300 sacs reçus, une grande enseigne de distribution constate un poids moyen maximal de rupture égal à 4,91 kg.

1-a) Trouver un intervalle dans lequel la moyenne des poids maximaux constatés sur un échantillon de taille 300, se trouve comprise dans au moins 99% des cas.

1-b) Qu'en conclure pour ce qui est de l'observation constatée ci-dessus, à savoir la valeur 4,91 kg ?

2°) Déterminer le poids moyen dépassé dans 97% des cas, sur un échantillon de taille $n = 300$.

Solution : 1-a) Désignant par X le poids maximal aléatoire qu'un sac en plastique est susceptible de supporter, il résulte de l'énoncé, que X suit la *loi normale* de moyenne $m = 5$ kg et $\sigma = 1$ kg, soit la loi $N(m = 5, \sigma = 1)$.

Formant la moyenne $\overline{X} = \frac{\sum_{i=1}^{i=n} X_i}{n}$ des limites maximales de résistance constatées dans un échantillon de taille n ($n = 300$), on sait que $\overline{X}$ suit la loi normale $N(m, \frac{\sigma}{\sqrt{n}})$ (cf. rappels de cours). Numériquement, on obtient donc pour $\overline{X}$, la loi normale $N(m = 5, \sigma = 0,0577)$.

- Considérant la variable aléatoire, normale, centrée, réduite, associée à X, soit $\xi = \frac{X - 5}{0,0577}$, on sait associer à tout $\alpha / 0 \prec \alpha \prec 1$, le nombre $t_\alpha / \Pr ob(-t_\alpha \le \xi \le t_\alpha) = \alpha$. Plus précisément et en utilisant la fonction de répartition $\Pi(t) = \Pr ob(\xi \le t)$ dont la table des valeurs est annexée, on a $\Pr ob(-t_\alpha \le \xi \le t_\alpha) = \Pi(t_\alpha) - \Pi(-t_\alpha) = 2.\Pi(t_\alpha) - 1$.

Il s'ensuit pour $\overline{X}$, l'encadrement $5-t_\alpha.0,0577 \le \overline{X} \le 5+t_\alpha.0,0577$. Dans le cas particulier où $\alpha = 0,99$, on obtient $2.\Pi(t_\alpha)-1=0,99 \Rightarrow \Pi(t_\alpha)=0,995$, ce qui entraîne par lecture dans la table, $t_\alpha = 2,57$. Ainsi a-t-on dans au moins 99% des cas, l'encadrement ci-dessous de $\overline{X}$:

$$4,852 \le \overline{X} \le 5,148 \quad (\textit{intervalle dit} \text{ « } \textbf{de confiance} \text{ » au seuil 99\%}).$$

1-b) Pour ce qui est de la valeur constatée, à savoir 4,91 kg, elle se trouve comprise dans l'intervalle de confiance susmentionné au seuil 99%. Autrement dit, l'échantillon en question se trouve être conforme aux attentes.

2°) Dans cette question, on cherche le poids moyen c qui, dans 97% des cas, est dépassé, la base du calcul étant toujours celle d'échantillons de taille $n=300$. Il faut donc résoudre l'équation $\Pr ob(\overline{X} \succ c)=0,97$ où $\overline{X}$ suit la loi normale $N(m=5,\sigma=0,0577)$.

Par passage à la variable aléatoire normale, centrée, réduite, associée à $\overline{X}$, soit $\xi = \dfrac{\overline{X}-5}{0,0577}$, il vient $\Pr ob(\xi \succ \dfrac{c-5}{0,0577}) = 0,97$. Désignant par t_α le nombre tel que $\Pr ob(\xi \succ t_\alpha)=0,97$, on obtient par introduction de la fonction de répartition $\Pi(t)=\Pr ob(\xi \le t)$, $\Pr ob(\xi \le t_\alpha)=0,03=\Pr ob(\xi \ge -t_\alpha)=1-\Pi(-t_\alpha)$.

La lecture dans la table des valeurs de $\Pi(t)$ du nombre $-t_\alpha$ vérifiant $\Pi(-t_\alpha)=0,97$ fournit immédiatement $-t_\alpha = 1,88$. En définitive, $\dfrac{c-5}{0,0577}=-1,88 \Rightarrow c=4,891$.

Le poids moyen des échantillons qui est dépassé dans 97% des cas est donc 4,891 kg.

4. Les résultats d'une enquête sur une population de 1000 salariés d'une entreprise a montré que dans 60% des cas, les agents employés avaient au moins un crédit en cours.

Trouver la probabilité pour que deux échantillons exhaustifs de 300 agents chacun, indiquent plus de 5 points (c'est-à-dire 5%) d'écart pour ce qui est de la proportion de des employés qui ont au moins un crédit en cours.

Solution : 1-a) L'exercice proposé est très proche de l'application 1.5 du présent chapitre. Toutefois, on note ici que la taille N de la population de référence n'est pas grande par rapport à la taille de l'échantillon n (en tout cas, on est loin de vérifier $\dfrac{N}{n}=\dfrac{1000}{300} \ge 10$, qui est le critère de convergence de la loi hypergéométrique vers la loi binomiale proposé en rappels de cours).

Dans la mesure où on part *d'échantillons exhaustifs*, il faut donc utiliser le **coefficient correcteur d'exhaustivité** $\dfrac{N-n}{N-1}$ portant sur les variances, coefficient mis en évidence dans les rappels de cours et établi dans l'application 2.1.

Reprenant les résultats des rappels de cours (paragraphe 5), la proportion des personnes qui, dans un échantillon de taille $n=300$, ont au moins un crédit en cours, est décrite par la statistique $F_n=\dfrac{X}{n}$ où, puisque n est grand ($n \ge 30$), la loi de F_n converge vers la loi normale de moyenne p et de variance $\dfrac{p.q}{n}$, ou plutôt $\dfrac{N-n}{N-1}.\dfrac{p.q}{n}$ (puisqu'il faut appliquer le *coefficient correcteur d'exhaustivité*).

Numériquement, F_n converge vers la loi normale de moyenne 0,6 et de variance égale à $\frac{1000-300}{1000-1}.\frac{0,6\times 0,4}{300}=0,00056056$, soit la loi normale $N(0,6,\sigma=0,023676)$.

• *Comparant deux échantillons de ce type* et considérant les statistiques associées $F_{n,1}$ et $F_{n,2}$, il en résulte, pour la caractérisation de la loi de la différence $F_{n,1}$ - $F_{n,2}$, *la loi normale* de moyenne $E(F_{n,1}-F_{n,2})=E(F_{n,1})-E(F_{n,2})=0$ et de variance $Var(F_{n,1}-F_{n,2})=Var(F_{n,1})+Var(F_{n,2})$ soit numériquement $Var(F_{n,1}-F_{n,2})=0,00112112$, ou encore $\sigma_{F_{n,1}-F_{n,2}}=0,033483$ (on se reportera aux résultats de l'application 1.5 pour plus de justifications sur cette conclusion).

• La question posée, consiste à évaluer la probabilité $\Pr ob(|F_{n,1}-F_{n,2}|\succ 0,05)$. Considérant la variable normale $F_{n,1}$ - $F_{n,2}$ de loi $N(0,\sigma=0,033483)$ et formant la *variable normale, centrée, réduite*, associée $\xi=\frac{F_{n,1}-F_{n,2}-0}{0,033483}$, il s'agit donc d'évaluer $\Pr ob(|\xi|\succ \frac{0,05}{0,033483}=1,493)$, soit $1-\Pr ob(|\xi|\leq 1,493)$. Se référant à la fonction de répartition $\Pi(t)=\Pr ob(\xi\leq t)$ et à sa table de valeurs (cf. annexes), il vient $\Pr ob(|\xi|\leq 1,493)=2.\Pi(1,493)-1=0,864$.

D'où, en conclusion, l'évaluation cherchée, $\Pr ob(|F_{n,1}-F_{n,2}|\succ 0,05)=\Pr ob(|\xi|\succ 1,493)$. soit 1-0,864= 0, 136.

5. Sur les 20000 agents d'une grande collectivité, on souhaite connaître la proportion p d'entre eux qui possèdent au moins un bien immobilier. Pour chaque individu de la base de sondage, on dispose de la valeur du revenu. On décide alors de constituer trois strates dans la population de référence, strates formées par les cadres A (strate n°1), B (strate n°2), et C (strate n°3).

On note par N_h la taille de la strate h et par $\widehat{p_h}$ l'estimateur de la proportion d'agents qui possèdent au moins un bien immobilier dans la strate h. Les données sont ainsi rassemblées ci-dessous :

	Strate 1	Strate 2	Strate 3
N_h	2000	4000	14000
n_h	150	150	700
$\widehat{p_h}$	0,60	0,45	0,30

1°) Quel estimateur $\widehat{p}$ de p peut-on choisir ici ? Est-il sans biais ($E(\widehat{p})=p$) ?

2°) Déterminer la précision de $\widehat{p}$.

3°) En déduire un intervalle dans lequel p est contenu dans au moins 95% des cas (intervalle dit « de confiance »).

Solution : 1°) Se référant aux résultats de l'application 2.2 du présent chapitre et assimilant la proportion p à une moyenne, il est immédiat que l'estimateur $\widehat{p}$ de p est fourni par la statistique $\widehat{p}=\sum_{h=1}^{h=3}\frac{N_h}{N}.\widehat{p_h}$.

Formant $E(\widehat{p})$, on obtient par *linéarité de l'espérance mathématique*, $E(\widehat{p}) = \sum_{h=1}^{h=3} \frac{N_h}{N} . E(\widehat{p_h})$.

Or $E(\widehat{p_h}) = p_h$ (suivant rappels de cours- paragraphe 5). Ainsi $E(\widehat{p}) = \sum_{h=1}^{h=3} \frac{N_h}{N} . p_h = p$, ce qui montre que l'estimateur $\widehat{p}$ est **sans biais**.

2°) La précision de cet estimateur est fournie par l'écart entre ce dernier et la valeur inconnue p qu'il approche, soit au sens de la **norme** habituelle $\left\| \widehat{p} - p \right\| = \sqrt{E\left[(\widehat{p} - p)^2 \right]}$. Ainsi, au carré près, cette précision est-elle décrite par la variance $Var(\widehat{p})$.

- L'indépendance entre strates entraîne immédiatement le développement :

$$Var(\widehat{p_h}) = \sum_{h=1}^{h=3} (\frac{N_h}{N})^2 . Var(\widehat{p_h}) = \sum_{h=1}^{h=3} (\frac{N_h}{N})^2 . \frac{N_h - n_h}{N_h - 1} . \frac{p_h . q_h}{n_h} \text{ (avec } q_h = 1 - p_h).$$

On ne précise pas ici, l'exhaustivité ou non des prélèvements, mais au regard de la satisfaction des conditions $\frac{N_h}{n_h} \geq 10 (\forall h / 1 \leq h \leq 3)$, on peut admettre la *convergence* de $\frac{N_h - n_h}{N_h - 1}$ *vers* 1 (*hypothèse des tirages non exhaustifs*, régis par la loi binomiale). Ainsi, pourra-t-on admettre :

$$Var(\widehat{p}) = \sum_{h=1}^{h=3} (\frac{N_h}{N})^2 . \frac{p_h . q_h}{n_h} .$$

A défaut de connaître précisément les (p_h, q_h), on pourra utiliser entre autres, leur approximation par les estimateurs $(\widehat{p_h}, \widehat{q_h})$. Numériquement, on obtient :

$$Var(\widehat{p}) = (\frac{2000}{20000})^2 . \frac{0,6 \times 0,4}{150} + (\frac{4000}{20000})^2 . \frac{0,45 \times 0,55}{150} + (\frac{14000}{20000})^2 . \frac{0,3 \times 0,7}{700},$$

soit après calculs, $Var(\widehat{p}) = 0,000229$.

3°) On peut admettre la convergence de chacun des estimateurs $\widehat{p_h}$ vers la loi normale (car les n_h sont grands). Il en résulte que $\widehat{p}$ suit aussi *la loi normale* dont en fonction des questions précédentes, espérance mathématique et variance sont respectivement égaux à p et 0,000229 $\Rightarrow \sigma = 0,01513$.

Considérant la variable normale, centrée, réduite, associée à $\widehat{p}$, soit $\xi = \frac{\widehat{p} - p}{0,01513}$, on peut écrire pour tout $\alpha / 0 \prec \alpha \prec 1$, $\Pr ob(-t_\alpha \leq \xi \leq t_\alpha) = 2.\Pi(t_\alpha) - 1 = \alpha$. Choisissant ainsi la valeur $\alpha = 0,95 \Rightarrow \Pi(t_\alpha) = 0,975 \Rightarrow t_\alpha = 1,96$, on obtient **l'intervalle de confiance** de ξ au seuil de confiance $\alpha = 95\%$, à savoir $-1,96 \leq \xi = \frac{\widehat{p} - p}{0,01513} \leq +1,96$. D'où, l'encadrement cherché de p, défini par $\widehat{p} - 1,96 \times 0,01513 \leq p \leq \widehat{p} + 1,96 \times 0,01513$ avec $\widehat{p} = \sum_{h=1}^{h=3} \frac{N_h}{N} . \widehat{p_h}$, c'est-à-dire $\widehat{p} = (\frac{2000}{20000}) \times 0,60 + (\frac{4000}{20000}) \times 0,45 + (\frac{14000}{20000}) \times 0,30 = 0,36$. Nunériquement, *l'intervalle de confiance* cherché pour p est donc $0,33 \leq p \leq 0,39$.

6. Dans cet exercice, il est montré qu'une stratégie optimale pour estimer une quantité inconnue dans l'ensemble d'une population stratifiée, ne l'est plus tout à fait si l'objectif est autre, tel comparer les strates entre elles. Il convient donc de définir précisément en amont ce qu'on recherche avant d'opter pour la technique employée.

On considère une population de taille N formée de deux strates de tailles N_1 et N_2 et on s'intéresse ici à la moyenne inconnue m d'un caractère X, les moyennes inconnues au sein de chacune des deux strates étant respectivement égales à m_1 et m_2. On note par $\overline{X_1}$ et par $\overline{X_2}$ les estimateurs respectifs de m_1 et m_2, l'estimateur de m étant noté $\overline{X}$, quant à lui.

On suppose enfin que la variance de X est la même dans chaque strate, à savoir $\sigma_1^2 = \sigma_2^2 = \sigma^2$.

On dispose d'un budget C et on suppose que les prélèvements effectués sont des sondages aléatoires élémentaires sans remise, respectivement de tailles n_1 et n_2 au sein des deux strates considérées. On note par $C_1.n_1 + C_2.n_2$ la fonction de coût du sondage, C_1 et C_2 désignant les coûts unitaires respectifs associés aux deux strates.

1°) Exprimer $\overline{X}$ en fonction de $\overline{X_1}$ et de $\overline{X_2}$ et évaluer $E(\overline{X})$ et $Var(\overline{X})$.

2°) On construit ici un échantillon optimum à coût constant.

2-a) Quelle répartition (n_1^*, n_2^*) de l'échantillon donne une variance $Var(\overline{X})$ minimale ?

2-b) Calculer $Var(\overline{X})$ dans ces conditions optimales, pour $N_1 = 10000, N_2 = 20000$, $S_1 = 2, S_2 = 2, C_1 = 4, C_2 = 9, C = 1000$.

3°) On reprend ici le même problème, mais suivant une allocation proportionnelle, c'est-à-dire $N_h/n_h = N/n, \forall h$ (taux de sondage uniforme).

3-a) Déterminer n_1, n_2, et n.

3-b) Reprenant les données numériques précédentes, calculer $Var(\overline{X})$ et évaluer la perte relative de précision par raport à l'échantillon optimum de la 2ème question.

4°) En fait, on se propose plutôt d'évaluer l'écart entre les moyennes m_1 et m_2 des deux strates.

4-a) Montrer que $\overline{X_1} - \overline{X_2}$ est un estimateur sans biais de $m_1 - m_2$ et calculer $Var(\overline{X_1} - \overline{X_2})$.

4-b) Déterminer la répartition optimale (n_1^*, n_2^*) de l'échantillon pour que $Var(\overline{X_1} - \overline{X_2})$ soit minimale toujours avec la même contrainte de budget.

4-c) Reprenant les données numériques des questions antérieures, calculer $Var(\overline{X})$ pour la répartition optimale de la question 4-b) et comparer avec les résultats obtenus par les méthodes antérieures.

Solution : 1°) Reprenant les résultats de l'application 2.2 relative aux sondages par stratification, on a immédiatement $\overline{X} = \frac{N_1}{N}.\overline{X_1} + \frac{N_2}{N}.\overline{X_2}$. Il s'ensuit, par pseudo- linéarité de la variance,

$$Var(\overline{X}) = (\frac{N_1}{N})^2.Var(\overline{X_1}) + (\frac{N_2}{N})^2.Var(\overline{X_2}) = (\frac{N_1}{N})^2.\frac{N_1 - n_1}{N_1 - 1}.\frac{\sigma^2}{n_1} + (\frac{N_2}{N})^2.\frac{N_2 - n_2}{N_2 - 1}.\frac{\sigma^2}{n_2}.$$

2-a) Considérant que N_1 et N_2 sont grands devant n_1 et n_2 et qu'on a donc $\frac{N-n}{N-1} \to 1$, les résultats obtenus dans l'application 2.2 conduisent, pour la recherche des échantillons de tailles optimales à coût total constant, aux conditions :

$$\frac{n_1^*}{N_1.\sigma/\sqrt{C_1}} = \frac{n_2^*}{N_2.\sigma/\sqrt{C_2}} = \lambda = \text{constante.}$$

Posant $K = \lambda.\sigma$, on en déduit $n_1^* = K.\frac{N_1}{\sqrt{C_1}}, n_2^* = K.\frac{N_2}{\sqrt{C_2}}$, la relation $C_1.n_1^* + C_2.n_2^* = C$ entraînant immédiatement $K = \frac{C}{N_1.\sqrt{C_1} + N_2.\sqrt{C_2}}$. D'où, la valeur de K et à fortiori, les valeurs des tailles optimales n_1^* et n_2^*.

2-b) Suivant l'application numérique proposée, on a successivement $K = \frac{1000}{10000.\sqrt{4} + 20000.\sqrt{9}} = \frac{1}{80}$, $n_1^* = \frac{1}{80} \times \frac{10000}{\sqrt{4}} = 62,5$, et $n_2^* = \frac{1}{80} \times \frac{20000}{\sqrt{9}} = 83,3$, soit en arrondissant et tout en respectant la contrainte d'un coût total inférieur à 1000, les valeurs $n_1^* = 62, n_2^* = 83$.

Pour ces valeurs, on a donc en supposant $\frac{N_1 - n_1^*}{N_1 - 1}$ et $\frac{N_2 - n_2^*}{N_2 - 1}$ équivalents à l'unité, l'évaluation $Var(\overline{X}) = (\frac{1}{3})^2.\frac{\sigma^2}{62} + (\frac{2}{3})^2.\frac{\sigma^2}{83} = (\frac{1}{9 \times 62} + \frac{4}{9 \times 83}).\sigma^2 = 0,007147.\sigma^2$.

3-a) *L'allocation proportionnelle* conduit à la relation $\frac{n_1}{N_1} = \frac{n_2}{N_2} = K \Rightarrow n_1 = K.N_1, n_2 = K.N_2$. Ecrivant que $C_1.n_1 + C_2.n_2 = C$, il vient $K.(C_1.N_1 + C_2.N_2) = C$, d'où les valeurs de $K, n_1, n_2, ..., Var(\overline{X})$.

3-b) Numériquement, on a successivement $K = \frac{1000}{(4 \times 10000) + (9 \times 20000)} = \frac{1}{220}$, puis $n_1 = \frac{10000}{220} = 45,45$ et $n_2 = \frac{20000}{220} = 90,90$. Suivant arrondis et respectant de nouveau la contrainte d'un coût total inférieur à 1000, on obtient $n_1 = 45, n_2 = 91$.

Dans ces conditions, $Var(\overline{X}) = (\frac{1}{3})^2.\frac{\sigma^2}{45} + (\frac{2}{3})^2.\frac{\sigma^2}{91} = (\frac{1}{9 \times 45} + \frac{4}{9 \times 91}).\sigma^2 = 0,007353.\sigma^2$.

La *perte de précision* est relativement faible ici puisqu'entre l'échantillon optimal de la 2ème question et l'allocation proportionnelle, elle est de $\frac{7353 - 7147}{7147} = 2,9\%$.

4-a) Il est bien évident que $E\left[\overline{X_1}-\overline{X_2}\right]=E(\overline{X_1})-E(\overline{X_2})=m_1-m_2$. Par ailleurs, on a aussi $Var(\overline{X_1}-\overline{X_2})=Var(\overline{X_1})+Var(\overline{X_2})=\frac{\sigma^2}{n_1}+\frac{\sigma^2}{n_2}$ (toujours en supposant $\frac{N_1-n_1}{N_1-1}$ et $\frac{N_2-n_2}{N_2-1}$ proches de 1).

4-b) Il s'agit donc de trouver les valeurs n_1^* et n_2^* de n_1 et de n_2 qui rendent minimale, la fonction $\sigma^2.(\frac{1}{n_1}+\frac{1}{n_2})$, les variables n_1 et n_2 satisfaisant la contrainte $C_1.n_1+C_2.n_2=C$. La *méthode des multiplicateurs de* LAGRANGE, conduit à la fonction auxiliaire :

$$V(n_1,n_2)=\sigma^2.(\frac{1}{n_1}+\frac{1}{n_2})+\lambda.\left[C_1.n_1+C_2.n_2-C\right]$$

dont la nullité des dérivées partielles $\frac{\partial V}{\partial n_1}$ et $\frac{\partial V}{\partial n_2}$ s'écrit suivant les équations :

$$\begin{cases}-\frac{\sigma^2}{n_1^2}+\lambda.C_1=0\\ -\frac{\sigma^2}{n_2^2}+\lambda.C_2=0\end{cases}\Rightarrow n_1^*=\frac{K}{\sqrt{C_1}},n_2^*=\frac{K}{\sqrt{C_2}}.$$

L'écriture de la condition $C_1.n_1^*+C_2.n_2^*=C$, conduit immédiatement à l'expression de K. à savoir $K=\frac{C}{\sqrt{C_1}+\sqrt{C_2}}$ et à fortiori aux valeurs cherchées optimales n_1^* et n_2^*.

4-c) Numériquement, $K=\frac{1000}{5}=200$, $n_1^*=\frac{200}{2}=100$, et $n_2^*=\frac{200}{3}=66,66$. Par arrondi et respectant la contrainte d'un coût maximal égal à 1000, on obtient $n_1^*=100,n_2^*=66$.

Pour ces valeurs, $Var(\overline{X})=(\frac{1}{3})^2.\frac{\sigma^2}{100}+(\frac{2}{3})^2.\frac{\sigma^2}{66}=(\frac{1}{9\times100}+\frac{4}{9\times66}).\sigma^2=0{,}0078451.\sigma^2$.
Cette fois, la perte de précision est plus importante puisque par rapport aux échantillons de tailles optimales de la 2ème question, cette perte est ici de 9,7%. On voit donc que la méthode en question est moins appropriée à l'estimation de m.

7. On considère un échantillon de taille n, soit $(X_1,X_2,....,X_n)$, de n variables aléatoires indépendantes et équidistribuées (loi « parente » X de densité de probabilité $f(x)$).

Exprimer en fonction de $f(x)$ et de la fonction de répartition $F(x)=\Pr ob(X\le x)$, les densités de probabilités des statistiques $U=\inf_{1\le i\le n}X_i$ et $V=\sup_{1\le i\le n}X_i$.

Solution : 1°) Pour tout u on a $\Pr ob(U\ge u)=\Pr ob(X_1\ge u,X_2\ge u,....,X_n\ge u)$. Désignant respectivement par $G(u)$ et $g(u)$, la fonction de répartition et la densité de probabilité de U, il s'ensuit, compte tenu de l'indépendance des variables aléatoires X_i, l'expression :

$$\Pr ob(U\ge u)=1-G(u)=\prod_{i=1}^{i=n}\Pr ob(X_i\ge u)=(1-F(u))^n$$

Rappelant qu'entre fonction de répartition $F(x)$ et densité de probabilité $f(x)$, on a la relation $F(x) = \int_{-\infty}^{x} f(u).du$, soit par dérivation, $\frac{dF(x)}{dx} = f(x)$, la dérivation terme à terme de la relation précédente, conduit immédiatement au résultat :

$$\frac{d\,\mathrm{Prob}(U \geq u)}{du} = -\frac{dG(u)}{du} = -n.(1-F(u))^{n-1}.\frac{dF(u)}{du}$$

soit, après simplifications, l'expression cherchée de la densité de probabilité de U, à savoir :

$$g(u) = n.(1-F(u))^{n-1}.f(u).$$

• D'autre part, pour tout v, on peut écrire $\mathrm{Prob}(V \leq v) = \mathrm{Prob}(X_1 \leq v, X_2 \leq v, ..., X_n \leq v)$, soit compte tenu de l'indépendance des variables aléatoires X_i, l'expression :

$$\mathrm{Prob}(V \leq v) = \prod_{i=1}^{i=n} \mathrm{Prob}(X_i \leq v)$$

Ainsi a-t-on en désignant respectivement par $H(v)$ et $h(v)$ la fonction de répartition et la densité de probabilité de V, la relation $H(v) = \left[F(v)\right]^n$. Par dérivation, il en découle l'expression cherchée de la densité de probabilité $h(v)$ de la variable aléatoire V :

$$h(v) = \frac{dH(v)}{dv} = n.\left[F(v)\right]^{n-1}.\frac{dF(v)}{dv} = n.\left[F(v)\right]^{n-1}.f(v).$$

CHAPITRE II

ESTIMATION

A - Rappels de cours

1. La problématique de l'estimation statistique

Comme cela a été déjà indiqué en avant-propos, **l'estimation statistique** intervient lorsque le modèle probabiliste qui est susceptible d'avoir généré les observations effectuées a été choisi. Cependant son impact est déterminant dans la qualité de la modélisation mise en œuvre et l'estimation forme, à cet égard, une *partie essentielle* de la statistique mathématique.

Dans le chapitre précédent, le lecteur a déjà pu constater la place majeure que tiennent les *moyennes* « m », *proportions* « p », et *variances* « σ^2 » dans *l'estimation* « $\hat{\theta}$ » des paramètres « θ » des lois de probabilités usuelles. Il y a aussi été mis en évidence, des *estimateurs* $\overline{X}$, F_n, $\hat{S}^2$, dont les propriétés sont séduisantes puisqu'on a montré qu'on avait $E(\overline{X}) = m, E(F_n) = p, E(\hat{S}^2) = \sigma^2$ (bref, des estimateurs qualifiés plus loin, de « *sans biais* »).

• Mais **sont-ils les estimateurs les meilleurs** ? Par ailleurs, que dire des nombreux **autres paramètres et estimateurs** également rencontrés en statistique ? Telles sont les questions qui font l'objet de la formalisation présentée dans ce chapitre et à travers lequel sont développés les **propriétés des estimateurs**, la **comparaison de leurs qualités**, et leurs **modes de construction**.

• C'est **l'estimation ponctuelle** (θ approximé par $\hat{\theta}$) qui est principalement visée ci-dessus. Or l'exemple de l'application 1.6 du chapitre I montre, par le biais d'échantillons de taille $n = 2$, toute la *disparité* des estimations de m qu'on peut obtenir en recourant à la moyenne empirique $\overline{X} = \frac{\sum_{i=1}^{i=n} X_i}{n}$, dont les valeurs varient ici entre 3 et 12.

• En fait, la taille n de l'échantillon est un élément incontournable de la validité de l'estimation obtenue et c'est la raison pour laquelle on préférera fréquemment à l'estimation ponctuelle, un *encadrement* de la valeur inconnue à estimer, θ, *suivant un seuil de confiance* fixé à priori, $1-\alpha$ (en pratique, 90%, 95%, ou 99%). Cette construction constitue **l'estimation par intervalle de confiance**. Autrement dit, on cherche $a, b / \Pr ob(a \le \theta \le b) = 1-\alpha$, α donné/$0 \prec \alpha \prec 1$. L'estimation par intervalle de confiance est également traitée dans ce chapitre.

2. Propriétés des estimateurs ponctuels

2.1 Qualités d'un bon estimateur

Le contexte classique est celui d'un **échantillon** de n variables aléatoires indépendantes $(X_1, X_2, ..., X_n)$ suivant la même loi, à savoir celle d'une variable aléatoire X (variable dite « *parente* »), de densité de probabilité $f(x,\theta)$ (resp. de loi de probabilité $\Pr ob(X = x) = p(x,\theta)$ dans le cas discret), θ étant un *paramètre unidimensionnel ou multidimensionnel* dont on se propose d'évaluer l'estimation $\hat{\theta}$ à travers une **statistique** $\hat{\theta} = T_n(X_1, X_2, ..., X_n)$ (dite encore « **estimateur** »).

• En premier lieu, on peut attendre d'un estimateur « *correct* », soit $\hat{\theta} = T_n$, qu'il vérifie deux conditions :

- être **convergent**, c'est à dire vérifier $\hat{\theta} \to \theta$ lorsque le nombre des observations devient grand $(n \to +\infty)$ (on parlera également *d'estimateur consistant* pour caractériser une telle propriété) ;
- être **sans biais**, c'est-à-dire vérifier que, sur la population de la totalité des échantillons de taille n qu'il est possible d'extraire de la population considérée, on a $E(T_n) = \theta$ (on parlera *d'estimateur asymptotiquement sans biais* si on a seulement $\lim_{n \to +\infty} E(T_n) = \theta$. Cette notion s'étend aux estimateurs T_n d'une fonction $g(\theta)$ à travers la relation $E(T_n) = g(\theta)$.

La propriété de *convergence* dont il est précisé qu'il s'agit d'une *convergence en probabilité* ($\forall \varepsilon \succ 0, \lim_{n \to +\infty} \Pr ob(|T_n - \theta| \succ \varepsilon) = 0$), est le moins qu'on puisse exiger d'une statistique dont on attend qu'elle soit un estimateur acceptable. On peut en dire de même de la condition d'être *sans biais* encore qu'il existe des estimateurs biaisés tout à fait satisfaisants. Ainsi préférera-t-on un estimateur biaisé mais de variance faible à un estimateur sans biais de grande variance.

La *convergence en probabilité* est satisfaite lorsque les conditions $\lim_{n \to +\infty} E(T_n) = \theta$ et $\lim_{n \to +\infty} Var(T_n) = 0$ sont remplies simultanément. Autrement dit, *un estimateur asymptotiquement sans biais et de variance tendant vers 0 est convergent.* C'est immédiat puisque suivant *l'inégalité triangulaire*. $|T_n - \theta| \le |T_n - E(T_n)| + |E(T_n) - \theta|$.

En pratique, presque toujours la distribution de T_n devient de plus en plus étroite lorsque $n \to +\infty$ et converge, qui plus est, vers la loi normale. Les hypothèses ci-dessus sont donc le plus souvent satisfaites.

2.2 Comparaison des estimateurs

La *comparaison* des estimateurs repose sur la notion de **risque**, ce dernier étant tout naturellement défini par l'écart en norme quadratique $R(T_n) = \|T_n - \theta\|^2 = E\left[(T_n - \theta)^2\right]$.

- Ainsi l'estimateur $T_n^{'}$ est-il plus *efficace* que l'estimateur T_n^{*} si $R(T_n^{'}) \le R(T_n^{*})$.
- S'agissant d'estimateur biaisé ($E(T_n) = \theta + B$ où B désigne le biais de T_n), on a l'expression $R(T_n) = Var(T_n) + B^2$. Bien entendu, lorsqu'on a un estimateur sans biais ($B = 0$), le risque et les comparaisons qui en découlent, se résume à $Var(T_n)$.

En effet, lorsque $E(T_n)=\theta+B$, le développement du risque $R(T_n)=\|T_n-\theta\|^2$ s'écrit $E\left[(T_n-\theta)^2\right]=E\left[(T_n-E(T_n)+E(T_n)-\theta)^2\right]=E\left[(T_n-E(T_n)+B)^2\right]$, soit suivant linéarité de l'espérance mathématique :

$$E\left[(T_n-\theta)^2\right]=E\left[(T_n-E(T_n))^2+E(B^2)+2.E\left[(T_n-E(T_n).B\right]\right].$$

Mais d'une part, $E(B^2)=B^2$, et d'autre part, $E\left[(T_n-E(T_n).B\right]=B.E\left[T_n-E(T_n)\right]$, soit $B.E(T_n)-B.E\left[E(T_n)\right]=0$. Ainsi obtient-on, en conclusion, le résultat annoncé :

$$E\left[(T_n-\theta)^2\right]=Var(T_n)+B^2$$

- L'estimateur T_n^* pour lequel $R(T_n^*)\leq R(T_n)$, $\forall$ l'estimateur T_n considéré, est dit *optimal*. Dans le cas *d'estimateurs sans biais*, l'estimateur optimal s'apparente à *l'estimateur de variance minimale* (on parle d'**estimateur efficace**).

- Comme il sera montré plus loin, le meilleur estimateur qui est sans biais et de variance minimale n'existe pas toujours. On pourra lui préférer un estimateur biaisé de risque plus faible comme cela a déjà été mentionné précédemment.

- Enfin, ce n'est pas parce que $\hat{\theta}$ est un bon estimateur de θ que $g(\hat{\theta})$ est un bon estimateur de $g(\theta)$. Par exemple, la constatation simple $E(X^2)=Var(X)+E(X)^2$ montre que $E(g(\theta))$ est différent généralement de $g(E(\theta))$ et qu'on peut donc avoir $g(\hat{\theta})$ biaisé, alors que $\hat{\theta}$ est sans biais.

2.3 Information de FISHER

Développée dans les années 1920 et anticipant d'une vingtaine d'années la fonction « **entropie** » de SHANNON, base de la *théorie de l'information*, l'information de FISHER dont l'intérêt est souligné par l'inégalité de CRAMER RAO présentée ci-après, a pour objet la **quantification de l'information** pertinente contenue dans les données.

• Une *statistique* est une transformation des données de l'échantillon qui *résume* et *simplifie* ce dernier. Dans ce cadre une statistique est dite « **exhaustive** » (cf. paragraphe 2.5), si ce résumé ne supprime pas de l'information contenue dans l'échantillon en question. Ainsi, les notions d'information et de statistiques exhaustives sont-elles liées.

• Mathématiquement, X étant une variable aléatoire dont la loi $f(x,\theta)$ (on supposera $f(x,\theta)\succ 0$) dépend d'un paramètre réel θ, on définit la **quantité d'information de FISHER** *fournie par la variable aléatoire* X *sur le paramètre* θ, par la quantité :

$$I(\theta)=E\left[\left(\frac{\partial}{\partial\theta}\ln f(x,\theta)\right)^2\right]$$

(le calcul de l'espérance mathématique étant effectué par rapport à X).

• Lorsque $f(x,\theta)$ est *au moins deux fois dérivable* et qu'on a de plus le domaine de définition de $f(x,\theta)$ (dit *« support »*) *indépendant de* θ, on obtient pour $I(\theta)$, une seconde expression fort utile pour les calculs, à savoir :

$$I(\theta)=-E\left[\frac{\partial^2}{\partial\theta^2}\ln f(x,\theta)\right].$$

En effet, supposant par exemple X à valeurs sur R et y satisfaisant les conditions $f(x,\theta)$ strictement positive, deux fois dérivable. et de support indépendant de θ, on a par définition, $\int_R f(x,\theta).dx = 1, \forall\theta$. Dérivant cette relation par rapport à θ, il vient successivement, $\frac{d}{d\theta}(\int_R f(x,\theta).dx = 0 = \int_R \frac{\partial}{\partial\theta} f(x,\theta).dx = \int_R (\frac{\partial}{\partial\theta}\ln f(x,\theta)).f(x,\theta).dx$, soit l'espérance mathématique $E\left[\frac{\partial}{\partial\theta}\ln f(x,\theta)\right]$.

Dérivant de nouveau, la relation obtenue ci-dessus, $E\left[\frac{\partial}{\partial\theta}\ln f(x,\theta)\right]$=0, il vient :

$\frac{d}{d\theta}\int_R (\frac{\partial}{\partial\theta}\ln f(x,\theta)).f(x,\theta).dx = 0$, soit en développant la dérivée en question :

$$\int_R (\frac{\partial^2}{\partial\theta^2}\ln f(x,\theta)).f(x,\theta).dx + \int_R (\frac{\partial}{\partial\theta}\ln f(x,\theta))^2.f(x,\theta).dx = 0$$

Ainsi, $I(\theta)$ qui est égale à $E\left[(\frac{\partial}{\partial\theta}\ln f(x,\theta))^2\right] = \int_R (\frac{\partial}{\partial\theta}\ln f(x,\theta))^2.f(x,\theta).dx$ est aussi égale à $-\int_R (\frac{\partial^2}{\partial\theta^2}\ln f(x,\theta)).f(x,\theta).dx = -E\left[\frac{\partial^2}{\partial\theta^2}\ln f(x,\theta)\right]$ ce qui prouve le résultat annoncé.

• Le *modèle uniforme* présenté en application 1.3 du présent chapitre, est un exemple usuel où les conditions du calcul précédent ne sont pas remplies puisque le support de $f(x,\theta)$ dépend alors de θ.

• Plus largement et toujours sous les hypothèses de validité de la seconde expression de $I(\theta)$, pour deux variables aléatoires indépendantes de lois respectives $f(x,\theta)$ et $g(x,\theta)$, la *quantité d'information* $I(\theta)$ *liée au couple* (X,Y), soit $I_{(X,Y)}(\theta)$, *est égale à la somme* $I_X(\theta)+I_Y(\theta)$ *des quantités d'informations* liées respectivement à X et à Y.

La démonstration est immédiate puisque X et Y étant indépendantes, le couple (X,Y) a pour densité de probabilité $f(x,\theta).g(y,\theta)$. Dans ces conditions :

$I_{(X,Y)}(\theta) = -E\left[\frac{\partial^2}{\partial\theta^2}\ln(f(x,\theta).g(y,\theta))\right] = -E\left[\frac{\partial^2}{\partial\theta^2}(\ln f(x,\theta)+\ln(g(y,\theta))\right]$, soit par linéarité de l'espérance mathématique, $I_{(X,Y)}(\theta) = -E\left[\frac{\partial^2}{\partial\theta^2}\ln f(x,\theta)\right] - E\left[\frac{\partial^2}{\partial\theta^2}g(y,\theta)\right]$.

Ceci établit le résultat annoncé.

• Par extension, il s'ensuit que pour un *échantillon* $(X_1,X_2,...,X_n)$ de n variables aléatoires indépendantes X_i de loi parente X, *la quantité d'information contenue dans l'échantillon relativement à* θ, soit $I_{(X_1,X_2,...,X_n)}(\theta)$ *est égale à* $n.I_X(\theta)$.

C'est une évidence en itérant le résultat précédent établi pour le cas de deux variables.

• Désignant par $L(X_1,X_2,...,X_n,\theta)$ la densité de probabilité du n-uplet $(X_1,X_2,...,X_n)$, fonction essentielle en statistique et dite « **fonction de vraisemblance** », on a :

$$I_{(X_1,X_2,...,X_n)}(\theta) = E\left[(\frac{\partial}{\partial\theta}\ln L(X_1,X_2,...,X_n,\theta))^2\right]$$

Ici encore, le résultat est immédiat dans la mesure où, considérant les variables aléatoires X_i indépendantes de densités de probabilités respectives $f(x_i,\theta)$, la densité du n-uplet $(X_1,X_2,...,X_n)$ est égale au produit des densités $L(X_1,X_2,...,X_n,\theta)=\prod_{i=1}^{i=n} f(X_i,\theta)$.

2.4 Inégalité de CRAMER - RAO

Cette inégalité est fondamentale puisqu'elle fournit dans les conditions précisées ci-après, une *borne inférieure* de la qualité des estimateurs possibles pour un paramètre θ donné. Ce résultat est le suivant :

• Considérant un échantillon $(X_1,X_2,...,X_n)$ de n variables aléatoires indépendantes équidistribuées de même loi, c'est-à-dire de variable « parente » X caractérisée par la loi de densité de probabilité $f(x,\theta)$ (resp. $Prob(X=x)=p(x,\theta)$ dans le cas discret), pour tout estimateur $T_n(X_1,X_2,...,X_n)$ du paramètre θ (voir plus largement d'une fonction $g(\theta)$), on a $Var(T_n)\geq \dfrac{(\frac{d}{d\theta}E(T_n))^2}{n.I_X(\theta)}$.

L'inégalité de CRAMER - RAO nécessite la satisfaction des conditions d'applicabilité de la seconde expression de la quantité d'information de FISHER, c'est-à-dire $f(x,\theta)$ au moins deux fois dérivable, strictement positive, et possédant un support indépendant de θ.

Posant $L(v,\theta)=L(x_1,x_2,...,x_n,\theta)$ et $dv=\prod_{i=1}^{i=n} dx_i$ pour simplifier les notations. on a par définition $\int_{R^n} L(v,\theta).dv=1$. Par dérivation, $\frac{d}{d\theta}\int_{R^n} L(v,\theta).dv=0$ entraîne les relations $\int_{R^n}\frac{d}{d\theta}L(v,\theta).dv=\int_{R^n}(\frac{d}{d\theta}\ln L(v,\theta)).L(v,\theta).d\theta=0$ (dérivation sous le signe somme admissible parce que le support de $L(v,\theta)$ est indépendant de θ).

En conclusion, on a le premier résultat $E\left[\frac{\partial}{\partial\theta}\ln(v,\theta)\right]=0$ (1)

D'autre part, formant $E(T_n)=\int_{R^n} T_n(v).L(v,\theta).dv$, on a, par dérivation, les relations $\frac{d}{d\theta}E(T_n)=\int_{R^n} T_n(v).\frac{\partial}{\partial\theta}L(v,\theta).dv=\int_{R^n} T_n(v).(\frac{\partial}{\partial\theta}\ln L(v,\theta)).L(v,\theta).dv$. Ainsi, obtient-on le second résultat $\frac{d}{d\theta}E(T_n)=E\left[T_n.\frac{\partial}{\partial\theta}(\ln L(v,\theta))\right]$ (2)

Dès lors, écrivant (1) sous la forme $E\left[E(T_n).\frac{\partial}{\partial\theta}\ln(v,\theta)\right]=0$ (puisque $E(T_n)$ est constant), et développant (2)-(1), il vient $\frac{d}{d\theta}E(T_n)=E\left[(T_n-E(T_n)).\frac{\partial}{\partial\theta}(\ln L(v,\theta))\right]$.

L'inégalité de SCHWARTZ à savoir $\prec U,V\succ\leq \|U\|.\|V\|,\forall(U,V)$, permet de conclure au résultat $VarT_n.E\left[(\frac{\partial}{\partial\theta}\ln L(v,\theta))^2\right]\geq(\frac{d}{d\theta}E(T_n))^2$ d'où l'inégalité de CRAMER RAO, à savoir $Var(T_n)\geq\dfrac{(\frac{d}{d\theta}E(T_n))^2}{I_v(\theta)}$ avec $I_v(\theta)=n.I_X(\theta)$.

• *Dans le cas très classique d'estimateurs* T_n *sans biais* d'une fonction $g(\theta)$ du paramètre θ, on a immédiatement $Var(T_n) \geq \frac{g'(\theta)^2}{n.I_X(\theta)}$. Ainsi, pour l'ensemble des estimateurs sans biais de θ (cas où $g(\theta) = \theta$), a-t-on pour évaluation de la *borne inférieure de la précision desdits estimateurs,* la valeur minimale $\frac{1}{n.I_X(\theta)}$. Lorsqu'elle est atteinte, cette borne notée couramment « **borne F.D.C.R** » (pour FRECHET, DARMOIS, CRAMER, RAO), correspond à l'*estimateur* dit « **efficace** ».

L'inégalité de CRAMER - RAO susmentionnée fixe la meilleure précision qu'on puisse espérer. Mais, il est rappelé que cela ne prouve pas l'existence systématique d'estimateurs sans biais de θ (ou $g(\theta)$ qui atteignent cette borne minimale F.D.C.R.

S'agissant d'estimateurs baisés ($E(T) = g(\theta) + B$), on sait d'après un calcul antérieur que $E\left[(T - g(\theta))^2\right] = VarT + B^2$. On pourra donc écrire :

$$E\left[(T - g(\theta))^2\right] \geq B^2 + \frac{g'(\theta)^2}{n.I_X(\theta)},$$

résultat, constituant **l'inégalité de CRAMER - RAO généralisée** et se ramenant lorsque $B = 0$, à l'énoncé portant sur les estimateurs sans biais.

• Toujours, dans le cas d'estimateurs sans biais, on montre qu'une *condition nécessaire et suffisante* pour qu'un estimateur T de $g(\theta)$ soit *efficace*, c'est-à-dire de variance minimale (égale à la borne F.D.C.R), est qu'il existe trois fonctions $\alpha(\theta), \beta(\theta)$, et $\gamma(x)$ telles que $\ln f(x,\theta) = \alpha(\theta).T(x) + \beta(\theta) + \gamma(x)$. Les modèles appartenant à la **famille exponentielle** (cf. ci-après) sont un exemple classique d'illustration de cette condition nécessaire et suffisante.

• Enfin, le cas où σ est *multidimensionnel* n'est traité que très partiellement dans cet ouvrage car conduisant à des calculs assez lourds. Mais il donne lieu à des développements comparables à ce qui vient d'être exposé.

2.5 Statistiques exhaustives

Comme cela a été exposé dans le paragraphe 2.3 précédent, il s'agit de statistiques intéressantes puisque tout en le simplifiant, *elles ne suppriment en rien l'information contenue dans l'échantillon.* En d'autres termes, notant par $I_{(X_1,X_2,...,X_n)}(\theta)$ l'information fournie sur θ par un échantillon de taille n et par $I_S(\theta)$ l'information fournie par toute statistique S sur θ, on a $I_S(\theta) \leq I_{(X_1,X_2,...,X_n)}(\theta)$. *Lorsque* S *est exhaustive,* $I_S(\theta)$ *est maximale* et atteint $I_{(X_1,X_2,...,X_n)}(\theta)$.

• Au plan formel, une statistique S est dite *exhaustive* pour le paramètre θ, si la loi de probabilité de l'échantillon $(X_1, X_2, ..., X_n)$ conditionnellement à la valeur $S = s$ ne dépend pas de la valeur de θ.

Par exemple, afin d'évaluer la proportion inconnue de pièces défectueuses dans une fabrication, soit p, on effectue un prélèvement (avec remise) de n pièces qui conduit à l'échantillon $(X_1, X_2, ..., X_n)$ des n variables de BERNOULLI égales à 1 si la pièce est défectueuse et à 0 sinon.

Il est intuitif de considérer que la connaissance d'un n-uplet de 0 et de 1 induite par un prélèvement se résume pour ce qui est de p, à la détermination du nombre de pièces défectueuses dans l'échantillon, soit $S=\sum_{i=1}^{i=n} X_i$. Or, l'expression de la loi conditionnelle $\Pr ob(X_1=x_1, X_2=x_2,...,X_n=x_n / S=x)$ conduit au rapport :

$$\Pr ob(X_1=x_1, X_2=x_2,...,X_n=x_n / S=x)=\frac{\prod_{i=1}^{i=n}\Pr ob(X_i=x_i)}{\Pr ob(S=x)}=\frac{p^x.(1-p)^{n-x}}{C_n^x.p^x.(1-p)^{n-x}}=\frac{1}{C_n^x}$$

Cette loi conditionnelle qui est la loi uniforme sur $\{0,1\}^n$ ne dépend pas de p et S forme donc ici un exemple de *statistique exhaustive*.

• En pratique, les calculs pour expliciter la loi conditionnelle sont loin de présenter la simplicité ci-dessus. Le théorème suivant qu'on admettra et qui est connu sous le nom de **théorème de factorisation (théorème de NEYMAN - FISHER)** permet *d'identifier plus facilement les statistiques exhaustives*. Son énoncé est le suivant :

- S est une statistique exhaustive si et seulement si on peut trouver deux fonctions g et h telles que la loi $f(X,\theta)$ de l'échantillon $X=(X_1,X_2,...,X_n)$ s'écrit sous la forme $f(X,\theta)=g(S(X),\theta)\times h(X)$, c'est-à-dire le produit d'une fonction qui dépend de S et de θ mais pas explicitement de X par une fonction qui dépend seulement de l'échantillon $X=(X_1,X_2,...,X_n)$ et pas du paramètre θ.

Supposant par exemple la variable « parente » de *loi exponentielle* de paramètre inconnu λ et considérant la statistique $S=\sum_{i=1}^{i=n} X_i$, on a $f(X,\theta)=\prod_{i=1}^{i=n}\lambda.e^{-\lambda.X_i}=\lambda^n.e^{-\lambda.S}$. On constate ainsi que $f(X,\theta)$ est le produit de la fonction $g(S,\theta)=\lambda^n.e^{-\lambda.S}$ et de la fonction $h(X)=1$. Il s'agit bien d'une *statistique exhaustive* pour le paramètre λ, d'après le *théorème de factorisation*.

2.6 Le cas particulier de la famille exponentielle

C'est un cas particulièrement usité parce que les *modèles courants* que *sont la loi binomiale, la loi de POISSON, la loi Gamma – n, la loi exponentielle, la loi normale,* appartiennent entre autres à la **famille exponentielle**.

• Cette dernière recouvre les modèles pour lesquels la variable *« parente »* X a une densité de probabilité $f(x,\theta)$ (resp. une loi $p(x,\theta)$ dans le cas discret), qui peut s'écrire sous la forme $f(x,\theta)=\exp\left[Q(\theta).T(x)+\Psi(\theta)+C(x)\right]$, soit encore suivant la forme logarithmique, $\ln f(x,\theta)=Q(\theta).T(x)+\beta(\theta)+\gamma(x)$.

• On montre que sous certaines conditions (généralement vérifiées pour les lois courantes susmentionnées), $T=\sum_{i=1}^{i=n} T(X_i)$ constitue une **statistique exhaustive**, qui plus est **efficace**, compte tenu de la condition nécessaire et suffisante précédemment mentionnée au paragraphe 2.4.

Par exemple, pour la loi de POISSON caractérisée par $p(x,\theta)=\frac{e^{-\theta}.\theta^x}{x!}$, on a $\ln p(x,\theta)=-\theta+x.\ln\theta-\ln(x!)$.

Il s'agit bien d'une loi appartenant à la *famille exponentielle* pour laquelle $Q(\theta)=\ln\theta, T(x)=x, \beta(\theta)=-\theta, \gamma(x)=-\ln(x!)$. $T(X)=\sum_{i=1}^{i=n} X_i$ constitue donc une *statistique exhaustive* pour θ ce qui était prévisible.

3. Construction des estimateurs

3.1 Construction suivant statistique exhaustive (théorèmes de RAO – BLACKWELL et LEHMANN – SCHEFFE)

Le théorème ci-dessous permet *d'améliorer un estimateur à l'aide d'une statistique exhaustive* :

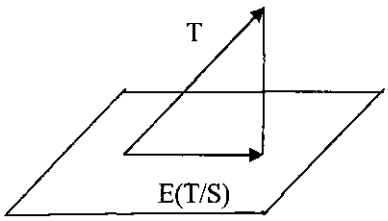
• Si T est un *estimateur sans biais* de $g(\theta)$ et si S est une *statistique exhaustive* pour $g(\theta)$, *l'estimateur* $Z(S)=E(T/S)$ est un **estimateur sans biais préférable à T (théorème de RAO – BLACKWELL)**.

En effet, $Z(S)$ à pour risque $R(Z)=E\left[(Z-g(\theta))^2\right]=E(Z^2)-2g(\theta).E(Z)+g(\theta)^2$. La différence des risques associés respectivement à T et à Z est donc égale à $R(T)-R(Z)=E(T^2)-E(Z^2)-2g(\theta).\left[E(T)-E(Z)\right]$.

Or, suivant le *théorème de l'espérance totale* portant sur les espérances conditionnelles, $E(Z)=E\left[E(T/S=s)\right]=E(T)$. Ainsi, $R(T)-R(Z)$ se résume t-elle à $E(T^2)-E(Z^2)$.

Formant parallèlement $E\left[(T-Z)^2\right]=E(T^2)-2.E(T.Z)+E(Z^2)$, il est proposé ci-après d'expliciter $E(T.Z)=E\left[T.E(T/S)\right]$.

En ce sens, il est rappelé que $\forall(X,Y)$, l'espérance conditionnelle $E(Y/X)$ représente la projection orthogonale de Y sur le sous-ensemble engendré par la variable aléatoire X, tout cela suivant le produit scalaire habituel $\prec U,V\succ=E(U.V)$.



Le schéma ci-contre montre qu'on a $E\left[T.E(T/S)\right]=\prec T,E(T/S)\succ$, soit aussi, $\prec T-E(T/S),E(T/S)\succ+\prec E(T/S),E(T/S)\succ$, ce qui se résume à :
$\prec E(T/S),E(T/S)\succ=E(T/S)^2$ du fait que $T-E(T/S)\perp S\Rightarrow\prec T-E(T/S),E(T/S)\succ=0$.

En résumé, $E(T.Z)=E(Z^2)$ et $E\left[(T-Z)^2\right]=E(T^2)-E(Z^2)$. Ainsi la différence des risques $R(T)-R(Z)=E(T^2)-E(Z^2)=E\left[(T-Z)^2\right]$ est-elle positive ou nulle, ce qui établit le théorème de RAO-BLACKWELL suivant lequel Z est préférable à T.

L'exemple ci-dessous illustre les mécanismes du théorème de RAO- BLACKWELL.

On considère un échantillon indépendant $(X_1,X_2,...,X_n)$ de loi parente X de type POISSON ($\Pr ob(X=x)=\frac{e^{-\theta}.\theta^x}{x!}$). On se propose d'estimer ici $e^{-\theta}=\Pr ob(X=0)$.

Pour cela on va partir de la *statistique exhaustive* $S=\sum_{i=1}^{i=n} X_i$ précédemment exprimée dans le cadre d'un exemple de famille exponentielle et d'un *estimateur T très grossier* défini par :

$T=1$ si $X_1=0$ et $T=0$ sinon. Certes T est un estimateur sans biais puisque $E(T)=1\times \Pr ob(X_1=0)+0\times \Pr ob(X_1\neq 0)=e^{-\theta}$, mais, fonction de la seule variable X_1, il est loin de prendre en compte toute l'information contenue dans l'échantillon de taille n, $(X_1,X_2,...,X_n)$. Il constitue donc à cet égard un estimateur très rudimentaire.

Dans ces conditions, caractérisons la variable conditionnelle X_1/S. Il est immédiat en premier lieu que $S=\sum_{i=1}^{i=n}X_i$ suit la loi de POISSON de paramètre $n.\theta$ (cf. cours de probabilités). Dès lors :

$$\Pr ob(X_1=k/S=s)=\frac{\Pr ob(X_1=k,S=s)}{\Pr ob(S=s)}=\frac{\Pr ob(X_1=k,\sum_{i=2}^{i=n}X_i=s-k)}{\Pr ob(S=s)}.$$

$\sum_{i=2}^{i=n}X_i$ suivant pour sa part la loi de POISSON de paramètre $(n-1).\theta$, il vient en définitive :

$$\Pr ob(X_1=k/S=s)=\frac{\frac{\theta^k.e^{-\theta}}{k!}.\frac{\left[(n-1).\theta\right]^{s-k}.e^{-(n-1).\theta}}{(s-k)!}}{\frac{(n.\theta)^s}{s!}.e^{-n.\theta}}=\frac{s!}{k!.(s-k)!}.\frac{(n-1)^{s-k}}{n^s}.$$

On reconnaît ici *la loi binomiale* de type $B(s,\frac{1}{n})$.

Revenant à $E(T/S)$ qui est l'estimateur fourni par le théorème de RAO- BLACKWELL. il est égal à $1\times \Pr ob(X_1=0/S=s)+0\times \Pr ob(X_1\neq 0/S=s)$, soit $(1-\frac{1}{n})^s$ suivant le calcul précédent relatif à X_1/S.

On a donc là, à travers $(1-\frac{1}{n})^S$, un estimateur qui tout comme T forme un estimateur sans biais de $e^{-\lambda}$, mais qui, suivant le théorème de RAO- BLACKWELL, est beaucoup plus précis. C'est même un *estimateur optimal* comme le montre le théorème de LEHMANN – SCHEFFE ci-après.

• Une statistique est dite **complète** si $\forall\theta, E\left[h(S)\right]=0\Rightarrow h=0$ *presque partout* (c'est-à-dire partout sauf en un ensemble de points de mesure nulle).

• Dans ces conditions, si T^* est un *estimateur sans biais* de $g(\theta)$ *dépendant d'une statistique exhaustive complète*, il est alors **l'unique estimateur sans biais de variance minimale** de $g(\theta)$ **(théorème de LEHMANN – SCHEFFE).**

En particulier, si on possède déjà un estimateur T de $g(\theta)$ qui est sans biais, et si U est une statistique exhaustive complète, on a nécessairement $T^*=E(T/U)$.

Montrer qu'une statistique exhaustive est complète n'est pas aisé surtout si on survole les notions de théorie de la mesure et de l'intégration comme c'est le cas dans cet ouvrage. On admettre cependant que dans le cas particulier de la *famille exponentielle, toute statistique exhaustive est complète*, ce qui suffira pour la plupart des cas pratiques à commencer par l'exemple précédent relatif à la loi de POISSON. Ainsi $(1-\frac{1}{n})^S$ est-il bien l'estimateur optimal de $e^{-\lambda}$ puisque S est exhaustive et que la loi de POISSON appartient à la famille exponentielle.

3.2 Méthode des moments

C'est la méthode intuitivement *la plus naturelle* et elle est d'ailleurs antérieure aux autres techniques dont notamment la *méthode du maximum de vraisemblance* développée plus loin. Plus usitée dans le cas où θ est multidimensionnel, soit $\theta = (\theta_1, \theta_2, ..., \theta_p)$, son principe consiste à exprimer les θ_i $(1 \le i \le p)$ en fonction des p moments d'ordre h, $m_h = E(X^h)$ avec $1 \le h \le p$, ces derniers étant ensuite substitués par leurs estimations par les *moments empiriques* $\widehat{m_h} = \dfrac{\sum_{i=1}^{i=n} X_i^h}{n}$. La méthode exploite la propriété suivant laquelle les moments empiriques $\widehat{m_h}$ constituent des estimateurs sans biais et convergents des moments théoriques m_h $(1 \le h \le p)$.

Au plan pratique, on exprime les m_h en fonction des θ_h, soient :

$$\begin{cases} m_1 = g_1(\theta_1, \theta_2, ..., \theta_p) \\ m_2 = g_2(\theta_1, \theta_2, ..., \theta_p) \\ ..... \\ m_p = g_p(\theta_1, \theta_2, ..., \theta_p) \end{cases}$$

Puis, on résout, lorsque c'est possible, le système qui permet d'exprimer les θ_h en fonction des moments m_h, soient les expressions :

$$\begin{cases} \theta_1 = h_1(m_1, m_2, ..., m_p) \\ \theta_2 = h_2(m_1, m_2, ..., m_p) \\ ..... \\ \theta_p = h_p(m_1, m_2, ..., m_p) \end{cases}$$

En remplaçant les valeurs théoriques m_h par les moments empiriques $\widehat{m_h}$, on obtient donc ainsi les estimations cherchées $\widehat{\theta_h}$ des θ_h ($1 \le h \le p$), soient :

$$\begin{cases} \widehat{\theta_1} = h_1(\widehat{m_1}, \widehat{m_2}, ..., \widehat{m_p}) \\ \widehat{\theta_2} = h_2(\widehat{m_1}, \widehat{m_2}, ..., \widehat{m_p}) \\ ..... \\ \widehat{\theta_p} = h_p(\widehat{m_1}, \widehat{m_2}, ..., \widehat{m_p}) \end{cases}$$

Le **théorème de SLUTSKY** justifie la propriété de *convergence en probabilité* des estimateurs ainsi obtenus.

Le théorème de SLUTSKY exprime la convergence en loi (ou en probabilité) de toute fonction $f(X_n)$ vers $f(X)$ dès lors que la suite de variables aléatoires X_n converge en loi (ou en probabilité) vers X.

Par ailleurs, la méthode des moments s'applique tout autant en utilisant les moments centrés $Var(X_h) = E\left[(X - E(X))^h\right]$, voire même un mélange des deux (centrés et non centrés) comme cela est montré dans l'exemple ci- après.

Dans cet exemple, la *méthode des moments* est utilisée pour estimer les paramètres p et q de la *loi* $Beta(p,q)$, loi dont pour p et q positifs, la densité de probabilité est définie par :

$$f(x)=\frac{1}{B(p,q)}.x^{p-1}.(1-x)^{q-1} \text{ pour } 0\le x\le 1 \text{ et :}$$

$$B(p,q)=\int_0^{+\infty}\frac{u^{p-1}}{(1+u)^{p+q}}.du=\int_0^{+\infty}u^{p-1}.(1-u)^{q-1}.du=\frac{\Gamma(p).\Gamma(q)}{\Gamma(p+q)},\quad B(p,q) \text{ et } \Gamma(p)$$

désignant respectivement les fonctions *Béta* et d'*Euler*.

On montre par les calculs classiques que $E(X)=\frac{p}{p+q}$ et $Var(X)=\frac{p.q}{(p+q)^2.(p+q+1)}$.

On en déduit donc aisément que $p=E(X).\left[\frac{E(X).[1-E(X)]-Var(X)}{Var(X)}\right]$ et $q=[1-E(X)].\left[\frac{E(X).[1-E(X)]-Var(X)}{Var(X)}\right]$. Dès lors, en considérant les statistiques habituelles associées à $E(X)$ et $Var(X)$, soient $\overline{X}=\frac{\sum_{i=1}^{i=n}X_i}{n}$ et $\hat{S}^2=\frac{1}{n-1}.\sum_{i=1}^{i=n}(X_i-\overline{X})^2$, on obtient les estimateurs cherchés de p et q, soient $\hat{p}=\overline{X}.\left[\frac{\overline{X}.(1-\overline{X})-\hat{S}^2}{\hat{S}^2}\right]$ et $\hat{q}=(1-\overline{X}).\left[\frac{\overline{X}.(1-\overline{X})-\hat{S}^2}{\hat{S}^2}\right]$.

• Les estimateurs obtenus par la *méthode des moments* ne sont pas forcément sans biais et l'évaluation de leur qualité reste parfois délicate sans omettre, qui plus est, les difficultés analytiques que pose la résolution du système des m_h en fonction des θ_h. La **méthode du maximum de vraisemblance** exposée ci-dessous reste donc *la plus courante et la plus souhaitable.*

3.3 Méthode du maximum de vraisemblance

Due à FISHER, c'est comme indiqué précédemment la *méthode la plus utilisée*, notamment lorsque le paramètre θ est unidimensionnel (hypothèse la plus courante dans cet ouvrage). Elle permet en effet, de trouver des estimateurs qui le plus souvent sont **performants**.

Supposant vérifiée l'unicité de la valeur $\hat{\theta}$ qui rend **maximale** la *fonction de vraisemblance* $L(X_1,X_2,...,X_n,\theta)$ associée à l'échantillon $(X_1,X_2,...,X_n)$, on montre que $\hat{\theta}$ qui est appelé **l'estimateur du maximum de vraisemblance** de θ et est noté « **E.M.V** », forme relativement au paramètre inconnu θ, un estimateur qui est *convergent en probabilité (donc consistant), asymptotiquement sans biais, efficace, et de loi normale.*

• Un estimateur du maximum de vraisemblance n'est pas nécessairement unique car la fonction de vraisemblance peut avoir plusieurs maximas. Mais la *condition supplémentaire* $\frac{\partial^2 L}{\partial\theta^2}<0$ assure cette unicité, outre la condition d'optimalité $\frac{\partial L}{\partial\theta}=0$, et la validité de la méthode en question.

• Par ailleurs, comme l'illustre l'application 1.3 ci-après dans le cas du *modèle uniforme, l'estimateur E.M.V n'a pas obligatoirement une expression analytique explicite.*

• Enfin, il est à noter que les *propriétés attrayantes* de l'estimateur E.M.V sont avant tout **asymptotiques (convergence, normalité, et efficacité)**. Mais tous les estimateurs E.M.V ne sont pas sans biais, ni de variance minimale.

• Ces propriétés asymptotiques entraînent que $\hat{\theta}-\theta$ *converge, pour n grand, vers une loi normale centrée,* puisque $\lim_{n\to+\infty} E(\hat{\theta})=\theta$.

Plus précisément, on remarque tout d'abord que $Var(\hat{\theta}-\theta)=E\left[(\hat{\theta}-\theta)^2\right]-\left[E(\hat{\theta}-\theta)\right]^2$ converge vers $E\left[(\hat{\theta}-\theta)^2\right]=Var(\widehat{\theta})$ puisque, lorsque n est grand, $E(\hat{\theta}-\theta)=0$. Dès lors, l'efficacité de l'estimateur E.M.V, soit $\hat{\theta}$, implique que $Var(\hat{\theta})$ atteint la borne minimale F.D.C.R (de FRECHET, DARMOIS, CRAMER, RAO), borne dont on a montré précédemment qu'elle était égale à $\frac{1}{I_{(X_1,X_2,...,X_n)}(\theta)}$, $I_{(X_1,X_2,...,X_n)}(\theta)$ désignant l'information de FISHER associée à l'échantillon $(X_1,X_2,...,X_n)$. Ainsi, en définitive, $\hat{\theta}-\theta$ converge-t-il vers la loi normale $N(0,\sigma^2=\frac{1}{I_{(X_1,X_2,...,X_n)}(\theta)})$.

• **Au plan pratique**, l'indépendance des variables X_i qui constituent l'échantillon $(X_1,X_2,...,X_n)$ considéré et dont la loi parente est supposée être caractérisée par la densité de probabilité $f(x,\theta)$ (respectivement la loi $p(x,\theta)$ dans le cas discret), entraîne l'expression $L(X_1,X_2,...,X_n,\theta)=\prod_{i=1}^{i=n} f(X_i,\theta)$. Notant synthétiquement par $L(X,\theta)$ la fonction de vraisemblance en question, l'estimateur E.M.V qui est solution de l'équation $\frac{\partial}{\partial\theta}L(X,\theta)=0$, est aussi solution de $\frac{\partial}{\partial\theta}\ln L(X,\theta)=0$. *L'usage de la fonction* dite « **de log-vraisemblance** », soit $\ln L(X,\theta)$, **simplifie** à l'évidence *les calculs* compte tenu de l'expression sous forme de produit, de $L(X,\theta)$.

• Enfin, on montre, pour toute fonction $g(\theta)$, que si $\hat{\theta}$ est l'estimateur E.M.V de θ, $g(\hat{\theta})$ est l'estimateur E.M.V de $g(\theta)$.

La méthode du maximum de vraisemblance est largement illustrée dans la suite de ce chapitre. Néanmoins, une application des résultats ci-dessus est d'ores et déjà montrée ci-après.

Considérant par exemple, un échantillon de n variables aléatoires indépendantes de loi « parente » de type *géométrique* ($\Pr ob(X=x)=(1-\theta)^{x-1}.\theta$, la recherche de l'estimateur E.M.V de θ, conduit à former snccessivement $L(X,\theta)=\prod_{i=1}^{i=n}(1-\theta)^{X_i-1}.\theta$, soit $L(X,\theta)=(\frac{\theta}{1-\theta})^n.(1-\theta)^{\sum_{i=1}^{i=n}X_i}$, puis $\ln L(X,\theta)=n.\ln\theta-n.\ln(1-\theta)+(\sum_{i=1}^{i=n}X_i).\ln(1-\theta)$, et enfin $\frac{\partial}{\partial\theta}\ln L(X,\theta)=\frac{n}{\theta}+\frac{n}{1-\theta}-\frac{\sum_{i=1}^{i=n}X_i}{1-\theta}$.

Il s'ensuit que l'estimateur E.M.V cherché de θ, soit $\hat{\theta}$, est solution de l'équation $\frac{\partial}{\partial\theta}\ln L(X,\theta)=0\Rightarrow\hat{\theta}=\frac{n}{\sum_{i=1}^{i=n}X_i}$. La valeur en question est bien un maximum, puisqu'on constate aisément que $\frac{\partial^2}{\partial\theta^2}\ln L(X,\theta)=-\frac{n}{\theta^2}-(\sum_{i=1}^{i=n}X_i-n).\frac{1}{(1-\theta)^2}\prec 0$. Cette valeur ainsi trouvée pour ce qui est de l'E.M.V est cohérente avec le fait que $E(X)=\frac{1}{\theta}$.

Il résulte de la dernière des propriétés rappelées précédemment que $\frac{1}{\hat{\theta}}=\frac{\sum_{i=1}^{i=n}X_i}{n}$ forme également un E.M.V de $\frac{1}{\theta}$. S'agissant de ce dernier estimateur, le calcul de $E(\frac{1}{\hat{\theta}})=\frac{\sum_{i=1}^{i=n}E(X_i)}{n}=\frac{n.\frac{1}{\theta}}{n}=\frac{1}{\theta}$, montre qu'il s'agit d'un *estimateur sans biais*.

Par ailleurs, $Var(\frac{1}{\hat{\theta}})=Var(\frac{\sum_{i=1}^{i=n}X_i}{n})=\frac{1}{n^2}.\sum_{i=1}^{i=n}Var(X_i)$ où $Var(X_i)=\frac{1-\theta}{\theta^2}$ (se reporter aux propriétés de la loi géométrique rappelées au chapitre I). Ainsi, $Var(\frac{1}{\hat{\theta}})=\frac{1}{n}.\frac{1-\theta}{\theta^2}$, ce qui montre que $\frac{1}{\hat{\theta}}$ est un *estimateur convergent* de $\frac{1}{\theta}$ (puisque $\lim_{n\to+\infty}Var(\frac{1}{\hat{\theta}})=0$.

Enfin, le calcul direct de *l'information de* FISHER pour la loi « parente » conduit à $I_X(\frac{1}{\theta})=E\left[(\frac{\partial}{\partial(\frac{1}{\theta})}\ln p(X,\theta))^2\right]$ avec $p(X,\theta)=(1-\theta)^{X-1}.\theta$, soit suivant la log- vraisemblance, $\ln p(X,\theta)=(X-1).\ln(1-\theta)+\ln\theta$. Il s'ensuit, par dérivation, $\frac{\partial}{\partial(\frac{1}{\theta})}\ln p(X,\theta)=\frac{\partial}{\partial\theta}\ln p(X,\theta)\times\frac{\partial\theta}{\partial(\frac{1}{\theta})}=\left[-\frac{(X-1)}{1-\theta}+\frac{1}{\theta}\right]\times(-\frac{1}{(\frac{1}{\theta})^2})=\frac{\theta.(\theta.X-1)}{(1-\theta)}$.

Ainsi, $I_X(\frac{1}{\theta})=\frac{\theta^2}{(1-\theta)^2}E\left[(\theta.X-1)^2\right]$. Mais, $E\left[(\theta.X-1)^2\right]=Var(\theta.X-1)=\theta^2.Var(X)$ puisqu'on a $E\left[(\theta.X-1)\right]=\theta.E(X)-1=\theta.\frac{1}{\theta}-1=0$, tout cela suivant les propriétés habituelles de l'espérance mathématique et de la variance. En définitive et rappelant que $Var(X)=\frac{1-\theta}{\theta^2}$, on obtient $I_X(\frac{1}{\theta})=\frac{\theta^2}{(1-\theta)^2}.\frac{\theta^2.(1-\theta)}{\theta^2}=\frac{\theta^2}{1-\theta}$.

S'agissant de l'échantillon $(X_1,X_2,...,X_n)$, l'information de FISHER, $I_{(X_1,X_2,...,X_n)}(\theta)$ qui est égale à $n.I_X(\theta)$, a donc pour valeur $\frac{n.\theta^2}{1-\theta}$.

Or, on a précédemment trouvé pour $Var(\frac{1}{\hat{\theta}})$, le résultat $\frac{1-\theta}{n.\theta^2}$ dont on constate en l'occurrence qu'il correspond également à la borne F.D.C.R de l'inégalité de CRAMER RAO, soit $\frac{1}{n.I_X(\frac{1}{\theta})}$.

L'estimateur $\frac{1}{\hat{\theta}}$ de $\frac{1}{\theta}$ forme donc également un *estimateur efficace* c'est-à-dire de variance minimale.

Pour conclure quant à cette illustration, il est manifeste que la statistique $S=\sum_{i=1}^{i=n} X_i$ est *exhaustive* pour θ (ou $\frac{1}{\theta}$) puisque $L(X,\theta)=\theta^n.(1-\theta)^{S-n}$ vérifie le *théorème de factorisation* $L(X,\theta)=g(S,\theta).h(X)$ avec $g(S,\theta)=\theta^n.(1-\theta)^{S-n}$ et $h(X)=1$.

• La *méthode du maximum de vraisemblance* s'applique aussi lorsque θ est **multidimensionnel** comme on pourra le constater dans certaines des applications développées plus loin (notamment, s'agissant de la loi normale).

3.4 Méthode des moindres carrés

Très utilisée pour déterminer les coefficients $\widehat{a_0},\widehat{a_1},...,\widehat{a_p}$ de la meilleure approximation affine d'une variable aléatoire Y en fonction de n variables aléatoires $X_1,X_2,...,X_n$ voire même d'autres modèles plus complexes de régression, cette méthode qui consiste à minimiser algébriquement ou géométriquement l'écart $\left\|Y-\sum_{i=1}^{i=n} a_i.X_i\right\|^2$ (avec $X_0=1$) est développée dans le chapitre IV.

3.5 Espérance, proportion, variance, covariance

Comme cela était indiqué dans le paragraphe 1 d'introduction du présent chapitre, l'inventaire des lois de probabilités de base montre que dans la plupart des cas, le (ou les) paramètres desdites lois coïncide avec un ou plusieurs des paramètres représentatifs que sont l'espérance mathématique et la variance, ou à défaut, une fonction desdits paramètres.

Par exemple, dans la loi de POISSON, le paramètre a désigne à la fois $E(X)$ et $Var(X)$.

Dans la loi normale, $m=E(X)$ et . Dans la binomiale, $p=\frac{1}{n}.E(X)$, n étant fixé.....

• Dès lors, les statistiques $\overline{X},F_n,\hat{S}^2$ étudiées dans le chapitre I, tiennent une place privilégiée puisque constituant immédiatement des estimateurs **sans biais et convergents**, respectivement, de $m=E(X),p,\sigma^2=Var(X)$.

En effet, on a montré que par exemple que $E(\overline{X})=m$, $E(\hat{S}^2)=\sigma^2$ ce qui établit le caractère sans biais des estimateurs $\overline{X}$ et $\hat{S}^2$. D'autre part, $\lim_{n\to+\infty} Var(\overline{X})=\lim_{n\to+\infty}\frac{\sigma^2}{n}=0$ et $\lim_{n\to+\infty} Var(\hat{S}^2)=\lim_{n\to+\infty}\left[\frac{E\left[(X-E(X))^4\right]}{n}-\frac{(n-3)}{n.(n-1)}.\sigma^4\right]=0$ ce qui établit le caractère convergent des estimateurs en question.

- Par contre, si la propriété d'être **asymptotiquement efficace** est assurée pour les trois estimateurs $\overline{X}, F_n, \hat{S}^2$ concernés (ce qui est largement suffisant dès que n est grand), *il en est autrement de l'efficacité* qui n'est pas assurée (notamment, pour ce qui est de la variance).

Par exemple, considérant un modèle gaussien de loi « parente » X de type $N(m,\sigma)$ (m et σ inconnus), l'estimateur $\hat{S}^2$ de σ^2, soit $\hat{S}^2 = \frac{1}{n-1}.\sum_{i=1}^{i=n}(X_i - \overline{X})^2$ vérifie $E(\hat{S}^2) = \sigma^2$ et $Var(\hat{S}^2) = \frac{2.\sigma^4}{n-1}$ (cf. application 1.2 du chapitre I).

Intéressons nous dès lors, à la classe plus large des estimateurs de la forme $T(\alpha) = \alpha.\hat{S}^2$ au sein desquels on va rechercher celui qui est le plus efficace. Le risque associé est fourni par $R(T(\alpha)) = E\left[(\alpha.\hat{S}^2 - \sigma^2)^2\right]$, soit par linéarité $R(T(\alpha)) = \alpha^2.E(\hat{S}^4) - 2\alpha.\sigma^2.E(\hat{S}^2) + \sigma^4$.

Explicitant $R(T(\alpha))$, on a :

$$E(\hat{S}^2) = \sigma^2 \text{ et } Var(\hat{S}^2) = E(\hat{S}^4) - \left[E(\hat{S}^2)\right]^2 \Rightarrow E(\hat{S}^4) = \frac{2.\sigma^4}{n-1} + \sigma^4 = \frac{n+1}{n-1}.\sigma^4.$$

En résumé, $R(T(\alpha)) = (\frac{n+1}{n-1}.\sigma^4).\alpha^2 - 2.\alpha.\sigma^4 + \sigma^4$, trinôme du second degré en α. qui est minimal pour la valeur $\alpha^* = \frac{n-1}{n+1}$. En cet optimum, $R(T(\alpha^*)) = \frac{2.\sigma^4}{n+1}$.

On a donc mis ici en évidence un estimateur dont le risque associé est plus faible que $\hat{S}^2$.

- *Plus généralement*, et dans le **cas multidimensionnel** d'un vecteur aléatoire $(X_1, X_2, ..., X_p)$ dont un échantillon de taille n associé est fourni par les n p-uplets :

$$(X_{11}, X_{12}, ..., X_{1j}, ..., X_{1p})$$
$$(X_{21}, X_{22}, ..., X_{2j}, ..., X_{2p})$$
$$...$$
$$(X_{i1}, X_{i2}, ..., X_{ij}, ..., X_{ip})$$
$$...$$
$$(X_{n1}, X_{n2}, ..., X_{nj}, ..., X_{pn})$$

on utilisera comme estimateurs des paramètres représentatifs $E(X_i), Var(X_i), \text{cov}(X_i, X_k)$ $(1 \le i \le n, 1 \le k \le p)$, les statistiques définies ci-après :

$$E(X_i) \approx \frac{1}{n}.\sum_{j=1}^{j=n} X_{ij} \qquad Var(X_i) \approx \frac{1}{n-1}.\sum_{j=1}^{j=n}(X_{ij} - \overline{X_i})^2 = \frac{1}{n-1}.(\sum_{j=1}^{j=n} X_{ij}^2 - n.\overline{X_i}^2)$$

$$Cov(X_i, X_k) = \frac{1}{n-1}.\sum_{j=1}^{j=n}(X_{ij} - \overline{X_i}).(X_{kj} - \overline{X_k}) = \frac{1}{n-1}.(\sum_{j=1}^{j=n} X_{ij}.X_{kj} - n.\overline{X_i}.\overline{X_k})$$

4. Estimation par intervalle de confiance

4.1 Construction de l'intervalle de confiance

Partant d'un échantillon $(X_1, X_2, ..., X_n)$ de variable parente X de loi $f(x,\theta)$ (resp. $p(x,\theta)$ dans le cas discret), il s'agit de trouver les **bornes** (a,b) (fonctions des X_i) telles que $\Pr ob(a \le \theta \le b) = 1 - \alpha$ ($1-\alpha$ **seuil de confiance** fixé à priori et égal le plus souvent à 0,90, 0,95, ou 0,99).

• Pour construire un tel intervalle, on s'appuiera sur une fonction dite « **pivotale pour** θ », c'est-à-dire une fonction des X_i dont la loi ne dépend pas de θ. En effet, désignant par $h(X_1, X_2, ..., X_n, \theta)$ une telle fonction et supposant qu'on puisse déterminer numériquement u_1 et $u_2 / \Pr ob(u_1 \le h(X_1, X_2, ..., X_n) \le u_2) = \alpha$, la résolution en θ de la double inéquation $u_1 \le h(X_1, X_2, ..., X_n) \le u_2$ conduit immédiatement à l'encadrement cherché $g_1(X_1, X_2, ..., X_n) \le \theta \le g_2(X_1, X_2, ..., X_n)$.

Par exemple, en se référant aux distributions d'échantillonnage développées dans le chapitre I :

- $\dfrac{\overline{X} - m}{\sigma / \sqrt{n}}$ et $\dfrac{\overline{X} - m}{\hat{S} / \sqrt{n}}$ sont *pivotales pour* m lorsque X est *normale* de loi $N(m, \sigma^2)$;

- plus généralement, $\dfrac{\overline{X} - m}{\sigma / \sqrt{n}}$ est *asymptotiquement pivotale pour* m *pour toute variable aléatoire* X ;

- $\dfrac{(n-1).\hat{S}^2}{\sigma^2}$ est *pivotale pour* σ^2 lorsque X est normale de loi $N(m, \sigma^2)$.

• *Plus généralement,* considérant **l'estimateur E.M.V**, soit $\hat{\theta}$, fourni par la *méthode du maximum de vraisemblance,* la fonction $\dfrac{\hat{\theta} - \theta}{\sqrt{1 / n.I_X(\theta)}} = \sqrt{n}.\sqrt{I_X(\theta)}.\left[\hat{\theta} - \theta\right]$ est **asymptotiquement pivotale**.

• Des *illustrations de la construction de l'intervalle de confiance* sont données ci-dessous pour les cas les plus courants de la *moyenne,* de la *proportion,* et de la *variance,* ces calculs étant conduits suivant deux considérations, à savoir *lois symétriques ou non, grands ou petits échantillons.*

4.2 Le cas d'une moyenne

Le caractère symétrique des fonctions pivotales conduit ici à **centrer** l'intervalle de confiance (a, b) sur *l'information possédée* qui en l'occurrence est *l'estimateur ponctuel* $\overline{X}$. Ecrivant donc (a, b) sous la forme $(\overline{X} - \varepsilon, \overline{X} + \varepsilon)$, on cherche ε tel que :

$$\Pr ob(\overline{X} - \varepsilon \le m \le \overline{X} + \varepsilon) = 1 - \alpha \quad \text{(E)}$$

• **Pour** n **grand** $(n \ge 30)$, on peut admettre que $\dfrac{\overline{X} - m}{\sigma / \sqrt{n}}$ converge vers la loi normale centrée réduite $N(0,1)$, X de loi connue ou non. L'équation (E) se ramène donc à :

$$\Pr ob(\left|\overline{X} - m\right| \le \varepsilon) = \Pr ob(\left|\xi\right| \le \frac{\varepsilon}{\sqrt{n}}) = 1 - \alpha$$

Déterminant par lecture dans les tables des valeurs de la loi normale $N(0,1)$ le nombre t_α vérifiant $\Pr ob(\left|\xi\right| \le t_\alpha) = 1 - \alpha$, il vient immédiatement, l'intervalle cherché :

$$\overline{X} - t_\alpha . \frac{\sigma}{\sqrt{n}} \le m \le \overline{X} + t_\alpha . \frac{\sigma}{\sqrt{n}}$$

A noter que si σ est inconnu, on sera contraint d'utiliser son estimateur ponctuel $\hat{S}$, approximation néanmoins admissible pour n grand.

• **Pour n petit** ($n \prec 30$), la connaissance de la loi de X est indispensable pour mener les calculs à partir de $\overline{X}$. Dans le cas courant des échantillons gaussiens (X de loi normale $N(m,\sigma^2)$, $\dfrac{\overline{X}-m}{\sigma/\sqrt{n}}$ suit la loi $N(0,1)$ si bien qu'on retrouve le résultat précédent :

$$\overline{X} - t_\alpha \cdot \frac{\sigma}{\sqrt{n}} \le m \le \overline{X} + t_\alpha \cdot \frac{\sigma}{\sqrt{n}}$$

Toutefois, *si σ est inconnu*, l'approximation de σ par $\hat{S}$ n'est plus acceptable, n étant faible. Mais on peut faire appel alors à la *fonction pivotale* $\dfrac{\overline{X}-m}{\hat{S}/\sqrt{n}}$ dont on sait qu'elle suit la loi de STUDENT, soit $T(n-1)$ à $\nu = n-1$ degré de libertés (cf. chapitre I). Il s'ensuit le résultat :

$$\overline{X} - t_\alpha \cdot \frac{\hat{S}}{\sqrt{n}} \le m \le \overline{X} + t_\alpha \cdot \frac{\hat{S}}{\sqrt{n}}$$

t_α vérifiant $Prob(|T| \le t_\alpha) = 1-\alpha$ où T désigne la variable de STUDENT à $n-1$ degrés de libertés.

4.3 Le cas d'une proportion

Assimilant X à la loi de BERNOULLI et la proportion inconnue à la moyenne de ladite loi, on est ramené ici au cas précédent d'une moyenne dans lequel par analogie, $\overline{X} = \dfrac{\sum_{i=1}^{i=n} X_i}{n} \equiv F_n$, $m \equiv p$, et $\sigma^2 = Var(X_i) \equiv p.q$ (avec $q = 1-p$). Dans le cas de grands échantillons ($n \ge 30$), on obtient immédiatement l'intervalle :

$$F_n - t_\alpha \cdot \sqrt{\frac{p.q}{n}} \le p \le F_n + t_\alpha \cdot \sqrt{\frac{p.q}{n}}$$

Toutefois, p est inconnu ici. Trois solutions sont suggérées pour lever cette difficulté :

1°) Considérer par excès, l'encadrement :

$$F_n - \frac{t_\alpha}{2.\sqrt{n}} \le p \le F_n + \frac{t_\alpha}{2.\sqrt{n}}$$

2°) Utiliser l'approximation de p par F_n, ce qui conduit à l'encadrernent :

$$F_n - t_\alpha \cdot \sqrt{\frac{F_n.(1-F_n)}{n}} \le p \le F_n + t_\alpha \cdot \sqrt{\frac{F_n.(1-F_n)}{n}}$$

2°) Chercher les bornes p_1 et p_2 dans lesquelles p est compris lorsque :

$$(p-F_n)^2 \le t_\alpha^2 \cdot \frac{p.(1-p)}{n}$$

Pour ce qui est du 1°) on remarque que $p.(1-p)$ est maximal lorsque $p = \dfrac{1}{2}$.

Il est bien évident dans ces conditions que $|F_n - p| \le t_\alpha . \sqrt{\frac{p.q}{n}}$ est vérifié dès lors qu'on a $|F_n - p| \le t_\alpha . \underset{0 \le p \le 1}{Max} \sqrt{\frac{p.q}{n}} = \frac{t_\alpha}{2.\sqrt{n}}$, d'où le résultat cherché.

Par ailleurs, pour ce qui est du 3°), il faut signaler l'existence d'abaques qui sont susceptibles de simplifier les choses.

4.4 Le cas d'une variance

Cette fois *les fonctions pivotales* $\frac{n.S^2}{\sigma^2}$ (lorsque m est connu) et $\frac{(n-1).\hat{S}^2}{\sigma^2}$ (lorsque m est inconnu) *ne sont pas symétriques*. Plutôt que de centrer $[a,b]$ sur $\hat{\theta}$ (ici σ^2), c'est par rapport aux probabilités qu'on va chercher une symétrie en résumant la détermination de $(a,b) / \Pr ob(a \le \theta \le b) = 1 - \alpha$ à celles des bornes a,b vérifiant $\Pr ob(\theta \prec a) = \frac{\alpha}{2}$ et $\Pr ob(\theta \succ b) = \frac{\alpha}{2}$.

• Ainsi, *pour le cas de la variance* et lorsque m **est connu** :

$$\Pr ob(\sigma^2 \prec a) = \frac{\alpha}{2} \Rightarrow \Pr ob(\frac{n.S^2}{\sigma^2} \succ \frac{n.S^2}{a}) = \frac{\alpha}{2}$$

où la *fonction pivotale* $\frac{n.S^2}{\sigma^2}$ suit la *loi du chi-deux à n degrés de libertés*, soit $\chi^2(n)$ (cf. chapitre I). Lisant dans les tables de valeurs (cf. annexes), le nombre t_1 vérifiant $\Pr ob(\chi^2(n) \succ t_1) = \frac{\alpha}{2}$, il vient $a = \frac{n.S^2}{t_1}$. De même, la lecture du nombre t_2 tel que $\Pr ob(\chi^2(n) \prec t_2) = \frac{\alpha}{2}$ entraîne $b = \frac{n.S^2}{t_2}$, d'où en définitive, l'intervalle cherché :

$$\frac{n.S^2}{t_1} \le \sigma^2 \le \frac{n.S^2}{t_2}$$

• Lorsque m **est inconnu**, la *fonction pivotale* $\frac{(n-1).S^2}{\sigma^2}$ suit la *loi du chi-deux à n-1 degrés de libertés*, soit $\chi^2(n-1)$. On aura donc, pour ce cas, l'intervalle :

$$\frac{(n-1).\hat{S}^2}{t_1} \le \sigma^2 \le \frac{(n-1).\hat{S}^2}{t_2}$$

où t_1 et t_2 sont lus dans la table des valeurs de la loi $\chi^2(n-1)$ et non pas la loi $\chi^2(n)$ comme précédemment.

B - Applications

1. Exemples de modèles et propriétés des estimateurs

1.1 Modèle gaussien

Enoncé : L'étude des estimateurs de m et σ^2 est proposée ici dans le cadre d'un échantillon $(X_1, X_2, ..., X_n)$ de n variables aléatoires indépendantes équidistribuées de loi parente X de type « loi normale », soit $N(m,\sigma^2)$.

PARTIE I

On s'intéresse en premier lieu, dans cette partie, à la moyenne $m = E(X)$.

I-1°) Déterminer l'estimateur du maximum de vraisemblance de la moyenne m.

I-2°) Calculer l'information de FISHER, $I_{(X_1,X_2,...,X_n)}(m)$ pour le paramètre m. Qu'en conclure quant à l'efficacité de l'estimateur trouvé à la question précédente ?

PARTIE II

On étudie ici, un estimateur de la variance σ^2, m étant supposé connu.

II-1°) Déterminer l'estimateur du maximum de vraisemblance de σ^2.

II-2°) Calculer l'information de FISHER, $I_{(X_1,X_2,...,X_n)}(\sigma^2)$ pour le paramètre σ^2. Qu'en conclure quant à l'efficacité de l'estimateur ainsi obtenu en II-1°) ?.

II-3°) On suppose dans cette question que m est inconnue. On considère la statistique $\hat{S}^2 = \frac{1}{n-1}.\sum_{i=1}^{i=n}(X_i - \overline{X})^2$. Tout en étant sans biais, cet estimateur est-il efficace ?

Solution : I-1°) La *fonction de vraisemblance* associée à l'échantillon considéré est égale à $L(X_1, X_2, ..., X_n, m, \sigma^2) = \prod_{i=1}^{i=n} \frac{1}{\sigma.\sqrt{2\pi}}.\exp(-\frac{1}{2\sigma^2}(X_i - m)^2)$. Il s'ensuit, pour ce qui est de la *fonction de log-vraisemblance*, l'expression :

$$\ln L = -\frac{1}{2\sigma^2}.\sum_{i=1}^{i=n}(X_i - m)^2 - n.\ln\sigma - \frac{n}{2}.\ln(2\pi)$$

L'estimateur du maximum de vraisemblance de m, est solution de l'équation $\frac{\partial L}{\partial m} = 0$, soit $\frac{1}{\sigma^2}.\sum_{i=1}^{i=n}(X_i - m) = 0$, d'où la solution cherchée $m = \frac{\sum_{i=1}^{i=n} X_i}{n}$ qui coïncide avec la *moyenne empirique* $\overline{X}$.

Il s'agit bien d'un maximum puisque $\frac{\partial^2}{\partial m^2} L = -\frac{n}{\sigma^2} < 0$.

I-2°) L'information de FISHER fournie par la loi parente X de loi normale $N(m,\sigma^2)$ relativement au paramètre m est égale à $I_X(m) = -E\left[\frac{\partial^2}{\partial m^2}\ln f(X,m)\right]$ avec $f(X,m) = \frac{1}{\sigma.\sqrt{2\pi}}.\exp(-\frac{1}{2\sigma^2}.(X-m)^2)$.

On a $\ln f(X,m) = -\ln(\sigma.\sqrt{2\pi}) - \frac{1}{2\sigma^2}.(X-m)^2$, soit $\frac{\partial}{\partial m}\ln f(X,m) = \frac{1}{\sigma^2}.(X-m)$ et $\frac{\partial^2}{\partial m^2}\ln f(X,m) = -\frac{1}{\sigma^2}$.

En définitive $I_X(m) = -E(-\frac{1}{\sigma^2}) = \frac{1}{\sigma^2}$ et s'agissant de l'échantillon $(X_1, X_2, ..., X_n)$, la quantité d'information associée (égale à $n.I_X(m)$) est donc $I_{(X_1,X_2,...,X_n)}(m) = \frac{n}{\sigma^2}$.

Or $Var(\overline{X}) = \frac{\sigma^2}{n}$, ce qui correspond également à la *borne minimale* F.D.C.R de FRECHET, DARMOIS, CRAMER, RAO fournie par l'inverse de la quantité d'information de FISHER, soit $\frac{1}{I_{(X_1,X_2,...,X_n)}(m)}$. $\overline{X}$ forme est donc ici un *estimateur sans biais de variance minimale* (cf. paragraphe 2.3 du rappel de cours).

• D'ailleurs, *la loi normale appartient à la famille exponentielle* puisque la log-vraisemblance $\ln f(X,m) = -\frac{1}{2\sigma^2}.(X-m)^2 - \ln(\sigma.\sqrt{2\pi})$, soit en développant, $\ln f(X,m) = \frac{m}{\sigma^2}.X - \frac{m^2}{2\sigma^2} - \frac{X^2}{2\sigma^2} - \ln(\sigma.\sqrt{2\pi})$, c'est-à-dire une expression identifiable à la forme $Q(m).T(X) + \beta(m) + \gamma(X)$. La statistique $T(X) = \sum_{i=1}^{i=n} X_i$ forme donc une *statistique exhaustive*, $\overline{X} = \frac{\sum_{i=1}^{i=n} X_i}{n}$ qui est un estimateur sans biais, vérifiant la condition nécessaire et suffisante d'efficacité (cf. paragraphe 2.5 du rappel de cours). On retrouve donc par cette méthode le résultat du calcul précédent quant aux propriétés de $\overline{X}$.

II-1°) Partant de nouveau de la fonction de *log –vraisemblance* associée à l'échantillon $(X_1, X_2, ..., X_n)$, soit $\ln L = -\frac{1}{2\sigma^2}.\sum_{i=1}^{i=n}(X_i - m)^2 - n.\ln\sigma - \frac{n}{2}.\ln(2\pi)$, le calcul de $\frac{\partial}{\partial\sigma^2}\ln L$ conduit en utlisant le changement de variable $u = \sigma^2$, au résultat :

$$\frac{\partial}{\partial u}\ln L = \frac{\partial}{\partial\sigma}\ln L \times \frac{\partial\sigma}{\partial u} = \frac{1}{2.\sqrt{u}}.\frac{\partial}{\partial\sigma}\ln L = \frac{1}{2\sigma}.\frac{\partial}{\partial\sigma}\ln L$$

soit, $\frac{\partial}{\partial u}\ln L = \frac{1}{2\sigma}.\left[\frac{2}{2\sigma^3}.\sum_{i=1}^{i=n}(X_i - m)^2 - \frac{n}{\sigma}\right] = \frac{1}{2\sigma^4}.\sum_{i=1}^{i=n}(X_i - m)^2 - \frac{n}{2\sigma^2}$.

L'estimateur du maximum de vraisemblance de σ^2 qui est solution de $\frac{\partial}{\partial\sigma^2}\ln L = 0$ est donc égal à $\hat{\sigma}^2 = \frac{1}{n}.\sum_{i=1}^{i=n}(X_i - m)^2$. On retrouve donc ici la statistique S^2 définie et étudiée au chapitre I.

II-2°) Se référant à l'application 1.2 du chapitre I, on a $E(S^2) = \sigma^2$ et $Var(S^2) = \frac{2.\sigma^4}{n}$.

• L'information de FISHER fournie par l'échantillon $(X_1, X_2, ..., X_n)$ relativement au paramètre σ^2 est égale à $I_{(X_1,X_2,...,X_n)}(\sigma^2) = -E\left[\frac{\partial^2}{\partial(\sigma^2)^2}\ln L(X_1, X_2, ..., X_n, \sigma^2)\right]$.

On a $\ln L = -\frac{1}{2\sigma^2}.\sum_{i=1}^{i=n}(X_i - m)^2 - n.\ln\sigma - \frac{n}{2}.\ln(2\pi)$, puis par dérivations successives, $\frac{\partial}{\partial\sigma^2}\ln L = \frac{1}{2\sigma^4}.\sum_{i=1}^{i=n}(X_i - m)^2 - \frac{n}{2\sigma^2}$ et $\frac{\partial^2}{\partial\sigma^4}\ln L = -\frac{1}{\sigma^6}.\sum_{i=1}^{i=n}(X_i - m)^2 + \frac{n}{2\sigma^4}$,. cette dernière expression découlant ici encore du changement de variable $u = \sigma^2$, et de la relation $\frac{\partial}{\partial u}(\frac{\partial}{\partial u}\ln L) = -\frac{1}{u^3}.\sum_{i=1}^{i=n}(X_i - m)^2 + \frac{n}{2u^2}$.

Revenant à $I_{(X_1,X_2,...,X_n)}(\sigma^2) = -E\left[\frac{\partial^2}{\partial\sigma^4}\ln L\right]$, c'est donc, compte tenu de la linéarité de l'espérance mathématique et remarquant que $\sum_{i=1}^{i=n}(X_i - m)^2 = n.S^2$:

$$I_{(X_1,X_2,...,X_n)}(\sigma^2) = -\left[-\frac{1}{\sigma^6}.E(n.S^2) + \frac{n}{2\sigma^4}\right]$$

Mais, $E(n.S^2) = n.E(S^2) = n.\sigma^2$. En conclusion :

$I_{(X_1,X_2,...,X_n)}(\sigma^2) = -\left[-\frac{n.\sigma^2}{\sigma^6} + \frac{n}{2\sigma^4}\right] = \frac{n}{2\sigma^4}$, soit l'inverse de la variance $Var(S^2) = \frac{2\sigma^4}{n}$.

• S^2 qui est *sans biais* et dont *la variance atteint la borne minimale* F.D.C.R égale à l'inverse de la quantité d'information de FISHER, est donc un *estimateur efficace*.

II-3°) Supposant la moyenne m inconnue et considérant la statistique $\hat{S}^2 = \frac{1}{n-1}.\sum_{i=1}^{i=n}(X_i - \overline{X})^2$, les résultats obtenus dans l'application 1.2 du chapitre I montrent que $E(\hat{S}^2) = \sigma^2$ et $Var(\hat{S}^2) = \frac{2.\sigma^4}{n-1}$. Ansi $\hat{S}^2$ est-il un *estimateur sans biais et convergent* de σ^2 puisque $\lim_{n\to+\infty} Var(\hat{S}^2) = 0$.

Suivant l'inégalité de CRAMER-RAO, la borne F.D.C.R est égale à $\frac{2.\sigma^4}{n}$. Tout en étant sans biais, $\hat{S}^2$ *ne constitue donc pas ici un estimateur efficace* de σ^2 puisque de variance supérieure à la dite borne. En effet, $Var(\hat{S}^2) = \frac{2.\sigma^4}{n-1} \succ \frac{2.\sigma^4}{n}$.

1.2 Modèle de POISSON

On a le cas ici, d'un paramètre qui est à la fois espérance mathématique et variance et pour lequel il y a donc plusieurs estimateurs qu'il est proposé de comparer.

Enoncé : On considère un échantillon de taille n, soit $(X_1, X_2, ..., X_n)$, de n variables aléatoires équidistribuées et indépendantes, de même loi « parente » de type POISSON (de paramètre θ).

1°) A l'aide de la méthode des moments, en déduire deux estimateurs sans biais de θ, soient T_1 et T_2.

2°) Déterminer l'estimateur du maximum de vraisemblance de θ et retrouver ainsi l'un des résultats précédents.

3°) Comparer T_1 et T_2. Peut-on trouver un estimateur sans biais plus efficace ?

Solution : 1°) On a simultanément $\theta = E(X)$ et $\theta = Var(X) = E(X^2) - E(X)^2$. L'utilisation des *distributions empiriques d'échantillonnage* développées au chapitre I, conduit immédiatement aux deux estimateurs sans biais $T_1 = \frac{1}{n}.\sum_{i=1}^{i=n} X_i$ et $T_2 = \frac{1}{n-1}.\sum_{i=1}^{i=n}(X_i - \overline{X})^2$ de θ, avec $\overline{X} = \frac{\sum_{i=1}^{i=n} X_i}{n}$.

2°) La *méthode du maximum de vraisemblance* appliquée à θ et à la fonction de vraisemblance $L(X_1, X_2, ..., X_n, \theta) = \prod_{i=1}^{i=n} \frac{\theta^{X_i}.e^{-\theta}}{X_i!}$, conduit à rechercher le maximum de la log-vraisemblance, soit la solution $\hat{\theta}$ de l'équation $\frac{\partial}{\partial\theta}\ln L = 0$. Il s'ensuit successivement $\ln L = -n.\theta + (\sum_{i=1}^{i=n} X_i).\ln\theta - \sum_{i=1}^{i=n}\ln(X_i!)$ puis $\frac{\partial}{\partial\theta}\ln L = -n + (\sum_{i=1}^{i=n} X_i).\frac{1}{\theta}$.

D'où en définitive, la solution cherchée $\hat{\theta} = \frac{\sum_{i=1}^{i=n} X_i}{n}$. On remarquera qu'on a bien un maximum puisque $\frac{\partial^2}{\partial\theta^2}\ln L = -\frac{\sum_{i=1}^{i=n} X_i}{\theta^2} \prec 0$. On constate donc que l'estimateur E.M.V ainsi obtenu coïncide avec T_1.

3°) $Var(T_1) = \frac{1}{n^2}.\sum_{i=1}^{i=n} Var(X_i) = \frac{n.\theta}{n^2} = \frac{\theta}{n}$ (en effet, s'agissant de la loi de POISSON de paramètre θ, il est rappelé que $Var(X) = \theta$. D'autre part, et suivant un résultat établi dans l'application 1.2 du chapitre I, on a :

$$Var(T_2) = \frac{\mu_4}{n} - \frac{(n-3)}{n.(n-1)}.\sigma^4$$

où μ_4 désigne le moment d'ordre 4 de la variable centrée $X - E(X)$, soit $\mu_4 = E\left[(X-\theta)^4\right]$ puisque $E(X) = \theta$. Développant μ_4 suivant la linéarité de l'espérance mathématique, il vient $\mu_4 = E(X^4) - 4.\theta.E(X^3) + 6.\theta^2.E(X^2) - 4.\theta^3.E(X) + \theta^4$.

• Calculant les différents moments d'ordre $h(1 \le h \le 4)$ ci-dessus, il vient $E(X) = \theta, E(X^2) = Var(X) + E(X)^2 = \theta + \theta^2$. Par ailleurs, utilisant la *fonction génératrice* $\Psi_X(t) = E\left[t^X\right]$, il est rappelé que s'agissant de la loi de POISSON, on a $\Psi_X(t) = \sum_{x=0}^{x=+\infty} \frac{t^x.\theta^x.e^{-\theta}}{x!} = e^{\theta.(t-1)}$. Les dérivations successives de $\Psi_X(t)$ conduisent d'une part aux expressions de $E(X)$ et de $E(X^2)$ déjà mentionnées ci-dessus, mais également, à des ordres supérieurs, à l'expression de $E(X^h), h \ge 3$.

Ainsi :

$$\psi'_X(t) = E\left[X.t^{X-1}\right] \Rightarrow \Psi'_X(1) = E(X)$$

$$\Psi''_X(t) = E\left[X.(X-1).t^{X-2}\right] \Rightarrow \Psi''_X(1) = E(X^2) - E(X)$$

$$\Psi'''_X(t) = E\left[X.(X-1).(X-2).t^{X-3}\right] \Rightarrow \Psi'''_X(1) = E(X^3) - 3.E(X^2) + 2.E(X)$$

$$\Psi''''_X(t) = E\left[X.(X-1).(X-2).(X-3).t^{X-4}\right] \Rightarrow \Psi''''_X(1) = E(X^4) - 6.E(X^3) + 11.E(X^2) - 6.E(X)$$

Mais, $\Psi'_X(t) = \theta.e^{\theta.(t-1)} \Rightarrow \Psi'_X(1) = \theta$ et de même $\Psi''_X(1) = \theta^2, \Psi'''_X(1) = \theta^3, \Psi''''_X(1) = \theta^4$.

En conclusion, et compte tenu des équations ci-dessus, les moments d'ordre h cherchés ont pour expressions :

$$E(X^3) = \Psi'''_X(1) + 3.\left[\Psi''_X(1) + \Psi'_X(1)\right] - 2.\Psi'_X(1) = \theta^3 + 3.(\theta^2 + \theta) - 2.\theta = \theta^3 + 3.\theta^2 + \theta.$$

$$E(X^4) = \theta^4 + 6.(\theta^3 + 3.\theta^2 + \theta) - 11.(\theta + \theta^2) + 6.\theta = \theta^4 + 6.\theta^3 + 7.\theta^2 + \theta.$$

Rassemblant ces divers résultats autour du calcul de μ_4, on aboutit à :

$$\mu_4 = (\theta^4 + 6.\theta^3 + 7.\theta^2 + \theta) - 4.(\theta^4 + 3.\theta^3 + \theta^2) + 6.(\theta^3 + \theta^4) - 4.\theta^4 + \theta^4 = 3.\theta^2 + \theta$$

On en déduit $Var(T_2) = \dfrac{3.\theta^2 + \theta}{n} - \dfrac{(n-3).\theta^2}{n.(n-1)}$ (puisque $\theta = Var(X)$), soit après développement et simplifications, $Var(T_2) = \dfrac{\theta}{n} + \dfrac{2.\theta^2}{n-1}$.

La *comparaison des risques associés* aux deux estimateurs T_1 et T_2, soient $Var(T_1)$ et $Var(T_2)$, puisqu'il s'agit d'estimateurs sans biais, conduit à la relation :

$$Var(T_2) - Var(T_1) = \frac{2.\theta^2}{n-1} \succ 0$$

T_1 et T_2 sont tous deux *convergents en probabilité* puisque $\lim_{n\to+\infty} Var(T_1) = \lim_{n\to+\infty} Var(T_2) = 0$. Mais parmi ces deux estimateurs, *c'est donc T_1 qui est le plus efficace* (puisque de variance plus faible).

• En fait, comme cela est manifeste à travers le rappel de cours, T_1 est bien le *meilleur estimateur sans biais* de θ, $\sum_{i=1}^{i=n} X_i$ formant une *statistique exhaustive* et la loi de POISSON *faisant partie de la famille exponentielle.*

D'ailleurs, le calcul direct de la borne F.D.C.R de FRECHET, DARMOIS, CRAMER, RAO, conduit à $\dfrac{1}{I_{(X_1,X_2,\ldots,X_n)}(\theta)}$ avec $I_{(X_1,X_2,\ldots,X_n)}(\theta) = -E\left[-\dfrac{\sum_{i=1}^{i=n} X_i}{\theta^2}\right] = \dfrac{n}{\theta}$.

Cette borne F.D.C.R est bien atteinte par T_1 puisque $Var(T_1) = \dfrac{\theta}{n} = \dfrac{1}{I_{(X_1,X_2,\ldots,X_n)}(\theta)}$.

1.3 Modèle uniforme

Cet exemple met en évidence l'une des conditions de validité de l'inégalité de CRAMER RAO qui, pour le cas de la loi uniforme, n'est pas vérifiée.

Enoncé : Soit U une variable aléatoire uniforme sur $[0,\theta], \theta \succ 0$. La borne supérieure θ de l'intervalle de définition de U est inconnue et on cherche à l'estimer à l'aide d'un échantillon de la taille n, soit $(U_1,U_2,...,U_n)$.

PARTIE I

On se propose de constater ici que l'inégalité de CRAMER RAO n'est pas applicable comme d'ailleurs la seconde expression de la quantité d'information de FISHER, à savoir $I_n(\theta) = E\left[-\frac{\partial^2}{\partial\theta^2}\ln L\right]$.

I-1°) Exprimer la fonction de vraisemblance $L(U_1,U_2,...,U_n,\theta)$ pour le paramètre θ et un échantillon $(U_1,U_2,...,U_n)$ de n valeurs indépendantes prises par U.

I-2°) Calculer l'information de FISHER, $I_n(\theta)$, fournie par l'échantillon $(U_1,U_2,...,U_n)$ pour l'estimation de θ.

I-3°) Qu'en conclure ?

PARTIE II

On étudie ici les propriétés de la statistique $V = Max(U_1,U_2,...,U_n)$.

II-1°) Expliciter la loi de probabilité de V.

II-2°) Calculer $E(V)$ et $VarV$.

II-3°) Montrer que V est une statistique exhaustive.

PARTIE III

Dans cette partie, il est proposé d'expliciter des estimateurs convergents de θ suivant les méthodes des moments et du maximum de vraisemblance, soient respectivement T_1 et T_2 et d'énumérer leurs qualités respectives.

III-1°) Expliciter T_1 et montrer qu'il s'agit d'un estimateur sans biais et convergent.

III-2°) Montrer que T_2 est défini par la statistique V étudiée en partie II et indiquer les propriétés de T_2.

III-3°) Comparer l'efficacité des deux estimateurs T_1 et T_2 ainsi obtenus.

PARTIE IV

Il est proposé ici d'améliorer T_1 à l'aide du théorème de RAO-BLACKWELL pour aboutir à un estimateur optimal.

IV-1°) Expliciter $E(T_1/T_2)$ et en déduire que l'estimateur $T_2^* = \frac{n+1}{n}.T_2$ constitue un estimateur sans biais préférable à T_1.

IV-2°) Comparer directement T_1 et T_2^* pour vérifier le résultat précédent.

IV-3°) Comparer T_2 et T_2^*. Qu'en conclure ?

Solution : I-1°) La variable « parente » considérée ici est définie par la densité de probabilité $f(u,\theta)=\frac{1}{\theta}.1_{[0,\theta]}(u)$, $1_{[0,\theta]}(u)$ étant la *fonction indicatrice* associée à l'intervalle $[0,\theta]$.

La *fonction de vraisemblance* associée à l'échantillon $(U_1,U_2,...,U_n)$ est égale au produit des densités de probabilités soit :

$$L(U_1,U_2,...,U_n,\theta)=\prod_{i=1}^{i=n}\frac{1}{\theta}.1_{[0,\theta]}(U_i)=\frac{1}{\theta^n}$$

avec $0\le U_i\le\theta,\forall i/1\le i\le n\Rightarrow \underset{1\le i\le n}{Max}U_i\le\theta$). En bref :

$$L(U_1,U_2,...,U_n,\theta)=\frac{1}{\theta^n}.1_{v=\underset{1\le i\le n}{Max}U_i\le\theta}(v)$$

I-2°) Par définition, $I_n(\theta)=E\left[(\frac{\partial}{\partial\theta}\ln L)^2\right]$. Or, $\ln L=-n.\ln\theta$ et $\frac{\partial}{\partial\theta}\ln L=-\frac{n}{\theta}$. Ainsi, rappelant que $0\le U_i\le\theta,\forall i/1\le i\le n$, $I_n(\theta)=E\left[\frac{n^2}{\theta^2}\right]=\int_{[0,\theta]^n}\int...\int\frac{n^2}{\theta^2}.\frac{1}{\theta^n}.dU_1.dU_2...dU_n$. soit le produit des intégrales simples $I_n(\theta)=\frac{n^2}{\theta^2.\theta^n}.\prod_{i=1}^{i=n}\int_0^\theta du=\frac{n^2}{\theta^2}$.

I-3°) Considérant la *quantité d'information de* FISHER, soit $I_1(\theta)$ fournie par la variable « parente » U de loi uniforme sur $[0,\theta]$, on a immédiatement $L(U,\theta)=\frac{1}{\theta}$ pour $0\le U\le\theta$ et $\ln L(U,\theta)=-\ln\theta$. Ainsi, $I_1(\theta)=E\left[(\frac{\partial}{\partial\theta}\ln L(U,\theta))^2\right]=E\left[\frac{1}{\theta^2}\right]$, soit en explicitant, $I_1(\theta)=\int_0^\theta\frac{1}{\theta^2}.\frac{du}{\theta}=\frac{1}{\theta^2}$.

• On remarque que la relation $I_n(\theta)=n.I_1(\theta)$ n'est pas vérifiée ici puisque $I_n(\theta)=\frac{n^2}{\theta^2}$ et non pas $\frac{n}{\theta^2}$. D'ailleurs, le calcul de $I_n(\theta)$ suivant la deuxième expression de la quantité d'information, à savoir, $I_n(\theta)=-E\left[\frac{\partial^2}{\partial\theta^2}\ln L\right]=E\left[\frac{n}{\theta^2}\right]=\frac{n}{\theta^2}$. En fait, ni cette dernière expression, ni l'inégalité de CRAMER RAO sont applicables pour le modèle considéré dont le domaine de définition de la loi de probabilités de la variable « parente » n'est pas indépendant du paramètre θ à estimer (se reporter au rappel de cours, paragraphe 2.4, pour ce qui est de ladite condition).

II-1°) Notant par $H(V)$ la fonction de répartition de V et par $h(v)$ sa densité de probabilité, on a d'une part $H(v)=\Pr ob(V\le v)=\Pr ob(U_1\le v,U_2\le v,...,U_n\le v)$ et d'autre part $h(v)=\frac{dH(v)}{dv}$.

Explicitant les calculs, $H(v)=\prod_{i=1}^{i=n}\Pr ob(U_i\le v)=(\int_0^v\frac{1}{\theta}.du)^n=\frac{v^n}{\theta^n}$. D'où la densité de probabilité $h(v)$ de la variable V, soit $h(v)=\frac{n.v^{n-1}}{\theta^n}$.

II-2°) $E(V)=\int_0^\theta v.\frac{n.v^{n-1}}{\theta^n}.dv=\frac{n}{\theta^n}.\frac{\theta^{n+1}}{n+1}=\frac{n.\theta}{n+1}$.

De même, $E(V^2)=\int_0^\theta v^2.\frac{n.v^{n-1}}{\theta^n}.dv=\frac{n}{\theta^n}.\frac{\theta^{n+2}}{n+2}=\frac{n}{n+2}.\theta^2$. Il en résulte l'expression de la variance $Var(V)=E(V^2)-E(V)^2$, soit $Var(V)=\left[\frac{n}{n+2}-\frac{n^2}{(n+1)^2}\right].\theta^2=\frac{n.\theta^2}{(n+2).(n+1)^2}$.

II-3°) La statistique $L(U_1,U_2,...,U_n,\theta)=\frac{1}{\theta^n}$ pour $\underset{1\le i\le n}{MaxU_i}\le\theta$, se décompose sous la forme $h(v,\theta).g(u)$ où $h(v,\theta)=\frac{n.v^{n-1}}{\theta^n}$ et $g(u)=\frac{1}{n.(\underset{1\le i\le n}{MaxU_i})^{n-1}}$. Ainsi la statistique $V=\underset{1\le i\le n}{MaxU_i}$ est-elle *exhaustive* compte tenu du *théorème de factorisation de* FISHER énoncé précédemment (cf. paragraphe 2.5 des rappels de cours).

III-1°) Revenant à la variable U de loi uniforme sur $[0,\theta]$, on a immédiatement :

$$E(U)=\int_0^\theta u.\frac{du}{\theta}=\left[\frac{u^2}{2.\theta}\right]_0^\theta=\frac{\theta}{2}.$$

La **méthode des moments** conduit donc à estimer $\theta=2.E(U)$ par $T_1=2.\overline{U}$ avec $\overline{U}=\frac{\sum_{i=1}^{i=n}U_i}{n}$. Il est immédiat que $E(T_1)=\frac{2}{n}.\sum_{i=1}^{i=n}E(U_i)=\theta$ suivant la linéarité de l'espérance mathématique. Ainsi T_1 est-il un *estimateur sans biais* de θ.

D'autre part, $Var(T_1)=4.Var(\overline{U})=\frac{4}{n^2}.\sum_{i=1}^{i=n}Var(U_i)$ (suivant les propriétés classiques de la variance). Mais, $Var(U)=E(U^2)-E(U)^2$ avec $E(U^2)=\int_0^\theta\frac{u^2.du}{\theta}=\left[\frac{u^3}{3.\theta}\right]_0^\theta=\frac{\theta^2}{3}$.

Finalement, $Var(U)=\frac{\theta^2}{3}-\frac{\theta^2}{4}=\frac{\theta^2}{12}$ et par suite, $Var(T_1)=\frac{4}{n^2}.\frac{n.\theta^2}{12}=\frac{\theta^2}{3.n}$. T_1 est donc un *estimateur convergent* puisque $\lim_{n\to+\infty}Var(T_1)=0$.

III-2°) La fonction de vraisemblance $L(U_1,U_2,...,U_n,\theta)=\frac{1}{\theta^n}.1_{\underset{1\le i\le n}{MaxU_i}\le\theta}$ est maximale lorsque θ est minimal et atteint donc sa borne inférieure $T_2=\underset{1\le i\le n}{MaxU_i}$. L'estimateur E.M.V ainsi obtenu, soit T_2 ,coïncide donc avec la statistique $V=\underset{1\le i\le n}{MaxU_i}$ étudiée en partie II.

Il s'ensuit notamment $E(T_2)=\frac{n.\theta}{n+1}=\theta-\frac{\theta}{n+1}$ ce qui montre que T_2 constitue un *estimateur biaisé* de θ (le biais étant $B(\theta)=-\frac{\theta}{n+1}$). Par contre, T_2 est *asymptotiquement sans biais* puisque $\lim_{n\to+\infty} E(T_2)=\theta$.

• Par ailleurs, $Var(T_2)=\frac{n.\theta^2}{(n+2).(n+1)^2}$ (selon les résultats de la question II-2°)). T_2 est donc un *estimateur convergent* de θ.

III-3°) La comparaison de l'efficacité de T_1 et de T_2 conduit à former la différence $R(T_2)-R(T_1)$ des risques associés $R(T_1)=E\left[(T_1-\theta)^2\right]=Var(T_1)=\frac{\theta^2}{3.n}$ et $R(T_2)=E\left[(T_2-\theta)^2\right]=E(T_2^2)-2.\theta.E(T_2)+\theta^2$, soit toujours suivant les résultats précédents de la question II-2°), $R(T_2)=\frac{n}{n+2}.\theta^2-2.\theta.\frac{n.\theta}{n+1}+\theta^2=\frac{2.\theta^2}{(n+1).(n+2)}$.

En conclusion, $R(T_2)-R(T_1)=\left[\frac{2}{(n+1).(n+2)}-\frac{1}{3.n}\right].\theta^2=-\frac{(n-1).(n-2)}{3.n.(n+1).(n+2)}.\theta^2$, expression négative dès que $n\succ 2$. Comme $R(T_2)\prec R(T_1)$, T_2 *est plus efficace que* T_1.

IV-1°) T_1 est un *estimateur sans biais* de θ et par ailleurs T_2 est une *statistique exhaustive* pour θ (cf. question II-3° précédente). Dans ces conditions, l'estimateur $T_2^*=E(T_1/T_2)$ est un *estimateur sans biais* « **amélioré** » de T_1 par rapport au paramètre θ (c'est le théorème de RAO-BLACKWELL mentionné en rappels de cours).

Or $E(T_1/T_2)=\frac{2}{n}.E\left[(U_1+U_2+...+U_n)/T_2\right]$. Supposant $T_2=U_n$ (pour simplifier les écritures), la linéarité de l'espérance mathématique permet d'écrire *l'espérance conditionnelle* ci-dessous, sous la forme :

$$E(T_1/T_2)=\frac{2}{n}.E\left[(U_1+U_2+...+U_{n-1})/T_2\right]+\frac{2}{n}.E(U_n/U_n)$$

Mais, $E(U_n/U_n)=U_n$. D'autre part, les *variables conditionnelles* U_i/U_n sont uniformes sur $[0,U_n]$ et ont donc pour espérance $\frac{U_n}{2}$. En bref, l'espérance conditionnelle $E(T_1/T_2)$ est égale à $\frac{2}{n}.\left[U_n+(n-1).\frac{U_n}{2}\right]=\frac{n+1}{n}.U_n$.

• L'estimateur $T_2^*=E(T_1/T_2)=\frac{n+1}{n}.T_2$ est donc *l'estimateur sans biais amélioré* ainsi attendu comme le confirme d'ailleurs le calcul de comparaison d'efficacités ci-après.

IV-2°) En premier lieu, T_2^* est un estimateur sans biais puisque $E(T_2^*)=\frac{n+1}{n}.E(T_2)=\frac{n+1}{n}.\frac{n}{n+1}.\theta=\theta$. D'autre part, la comparaison d'efficacité de T_2^* avec T_1 conduit à étudier le signe de la différence $R(T_2^*)-R(T_1)$ égale en la circonstance à $Var(T_2^*)-Var(T_1)$.

D'une part, $Var(T_2^*)=(\frac{n+1}{n})^2.Var(T_2)=(\frac{n+1}{n})^2.\frac{n.\theta^2}{(n+2).(n+1)^2}=\frac{\theta^2}{n.(n+2)}$ et d'autre part, $Var(T_1)=\frac{\theta^2}{3.n}$. On obtient donc :

$$R(T_2^*)-R(T_1)=\left[\frac{1}{n.(n+2)}-\frac{1}{3.n}\right].\theta^2=-\frac{(n-1)}{3.n.(n+2)}.\theta^2$$

Pour $n\succ 1$, la différence $R(T_2^*)-R(T_1)$ est négative ce qui confirme l'amélioration résultant *du théorème de* RAO- BLACKWELL puisque T_2^* est plus efficace que T_1.

IV- 3°) La comparaison d'efficacité entre l'estimateur sans biais T_2^* et l'estimateur E.M.V (biaisé), soit T_2, conduit à rapprocher les risques $R(T_2)=\frac{2.\theta^2}{(n+1).(n+2)}$ et $R(T_2^*)=\frac{\theta^2}{n.(n+2)}$. Pour changer de la différence $R(T_2)-R(T_2^*)$, on peut aussi passer par le rapport $\frac{R(T_2)}{R(T_2^*)}=\frac{2.n}{n+1}$.

On constate que $n+1\prec 2.n,\forall n,\Rightarrow \frac{R(T_2)}{R(T_2^*)}\succ \frac{2.n}{2.n}=1$. Ainsi T_2^* est-il plus efficace que T_2. En fait, T_2^* constitue *l'estimateur sans biais le plus efficace de* θ, ceci en conformité avec le théorème de LEHMAN- SCHEFFE puisqu'on peut montrer par ailleurs que la statistique $T_2=\underset{1\le i\le n}{MaxU_i}$ est **complète**.

1.4 Modélisation d'une hauteur de crue et prédictions de catastrophe (loi de RAYLEIGH)

Enoncé : La hauteur maximale H de la crue annuelle d'un fleuve est suivie statistiquement, toute crue supérieure ou égale à 6m étant considérée comme catastrophique. On suppose que H est modélisée par une loi de probabilités de type RAYLEIGH, c'est-à-dire pour le cas présent, la loi de densité de probabilité $f(x,\theta)=\frac{x}{\theta}.\exp(-\frac{x^2}{2.\theta})$ pour $x\ge 0$ et θ paramètre inconnu.

L'observation des hauteurs maximales annuelles des crues observées durant dix ans conduit au tableau de valeurs ci-dessous :

2,6	2,4	2,5	1,8	2,9	1,0	2,5	2,6	1,5	2,2

1°) Déterminer l'estimateur du maximum de vraisemblance de θ, soit $\hat{\theta}$ et énumérer ses propriétés.

2°) Déterminer la probabilité qu'une catastrophe se produise durant une année déterminée. En déduire la fréquence d'une catastrophe.

3°) Un assureur table sur une fréquence d'une catastrophe en mille ans. Calculer la probabilité qu'il soit « perdant » de par cette hypothèse.

Solution : 1°) La fonction de vraisemblance $L(X_1,X_2,...,X_n,\theta)$ associée à l'échantillon $(X_1,X_2,...,X_n)$ de taille n s'écrit pour la loi « parente » proposée ici :

$$L(X_1,X_2,...,X_n,\theta)=\prod_{i=1}^{i=n} f(X_i,\theta)=\frac{X_1.X_2...X_n}{\theta^n}.\exp(-\frac{1}{2.\theta}.\sum_{i=1}^{i=n}X_i^2)$$

L'estimateur « E.M.V » cherché (*estimateur du maximum de vraisemblance*), soit $\hat{\theta}$, est solution de l'équation $\frac{\partial}{\partial\theta}L(X,\theta)=0$ ou encore, introduisant la log – vraisemblance $\ln L(X,\theta)$, de l'équation $\frac{\partial}{\partial\theta}\ln L(X,\theta)=0$.

Or, $\ln L(X,\theta)=\sum_{i=1}^{i=n}\ln X_i-n.\ln\theta-\frac{1}{2.\theta}.\sum_{i=1}^{i=n}X_i^2 \Rightarrow \frac{\partial}{\partial\theta}\ln L(X,\theta)=-\frac{n}{\theta}+\frac{1}{2.\theta^2}.\sum_{i=1}^{i=n}X_i^2$.

Il résulte de $\frac{\partial}{\partial\theta}\ln L(X,\theta)=0$, la solution $\hat{\theta}=\frac{\sum_{i=1}^{i=n}X_i^2}{2.n}$. On remarquera qu'on a bien un maximum puisqu'en $\hat{\theta}$, $\frac{\partial^2}{\partial\theta^2}\ln L(X,\theta)=\frac{1}{\hat{\theta}^2}.\left[n-\frac{\sum_{i=1}^{i=n}X_i^2}{\hat{\theta}}\right]=-\frac{n}{\hat{\theta}}\prec 0$.

Numériquement, la valeur obtenue ici pour $\hat{\theta}$ est immédiatement égale à :

$$\frac{1}{2\times10}.\left[2,6^2+...+2,2^2\right]=2,576$$

• Le calcul de $E(\hat{\theta})$ conduit au développement $E(\hat{\theta})=\frac{1}{2.n}.\sum_{i=1}^{i=n}E(X_i^2)$ avec, pour ce qui est de la variable « parente », $E(X^2)=\int_0^{+\infty}\frac{x^3}{\theta}.\exp(-\frac{x^2}{2.\theta}).dx$. Suivant intégration par parties $(U=x^2,dV=\frac{x}{\theta}.\exp(-\frac{x^2}{2.\theta})\Rightarrow dU=2x.dx,V=-\exp(-\frac{x^2}{2.\theta})$, on obtient :

$$E(X^2)=\left[-x^2.\exp(-\frac{x^2}{2.\theta})\right]_0^{+\infty}+2.\int_0^{+\infty}x.\exp(-\frac{x^2}{2.\theta}).dx=2.\left[-\theta.\exp(-\frac{x^2}{2.\theta}\right]_0^{+\infty}=2.\theta$$

$\hat{\theta}$ forme donc un *estimateur sans biais* de θ, puisque $E(\hat{\theta})=\frac{1}{2.n}.\sum_{i=1}^{i=n}2.\theta=\theta$.

• La loi de RAYLEIGH proposée ici appartient qui plus est à la **famille exponentielle**. En effet, $\ln f(x,\theta)=\ln x-\ln\theta-\frac{x^2}{2.\theta}$, ce qui est bien de la forme caractérisant ladite famille soit $Q(\theta).T(x)+\beta(\theta)+\gamma(x)$ avec notamment, $T(x)=x^2$. La statistique $\sum_{i=1}^{i=n}X_i^2$ est **exhaustive** pour θ et $\hat{\theta}$ qui est sans biais. Elle forme donc un **estimateur efficace** (cf. paragraphe 2.5 des rappels de cours). En outre, le *théorème central limite* assure la convergence de $\hat{\theta}$ vers la *loi normale*, lorsque n est grand.

2°) La probabilité qu'une catastrophe se produise est égale à :

$$\Pr ob(H \geq 6) = \int_6^{+\infty} \frac{x}{2,576}.\exp(-\frac{x^2}{2\times 2,576}).dx = \exp(-\frac{36}{2\times 2,576}) = 9,24\times 10^{-4}$$

La fréquence d'une catastrophe qui est égale à l'inverse de la probabilité susmentionnée est donc égale à 1083 ans.

3°) L'assureur est perdant quand plus d'une catastrophe est constatée en mille ans. Notant par N le nombre de catastrophes constatées durant mille ans et par p la probabilité d'occurrence annuelle calculée dans la question précédente ($p = 9,24\times 10^{-4}$), N suit la *loi binomiale* $B(1000, p)$, loi dont la convergence vers la loi de POISSON est évidente (p faible, n grand).

Dans ces conditions, on obtient $\Pr ob(N \succ 1) = 1 - \Pr ob(N = 0) - \Pr ob(N = 1)$ avec $\Pr ob(N = n) = \frac{e^{-a}.a^n}{n!}$ et $a = n.p = 0,924$. On trouve immédiatement $\Pr ob(N \succ 1) = 0,23$ (ce qui est un risque non moindre !).

1.5 Modélisation de la durée de vie de diodes (loi de WEIBULL)

Enoncé : On souhaite modéliser la durée de vie de diodes par une variable aléatoire Y de loi de WEIBULL de paramètres $(\lambda, 2)$, c'est-à-dire de fonction de répartition :

$$F_{(\lambda,2)}(t) = 0 \text{ si } t \leq 0 \text{ et } F_{(\lambda,2)}(t) = 1 - e^{-\lambda.t^2} \text{ si } t \succ 0$$

1°) Expliciter les densités de probabilité $f_{(\lambda,2)}$ et $g_{(\lambda,2)}$ des variables aléatoires Y et Y^2. Qu'en conclure pour ce qui est de la loi de Y^2 ?

2°) Exprimer la fonction de vraisemblance $L(Y_1, Y_2, ..., Y_n, \lambda)$ d'un échantillon de taille n, soit $(Y_1, Y_2, ..., Y_n)$, de la loi « parente » Y.

3°) Montrer qu'il existe un estimateur T du maximum de vraisemblance pour le paramètre λ.

4°) Calculer $E(T)$. Qu'en conclure ?

5°) Calculer le risque quadratique de T, c'est-à-dire l'espérance mathématique $E\left[(T-\lambda)^2\right]$.

Solution : 1°) $f_{(\lambda,2)}(t) = \frac{dF_{(\lambda,2)}(t)}{dt} = 2.\lambda.t.e^{-\lambda.t^2}$ (pour $t \succ 0$).

D'autre part, posant $Z = Y^2$, soit $Y = \sqrt{Z}$ (puisque $Y \succ 0$), et notant par $g_{(\lambda,2)}(z)$ la densité de probabilité de Z, le *théorème de la mesure image* entraîne $g_{(\lambda,2)}(z).dz = 2.\lambda.\sqrt{z}.e^{-\lambda.z}.\frac{dz}{2.\sqrt{z}} = \lambda.e^{-\lambda.z}.dz$.

Ainsi Y^2 dont la densité de probabilité $g_{(\lambda,2)}(z)$ est égale à $\lambda.e^{-\lambda.z}$ suit-elle la *loi exponentielle* de paramètre λ.

2°) La *fonction de vraisemblance* associée à l'échantillon $(Y_1,Y_2,...,Y_n)$ est égale au produit des densités de probabilité, soit $L(Y_1,Y_2,...,Y_n,\lambda)=2^n.\lambda^n.\prod_{i=1}^{i=n}Y_i.e^{-\lambda.Y_i^2}$. La fonction de *log -vraisemblance* est égale quant à elle, à $\ln L = n.\ln 2 + n.\ln\lambda + \sum_{i=1}^{i=n}\ln Y_i - \lambda.\sum_{i=1}^{i=n}Y_i^2$.

3°) L'estimateur E.M.V de λ, soit T, est solution de l'équation en λ, $\frac{\partial}{\partial\lambda}\ln L = 0 \Rightarrow \frac{n}{\lambda} - \sum_{i=1}^{i=n}Y_i^2 = 0$. On obtient donc pour T, l'expression $T=\frac{n}{\sum_{i=1}^{i=n}Y_i^2}$. Il s'agit bien d'un maximum puisque $\frac{\partial^2}{\partial\lambda^2}\ln L \prec 0$.

4°) On remarque préalablement que $\sum_{i=1}^{i=n}Y_i^2$, qui est une somme de n variables aléatoires exponentielles indépendantes, suit la loi *Gamma n* dont il est rappelé que la densité de probabilité est égale à $\frac{\lambda^n.e^{-\lambda.t}.t^{n-1}}{(n-1)!}.1_{t\geq 0}$ (cf. rappels de cours du chapitre I).

Dans ces conditions, $E(T)=E\left[\frac{n}{S}\right]$ avec $S=\sum_{i=1}^{i=n}Y_i^2$, a donc pour expression :

$$E(T)=\int_0^{+\infty}\frac{n}{s}.\frac{\lambda^n.e^{-\lambda.s}.s^{n-1}.ds}{(n-1)!}=\frac{n.\lambda^n}{(n-1)!}.\int_0^{+\infty}s^{n-2}.e^{-\lambda.s}.ds$$

Un calcul par récurrence, montre que l'intégrale $\int_0^{+\infty}s^{n-2}.e^{-\lambda.s}.ds$ (dite d'EULER) est égale à $\left[-s^{n-2}.\frac{e^{-\lambda.s}}{\lambda}\right]_0^{+\infty}+\frac{(n-2)}{\lambda}.\int_0^{+\infty}s^{n-3}.e^{-\lambda.s}.ds=\frac{(n-2)}{\lambda}.\int_0^{+\infty}s^{n-3}.e^{-\lambda.s}.ds$. De proche en proche, on a ainsi $\int_0^{+\infty}s^{n-2}.e^{-\lambda.s}.ds=\frac{(n-2)!}{\lambda^{n-2}}.I_0$ avec $I_0=\int_0^{+\infty}e^{-\lambda.s}.ds=\frac{1}{\lambda}$.

En conclusion, $E(T)=\frac{n.\lambda^n}{(n-1)!}.\frac{(n-2)!}{\lambda^{n-1}}=\frac{n}{n-1}.\lambda$.

• *T n'est donc pas un estimateur sans biais.* Par contre, *T est asymptotiquement sans biais* puisque $\lim_{n\to+\infty}E(T)=\lambda$. On notera cependant que la *correction* $T^*=\frac{n-1}{n}.T$ permettrait ici d'obtenir *un estimateur sans biais* de λ.

5°) Le risque quadratique de T est égal à $\|T-\lambda\|^2=E\left[(T-\lambda)^2\right]$ soit en développant suivant la linéarité de l'espérance mathématique $\|T-\lambda\|^2=E(T^2)-2.\lambda.E(T)+\lambda^2$.

Reprenant le calcul antérieur, $E(T^2)=\int_0^{+\infty}\frac{n^2}{s^2}.\frac{\lambda^n.e^{-\lambda.s}.s^{n-1}}{(n-1)!}.ds=\frac{n^2.\lambda^n}{(n-1)!}.\int_0^{+\infty}s^{n-3}.e^{-\lambda.s}.ds$, soit $E(T^2)=\frac{n^2.\lambda^n}{(n-1)!}.\frac{(n-3)!}{\lambda^{n-2}}=\frac{n^2.\lambda^2}{(n-1).(n-2)}$.

Revenant à $E\left[(T-\lambda)^2\right]$, on obtient donc l'expression :

$$E\left[(T-\lambda)^2\right]=\lambda^2.\left[\frac{n^2}{(n-1).(n-2)}-\frac{2.n}{n-1}+1\right]=\frac{(n+2)}{(n-1).(n-2)}.\lambda^2$$

T est *convergent en probabilité* puisque $\lim_{n\to+\infty} E\left[(T-\lambda)^2\right]=0$.

• Il est intéressant de constater que T *est asymptotiquement efficace*. En effet, pour n grand, $\frac{(n+2)}{(n-1).(n-2)}.\lambda^2\approx\frac{\lambda^2}{n}$. Or $\frac{\lambda^2}{n}$ est bien la borne F.D.C.R égale à l'inverse de la quantité d'information de FISHER, soit $I_{(Y_1,Y_2,\ldots,Y_n,\lambda)}(\lambda)=-E\left[\frac{\partial^2}{\partial\lambda^2}\ln L\right]=\frac{n}{\lambda^2}$.

1.6 Modèle de PARETO

Enoncé : La loi de PARETO comme la loi log- normale et les lois en puissance permet de modéliser convenablement des phénomènes comme l'occurrence des mots les plus usités dans une langue donnée, le nombre de connexions sur le réseau internet, le montant des revenus, le nombre d'habitants dans une ville.... De façon générale, cette loi est caractérisée par la densité de probabilité :

$$f_{(\alpha,\beta)}(x)=\alpha.\frac{\beta^\alpha}{x^{\alpha+1}}.1_{x\geq\beta}$$

(α et β étant des paramètres réels strictement positifs).

On s'intéresse dans ce problème à la loi dite « réduite » qui correspond au cas $\beta=1$, loi à un seul paramètre α, caractérisée par la densité de probabilité :

$$f_\alpha(x)=\frac{\alpha}{x^{\alpha+1}}.1_{x\geq1},(\alpha\succ0)$$

1°) Déterminer l'estimateur du maximum de vraisemblance de α, soit $\hat{\alpha}$.

2°) Montrer que si on effectue le changement de variable $U=\ln X$, U suit la loi exponentielle. En déduire une méthode de simulation d'un échantillon de taille n dont la loi « parente » est de type PARETO.

3°) Calculer $E(\hat{\alpha})$. Qu'en conclure ?.

4°) Montrer cependant que $\hat{\alpha}$ est convergent et asymptotiquement efficace.

5°) Utilisant la convergence de $\hat{\alpha}$ vers la loi normale, en déduire une estimation par intervalle de confiance de α.

6°) Montrer que $\frac{\alpha}{\hat{\alpha}}$ suit la loi Gamma n. En déduire une estimation par intervalle de confiance de α suivant une méthode exacte et non plus asymptotique comme à la question précédente.

Solution : 1°) L'estimateur E.M.V cherché de α, soit $\hat{\alpha}$, est la solution de l'équation $\frac{\partial}{\partial\alpha}\ln L=0$, soit $\frac{n}{\alpha}-\sum_{i=1}^{i=n}\ln(X_i)=0$. On a donc $\hat{\alpha}=\frac{n}{\sum_{i=1}^{i=n}\ln(X_i)}$.

A noter que la solution précédente $\hat{\alpha}$ est bien un maximum puisque $\frac{\partial^2}{\partial \alpha^2} \ln L = -\frac{n}{\alpha^2} \prec 0$.

2°) Le changement de variable $U = \ln X$ dans la densité de probabilité de X, soit $\frac{\alpha}{x^{\alpha+1}}.1_{x\geq 1}$, conduit à la densité de probabilité $g(u)$ vérifiant $g(u).du = \frac{\alpha}{e^{(\alpha+1).u}}.e^u.du$. Ainsi a-t-on $g(u) = \alpha.e^{-\alpha.u}$ qui est la densité de probabilité de la *loi exponentielle* de paramètre α.

• La *méthode de la transformation inverse* permet de simuler la loi exponentielle à partir de nombres aléatoires uniformément distribués sur $[0,1]$. Pour rappel, la loi exponentielle de paramètre α et de valeurs u_i et la loi uniforme sur $[0,1]$ de valeurs z_i sont liées par la relation $z_i = F(u_i)$ avec $F(u_i) = \Pr ob(U \leq u_i) = \int_0^{u_i} \alpha.e^{-\alpha.t}.dt = 1 - e^{-\alpha.u_i}$.

Bref, $u_i = -\frac{1}{\alpha}.\ln(1-z_i)$. La procédure de simulation proposée pour générer un échantillon $(X_1, X_2, ..., X_n)$ de n variables de PARETO à partir d'un échantillon $(Z_1, Z_2, ..., Z_n)$ de n variables aléatoires uniformes sur $[0,1]$, est donc simple comme indiqué ci-dessous :

Etape 1→ Générer un nombre aléatoire z_i uniforme sur $[0,1]$ à l'aide de l'ordinateur ou d'une calculatrice (fonctions RND, ALEA...).

Etape 2→ Former $u_i = -\frac{1}{\alpha}.\ln(1-z_i)$.

Etape 3→ Former $x_i = \exp(u_i)$. La valeur ainsi déterminée constitue une valeur simulée de la loi de PARETO réduite de paramètre α.

3°) $E(\hat{\alpha}) = E\left[\frac{n}{\sum_{i=1}^{i=n} \ln(X_i)}\right]$. Mais, suivant la question précédente, la somme $\sum_{i=1}^{i=n} \ln(X_i)$ suit la loi Gamma n de paramètre α puisque somme des n variables aléatoires exponentielles indépendantes $\ln(X_i)$.

On peut donc écrire $E(\hat{\alpha}) = \int_0^{+\infty} \frac{n}{u}.\frac{\alpha^n.e^{-\alpha.u}.u^{n-1}}{(n-1)!}.du$ (en utilisant l'expression de la densité de probabilité de la loi Gamma n rappelée au chapitre I, soit $\frac{\alpha^n.e^{-\alpha.u}.u^{n-1}}{(n-1)!}.1_{u\geq 0}$.

Finalement, $E(\hat{\alpha}) = \frac{n.\alpha^n}{(n-1)!}.\int_0^{+\infty} e^{-\alpha.u}.u^{n-2}.du$.

Introduisant la densité de probabilité de la loi Gamma $n-1$, soit $\frac{\alpha^{n-1}.e^{-\alpha.u}.u^{n-2}}{(n-2)!}.1_{u\geq 0}$ et écrivant que l'intégrale de cette densité de probabilité sur $[0,+\infty]$ est égale à l'unité, il vient immédiatement $\int_0^{+\infty} e^{-\alpha.u}.u^{n-2}.du = \frac{(n-2)!}{\alpha^{n-1}}$.

En définitive, $E(\hat{\alpha}) = \frac{n.\alpha^n}{(n-1)!}.\frac{(n-2)!}{\alpha^{n-1}} = \frac{n}{n-1}.\alpha$. L'estimateur $\hat{\alpha}$ *n'est pas sans biais* (il surestime la valeur de α). Cependant, $\hat{\alpha}$ est *asymptotiquement sans biais* puisque $\lim_{n\to+\infty} E(\hat{\alpha}) = \alpha$.

4°) Le risque $R(\alpha) = \left\|\hat{\alpha} - \alpha\right\|^2 = E\left[(\hat{\alpha} - \alpha)^2\right]$ est égal à $E(\hat{\alpha}^2) - 2.\alpha.E(\hat{\alpha}) + \alpha^2$. Reprenant les calculs antérieurs, on obtient $E(\hat{\alpha}^2) = n^2.\int_0^{+\infty} \frac{1}{u^2}.\frac{\alpha^n.e^{-\alpha.u}.u^{n-1}}{(n-1)!}.du = \frac{n^2.\alpha^n.(n-3)!}{(n-1)!.\alpha^{n-2}}$, soit, $E(\hat{\alpha}^2) = \frac{\alpha^2.n^2}{(n-1).(n-2)}$. En conclusion :

$$R(\alpha) = \left[\frac{n^2}{(n-1).(n-2)} - \frac{2.n}{n-1} + 1\right].\alpha^2 = \frac{(n+2).\alpha^2}{(n-1).(n-2)}.$$

$\hat{\alpha}$ est un *estimateur convergent* puisque $\lim_{n\to+\infty} R(\alpha) = 0$.

- D'autre part, *la quantité d'information de* FISHER fournie par l'échantillon $(X_1, X_2, ..., X_n)$ est égale à $I_{(X_1,X_2,...,X_n)}(\alpha) = -E\left[\frac{\partial^2}{\partial\alpha^2}\ln L\right] = -E\left[-\frac{n}{\alpha^2}\right] = \frac{n}{\alpha^2}$. Or, pour n grand, $R(\alpha) \approx \frac{\alpha^2}{n}$, c'est-à-dire la borne minimale F.D.C.R égale à l'inverse de $I_{(X_1,X_2,...,X_n)}(\alpha)$. L'estimateur $\hat{\alpha}$ est donc *asymptotiquement efficace.*

5°) La convergence de $\hat{\alpha}$ induite par *le théorème central limite*, entraîne pour la variable aléatoire $\hat{\alpha} - \alpha$ la convergence vers la *loi normale* de moyenne 0 (puisque $\hat{\alpha}$ est asymptotiquement sans biais) et de variance $\frac{\alpha^2}{n}$.

On en déduit donc, au *seuil de confiance* $1-\beta$ donné (par exemple, $1-\beta = 95\%$), l'encadrement $-t_\beta.\frac{\alpha}{\sqrt{n}} \le \hat{\alpha} - \alpha \le +t_\beta.\frac{\alpha}{\sqrt{n}}$ où t_β vérifie la relation $\Pr ob(|\xi| \le t_\beta) = 1-\beta$. (Par exemple, dans le cas $1-\beta = 95\%$, on a $t_\beta = 1{,}96$). En conclusion, on obtient pour α, l'encadrement $\hat{\alpha} - t_\beta.\frac{\alpha}{\sqrt{n}} \le \alpha \le \hat{\alpha} + t_\beta.\frac{\alpha}{\sqrt{n}}$.

- α étant inconnu, on pourra approximer sa valeur par $\hat{\alpha}$ dans les bornes de l'encadrement susmentionné, d'où *l'intervalle de confiance* :

$$\hat{\alpha}.\left[1 - \frac{t_\beta}{\sqrt{n}}\right] \le \alpha \le \hat{\alpha}.\left[1 + \frac{t_\beta}{\sqrt{n}}\right]$$

6°) $V = \sum_{i=1}^{i=n} \ln(X_i) = \frac{n}{\hat{\alpha}}$ suit la *loi Gamma n de paramètre* α, c'est-à-dire de densité de probabilité $\frac{\alpha^n.e^{-\alpha.v}.v^{n-1}}{(n-1)!}.1_{v\ge 0}$. On se propose de déterminer la loi de $Z = \frac{\alpha.V}{n} = \frac{\alpha}{\hat{\alpha}}$.

Le théorème de la mesure image conduit immédiatement après changement de variable $z=\frac{\alpha.v}{n}$ dans la densité élémentaire $\frac{\alpha^n.e^{-\alpha.v}.v^{n-1}}{(n-1)!}.1_{v\geq 0}.dv$, à la nouvelle densité élémentaire $\frac{\alpha^n.e^{-n.z}.n^{n-1}.z^{n-1}.n.dz}{(n-1)!.\alpha^n}$, d'où en définitive, la loi de $Z=\frac{\alpha}{\widehat{\alpha}}$ définie par la densité $\frac{n^n.e^{-n.z}.z^{n-1}}{(n-1)!}.1_{z\geq 0}$.

• Il s'agit en fait de *la loi Gamma n de paramètre n*. Dès lors, en considérant les *quantiles* d'ordre $\frac{\beta}{2}$, soient t_1 et t_2 solutions respectives des équations $\Pr ob(Z\leq t_1)=\frac{\beta}{2}$ et $\Pr ob(Z\geq t_2)=\frac{\beta}{2}$, on en déduit l'encadrement $t_1\leq\frac{\alpha}{\widehat{\alpha}}\leq t_2$, soit *l'estimation par intervalle de confiance au seuil* $1-\beta$ cherchée, relativement à α :

$$\widehat{\alpha}.t_1\leq\alpha\leq\widehat{\alpha}.t_2$$

La mise en œuvre de cette méthode exige cependant de disposer d'un bon calculateur.

1.7 Modèle exponentiel translaté

Enoncé : On considère la variable aléatoire X de loi exponentielle de paramètre θ translatée de α, c'est-à-dire de densité de probabilité :

$$f(x)=\frac{1}{\theta}.\exp\left[-(\frac{x-\alpha}{\theta})\right].1_{[\alpha,+\infty[}(x)$$

1°) Déterminer les estimateurs du maximum de vraisemblance de α et de θ, soient $\widehat{\alpha}$ et $\hat{\theta}$.

2°) Exprimer $\Pr ob(n.(\widehat{\alpha}-\alpha)\geq t)$ et en déduire la loi suivie par la variable aléatoire $n.(\widehat{\alpha}-\alpha)$.

3°) Montrer que $\sqrt{n}.(\hat{\theta}-\theta)$ se décompose suivant la différence $Z-\frac{W}{\sqrt{n}}$, où Z converge vers une loi normale et où W suit la loi exponentielle de paramètre $\frac{1}{\theta}$, et en conclure à la convergence de $\sqrt{n}.(\hat{\theta}-\theta)$ vers une loi normale que l'on précisera.

Solution : 1°) La fonction de vraisemblance $L(X_1,X_2,...,X_n,\theta,\alpha)$ de l'échantillon $(X_1,X_2,...,X_n)$ de loi «parente » X est égale à :

$$L(X_1,X_2,...,X_n,\theta,\alpha)=\prod_{i=1}^{i=n}\frac{1}{\theta}.\exp\left[-(\frac{X_i-\alpha}{\theta})\right] \text{ si } \inf_{1\leq i\leq n} X_i\geq\alpha$$

Développant $L(X_1,X_2,...,X_n,\theta,\alpha)$, on obtient l'expression :

$$L(X_1,X_2,...,X_n,\theta,\alpha)=\frac{1}{\theta^n}.\exp(\frac{n.\alpha}{\theta}).\exp(-\frac{\sum_{i=1}^{i=n}X_i}{\theta}) \text{ pour } \inf_{1\leq i\leq n} X_i\geq\alpha .$$

• Par rapport à α, $L(X_1,X_2,...,X_n,\theta,\alpha)$ qui est croissante avec α est donc maximale pour α maximal, soit la valeur $\hat{\alpha} = \inf_{1\le i\le n} X_i$. On a le cas ici d'un maximum de vraisemblance dont l'expression n'est pas fournie algébriquement par l'équation $\frac{\partial}{\partial\alpha} L = 0$.

• D'autre part, toujours à l'optimum, et cette fois relativement à θ, on a en considérant la fonction de log- vraisemblance définie par $\ln L = -n.\ln\theta + \frac{n.\hat{\alpha}}{\theta} - \frac{1}{\theta}.\sum_{i=1}^{i=n} X_i$, l'estimateur E.M.V cherché $\hat{\theta}$, solution de l'équation $\frac{\partial}{\partial\theta}\ln L = 0$, ce qui conduit à :

$$-\frac{n}{\theta} - \frac{n.\hat{\alpha}}{\theta^2} + \frac{1}{\theta^2}.\sum_{i=1}^{i=n} X_i = 0 \text{, d'où } \hat{\theta} = \frac{\sum_{i=1}^{i=n} X_i}{n} - \hat{\alpha} = \frac{\sum_{i=1}^{i=n} X_i}{n} - \underset{1\le i\le n}{MinX_i}.$$

2°) $\Pr ob(n.(\hat{\alpha}-\alpha)\ge t) = \Pr ob(\hat{\alpha} = \underset{1\le i\le n}{MinX_i} \ge \alpha + \frac{t}{n})$. Mais, cette dernière probabilité admet la décomposition $\Pr ob(\underset{1\le i\le n}{Min}\, X_i \ge \alpha + \frac{t}{n}) = \prod_{i=1}^{i=n} \Pr ob(X_i \ge \alpha + \frac{t}{n})$ puisque pour tout c, la condition $\underset{1\le i\le n}{Min}\, X_i \ge c \Leftrightarrow X_1 \ge c, X_2 \ge c, ..., X_n \ge c$.

Explicitant chacune de ces probabilités, il vient :

$$\Pr ob(X_i \ge \alpha + \frac{t}{n}) = \int_{\alpha+\frac{t}{n}}^{+\infty} \frac{1}{\theta}.\exp(-(\frac{x-\alpha}{\theta})).dx \text{ (en effet, pour } t\succ 0, \alpha+\frac{t}{n}\ge\alpha).$$

En résumé, $\Pr ob(X_i \ge \alpha + \frac{t}{n}) = -\left[\exp(-(\frac{x-\alpha}{\theta})\right]_{\alpha+\frac{t}{n}}^{+\infty} = \exp(-\frac{t}{n.\theta})$ et par suite, on obtient $\Pr ob(\underset{1\le i\le n}{Min}\, X_i \ge \alpha + \frac{t}{n}) = \left[\exp(-\frac{t}{n.\theta}\right]^n = \exp(-\frac{t}{\theta})$.

• Revenant à la loi de la variable aléatoire $n.(\hat{\alpha}-\alpha)$ dont la densité de probabilité $\varphi(t)$ vérifie $\varphi(t) = -\frac{dV(t)}{dt}$, équation dans laquelle $V(t)$ désigne la fonction de fiabilité exprimée par $V(t) = \Pr ob(n.(\hat{\alpha}-\alpha)\ge t) = \exp(-\frac{t}{\theta})$, il s'agit donc de *la loi exponentielle de paramètre* $\frac{1}{\theta}$ puisque $\varphi(t) = -\frac{d}{dt}\left[\exp(-\frac{t}{\theta})\right] = \frac{1}{\theta}.\exp(-\frac{t}{\theta})$.

• Des résultats ci-dessus, découle immédiatement une *procédure de construction d'un intervalle de confiance* pour α et ceci au seuil $1-\beta$. En effet, désignant t_1 et t_2 les nombres vérifiant pour la variable W de loi exponentielle de paramètre $\frac{1}{\theta}$, les relations $\Pr ob(W\le t_1) = \Pr ob(W\ge t_2) = \frac{\beta}{2}$ (on dit aussi les *quantiles d'ordre* $\frac{\beta}{2}$ de W), on obtient immédiatement, sachant que $W = n.(\hat{\alpha}-\alpha)$, l'intervalle $\hat{\alpha} - \frac{t_2}{n} \le \alpha \le \hat{\alpha} - \frac{t_1}{n}$.

3°) La normalité asymptotique de $\hat{\theta}$ suivant les résultats des rappels de cours portant sur *la méthode du maximum de vraisemblance* (cf. paragraphe 3.3) *n'est pas applicable ici* car on est dans un *cas multidimensionnel* de deux paramètres θ et α dont au demeurant, les estimateurs $\hat{\theta}$ et $\hat{\alpha}$ ne sont pas indépendants.

La décomposition proposée va lever cette difficulté. En effet :

$$\sqrt{n}.(\hat{\theta}-\theta)=\sqrt{n}.(\overline{X}-\hat{\alpha}-\theta)=\sqrt{n}.(\overline{X}-(\theta+\alpha)-(\hat{\alpha}-\alpha)) \text{ avec } \overline{X}=\frac{\sum_{i=1}^{i=n}X_i}{n}.$$

Or, suivant le *théorème central limite*, $\overline{X}=\frac{\sum_{i=1}^{i=n}X_i}{n}$ converge, pour n grand, vers la loi normale de moyenne $E(X_i)$ et de variance $\frac{Var(X_i)}{n}$. Le calcul des ces paramètres conduit aux développements suivants :

• $E(X_i)=\int_{\alpha}^{+\infty}\frac{x}{\theta}.\exp\left[-(\frac{x-\alpha}{\theta})\right].dx=\left[-x.\exp\left[-(\frac{x-\alpha}{\theta})\right]\right]_{\alpha}^{+\infty}+\int_{\alpha}^{+\infty}\exp\left[-(\frac{x-\alpha}{\theta})\right].dx=\alpha+\theta$

• $Var(X_i)=E(X_i^2)-E(X_i)^2$ avec $E(X_i^2)=\int_{\alpha}^{+\infty}\frac{x^2}{\theta}.\exp\left[-(\frac{x-\alpha}{\theta})\right].dx$. Intégrant cela par parties, en posant $U=x^2, dV=\frac{1}{\theta}.\exp\left[-(\frac{x-\alpha}{\theta})\right]$, on obtient immédiatement l'expression $E(X_i^2)=\left[-x^2.\exp\left[-(\frac{x-\alpha}{\theta})\right]\right]_{\alpha}^{+\infty}+\int_{\alpha}^{+\infty}2.x.\exp\left[-(\frac{x-\alpha}{\theta})\right].dx$, soit après calculs, $E(X_i^2)=\alpha^2+2.\theta.(\alpha+\theta)$. Ainsi $Var(X_i)=\alpha^2+2.\theta.(\alpha+\theta)-(\alpha+\theta)^2=\theta^2$.

$\overline{X}$ converge donc vers la loi normale de moyenne $\theta+\alpha$ et de variance $\frac{\theta^2}{n}$. En d'autres termes, $Z=\sqrt{n}.(\overline{X}-(\alpha+\theta))$ converge vers la loi normale centrée et de variance θ^2 puisqu'on a immédiatement, de par les propriétés habituelles de l'espérance mathématique et de la variance d'une fonction affine, $E(Z)=\sqrt{n}.\left[E(\overline{X})-(\alpha+\theta)\right]=0$ et $Var(Z)=n.Var(\overline{X})=\theta^2$.

D'autre part, on a montré dans la 2ème question que la variable aléatoire $n.(\hat{\alpha}-\alpha)$ suivait la loi exponentielle de paramètre $\frac{1}{\theta}$. On a donc bien, pour $\sqrt{n}.(\hat{\theta}-\theta)$ une décomposition sous la forme $\sqrt{n}.(\hat{\theta}-\theta)=Z-\sqrt{n}.(\hat{\alpha}-\alpha)=Z-\frac{n}{\sqrt{n}}.(\hat{\alpha}-\alpha)=Z-\frac{W}{\sqrt{n}}$ dans laquelle $Z=\sqrt{n}.(\overline{X}-(\alpha+\theta))$ suit asymptotiquement la loi $N(0,\theta^2)$ et $W=n.(\hat{\alpha}-\alpha)$ est exponentielle de paramètre $\frac{1}{\theta}$ (ce qui démontre le résultat proposé).

• Lorsque $n\to+\infty$, la loi de $\sqrt{n}.(\hat{\theta}-\theta)$ converge vers la loi de Z, c'est-à-dire à la limite, la loi normale $N(0,\theta^2)$, du moins en admettant certaines propriétés inhérentes aux modes de convergences dont notamment la *convergence en loi* et la *convergence en probabilité*.

On trouve donc ici encore une *méthode d'estimation par intervalle de confiance* (cette fois pour θ), puisque si t_β désigne le nombre tel que $\Pr ob(|\xi| \le t_\beta) = 1-\beta$ (ξ désignant la variable aléatoire normale centrée réduite $N(0,1)$), on a l'encadrement :

$$\hat{\theta} - t_\beta . \frac{\theta}{\sqrt{n}} \le \theta \le \hat{\theta} + t_\beta . \frac{\theta}{\sqrt{n}}$$

qu'on pourra approximer par $\hat{\theta}.(1-\frac{t_\beta}{\sqrt{n}}) \le \theta \le \hat{\theta}.(1+\frac{t_\beta}{\sqrt{n}})$.

• Revenant à l'aspect théorique susmentionné pour justifier la convergence de $\sqrt{n}.(\hat{\theta}-\theta)$, on distingue entre autres, la *convergence en loi* ($X_n \xrightarrow{L} X$ si $\lim_{n\to+\infty} F_n(x) = F(x)$ avec $F(x) = \Pr ob(X \le x)$) et la *convergence en probabilité* ($X_n \xrightarrow{P} X$ si $\lim_{n\to+\infty} \Pr ob(|X_n - X| \ge \varepsilon) = 0$), cette dernière convergence entraînant la convergence en loi. On montre alors que si $X_n \xrightarrow{L} X$ et $Y_n \xrightarrow{P} a$, (a constant), la somme $X_n + Y_n$ *converge en loi* vers la variable aléatoire $X + a$.

Ceci justifie la convergence susmentionnée de $\sqrt{n}.(\hat{\theta}-\theta)$, la constante a étant nulle en la circonstance puisque $\sqrt{n}.(\hat{\alpha}-\alpha) = \frac{1}{\sqrt{n}}.n.(\hat{\alpha}-\alpha) \xrightarrow{P} 0$.

2. Techniques d'estimation et modèles divers

2.1 Un modèle « mélangé » POISSON/Gamma en assurance automobile

Enoncé : Après avoir mis en évidence en partie I, les caractéristiques du processus de POISSON dit « mélangé », il est proposé en partie II de comparer trois méthodes d'estimation des paramètres dudit processus avec une illustration à partir de données portant sur le risque automobile.

PARTIE I

Pour rappel, le processus de POISSON décrit l'occurrence des évènements rares et est caractérisé par l'équation $\Pr ob(N(t) = n) = p_n(t) = \frac{e^{-\lambda . t}(\lambda . t)^n}{n!}$. Ce processus décrit très valablement le nombre de sinistres au cours du temps, par assuré, lorsque ces derniers sont égaux devant le risque et forment un groupe homogène. Pour simplifier les choses, on supposera raisonner dans cette application sur des périodes de temps unitaires, ce qui entraîne $t = 1$.

En pratique, l'homogénéité des assurés au regard du risque n'est que rarement vérifiée (du fait de raisons diverses comme, par exemple, la profession, l'âge, l'état de santé, le comportement au volant, les parcours empruntés...). On prend en compte cette situation en mélangeant le modèle de POISSON par une loi qui décrit ces particularités entre individus.

On considère ainsi λ comme une fonction de chaque assuré, c'est-à-dire en définitive, comme une variable aléatoire Λ (fonction dite « de structure ») dont on note par $u(\lambda)$ la densité de probabilité, la loi conditionnelle de $N(t)$ sachant « $\Lambda = \lambda$ » étant quant à elle, la loi de POISSON susmentionnée de paramètre λ (puisque $t = 1$).

I-1°) Ecrire l'expression de la probabilité $\Pr ob(N=n)$ en fonction de $u(\lambda)$ et notamment dans le cas où Λ suit la loi Gamma « généralisée » de paramètres α et β, c'est-à-dire la loi de densité de probabilité $u(\lambda)=\dfrac{\beta^{\alpha}.e^{-\beta.\lambda}.\lambda^{\alpha-1}}{\Gamma(\alpha)},(\alpha\succ 0,\beta\succ 0)$.

I-2°) Exprimer la fonction génératrice des moments de N.

I-3°) Soit X une variable aléatoire de loi binomiale négative de paramètres r et p. Déterminer la fonction génératrice des moments de X puis celle de la variable $Y=X-r$ dont on exprimera également la loi de probabilité.

I-4°) Comparant les fonctions génératrices de N et de Y, en déduire la loi de probabilité de N.

I-5°) Calculer $E(N)$ et $Var(N)$.

PARTIE II

Partant de la loi binomiale négative précédente de paramètres α et p, on se propose d'estimer ici ces deux paramètres suivant trois méthodes distinctes. On utilisera à cet effet, les notations suivantes :

* n, nombre total des assurés ;
* k, nombre total de sinistres par année et par assuré ;
* m, maximum des valeurs k observées ;
* k_i, nombre de sinistres par année pour un assuré donné i ;
* n_k, nombre d'assurés responsables de k sinistres durant l'année ;
* f_k, fréquence relative des assurés responsables de k sinistres durant l'année ;
* $\bar{n}$, moyenne empirique du nombre de sinistres par année et par assuré ;
* s^2, variance empirique du nombre de sinistres par année et par assuré ;

Des notations ci-dessus, résultent entre autres, les développements immédiats ci-dessous :

* $n=n_0+n_1\ldots+n_k\ldots+n_m$;
* $f_k=n_k/n$;
* $\sum_{i=1}^{i=n}k_i=\sum_{k=0}^{k=m}k.n_k=n.\bar{n}$;
* $\bar{n}=\sum_{i=1}^{i=n}k_i\Big/n=\sum_{k=0}^{k=m}k.n_k\Big/n=\sum_{k=0}^{k=m}k.f_k$;
* $s^2=\dfrac{n}{n-1}.\sum_{k=0}^{k=m}f_k.(k-\bar{n})^2$.

II-1°) On dispose des données ci-dessous issues de statistiques établies sur la base de $n=23589$ assurés (risque automobile).

k	0	1	2	3	4	5	6	$\succ 6$	Σ
n_k	20592	2651	297	41	7	0	1	0	23589

Calculer $\bar{n}$ et s^2.

II-2°) « *Méthode de la moyenne et de la fréquence zéro* »

Le principe en est simple et consiste à identifier d'une part l'espérance mathématique $E(N)$ du nombre total de sinistres N avec la moyenne empirique et d'autre part à estimer $\Pr ob(N=0)$ par la fréquence zéro, soit f_0.

Ecrire les équations ainsi obtenues et en déduire une méthode de calcul des estimateurs de p et α qu'on appliquera ensuite numériquement aux données proposées en II-1°).

II-3°) « *Méthode des moments* »

Exprimer $E(N)$ et $Var(N)$ en fonction des moyenne et variance empiriques et en déduire l'expression des estimateurs de p et α. A partir des données numériques proposées en II-1°), quelles valeurs obtient-on quant à ces estimateurs ?

II-4°) « *Méthode du maximum de vraisemblance (E.M.V)* »

Ecrire les équations qui résultent de l'application de la méthode du maximum de vraisemblance et en déduire une procédure pour déterminer les E.M.V de p et de α. Appliquer cette procédure dans le cas des données numériques susmentionnées en II-1°).

II-5°) Comparer tous les résultats antérieurs et en dresser les conclusions appropriées.

Solution : I-1°) On a d'une part Λ, continue et de densité de probabilité $u(\lambda)$, et d'autre part, la loi conditionnelle $N/\Lambda=\lambda$ de type POISSON de paramètre λ. Par extension de la *formule des probabilités totales* ($p(A)=\sum_{i=1}^{i=n} p(H_i).p(A/H_i)$), on a donc :

$$\Pr ob(N=n)=\int_0^{+\infty} p(n,1).u(\lambda).d\lambda=\int_0^{+\infty}\frac{\lambda^n.e^{-\lambda}}{n!}.u(\lambda).d\lambda$$

En particulier, pour Λ de loi Gamma (α,β), on a l'expression :

$$\Pr ob(N=n)=\int_0^{+\infty}\frac{\lambda^n.e^{-\lambda}}{n!}.\frac{\beta^\alpha.e^{-\beta\lambda}.\lambda^{\alpha-1}}{\Gamma(\alpha)}.d\lambda$$

II-2°) Pour rappel, le *théorème de l'espérance totale* s'écrit, pour tout couple de variables aléatoires X et Y, $E(Y)=E\left[E(Y/X)\right]$. Appliquant ce théorème au calcul de la fonction génératrice des moments de N, soit $\Psi_N(t)=E\left[e^{t.N}\right]$, il s'ensuit l'expression $\Psi_N(t)=E_\Lambda\left[E\left[e^{t.N}/\Lambda\right]\right]$.

• Mais d'une part, la variable conditionnelle N/Λ suivant la loi de POISSON de paramètre λ, on a $E\left[e^{t.N}/\Lambda\right]=\sum_{n=0}^{n=+\infty} e^{t.n}.\frac{\lambda^n.e^{-\lambda}}{n!}=e^{\lambda.(e^t-1)}$, soit $\Psi_N(t)=E_\Lambda\left[e^{\lambda.(e^t-1)}\right]$.

• D'autre part, la loi Gamma (α,β) a pour fonction génératrice des moments $\Psi_\Lambda(u)=E\left[e^{\lambda.u}\right]$, soit $\Psi_\Lambda(u)=\int_0^{+\infty}\frac{e^{\lambda u}}{\Gamma(\alpha)}.\beta^\alpha.e^{-\beta.\lambda}.\lambda^{\alpha-1}.d\lambda$, soit après transformation, $\Psi_\Lambda(u)=(\frac{\beta}{\beta-u})^\alpha.\int_0^{+\infty}e^{-\lambda.(\beta-u)}.\frac{(\beta-u)^\alpha}{\Gamma(\alpha)}.\lambda^{\alpha-1}.d\lambda=(\frac{\beta}{\beta-u})^\alpha$ (puisque par définition l'intégrale de la densité de probabilité de la loi Gamma $(\alpha,\beta-u)$ est égale à 1).

• Revenant à $\Psi_N(t)$, il s'agit de $\Psi_\Lambda(u)$ au point $u = e^t - 1$, d'où le résultat :

$$\Psi_N(t) = \left(\frac{\beta}{\beta + 1 - e^t}\right)^\alpha,$$

(fonction dont on notera cependant qu'elle suppose $\frac{\beta}{\beta - u} \succ 0 \Rightarrow u \prec \beta$, soit $t \prec \ln(\beta + 1)$.

I-3°) X étant une variable aléatoire binomiale négative de paramètres r et p, on a d'une part $\Pr ob(X = x) = C_{x-1}^{r-1} p^r . q^{x-r}$, pour $x \geq r$, et d'autre part, $E(X) = \frac{r}{p}$ et $Var(X) = r.\frac{q}{p^2}$, avec $q = 1 - p$ (cf. rappels de cours du chapitre I).

Concrètement, la loi binomiale négative décrit le nombre d'épreuves indépendantes de BERNOULLI nécessaires jusqu'à l'obtention r fois d'un caractère C donné et peut donc s'interpréter comme la somme de r variables aléatoires X_i indépendantes et de loi géométrique (loi de PASCAL) qui caractérisent le nombre d'épreuves nécessaires pour atteindre chacun des r résultats C ci-dessus.

• Or, pour cette loi de PASCAL définie par $\Pr ob(X_i = x) = (1-p)^x . p$, avec $x \geq 1$, on a immédiatement $\Psi_{X_i}(t) = \sum_{x=1}^{x=+\infty} e^{t.x} . (1-p)^{x-1} . p = p.e^t . \sum_{u=0}^{u=+\infty} e^{t.u} . (1-p)^u$ (en posant $u = x - 1$).

Finalement, $\Psi_{X_i}(t) = \frac{p.e^t}{1 - (1-p).e^t}$, par sommation de la série géométrique de raison $e^t.(1-p)$.

• Relativement à la somme $X = \sum_{i=1}^{i=r} X_i$, il s'ensuit $\Psi_X(t) = \prod_{i=1}^{i=r} \Psi_{X_i}(t)$, soit l'expression :

$$\Psi_X(t) = p^r . e^{r.t} . \left[\frac{1}{1 - (1-p).e^t}\right]^r$$

Formant la variable $Y = X - r$, il est immédiat que :

$$\Psi_Y(t) = E\left[e^{t.Y}\right] = E\left[e^{t.(X-r)}\right] = e^{-r.t} . \Psi_X(t) = p^r . \left[\frac{1}{1 - (1-p).e^t}\right]^r$$

• Le changement de variable $Y = X - r$ dans la loi de X conduit à l'équation $\Pr ob(Y = y) = C_{r+y-1}^{r-1} p^r . q^y$, pour $y \geq 0$. Concrètement, $Y = X - r$, représente dans le schéma qui conduit à la loi binomiale négative, le nombre d'échecs nécessaires jusqu'à l'obtention r fois du caractère C.

I-4°) Le rapprochement de $\Psi_N(t) = \left[\frac{\beta}{\beta + 1 - e^t}\right]^\alpha$ avec $\Psi_Y(t) = p^r . \left[\frac{1}{1 - (1-p).e^t}\right]^r$ montre l'analogie entre ces deux fonctions génératrices. En effet, $\Psi_N(t)$ s'écrit aussi sous la forme $\left[\frac{\beta}{\beta + 1}\right]^\alpha . \left[\frac{1}{1 - e^t/\beta + 1}\right]^\alpha$ qui est de la forme $p^r . \left[\frac{1}{1 - (1-p).e^t}\right]^r$ où $p = \frac{\beta}{\beta + 1}$, $r = \alpha$, et $q = 1 - p$.

Les lois de N et de Y sont donc de même nature et on peut définir la loi de N par l'équation $Prob(N=n)=C_{\alpha+n-1}^{\alpha-1}p^{\alpha}.q^{n}$, avec $n\geq 0, p=\dfrac{\beta}{\beta+1}, q=1-p$.

I-5°) N s'écrivant sous la forme $N=X-\alpha$ où X suit la loi binomiale négative de paramètres α et p, on a $E(N)=E(X)-\alpha=\dfrac{\alpha}{p}-\alpha$, soit en remplaçant p par $\dfrac{\beta}{\beta+1}$, $E(N)=\dfrac{\alpha.}{\beta}$. D'autre part, $Var(N)=Var(X)=\alpha.\dfrac{q}{p^2}=\dfrac{\alpha.(\beta+1)}{\beta^2}$.

II-1°) Le calcul de $\bar{n}$ et de s^2 avec les données et les notations proposées dans l'énoncé conduit aux expressions $\bar{n}=\dfrac{\sum_{k=0}^{k=6}k.n_k}{\sum_{k=0}^{k=6}n_k}$ et $s^2=\dfrac{1}{n-1}.\sum_{k=0}^{k=6}n_k.(k-\bar{n})^2$, avec $n=\sum_{k=0}^{k=6}n_k$.

Il s'ensuit le tableau de calculs ci-après.

k	n_k	$k.n_k$	$n_k.(k-\bar{n})^2$
0	20592	0	428,3000
1	2651	2651	1941,4859
2	297	594	1022,8443
3	41	123	334,3747
4	7	28	104,0693
5	0	0	0
6	1	6	34,2902
$\succ 6$	0	0	0
Σ	23589	3402	3865,3644

Numériquement, il en résulte $\bar{n}=0,14422$ et $s^2=0,16386$.

II-2°) L'identification de $E(N)$ avec $\bar{n}$ conduit à la relation $\alpha.\dfrac{1-p}{p}=\bar{n}$ puisque $E(N)=\alpha.\dfrac{1-p}{p}$. Par ailleurs, l'identification de $Prob(N=0)$ avec la fréquence f_0, conduit à la relation $f_0=p^{\alpha}$. En effet, la probabilité de ne pas avoir de sinistre est fournie par l'équation $Prob(N=0)=C_{\alpha-1}^{0}p^{\alpha}.q^{0}=p^{\alpha}$.

La résolution en α et p, des deux équations $\bar{n}=\dfrac{\alpha.(1-p)}{p}$ et $f_0=p^{\alpha}$ conduit d'une part à $\alpha=\dfrac{\bar{n}.p}{1-p}$ et d'autre part à $\alpha=\dfrac{\ln f_0}{\ln p}$, ce qui entraîne $p=\dfrac{1-p}{\bar{n}}.\dfrac{\ln f_0}{\ln p}$. Une méthode par approximations (de type NEWTON ou « du point fixe ») est nécessaire pour résoudre l'équation précédente. La mise en œuvre de ces calculs conduit au valeurs numériques $p=0,8887$ et $\alpha=1,153$.

II-3°) Pour rappel, la méthode des moments consiste à exprimer α et p en fonction de $E(N)$ et de $E(N^2)$, ou encore plus simplement, de $E(N)$ et de $Var(N)$.

L'identification de ces deux derniers paramètres représentatifs avec $\bar{n}$ et s^2 conduit aux expressions cherchées de α et p. Ainsi de $E(N)=\alpha.\frac{1-p}{p}$ et $Var(N)=\frac{\alpha.(1-p)}{p^2}$ déduit-on $\bar{n}=\frac{\alpha.(1-p)}{p}$ et $s^2=\frac{\alpha.(1-p)}{p^2}$, d'où $p=\frac{\bar{n}}{s^2}$ et $\alpha=\frac{\bar{n}^2}{s^2-\bar{n}}$. Il en résulte numériquement, les valeurs $p=0,88013$ et $\alpha=1,0590$.

II-4°) Utilisant l'expression trouvée en I-4°) pour la loi de N, soit $Prob(N=n)=C_{\alpha+n-1}^{n}p^{\alpha}.q^{n},(n\in N)$, la fonction de vraisemblance associée à un échantillon de taille n de variable parente N, soit $(N_1,N_2,...,N_n)$, est définie par l'équation $L(N_1,N_2,...,N_n,\alpha,p)=\prod_{i=1}^{i=n}C_{\alpha+n_i-1}^{n_i}p^{\alpha}.(1-p)^{n_i}$.

Mais $C_{\alpha+n_i-1}^{n_i}=\frac{(\alpha+n_i-1)!}{n_i!.(\alpha-1)!}=\frac{1}{n_i!}.(\alpha+n_i-1).(\alpha+n_i-2)...\alpha=\prod_{j=1}^{j=n_i}\frac{\alpha+n_i-j}{n_i!}$. On a donc, s'agissant de la vraisemblance, $L(N_1,N_2,...,N_n,\alpha,p)=\prod_{i=1}^{i=n}(\prod_{j=1}^{j=n_i}\frac{\alpha+n_i-j}{n_i!}).p^{\alpha}.(1-p)^{n_i}$.

Le passage à la log- vraisemblance, soit $\ln L$, conduit à :

$$\ln L=\left[\sum_{i=1}^{i=n}(\sum_{j=1}^{j=n_i}\ln(\alpha+n_i-j)-\ln(n_i!))\right]+n.\alpha.\ln p+(\sum_{i=1}^{i=n}n_i).\ln(1-p).$$

- Les estimateurs E.M.V cherchés, soient α^* et p^*, sont solutions du système d'équations :

$$\begin{cases}\frac{\partial \ln L}{\partial p}=\sum_{i=1}^{i=n}(\frac{\alpha}{p}-\frac{n_i}{1-p})=0\\ \frac{\partial \ln L}{\partial \alpha}=\sum_{i=1}^{i=n}\sum_{j=1}^{j=n_i}\frac{1}{\alpha+n_i-j}+n.\ln p=0\end{cases}$$

Ici encore, la résolution numérique d'un tel système exige l'utilisation d'une méthode d'approximation. La mise en œuvre de l'une de ces méthodes conduit pour les données proposées aux résultats numériques $p=0,8857$ et $\alpha=1,1181$.

II-5°) Le tableau récapitulatif ci-dessous souligne le faible écart entre les trois méthodes utilisées, comparaison qui reste cependant à justifier par le calcul des risques quadratiques associés.

Méthode utilisée	α	p	$E(N)$	Var(N)
Méthode de la moyenne et de la fréquence zéro	1,153	0,8887	0,14440	0,1625
Méthode des moments	1,0590	0,88013	0,14423	0,1639
Méthode du maximum de vraisemblance	1,1181	0,8857	0,14429	0,1629

On remarque également que les espérances et variances ainsi obtenues sont très proches de la moyenne et de la variance empirique calculées en II-1°).

2.2 Comment estimer un paramètre « intime »

Pour certains sujets (frauder le code du travail, par exemple), on court le risque que les personnes sondées ne répondent pas franchement aux questions de peur d'être poursuivies, ce qui fausse le résultat de l'enquête. Dans la méthode suggérée ici, les réponses sont aléatoirement inversées ce qui garantit une certaine confidentialité.

Enoncé : Notant par p la proportion inconnue du caractère « intime » à estimer, l'enquêteur demande à chaque personne interrogée de lancer un dé. Si le résultat est six, la personne doit donner sa réponse sans mentir, sinon elle doit donner une réponse contraire à la vérité. Le tirage est effectué de sorte que l'enquêteur ignore le résultat du dé, si bien que la personne sondée est bien assurée du secret qui cache son réel comportement.

1-a) Déterminer en fonction de p, la probabilité π qu'une réponse recueillie par l'enquêteur, soit positive.

1-b) Fournir un estimateur sans biais et convergent de π et en déduire un estimateur de p, soit T_n.

1-c) Montrer que T_n est sans bais et convergent.

1-d) Comparer $Var(T_n)$ à la variance de l'estimateur obtenu par le sondage simple classique et en déduire le « prix à payer »en termes d'efficacité, généré par la méthode développée ici

2°) On généralise les résultats précédents à des épreuves pour lesquelles chaque personne doit répondre franchement avec la probabilité α et non sincèrement avec la probabilité complémentaire $1-\alpha$.

2-a) Exprimer $T_n, E(T_n)$ et $Var(T_n)$.

2-b) Discuter suivant les valeurs de α, l'efficacité de l'estimateur obtenu.

Solution : 1-a) Désignant par A l'évènement « réponse positive » pour une personne interrogée et par H_1 et H_2 les hypothèses respectives d'un résultat « six » pour le dé jeté et d'un autre résultat à l'issue dudit jet, la formule des probabilités totales ($p(A)=\sum_{i=1}^{i=n} p(H_i).p(A/H_i)$) conduit immédiatement à la relation :

$$\pi = \frac{1}{6}.p + \frac{5}{6}.(1-p) = \frac{5}{6} - \frac{2}{3}.p.$$

En effet, compte tenu des notations, on a immédiatement, $p(H_1)=\frac{1}{6}, p(H_2)=\frac{5}{6}$, $P(A/H_1)$=Prob(répondre « oui sans mentir »)= p, $P(A/H_2)$ = Prob(répondre « oui en mentant »)= $1-p$.

1-b) La fréquence empirique F_n des « oui » collectés au sein de l'échantillon de taille n considéré, constitue un estimateur sans biais et convergent de π (cf. rappels de cours du présent chapitre – paragraphe 3.5).

Or la relation susmentionnée entre π et p, s'écrit aussi $p=\frac{5}{4}-\frac{3\pi}{2}$. On retiendra donc pour p, l'estimateur $T_n=\frac{5}{4}-\frac{3.F_n}{2}$.

1-c) F_n étant sans biais ($E(F_n)=\pi$), la linéarité de l'espérance mathématique appliquée à T_n entraîne $E(T_n)=\frac{5}{4}-\frac{3}{2}.E(F_n)=\frac{5}{4}-\frac{3}{2}.\pi=\frac{5}{4}-\frac{3}{2}.(\frac{5}{6}-\frac{2p}{3})=p$. T_n est donc un estimateur sans biais.

D'autre part, la relation $Var(a.X+b)=a^2.Var(X),\forall(a,b)$, entraîne $Var(T_n)=\frac{9}{4}.Var(F_n)$, relation dans laquelle, par ailleurs, $Var(F_n)=\frac{\pi.(1-\pi)}{n}$ (se reporter aux rappels de cours du chapitre I –paragraphe 2.3). En résumé, $Var(T_n)=\frac{9}{4}.\frac{\pi.(1-\pi)}{n}$, soit en remplaçant π par son expression en fonction de p, le résultat $Var(T_n)=\frac{1}{16.n}.(5-4.p).(1+4.p)$

On constate que $\lim_{n\to+\infty} Var(T_n)=0$ pour ce qui est de cet estimateur sans biais qui, en conséquence, est convergent.

1-d) Dans le cas d'un sondage classique, le risque associé à l'estimateur F_n de p est égal à $Var(F_n)=\frac{p.(1-p)}{n}$. En effet, F_n étant sans biais, $R(F_n)=E\left[(F_n-p)^2\right]=Var(F_n)$.

Par ailleurs, désignant par $R(T_n)$ le risque associé à T_n, on a $R(T_n)=Var(T_n)$ puisque T_n constitue lui aussi un estimateur sans biais de p. La comparaison entre les deux risques en question, conduit donc à $R(T_n)-R(F_n)=\frac{1}{16.n}.(5-4.p).(1+4.p)-\frac{p.(1-p)}{n}$, soit $R(T_n)-R(F_n)=\frac{5}{16.n}$.

Ainsi $R(T_n)=R(F_n)+\frac{5}{16.n}$, la perte d'efficacité résultant du dispositif de confidentialité, « le prix à payer », devenant négligeable dès que n est grand.

2-a) Dans le cas général, la formule des probabilités totales entraîne la relation $\pi=\alpha.p+(1-\alpha).(1-p)$, soit $p=\frac{\pi-1+\alpha}{2.\alpha-1}$. Partant de la fréquence empirique F_n des « oui » recueillis et considérant l'estimateur T_n de p, on a donc $T_n=\frac{F_n+\alpha-1}{2.\alpha-1}$.

- Suivant la linéarité de l'espérance mathématique, $E(T_n)=\frac{1}{2.\alpha-1}.E(F_n)+\frac{\alpha-1}{2.\alpha-1}$, avec $E(F_n)=\pi$. On a donc $E(T_n)=\frac{1}{2.\alpha-1}.\left[(2.\alpha-1).p+1-\alpha\right]+\frac{\alpha-1}{2.\alpha-1}=p$.

Ainsi l'estimateur T_n est-il un estimateur sans biais de p.

- D'autre part, $Var(T_n)=\frac{1}{(2.\alpha-1)^2}.Var(F_n)=\frac{\pi.(1-\pi)}{(2.\alpha-1)^2.n}$, soit en remplaçant π par son expression en fonction de p, $Var(T_n)=\frac{p.(1-p)}{n}+\frac{\alpha.(1-\alpha)}{n.(2.\alpha-1)^2}$.

2-b) Bien entendu, T_n reste convergent ici puisque $\lim_{n\to+\infty} Var(T_n) = 0$. Par ailleurs, pour n fixé, on remarque que $Var(T_n)$ tend vers l'infini lorsque α tend vers $1/2$. A contrario, $Var(T_n)$ est minimale pour $\alpha = 0$ ou $\alpha = 1$, ce qui n'a pas d'intérêt puisqu'on est ramené alors à un sondage classique sans confidentialité.

Le problème est donc de choisir α suffisamment grand pour que la confidentialité soit crédible mais d'une valeur éloignée de la valeur $1/2$ qui est le cas de la confidentialité maximale où on perd toute précision d'estimation. La valeur $\alpha = 1/6$ qui correspond à l'épreuve de la 1ère question semble répondre ici à ce souci de compromis.

2.3 Le comptage des poissons dans un lac (méthode de capture et recapture)

Enoncé : Pour estimer le nombre total inconnu N de poissons dans un lac, on prélève dans un premier temps un ensemble de m poissons que l'on bague et que l'on rejette à l'eau (m fixé). Puis laissant les poissons se mélanger totalement parmi leurs congénères, on prélève avec remise, un nouveau groupe de n poissons au sein duquel on observe la variable aléatoire X égale au nombre de poissons bagués dans l'échantillon ainsi prélevé.

1°) Déterminer l'estimateur du maximum de vraisemblance de N dont on montrera qu'il est égal à $\widehat{N} = \frac{n.m}{X}$.

2°) Déterminer la variance asymptotique de N et en déduire un intervalle de confiance au seuil $1-\alpha = 95\%$.

3°) On suppose ici $m = n = 100$. Par ailleurs huit des poissons parmi les 100 recapturés se sont avérés présenter une bagne. Déterminer dans ces conditions, une estimation par intervalle de confiance au seuil $1-\alpha = 95\%$ de la population de poissons dans le lac considéré.

Solution : 1°) Après marquage des poissons prélevés à la première capture, la probabilité des poissons bagués au sein du lac est égale à $\frac{m}{N}$.

Pour ce qui est de la recapture, le modèle statistique correspondant est de type BERNOULLI, la variable X_i qui caractérise le baguage ou non du $i^{ème}$ poisson prélevé ayant pour loi $(\frac{m}{N})^{X_i}.(1-\frac{m}{N})^{1-X_i}$. En effet, $\Pr ob(X_i = 1) = \frac{m}{N}$ et $\Pr ob(X_i = 0) = 1 - \frac{m}{N}$.

- La fonction de vraisemblance associée à l'échantillon $(X_1, X_2, ..., X_n)$ des poissons prélevés par recapture s'écrit $L(X_1, X_2, ..., X_n, N) = (1-\frac{m}{N})^n . \prod_{i=1}^{i=n} (\frac{m}{N})^{X_i}.(1-\frac{m}{N})^{-X_i}$, du moins en supposant l'indépendance des X_i ce qui est le cas ici puisque la pêche a lieu avec remise.

La recherche de l'estimateur E.M.V de N (estimateur du maximum de vraisemblance) conduit à rechercher la solution de $\frac{\partial}{\partial N} \ln L = 0$ en considérant la fonction de log- vraisemblance $\ln L = n.\ln(1-\frac{n}{N}) + (\sum_{i=1}^{i=n} X_i).\left[\ln\frac{m}{N} - \ln(1-\frac{m}{N})\right]$.

Il s'ensuit $\frac{\partial}{\partial N}\ln L = \frac{n.m}{N^2}.\frac{1}{1-\frac{m}{N}} - \frac{m}{N^2}.\frac{N}{m}.\sum_{i=1}^{i=n} X_i - \frac{m}{N^2}.\frac{1}{1-\frac{m}{N}}.\sum_{i=1}^{i=n} X_i$, soit après réduction au même dénominateur et simplifications, $\frac{\partial}{\partial N}\ln L = \frac{n.m - N.\sum_{i=1}^{i=n} X_i}{N.(N-m)}$. En résumé, l'estimateur E.M.V cherché, soit $\widehat{N}$, solution de $\frac{\partial}{\partial N}\ln L = 0$, est défini par $\widehat{N} = \frac{n.m}{\sum_{i=1}^{i=n} X_i} = \frac{n.m}{X}$ (en effet, $X = \sum_{i=1}^{i=n} X_i$).

• On remarquera qu'il s'agit bien d'un maximum puisqu'on a immédiatement après calculs, $\frac{\partial^2}{\partial N^2}\ln L = \frac{N^2.\sum_{i=1}^{i=n} X_i - 2.N.n.m + n.m^2}{N^2.(N-m)^2}$, soit pour la valeur de N égale à $\widehat{N}$, l'expression $\frac{\frac{n^2.m^2}{X^2}.X - \frac{2.n^2.m^2}{X} + n.m^2}{\frac{n^2.m^2}{X^2}.(\frac{n.m}{X} - m)^2}$, c'est-à-dire, $\frac{\partial^2}{\partial N^2}\ln L = -\frac{X^3}{n.m^2.(n-X)}$. Ainsi cette dérivée seconde $\frac{\partial^2}{\partial N^2}\ln L$ est-elle bien négative ou nulle au point $N = \widehat{N}$, ce qui assure le caractère maximal susmentionné.

2°) Estimateur E.M.V, $\widehat{N}$ est asymptotiquement sans biais, efficace, et de loi normale, sa variance étant égale à l'inverse de la quantité d'information de FISHER. Le calcul de cette dernière égale à $I_{(X_1,X_2,\ldots,X_n)}(N) = -E\left[\frac{\partial^2}{\partial N^2}\ln L\right]$ conduit à partir de l'expression de $\frac{\partial^2}{\partial N^2}\ln L$ calculée ci-dessus au résultat $E\left[\frac{\partial^2}{\partial N^2}\ln L\right] = E\left[\frac{N^2.\sum_{i=1}^{i=n} X_i - 2.N.m + n.m^2}{N^2.(N-m)^2}\right]$, soit par linéarité de l'espérance mathématique, $\frac{N^2}{N^2.(N-m)^2}.\sum_{i=1}^{i=n} E(X_i) + \frac{n.m^2 - 2.N.n.m}{N^2.(N-m)^2}$.

Or, $E(X_i) = \frac{m}{N}$ (il s'agit de l'espérance de la variable de BERNOULLI égale à $1\times\frac{m}{N} + 0\times(1-\frac{m}{N}) = \frac{m}{N}$). On obtient donc finalement :

$$E\left[\frac{\partial^2}{\partial N^2}\ln L\right] = \frac{N^2}{N^2.(N-m)^2}.\frac{n.m}{N} + \frac{n.m^2 - 2.N.m}{N^2.(N-m)^2} = -\frac{n.m}{N^2.(N-m)},$$

ce qui entraîne $I_{(X_1,X_2,\ldots,X_n)}(N) = \frac{n.m}{N^2.(N-m)}$. En conclusion, $\widehat{N}$ converge vers la loi normale de moyenne N et de variance $\frac{N^2.(N-m)}{n.m}.N$

• Désignant par t_α le nombre vérifiant $Prob(|\xi| \le t_\alpha) = 1-\alpha = 2.\Pi(t_\alpha) - 1$, avec $\Pi(t) = Prob(\xi \le t)$, il vient pour le seuil $1-\alpha = 95\%$ et par lecture dans la table des valeurs annexées de la fonction $\Pi(t)$, le résultat $t_\alpha = 1,96$.

L'intervalle de confiance de N s'écrit donc :

$$\widehat{N} - 1,96.N.\sqrt{\frac{N-m}{n.m}} \le N \le \widehat{N} + 1,96.N.\sqrt{\frac{N-m}{n.m}}$$

qu'on pourra approcher par :

$$\widehat{N} - 1,96.\widehat{N}.\sqrt{\frac{\widehat{N}-m}{n.m}} \le N \le \widehat{N} + 1,96.\widehat{N}.\sqrt{\frac{\widehat{N}-m}{n.m}}$$

puisque la valeur exacte de N est inconnue.

3°) Suivant l'application numérique proposée, on a $m = 100, n = 100, X = 8$. Il en résulte d'une part, $\widehat{N} = \frac{100 \times 100}{8} = 1250$ et d'autre part, l'estimation par intervalle de confiance, $420 \le \widehat{N} \le 2080$.

- • L'inconvénient de l'estimateur E.M.V, $\widehat{N} = \frac{n.m}{X}$, est qu'il n'est pas défini lorsque $X = 0$. Pour lever cette difficulté on pourra par exemple, faire appel à l'estimateur $\widehat{N} = \frac{n.m}{X+1}$ en partant du principe que les résultats ne sont pas trop faussés si n, m, et X sont suffisamment importants.

2.4 Estimation du nombre de fraudeurs dans un transport collectif (loi géométrique)

Enoncé : On considère la loi géométrique (ou de PASCAL) de paramètre p caractérisée par $Prob(X = x) = (1-p)^{x-1}.p, x \in N^*$.

1°) Rappeler l'expression de l'estimateur du maximum de vraisemblance de p, soit $\widehat{p}$.

2°) En déduire, pour n assez grand, un intervalle de confiance de p au seuil $1-\alpha = 95\%$.

3°) Pour évaluer le nombre de passagers en situation irrégulière sur une ligne ferrovière donnée, il est procédé à un jour déterminé, à une grande enquête suivant laquelle les contrôleurs notent le nombre de billets qu'ils valident jusqu'à rencontrer un fraudeur, celui-ci étant inclus dans le nombre constaté, et ainsi de suite....

Les données ainsi recueillies sont les suivantes :

67	101	63	58	16	37	98	51	107	151
111	28	6	76	63	58	91	25	12	73
208	93	131	37	55	72	29	35	16	93

3-a) En déduire une estimation de la probabilité de fraude et son intervalle de confiance au seuil $1-\alpha = 95\%$.

3-b) Quelle estimation du nombre de fraudeurs peut-on prévoir sur une population de 10000 voyageurs ?

Solution : 1°) Comme cela a été indiqué en rappels de cours, l'estimateur E.M.V cherché est défini par la valeur $\hat{p}$ de p qui est solution de l'équation $\frac{\partial}{\partial p}\ln L = 0$ avec

$L(X_1, X_2, ..., X_n, p) = \prod_{i=1}^{i=n}(1-p)^{X_i}.p = (\frac{p}{1-p})^n.(1-p)^{\sum_{i=1}^{i=n}X_i}$.

On a donc $\frac{\partial}{\partial p}\left[n.\ln p - n.\ln(1-p) + (\sum_{i=1}^{i=n}X_i).\ln(1-p)\right] = 0$, soit après dérivation et simplifications, l'estimateur $\hat{p} = \frac{n}{\sum_{i=1}^{i=n}X_i}$ (on a immédiatement $\frac{\partial^2}{\partial p^2}L \prec 0$ en $p = \hat{p}$ ce qui confirme le caractère maximal de la solution trouvée).

2°) Estimateur du maximum de vraisemblance, $\hat{p}$ est asymptotiquement sans biais, efficace, et de loi normale. Plus précisément, pour n assez grand, $\hat{p} - p$ converge vers la loi normale de moyenne 0 et de variance $\frac{1}{n.I_X(p)}$, $I_X(p)$ désignant la quantité d'information de FISHER relative à la variable « parente » X, c'est-à-dire la variable de loi $\Pr ob(X = x) = (1-p)^{x-1}.p$.

Calculant $I_X(p)$ et remarquant que la fonction de log- vraisemblance $\ln L(X, p)$ est deux fois dérivable, on obtient $I_X(p) = -E\left[\frac{\partial^2}{\partial p^2}\ln L(X, p)\right]$, soit en développant les calculs, $\ln L(X, p) = (X-1).\ln(1-p) + \ln p$, puis en dérivant $\frac{\partial}{\partial p}\ln L(X, p) = \frac{X-1}{1-p} + \frac{1}{p}$, et $\frac{\partial^2}{\partial p^2}\ln L(X, p) = -\frac{X-1}{(1-p)^2} - \frac{1}{p^2}$, et enfin $-E\left[\frac{\partial^2}{\partial p^2}\ln L\right] = \frac{1}{(1-p)^2}.[E(X)-1] + \frac{1}{p^2}$.

Posant $q = 1-p$ et remarquant que $E(X) = \frac{1}{p}$, on a donc $I_X(p) = \frac{1}{q^2}.(\frac{1}{p}-1) + \frac{1}{p^2}$, soit $I_X(p) = \frac{p - p^2 + q^2}{p^2.q^2} = \frac{p.(1-p) + q^2}{p^2.q^2} = \frac{p.q + q^2}{p^2.q^2} = \frac{1}{p^2.q}$.

En conclusion, $\frac{1}{n.I_X(p)} = \frac{p^2.q}{n}$, ce qui conduit, pour n grand à la convergence de $\hat{p} - p$ vers la loi normale centrée (de moyenne nulle) et de variance $\frac{p^2.q}{n}$.

- Désignant par t_α le nombre vérifiant $\Pr ob(|\xi| \le t_\alpha) = 2.\Pi(t_\alpha) - 1 = 1 - \alpha$ (où $\Pi(t) = \Pr ob(\xi \le t)$, on en déduit l'intervalle de confiance :

$$\hat{p} - t_\alpha.\sqrt{\frac{p^2.q}{n}} \le p \le \hat{p} + t_\alpha.\sqrt{\frac{p^2.q}{n}}, \text{ avec } q = 1-p.$$

La valeur de p étant inconnue, on assimilera p à son estimateur $\hat{p}$ pour déterminer pratiquement les bornes de l'intervalle en question, ce qui conduit au résultat ci-après.

$$\hat{p}-t_\alpha\cdot\sqrt{\frac{\hat{p}^2.(1-\hat{p})}{n}}\le p\le\hat{p}+t_\alpha\cdot\sqrt{\frac{\hat{p}^2.(1-\hat{p})}{n}}$$

3-a) Le calcul effectué à partir des données de l'énoncé conduit à $\sum_{i=1}^{i=30}X_i=2061$ et $\hat{p}=0,01456$. Admettant la convergence vers la loi normale $(n=30)$, et partant d'un seuil de confiance $1-\alpha=95\%$, on en déduit $t_\alpha=1,96$.

En effet, $2.\Pi(t_\alpha)-1=0,95\Rightarrow\Pi(t_\alpha)=0,975\Rightarrow t_\alpha=1,96$ par lecture dans la table des valeurs annexées de la fonction $\Pi(t)=\Pr ob(\xi\le t)$.

Compte tenu de l'expression obtenue dans la 2ème question, l'intervalle de confiance de p est donc $0,01456\pm1,96\times0,01456\times\sqrt{\frac{1-0,01456}{30}}$, soit $0,0094\le p\le0,0197$.

3-b) Sur un ensemble de 10000 voyageurs, le nombre de fraudeurs est donc compris entre 94 et 197, ceci au seuil 95%.

2.5 Evaluation d'une contamination (méthode « most powerful number – M.P.N »)

Utilisée en agroalimentaire, environnement, pharmacologie...., cette méthode vise à estimer le degré de contamination d'une population par une bactérie suivant le seul mode d'investigation qui est celui d'un indicateur de présence/absence (on ne sait pas observer le nombre de bactéries présentes).

Enoncé : La méthode repose sur un principe de dilutions, le but étant d'estimer la densité en bactéries (λ nombre de bactéries par unité de volume). On réalise ainsi n prélèvements indépendants de même volume unitaire ($V=1$) et on note :

- Z_k, la variable aléatoire qui décrit le nombre (non observé) de bactéries présentes dans le tube k ;

- X_k, la variable indicatrice égale à 1 s'il n'y a pas de bactérie dans le tube k et 0 sinon.

On suppose que Z_k suit une loi de POISSON de paramètre λ_k. Soit $Y=\sum_{k=1}^{k=n}X_k$ le nombre de tubes ne contenant pas la bactérie considérée (résultats négatifs).

PARTIE I

I-1°) Exprimer en fonction de λ, la probabilité $\pi=\Pr ob(X_k=1)$, puis caractériser les lois de X_K et de Y.

I-2°) Exprimer l'estimateur du maximum de vraisemblance (E.M.V) de π et en déduire l'estimateur E.M.V de λ.

I-3°) Exprimer au seuil de confiance $1-\alpha=95\%$ un intervalle de confiance pour π et pour λ. Expliciter numériquement le résultat obtenu lorsque, parmi 30 prélèvements effectués, 21 d'entre eux sont négatifs.

PARTIE II

Des densités extrêmes (λ proche de 0 ou élevé) induisent des problèmes d'estimation. On pallie à ce travers en utilisant la méthode de dilution exposée ci-après.

- Si on craint de n'avoir que des prélèvements positifs, on dilue ceux-ci d fois, la densité de bactéries étant alors réduite à λ/d et le nombre de bactéries à une loi de POISSON de paramètre λ/d.
- Si on craint de n'avoir que des prélèvements négatifs, on augmente le volume des prélèvements étant entendu que dans un volume d fois plus grand, le nombre de bactéries suit la loi de POISSON de paramètre $\lambda.d$.

La mise en oeuvre de la méthode exige donc une idée à priori de l'ordre de grandeur de λ pour choisir les niveaux de dilutions. On considère ainsi N échantillons formés chacun de n_i prélèvements opérés suivant le taux de dilution $d_i, i \in \{1,2,...,N\}$. On note par Y_i le nombre de prélèvements négatifs au sein du $i^{ème}$ échantillon.

II-1°) Exprimer la loi de Y_i.

II-2°) Ecrire l'équation qui détermine l'estimateur E.M.V de λ, estimateur dit « nombre le plus probable (N.P.P) » ou encore « most powerful number (M.P.N) ».

II-3°) Expliciter la variance asymptotique de l'estimateur N.P.P ci-dessus.

II-4°) On suppose que $N=2, d_1=1, d_2=2, n_1=n_2=30, y_1=27, y_2=21$. Estimer λ ponctuellement et par intervalle de confiance au seuil 95%.

Solution : I-1°) $\Pr ob(X_k=1)=\Pr ob(Z_k=0)$ puisqu'il s'agit d'un prélèvement sans bactérie. Or, Z_k suit la loi de POISSON caractérisée par $\Pr ob(Z_k=x)=\frac{\lambda^x.e^{-\lambda}}{x!}$. Il est bien évident dans ces conditions, que $\pi=\Pr ob(Z_k=0)=e^{-\lambda}$.

- Pour sa part, X_k est la variable de BERNOULLI dont la loi est caractérisée par $\Pr ob(X_k=1)=\pi=e^{-\lambda}$ et $\Pr ob(X_k=0)=1-e^{-\lambda}$. Plus généralement et par définition, $Y=\sum_{i=1}^{i=n} X_i$ suit immédiatement la loi binomiale $B(n,\pi)$ avec $\pi=e^{-\lambda}$.

I-2°) La fonction de vraisemblance $L(X_1,X_2,...,X_n,\pi)$ associée à l'échantillon de taille $(X_1,X_2,...,X_n)$ résultant de n prélèvements, est égale à :

$$L(X_1,X_2,...,X_n,\pi)=\prod_{k=1}^{k=n}\pi^{X_k}.(1-\pi)^{1-X_k}\ ,X_k\in\{0,1\},1\le k\le n.$$

En effet, $\Pr ob(X_k=x_k)$ peut s'écrire sous la forme synthétique $\pi^{X_k}.(1-\pi)^{1-X_k}$, les valeurs prises par X_k étant 0 ou 1.

- Il s'ensuit, pour ce qui est de la log- vraisemblance, l'expression :

$$\ln L=(\sum_{k=1}^{k=n}X_k).\ln\pi+(n-\sum_{k=1}^{k=n}X_k).\ln(1-\pi).$$

La dérivation par rapport à π de la fonction susmentionnée, fournit l'estimateur E.M.V cherché, à savoir la solution de l'équation $\frac{\partial}{\partial\pi}\ln L=\frac{1}{\pi}.\sum_{k=1}^{k=n}X_k-\frac{1}{1-\pi}.(n-\sum_{k=1}^{k=n}X_k)=0$.

Il en résulte l'estimateur E.M.V de π, soit $\hat{\pi} = \frac{\sum_{k=1}^{k=n} X_k}{n} = \frac{Y}{n}$.

• On sait que si $\hat{\theta}$ est E.M.V de θ, $g(\hat{\theta})$ est E.M.V de $g(\theta)$(cf. rappels de cours du présent chapitre). De la relation $\pi = e^{-\lambda}$ qui équivaut à $\lambda = -\ln \pi$, découle pour ce qui est de l'estimateur E.M.V de λ, soit $\hat{\lambda}$, le résultat $\hat{\lambda} = -\ln \hat{\pi} = -\ln(\frac{Y}{n})$.

I-3°) L'estimation par intervalle de confiance au seuil $1-\alpha$ de :la proportion π s'écrit :

$$\hat{\pi} - t_\alpha \cdot \sqrt{\frac{\pi.(1-\pi)}{n}} \le \pi \le \hat{\pi} + t_\alpha \cdot \sqrt{\frac{\pi.(1-\pi)}{n}} \text{ où } t_\alpha \text{ vérifie } Prob(|\xi| \le t_\alpha) = 1-\alpha .$$

Pour l'application proposée $1-\alpha = 95\%$. Par ailleurs, le résultat ci-dessus suppose la validité de la convergence vers la loi normale, ce qu'on peut tout juste admettre ici puisque $n = 30$. Utilisant l'approximation de π par $\hat{\pi}$ pour ce qui est du facteur $\frac{\pi.(1-\pi)}{n}$, il vient finalement l'encadrement $\hat{\pi} - t_\alpha \cdot \sqrt{\frac{\hat{\pi}.(1-\hat{\pi})}{n}} \le \pi \le \hat{\pi} + t_\alpha \cdot \sqrt{\frac{\hat{\pi}.(1-\hat{\pi})}{n}}$.

Numériquement, $Prob(|\xi| \le t_\alpha) = 2.\Pi(t_\alpha) - 1 = 0,95 \Rightarrow \Pi(t_\alpha) = 0,975$. Par lecture dans la table des valeurs annexée de la fonction $\Pi(t) = Prob(\xi \le t)$, il vient $t_\alpha = 1,96$. D'autre part, $\hat{\pi} = \frac{21}{30} = 0,7$ et $n = 30$. On obtient donc l'intervalle $0,54 \le \pi \le 0,86$.

• La transformation $\lambda = -\ln \pi$, induit pour λ, l'intervalle de confiance :

$$-\ln(\hat{\pi} + t_\alpha \cdot \sqrt{\frac{\hat{\pi}.(1-\hat{\pi})}{n}}) \le \lambda \le -\ln(\hat{\pi} - t_\alpha \cdot \sqrt{\frac{\hat{\pi}.(1-\hat{\pi})}{n}})$$

soit, numériquement, le résultat $0,15 \le \lambda \le 0,62$. On remarque que l'intervalle en question n'est plus centré sur l'estimateur ponctuel $\hat{\lambda}$ qui, par ailleurs, est égal à $-\ln \hat{\pi} = 0,36$.

II-1°) Tout d'abord, il est manifeste que l'encadrement précédent de λ est inopérant dans l'hypothèse des densités extrêmes $\lambda = 0$ et $\lambda = 1$. En effet, $\lambda = 0 \Rightarrow \pi = 1$, ce qui conduit à la non définition de $\frac{\partial}{\partial \pi} \ln L$ et à l'impossibilité d'estimer π et λ. C'est la même impossibilité qui a lieu quand $\lambda = 1$ puisqu'on a alors $\pi = 0$.

Par ailleurs, on assimilera dans la suite, les deux cas de figure induits par le principe de dilution au cas $\lambda \to \lambda.d$, avec $d \succ 1$ lorsqu'il s'agit d'une augmentation du volume prélevé et $d \prec 1$ lorsqu'il s'agit de diluer la solution prélevée.

Dans ces conditions et au sein du $i^{ème}$ échantillon, Y_i suit la loi binomiale $B(n_i, \pi_i)$ où π_i désigne ici la probabilité d'obtenir un résultat négatif (absence de bactérie) lors de chacun des prélèvements « k » $(1 \le k \le n_i)$ effectués ainsi dans une solution de taux de dilution d_i. Mais notant par Z_k^i le nombre de bactéries présentes dans un prélèvement de solution diluée au taux d_i, π_i est aussi est égale à $Prob(Z_k^i = 0)$, Z_k^i suivant la loi de POISSON de paramètre $\lambda.d_i$.

En conclusion et compte tenu de l'expression de la loi de POISSON, Y_i suit la loi binomiale $B(n_i, e^{-\lambda.d_i})$, π_i étant égal à $\Pr ob(Z_k^i = 0) = e^{-\lambda.d_i}.\frac{(\lambda.d_i)^0}{0!} = e^{-\lambda.d_i}$.

II-2°) La fonction de vraisemblance associée à la suite $(Y_1, Y_2, ..., Y_N)$ des variables Y_i représentant le nombre de prélèvements négatifs parmi les n_i prélèvements effectués dans la solution de taux de dilution d_i, $1 \le i \le N$, est égale au produit des lois des Y_i, soit la fonction $L(Y_1, Y_2, ..., Y_n, \lambda) = \prod_{i=1}^{i=N} C_{n_i}^{Y_i} e^{-\lambda.d_i.Y_i}.(1-e^{-\lambda.d_i})^{n_i - Y_i}$. L'appel à la log- vraisemblance conduit à $\ln L = \sum_{i=1}^{i=N} \ln(C_{n_i}^{Y_i}) - \sum_{i=1}^{i=N} \lambda.d_i.Y_i + \sum_{i=1}^{i=N} (n_i - Y_i).\ln(1-e^{-\lambda.d_i})$.

La recherche de l'estimateur E.M.V de λ, solution de l'équation $\frac{\partial}{\partial \lambda} \ln L = 0$ entraîne la relation $-\sum_{i=1}^{i=N} d_i.Y_i + \sum_{i=1}^{i=N} (n_i - Y_i).\frac{d_i.e^{-\lambda.d_i}}{1-e^{-\lambda.d_i}} = 0$. On obtient donc pour ce qui est de l'estimateur E.M.V de λ, la valeur $\hat{\lambda}$ vérifiant la relation :

$$\sum_{i=1}^{i=N} d_i.Y_i = \sum_{i=1}^{i=N} (n_i - Y_i).\frac{d_i.e^{-\hat{\lambda}.d_i}}{1-e^{-\hat{\lambda}.d_i}}$$

C'est aussi l'estimateur M.P.N (most powerful number) ou encore N.P.P (nombre le plus probable).

II-3°) La variance de l'estimateur $\hat{\lambda}$ susmentionné obtenu par la méthode du maximum de vraisemblance est asymptotiquement efficace et est donc égale à l'inverse de la quantité d'information de FISHER relative à la suite $(Y_1, Y_2, ..., Y_N)$ et portant sur le paramètre λ (cf. rappels de cours du présent chapitre – paragraphe 3.2).

Calculant cette quantité d'information, soit $I_{(Y_1, Y_2, ..., Y_N)}(\lambda)$, la propriété d'additivité démontrée en rappels de cours du présent chapitre permet d'écrire l'expression $I_{(Y_1, Y_2, ..., Y_N)}(\lambda) = \sum_{i=1}^{i=N} I_{Y_i}(\lambda)$. D'autre part, notant par $p(Y_i, \lambda)$ la loi de Y_i, on a, $p(Y_i, \lambda)$ étant manifestement deux fois dérivable :

$$I_{Y_i}(\lambda) = E\left[(\frac{\partial}{\partial \lambda} \ln p(Y_i, \lambda))^2 \right] = -E\left[\frac{\partial^2}{\partial \lambda^2} \ln p(Y_i, \lambda) \right]$$

Concrètement, $p(Y_i, \lambda) = C_{n_i}^{Y_i} e^{-\lambda.d_i.Y_i}.(1-e^{-\lambda.d_i})^{n_i - Y_i}$, soit par passage au logarithme, $\ln p(Y_i, \lambda) = \ln(C_{n_i}^{Y_i}) - \lambda.d_i.Y_i + (n_i - Y_i).\ln(1-e^{-\lambda.d_i})$. Il s'ensuit, par dérivation :

$$\frac{\partial}{\partial \lambda} \ln p(Y_i, \lambda) = -d_i.Y_i + \frac{d_i.e^{-\lambda.d_i}.(n_i - Y_i)}{1-e^{-\lambda.d_i}} = \frac{-d_i.Y_i + d_i.Y_i.e^{-\lambda.d_i} + n_i.d_i.e^{-\lambda.d_i} - d_i.Y_i.e^{-\lambda.d_i}}{1-e^{-\lambda.d_i}}.$$

Après simplifications, $\frac{\partial}{\partial \lambda} p(Y_i, \lambda) = \frac{-d_i.Y_i + n_i.d_i.e^{-\lambda.d_i}}{1-e^{-\lambda.d_i}}$. Dérivant de nouveau et développant, il s'ensuit $\frac{\partial^2}{\partial \lambda^2} \ln p(Y_i, \lambda) = \frac{-n_i.d_i^2.e^{-\lambda.d_i} + d_i^2.Y_i.e^{-\lambda.d_i}}{(1-e^{-\lambda.d_i})^2}$.

- Passant à l'espérance mathématique, on obtient par linéarité :

$$E\left[\frac{\partial^2}{\partial\lambda^2}\ln p(Y_i,\lambda)\right]=\frac{d_i^2.e^{-\lambda.d_i}}{(1-e^{-\lambda.d_i})^2}.E(Y_i)-\frac{n_i.d_i^2.e^{-\lambda.d_i}}{(1-e^{-\lambda.d_i})^2}$$

Mais $E(Y_i)=n_i.e^{-\lambda.d_i}$ (puisque Y_i suit la loi binomiale $B(n_i,e^{-\lambda.d_i})$. En définitive, on obtient $E\left[\frac{\partial^2}{\partial\lambda^2}\ln p(Y_i,\lambda)\right]=\frac{n_i.d_i^2.e^{-2\lambda.d_i}}{(1-e^{-\lambda.d_i})^2}-\frac{n_i.d_i^2.e^{-\lambda.d_i}}{(1-e^{-\lambda.d_i})^2}$, ce qui conduit, après factorisation suivant $1-e^{-\lambda.d_i}$, au résultat $I_{Y_i}(\lambda)=-E\left[\frac{\partial^2}{\partial\lambda^2}\ln p(Y_i,\lambda)\right]=\frac{n_i.d_i^2.e^{-\lambda.d_i}}{(1-e^{-\lambda.d_i})}$.

- Ainsi, revenant à $I_{(Y_1,Y_2,\ldots,Y_N)}(\lambda)=\sum_{i=1}^{i=N}I_{Y_i}(\lambda)$, a-t-on $I_{(Y_1,Y_2,\ldots,Y_N)}(\lambda)=\sum_{i=1}^{i=N}\frac{n_i.d_i^2.e^{-\lambda.d_i}}{(1-e^{-\lambda.d_i})}$.

La variance asymptotique de l'estimateur « N.P.P », soit $\hat{\lambda}$, dont il est rappelé que c'est l'inverse de la quantité d'information de FISHER, est donc égale à :

$$\frac{1}{I_{(Y_1,Y_2,\ldots,Y_n)}(\lambda)}=\left[\sum_{i=1}^{i=N}\frac{n_i.d_i^2.e^{-\lambda.d_i}}{1-e^{-\lambda.d_i}}\right]^{-1}$$

II-4°) La relation qui fournit l'estimateur M.P.N (ou N.P.P) s'écrit, pour les données numériques proposées, $y_1+2.y_2=(30-y_1).\frac{e^{-\lambda}}{1-e^{-\lambda}}+(30-y_2).\frac{2.e^{-2.\lambda}}{1-e^{-2.\lambda}}$.

Remplaçant y_1 et y_2 par leurs valeurs respectives 27 et 21, on aboutit donc à l'équation en λ, $69=\frac{3.e^{-\lambda}}{1-e^{-\lambda}}+\frac{18.e^{-2.\lambda}}{1-e^{-2.\lambda}}$, équation qui, en posant $z=e^{-\lambda}$, s'écrit, après développement, $90z^2+3z-69=0$. Il en résulte les solutions, $z_1=0,859$ et $z_2=-0,8924$ desquelles on déduit, pour la seule valeur $\hat{\lambda}$ admissible, le résultat $\hat{\lambda}=-\ln z_1=0,152$.

- Quant à la variance asymptotique, elle est égale ici à $\left[\frac{30.e^{-\hat{\lambda}}}{1-e^{-\hat{\lambda}}}+\frac{120.e^{-2.\hat{\lambda}}}{1-e^{-2\hat{\lambda}}}\right]^{-1}$, soit la valeur 0,00192. Admettant la convergence de $\hat{\lambda}$ vers la loi normale $N(\hat{\lambda},\ 1/I(\hat{\lambda}))$, on obtient donc pour λ et au seuil de confiance $1-\alpha=95\%$, l'intervalle de confiance :

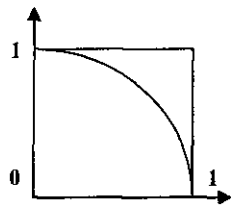
$\hat{\lambda}-1,96.\sqrt{Var(\hat{\lambda})}\le\lambda\le\hat{\lambda}+1,96.\sqrt{Var(\hat{\lambda})}$, soit numériquement, $0,106\le\lambda\le0,194$.

2.6 Estimations du nombre π à l'aide de méthodes de MONTE CARLO

Les méthodes de MONTE CARLO visent à trouver des solutions approchées de nombreux problèmes (calculs d'aires et de volumes, résolution d'équations...) à l'aide de techniques probabilistes fondées sur la simulation de variables aléatoires, dont notamment la loi uniforme. Elles trouvent un large champ d'applications en physique, biologie, chimie, transport, finances....

Enoncé : A travers la présente application, deux méthodes d'estimation du nombre π sont proposées.

PARTIE I



Dans le carré ci-dessous de côté 1, soit $[0,1]\times[0,1]$, on génère des points de façon aléatoire et uniforme et on dénombre ceux qui sont à l'intérieur du quart de disque de centre O et de rayon 1.

Soit X_n la variable de comptage ainsi générée.

I-1°) Quelle est la loi suivie par X_n ?

I-2°) En déduire un estimateur T_1 de π dont on montrera qu'il est sans biais et convergent.

I-3°) Exprimer un intervalle de confiance asymptotique de π suivant cette méthode (on prendra un seuil de confiance $1-\alpha$ égal à 95%).

I-4°) Combien de points aléatoires faut-il générer au minimum pour obtenir, suivant cette méthode d'estimation, une erreur absolue n'excédant pas 3%.

PARTIE II

On se propose d'utiliser la variable $V=\sqrt{1-U^2}$, où U suit la loi uniforme sur $[0,1]$.

II-1°) Montrer que $E(V)=\frac{\pi}{4}$.

II-2°) En déduire, une méthode d'estimation de π à partir d'un échantillon $(U_1,U_2,...,U_n)$ de « loi parente » U, uniforme sur $[0,1]$, l'estimateur ainsi défini étant noté T_2.

II-3°) Montrer que T_2 est sans biais et convergent.

II-4°) Comparer T_2 à T_1.

II-5°) De nouveau à partir d'un intervalle de confiance asymptotique de π obtenu par la présente méthode (toujours au seuil de confiance 95%), déterminer le nombre minimal de variables uniformes à générer pour obtenir une précision absolue d'au moins 3%.

Solution : I-1°) Suivant la probabilité uniforme sur l'espace de probabilités $\Omega=[0,1]^2$, associant à tout évènement aléatoire A la probabilité égale au rapport des aires respectives de A et de Ω, chaque point aléatoire qui est généré est contenu (ou non) dans le quart de disque considéré dans l'énoncé avec la probabilité $p=\frac{\pi}{4}$ (resp. $q=1-\frac{\pi}{4}$).

Dans ces conditions, il est immédiat que la variable aléatoire X_n égale au nombre de points contenus dans le quart de disque de rayon 1, parmi les n points générés par simulation aléatoire, suit la loi binomiale $B(n,\frac{\pi}{4})$.

I-2°) Se référant aux rappels de cours du présent chapitre (paragraphe 3.5), $F_n=\frac{X_n}{n}$ est un estimateur ponctuel de p dont on montre qu'il vérifie $E(F_n)=p$ et $Var(F_n)=\frac{p.(1-p)}{n}$ (cf. chapitre I).

Ainsi, $T_1 = 4.F_n$ constitue-t-il un estimateur ponctuel de π, vérifiant $E(T_1) = 4.E(F_n) = 4.p = \pi$ et $Var(T_1) = 16.Var(F_n) = \dfrac{\pi.(4-\pi)}{n}$. T_1 est donc bien un estimateur sans biais, qui plus est convergent, puisque $\lim\limits_{n\to+\infty} Var(T_1) = 0$.

I-3°) Le théorème de MOIVRE LAPLACE appliqué à X_n entraîne la convergence de F_n et à fortiori de T_1 vers la loi normale (pour n grand), loi caractérisée par la moyenne $E(T_1) = \pi$ et la variance $Var(T_1) = \dfrac{\pi.(4-\pi)}{n}$.

Il en résulte, pour la fonction pivotale $\xi = \dfrac{T_1 - \pi}{\sqrt{\dfrac{\pi.(4-\pi)}{n}}}$ de loi normale centrée réduite $N(0,1)$, l'encadrement $-t_\alpha \le \xi \le t_\alpha$, où t_α vérifie, pour le seuil de confiance de 95% proposé ici, $\Pr ob(|\xi| \le t_\alpha) = 95\%$. Par lecture dans la table des valeurs de la fonction $\Pi(t) = \Pr ob(\xi \le t)$, on obtient immédiatement $t_\alpha = 1,96$.

En conclusion, on obtient pour π, l'intervalle de confiance :

$$T_1 - 1,96.\sqrt{\frac{\pi.(4-\pi)}{n}} \le \pi \le T_1 + 1,96.\sqrt{\frac{\pi.(4-\pi)}{n}}.$$

Approximant π par son estimateur ponctuel T_1 dans les bornes de l'intervalle ci-dessus, l'intervalle de confiance cherché s'écrit finalement :

$$T_1 - 1,96.\sqrt{\frac{T_1.(4-T_1)}{n}} \le \pi \le T_1 + 1,96.\sqrt{\frac{T_1.(4-T_1)}{n}}$$

I-4°) On cherche la valeur minimale à partir de laquelle $1,96.\sqrt{\dfrac{T_1.(4-T_1)}{n}} \le 0,03$ ce qui entraine $n \ge \dfrac{(1,96)^2.T_1.(4-T_1)}{(0,03)^2}$ (E). Comme cela est indiqué dans l'application 3.1 ci-après, la méthode la plus simple est de considérer la fonction $T_1.(4-T_1)$ dont le maximum, de valeur 4, est atteint lorsque $T_1 = 2$.

La condition minimale $n \ge n^* = \dfrac{(1,96)^2}{(0,03)^2}.Max\left[T_1.(1-T_1)\right] \simeq \dfrac{16}{(0,03)^2} = 17800$ garantit la satisfaction de la relation (E) susmentionnée, quelque soit la valeur de T_1.

• On remarquera cependant que le nombre des simulations nécessaires pour obtenir une bonne précision reste très élevé ici.

II- 1°) U ayant pour densité de probabilité $f(u) = 1_{[0,1]}(u)$, on a immédiatement, par définition de l'espérance mathématique $E(V) = \int_R \sqrt{1-u^2}.f(u).du = \int_0^1 \sqrt{1-u^2}.du$, c'est-à-dire, suivant le changement de variable $u = \sin\varphi$:

$$E(V) = \int_0^{\pi/2} \cos^2\varphi.d\varphi = \int_0^{\pi/2} \left[\frac{1+\cos 2\varphi}{2}\right].d\varphi = \frac{\pi}{4}.$$

II-2°) Comme indiqué en rappels de cours du présent chapitre (cf. paragraphe 3.5), la moyenne empirique $\overline{V} = \frac{\sum_{i=1}^{i=n} V_i}{n}$ forme un estimateur sans biais de $E(V)$. En d'autres termes, la statistique $T_2 = 4.\overline{V}$ est un estimateur sans biais de π puisque $E(T_2) = 4.E(\overline{V}) = 4.E(V) = 4.\frac{\pi}{4} = \pi$, estimateur dont la construction s'opère comme suit :

Etape 1 → Générer n valeurs uniformes et indépendantes sur $[0,1]$, soient $U_i, 1 \le i \le n$.

Etape 2 → Former les nombres $V_i = \sqrt{1-U_i^2}, 1 \le i \le n$.

Etape 3 → Calculer $T_2 = \frac{4.\sum_{i=1}^{i=n} V_i}{n}$, estimateur cherché de π.

II-3°) Outre la propriété d'être sans biais démontrée ci-dessus, l'estimateur T_2 vérifie $Var(T_2) = 16.Var(\overline{V}) = 16.\frac{\sum_{i=1}^{i=n} Var(V_i)}{n^2}$, avec $Var(V_i) = E(V_i^2) - E(V_i)^2$.

Or, $E(V_i^2) = \int_0^1 (1-u^2).du = \frac{2}{3}$. Ainsi, $Var(V_i) = \frac{2}{3} - (\frac{\pi}{4})^2$ et $Var(T_2) = \frac{16.n}{n^2}.(\frac{32-3.\pi^2}{48})$

soit en conclusion, $Var(T_2) = \frac{32-3.\pi^2}{3.n}$. Il s'agit bien d'un estimateur convergent puisque $\lim_{n \to +\infty} Var(T_2) = 0$.

II-4°) Pour comparer T_1 et T_2, on considère la différence des risques associés $R(T_1) - R(T_2)$, différence qui, s'agissant d'estimateurs sans biais, est égale à la différence des variances $Var(T_1) - Var(T_2)$. Plus précisément, le développement de cette dernière s'écrit $\frac{\pi.(4-\pi)}{n} - \frac{32-3.\pi^2}{3.n} = \frac{1}{n}.\left[4.\pi - \frac{32}{3}\right] > 0$. On en conclut que l'estimateur T_2 est préférable à T_1 puisque de variance plus faible.

II-5°) Réitérant les développements des questions I-3°) et I-4°), on a immédiatement pour π et au seuil de confiance 95%, l'encadrement par intervalle de confiance suivant :

$$T_2 - 1,96.\sqrt{\frac{32-3.\pi^2}{3.n}} \le \pi \le T_2 + 1,96.\sqrt{\frac{32-3.\pi^2}{3.n}}$$

intervalle qu'on peut approcher par :

$$T_2 - 1,96.\sqrt{\frac{32-3.T_2^2}{3.n}} \le \pi \le T_2 + 1,96.\sqrt{\frac{32-3.T_2^2}{3.n}}$$

- La condition minimale sur n pour obtenir une précision absolue d'au moins 3%, s'écrit cette fois, $n \ge n^* = (1,96)^2.\frac{32-3.T_2^2}{3.(0,03)^2}$. Il s'agit là, d'une méthode plus efficace que celle développée en partie I, puisqu'en minorant T_2 par 0, on obtient $n^* = 11850$ (au lieu de la valeur 17800).

Le lecteur est invité à ce sujet à utiliser des simulateurs proposés sur internet pour constater de visu les résultats obtenus à travers diverses séries de simulations notamment pour la première des deux méthodes exposées dans cette application (méthode « du tir à la cible »).

3. Intervalles de confiance

3.1 Comparaison des méthodes d'approximation pour une proportion

Enoncé : On considère une population de grande taille au sein de laquelle on s'intéresse à la proportion des personnes qui ont les yeux bleus, soit p. L'observation du caractère en question sur un échantillon de 500 personnes conduit à une fréquence « empirique », F_n, égale à 37%.

En déduire une estimation ponctuelle par intervalle de confiance de p, au seuil $1-\alpha = 90\%$, et ceci en utilisant trois méthodes :

a) la majoration de $p.(1-p)$ par $1/4$;

b) l'approximation de p par F_n ;

c) un calcul exact.

Solution : Les éléments présentés en rappels de cours du présent chapitre conduisent, n étant supposé grand (ce qui est le cas ici), à la fonction pivotale $\xi = \dfrac{F_n - p}{\sqrt{\dfrac{p.(1-p)}{n}}}$ de loi limite $N(0,1)$, et en conséquence, à l'encadrement par intervalle de confiance de p, sous la forme :

$$F_n - t_\alpha.\sqrt{\frac{p.(1-p)}{n}} \le p \le F_n + t_\alpha.\sqrt{\frac{p.(1-p)}{n}},$$

t_α vérifiant $\Pr ob(|\xi| \le t_\alpha) = 2.\Pi(t_\alpha) - 1 = 1-\alpha$, avec $\Pi(t) = \Pr ob(\xi \le t)$.

• Pour chacune des méthodes proposées, il s'ensuit numériquement et tenant compte des valeurs $n = 500, F_n = 0,37, \alpha = 90\% \Rightarrow t_\alpha = 1,645$ (solution de $2.\Pi(t_\alpha) - 1 = 0,90$, soit $\Pi(t_\alpha) = 0,95$), les bornes suivantes de l'intervalle de confiance de p :

Méthode a) : $F_n - 1,645.\dfrac{1}{2.\sqrt{n}} \le p \le F_n + 1,645.\dfrac{1}{2.\sqrt{n}}$, soit numériquement , l'intervalle $0,333 \le p \le 0,407$;

Méthode b) : $F_n - 1,645 \times \sqrt{\dfrac{0,37 \times 0,63}{500}} \le p \le F_n + 1,645 \times \sqrt{\dfrac{0,37 \times 0,63}{500}}$, soit l'intervalle $0,334 \le p \le 0,406$;

Méthode c) : les bornes de l'intervalle de confiance sont ici les solutions p_1 et p_2 telles que $|p - F_n| \le 1,645.\sqrt{\dfrac{p.(1-p)}{n}}$. Il en résulte l'inéquation du second degré en p :

$$(p - F_n)^2 \le (1,645)^2.\frac{p.(1-p)}{n}$$

Numériquement, l'inégalité ci-dessus s'écrit après développement :

$$(1+\frac{1,645^2}{500}).p^2-(0,74+\frac{1,645^2}{500}).p+0,37^2\le 0$$

soit, $1,0054.p^2-0,7454.p+0,1369\le 0$. Par résolution, il vient immédiatement, l'encadrement $0,338\le p\le 0,402$.

On obtient ainsi des résultats de plus en plus précis, mais dont on notera qu'ils restent cependant très proches les uns des autres (du moins, pour n assez grand).

3.2 Sondages de popularité

Enoncé : Un institut de sondages souhaite estimer avec une précision de 3 points (à droite et à gauche), la probabilité qu'un individu vote pour le Maire actuel à une prochaine élection.

1°) Combien de personnes est-il nécessaire de sonder ?

2°) Sur un échantillon représentatif de 1000 personnes, on étudie l'évolution des avis favorables pour l'élu en question. En novembre, il y avait 43% d'avis favorables et en décembre, la côte baisse à 41%. Un journaliste prend très au sérieux cette chute de popularité du candidat. Peut-on infirmer ou non la position du journaliste quant à une modification de la popularité dudit candidat ?

On traitera les deux questions avec un seuil de confiance $1-\alpha$ égal à 95%.

Solution : 1°) Désignant par p la probabilité inconnue qu'un électeur vote pour le candidat considéré et par F_n la fréquence empirique des intentions de vote pour ce candidat à travers l'échantillon de taille n, l'intervalle de confiance de p au seuil $1-\alpha$. s'écrit :

$$F_n-t_\alpha.\sqrt{\frac{p.(1-p)}{n}}\le p\le F_n+t_\alpha.\sqrt{\frac{p.(1-p)}{n}}\text{, où } t_\alpha \text{ vérifie } \Pr ob(|\xi|\le t_\alpha)=1-\alpha$$

(ξ désignant la variable de loi normale centrée réduite $N(0,1)$).

- La question qui est posée ici est de trouver n tel que la largeur de l'intervalle de confiance, soit $2t_\alpha.\sqrt{\frac{p.(1-p)}{n}}$ est inférieure à 0,06 (un point représente 1%). On ne connaît pas p, mais on peut se satisfaire en première approximation de la majoration $p.(1-p)\le\frac{1}{4}$ pour $0\le p\le 1$.

Ainsi la condition $2t_\alpha.\sqrt{\frac{p.(1-p)}{n}}\le 0,06$ ou encore $4t_\alpha^2.\frac{p.(1-p)}{n}\le 0,06^2$, est-elle satisfaite dès que $4t_\alpha^2.\underset{0\le p\le 1}{Max}\frac{p.(1-p)}{n}\le 0,06^2$, soit $4t_\alpha^2.\frac{1}{4.n}\le 0,06^2\Rightarrow n\ge\frac{t_\alpha^2}{0,06^2}$.

Pour un risque α égal à 5%, on a donc $\Pr ob(|\xi|\le t_\alpha)=0,95\Rightarrow 2.\Pi(t_\alpha)-1=0,95$, $\Pi(t)$ désignant la fonction de répartition de ξ, soit $\Pi(t)=\Pr ob(\xi\le t)$. Par lecture dans la table des valeurs correspondante (cf. annexes) du nombre $t_\alpha/\Pi(t_\alpha)=0,975$, il en résulte $t_\alpha=1,96$. On obtient donc, en conclusion, la condition $n\ge\left[\frac{1,96}{0,06}\right]^2=1067,1$, c'est-à-dire, après arrondi à l'entier immédiatement supérieur, la condition cherchée $n\ge 1068$.

2°) Se référant à l'application 1.5 du chapitre I, et notant par $F_{n,1}$ et $F_{n,2}$ les fréquences empiriques constatées lors des sondages respectivement de novembre et de décembre, la statistique $\dfrac{F_{n,1}-F_{n,2}-(p_1-p_2)}{\sqrt{\dfrac{p_1.(1-p_1)}{n}+\dfrac{p_2.(1-p_2)}{n}}}$ converge vers la loi normale centrée réduite $N(0,1)$ puisque la taille commune des échantillons (égale à n, soit numériquement $n=1000$) est grande.

Il est précisé que p_1 et p_2 désignent ici les proportions inconnues des avis favorables au candidat considéré, respectivement en novembre et en décembre, et qu'il convient donc ici de tester la stabilité ou non de p, c'est à dire l'hypothèse $p_1 = p_2$. Lorsque cette dernière est vérifiée ($p_1 = p_2 = p$), on a en désignant par t_α le nombre vérifiant la relation $\Pr ob(|\xi| \le t_\alpha) = 1-\alpha$, l'encadrement $-t_\alpha \le \dfrac{F_{n,1}-F_{n,2}}{\sqrt{\dfrac{2.p.(1-p)}{n}}} \le t_\alpha$.

Ainsi, dans le cas particulier $1-\alpha = 95\% \Rightarrow t_\alpha = 1,96$, peut-on écrire que dans 95% des cas, au moins, $|F_{n_1}-F_{n_2}| \le 1,96.\sqrt{\dfrac{2.p.(1-p)}{n}}$ (du moins, sous l'hypothèse $p_1 = p_2 = p$).

• On ne connaît pas p, mais on pourra se contenter ici du majorant $\underset{0\le p\le 1}{Max}[p.(1-p)] = \dfrac{1}{4}$.

D'où, la borne $|F_{n_1}-F_{n_2}| \le \dfrac{1,96}{2.\sqrt{n}} = 0,0438$.

• Plus précisément, on peut aussi remplacer p par son estimateur $\hat{p} = \dfrac{n_1.F_{n,1}+n_2.F_{n_2}}{n_1+n_2}$, soit en tenant compte qu'on a ici, $n_1 = n_2 = 1000$, l'estimateur $\hat{p} = \dfrac{F_{n_1}+F_{n_2}}{2}$. Numériquement, on obtient donc la borne $|F_{n_1}-F_{n_2}| \le 1,96.\sqrt{\dfrac{2\times 0,42\times 0,58}{1000}} = 0,0432$.

• Enfin, on peut dans le facteur $\sqrt{\dfrac{p_1.(1-p_1)}{n_1}+\dfrac{p_2.(1-p_2)}{n_2}}$ remplacer p_1 et p_2 par leurs estimateurs F_{n_1} et F_{n_2}, d'où la borne $|F_{n_1}-F_{n_2}| \le 1,96.\sqrt{\dfrac{0,43\times 0,57}{1000}+\dfrac{0,41\times 0,59}{1000}} = 0,0432$.

Bref, tout un ensemble d'approximations, d'autant plus proches, que n est grand.

Pour en revenir à la question posée dans l'énoncé, la différence des fréquences empiriques relevées, soient 43% et 41%, reste comprise dans l'intervalle de confiance de $|F_{n_1}-F_{n_2}|$ lorsque $p_1 = p_2$, puisqu'on a en effet, $|F_{n_1}-F_{n_2}| = 2\% \prec 4,3\%$. On ne peut donc pas accréditer l'hypothèse suivant laquelle la proportion des électeurs favorables au Maire sortant considéré aurait évolué (hypothèse $p_2 \ne p_1$).

• En fait, le raisonnement ci-dessus, fait référence, par anticipation, aux éléments de théorie des tests développés dans le chapitre III suivant. On peut présenter cela autrement.

• Le raisonnement équivalent ci-après est préférable au sein de ce chapitre, car ne débordant pas du cadre de l'estimation par intervalle de confiance. Considérant de nouveau la fonction pivotale $\dfrac{F_{n,1}-F_{n,2}-(p_1-p_2)}{\sqrt{\dfrac{p_1.(1-p_1)}{n}+\dfrac{p_2.(1-p_2)}{n}}}$ (mais cette fois sans supposer $p_1 = p_2$), on a immédiatement l'encadrement :

$$F_{n_1}-F_{n_2}-t_\alpha.\sqrt{\frac{p_1.(1-p_1)}{n_1}+\frac{p_2.(1-p_2)}{n_2}} \le p_1-p_2 \le F_{n_1}-F_{n_2}+t_\alpha.\sqrt{\frac{p_1.(1-p_1)}{n_1}+\frac{p_2.(1-p_2)}{n_2}}$$

Numériquement et compte tenu des approximations possibles de p_1 et p_2 (voir ci-dessus), on a donc l'encadrement $0,02-0,0432 \le p_1-p_2 \le 0,02+0,0432$, soit $0,0232 \le p_1-p_2 \le 0,0632$. La valeur 0 qui correspond à l'hypothèse $p_1 = p_2$ est bien contenue dans cet intervalle. On ne peut donc pas conclure à une évolution à la baisse de la popularité du candidat considéré.

• La grande difficulté réside cependant dans le fait que ce faible risque de 5% à partir duquel on est parti, décrit la probabilité α de conclure à tort qu'il y ait un changement de popularité et non pas la probabilité β de conclure à la stabilité de cette popularité alors qu'en réalité elle se serait dégradée. Or, ces deux risques α et β sont antinomiques (β st d'autant plus grand que α est faible), et c'est là un point majeur de la décision statistique (théorie des tests) développée au chapitre III.

• Enfin, c'est un encadrement bilatéral qui a été considéré ici et qui traite le cas du test de l'hypothèse $p_1 = p_2$ contre l'hypothèse $p_1 \neq p_2$. Un raisonnement unilatéral pour tester $p_1 = p_2$ contre $p_1 \succ p_2$ (chute de la popularité) serait plus approprié ici. Il conduit à $t_\alpha = 1,645$ (au lieu de 1,96), ce qui ne modifie pas la conclusion retenue précédemment.

3.3 Contrôle de fabrication par mesures (loi normale)

Enoncé : Les poids en grammes de 1000 pots de confiture sortis d'une machine à conditionner sont les suivants (les résultats étant fournis par classes de longueur 2 grammes, l'origine de la première des classes étant 2000 et l'extrémité de la demière des classes étant 2022).

Classes	1	2	3	4	5	6	7	8	9	10	11
Effectifs	9	21	58	131	204	213	185	110	50	16	3

1°) Utilisant la méthode graphique de la droite de HENRY, montrer que la loi du poids des pots peut être assimilée à une loi normale dont on estimera graphiquement la moyenne et l'écart- type.

2°) Estimer ponctuellement par les calculs habituels, la moyenne et l'écart- type susmentionnés puis à l'aide d'un intervalle de confiance au seuil $1-\alpha = 95\%$.

3°) En admettant que l'écart- type de la machine à conditionner est invariable dans le temps (égal à celui estimé dans la 2ème question) et que le réglage n'a d'influence que sur la moyenne, quelle valeur doit-on choisir quant audit réglage si l'on veut que la probabilité pour qu'un pot pèse moins de 2000 grammes (seuil d'infraction fixé par la législation en cours) soit inférieur à 10^{-4} ?

4°) La machine ayant ainsi été réglée, on pèse huit pots simultanément en cours de fabrication, pour contrôler le réglage. Dans quels cas, décidera-t-on de modifier ce réglage si on admet un risque ne dépassant pas 1% ?

Solntion : 1°) Pour rappel, la méthode de la droite de HENRY est une technique d'ajustement graphique justifiée par le fait que si X est de loi normale $N(m,\sigma)$, la variable $T=\dfrac{X-m}{\sigma}$ suit la loi normale centrée réduite $N(0,1)$ dont, pour la fonction de répartition $\Pi(t)=\Pr ob(T\le t)$, on dispose d'une table des valeurs annexée. Ainsi, pour tout x, a-t-on $\Pr ob(X\le x)=\Pr ob(T\le \dfrac{x-m}{\sigma}=t)=\Pi(t)$.

Une autre approche est d'associer aux fréquences cumulées observées $F(x)=\Pr ob(X\le x)$ les nombres t définis par $\Pi(t)=F(x)$. Si X suit la loi normale, la relation qui lie t à x doit être affine puisqu'en théorie $t=\dfrac{x-m}{\sigma}$. Ainsi construit-on la méthode connue sous le nom de droite de HENRY et dont le mode opératoire est le suivant :

Etape 1 → Dresser la liste des fréquences cumulées observées pour chaque valeur x_i.

Etape 2 → Déterminer les nombres t_i définis par les relations $\Pi(t_i)=F(x_i)$.

Etape 3 → Tracer dans un repère (x,t), les points $M_i(x_i,t_i)$.

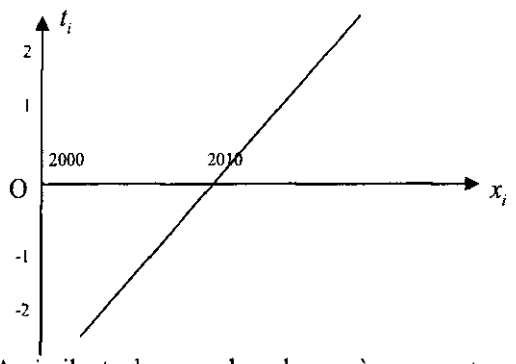
Etape 4 → Si ces points sont sensiblement alignés, la normalité de la distribution de X est vérifiée.

On remarque, en outre, que la droite susmentionnée coupe l'axe des x au point de coordonnées $(m,0)$. En effet, $x=m$ pour $t=0$. L'abscisse de ce point d'intersection fournit donc une estimation graphique de $m=E(X)$. Par ailleurs, la pente de la droite en question est l'inverse de σ, d'où ici encore, une estimation graphique de σ, cette fois.

• L'application de la méthode de la droite de HENRY à la distribution proposée, conduit par lecture dans la table des valeurs de la loi normale centrée réduite, et compte tenu des bornes des classes mentionnées dans l'énoncé (11 classes de 2002 à 2022), aux calculs ci-dessous :

Classe	1	2	3	4	5	6	7	8	9	10	11
x_i (en grammes)	2002	2004	2006	2008	2010	2012	2014	2016	2018	2020	2022
Effectif	9	21	58	131	204	213	185	110	50	16	3
Effectif cumulé	9	30	88	219	423	636	821	931	981	997	1000
Fréquence cumulée $F(x_i)$	0,009	0,030	0,088	0,219	0.423	0,636	0.821	0,931	0,981	0,997	1,000
t_i	-2,36	-1,88	-1,36	-0,78	-0,20	+0,35	+0,92	+1,48	+2,08	+2,75	$+\infty$

La linéarité de la représentation graphique ci-après des t_i en fonction des x_i permet de conclure ici à la validation d'une modélisation par la loi normale.



L'abscisse de l'intersection de cette droite avec l'axe $t_i = 0$ fournit pour m, l'estimation $m \simeq 2010,8$.

Quant à σ qui est l'inverse de la pente de ladite droite, sa valeur est approximativement 3,48.

2°) Assimilant chacune des classes à son centre, le calcul des estimateurs ponctuels $\overline{X} = \dfrac{\sum_{i=1}^{i=n} X_i}{n}$ et $\hat{S}^2 = \dfrac{1}{n-1}.\sum_{i=1}^{i=n}(X_i - \overline{X})^2 = \dfrac{1}{n-1}.\left[\sum_{i=1}^{i=n} X_i^2 - n.\overline{X}^2\right]$ conduit aux résultats numériques rassemblés dans le tableau ci-dessous :

Classe	1	2	3	4	5	6	7	8	9	10	11
Valeurs x_i (*)	2001	2003	2005	2007	2009	2011	2013	2015	2017	2019	2021
Valeurs y_i (*)	1	3	5	7	9	11	13	15	17	19	21
Effectifs f_i	9	21	58	131	204	213	185	110	50	16	3
$f_i.y_i$	9	63	290	917	1836	2343	2405	1650	850	304	63
$f_i.y_i^2$	9	189	1450	6419	16524	25773	31265	24750	14450	5776	1323

(*) Pour ce qui est des valeurs x_i, ce sont les valeurs centrales de chacune des classes qui ont été considérées ici comme habituellement. D'autre part, pour alléger les écritures, il a été fait appel à la variable auxiliaire $Y = X - 2000$ dont espérance et variance sont liées à $E(X)$ et $Var(X)$ par les relations $E(X) = E(Y) + 2000$ et $Var(X) = Var(Y)$.

• Des calculs ci-dessus résultent, relativement à Y, les estimations :

$E(Y) \simeq \overline{y} = \dfrac{\sum_{i=1}^{i=n} f_i.y_i}{\sum_{i=1}^{i=n} f_i} = 10,73$ et $Var(Y) \simeq \dfrac{1}{n-1}.\left[\sum_{i=1}^{i=n} f_i^2.y_i - n.\overline{y}^2\right]$, soit numériquement,

$Var(Y) \simeq \dfrac{1}{999}.\left[127928 - 1000 \times 10,73^2\right] = 12,808$. Il s'ensuit, pour X, les estimations $E(X) \simeq \overline{x} = 2010,73$ et $Var(X) \simeq 12,808 \Rightarrow s_X = 3,58$.

• Quant aux estimations par intervalle de confiance de la moyenne et de la variance de X, elles sont fournies, compte tenu des résultats des rappels de cours des chapitres I et II, selon le mode procédural indiqué comme suit :

→ Pour la moyenne, σ étant inconnu, on considère la fonction pivotale $\dfrac{\overline{X} - m}{\hat{S}/\sqrt{n}}$ de loi de STUDENT à $\nu = n - 1$ degrés de libertés.

Mais n étant grand, cette fonction pivotale converge vers la variable aléatoire ξ de loi normale centrée réduite $N(0,1)$. Ainsi, t_α désignant le nombre vérifiant $\Pr ob(|\xi| \le t_\alpha) = 95\% \Rightarrow t_\alpha = 1,96$, obtient-on en définitive, l'intervalle de confiance :

$$\overline{X} - 1,96.\hat{S}/\sqrt{n} \le m \le \overline{X} + 1,96.\hat{S}/\sqrt{n},$$

soit numériquement, $2010,73 - \frac{1,96 \times 3,58}{\sqrt{1000}} \le m \le 2010,73 + \frac{1,96 \times 3,58}{\sqrt{1000}}$, c'est-à-dire, $2010,51 \le m \le 2010,95$.

→ Pour la variance, m étant inconnue, on considère la fonction pivotale $V = (n-1).\frac{\hat{S}^2}{\sigma^2}$ de loi du chi- deux à $\nu = n-1$ degrés de libertés, loi dont ici encore, la convergence vers la loi normale est satisfaite compte tenu de la valeur très élevée de n et de l'application du *théorème centrale limite.* Plus précisément et se référant au résultat établi au chapitre I (cf. paragraphe 2.4 des rappels de cours), la variable $\chi^2(n)$ converge vers la loi normale de moyenne n et de variance $2n$. On obtient donc ainsi et toujours pour le seuil de confiance $1-\alpha = 95\%$, les encadrements successifs $-1,96 \le \frac{V-999}{\sqrt{1998}} \le +1,96$, soit $999 - 1,96.\sqrt{1998} \le V \le 999 + 1,96.\sqrt{1998} \Rightarrow 911,39 \le V \le 1086,61$.

Mais $V = \frac{999 \times 3,58^2}{\sigma^2}$. Il s'ensuit donc $0,0712 \le \frac{1}{\sigma^2} \le 0,0849$, ce qui entraîne $11,775 \le \sigma^2 \le 14,039$ et à fortiori l'estimation par intervalle de confiance de σ, soit $3,43 \le \sigma \le 3,74$.

• Complétant les développements précédents, on pourra noter ici la qualité relativement correcte de l'estimation sommaire obtenue graphiquement et dont les résultats sont assez proches de ceux émanant de l'emploi des formules habituelles d'estimation.

3°) Supposant $\sigma = 3,58$, valeur trouvée précédemment, on cherche la valeur du réglage à programmer de façon à ce que le poids aléatoire X d'un pot de confiture soit inférieure à 2000 grammes et ceci suivant une probabilité n'excédant pas 10^{-4}, X suivant par ailleurs la loi normale de type $N(m,\sigma)$.

Mathématiquement et introduisant la variable normale centrée réduite ξ de loi $N(0,1)$, on cherche donc $m / \Pr ob(\xi \le \frac{2000-m}{3,58}) \le 10^{-4}$. La lecture dans les tables de valeurs de la fonction de répartition $\Pi(t) = \Pr ob(\xi \le t)$, du nombre t vérifiant $\Pi(t) = 10^{-4}$ conduit à $t = -3,71$. De la condition $\frac{2000-m}{3,58} \le -3,71$, on déduit $m \ge 2013,28$.

4°) Comme le lecteur le constatera à travers le chapitre suivant, la question posée ici relève d'un test de décision statistique de type paramétrique et bilatéral dans lequel on confronte l'hypothèse H_0 qui est celle d'un bon réglage ($m = m_0 = 2013,28$) à celle d'un dérèglement, soit H_1 ($m \ne m_0$). Un encadrement approprié suivant intervalle de confiance permet de résoudre ce problème comme indiqué ci-après.

En effet, sous l'hypothèse H_0 ($m = m_0 = 2013,28$) et supposant que $\sigma = 3,58$ (valeur constante supposée connue et admise), la moyenne $\overline{X}$ des poids observés à travers des échantillons de taille $n = 8$, suit la loi normale de moyenne $E(\overline{X}) = \frac{1}{8}.\sum_{i=1}^{i=8} E(X_i) = m_0$ et de variance $Var(\overline{X}) = \frac{1}{8^2}.\sum_{i=1}^{i=8} Var(X_i) = \frac{Var(X_i)}{8} = \frac{3,58^2}{8} = 1,602 \Rightarrow \sigma(\overline{X}) = 1,2657$.

Ainsi la variable $\xi = \frac{\overline{X} - 2013,28}{1,2657}$ se trouve-t-elle comprise dans 99% des cas dans l'intervalle $-t_\alpha \le \xi \le t_\alpha$, t_α étant solution de l'équation $\Pr ob(|\xi| \le t_\alpha) = 0,99$. Explicitant $\Pr ob(|\xi| \le t_\alpha) = 2.\Pi(t_\alpha) - 1$, avec $\Pi(t) = \Pr ob(\xi \le t)$, il vient $\Pi(t_\alpha) = 0,995$, soit $t_\alpha = 2,57$.

• On a donc dans 99% des cas et lorsque la machine est bien réglée, l'encadrement $2013,28 - 2,57 \times 1,2657 \le \overline{X} \le 2013,28 + 2,57 \times 1,2657$, soit $2010,03 \le \overline{X} \le 2016,53$. Sous cette dernière hypothèse, la règle de décision qui consiste à laisser le réglage inchangé si $\overline{X} \in [2010,03 - 2016,53]$ et à modifier ce dernier dans le cas contraire $\overline{X} \notin [2010,03 - 2016,53]$ conduit bien à un risque d'erreur au plus égal à 1% (à savoir la probabilité conditionnelle Prob(modifier le réglage/réglage correct) dont l'expression est $\Pr ob(\overline{X} \notin [2010,03 - 2016,53] / m = 2013,28)$. C'est la règle cherchée.

• A noter cependant qu'on omet dans ce raisonnement l'autre risque qui est de ne pas modifier le réglage d'une machine qui en réalité serait déréglée. Ce risque, caractérisé par la probabilité conditionnelle Prob(ne pas modifier le réglage/réglage incorrect) est antinomique du précédent et c'est bien là toute la problématique de la théorie statistique développée dans le chapitre III.

3.4 Intervalles de confiance d'une moyenne dans le cas d'nn modèle de POISSON

En dehors du modèle gaussien et des intervalles de confiance asymptotiques justifiés par le théorème central limite, le recours à des méthodes exactes s'impose pour les petits échantillons, telle celle présentée ci-dessous dans le cadre d'un modèle de type POISSON.

Enoncé : Dans un secteur géographique donné, on recense le nombre journalier d'accidents de la route avec corporel, variable aléatoire X dont on admettra qu'elle suit la loi de POISSON de paramètre θ.

PARTIE I

On dispose dans cette partie, des données suivantes établies sur une semaine :

Jour	1	2	3	4	5	6	7
Nombre d'accidents	0	2	1	0	3	5	3

La taille réduite de l'échantillon conduit à se reporter sur une *méthode exacte* à partir de la loi de la variable Z égale au nombre total d'accidents à l'issue des sept jours. Pour cela, on cherche l'ensemble des lois possibles qui, au seuil de confiance donné $1-\alpha$, contiennent l'observation $Z = z_0$, ce qui entraîne pour θ, un intervalle de confiance $[\theta_{min}, \theta_{max}]$ au niveau $1-\alpha$.

Concrètement, on mène ici, à l'envers, le test bilatéral $H_0 : \theta_{min} \leq \theta \leq \theta_{max}$ contre l'hypothèse $H_1 : \theta \prec \theta_{min}$ ou $\theta \succ \theta_{max}$ (se reporter au chapitre III su sujet de ces tests). Bref, à partir des quantiles d'ordre $\alpha/2$ de Z, il est proposé de déterminer les bornes θ_{min} et θ_{max} telles que $\Pr ob(Z \prec z_0) \leq \frac{\alpha}{2}$ et $\Pr ob(Z \succ z_0) \leq \frac{\alpha}{2}$ (relations « E »).

I-1°) Caractériser la loi de Z et écrivant les relations ci-dessus, en déduire les équations qui, à partir de la fonction de répartition $F(z) = \Pr ob(Z \leq z)$ permettent de déterminer l'encadrement cherché de θ au seuil de confiance $1-\alpha$.

I-2°) Expliciter les résultats numériques obtenus dans le cadre des données proposées ici.

PARTIE II

On dispose dans cette partie, d'informations plus complètes qui portent sur 100 jours d'observations, la fréquence f_i du nombre d'accidents journaliers observés durant cette période étant décrite ci-dessous :

Nombre d'accidents x_i	0	1	2	3	4	5
Fréquence observée f_i	10	24	34	23	6	3

II-1°) Utilisant le théorème central limite, en déduire un encadrement par intervalle de confiance au seuil $1-\alpha = 95\%$, de la somme $Z = \sum_{i=1}^{i=100} X_i$ du nombre total d'accidents sur la période considérée.

II-2°) Appliquer cette méthode asymptotique au cas du petit échantillon de la partie I. Qu'en conclure ?

Solution : I-1°) La loi de POISSON de paramètre θ est caractérisée par l'équation $\Pr ob(Z = z) = \frac{e^{-\theta}.\theta^z}{z!}$ pour $z \in N$. On remarque tout d'abord que $\Pr ob(Z \prec z_0) = \sum_{u=0}^{u=z_0-1} \frac{e^{-\theta}.\theta^u}{u!}$ décroît lorsque θ augmente comme on peut aisément le vérifier à travers les tables de valeurs (cf. annexes). C'est pourquoi la condition $\Pr ob(Z \prec z_0) \leq \frac{\alpha}{2}$ est vérifiée pour toutes les valeurs de $\theta \geq \theta_{max}$, ce dernier seuil étant caractérisé par la relation $\Pr ob(Z \prec z_0 / \theta = \theta_{max}) = \frac{\alpha}{2}$.

De même, la relation $\Pr ob(Z \succ z_0) \leq \frac{\alpha}{2}$ qui équivaut à $\Pr ob(Z \leq z_0) \geq 1 - \frac{\alpha}{2}$ est-elle satisfaite pour toute les valeurs de $\theta \leq \theta_{min}$, où θ_{min} est caractérisé par la relation $\Pr ob(Z \leq z_0) = 1 - \frac{\alpha}{2}$. En résumé, les relations (E) sont satisfaites dès que sont vérifiées les équations $\Pr ob(Z \prec z_0 / \theta = \theta_{max}) = \frac{\alpha}{2}$ et $\Pr ob(Z \leq z_0 / \theta = \theta_{min}) = 1 - \frac{\alpha}{2}$.

I-2°) La variable $Z=\sum_{i=1}^{i=7} X_i$, où X_i désigne le nombre d'accidents constatés au $i^{ème}$ jour, suit la loi de POISSON de paramètre $7.\theta$ (puisqu'on sait que la somme de variables aléatoires de POISSON indépendantes reste elle-même une variable de POISSON).

Les équations obtenues dans la 1ère question, permettent donc de déterminer les bornes $7.\theta_{\min}$ et $7.\theta_{\max}$ suivant les relations :

$$\sum_{u=0}^{u=z_0-1} \frac{e^{-7\theta_{\max}}.(7.\theta_{\max})^u}{u!}=\frac{\alpha}{2} \text{ et } \sum_{u=0}^{u=z_0} \frac{e^{-7\theta_{\min}}.(7.\theta_{\min})^u}{u!}=1-\frac{\alpha}{2}.$$

Tables, abaques, ou méthodes par approximations successives développés sur calculateur sont nécessaires ici pour trouver une solution numérique. Utilisant EXCEL, l'application de ces formules lorsque $z_0=14$ et $1-\alpha=95\% \Rightarrow \alpha=5\%$, conduit aisément aux bornes $7.\theta_{\max}=22,23$ et $7.\theta_{\min}=8,39$, ce qui entraîne pour θ, l'intervalle de confiance $1,20 \le \theta \le 3,17$.

II-1°) Le nombre total d'accidents observés à travers le tableau proposé est immédiatement $z_0=\sum_{i=1}^{i=8} f_i.x_i=200$. D'autre part, désignant par X_i le nombre d'accidents relevé pour le $i^{ème}$ jour $(1 \le i \le 100)$, la variable $Z=\sum_{i=1}^{i=100} X_i$ peut être considérée comme convergent vers la loi normale de moyenne $E(Z)=\sum_{i=1}^{i=100} E(X_i)=100.\theta$ et de variance $Var(Z)=\sum_{i=1}^{i=100} Var(X_i)=100.\theta$. En effet, l'hypothèse d'un grand échantillon $(n \ge 100)$, autorise l'usage du théorème central limite.

- Ainsi $\frac{Z-100.\theta}{\sqrt{100.\theta}}$ converge-t-elle vers la variable normale centrée réduite, soit ξ de loi $N(0,1)$. Centrant l'intervalle de confiance $[\theta_{\min},\theta_{\max}]$ sur l'information possédée relativement à Z, soit z_0 (avec $z_0=200$), il s'agit en définitive de trouver ε tel que :

$$\Pr ob(|\xi|=\left|\frac{z_0-100.\theta}{\sqrt{100.\theta}}\right| \le \varepsilon) \ge 1-\alpha$$

Pour $1-\alpha=95\%$, la lecture dans les tables de valeurs de la fonction de répartition $\Pi(t)=\Pr ob(\xi \le t)$ du nombre t_α vérifiant $\Pr ob(|\xi| \le t_\alpha)=2.\Pi(t_\alpha)-1=0,95$, conduit à $\Pi(t_\alpha)=0,975 \Rightarrow t_\alpha=1,96$.

- On a donc au seuil 95%, l'encadrement $|z_0-100.\theta| \le 1,96.\sqrt{100.\theta}$, ce qui par élévation au carré conduit à l'inéquation en θ, $100^2.\theta^2-200.\theta.(z_0+\frac{1,96^2}{2})+z_0^2 \le 0$. Il s'ensuit par résolution, la solution $\theta_{\min} \le \theta \le \theta_{\max}$, avec :

$$\theta_{\min}=\frac{1}{100}.\left[z_0+\frac{1,96^2}{2}-1,96.\sqrt{z_0+\frac{1,96^2}{4}}\right] \text{ et } \theta_{\max}=\frac{1}{100}.\left[z_0+\frac{1,96^2}{2}+1,96.\sqrt{z_0+\frac{1,96^2}{4}}\right]$$

Numériquement, on obtient ainsi l'intervalle de confiance $1,74 \le \theta \le 2,30$, intervalle qui, fort logiquement, est de qualité bien meilleure que lors de la partie I, l'échantillon étant tout autrement plus important et significatif.

II-2°) Revenant sur la précision de l'estimation susmentionnée, l'application de la méthode asymptotique aux données de la partie I, conduit au résultat $1,19 \le \theta \le 3,36$. On remarque que, même pour une valeur très faible de la taille de l'échantillon (en l'occurrence $n = 7$), la méthode asymptotique fournit finalement un résultat qui n'est pas très éloigné de celui obtenu par méthode exacte, ce qui en confirme tout l'intérêt.

3.5 Une méthode par simulation, le bootstrap

Développée par EFRON (1979), cette méthode consiste à régénérer, à partir d'un échantillon de données proveuant d'une distribution aléatoire donnée, un ensemble d'échantllons de même loi (principe de rééchantillounage), offrant ainsi un mode de traitement par simulatiou, des problèmes usuels de statistique inférentielle (estimateurs, biais et risques, intervalles de confiance...). Exigeant souvent un nombre élevé de simulations, cette méthode trouve notamment tout son intérêt lorsque la distribution des données n'est pas de loi connue ou lorsque l'application des méthodes statistiques usuelles donne lien à des calcnls inextricables.

Enoncé : On considère un n- échantillon $(X_1, X_2, ..., X_n)$ de variables aléatoires indépendantes de loi parente X de distribution inconnue, soit $F(x) = \Pr ob(X \le x)$ (fonction de répartition).

PARTIE I

Dans cette partie, on se propose d'approcher la distribution d'échantillonnage de X et d'estimer sa moyenne.

I-1°) Approchant $F(x)$ par la fonction de répartition empirique $\widehat{F_n}(x) = \sum_{j=1}^{j=n} 1_{X_j}(n)$, montrer que tout échantillon $(Y_1, Y_2, ..., Y_n)$ constitué des variables Y_i obtenues par tirage avec replacement au sein du n- échantillon $(X_1, X_2, ..., X_n)$, forme un nouvel échantillon dont la distribution empirique est $\widehat{F_n}$.

I-2°) En déduire un mode de construction de B échantillons de taille n et de distribution empirique $\widehat{F_n}(x)$, soient $(X_1^k, X_2^k, ..., X_n^k), 1 \le k \le B$, échantillons dits « bootstrap ».

I-3°) Ecrire les formules qui, à partir des échantillons ci-dessus, fournissent une estimation $\widehat{\theta^*}$ de $E(X)$ et de sa variance $Var(\widehat{\theta^*})$.

I-4°) On souhaite appliquer les résultats ci-dessus à un échantillon de dix poids d'adultes, données qui, en réalité, sont des valeurs simulées d'une loi normale, plus précisément la loi normale de moyenne 70kg et d'écart- type 10kg.

Poids (x) en kg	71	76	70	71	72	57	54	73	69	83

Bien évidemment, l'expérimentateur ne connaît pas ces éléments quant à la distribution théorique des poids, car dans le cas contraire, il n'y aurait pas de nécessité d'estimation.

a) Sur la base de l'hypothèse de normalité du caractère étudié, estimer ponctuellement $E(X)$ et $Var(X)$, puis expliciter un intervalle de confiance de $E(X)$ au seuil $1 - \alpha = 90\%$.

b) Ignorant totalement toute hypothèse sur la nature de la distribution de X, utiliser des échantillons « bootstrap » (on supposera $B = 40$) pour obtenir une estimation $\widehat{\theta^*}$ de $E(X)$ et de sa variance $Var(\widehat{\theta^*})$.

c) Comparer les résultats obtenus précédemment quant aux estimations de $E(X)$ et de $Var(X)$.

PARTIE II

Dans cette partie on se propose de construire, suivant la méthode la plus élémentaire qui est celle des « percentiles », un intervalle de confiance de $E(X)$ au seuil $1-\alpha$, l'algorithme présenté étant à mettre en œuvre dans le cadre de l'exemple numérique précédent, avec $1-\alpha = 90\%$ et toujours avec $B = 40$.

II-1°) On rappelle que le percentile α d'une distribution empirique donnée, soit t_α, sépare cette distribution dans les proportions respectives α % et $(1-\alpha)$%. Dans cette méthode des « percentiles simples », c'est directement sur les estimateurs $\widehat{\theta_k^*}$ obtenus pour chacun des échantillons bootstrap $(1 \le k \le B)$, qu'on applique un encadrement par les percentiles $\alpha/2$ et $1-\alpha/2$, soient respectivement $\hat{\theta}_{\alpha/2}$ et $\hat{\theta}_{1-\alpha/2}$, l'intervalle de confiance cherché s'écrivant $\hat{\theta}_{\alpha/2} \le \theta \le \hat{\theta}_{1-\alpha/2}$.

Appliquer la méthode en question au cas numérique considéré ici.

II-2°) Comparer les résultats obtenus avec l'estimation par intervalle de confiance obtenue dans la question I-4-a) suivant l'hypothèse de normalité de l'échantillon et une méthode analytique.

Solution : I-1°) Le tirage des Y_i s'effectuant avec remise au sein des valeurs de l'ensemble $\{X_1, X_2, ..., X_n\}$, ces derniers sont indépendants et de même loi. Ainsi :

$$\Pr ob(Y_i = X_j / (X_1, X_2, ..., X_n)) = 1/n, \forall (i, j)$$

D'autre part, considérant l'événement « $Y_1 \le x / (X_1, X_2, ..., X_n)$ », et les causes incompatibles et d'union certaine « $Y_i = X_j / (X_1, X_2, ..., X_n)$ », $1 \le j \le n$, la formule des probabilités totales permet d'écrire :

$$\Pr ob(Y_1 \le x / (X_1, X_2, ..., X_n) = \sum_{j=1}^{j=n} \Pr ob(Y_i = X_j / (X_1, X_2, ..., X_n). \Pr ob(X_j \le x)$$

soit, dans le présent contexte « empirique », la relation :

$$\Pr ob(Y_1 \le x / (X_1, X_2, ..., X_n)) = \frac{1}{n} . \sum_{j=1}^{j=n} 1_{X_j \le x} = \widehat{F_n}(x)$$

I-2°) Bien qu'il n'y ait pas de raison d'espérer des informations supplémentaires d'un rééchantillonnage qui n'utilise pas d'autres sources que l'échantillon initial, le résultat obtenu en I-1°) permet de développer des échantillons de même distribution que cet échantillon initial et de recourir à l'inférence statistique. Il suffit ainsi de construire chaque échantillon dit « bootstrap », soit $(X_1^k, X_2^k, ..., X_n^k), 1 \le k \le B$, par tirage au sort avec remise au sein de l'échantillon initial $(X_1, X_2, ..., X_n)$.

I-3°) Utilisant les formules habituelles d'estimation ponctuelle, on pourra donc :

→ A partir de chaque échantillon « bootstrap », $(X_1^k, X_2^k, ..., X_n^k)$, calculer la moyenne $\widehat{\theta_k^*} = \dfrac{\sum_{i=1}^{i=n} X_i^k}{n}$;

→ puis à partir des $\widehat{\theta_k^*}$, calculer la moyenne $\widehat{\theta^*} = \dfrac{1}{B}.\sum_{k=1}^{k=B} \widehat{\theta_k^*}$;

→ et enfin, calculer la variance de cette moyenne $\widehat{\theta^*}$ à l'aide de la formule usuelle, $Var(\widehat{\theta^*}) = \dfrac{1}{(B-1)}.\sum_{k=1}^{k=B} (\widehat{\theta_k^*} - \widehat{\theta^*})^2$.

I-4-a) L'utilisation de la statistique $\overline{X}$ et des formules usuelles rappelées dans le chapitre I (rappels de cours, paragraphe 2.2), quant aux estimateurs de $E(X)$ et de $Var(X)$, conduit aux résultats ci-dessous, pour les données proposées :

→ $E(X) \approx \overline{X} = \dfrac{\sum_{i=1}^{i=n} X_i}{n}$, soit numériquement $\overline{X} = 69,6$;

→ $Var(X) \approx s_X^2 = \dfrac{1}{n-1}.\left[\sum_{i=1}^{i=n} X_i^2 - n.\overline{X}^2\right]$, soit numériquement $s_X^2 = 71$.

Supposant vérifiée la normalité de la distribution en question, l'application des résultats mentionnés en rappels de cours du présent chapitre (paragraphe 4.2) quant à l'intervalle de confiance de la moyenne pour un petit échantillon, conduit à utiliser la fonction pivotale $T = \dfrac{\overline{X} - m}{s_X / \sqrt{n}}$ de loi de STUDENT à $\nu = n-1$ degrés de libertés.

Désignant par t_α le nombre vérifiant $\Pr ob(|T| \le t_\alpha) = 1-\alpha$, on obtient donc l'encadrement $-t_\alpha \le T \le t_\alpha$, soit $\overline{X} - t_\alpha.\dfrac{s_X}{\sqrt{n}} \le m \le \overline{X} + t_\alpha.\dfrac{s_X}{\sqrt{n}}$. Numériquement, $1-\alpha = 90\%$ et $\nu = 10-1 = 9$ entraînent $t_\alpha = 1,833$ (cf. tables de valeurs en annexe). D'où l'intervalle de confiance cherché de $m = E(X)$, soit après calculs l'encadrement :

$$64,72 \le m \le 74,48.$$

I-4-b) Reprenant chacune des observations de l'échantillon initial repérées par leur rang et utilisant un simulateur de nombres aléatoires uniformes u sur $[0,1]$ pour choisir dans l'échantillon initial la variable de rang $n = E[u]+1$ ($E[.]$ désignant la partie entière), la répétition de cette procédure conduit en définitive aux échantillons « bootstrap », $(X_1^k, X_2^k, ..., X_n^k), 1 \le k \le 40$, dont la liste est partiellement présentée ci-après, l'essentiel demeurant pour la suite des calculs, le B - uplet formé des estimateurs $\widehat{\theta_k^*}$ générés par chacun des échantillons.

Ces $\widehat{\theta_k^*}$ sont les moyennes empiriques des valeurs de chacun des échantillons « bootstrap » puisqu'il s'agit ici d'estimer une moyenne.

Rang	1	2	3	4	5	6	7	8	9	10
Echantillon de base	71	76	70	71	72	57	54	73	69	83
Echantillons Bootstrap										
1	57	73	72	70	72	54	83	71	76	69
2	70	70	69	71	70	70	73	76	83	83
...										
40	69	76	70	54	72	83	76	71	83	57

Moyenne empirique observée pour l'échantillon de base : $\hat{\theta} = 69,6$

Moyennes empiriques observées pour les échantillons « bootstrap » : $\widehat{\theta_k^*}$, $1 \le k \le 40$

69,7	73,5	68	72	71,1	66,6	65,9	67,7	71	72,4
66,7	74	70,7	69	67,6	66,4	71,3	72,1	74,3	70,1
59,6	68,7	67,3	72,4	65,3	66,5	70,7	73,3	72	69,2
71,9	74,7	70,5	70,6	70	69,1	67,4	73,8	69,2	71,1

• L'application des formules $\widehat{\theta^*} = \frac{1}{B}.\sum_{k=1}^{k=B} \widehat{\theta_k^*}$ et $Var(\widehat{\theta^*}) = \frac{1}{(B-1)}.\sum_{k=1}^{k=B} (\widehat{\theta_k^*} - \widehat{\theta^*})^2$ conduit, à partir de la série en question, aux résultats numériques $\widehat{\theta^*} = 69,83$ et $Var(\widehat{\theta^*}) = 9,13$.

I-4-c) Les résultats obtenus par la méthode « bootstrap » sont proches des estimations fournies par les calculs classiques et même légèrement meilleurs puisque les vraies valeurs (inconnues) de $E(X)$ et de $\sigma(X)$ sont respectivement 70 et 10.

Il faut remarquer à ce sujet que la variance de la moyenne est égale à σ^2/n où $\sigma^2 = Var(X)$. Aussi l'estimation de σ^2 que peut fournir $Var(\widehat{\theta^*})$ est-elle égale à $n.Var(\widehat{\theta^*}) = 91,3$ (et non 9,13 !). En résumé, le tableau comparatif des résultats qui viennent d'être obtenus est le suivant :

	Estimateur de $E(X)$	Estimateur de $\sigma(X)$
Valeur théorique	70	10
Estimation par le calcul	69,6	8,43
Estimation « bootstrap »	69,83	9,55

II-1°) Le classement par ordre croissant des $\widehat{\theta_k^*}$ et l'identification des 5ème et 95ème percentiles, conduit immédiatement, à partir de la série des valeurs $\widehat{\theta_k^*}$ en question, aux bornes « percentiles », $\widehat{\theta_{0,05}} = 66,5$ et $\widehat{\theta_{0,95}} = 73,5$.

On obtient donc, au seuil $1-\alpha = 90\%$, l'intervalle de confiance de $E(X)$ suivant :

$$66,5 \le \theta \le 73,5$$

II-2°) Ici encore, le résultat obtenu est proche, voire meilleur, que celui obtenu par le calcul et dont il est rappelé qu'il conduit à l'intervalle de confiance $64,70 \le \theta \le 74,50$.

La méthode « bootstrap » simple utilisée ici reste cependant restreinte, dans son champ d'utilisation, aux distributions symétriques et proches d'une loi normale comme pour le cas présent.

De façon plus générale, on devra recourir à d'autres procédés « bootstrap », basés toujours sur le principe d'un rééchantillonnage, et dont on peut citer, pour le plus courant, la méthode du « bootstrap-t ».

• Cette dernière vise, à partir de l'encadrement $\hat{\theta} - t_\alpha.\sigma(\hat{\theta}) \le \theta \le \hat{\theta} + t_\alpha.\sigma(\hat{\theta})$ où $T = \dfrac{\hat{\theta} - \theta}{\sigma(\hat{\theta})}$ suit la loi de STUDENT, à associer à chaque échantillon « bootstrap » $k, (1 \le k \le B)$, la quantité de STUDENT, $t_k^* = \dfrac{\widehat{\theta_k^*} - \widehat{\theta^*}}{\sigma(\widehat{\theta_k^*})}$. La détermination des percentiles $\dfrac{\alpha}{2}$ et $1 - \dfrac{\alpha}{2}$ par rapport à cette série des valeurs $t_k^*, (1 \le k \le B)$, soient $t_{\alpha/2}$ et $t_{1-\alpha/2}$, permet de conclure à l'intervalle de confiance cherché, à savoir $\widehat{\theta^*} - t_\alpha.\sigma(\widehat{\theta^*}) \le \theta \le \widehat{\theta^*} + t_\alpha.\sigma(\widehat{\theta^*})$.

Néanmoins, le calcul des écarts-type $\sigma(\widehat{\theta_k^*})$ exige, pour chaque échantillon « bootstrap », $(X_1^k, X_2^k, ..., X_n^k)$, un rééchantillonnage qui augmente considérablement la lourdeur des calculs et le nombre des itérations.

• En résumé, bien que séduisante dans le cas d'école d'une distribution normale, les techniques « bootstrap » exigent rapidement un recours à l'informatique et sont donc à réserver aux cas où on ne sait pas expliciter la distribution considérée, les méthodes de calcul classiques restant à privilégier dans le cas contraire.

C - Exercices complémentaires

1. On considère une variable aléatoire discrète X dont la loi de probabilité est définie par $\Pr ob(X = k) = \dfrac{\theta^{k-1}}{(1+\theta)^k}$, pour $k \in \{1, 2, 3, ...\}$, θ désignant un paramètre inconnu et strictement positif.

1°) Montrer que $E(X) = \theta + 1$ et $Var(X) = \theta.(1+\theta)$.

2°) On se propose d'estimer le paramètre θ à partir d'une réalisation numérique $(x_1, x_2, ..., x_n)$ d'un n- échantillon $(X_1, X_2, ..., X_n)$ de la loi de X.

a) Trouver l'estimateur du maximum de vraisemblance $\widehat{\theta_n}$ du paramètre θ.

b) Cet estimateur $\widehat{\theta_n}$ est-il sans biais ?

c) Cet estimateur $\widehat{\theta_n}$ est-il convergent ?

Solution : 1°) Se référant aux lois rappelées dans le chapitre I (cf. rappels de cours, paragraphe 1.1), la loi géométrique est définie par $\Pr ob(X = k) = q^{k-1}.p, (k \in N^*, q = 1 - p)$.

Il s'agit d'une loi d'espérance mathématique $E(X)=\frac{1}{p}$ et de variance $Var(X)=\frac{q}{p^2}$.

• Or, la définition proposée ici à travers l'équation $\Pr ob(X=k)=\frac{\theta^{k-1}}{(1+\theta)^k}$ correspond précisément à la loi géométrique susmentionnée où $q=\frac{\theta}{1+\theta}$ et $p=1-\frac{\theta}{1+\theta}=\frac{1}{1+\theta}$. On peut donc en déduire sans calcul, $E(X)=\frac{1}{p}$, soit $E(X)=1+\theta$, et $Var(X)=\frac{q}{p^2}$, soit $Var(X)=\frac{\theta}{1+\theta}.(1+\theta)^2=\theta.(1+\theta)$.

• Optant pour un calcul direct, on aurait $E(X)=\sum_{k=1}^{k=+\infty}k.\frac{\theta^{k-1}}{(1+\theta)^k}=\frac{1}{1+\theta}.\sum_{k=1}^{k=+\infty}k.\left[\frac{\theta}{1+\theta}\right]^{k-1}$. Mais, d'une part $\frac{d}{d\theta}\sum_{k=0}^{+\infty}\left[\frac{\theta}{1+\theta}\right]^k=\sum_{k=1}^{k=+\infty}k.\frac{d}{d\theta}(\frac{\theta}{1+\theta}).\left[\frac{\theta}{1+\theta}\right]^{k-1}=\frac{1}{(1+\theta)^2}.\sum_{k=1}^{k=+\infty}k.\left[\frac{\theta}{1+\theta}\right]^{k-1}$, et d'autre part, $\frac{d}{d\theta}\sum_{k=0}^{k=+\infty}\left[\frac{\theta}{1+\theta}\right]^k=\frac{d}{d\theta}\left[\frac{1}{1-\frac{\theta}{1+\theta}}\right]=\frac{d}{d\theta}(1+\theta)=1$. En conclusion, on retrouve bien le résultat $E(X)=\frac{1}{1+\theta}.\sum_{k=1}^{k=+\infty}k.\frac{\theta^{k-1}}{(1+\theta)^k}=1+\theta$.

• De même, pour la variance, on a $E(X^2)=\sum_{k=1}^{k=+\infty}k.(k-1)\frac{\theta^{k-1}}{(1+\theta)^k}+\sum_{k=1}^{k=+\infty}k.\frac{\theta^{k-1}}{(1+\theta)^k}$, soit $E(X^2)=\frac{\theta}{(1+\theta)^2}.\sum_{k=1}^{k=+\infty}k.(k-1).\left[\frac{\theta}{1+\theta}\right]^{k-2}+E(X)$. Or d'une part, $\frac{d}{d\theta}\sum_{k=1}^{k=+\infty}k.\left[\frac{\theta}{1+\theta}\right]^{k-1}$ est égal à $\sum_{k=2}^{k=+\infty}k.(k-1).\frac{d}{d\theta}(\frac{\theta}{1+\theta}).\left[\frac{\theta}{1+\theta}\right]^{k-2}$, soit $\frac{1}{(1+\theta)^2}.\sum_{k=2}^{k=+\infty}k.(k-1).\left[\frac{\theta}{1+\theta}\right]^{k-2}$ et d'autre part, on a $\frac{d}{d\theta}\sum_{k=1}^{k=+\infty}k.\left[\frac{\theta}{1+\theta}\right]^{k-1}=\frac{d}{d\theta}(1+\theta)^2=2.(1+\theta)$.

Finalement $E(X^2)=2.(1+\theta)^3.\frac{\theta}{(1+\theta)^2}+1+\theta$, ce qni entraîne, pour la variance, le résultat déjà connu $Var(X)=\frac{2.(1+\theta)^3.\theta}{(1+\theta)^2}+1+\theta-(1+\theta)^2=\theta.(1+\theta)$.

2-a) La fonction de vraisemblance $L(X_1,X_2,...,X_n,\theta)$ associée à l'échantillon de taille n, $(X_1,X_2,...,X_n)$ est égale au produit des lois, c'est-à-dire :

$$L(X_1,X_2,...,X_n,\theta)=\prod_{i=1}^{i=n}\frac{\theta^{X_i}}{\theta.(1+\theta)^{X_i}}=\theta^{-n}.\left[\frac{\theta}{1+\theta}\right]^{\sum_{i=1}^{i=n}X_i}$$

Considérant la log- vraisemblance $\ln L=\ln(\frac{\theta}{1+\theta}).\sum_{i=1}^{i=n}X_i-n.\ln\theta$, l'estimateur du maximum de vraisemblance cherché, soit $\widehat{\theta}_n$, est solution de l'équation $\frac{\partial}{\partial\theta}\ln L=0$.

De $\frac{\partial}{\partial\theta}\ln L = 0$, soit $\frac{1}{(1+\theta)^2}.\frac{(1+\theta)}{\theta}.\sum_{i=1}^{i=n} X_i - \frac{n}{\theta} = 0$, il résulte après simplifications, l'estimateur cherché $\widehat{\theta}_n = \frac{\sum_{i=1}^{i=n} X_i}{n} - 1$.

2-b) $\widehat{\theta}_n$ est sans biais car il est immédiat que $E(\widehat{\theta}_n) = E\left[\frac{\sum_{i=1}^{i=n} X_i}{n} - 1\right] = \frac{\sum_{i=1}^{i=n} E(X_i)}{n} - 1$, soit

$E(\widehat{\theta}_n) = \frac{n.(1+\theta)}{n} - 1 = \theta$.

2-c) $Var(\widehat{\theta}_n) = \frac{1}{n^2}.\sum_{i=1}^{i=n} Var(X_i)$ (en effet, il est rappelé que $Var(a.X+b) = a^2.Var(X)$ pour tout $(a,b) \in R^2$). Concrètement et pour le cas présent, $Var(\widehat{\theta}_n) = \frac{n.\theta.(1+\theta)}{n^2} = \frac{\theta.(1+\theta)}{n}$. Il s'agit donc bien d'un estimateur convergent puisque $\lim_{n\to+\infty} Var(\widehat{\theta}_n) = 0$.

• On a même en $\widehat{\theta}_n$ un estimateur efficace dont la variance atteint la borne minimale F.D.C.R caractérisée par $\frac{1}{n.I_X(\theta)}$ où la quantité d'information de FISHER, $I_X(\theta)$, est égale à $-E\left[\frac{\partial^2}{\partial\theta^2}\ln p(X,\theta)\right]$.

En effet, on a par définition $p(X,\theta) = \frac{\theta^{X-1}}{(1+\theta)^X}$, ce qui entraîne en considérant la log- vraisemblance et par dérivations, $\frac{\partial}{\partial\theta}\ln p(X,\theta) = \frac{X-1}{\theta} - \frac{X}{(1+\theta)}$, puis au second ordre, $\frac{\partial^2}{\partial\theta^2}\ln p(X,\theta) = -\frac{(X-1)}{\theta^2} + \frac{X}{(1+\theta)^2} = X.\left[\frac{1}{(1+\theta)^2} - \frac{1}{\theta^2}\right] + \frac{1}{\theta^2}$. On a donc, pour ce qui est de la quantité d'information, $I_X(\theta) = -E\left[\frac{\partial^2}{\partial\theta^2}\ln p(X,\theta)\right] = -\left[\frac{1}{(1+\theta)^2} - \frac{1}{\theta^2}\right].E(X) - \frac{1}{\theta^2}$, soit $I_X(\theta) = \frac{(1+2\theta)}{\theta^2.(1+\theta)^2}.(1+\theta) - \frac{1}{\theta^2} = \frac{1}{\theta.(1+\theta)}$.

La borne F.D.C.R est donc égale ici à $\frac{1}{n.I_X(\theta)} = \frac{\theta.(1+\theta)}{n}$ qui est bien aussi la valeur de $Var(\widehat{\theta}_n)$. Ainsi est donc établi le résultat d'efficacité annoncé précédemment.

2. Dans cet exercice, il est proposé une variante de la méthode de comptage par capture et recapture présentée dans l'application 2.3 du présent chapitre.

Ici encore, on cherche à évaluer le nombre N de poissons contenus dans un étang.

La méthode mise en œuvre consiste à prélever dans l'étang en question, un ensemble de m poissons (m, fixé) que l'on bague et que l'on remet à l'eau. Puis, on prélève au hasard et avec remise des poissons dans l'étang jusqu'à l'obtention de n poissons bagués (n fixé également).

Soit X_n la variable aléatoire égale au nombre de prélèvements ainsi nécessaires.

1°) On considère les variables D_i définies par $D_1 = X_1$, $D_i = X_i - X_{i-1}, (2 \le i \le n)$.

a) Interpréter D_i et en caractériser la loi de probabilité.

b) Exprimer $E(D_i)$ et $Var(D_i)$ et en déduire l'espérance et la variance de la variable aléatoire X_n.

c) On pose $A_n = \frac{m}{n}.X_n$. Montrer que A_n est un estimateur sans biais de N et en déterminer son risque quadratique $R(N) = E\left[(A_n - N)^2\right]$.

2°) On suppose que n est assez grand.

a) Approximer la loi de la moyenne $\overline{X_n} = \frac{X_n}{n}$.

On a marqué 200 poissons puis effectué 450 prélèvements pour obtenir 50 poissons marqués. En déduire un intervalle de confiance de N au seuil $1-\alpha$ =95%.

Solution : 1-a) A compter de la prise du $(i-1)^{ème}$ poisson marqué, D_i représente le nombre de prises nécessaires jusqu'à l'obtention d'un nouveau poisson marqué, la probabilité de tirage d'un tel poisson à chaque prise étant égale à $\frac{m}{N}$, puisqu'il s'agit de prélèvements supposés indépendants.

C'est la définition même de la loi géométrique, caractérisée en l'occurrence par l'équation $\Pr ob(D_i = k) = (1-\frac{m}{N})^{k-1}.\frac{m}{N}$.

1-b) Utilisant les résultats rappelés au précédent chapitre quant à l'espérance et à la variance de la loi géométrique (cf. paragraphe 1.1 des rappels de cours), on a immédiatement $E(D_i) = \frac{N}{m}$ et

$$\text{var}(D_i) = \frac{1 - m/N}{(m/N)^2} = \frac{N.(N-m)}{m^2}.$$

Quant à X_n, c'est la somme $\sum_{i=1}^{i=n} D_i$, c'est-à-dire une variable de loi binomiale négative (cf. rappels de cours 1.1 du chapitre I). La linéarité de l'espérance mathématique entraîne immédiatement $E(X_n) = \sum_{i=1}^{i=n} E(D_i) = \frac{n.N}{m}$. Par ailleurs, l'indépendance des D_i et la pseudo linéarité de la variance entraînent $Var(X_n) = \sum_{i=1}^{i=n} Var(D_i) = \frac{n.N.(N-m)}{m^2}$.

1-c) Posant $A_n = \frac{m}{n}.X_n$, on a $E(A_n) = \frac{m}{n}.E(X_n)$, soit $E(A_n) = \frac{m}{n}.\frac{n}{m}.N = N$. Il s'agit bien d'un estimateur sans biais.

D'autre part, le risque quadratique $R(N)=E\left[(A_n-N)^2\right]$ qui, s'agissant d'un estimateur sans biais est exprimé par $Var(A_n)$, est égal à $\frac{m^2}{n^2}.Var(X_n)$ (car $Var(a.X)=a^2.Var(X)$). Ainsi, a-t-on, $Var(A_n)=\frac{N.(N-m)}{n}$.

2-a) $X_n=\sum_{i=1}^{i=n}D_i$ converge, pour n grand, vers la loi normale puisqu'il s'agit d'une somme de variables aléatoires indépendantes et équidistribuées et que le théorème central- limite s'applique en conséquence. Dans ces conditions, la moyenne $\overline{X_n}=\frac{X_n}{n}$ converge vers la loi normale de moyenne $\frac{1}{n}.E(X_n)=\frac{N}{m}$ et de variance $\frac{1}{n^2}.Var(X_n)=\frac{N.(N-m)}{n.m^2}$.

2-b) $A_n=m.\overline{X_n}$ suit également la loi asymptotique normale de moyenne $m.E(\overline{X_n})=N$ et de variance $m^2.Var(\overline{X_n})=\frac{N.(N-m)}{n}$.

Désignant par t_α le nombre vérifiant $Prob(|\xi|\le t_\alpha)=1-\alpha$, il vient immédiatement, pour $1-\alpha=95\%$, la valeur $t_\alpha=1,96$. Considérant la variable normale centrée réduite $\xi=\frac{A_n-N}{\sqrt{\frac{N.(N-m)}{n}}}$, on a donc au seuil de confiance 95%, l'encadrement :

$$-1,96\le\frac{A_n-N}{\sqrt{\frac{N.(N-m)}{n}}}\le 1,96$$

Les données numériques proposées dans l'énoncé conduisent à $m=200, n=50, A_n=\frac{200}{50}.450=1800$. De l'inégalité $\frac{(1800-N)^2.50}{N.(N-200)}\le(1,96)^2$ résulte par développement l'inéquation $0,9232.N^2-3584,6336.N+3240000\le 0$, d'où l'intervalle de confiance suivant de N :

$$1431\le N\le 2451.$$

A noter que l'estimateur ponctuel de N est égal quant à lui à $A_n=\frac{m}{n}.X_n=\frac{200}{50}\times 450=1800$.

3. On considère une variable aléatoire positive ou nulle et distribuées sur R^+ suivant la loi Gamma (ou encore loi d'ERLANG) de paramètres 3 et θ $(\theta\succ 0)$, c'est-à-dire la loi de densité de probabilité $f(x,\theta)=\frac{x^2.e^{-\frac{x}{\theta}}}{2.\theta^3}, x\in R^+$.

1°) Calculer $E(X)$ et $Var(X)$.

2°) Montrer que l'estimateur du maximum de vraisemblance du paramètre θ est défini par $\hat{\theta}=\frac{\overline{X_n}}{3}$ où $\overline{X_n}$ désigne la moyenne empirique $\frac{\sum_{i=1}^{i=n}X_i}{n}$ des valeurs d'un n échantillon indépendant $(X_1,X_2,...,X_n)$ de variable parente X.

3°) Montrer que $\hat{\theta}$ est sans biais, convergent, et efficace.

4°) Utilisant la loi asymptotique de $\hat{\theta}$ lorsque n est grand, en déduire un intervalle de confiance de $\hat{\theta}$ au seuil $1-\alpha = 95\%$. Quel encadrement obtient-on lorsque $N = 100$ et $\overline{x_n} = 3,3$ (à travers l'observation des valeurs d'un échantillon).

5°) On considère pour θ, l'autre estimateur T_n défini ci-dessous :

$$T_n = \frac{1}{3.(n-1)}.\sum_{i=1}^{i=n} X_i$$

Comparer $\hat{\theta}$ et T_n.

Solution : 1°) Par définition de l'espérance mathématique ($E(X) = \int_R x.f(x).dx$), on a pour le cas présent $E(X) = \int_0^{+\infty} \frac{x^3.e^{-\frac{x}{\theta}}}{2.\theta^3}.dx = \frac{1}{2.\theta^3}.\int_0^{+\infty} \theta^3.u^3.e^{-u}.\theta.du$ (après avoir posé $x = \theta.u$). En bref, $E(X) = \frac{\theta}{2}.\Gamma(4)$ où $\Gamma(n)$ représente l'intégrale d'EULER, $\int_0^{+\infty} t^{n-1}.e^{-t}.dt$ égale à $(n-1)!$.

Concrètement, $E(X) = \frac{\theta}{2}.3! = 3.\theta$. Par ailleurs, $E(X^2) = \frac{1}{2.\theta^3}\int_0^{+\infty} x^4.e^{-\frac{x}{\theta}}.dx$, soit en posant de nouveau $x = \theta.u$, $E(X^2) = \frac{1}{2.\theta^3}\int^{+\infty} \theta^4.u^4.e^{-u}.\theta.du = \frac{\theta^2}{2}.\Gamma(5) = \frac{\theta^2}{2}.4! = 12.\theta^2$. On en déduit, $Var(X) = E(X^2) - E(X)^2 = 12.\theta^2 - (3.\theta)^2 = 3.\theta^2$.

2°) La fonction de vraisemblance $L(X_1, X_2, ..., X_n, \theta)$ associée à un n échantillon $(X_1, X_2, ..., X_n)$ est définie par $L(X_1, X_2, ..., X_n, \theta) = \frac{(X_1.X_2....X_n)^2}{(2.\theta^3)^n}.e^{-\frac{\sum_{i=1}^{i=n} X_i}{\theta}}$. Passant à la log- vraisemblance, $\ln L = 2.\sum_{i=1}^{i=n} \ln(X_i) - \frac{\sum_{i=1}^{i=n} X_i}{\theta} - n.\ln 2 - 3.n.\ln\theta$, l'estimateur E.M.V, soit $\hat{\theta}$, (estimateur du maximum de vraisemblance), est solution de l'équation $\frac{\partial}{\partial\theta}\ln L = 0$, soit :

$$\frac{\sum_{i=1}^{i=n} X_i}{\theta^2} - \frac{3.n}{\theta} = 0 \Rightarrow \hat{\theta} = \frac{\sum_{i=1}^{i=n} X_i}{3.n}$$

C'est bien un maximum, puisque $\frac{\partial^2}{\partial\theta^2}\ln L = \frac{-2}{\theta^3}.\sum_{i=1}^{i=n} X_i + \frac{3.n}{\theta^2}$, soit en $\hat{\theta}$, l'expression $\frac{\partial^2}{\partial\theta^2}\ln L = \frac{1}{\hat{\theta}^3}.\left[-2.\sum_{i=1}^{i=n} X_i + 3.n.\hat{\theta}\right] = -\frac{3.n}{\hat{\theta}^2}$ qui est bien négative.

3°) $E(\widehat{\theta}) = \frac{1}{3.n}.\sum_{i=1}^{i=n} E(X_i) = \frac{3.n.\theta}{3.n} = \theta$. L'estimateur $\hat{\theta}$ est donc sans biais. D'autre part, $Var(\hat{\theta}) = \frac{1}{9.n^2}.\sum_{i=1}^{i=n} Var(X_i) = \frac{3.n.\theta^2}{9.n^2} = \frac{\theta^2}{3.n}$. $\hat{\theta}$ est donc bien convergent puisque $\lim_{n\to+\infty} Var(\widehat{\theta}) = 0$.

Enfin, $\hat{\theta}$ est efficace car le modèle proposé appartient à la famille exponentielle. En effet, $\ln f(x,\theta) = 2.\ln x - \frac{x}{\theta} - 3.\ln\theta - \ln 2$ a pour forme $Q(\theta).T(x) + \beta(\theta) + \gamma(x)$ avec $T(x) = x, Q(\theta) = -\frac{1}{\theta}, \gamma(x) = 2.\ln x$, et $\beta(\theta) = -3.\ln\theta - \ln 2$.

- Suivant un calcul direct, on a successivement, $\ln f(x,\theta) = 2.\ln x - \frac{x}{\theta} - 3.\ln\theta - \ln 2$, puis $\frac{\partial}{\partial\theta}\ln L = \frac{x}{\theta^2} - \frac{3}{\theta}$, et enfin $\frac{\partial^2}{\partial\theta^2}\ln L = -\frac{2.x}{\theta^3} + \frac{3}{\theta^2}$. Il s'ensuit, pour ce qui est de la quantité d'information de FISHER, $I_X(\theta) = -E\left[\frac{\partial^2}{\partial\theta^2}\ln L\right]$, l'expression $I_X(\theta) = \frac{2}{\theta^3}.E(X) - \frac{3}{\theta^2}$, soit $I_X(\theta) = \frac{6.\theta}{\theta^3} - \frac{3}{\theta^2} = \frac{3}{\theta^2}$. La borne F.D.C.R qui correspond à la variance minimale des estimateurs sans biais de θ, est donc égale ici à $\frac{1}{n.I_X(\theta)}$, soit $\frac{\theta^2}{3.n}$. Il s'agit bien en l'occurrence, de la variance de l'estimateur $\hat{\theta} = \frac{\sum_{i=1}^{i=n} X_i}{3.n}$ qui vérifie donc la propriété d'efficacité.

4°) Pour n assez grand, le théorème central- limite entraîne la convergence de $\hat{\theta} = \frac{\sum_{i=1}^{i=n} X_i}{3.n}$ vers la loi normale de moyenne $E(\hat{\theta}) = \theta$ et de variance $Var(\hat{\theta}) = \frac{\theta^2}{3.n}$. Désignant par t_α le nombre vérifiant $Prob(|\xi| \le t_\alpha) = 1 - \alpha$ (ξ étant la variable normale centrée réduite), il s'ensuit relativement à $\hat{\theta}$ et à sa variable centrée réduite associée $\xi = \frac{\hat{\theta} - \theta}{\sqrt{\frac{\theta^2}{3.n}}}$, l'encadrement $-t_\alpha \le \xi \le t_\alpha$ qui s'écrit $-t_\alpha \le \frac{\hat{\theta} - \theta}{\sqrt{\frac{\theta^2}{3.n}}} \le +t_\alpha$. Ainsi, obtient-on pour θ, l'estimation par intervalle de confiance :

$$\hat{\theta} - t_\alpha.\sqrt{\frac{\theta^2}{3.n}} \le \theta \le \hat{\theta} + t_\alpha.\sqrt{\frac{\theta^2}{3.n}}$$

Séparant θ (qui est inconnu) et son estimateur $\hat{\theta}$, il vient successivement $\hat{\theta} - t_\alpha.\theta/\sqrt{3.n} \le \theta \le \hat{\theta} + t_\alpha.\theta/\sqrt{3.n}$, puis $\frac{\hat{\theta}}{\theta} - \frac{t_\alpha}{\sqrt{3.n}} \le 1 \le \frac{\hat{\theta}}{\theta} + \frac{t_\alpha}{\sqrt{2.n}}$, et enfin l'encadrement $1 - \frac{t_\alpha}{\sqrt{3.n}} \le \frac{\hat{\theta}}{\theta} \le 1 + \frac{t_\alpha}{\sqrt{3.n}}$ qui s'écrit aussi $\frac{1}{1 + t_\alpha/\sqrt{3.n}} \le \frac{\theta}{\hat{\theta}} \le \frac{1}{1 - t_\alpha/\sqrt{3.n}}$. En conclusion, l'intervalle de confiance cherché de θ s'écrit $\frac{1}{1 + t_\alpha/\sqrt{3.n}}.\hat{\theta} \le \theta \le \frac{1}{1 - t_\alpha/\sqrt{3.n}}$.

Numériquement, $1-\alpha = 95\% \Rightarrow t_\alpha = 1,96$. D'autre part, $\overline{x_n} = \frac{\sum_{i=1}^{i=n} x_i}{n} = 3,3 \Rightarrow \hat{\theta} = \frac{\overline{x_n}}{3} = 1,1$.

D'où le résultat $0,988 \leq \theta \leq 1,240$.

5°) $E(T_n) = \frac{1}{3.(n-1)} . \sum_{i=1}^{i=n} E(X_i) = \frac{3.n.\theta}{3.(n-1)} = \frac{n}{n-1} . \theta$. Contrairement à $\hat{\theta}$, on voit ici que T_n n'est pas un estimateur sans biais de θ. C'est un estimateur probablement moins performant que $\hat{\theta}$, d'autant que ce dernier est le meilleur de tous les estimateurs sans biais de θ puisque de variance minimale (égale à la borne F.D.C.R).

La comparaison des risques $R(T_n) = E\left[(T_n - \theta)^2\right]$ et $R(\hat{\theta}) = E\left[(\hat{\theta} - \theta)^2\right] = Var(\hat{\theta})$ confirme cette analyse. D'une part, $E\left[(T_n - \theta)^2\right] = E(T_n^2) - 2.\theta.E(T_n) + \theta^2$, soit également, $E\left[(T_n - \theta)^2\right] = Var(T_n) + E(T_n)^2 - 2.\theta.E(T_n) + \theta^2$ dont le développement en fonction des résultats antérieurs s'écrit $E\left[(T_n - \theta)^2\right] = \frac{3.n.\theta^2}{9.(n-1)^2} + \frac{n^2.\theta^2}{(n-1)^2} - \frac{2.\theta^2.n}{n-1} + \theta^2 = \frac{n+3}{3.(n-1)^2} . \theta^2$.

D'autre part, $Var(\hat{\theta}) = \frac{\theta^2}{3.n}$. Dans ces conditions, le rapport des risques $\frac{R(T_n)}{R(\hat{\theta})} = \frac{(n+3).n}{(n-1)^2}$ est manifestement supérieur à 1 puisque $n+3 \succ n-1$ et $n \succ n-1 \Rightarrow n.(n+3) \succ (n-1)^2$. Ceci établit le résultat annoncé.

4. Afin de mieux gérer les demandes de crédits des clients, un directeur d'agence bancaire réalise une étude relative à la durée de traitement des dossiers, supposée suivre une loi normale. Ces données sont résumées ci-dessous :

Durée de traitement (en mn)	0-10	10-20	20-30	30-40	40-50	50-60
Effectif	3	6	10	7	3	1

1°) Calculer la moyenne et l'écart- type des durées de traitement des dossiers de l'échantillon de taille $n = 30$ dont il est question ci-dessus.

2°) En déduire les estimations ponctuelles de la moyenne m et de l'écart- type σ de la population des dossiers.

3°) Construire des estimations par intervalles de confiance de m et de σ, ceci au seuil de confiance $1-\alpha = 90\%$.

Solution : 1°) Utilisant la méthode classique de statistique descriptive qui consiste à assimiler chaque classe à son centre x_i, le calcul de la moyenne et de l'écart- type des durées de traitement relevées dans l'échantillon proposé, conduit au tableau de calculs ci-dessous :

Durée x_i	5	15	25	35	45	55	Σ
Effectif f_i	3	6	10	7	3	1	30
$f_i.x_i$	15	90	250	245	135	55	790
$f_i.x_i^2$	75	1350	6250	8575	6075	3025	25350

On en déduit, pour ce qui est de la moyenne, la valeur $\overline{x} = \frac{\sum_{i=1}^{i=p} f_i.x_i}{\sum_{i=1}^{i=p} f_i}$, soit $\overline{x} = 26,33$

De même, pour ce qui est de la variance, on obtient $\sigma_{ech}^2 = \dfrac{\sum_{i=1}^{i=p} f_i.x_i^2}{\sum_{i=1}^{i=p} f_i} - \bar{x}^2$, soit numériquement, $\sigma_{ech}^2 = 151,55 \Rightarrow \sigma_{ech} = 12,31$.

2°) Des calculs précédents sur l'échantillon de taille $n = 30$, il ressort les estimateurs ponctuels de m et de σ^2, à savoir respectivement, $\bar{x} = 26,33$ et $\hat{S}^2 = \dfrac{n}{n-1}.\sigma_{ech}^2$ (avec $n = \sum_i f_i$), soit numériquement, $\hat{S}^2 = \dfrac{30}{29} \times 151,55 = 156,78 \Rightarrow \hat{S} = 12,52$.

3°) La distribution proposée étant supposée être de loi normale, la fonction pivotale $\dfrac{\hat{X} - m}{\hat{S}/\sqrt{n}}$ suit la loi de STUDENT à $\nu = n-1$ degrés de libertés (cf. rappels de cours du présent chapitre, paragraphe 3.2). La lecture dans la table annexée du nombre t_α vérifiant $\Pr ob(|T| \le t_\alpha) = 1-\alpha$, où T désigne la variable de STUDENT à $n-1$ degrés de liberté, conduit pour $n = 30, 1-\alpha = 90\%$, à la valeur $t_\alpha = 1,699$. Il s'ensuit donc, au seuil $1-\alpha = 90\%$, l'intervalle :

$$-1,699 \le \frac{\bar{X} - m}{\hat{S}/\sqrt{n}} \le 1,699, \text{ soit } \bar{X} - 1,699.\hat{S}/\sqrt{n} \le m \le \bar{X} + 1,699.\hat{S}/\sqrt{n}.$$

Numériquement, l'intervalle obtenu est donc $22,449 \le m \le 30,217$.

• On remarquera que n étant supérieur ou égal à 30, l'approximation de la loi de STUDENT par la loi normale est envisageable. Le recours à la fonction pivotale $\xi = \dfrac{\bar{X} - m}{\sigma/\sqrt{n}}$ où ξ suit la loi normale centrée réduite, conduit pour t_α / $\Pr ob(|\xi| \le t_\alpha) = 1-\alpha$, à la valeur $t_\alpha = 1,645$ lorsque $\alpha = 10\%$ (cf. table annexée des valeurs de la fonction de répartition $\Pi(t) = \Pr ob(\xi \le t)$).

On a donc l'encadrement $-1,645 \le \dfrac{\bar{X} - m}{\sigma/\sqrt{n}} \le 1,645$ qui conduit à l'intervalle de confiance $\bar{X} - 1,645.\sigma/\sqrt{n} \le m \le \bar{X} + 1,645.\sigma/\sqrt{n}$, soit en approchant σ par $\hat{S}$, l'estimation :

$$22,573 \le m \le 30,094.$$

Comme on pourra le constater, les résultats sont très proches, ce qui valide l'approximation préconisée.

• Enfin, pour le cas de la variance, la moyenne m étant inconnue puisque estimée à partir de l'échantillon proposé de taille $n = 30$, la fonction pivotale à considérer est $V = \dfrac{(n-1).\hat{S}^2}{\sigma^2}$ de loi du χ^2 à $n-1$ degrés de libertés. La recherche des nombres a et b tels que $\Pr ob(V \prec a) = \alpha/2$ et $\Pr ob(V \succ b) = \alpha/2$, conduit pour $n = 30$ et $1-\alpha = 90\%$, aux valeurs $a = 17,71$ et $b = 42,56$ (cf. tables des valeurs de la distribution de χ^2 consultables sur site internet).

Il en ressort, au seuil de confiance $1-\alpha = 90\%$, l'encadrement $17,71 \le \frac{(n-1).\hat{S}^2}{\sigma^2} \le 42,56$,

soit $\frac{(n-1).\hat{S}^2}{42,56} \le \sigma^2 \ge \frac{(n-1).\hat{S}^2}{17,71}$, c'est-à-dire, numériquement :

$$106,83 \le \sigma^2 \le 256,73 \Rightarrow 10,34 \le \sigma \le 16,02$$

• La valeur $n = 30$ constituant ici la borne inférieure à partir de laquelle l'approximation de la loi de χ^2 par la loi normale est envisageable, on peut rechercher ici le résultat auquel conduit cette méthode approchée. En ce sens, on préférera utiliser la convergence de $\sqrt{2.\chi_n^2}$, (χ_n^2 variable de chi- deux à n degrés de libertés) vers la loi normale de moyenne $\sqrt{2.n-1}$ et de variance 1, qui est plus rapide que la convergence de χ_n^2 vers la loi normale de moyenne n et d'écart- type $\sqrt{2.n}$.

Ainsi, $\sqrt{2.\chi_{n-1}^2}$ avec $\chi_{n-1}^2 = \frac{(n-1).\hat{S}^2}{\sigma^2}$ converge-t-elle vers la loi normale de moyenne $\sqrt{2.n-3}$ et de variance 1, ce qui, au seuil $1-\alpha = 90\%$, entraîne l'encadrement :

$$-1,645 \le \frac{\sqrt{\frac{2.(n-1).\hat{S}^2}{\sigma^2}} - \sqrt{2.n-3}}{1} \le 1,645.$$

En conclusion et après simplifications, on obtient l'intervalle de confiance :

$$\frac{\hat{S}.\sqrt{2.(n-1)}}{\sqrt{2.n-3}+1,645} \le \sigma \le \frac{\hat{S}.\sqrt{2.n-1}}{\sqrt{2.n-3}-1,645}$$

soit, numériquement, $10,37 \le \sigma \le 16,15$.

Le résultat obtenu est très proche de celui fourni par la méthode exacte.

5. On considère un ensemble $(X_1, X_2, ..., X_n)$ de n variables aléatoires indépendantes équidistribuées de loi normale $N(\mu, \sigma^2)$ et les variables $\overline{X} = \frac{1}{n}.\sum_{i=1}^{i=n} X_i, S_X^2 = \frac{1}{n}.\sum_{i=1}^{i=n}(X_i - \overline{X})^2$, et $\widehat{S_X}^2 = \frac{1}{(n-1)}.\sum_{i=1}^{i=n}(X_i - \overline{X})^2$.

1°) Expliciter les propriétés des statistiques S_X^2 et $\widehat{S_X}^2$ considérées comme des estimateurs du paramètre σ^2.

2°) Entre S_X^2 et $\widehat{S_X}^2$, quel est le meilleur estimateur de σ^2 ?

Solution : 1°) Se référant aux résultats de l'application 1.2 du chapitre I relative aux distributions d'échantillonnage du modèle gaussien, il est rappelé, avec les notations du présent énoncé, que $E(\widehat{S_X}^2) = \sigma^2$ et $Var(\widehat{S_X}^2) = \frac{2.\sigma^4}{n-1}$.

De la relation $S_X^2 = \frac{n-1}{n}.\widehat{S_X}^2$, il en résulte par ailleurs, $E(S_X^2) = \frac{n-1}{n}.E(\widehat{S_X}^2) = \frac{n-1}{n}.\sigma^2$

et $Var(S_X^2) = (\frac{n-1}{n})^2.Var(\widehat{S_X}^2) = \frac{2.(n-1).\sigma^4}{n^2}$.

En résumé, S_X^2 et $\widehat{S_X}^2$ sont convergents puisque $\lim_{n\to+\infty} Var(S_X^2) = \lim_{n\to+\infty} Var(\widehat{S_X}^2) = 0$, seul $\widehat{S_X}^2$ étant sans biais ici (on a en effet, $E(S_X^2) \neq \sigma^2$).

2°) La comparaison des risques quadratiques associés aux estimateurs S_X^2 et $\widehat{S_X}^2$, soient $R(S_X^2)$ et $R(\widehat{S_X}^2)$ conduit à former $R(S_X^2) = E\left[(S_X^2 - \sigma^2)^2\right]$ et $R(\widehat{S_X}^2) = E\left[(\widehat{S_X}^2 - \sigma^2)^2\right] = Var(\hat{S})$.

De la linéarité de l'espérance mathématique, il vient $R(S_X^2) = E(S_X^4) - 2.\sigma^2.E(S_X^2) + \sigma^4$, soit en remarquant que $E(S_X^4) = Var(S_X^2) + E(S_X^2)^2$ et en utilisant les résultats de la 1ère question :

$$R(S_X^2) = \left[\frac{2.(n-1)}{n^2} + \frac{(n-1)^2}{n^2} - \frac{2.(n-1)}{n} + 1\right].\sigma^4 = \frac{2.n-1}{n^2}.\sigma^4$$

La comparaison entre $R(\widehat{S_X}^2) = \frac{2.\sigma^4}{n-1}$ et $R(S_X^2) = \frac{2.n-1}{n^2}.\sigma^4$ conduit à la différence $\left[\frac{2}{n-1} - \frac{2.n-1}{n^2}\right].\sigma^4 = \frac{3.n-1}{(n-1).n^2}.\sigma^4$ dont le signe est positif dès que $n \succ 1$.

• Ainsi, entre S_X^2 qui n'est pas sans biais et $\widehat{S_X}^2$ qui constitue un estimateur sans biais de σ^2, c'est S_X^2 qui présente la meilleure efficacité (risque $R(S_X^2) \prec R(\widehat{S_X}^2)$) et qui, de ce fait, est préférable à $\widehat{S_X}^2$ contrairement à ce que l'on pourrait penser et qui, le plus souvent, est admis.

6. On place un compteur devant une source de rayonnement. Le nombre d'impulsions que l'on peut enregistrer pendant une minute est une variable de POISSON d'espérance μ (constante, pendant la durée de l'expérience) et dont la valeur mesure l'intensité de la source.

1°) On effectue un comptage d'une minute et on observe 6400 impulsions. En déduire, un intervalle de confiance de μ au coefficient de sécurité 90% (seuil de confiance).

2°) Pour améliorer la précision de cette estimation, on répète n fois la mesure en question. Sachant que la moyenne de ces n mesures est de l'ordre de 6400, quelle valeur faut-il donner à n pour estimer μ à dix unités près, ceci toujours au seuil de confiance 90% ?

3°) Améliorerait-on l'estimation en procédant à un comptage de n minutes plutôt que n comptages d'une minute ?

Solution : 1°) Notant par X la variable représentant le nombre d'impulsions enregistrées durant une minute, il s'agit selon les indications de l'énoncé, d'une variable de POISSON caractérisée par la loi $\Pr ob(X = x) = \frac{e^{-\mu}.\mu^x}{x!}$. Entre autres, on a $E(X) = \mu$ et $Var(X) = \mu$.

Toutefois, la grande valeur de μ autorise ici d'admettre la convergence de X vers la loi normale de moyenne μ et de variance μ. Dans ces conditions, l'encadrement par intervalle de confiance au seuil $1-\alpha = 90\%$ de la variable normale centrée réduite $\xi = \frac{X-\mu}{\sqrt{\mu}}$ conduit à $-t_\alpha \leq \frac{X-\mu}{\sqrt{\mu}} \leq t_\alpha$ où t_α vérifie $\Pr ob(|\xi| \leq t_\alpha) = 1-\alpha \Rightarrow t_\alpha = 1,645$ (lorsque $1-\alpha = 90\%$).

On a donc finalement l'encadrement $X - t_\alpha.\sqrt{\mu} \le \mu \le X + t_\alpha.\sqrt{\mu}$ avec $X = 6400$, soit l'intervalle de confiance $6400 - t_\alpha.\sqrt{\mu} \le \mu \le 6400 + t_\alpha.\sqrt{\mu} \Rightarrow |6400 - \mu| \le t_\alpha.\sqrt{\mu}$.

• Pour ce qui est d'expliciter numériquement l'intervalle susmentionné, l'approximation la plus simple est d'assimiler $\sqrt{\mu}$ à $\sqrt{X} = \sqrt{6400}$, d'où le résultat $6268,4 \le \mu \le 6531,6$.

A défaut de cette hypothèse simplificatrice, la méthode exacte conduit, par élévation au carré, à l'inégalité $(6400 - \mu)^2 \le 1,645^2.\mu$, d'où le trinôme $\mu^2 - 12802,706.\mu + 6400^2$ de racines $\mu_1 = 6269,75$ et $\mu_2 = 6532,96$. Ce trinôme est négatif ou nul lorsque $6269,75 \le \mu \le 6532,96$ qui forme l'intervalle de confiance cherché et dont on peut remarquer qu'il reste très proche du résultat suivant hypothèse simplificatrice.

2°) Lorsqu'on répète n fois l'expérience, soit $(X_1, X_2, ..., X_n)$ l'échantillon de n valeurs indépendantes ainsi obtenues, la moyenne μ est estimée ponctuellement par la moyenne empirique $\overline{X} = \dfrac{\sum_{i=1}^{i=n} X_i}{n}$. Pour n assez grand, le théorème central- limite entraîne la convergence de $\overline{X}$ vers la loi normale encore que cette propriété est vérifiée de fait ici puisqu'on a vu dans la question précédent qu'on pouvait admettre la convergence des X_i vers la loi normale et que dans ces conditions $\overline{X}$ est une variable normale. Plus précisément, il s'agit de la loi normale $N(m, \dfrac{\sigma}{\sqrt{n}})$ avec $\sigma = \sqrt{\mu}$.

• La question posée revient à chercher la valeur minimale de n, soit n^*, à partir de laquelle $|\overline{X} - \mu| \le 10$ au seuil de confiance $1 - \alpha = 90\%$. Passant à la variable centrée réduite associée à $\overline{X}$, soit $\xi = \dfrac{\overline{X} - \mu}{\sqrt{\mu/n}}$, on peut écrire que $\Pr ob(|\overline{X} - \mu| \le 10) = 0,90$ est équivalent à la relation $\Pr ob(|\xi| \le \dfrac{10}{\sqrt{\mu/n}}) = 0,90$. Désignant ici encore le nombre t_α vérifiant $\Pr ob(|\xi| \le t_\alpha) = 1 - \alpha$ et considérant la fonction de répartition $\Pi(t) = \Pr ob(\xi \le t)$, il vient immédiatement, pour $1 - \alpha = 90\%$, $\Pr ob(|\xi| \le t_\alpha) = 2.\Pi(t_\alpha) - 1 = 1 - \alpha \Rightarrow \Pi(t_\alpha) = 0,95 \Rightarrow t_\alpha = 1,645$.

Il suffit de rendre $\dfrac{10}{\sqrt{\mu/n}}$ supérieur à $t_\alpha = 1,645$ pour que la relation susmentionnée soit satisfaite. Par élévation au carré, on a donc $\dfrac{100.n}{\mu} \ge (1,645)^2$, soit $n \ge \dfrac{(1,645)^2.\mu}{100}$. Toutefois, μ étant inconnue, on utilisera son approximation par $\overline{X}$ ou, à défaut, par l'information qui résulte du seul comptage connu, soit 6400 pour le cas présent (cf. 1ère question).

En conclusion, la valeur minimale cherchée n^* est égale à $\dfrac{(1,645)^2 \times 6400}{100} = 174$ (par excès).

3°) Lorsqu'on procède à un seul comptage durant n minutes, on génère une variable aléatoire $S_n = \sum_{i=1}^{i=n} X_i$ dont la loi suit une loi de POISSON de paramètre $n.\mu$ puisqu'on sait que la loi de POISSON est stable par rapport à l'addition. Par convergence vers la loi normale, on a donc $S_n = \sum_{i=1}^{i=n} X_i = n.\overline{X}$ qui suit approximativement la loi normale $N(n.\mu, \sqrt{n.\mu})$.

L'encadrement de la variable normale centrée réduite $\xi = \dfrac{S_n - n.\mu}{\sqrt{n.\mu}}$, soit $-t_\alpha \le \xi \le t_\alpha$, conduit à l'intervalle de confiance $S_n - t_\alpha.\sqrt{n.\mu} \le \mu \le S_n + t_\alpha.\sqrt{n.\mu}$ qui s'écrit encore $\overline{X} - t_\alpha.\sqrt{\mu/n} \le \mu \le \overline{X} + t_\alpha.\sqrt{\mu/n}$. Cet intervalle est tout à fait identique à celui obtenu dans la 2ème question à partir de l'observation de n comptages de une minute. En d'autres termes, les deux méthodes sont rigoureusement équivalentes ici.

7. Soit un échantillon $(Y_1, Y_2, ..., Y_n)$ de n variables aléatoires (avec $Y_i > 0$ pour tout i) tel que $\ln(Y_i)$ suit la loi normale $N(\theta, 1)$ avec $\theta \in R$.

1°) Calculer $E\left[(Y_1)^\alpha\right]$ pour $\alpha \in R$.

2°) On estime e^θ par Y_1. Calculer le biais et le risque quadratique de cet estimateur.

3°) Calculer l'espérance mathématique de la moyenne arithmétique $\overline{Y} = \dfrac{1}{n}.\sum_{i=1}^{i=n} Y_i$ en fonction de θ. En déduire un estimateur sans biais Z de e^θ. Quel est son risque quadratique ?

4°) Calculer l'espérance mathématique de la moyenne géométrique $(\prod_{i=1}^{i=n} Y_i)^{\frac{1}{n}}$ en fonction de θ. En déduire un nouvel estimateur sans biais V de e^θ. Calculer son risque quadratique.

5°) Que conclure des résultats précédents ?

Solution : 1°) Tout d'abord, il est rappelé que la variable normale centrée réduite ξ, a pour fonction caractéristique $\phi_\xi(t) = e^{-\frac{t^2}{2}}$. En effet, $\phi_\xi(t) = E\left[e^{i.t.\xi}\right] = \dfrac{1}{\sqrt{2.\pi}}.\int_{-\infty}^{+\infty} e^{i.t.x}.e^{-\frac{x^2}{2}}.dx$, soit $\phi_\xi(t) = \dfrac{e^{-\frac{t^2}{2}}}{\sqrt{2.\pi}}.\int_{-\infty}^{+\infty} e^{-\frac{u^2}{2}}.du$ avec $u = x - i.t$. En résumé, $\phi_\xi(t) = e^{-\frac{t^2}{2}}$ puisque, par définition de la densité de probabilité de la loi normale centrée réduite $N(0,1)$, on a $\dfrac{1}{\sqrt{2.\pi}}.\int_{-\infty}^{+\infty} e^{-\frac{u^2}{2}}.du = 1$.

Plus généralement, si on considère la variable aléatoire X de loi normale $N(m, \sigma)$, la relation $X = \sigma.\xi + m$ entraîne $\phi_X(t) = E\left[e^{i.t.(\sigma.\xi + m)}\right] = e^{i.t.m}.e^{-\frac{t^2.\sigma^2}{2}}$.

• Revenant à $E\left[(Y_1)^\alpha\right]$, c'est aussi $E\left[\exp(\alpha.\ln(Y_1))\right]$, soit $E\left[\exp(\alpha.U)\right]$ où $U = \ln(Y_1)$ suit la loi normale $N(\theta, 1)$.

On reconnaît dans l'espérance mathématique précédente la fonction caractéristique de U prise au point $-i.\alpha$, soit $\phi_U(-i.\alpha)$. Il vient donc $E\left[(Y_1)^\alpha\right] = e^{i.\theta(-i.\alpha)}.e^{-(-i.\alpha)^2/2} = e^{\alpha\theta}.e^{\alpha^2/2}$.

2°) Estimant e^θ par Y_1, le calcul de $E(Y_1) = e^\theta.e^{1/2}$ (c'est le résultat général obtenu dans la 1ère question auquel on applique la valeur particulière $\alpha = 1$), montre qu'on obtient un estimateur de θ qui n'est pas sans biais puisque $E(Y_1) = \sqrt{e}.e^\theta \neq e^\theta$. Plus précisément, le biais de cet estimateur est $B(\theta) = e^\theta.(\sqrt{e}-1)$.

• Quant au risque quadratique $R(Y_1) = E\left[(Y_1 - e^\theta)^2\right]$, c'est suivant la linéarité de l'espérance mathématique et par développement, $R(Y_1) = E(Y_1^2) - 2.e^\theta.E(Y_1) + e^{2.\theta}$. Du résultat de la 1ère question décliné cette fois pour le cas particulier $\alpha = 2$, il ressort $E\left[(Y_1)^2\right] = e^{2.\theta}.e^2$, puis en conséquence, $R(Y_1) = (e^2 - 2.\sqrt{e} + 1).e^{2.\theta}$.

3°) Considérant cette fois, l'estimateur $\overline{Y} = \frac{1}{n}.\sum_{i=1}^{i=n} Y_i$, les propriétés classiques de l'espérance mathématique et de la variance entraînent immédiatement, pour $E(\overline{Y})$, le résultat $E(\overline{Y}) = \frac{1}{n}.\sum_{i=1}^{i=n} E(Y_i) = \frac{n.\sqrt{e}.e^\theta}{n} = \sqrt{e}.e^\theta$. De même, $Var(\overline{Y}) = \frac{1}{n^2}.\sum_{i=1}^{i=n} Var(Y_i) = \frac{1}{n^2}.n.Var(Y_1)$. Mais, $Var(Y_1) = E(Y_1^2) - E(Y_1)^2$, soit $Var(Y_1) = (e^2 - e).e^{2.\theta}$.

Finalement, $Var(\overline{Y}) = \frac{(e^2 - e).e^{2.\theta}}{n}$.

• Pour corriger le biais susmentionné de $E(\overline{Y})$, l'introduction de l'estimateur $Z = \frac{1}{\sqrt{e}}.\overline{Y}$ conduit à un estimateur sans biais de e^θ (puisque $E(Z) = \frac{1}{\sqrt{e}}.E(\overline{Y}) = e^\theta$) et de risque quadratique $E\left[(Z - e^\theta)^2\right] = Var(Z)$ avec $Var(Z) = \frac{1}{e}.Var(\overline{Y}) = \frac{e-1}{n}.e^{2.\theta}$.

4°) Utilisant la relation $(\prod_{i=1}^{i=n} Y_i)^{1/n} = \exp(\frac{1}{n}.\sum_{i=1}^{i=n} \ln(Y_i))$, l'espérance mathématique de la moyenne géométrique est celle de la variable $\exp(\frac{S}{n})$, où $S = \sum_{i=1}^{i=n} \ln(Y_i)$ suit la loi normale de moyenne $n.E\left[\ln(Y_1)\right] = n.\theta$ et de variance $n.Var\left[\ln(Y_1)\right] = n$ (puisque la loi normale est stable pour l'addition, du moins lorsque les variables considérées sont indépendantes).

En conclusion et compte tenu des résultats de la 1ère question appliqués au cas $\alpha = \frac{1}{n}$, on a :

$$E\left[(\prod_{i=1}^{i=n} Y_i)^{1/n}\right] = \phi_S(-\frac{i}{n}) = e^{i.(n.\theta).(\frac{-i}{n})}.e^{-(-\frac{i}{n})^2.\frac{n}{2}} = e^\theta.e^{\frac{1}{2.n}}.$$

• Ici encore, le recours à l'estimateur $V = e^{-\frac{1}{2.n}}.(\prod_{i=1}^{i=n} Y_i)^{1/n}$ corrige le biais puisque $E(V) = e^{-\frac{1}{2.n}}.E\left[(\prod_{i=1}^{i=n} Y_i)^{1/n}\right] = e^\theta$.

D'autre part, pour ce qui est du risque quadratique $E\left[(V-e^{\theta})^2\right]=Var(V)$, on a $Var(V)=e^{-\frac{1}{n}}.Var\left[(\prod_{i=1}^{i=n}Y_i)^{1/n}\right]$.

Mais, $E\left[((\prod_{i=1}^{i=n}Y_i)^{1/n})^2\right]=\phi_S(-\frac{2.i}{n})=e^{i.(n.\theta).(-\frac{2.i}{n})}.e^{-(\frac{2.i}{n})^2.\frac{n}{2}}=e^{2.\theta}.e^{2/n}$. Il en résulte pour la variance, $Var\left[(\prod_{i=1}^{i=n}Y_i)^{1/n}\right]=E\left[((\prod_{i=1}^{i=n}Y_i)^{1/n})^2\right]-E\left[(\prod_{i=1}^{i=n}Y_i)^{1/n}\right]^2$, soit compte tenu des calculs précédents, l'expression $(e^{\frac{2}{n}}-e^{\frac{1}{n}}).e^{2.\theta}=e^{\frac{1}{n}}.(e^{\frac{1}{n}}-1).e^{2.\theta}$. Revenant à $Var(V)$, on aboutit en définitive au résultat $Var(V)=(e^{\frac{1}{n}}-1).e^{2.\theta}$.

5°) La comparaison des risques quadratiques associés aux estimateurs $Y_1, Z,$ et V conduit immédiatement au classement $R(Y_1)\geq R(Z)\geq R(V)$.

En effet, $R(Y_1)=(e^2-2.\sqrt{e}+1).e^{2.\theta}=5,09.e^{2.\theta}$ est bien supérieur à $R(Z)=\frac{e-1}{n}.e^{2.\theta}$, soit $R(Z)=\frac{1,72}{n}.e^{2.\theta}$ dès que $n\geq 1$, ce qui établit la première partie du classement ci-dessus.

Par ailleurs, le rapport $\frac{R(Z)}{R(V)}=\frac{1}{n}.\frac{e-1}{e^{1/n}-1}$ s'écrit aussi, en minorant la série géométrique de raison $e^{1/n}$, $\frac{R(Z)}{R(V)}=\frac{1}{n}.\frac{(e^{1/n})^n-1}{e^{1/n}-1}\geq\frac{1}{n}.\left[1+e^{1/n}+e^{2/n}+...+e^{n-1/n}\right]$. Or, pour tout $u\geq 0$, l'inégalité $e^u\geq 1+u$ entraîne ici la minoration $\frac{R(Z)}{R(V)}\geq\frac{1}{n}.\left[1+(1+\frac{1}{n})+...(1+\frac{n-1}{n})\right]$, c'est à dire $\frac{R(Z)}{R(V)}\geq\frac{1}{n}.\left[n+\frac{1+2+...+(n-1)}{n}\right]=\frac{1}{n}.\left[n+\frac{n.(n-1)}{2.n}\right]\succ 1$. Ainsi a-t-on bien le résultat attendu $R(Z)\succ R(V)$.

- En conclusion, les formules proposées pour $Y_1, Z,$ et V, ont conduit à construire trois estimateurs sans biais de e^{θ} de qualités croissantes, V étant le meilleur estimateur ainsi obtenu.

8. On considère un échantillon de n variables aléatoires indépendantes $(X_1, X_2, ..., X_n)$ de variable parente de densité de probabilité $f_{\theta}(x)=\theta.x^{\theta-1}.1_{0<x<1}$ avec $\theta\succ 0$.

1°) Déterminer l'estimateur du maximum de vraisemblance (E.M.V) de θ, soit $\hat{\theta}$.

2°) Montrer que $\frac{n}{\hat{\theta}}$ suit la loi de densité de probabilité $g(y)=\frac{\theta^n}{(n-1)!}.y^{n-1}.\exp(-\theta.y)$ où $y\in R^+$.

3°) En déduire que $\frac{1}{\hat{\theta}}$ est un estimateur sans biais de $\frac{1}{\theta}$ et que la variance $Var(\frac{1}{\hat{\theta}})$ qu'on exprimera, atteint la borne minimale F.D.C.R de CRAMER RAO .

Solution : 1°) La fonction de vraisemblance associée à l'échantillon de taille n considéré et de variable parente X est égale au produit des densités de probabilité, soit :

$$L(X_1, X_2, ..., X_n, \theta) = \theta^n . \prod_{i=1}^{i=n} X_i^{\theta-1} . 1_{0 \prec X_i \prec 1}$$

Considérant la fonction de log- vraisemblance $\ln L = n.\ln\theta + (\theta - 1).\sum_{i=1}^{i=n} \ln(X_i)$, pour $0 \prec X_i \prec 1, 1 \le i \le n$, l'estimateur E.M.V cherché, soit $\hat{\theta}$, est la solution de l'équation $\frac{\partial}{\partial\theta} \ln L = 0$, ce qui entraîne, $\frac{n}{\theta} + \sum_{i=1}^{i=n} \ln(X_i) = 0$.

Il s'ensuit $\hat{\theta} = -\frac{n}{\sum_{i=1}^{i=n} \ln(X_i)}$ qui est bien un maximum puisque $\frac{\partial^2}{\partial\theta^2} \ln L = -\frac{n}{\hat{\theta}^2} \prec 0$.

2°) Remarquant que $\frac{n}{\hat{\theta}} = -\sum_{i=1}^{i=n} \ln(X_i)$, il est proposé ici d'identifier la loi des variables $U_i = -\ln(X_i)$. Désignant par $h(u_i)$ la densité de probabilité de la variable U_i, l'application du théorème de la mesure image qui consiste à opérer dans la densité de probabilité élémentaire de X_i le changement de variable $u_i = -\ln(x_i) \Rightarrow x_i = e^{-u_i}$, conduit au résultat :

$$h(u_i).du_i = \theta.e^{-(\theta-1).u_i}.e^{-u_i}.du_i = \theta.e^{-\theta.u_i}.du_i, u_i \in R^+$$

On constate donc que U_i dont la densité de probabilité est $h(u_i) = \theta.e^{-\theta.u_i}$, suit la loi exponentielle de paramètre θ.

• Dans ces conditions, $\frac{n}{\hat{\theta}}$ qui est défini par la somme $\sum_{i=1}^{i=n} U_i$, suit la loi Gamma n (somme de n variables aléatoires exponentielles indépendantes et équidistribuées), loi dont la densité de probabilité est $g(y) = \frac{\theta^n}{(n-1)!} . y^{n-1} . e^{-\theta.y}, y \in R^+$ (cf. rappels de cours du chapitre I, paragraphe 1.1).

3°) De $E(U_i) = \frac{1}{\theta}$ (espérance mathématique de la loi exponentielle), découle par linéarité, $E(\frac{n}{\hat{\theta}}) = \frac{n}{\theta}$. Ainsi, $E(\frac{1}{\hat{\theta}}) = \frac{1}{\theta}$ et $\frac{1}{\hat{\theta}}$ forme-t-il un estimateur sans biais de $\frac{1}{\theta}$.

De même, $Var(U_i) = \frac{1}{\theta^2}$ (variance de la loi exponentielle), ce qui entraîne $Var(\frac{n}{\hat{\theta}}) = \frac{n}{\theta^2}$. On en déduit, notamment, $Var(\frac{1}{\hat{\theta}}) = \frac{1}{n^2} . Var(\frac{n}{\hat{\theta}}) = \frac{1}{n.\theta^2}$.

Ainsi, $\frac{1}{\hat{\theta}}$ est-il un estimateur convergent puisque $\lim_{n \to +\infty} Var(\frac{1}{\hat{\theta}}) = 0$. Plus encore, la loi parente de densité de probabilité $f(x, \theta) = \theta.x^{\theta-1}.1_{0 \prec x \prec 1}$ appartient à la famille exponentielle puisque $\ln f(x, \theta) = (\theta - 1).\ln x + \ln\theta + \ln 1_{0 \prec x \prec 1}$ a pour forme $Q(\theta).T(x) + \beta(\theta) + \gamma(x)$.

Par ailleurs, $T(x) = -\sum_{i=1}^{i=n} \ln(X_i)$ forme bien une statistique exhaustive puisque vérifiant le théorème de factorisation $L(X_1, X_2, ..., X_n, \theta) = \varphi(T(X), \theta).\psi(X)$.

En effet, on a bien $L(X_1,X_2,...,X_n,\theta)=\theta^n.\exp\left[(\theta-1).\sum_{i=1}^{i=n}\ln(X_i)\right].\prod_{i=1}^{i=n}1_{0<X_i<1}$ qui s'écrit sous la forme précédente avec $\psi(X)=\prod_{i=1}^{i=n}1_{0<X_i<1}$.

• En conclusion, l'estimateur sans biais de $\frac{1}{\theta}$ que constitue $\frac{1}{\hat{\theta}}$ atteint la borne minimale F.D.C.R puisque la condition nécessaire et suffisante qui est vérifiée pour le cas des modèles appartenant à la famille exponentielle et pour les statistiques exhaustives, est bien remplie ici (cf. rappels de cours du présent chapitre, paragraphe 2.4).

• Le calcul direct ci-dessous confirme d'ailleurs ces conclusions. On a successivement (en se méfiant du fait que c'est le paramètre $1/\theta$ qui est considéré ici et non pas θ) :

$$I_X(1/\theta)=E\left[\frac{\partial^2}{\partial(1/\theta)^2}\ln f(X,\theta)\right], \text{ avec } \ln f(X,\theta)=\ln\theta+(\theta-1).\ln X+\ln 1_{0<X<1},$$

puis par dérivation, $\frac{\partial}{\partial(1/\theta)}\ln f(X,\theta)=\frac{\partial}{\partial\theta}\ln f(X,\theta).\frac{\partial\theta}{\partial u}$, avec $u=\frac{1}{\theta}$, soit $-\theta^2.(\frac{1}{\theta}+\ln X)$, c'est-à-dire, finalement, $\frac{\partial}{\partial(1/\theta)}\ln f(X,\theta)=-\theta-\theta^2\ln X$. Enfin, dérivant de nouveau, $\frac{\partial^2}{\partial(1/\theta)^2}\ln f(X,\theta)=-\theta^2.(-1-2.\theta.\ln X)=\theta^2+2.\theta^3.\ln X$.

• Ainsi $I_X(1/\theta)=-E\left[\theta^2+2.\theta^3.\ln X\right]=-\theta^2-2.\theta^3.E\left[\ln X\right]$, avec $E\left[\ln X\right]=-\frac{1}{\theta}$ (en effet, $-\ln X$ est exponentielle de paramètre θ). Finalement, $I_X(1/\theta)=-\theta^2+2.\theta^2=\theta^2$, la borne minimale F.D.C.R étant, quant à elle, égale à $\frac{1}{n.I_X(1/\theta)}$, soit $\frac{1}{n.\theta^2}$.

C'est bien la valeur de $Var(\frac{1}{\hat{\theta}})$ précédemment calculée qui est donc bien le meilleur estimateur sans biais de $\frac{1}{\theta}$.

CHAPITRE III

DECISION

A - Rappels de cours

1. Les principes généraux de la décision statistique

1.1 L'objet des tests d'hypothèse

Les développements présentés dans ce chapitre portent sur les *tests entre deux hypothèses*, le problème posé étant le choix, au vu d'observations fournies par un échantillon de taille n, entre une **hypothèse nulle**, notée H_0, et une **hypothèse alternative**, notée H_1 (on supposera qu'un seule de ces deux hypothèses est vraie).

Le contrôle de qualité, l'évaluation de l'efficacité d'un nouveau procédé, la validation ou non d'un ajustement dans le cadre d'une modélisation, le prononciation d'une peine, sont entre autres, des exemples de situations se résumant au choix «binaire » ci-dessus.

1.2 Les risques associés

Décision\Etats de la nature ↓ →	H_0	H_1
H_0	0	β
H_1	α	0

Comme le montre le tableau des enjeux ci-contre, deux erreurs menacent le décideur dans le contexte considéré :

- le risque noté « α » de décider H_1 alors que H_0 est vraie, dit « **erreur de première espèce** » ;
- le risque noté « β » de décider H_0 alors que H_1 est vraie, dit « **erreur de seconde espèce** ».

L'idéal serait de rendre les deux erreurs α et β, toutes les deux faibles. Mais, malheureusement, lorsqu'on dispose d'un nombre restreint d'observations, ce n'est pas possible, d'autant que les deux risques en question sont *antinomiques*. Concrètement, plus on cherche à diminuer α, plus on augmente la valeur de β et vice-versa.

Les enjeux de la justice illustrent simplement ce résultat, à travers les deux hypothèses H_0 « *être innocent* » et H_1 « *être coupable* », et les deux décisions « *acquitter* » et « *condamner* ». Il est bien manifeste que le souci de ne pas condamner un innocent augmente le risque de ne pas condamner un coupable.

Pratiquement, dans le cadre des tests classiques, c'est à partir de α qu'on va bâtir la **règle de décision**.

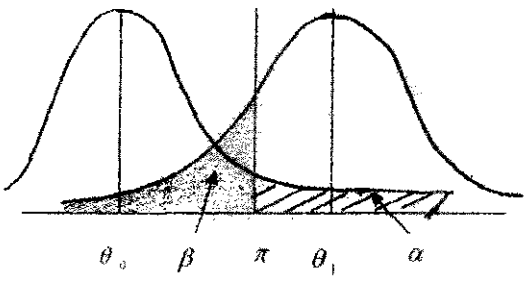
Puisque c'est le risque α qui est maîtrisé, on choisira donc les hypothèses H_0 et H_1 de sorte que ce qui soit vraiment intéressant, c'est rejeter H_0. En d'autres termes, H_0 est l'hypothèse envers laquelle on a le plus de confiance ou de sécurité.

Reprenant l'exemple précédent de la justice et procédant par analogie, on fait valoir le bénéfice du doute → tout accusé est présumé innocent, il fait prouver sa culpabilité. On parle aussi de test « *conservatif* » dans la mesure où on conserve l'hypothèse H_0, sauf si les données observées conduisent à la rejeter.

Autre exemple, considérant le taux d'une certaine substance dans le sang caractéristique d'une maladie peu grave dont le traitement peut cependant générer des effets secondaires fâcheux, on préférera contenir le risque α de traiter un malade en bonne santé plutôt que le risque β de ne pas diagnostiquer la maladie chez un sujet atteint (puisque cette maladie est peu grave alors que le traitement peut causer des effets secondaires fâcheux).

A l'inverse, s'il s'agissait d'une maladie potentiellement très grave mais facilement soignable, le risque majeur serait de ne pas la détecter. Cette fois, c'est sur la base « *être atteint par la maladie* » qu'il faudrait construire l'hypothèse nulle H_0.

• Il convient néanmoins de signaler que l'augmentation de la taille de l'échantillon permet, α étant fixé, de diminuer β. Très couramment est utilisée à ce sujet, la fonction $P=1-\beta$ dite « **puissance du test** » qui est égale à Prob(décider H_0 sachant H_1 vraie) et qui mesure en quelque sorte le caractère discriminatoire du test entre les deux hypothèses.



Le schéma ci-contre dans lequel les valeurs de θ sont portées en abscisses et celles de $\overline{X}$ en ordonnées, et qui porte sur une moyenne de loi normale et sur un test d'hypothèses $H_0 : \theta = \theta_0$ contre $H_1 : \theta = \theta_1$ (avec $\theta_1 \succ \theta_0$), la statistique discriminante considérée étant $\overline{X} = \frac{\sum_{i=1}^{i=n} X_i}{n}$, illustre le sens de α, β, et $P=1-\beta$, la règle de décision étant ici de décider H_1 à droite du seuil π et H_0 dans le cas contraire.

On y distingue ainsi α = Prob(décider H_1 / H_0 vraie) en zone hachurée et l'erreur de seconde espèce β = Prob(décider H_0 / H_1 vraie) en zone grisée. Diminuer α, c'est faire glisser π vers la droite et donc augmenter β. D'autre part, l'augmentation de n conduit à « amincir » la densité de probabilité de $\overline{X}$ puisque $\sigma(\overline{X}) = \sigma / \sqrt{n}$ décroît lorsque n augmente. Cela tend donc à réduire la partie en grisée et à fortiori à accroître la capacité du test à séparer les hypothèses H_0 et H_1, c'est-à-dire sa puissance $P=1-\beta$.

1.3 La classification des tests

Très vaste est la palette des tests qui peuvent être mis en œuvre pour traiter des problèmes ayant pour objet :

- la **conformité** d'un paramètre à une valeur standard donnée, les cas les plus courants étant ceux d'une moyenne, d'une proportion, et d'une variance ;
- la **comparaison** d'un paramètre ou plus généralement d'une distribution de probabilités entre K groupes (populations, échantillons...), le cas $K = 2$ étant plus particulièrement développé ici ;
- l'**ajustement** d'une distribution théorique donnée aux données observées ;
- l'**indépendance** entre variables aléatoires ;

....

- A cet effet, les deux grandes familles à considérer sont celles :
  - des **tests paramétriques** qui portent sur le paramètre de la distribution associée aux données considérées ;
  - des **tests non paramétriques** qui ne font pas d'hypothèse sur ladite distribution.
- Le choix du test à utiliser est également fonction de la nature des données proposées :
  - *paramétrique ou non*, dans le cas de valeurs représentatives des écarts, telles des mesures (variables dites **quantitatives**) ;
  - *non paramétrique*, dans le cas contraire de variables **qualitatives** ou **ordinales** à valeurs en nombre fini, telles oui-non, homme-femme, peu satisfait-satisfait-très satisfait...
- Enfin, dans le cadre des problèmes de comparaison, on distinguera :
  - les **échantillons indépendants** dans lesquels les observations faites sont indépendantes à l'intérieur d'un groupe et entre les groupes considérés ;
  - les **échantillons appariés** dans lesquels d'un groupe à l'autre les données sont liées, tel le cas le plus courant où il est procédé à des mesures répétées sur les mêmes sujets (par exemple, le poids d'une personne avant et après un régime).

2. Les tests paramétriques

2.1 Hypothèses simples et multiples

Portant sur les observations $(x_1, x_2, ..., x_n)$ qui sont les réalisations de n variables aléatoires indépendantes $(X_1, X_2, ..., X_n)$ de même loi dépendant d'un paramètre θ, ces tests dont les finalités sont principalement la **conformité** et l'**homogénéité**, ont pour forme générale $H_0 : \theta \in \Theta_0$ contre $H_1 : \theta \in \Theta_1$, les sous-ensembles Θ_0 et Θ_1 étant disjoints et recouvrant toutes les valeurs possibles de θ.

- Dans ce cadre, on distingue les **hypothèses simples** « $\theta = \theta_0$ » et les **hypothèses multiples** (ou composites) pour lesquelles Θ_0 n'est pas réduit à une seule valeur, telles $\theta \succ \theta_0$, $\theta \prec \theta_0, \theta \neq \theta_0$ ...., les termes **unilatéral** et **bilatéral** étant utilisés suivant qu'on se trouve d'un seul côté ou non de la valeur ciblée θ_0.

- Le plus souvent l'une des deux hypothèses au moins sera simple (en principe, c'est H_0), les tests ainsi rencontrés étant :

$\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta = \theta_1 \end{cases}$ (*test entre deux hypothèses simples*)

$\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta \succ \theta_0 \end{cases}$ (ou $H_1 : \theta \prec \theta_0$) (*test unilatéral*)

$\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta \neq \theta_0 \end{cases}$ (*test bilatéral*)

Mais d'autres schémas sont possibles comme :

$\begin{cases} H_0 : \theta \leq \theta_0 \\ H_1 : \theta \succ \theta_0 \end{cases}$ (ou vice-versa)

$\begin{cases} H_0 : \theta_1 \leq \theta \leq \theta_2 \\ H_1 : \theta \prec \theta_1 ou \theta \succ \theta_2 \end{cases}$

S'agissant des tests d'hypothèses multiples, on notera que, lorsque H_0 n'est pas une hypothèse simple, l'erreur de première espèce α =Prob (décider H_1 / H_0 vraie) devient une fonction de θ, sa borne supérieure $\underset{\theta \in \Theta_0}{Sup} \alpha(\theta)$ étant appelée « **niveau du test** ».

2.2 La constrnction de la règle de décision

Comme cela a déjà été indiqué au paragraphe 1.2 précédent, c'est à partir de l'erreur de première espèce α dont on se fixe la valeur (en principe 5% ou 10%), qu'on établit la *règle de décision* qui va permettre de trancher entre les hypothèses H_0 et H_1, et plus particulièrement la **région** dite « **critique** » formée de l'ensemble des valeurs $(x_1, x_2, ..., x_n)$ pour lesquelles on décide H_1, soit W (*principe de* NEYMAN *et* PEARSON).

A travers l'application 1.4 du chapitre I entre autres, on a pu constater le *lien étroit qui existe entre l'estimation par intervalle de confiance et la décision statistique.* Il est donc naturel ici, pour caractériser la région critique, de s'appuyer sur un estimateur ponctuel de θ dont plus particulièrement l'estimateur E.M.V du maximum de vraisemblance, surtout si cela permet d'utiliser une *fonction pivotale* telles celles déjà décrites en estimation statistique pour moyennes, variances, et proportions.

Bien évidemment, plusieurs seuils et régions critiques peuvent ainsi être retenus pour un même problème de décision statistique, mais l'*enjeu* reste de déterminer la méthode qui, pour α donné, permet d'aboutir au risque β le plus faible, c'est-à-dire au *test le plus puissant.* A cet égard, le **théorème de NEYMAN et PEARSON** offre une *solution optimale* dans le cadre du test entre deux hypothèses simples $H_0 : \theta = \theta_0$ contre $H_1 : \theta = \theta_1$, la région critique définie à partir du **rapport des fonctions de vraisemblance** $\frac{L(x_1, x_2, ..., x_n, \theta_0)}{L(x_1, x_2, ..., x_n, \theta_1)}$, soit $W = \left\{ (x_1, x_2, ..., x_n) / \frac{L(x_1, x_2, ..., x_n, \theta_0)}{L(x_1, x_2, ..., x_n, \theta_1)} \leq k \right\}$, conduisant à un *test de puissance maximale.*

De la donnée de l'erreur de première espèce α résulte la valeur de k, à travers l'équation $\Pr ob\left(\frac{L(x_1, x_2, ..., x_n, \theta_0}{L(x_1, x_2, ..., x_n, \theta_1} \leq k / \theta = \theta_0 \right) = \alpha$.

- Le théorème de NEYMAN et PEARSON et l'emploi du rapport de vraisemblance s'étendent aux *tests entre hypothèses multiples* à partir de la statistique $\lambda = \frac{\underset{\theta \in \Theta_0}{Sup} L(x, \theta)}{\underset{\theta \in \Theta_1}{Sup} L(x, \theta)}$.

Néanmoins, lorsque l'hypothèse H_0 n'est plus une hypothèse simple, la construction de la règle de décision devient plus complexe, quelques illustrations étant exposées dans les applications ci-après du présent chapitre.

• Enfin, tout en étant la technique la plus usuelle, le critère de NEYMAN et PEARSON n'est pas la seule approche de résolution des tests paramétriques, la **méthode de WALD** et les **procédures « bootstrap »** (par simulation) étant à citer ici.

2.3 Tests de conformité à une valeur standard

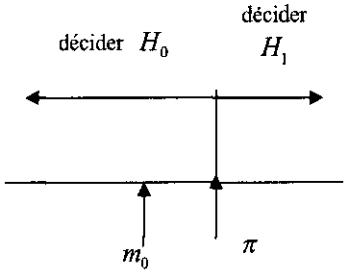
Particulièrement usités en *contrôle de qualité* et dans *l'exploitation des mesures*, ils portent sur la moyenne, proportion, et variance, mais aussi, moins couramment, sur le coefficient de corrélation, l'étendue, la médiane…

A l'instar des raisonnements tenus en estimation statistique par intervalle de confiance, il convient de distinguer le cas des **grands échantillons** ($n \geq 30$) pour lesquels le *théorème central limite* s'applique le plus souvent (ce qui autorise l'usage de la **loi normale**) et celui des **petits échantillons** ($n < 30$) dont le modèle le plus courant suppose que les données vérifient une *distribution normale* (loi de GAUSS), mais qui peuvent également être traités, par un calcul approprié, dans le cas d'autres types de distribution.

Les développements les plus classiques sont exposés ci-dessous :

a) Le cas d'une moyenne

On suppose raisonner sur un échantillon $(X_1, X_2, ..., X_n)$ de n variables aléatoires indépendantes équidistribuées de **loi « parente »** de type **« loi normale »**, soit $N(m, \sigma^2)$. La fonction **« discriminante »** à partir de laquelle on construit la *règle de décision* est ici la statistique habituelle $\overline{X} = \frac{\sum_{i=1}^{i=n} X_i}{n}$ ($\bar{x} = \frac{\sum_{i=1}^{i=n} x_i}{n}$ pour ce qui est des données observées).



Ainsi, pour les tests $\begin{cases} H_0 : m = m_0 \\ H_1 : m = m_1 \end{cases}$ (avec $m_1 \succ m_0$), voire plus largement $\begin{cases} H_0 : m = m_0 \\ H_1 : m \succ m_0 \end{cases}$, la *règle de décision* (choix de H_1 ou encore région dite « *critique* »), est-elle définie par $\bar{x} \geq \pi$, le seuil π étant déterminé à partir de l'erreur de première espèce α suivant α =Prob(décider H_1 / H_0 vraie), soit $\alpha = \Pr ob(\bar{x} \geq \pi / m = m_0)$.

Le passage à la variable normale centrée réduite $\xi = \frac{\bar{x} - m}{\sigma / \sqrt{n}}$ permet d'écrire immédiatement $\alpha = \Pr ob(\xi \geq \frac{\pi - m_0}{\sigma / \sqrt{n}})$ d'où le seuil cherché $\pi = m_0 + t_\alpha . \frac{\sigma}{\sqrt{n}}$ (t_α, solution de l'équation $\Pr ob(\xi \geq t_\alpha) = \alpha$.

Pour le cas considéré (loi normale), le choix de $\bar{x}$ comme *fonction discriminante* n'est pas seulement intuitif puisque correspondant au *test de puissance maximale* résultant du *rapport des vraisemblances et du théorème de* NEYMAN *et* PEARSON.

En effet, ce dernier est égal, pour ce qui est du cas gaussien, à :

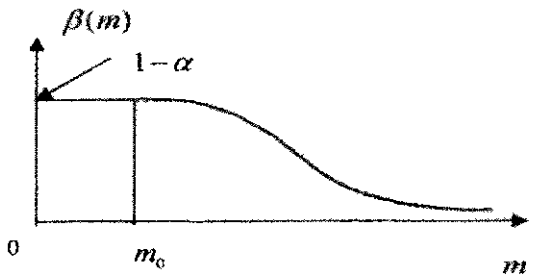
$$\frac{L(x_1,x_2,...,x_n,m_0)}{L(x_1,x_2,...,x_n,m_1)} = \exp\left[\frac{1}{2.\sigma^2}.\sum_{i=1}^{i=n}(x_i-m_1)^2 - \frac{1}{2.\sigma^2}.\sum_{i=1}^{i=n}(x_i-m_0)^2\right].$$

La région critique $W=\left\{(x_1,x_2,...,x_n)/\dfrac{L(x_1,x_2,...,x_n,m_0)}{L(x_1,x_2,...,x_n,m_1)}\le k\right\}$ est donc caractérisée (en passant aux logarithmes népériens) par la condition :

$$\frac{1}{2.\sigma^2}.\left[\sum_{i=1}^{i=n}(x_i-m_1)^2-\sum_{i=1}^{i=n}(x_i-m_0)^2\right]\le \ln k,$$

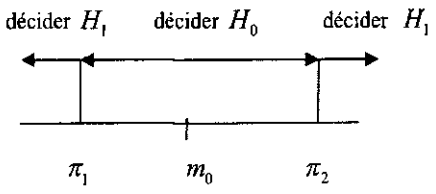
soit $(m_0-m_1).\sum_{i=1}^{i=n}(2x_i-(m_0+m_1))\le 2.\sigma^2.\ln k$, ou encore, après sommation, la relation $(m_0-m_1).\left[2n.\bar{x}-n.(m_0+m_1)\right]\le 2.\sigma^2.\ln k$.

Tenant compte de l'inégalité $m_1 \succ m_0$, il vient $\bar{x}\ge -\dfrac{2.\sigma^2.\ln k}{n.(m_1-m_0)}+\dfrac{m_0+m_1}{2}$, ce qui introduit bien ici la *fonction discriminante* $\bar{x}$, et la règle de décision $\bar{x}\ge\pi$.



Pour le test $\begin{cases}H_0 : m = m_0\\ H_1 : m \succ m_0\end{cases}$, la *courbe d'efficacité* du test demeure quant à elle une fonction de $m\,(m\succ m_0)$, soit $\beta(m)=\Pr ob(\bar{x}\prec\pi/m\succ m_0)$, avec $\beta(m_0)=1-\alpha$.

• Pour ce qui est du **test bilatéral** $\begin{cases}H_0 : m = m_0\\ H_1 : m \ne m_0\end{cases}$ dont la construction est celle d'un *intervalle de confiance* centré sur m_0 (du moins pour une loi à distribution symétrique (comme la loi normale), on pourra se ramener au cas antérieur par union des deux tests unilatéraux $\begin{cases}H_0 : m = m_0\\ H_1 : m \succ m_0\end{cases}$ et $\begin{cases}H_0 : m = m_0\\ H_1 : m \prec m_0\end{cases}$, chacun avec l'erreur de première espèce $\alpha' = \alpha/2$.



Pratiquement, on construira la région critique à partir des seuils π_1 et π_2 symétriques par rapport à m_0 et tels que : $\Pr ob(\bar{x}\le\pi_1/m=m_0)=\Pr ob(\bar{x}\ge\pi_2/m=m_0)=\alpha/2$

Il s'ensuit en posant $\pi_1=m_0-\varepsilon$ et $\pi_2=m_0+\varepsilon$, la relation $\Pr ob\left(\left|\bar{x}-m_0\right|\ge\varepsilon\right)=\alpha$.

Le passage à la variable normale centrée réduite $\xi = \dfrac{\bar{x} - m}{\sigma/\sqrt{n}}$ entraîne la *région critique* $\bar{x} \notin \left]m_0 - t_\alpha \cdot \dfrac{\sigma}{\sqrt{n}}, m_0 + t_\alpha \cdot \dfrac{\sigma}{\sqrt{n}}\right[$ où t_α vérifie $\Pr ob(|\xi| \geq t_\alpha) = \alpha$.

• Enfin , les résultats précédents peuvent être repris dans le cas où σ est **inconnu**. Il suffit d'y remplacer la fonction pivotale $\dfrac{\bar{x} - m}{\sigma/\sqrt{n}}$ par $\dfrac{\bar{x} - m}{\hat{S}/\sqrt{n}}$ dont la loi est de type STUDENT à $\nu = n - 1$ degrés de liberté (**test t de STUDENT à un échantillon**).

b) Le cas d'une proportion

Considérant les tests $\begin{cases} H_0 : p = p_0 \\ H_1 : p = p_1 \end{cases}$ (avec $p_1 \succ p_0$) voire $\begin{cases} H_0 : p = p_0 \\ H_1 : p \succ p_0 \end{cases}$, la *fonction discriminante* est représentée par la variable X « *occurrence du caractère C considéré dans l'échantillon de taille n prélevé* » ou encore la fréquence $F_n = \dfrac{X}{n}$ (c'est équivalent).

Il s'agit donc de trouver le seuil $C / \Pr ob(X \geq C / p = p_0) = \alpha$, relation dans laquelle X suit la loi $B(n, p)$.Bref, C est le plus petit entier à partir duquel la quantité $\Pr ob(X \geq C) = 1 - \Pr ob(X \prec C) = 1 - \sum_{k=0}^{k=C-1} C_n^k p_0^k \cdot (1 - p_0)^{n-k}$ est inférieure à α.

• En pratique, on se référera fréquemment aux *convergences de la loi binomiale* soit vers la loi de POISSON, soit vers la loi de GAUSS (loi normale), le seuil C obtenu dans ce dernier cas, étant défini par $C = n.p_0 + t_\alpha \cdot \sqrt{n.p_0.(1 - p_0)}$, t_α vérifiant $\Pr ob(\xi \geq t_\alpha) = \alpha$.

Dans le cas où on travaille sur la fréquence $F_n = \dfrac{X}{n}$ et non pas sur X, la région critique ayant pour forme $F_n \geq c$, le seuil c sera défini par $c = p_0 + t_\alpha \cdot \sqrt{\dfrac{p_0.(1 - p_0)}{n}}$.

• Enfin, une transcription immédiate des résultats ci-dessus pour le test **bilatéral** $\begin{cases} H_0 : p = p_0 \\ H_1 : p \neq p_0 \end{cases}$ conduit à la région critique :

$$W = \left\{ (x_1, x_2, \ldots, x_n) / F_n \notin \left]p_0 - t_\alpha \cdot \sqrt{\frac{p_0.(1 - p_0)}{n}}, p_0 + t_\alpha \cdot \sqrt{\frac{p_0.(1 - p_0)}{n}}\right[\right\}$$

où t_α vérifie $\Pr ob(|\xi| \leq t_\alpha) = \alpha$.

c) Le cas d'une variance

Supposant ici encore raisonner sur un **population normale** (loi de GAUSS), et utilisant les résultats des rappels de cours du chapitre I (paragraphe 2.4), la fonction « *discriminante* » est ici $S^2 = \dfrac{1}{n} \cdot \sum_{i=1}^{i=n} (X_i - m)^2$ ou $\hat{S}^2 = \dfrac{1}{n-1} \cdot \sum_{i=1}^{i=n} (X_i - \bar{X})^2$ selon les hypothèses « *m connue* », ou « *m inconnue* ».

Il est rappelé, relativement aux statistiques précédentes, que $\frac{n.S^2}{\sigma^2}$ (resp. $\frac{(n-1).\hat{S}^2}{\sigma^2}$) suit la **loi du chi-deux** à $v = n$ degrés de liberté(resp. $v = n-1$ degrés de liberté).

- Ainsi pour les tests $\begin{cases} H_0 : \sigma^2 = \sigma_0^2 \\ H_1 : \sigma^2 = \sigma_i^2 \end{cases}$ (avec $\sigma_1 \succ \sigma_0$), voire $\begin{cases} H_0 : \sigma^2 = \sigma_0^2 \\ H_1 : \sigma^2 \succ \sigma_0^2 \end{cases}$, la région critique est-elle définie par $W = \left\{(x_1, x_2, ..., x_n) / S^2 \geq \pi / \sigma^2 = \sigma_0^2\right\} \Rightarrow V = \frac{n.S^2}{\sigma^2} \geq a = \frac{\pi.n}{\sigma_0^2}$ où V suit la loi $\chi^2(n)$. Lisant dans les *tables de valeurs de la loi de chi- deux*, le nombre χ_α^2 vérifiant $\Pr ob(V \geq \chi_\alpha^2) = \alpha$, on obtient $\pi = \chi_\alpha^2 . \frac{\sigma_0^2}{n}$ (resp. $\pi = \chi_\alpha^2 . \frac{\sigma_0^2}{n-1}$ où χ_α^2 vérifie la relation $\Pr ob(\chi^2(n-1) \geq \chi_\alpha^2) = \alpha$ lorsque m est inconnue).

d) Autres tests de conformité

Par exemple, dans le cas des *dépendances statistiques* plus particulièrement développées dans le chapitre IV, et considérant le **coefficient de corrélation linéaire de PEARSON**, $r_{X,Y} = \frac{\text{cov}(X,Y)}{\sqrt{VarX.VarY}}$, on peut traiter sous certaines conditions le test $\begin{cases} H_0 : r = r_0 \\ H_1 : r = r_1 \end{cases}$, voire les tests unilatéraux $\begin{cases} H_0 : r = r_0 \\ H_1 : r \succ r_0 \end{cases}$, $\begin{cases} H_0 : r = r_0 \\ H_1 : r \prec r_0 \end{cases}$, ou le test bilatéral $\begin{cases} H_0 : r = r_0 \\ H_1 : r \neq r_0 \end{cases}$.

On montre en effet que, pour *n assez grand*, la **transformation dite « de FISHER »**, définie par $z = \frac{1}{2}.\ln(\frac{1+r}{1-r})$ restaure la symétrie de la distribution de r tout en étirant l'étendue de variation des valeurs, *la variable ainsi obtenue convergeant lorsque l'hypothèse H_0 est vérifiée vers* **la loi normale** de moyenne $z_0 = \frac{1}{2}.\ln(\frac{1+r_0}{1-r_0})$ et de variance $1/n-3$. Si on considère, par exemple, le test bilatéral, et la fonction $\frac{z - z_0}{1/\sqrt{n-3}}$, on aboutit à la région critique définie par $z \notin \left]z_0 - \frac{t_\alpha}{\sqrt{n-3}}, z_0 + \frac{t_\alpha}{\sqrt{n-3}}\right[$, t_α vérifiant $\Pr ob(|\xi| \geq t_\alpha) = \alpha$, soit en *utilisant la transformation inverse* $r = \frac{e^{2.z} - 1}{e^{2.z} + 1} = th(z)$ *« tangente hyperbolique », la règle de décision en r.*

- Autre variante, plus utilisée, le *test de signification du degré de corrélation* suivant $\begin{cases} H_0 : r = 0 \\ H_1 : r \neq 0 \end{cases}$ **(test du r de PEARSON)**. On montre que pour le cas d'observations $(x_1, x_2, ..., x_n)$ et $(y_1, y_2, ..., y_n)$ *indépendantes et de distribution normale (plus précisément de loi du couple de type binormale)*, la loi de la variable $T = \frac{r.\sqrt{n-2}}{\sqrt{1-r^2}}$ est de type **STUDENT** à $v = n-2$ degrés de libertés.

Désignant par t_α le nombre vérifiant $Prob(|T| \geq t_\alpha) = \alpha$, où T est la variable de STUDENT en question, on a donc une *région critique* définie par $W = \{(x_1, x_2, ..., x_n) / T \notin]-t_\alpha, +t_\alpha[\}$. Il s'ensuit immédiatement le seuil $r \geq \dfrac{t_\alpha}{\sqrt{n-2+t_\alpha^2}}$.

e) Le cas des hypothèses composites

On se ramènera aux cas précédents pour lesquels H_0 est une hypothèse simple.

• Considérant ainsi le test $\begin{cases} H_0 : m \prec m_0 \\ H_1 : m \succ m_0 \end{cases}$, m étant la moyenne d'une loi normale $N(m, \sigma^2)$, on assimile le test en question à l'union des tests $\begin{cases} H_0' : m = m' \\ H_1 : m \succ m_0 \end{cases}$ (avec $m' \prec m_0$). La région critique ayant pour forme $\bar{x} \geq \pi$, le risque de première espèce $\alpha = Prob(\bar{x} \geq \pi / m = m')$ varie avec m', soit $\alpha(m')$ la fonction en question. Plus précisément, ce risque $\alpha(m')$ est maximal pour $m' = m_0$, soit $\alpha_0 = \underset{m' \prec m_0}{Sup} \alpha(m')$ la valeur ainsi obtenue. Dans ces conditions, la règle du test consiste à déterminer la région critique à partir de l'hypothèse nulle $H_0 : m = m_0$ pour laquelle l'erreur de première espèce α_0 est maximale. Cette règle sera admise comme valable pour les tests $\begin{cases} H_0 : m = m' \\ H_1 : m \succ m_0 \end{cases}$ (avec $m' \prec m_0$) puisque pour ces derniers, l'erreur de première espèce $\alpha(m')$ est en tout état de cause plus faible que α_0.

• Plus généralement, et toujours suivant l'hypothèse du test de la moyenne d'une loi normale, s'il s'agit d'un test *composite* $\begin{cases} H_0 : m_0 - \delta \leq m \leq m_0 + \delta \\ H_1 : m \notin]m_0 - \delta, m_0 + \delta[\end{cases}$, on construira la *région critique* suivant un *principe de symétrie*, à partir des bornes $m_0 - \lambda$ et $m_0 + \lambda$, l'erreur de première espèce α étant fonction quant à elle, des valeurs de m dans l'intervalle $[m_0 - \delta, m_0 + \delta]$. Or cette erreur est maximale aux bornes du domaine.

Cherchant donc λ, par exemple, à partir du test $\begin{cases} H_0 : m = m_0 + \delta \\ H_1 : m \notin [m_0 - \delta, m_0 + \delta] \end{cases}$, la donnée de l'erreur de première espèce α conduit à la relation :

$$\alpha = Prob(\bar{x} \geq m_0 + \lambda / m = m_0 + \delta) + Prob(\bar{x} \leq m_0 - \lambda / m = m_0 + \delta).$$

Par passage à la variable centrée réduite $\xi = \dfrac{\bar{x} - m}{\sigma / \sqrt{n}}$ et utilisant la fonction de répartition $\Pi(t) = Prob(\xi \leq t)$, on obtient le résultat :

$$\alpha = 1 - \Pi\left(\frac{m_0 + \lambda - m_0 - \delta}{\sigma / \sqrt{n}}\right) + \Pi\left(\frac{m_0 - \lambda - m_0 - \delta}{\sigma / \sqrt{n}}\right) \Rightarrow 1 - \alpha = \Pi\left(\frac{\sqrt{n}}{\sigma}.(\lambda - \delta)\right) + \Pi\left(-\frac{\sqrt{n}}{\sigma}.(\lambda + \delta)\right)$$

Une **méthode par approximations successives** fournit immédiatement la solution cherchée ici.

2.4 Tests de comparaison entre deux échantillons indépendants

La comparaison des résultats de deux groupes est naturellement une tâche fondamentale dans les études statistiques, épidémiologiques, sociologiques....

A cet effet, le schéma le plus classique, est de s'appuyer sur un critère quantitatif donné (par exemple, les résultats sportifs entre hommes et femmes) et de raisonner à travers deux échantillons $(X_1, X_2, ..., X_{n_X})$ et $(Y_1, Y_2, ..., Y_{n_Y})$ issus respectivement des deux populations considérées et formés pour chacun de variables aléatoires indépendantes et de même loi. En outre, on suppose dans cette partie que les X_i et les Y_j sont indépendantes deux à deux, c'est à dire l'*indépendance des observations entre les deux populations* en cause.

A la question de l'identité ou non des distributions des variables parentes X et Y dans chacune des deux populations considérées, les *tests non paramétriques* offrent une réponse sans faire d'hypothèse spécifique sur le type de loi considéré.

Autrement, la *comparaison de moyennes, proportions, et variances*, par le biais de *tests paramétriques*, reste fondamentale ici, *l'hypothèse d'échantillons de type gaussien* (**loi normale**) étant supposée vérifiée ci-après (encore que ce champ d'application peut être élargi lorsqu'on travaille sur de grands échantillons, $n_X \geq 30, n_Y \geq 30$, et que le *théorème central limite* est applicable).

Ainsi part-on de deux échantillons $(X_1, X_2, ..., X_{n_X})$ et $(Y_1, Y_2, ..., Y_{n_Y})$ de tailles respectives n_X et n_Y, et de lois parentes de type « **gaussien** », soient respectivement $N(m_X, \sigma_X)$ et $N(m_Y, \sigma_Y)$.

• Comme on va le constater dans le test t de STUDENT relatif aux moyennes, l'égalité ou non des variances σ_X^2 et σ_Y^2 (hypothèse dite « **d'homoscédasticité** ») tient une importance prépondérante, surtout si on ne dispose pas de grands échantillons. C'est donc plutôt par la comparaison des variances qu'il convient logiquement de commencer, en principe.

a) La comparaison de variances (test de FISHER-SNEDECOR)

Partant donc de deux échantillons $(X_1, X_2, ..., X_{n_X})$ et $(Y_1, Y_2, ..., Y_{n_Y})$ de tailles respectives n_X et n_Y, et de lois parentes $N(m_X, \sigma_X)$ et $N(m_Y, \sigma_Y)$, on va développer ici, par exemple, le test bilatéral $\begin{cases} H_0 : \sigma_X^2 = \sigma_Y^2 \\ H_1 : \sigma_X^2 \neq \sigma_Y^2 \end{cases}$.

A cet effet, la *fonction discriminante* à retenir est le rapport S_X^2 / S_Y^2 (resp. $S_X'^2 / S_Y'^2$ lorsque m_X et m_Y sont inconnues), les statistiques S_X^2 (resp. S_Y^2) et $S_X'^2$ (resp. $S_Y'^2$) étant définies par $S_X^2 = \frac{1}{n_X} \cdot \sum_{i=1}^{i=n_X} (X_i - m_X)^2$ et $S_X'^2 = \frac{1}{n_X} \cdot \sum_{i=1}^{i=n_X} (X_i - \overline{X})^2$ (idem pour S_Y^2 et $S_Y'^2$).

Les éléments justificatifs de ce choix sont exposés ci-après, à partir des résultats démontrés dans l'application 1.3 du chapitre I.

On montre en effet, que les statistiques $F=\dfrac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2}$ et $F=\dfrac{n_Y.S_X'^2/(n_X-1).\sigma_X^2}{n_Y.S_Y'^2/(n_Y-1).\sigma_Y^2}$ suivent respectivement les lois de FISHER SNEDECOR, $F(\nu_1,\nu_2)$ à $\nu_1=n_X,\nu_2=n_Y$ et $\nu_1=n_X-1,\nu_2=n_Y-1$ degrés de libertés suivant que (m_X,m_Y) sont connus ou non.

Sous l'hypothèse $H_0(\sigma_X^2=\sigma_Y^2)$, les rapports $F=\dfrac{S_X^2}{S_Y^2}$ et $F=\dfrac{S_X'^2}{S_Y'^2}$ suivent donc les lois $F(n_X,n_Y)$ et $F(n_X-1,n_Y-1)$.

• Dans les conditions susmentionnées et par exemple pour le cas où m_X et m_Y sont connues, la *région critique du test* s'écrit :

$$W=\left\{(x_1,x_2,...,x_n),(y_1,y_2,...,y_n)/F=\frac{S_X^2}{S_Y^2}\notin\left]\pi_1,\pi_2\right[\right\},$$

π_1 et π_2 étant définis par $\alpha=\Pr ob(F\le\pi_1 \text{ ou } F\ge\pi_2)$, soit $1-\alpha=\Pr ob(\pi_1\prec F\prec\pi_2)$.

Comme on peut le constater à travers les *tables de valeurs jointes* en annexes, la loi de F *n'est pas symétrique.* Mais, dans le cas du *test bilatéral* considéré ici, on se contentera pour trouver π_1 et π_2 de raisonner suivant la **symétrie des risques** (et non des valeurs), les risques $\Pr ob(F\le\pi_1)$ et $\Pr ob(F\ge\pi_2)$ étant donc supposés être égaux à $\dfrac{\alpha}{2}$.

La lecture dans la table des valeurs de $\pi_1/\Pr ob(F\le\pi_i)=\dfrac{\alpha}{2}$, et de $\pi_2/\Pr ob(F\ge F_2)=\dfrac{\alpha}{2}$ permet de conclure quant à la détermination de cette région critique et à la décision à retenir en conséquence.

• Bien évidemment, l'exposé ci-dessus est aisément transposable aux *tests unilatéraux* tels $\begin{cases}H_0:\sigma_X^2=\sigma_Y^2\\H_1:\sigma_X^2\succ\sigma_Y^2\end{cases}$ ou $\begin{cases}H_0:\sigma_X^2=\sigma_Y^2\\H_1:\sigma_X^2\succ\sigma_Y^2\end{cases}$, le seuil critique étant défini alors par $\alpha=\Pr ob(F\ge\pi)$ (resp. $\alpha=\Pr ob(F\le\pi)$).

b) La comparaison de moyennes (test t de STUDENT à deux échantillons)

Le cadre est toujours le même, à savoir deux échantillons indépendants $(X_1,X_2,...,X_{n_X})$ et $(Y_1,Y_2,...,Y_{n_Y})$, de lois parentes $N(m_X,\sigma_X)$ et $N(m_Y,\sigma_Y)$, le test considéré étant par exemple, le *test bilatéral* $\begin{cases}H_0:m_X=m_Y\\H_1:m_X\ne m_Y\end{cases}$.

S'inspirant des résultats de l'application 1.4 du chapitre I, on est conduit à choisir ici, la différence $\bar{x}-\bar{y}$ comme *fonction discriminante*, la *région critique* ayant donc pour forme $W=\left\{(x_1,x_2,...,x_n),(y_1,y_2,...,y_n)/\left|\bar{x}-\bar{y}\right|\ge\pi\right\}$ avec $\alpha=\Pr ob(\left|\bar{x}-\bar{y}\right|\ge\pi/m_X=m_Y)$.

• Lorsque σ_X **et** σ_Y **sont connus**, il est rappelé (suivant ladite application 1.4) que la *fonction pivotale* $\dfrac{\overline{X}-\overline{Y}-(m_X-m_Y)}{\sqrt{\dfrac{\sigma_X^2}{n_X}+\dfrac{\sigma_Y^2}{n_Y}}}$ suit la loi normale $N(0,1)$ ce qui permet de déterminer facilement π puisque, sous l'hypothèse H_0, on a $\pi = t_\alpha.\sqrt{\dfrac{\sigma_X^2}{n_X}+\dfrac{\sigma_Y^2}{n_Y}}$ où t_α est solution de l'équation $Prob(|\xi| \geq t_\alpha) = \alpha$ (ξ de loi $N(0,1)$).

• Lorsque σ_X **et** σ_Y **ne sont pas connus**, on est contraint de supposer $\sigma_X = \sigma_Y$ (*homoscédasticité*), du moins dans le *cas des petits échantillons* puisque, dans le cas contraire, le *théorème central limite* permet de se ramener au cas précédent avec $\pi = t_\alpha.\sqrt{\dfrac{\widehat{S_X}^2}{n_X}+\dfrac{\widehat{S_Y}^2}{n_Y}}$ (t_α lu dans la table des valeurs de la loi $N(0,1)$).

En effet, toujours suivant l'application 1.4 – chapitre I susmentionnée, la variable $Z = (n_X-1).\dfrac{\widehat{S_X}^2}{\sigma_X^2}+(n_Y-1).\dfrac{\widehat{S_Y}^2}{\sigma_Y^2}$ suit la loi du χ^2 à n_X+n_Y-2 degrés de libertés.

Il s'ensuit selon le théorème de FISHER dont la démonstration est faite dans l'application 1.1 du chapitre I, que la variable $T = \dfrac{U}{\sqrt{\dfrac{Z}{n_X+n_Y-2}}}$ où $U = \dfrac{\overline{X}-\overline{Y}-(m_X-m_Y)}{\sqrt{\dfrac{\sigma_X^2}{n_X}+\dfrac{\sigma_Y^2}{n_Y}}}$ est de loi normale centrée réduite $N(0,1)$, suit en définitive la loi de STUDENT à $\nu = n_X+n_Y-2$ degrés de libertés .

Or, dans $T = \dfrac{(\overline{X}-\overline{Y}-(m_X-m_Y)).\sqrt{n_X+n_Y-2}}{\sqrt{\dfrac{\sigma_X^2}{n_X}+\dfrac{\sigma_Y^2}{n_Y}}.\sqrt{\dfrac{(n_X-1).\widehat{S_X}^2}{\sigma_X^2}+\dfrac{(n_Y-1).\widehat{S_Y}^2}{\sigma_Y^2}}}$ les paramètres inconnus σ_X^2 et σ_Y^2 subsistent.

La façon de les supprimer est de supposer $\sigma_X^2 = \sigma_Y^2 = \sigma^2$, la statistique T se réduisant alors à $\dfrac{((\overline{X}-\overline{Y})-(m_X-m_Y)).\sqrt{n_X+n_Y-2}}{\sqrt{\dfrac{1}{n_X}+\dfrac{1}{n_Y}}.\sqrt{(n_X-1).\widehat{S_X}^2+(n_Y-1).\widehat{S_Y}^2}}$ et constituant la fonction pivotale cherchée.

• Rappelant que $m_X = m_Y$ sous l'hypothèse H_0 et désignant par t_α le nombre qui vérifie $Prob(|T| \geq t_\alpha) = \alpha$ (lu dans la table de STUDENT à $\nu = n_X+n_Y-2$ degrés de libertés → cf. annexes), on a donc immédiatement lorsque $\sigma_X = \sigma_Y$, le *seuil critique* π égal à la valeur $\dfrac{t_\alpha}{\sqrt{n_X+n_Y-2)}}.\sqrt{\dfrac{1}{n_X}+\dfrac{1}{n_Y}}.\sqrt{(n_X-1).\widehat{S_X}^2+(n_Y-1).\widehat{S_Y}^2}$.

• C'est donc préalablement sur un test de FISHER SNEDECOR qu'on s'appuiera pour vérifier l'hypothèse *d' homoscédasticité* ($\sigma_X^2 = \sigma_Y^2$), et appliquer le test ci-dessus (dit « **test t de STUDENT à deux échantillons** »), tout en précisant cependant qu'on admet ici une faute de raisonnement puisqu'il a déjà été souligné que le non rejet de l'hypothèse $\sigma_X^2 = \sigma_Y^2$ est loin de garantir l'égalité $\sigma_X^2 = \sigma_Y^2$. En fait, c'est au test $\begin{cases} H_0 : \sigma_X^2 \neq \sigma_Y^2 \\ H_1 : \sigma_X^2 = \sigma_Y^2 \end{cases}$ auquel il faudrait recourir pour s'assurer de l'hypothèse d'homoscédasticité, mais il est manifeste que sa construction serait très problématique. Bien entendu, les raisonnements précédents s'adaptent aisément aux tests unilatéraux $\begin{cases} H_0 : m_X = m_Y \\ H_1 : m_X \succ m_Y \end{cases}$ ou $\begin{cases} H_0 : m_X = m_Y \\ H_1 : m_X \prec m_Y \end{cases}$.

c) La comparaison de proportions

On considère une caractéristique donnée « C » en proportions respectives p_1 et p_2 au sein de deux populations P_1 et P_2 et deux échantillons de tailles respectives n_1 et n_2 extraits de ces populations. Notant par X_1 et X_2 les variables qui représentent l'occurrence de C dans ces deux échantillons, la question posée ici est de tester $H_0 : p_1 = p_2$ contre $H_1 : p_1 \neq p_2$ (voire unilatéralement $p_1 \succ p_2$ ou $p_1 \prec p_2$).

Se référant à l'application 1.5 du chapitre I, on choisit ici comme *fonction discriminante*, la différence des fréquences relatives $F_1 - F_2$ (avec $F_1 = \frac{X_1}{n_1}$ et $F_2 = \frac{X_2}{n_2}$). La *région critique* étant immédiatement définie par $|F_1 - F_2| \geq \pi / p_1 = p_2$, la statistique $\frac{F_1 - F_2 - (p_1 - p_2)}{\sqrt{\frac{p_1.(1-p_1)}{n_1} + \frac{p_2.(1-p_2)}{n_2}}}$ converge vers la *loi normale centrée réduite* dès que les conditions du théorème de MOIVRE – LAPLACE sont réunies (pratiquement, $n_1.p_1 \succ 5$ et $n_1.(1-p_1) \succ 5$, $n_2.p_2 \succ 5$ et $n_2.(1-p_2) \succ 5$).

Sous l'hypothèse $H_0 : p_1 = p_2 = p$, il s'ensuit immédiatement la valeur du seuil critique $\pi = t_\alpha . \sqrt{p.(1-p).(\frac{1}{n_1} + \frac{1}{n_2})}$ où t_α vérifie $Prob(|\xi| \geq t_\alpha) = \alpha$. Toutefois p étant inconnue, π reste indéfini ici. Mais dans la mesure où on suppose vérifiée la condition de convergence vers la loi normale et travailler sur de **grands échantillons** ($n_1 \geq 30, n_2 \geq 30$), on admettra d'approcher p par son *estimateur* $\hat{p} = \frac{X_1 + X_2}{n_1 + n_2}$.

• Dans le cas des **petits échantillons**, on utilisera une **méthode exacte** ou un *test non paramétrique* tel le test de χ^2 (cf. paragraphe 3 ci-après).

2.5 Tests de comparaison entre deux échantillons appariés

Dans ce paragraphe, on reprend les notations du paragraphe antérieur avec notamment les deux échantillons $(X_1, X_2, ..., X_{n_X})$ et $(Y_1, Y_2, ..., Y_{n_Y})$, mais c'est désormais sur les mêmes individus que portent les comparaisons dans les deux échantillons, ce qui suppose $n_X = n_Y = n$ et les variables X_i et Y_i non indépendantes.

Par exemple, un régime testé sur n individus est-il efficace ou non, le caractère mesuré avant régime (X) et après régime (Y) étant le poids de la personne considérée.

On utilisera le plus souvent un *test non paramétrique* qui offre la facilité d'être peu exigeant en hypothèses préalables mais dont en contrepartie la puissance est nettement inférieure aux *tests paramétriques*. Ces derniers restent cependant utilisables dans le cas d'**échantillons gaussiens** et d'une **comparaison de moyenne**, le **test t de STUDENT** offrant ainsi une réponse aisée à mettre en œuvre.

• Considérant par exemple, le test bilatéral $\begin{cases} H_0 : m_X = m_Y \\ H_1 : m_X \neq m_Y \end{cases}$, le principe en est ici de raisonner sur les différences $D_i = X_i - Y_i, (1 \leq i \leq n)$, variables aléatoires qu'on supposera indépendantes et normales, leur loi commune ayant pour moyenne $E(D_i) = E(X_i) - E(Y_i) = m_X - m_Y$.

Le test conduit donc à partir de *l'échantillon des différences* $d_i = x_i - y_i$, soit $(d_1, d_2, ..., d_n)$, à la *fonction discriminante* $\overline{d} = \dfrac{\sum_{i=1}^{i=n} d_i}{n}$ et à la statistique $\dfrac{\overline{D} - (m_X - m_Y)}{\widehat{S_D}/\sqrt{n}}$ dont la loi est de type STUDENT à $\nu = n - 1$ degrés de libertés. Il en résulte immédiatement $\pi = t_\alpha . \dfrac{\widehat{S_D}}{\sqrt{n}}$ où t_α vérifie $\Pr ob(|T| \geq t_\alpha) = \alpha$ (T étant la variable de STUDENT en question).

Bien évidemment, on peut décliner de façon semblable les *tests unilatéraux* $\begin{cases} H_0 : m_X = m_Y \\ H_1 : m_X \prec m_Y \end{cases}$ et $\begin{cases} H_0 : m_X = m_Y \\ H_1 : m_X \succ m_Y \end{cases}$. D'autre part, dans le cas des *grands échantillons*, on pourra se ramener à la loi normale par application du *théorème central limite*.

Enfin, il est précisé que la normalité et l'indépendance des D_i font appel à la loi normale bidimensionnelle non traitée dans cet ouvrage. Ce qui est certain, c'est qu'on n'a pas $Var(D_i) = Var(X_i) + Var(Y_i)$ (puisque X_i et Y_i ne sont plus indépendantes) et on peut même noter à cet égard que la variance de la différence $D_i = X_i - Y_i$ est d'autant plus faible que la relation entre X_i et Y_i est forte.

2.6 Tests de comparaison entre K échantillons indépendants ($K \succ 2$)

L'analyse de la variance (test ANOVA) relative à la **comparaison de moyennes**, et le **test de BARTLETT** portant sur la **comparaison des variances**, sont principalement présentés ici pour ce qui est des *tests paramétriques*, les autres méthodes faisant essentiellement appel quant à elles à des tests de mise en œuvre plus complexe ou à des tests non paramétriques.

a) L'analyse de la variance

Imaginée par FISHER et s'appuyant sur le test F (cf. paragraphe 2.3 précédent), la méthode en question vise, du moins sous sa forme simple à un facteur contrôlé, à tester l'hypothèse $H_0 : m_1 = m_2 = ... = m_K$ contre l'hypothèse $H_1 : \exists (i, j) / m_i \neq m_j$.

Dans ce test les $m_i (1 \le i \le K)$ désignent les moyennes inconnues d'une variable aléatoire donnée X (c'est le facteur en question) dans K populations différentes. Pour répondre à la question posée, on part de K échantillons $(x_{1,j}, x_{2,j}, ..., x_{n_j,j})$ de tailles respectives $n_j, (1 \le j \le K)$, les hypothèses de la **normalité** de X et de l'égalité des variances de X au sein des K populations considérées (**homoscédasticité**) étant supposées être *vérifiées* ici.

Ceci impose donc en amont l'usage d'un test $\begin{cases} H_0 : \sigma_1^2 = \sigma_2^2 = ... = \sigma_K^2 \\ H_1 : \exists (i,j) / \sigma_i^2 \neq \sigma_j^2 \end{cases}$ tel le test de BARTLETT développé ci-après ou d'autres tests (LEVENE, HARTLEY, ...).

- Le *principe de l'analyse de la variance* en est de décomposer la somme S_T des carrés des écarts entre les observations et la moyenne générale $\bar{x} = \frac{1}{N} . \sum_{i=1}^{i=K} \sum_{j=1}^{j=n_i} x_{i,j}$ où $N = \sum_{i=1}^{i=K} n_i$, suivant la somme :
  - des *écarts au sein d'une même population* (écarts dits « **intraclasses** »), soit $S_W = \sum_{i=1}^{i=K} \sum_{j=1}^{j=n_i} (x_{i,j} - \bar{x_i})^2$;
  - des *écarts entre la moyenne au sein de chacune des populations et la moyenne générale* $\bar{x}$ (écarts dits « **interclasses** »), soit $S_B = \sum_{i=1}^{i=K} n_i . (\bar{x_i} - \bar{x})^2$.

  En effet, la décomposition $x_{i,j} - \bar{x} = x_{i,j} - \bar{x_i} + \bar{x_i} - \bar{x}$ entraîne pour ce qui est de S_T la décomposition :

  $$S_T = \sum_{i=1}^{i=K} \sum_{j=1}^{j=n_i} (x_{i,j} - \bar{x})^2 = \sum_{i=1}^{i=K} \sum_{j=1}^{j=n_i} (x_{i,j} - \bar{x_i})^2 + \sum_{i=1}^{i=K} \sum_{j=1}^{j=n_i} (\bar{x_i} - \bar{x})^2 + 2. \sum_{i=1}^{i=K} \sum_{j=1}^{j=n_i} (x_{i,j} - \bar{x_i}) . (\bar{x_i} - \bar{x}) .$$

  Mais, $\sum_{i=1}^{i=K} \sum_{j=1}^{j=n_i} (\bar{x_i} - \bar{x})^2 = \sum_{i=1}^{i=K} n_i . (\bar{x_i} - \bar{x})^2$ et d'autre part $\sum_{i=1}^{i=K} \sum_{j=1}^{j=n_i} (x_{i,j} - \bar{x_i}) . (\bar{x_i} - \bar{x})$ est égal à $\sum_{i=1}^{i=K} (\bar{x_i} - \bar{x}) \sum_{j=1}^{j=n_i} (x_{i,j} - \bar{x_i})$, soit la valeur nulle puisque $\sum_{j=1}^{j=n_i} (x_{i,j} - \bar{x_i}) = n_i . \bar{x_i} - n_i . \bar{x_i} = 0$.

  Ainsi obtient-on la décomposition attendue $S_T = S_W + S_B$.

- Or, on montre d'une part que $\frac{S_W}{N-K}$ (avec $N = \sum_{i=1}^{i=K} n_i$) forme un estimateur sans biais de σ^2 quelque soit l'hypothèse considérée et d'autre part que $\frac{S_B}{K-1}$ est également un estimateur sans biais de σ^2 lorsque l'hypothèse H_0 est vérifiée.

Dans ces conditions, *sous l'hypothèse* H_0, *le rapport* $F = \frac{S_B / K-1}{S_W / N-K}$ *est proche de* 1, ce même quotient étant supérieur à 1 dans l'hypothèse altemative H_1.

En outre, on montre que F suit **la loi de FISHER SNEDECOR**, à $\nu_1 = K-1$ et $\nu_2 = N-K$ degrés de libertés.

Ainsi désignant par F_α le seuil vérifiant $Prob(F \geq F_\alpha) = \alpha$ pour la loi en question $F(K-1, N-K)$, peut-on résumer la règle de décision du test considéré au choix de H_1 si $F_{obs} \geq F_\alpha$ (*région critique*) et de H_0 sinon. Pour rappel, le test se résume à $\begin{cases} H_0 : F = 1 \\ H_1 : F > 1 \end{cases}$.

S'inspirant de l'application 1.4 du chapitre I, les quantités $s_i^2 = \frac{1}{n_i - 1} . \sum_{j=1}^{j=n_i} (x_{i,j} - \bar{x}_i)^2$ constituent des estimateurs sans biais des variances σ_i^2 de X dans chacune des K populations considérées. La statistique $\frac{S_W}{N-K} = \sum_{i=1}^{i=K} \frac{(n_i - 1).s_i^2}{N-K}$ (avec $N = \sum_{i=1}^{i=K} n_i$) vérifie immédiatement par linéarité $E\left[\frac{S_W}{N-K}\right] = \sum_{i=1}^{i=K} \frac{(n_i - 1)}{N-K} . E(s_i^2) = \sum_{i=1}^{i=K} \frac{(n_i - 1).\sigma_i^2}{N-K}$ (les estimateurs corrigés s_i^2 étant sans biais).

Suivant l'hypothèse d'homoscédasticité ($\sigma_1^2 = \sigma_2^2 .... = \sigma_K^2 = \sigma^2$), on a donc immédiatement $E\left[\frac{S_W}{N-K}\right] = \frac{\sum_{i=1}^{i=K} (n_i - 1).\sigma^2}{N-K} = \sigma^2$ ce qui établit la première partie des résultats présentés ci-dessus.

Quant à la somme S_B des écarts « interclasses » qui est égale à $\sum_{i=1}^{i=K} n_i . (\bar{x}_i - \bar{x})^2$, soit par développement, $\sum_{i=1}^{i=K} n_i . \bar{x}_i^2 - 2.\bar{x} . \sum_{i=1}^{i=K} n_i . \bar{x}_i + \sum_{i=1}^{i=K} n_i . \bar{x}^2 = \sum_{i=1}^{i=K} n_i . \bar{x}_i^2 - 2.\bar{x} . (N.\bar{x}) + N.\bar{x}^2$, soit en définitive $S_B = \sum_{i=1}^{i=K} n_i . \bar{x}_i^2 - N.\bar{x}^2$, son espérance mathématique s'écrit, par linéarité, $E(S_B) = \sum_{i=1}^{i=p} n_i . E(\bar{x}_i^2) - N.E(\bar{x}^2)$.

Mais, de la relation $Var(\bar{x}_i) = \frac{\sigma_i^2}{n_i} = E(\bar{x}_i^2) - \left[E(\bar{x}_i)\right]^2$, $1 \leq i \leq K$, résulte l'expression $E(\bar{x}_i^2) = \frac{\sigma_i^2}{n_i} + m_i$, soit compte tenu de *l'hypothèse d'homoscédasticité* ci-dessus, $E(\bar{x}_i^2) = \frac{\sigma^2}{n_i} + m_i$.

De même $E(\bar{x}^2) = Var(\bar{x}) + E(\bar{x})^2$ avec d'une part, $Var(\bar{x}) = Var(\sum_{i=1}^{i=K} \frac{n_i . \bar{x}_i}{N})$, soit par *pseudo linéarité* de la variance, $Var(\bar{x}) = \frac{1}{N^2} . \sum_{i=1}^{i=K} n_i^2 . Var(\bar{x}_i) = \frac{1}{N^2} . \sum_{i=1}^{i=K} n_i^2 . \frac{\sigma^2}{n_i} = \sigma^2 . \frac{\sum_{i=1}^{i=K} n_i}{N^2} = \frac{\sigma^2}{N}$; et d'autre part $E(\bar{x}) = \sum_{i=1}^{i=K} \frac{n_i . m_i}{N} = m$ (toujours suivant l'*homoscédasticité* « égalité des variances σ_i^2 »).

Des développements précédents, résulte à partir de $E(S_B)=\sum_{i=1}^{i=K} n_i.E(\overline{x_i}^2)-N.E(\overline{x}^2)$, la relation $E(S_B)=\sum_{i=1}^{i=K} n_i.\frac{\sigma^2}{n_i}+\sum_{i=1}^{i=K} n_i.m_i^2-N.\frac{\sigma^2}{N}-N.m^2$. Lorsque l'hypothèse H_0 est vérifiée ($m_1=m_2...=m_K=m$), il est immédiat que $E(S_B)=(K-1).\sigma^2$ ce qui établit la propriété annoncée d'être sans biais pour ce qui est de l'estimateur de $\sigma^2 \rightarrow \frac{S_B}{K-1}$.

Revenant à $S_W=\sum_{i=1}^{i=K}(n_i-1).s_i^2$ et utilisant les résultats de l'application 1.4 du chapitre I, on remarque que les variables indépendantes $V_i=\frac{(n_i-1).s_i^2}{\sigma_i^2}$ suivent des lois de χ^2 à $\nu_i=n_i-1$ degrés de libertés ($1\le i\le K$). Ainsi, la somme $\sum_{i=1}^{i=K} V_i$ suit-elle la loi du χ^2 à $\nu=\sum_{i=1}^{i=K}\nu_i$ degrés de libertés, soit $\nu=N-K$ degrés de libertés. Rappelant l'hypothèse d'homoscédaticité ($\sigma_1^2=\sigma_2^2...=\sigma_K^2=\sigma^2$), la loi de $\frac{S_W}{\sigma^2}=\sum_{i=1}^{i=K}\frac{(n_i-1).s_i^2}{\sigma^2}$ est donc la loi $\chi^2(N-K)$. De même, suivant les résultats de l'application 1.1 du chapitre I (théorème de FISHER), le rapport $\frac{S_B}{\sigma^2}$ suit la loi $\chi^2(K-1)$ (loi du chi- deux à $K-1$ degrés de libertés).

Par quotient, le rapport $F=\frac{S_B/(K-1).\sigma^2}{S_W/(N-K).\sigma^2}$ suit la loi de FISHER SNEDECOR, $F(K-1,N-K)$ (cf. application 1.3 du chapitre I), ce qui justifie la fonction $F=\frac{S_B/K-1}{S_W/N-K}$ employée dans le présent test ANOVA et la procédure de construction de la région critique précédemment exposée.

• On remarquera que le rejet de l'hypothèse $H_0: m_1=m_2...=m_K$, ne fournit pas **l'identification des moyennes qui sont différentes**. Plusieurs méthodes permettent de traiter ces comparaisons multiples à commencer par un *test t* de STUDENT appliqué à *chaque comparaison deux à deux des moyennes* m_i.

Entre autres, la **méthode des contrastes de SCHEFFE** offre une solution souple dans le prolongement de l'analyse de la variance. On appelle **contraste** entre les moyennes $m_i(1\le i\le K)$, toute combinaison linéaire $C=\sum_{i=1}^{i=K} C_i.m_i$ dans laquelle la somme des coefficients C_i est nulle ($\sum_{i=1}^{i=K} C_i=0$). En outre, on rajoutera fréquemment la condition $\sum_{i=1}^{i=K}|C_i|=2$ pour homogénéiser les coefficients).

Par exemple, la série des C_i, $(\frac{1}{2},\frac{1}{2},-\frac{1}{2},-\frac{1}{2},0,...,0)$ permet de comparer m_1+m_2 à $m_3+m_4,...$

Le contraste est jugé *significatif* si $\left|\widehat{C}\right| \geq \pi = \sqrt{(K-1).F_\alpha.\frac{S_W}{N-K}.(\sum_{i=1}^{i=K}\frac{C_i^2}{n_i})}$ où $\widehat{C} = \sum_{i=1}^{i=K} C_i.\overline{x_i}$, F_α étant le seuil vérifiant $\Pr ob(F \geq F_\alpha) = \alpha$ où F suit la loi de FISHER SNEDECOR, $F(K-1, N-K)$ et où n_i désigne la taille de l'échantillon considéré dans la population $i, 1 \leq i \leq K$.

b) Le test de comparaison des variances de BARTLETT

Applicable dans les conditions du paragraphe précédent, à savoir la **normalité** des échantillons extraits dans chacune des K populations, échantillons de tailles n_i $(1 \leq i \leq K)$ égales ou non, le **test de BARTLETT** qui répond au choix de $H_0 : \sigma_1^2 = \sigma_2^2 ... = \sigma_K^2$ contre $H_1 : \exists (i,j) / \sigma_i^2 \neq \sigma_j^2$, repose sur la statistique :

$$B = \frac{(N-K).\ln(\frac{S_W}{N-K}) - \sum_{i=1}^{i=K}(n_i - 1)\ln S_i^2}{1 + \frac{1}{3.(K-1)}.\left[\sum_{i=1}^{i=K}\frac{1}{n_i - 1} - \frac{1}{N-K}\right]}$$

(avec les notations du paragraphe antérieur et $S_i^2 = \frac{1}{n_i - 1}.\sum_{j=1}^{j=n_i}(X_{i,j} - \overline{X_i})^2, 1 \leq i \leq K$.

En effet, on montre que, sous l'hypothèse H_0, la variable *la variable B suit sensiblement une loi du* χ^2 à $\nu = K-1$ degrés de libertés. Dès lors, la mise en œuvre du test est simple et se résume à déterminer, dans les tables de valeurs annexées, le seuil χ_α^2 vérifiant $\Pr ob(\chi^2 \geq \chi_\alpha^2) = \alpha$,(α erreur de première espèce donnée). Notant par b, la valeur calculée de B, la *règle de décision* est donc de décider H_1 si $b \geq \chi_\alpha^2$ et H_0 sinon.

2.7 Tests progressifs

Dans ces tests, la *taille n de l'échantillon* est supposée **aléatoire** en fonction des observations effectuées, pratique intéressante au regard des économies qu'elle peut permettre de faire sur l'échantillonnage.

En effet, considérons par exemple, un test de conformité qui, portant sur une proportion et des échantillons de taille 25, conduirait à rejeter ladite conformité si plus de 10 pièces défectueuses sont constatées parmi les 25 pièces prélevées.

Supposant ces pièces tirées une à une et examinées aussitôt, il est bien évident que si un tirage conduisait à 9 pièces incorrectes au bout de 9 prélèvements, il y aurait fort à parier qu'il soit inutile d'attendre le 25ème tirage pour conclure à la non conformité de la fabrication.

Ainsi, la taille de l'échantillon est-elle guidée par les résultats obtenus.

Une **méthode progressive** relativement au test d'une hypothèse "H" est donc caractérisée par une règle de décision qui, après chaque observation, permet de statuer entre :

- accepter H;
- rejeter H ;
- effectuer une observation supplémentaire.

• Mis au point par **WALD**, le test considéré ici, relativement au paramètre θ d'une distribution donnée, a pour forme $\begin{cases} H_0 : \theta \leq \theta_0 \\ H_1 : \theta \geq \theta_1 \end{cases}$, avec $\theta_1 \succ \theta_0$, test s'appuyant sur le test entre deux hypothèses simples $\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta = \theta_1 \end{cases}$.

En effet, il est montré que R_n désignant le *rapport des fonctions de vraisemblance* égal à $\dfrac{L(x_1, x_2, ..., x_n, \theta_1)}{L(x_1, x_2, ..., x_n, \theta_0)}$, on a relativement au test composite $\begin{cases} H_0 : \theta \leq \theta_0 \\ H_1 : \theta \geq \theta_1 \end{cases}$, les résultats :

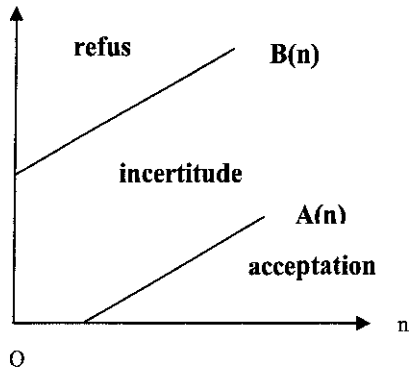
- pour $R_n \prec \dfrac{\beta}{1-\alpha}$, un risque d'erreur inférieur à β si on retient l'hypothèse H_0 alors que H_1 est vraie ;
- pour $R_n \succ \dfrac{1-\beta}{\alpha}$, un risque d'erreur inférieur à α, si on retient l'hypothèse H_1 alors que H_0 est vraie.

Concrètement, on calculera successivement les rapports $R_1, R_2, ..., R_n$ après chaque observation et on les comparera aux seuils ci-dessus pour conclure entre H_0 et H_1, et la poursuite des prélèvements. Cette procédure est imagée ci-dessous dans le cadre d'un test entre deux hypothèses simples, portant sur la moyenne d'une loi normale (σ supposé connu).

Dans les conditions ci-dessus et pour le test $\begin{cases} H_0 : m = m_0 \\ H_1 : m = m_1 \end{cases}$ (avec $m_1 \succ m_0$), le rapport R_n a pour expression $\exp\left[-\dfrac{1}{2.\sigma^2}.\sum_{i=1}^{i=n}\left[(x_i - m_1)^2 - (x_i - m_0)^2\right]\right]$, soit par développement, $R_n = \exp\left[\dfrac{(m_1 - m_0)}{\sigma^2}.(\sum_{i=1}^{i=n} x_i - n.\dfrac{m_0 + m_1}{2})\right]$. La double inégalité $\dfrac{\beta}{1-\alpha} \leq R_n \leq \dfrac{1-\beta}{\alpha}$ qui correspond à la zone d'incertitude, s'écrit donc, eu égard à l'expression de R_n :

$$n.q - h_1 \leq \sum_{i=1}^{i=n} x_i \leq n.q + h_2$$

avec $q = \dfrac{m_0 + m_1}{2}, h_1 = \dfrac{\sigma^2}{(m_1 - m_0)}.\ln(\dfrac{1-\alpha}{\beta}), h_2 = \dfrac{\sigma^2}{(m_1 - m_0)}.\ln(\dfrac{1-\beta}{\alpha})$.



Posant $A(n) = n.q - h_1$ et $B(n) = n.q + h_2$, la mise en œuvre du test consiste à tracer, dans un repère où les valeurs de n sont portées en abscisse, les droites d'équations $A(n)$ et $B(n)$.

Considérant, par exemple, le point représentatif du $p^{ième}$ prélèvement, point de coordonnées $(p, \sum_{i=1}^{i=p} x_i)$ dans le repère ci-contre, on conclura à :

- H_0 : si le point en question vient en dessous de la droite $A(n)$;

- H_1 : si le point en question vient au dessus de la droite $B(n)$;

- l'incertitude, c'est-à-dire la poursuite des prélèvements dans les autres cas.

3. Les tests non paramétriques

Peu exigeants quant aux hypothèses de validité hormis *l'indépendance des observations* (ce n'est pas cependant une condition systématique), ils font *abstraction de la condition de normalité* très présente dans les tests paramétriques pour le cas des petits échantillons et ils trouvent donc en ce domaine ($n \prec 30$), un terrain d'applications privilégié.

D'une grande variété de choix, les **tests non paramétriques** dont la mise en œuvre est, en outre, souvent simple, touchent un *champ très large de domaines*, étant notamment incontournables lorsque les *données ne sont pas quantitatives* (tests qualitatifs).

A noter néanmoins, que leur puissance est inférieure à celle des tests paramétriques lorsque ceux-ci sont applicables et que l'hypothèse de normalité est satisfaite.

3.1 Tests d'adéquation

A partir d'une *distribution empirique donnée* caractérisée par les valeurs x_i d'une variable aléatoire X, regroupées en classes ou non, et leurs fréquences absolues associées N_i, la question qui est posée ici est d'admettre ou non *l'ajustement de la variable X par une loi théorique donnée,* discrète ou continue.

Le **test de chi- deux** constitue le plus connu des outils utilisables ici. Toutefois, dans le cas de données quantitatives, il pourra utilement être substitué par d'autres tests tels le test de KOLMOGOROV, tous deux construits autour de la notion de *distance entre distributions empiriques et théoriques*, ou s'agissant de la loi normale, le test de SHAPIRO et WILK et la méthode graphique de la droite de HENRY.

a) Le test d'adéquation du chi- deux(χ^2)

Les données relatives à X sont réparties en *classes* "i", ($1 \le i \le m$) comme le montre le tableau ci-dessous, chacune d'entre elles correspondant à une valeur x_i ou à un intervalle $[x_i, x_{i+1}[$. Les *fréquences absolues empiriques* (observées), soient N_i, et les *fréquences théoriques* induites par la loi par laquelle on teste l'ajustement ($N.p_i$ avec $p_i = \Pr ob(X = x_i)$ ou $p_i = \Pr ob(x_i \prec X \le x_{i+1})$ et $N = \sum_{i=1}^{i=p} N_i$) sont mentionnées face à face dans ce tableau.

Classe ($1 \le i \le m$)	Effectif observé	Probabilité théorique	Effectif théorique
1	N_1	p_1	$N.p_1$
.	.	.	.
i	N_i	p_i	$N.p_i$
.	.	.	.
m	N_m	p_m	$N.p_m$
Total	N	1	N

Le test est donc ici : $\begin{cases} H_0 : \text{ on accepte l'ajustement} \\ H_1 : \text{ on refuse l'ajustement} \end{cases}$.

• A cet effet, **PEARSON** mesure la *distance D entre la loi théorique et la loi empirique* (observée) par la quantité $D = \sum_{i=1}^{i=m} \frac{(N_i - N.p_i)^2}{N.p_i}$, dite « **distance du chi- deux** » et dont on montre qu'elle *converge, pour N grand et lorsque H_0 est vraie*, vers la *loi du chi- deux* à $m-1$ degrés de libertés, soit $\chi^2(m-1)$.

Dans ces conditions, et définissant la *région critique* (zone de choix de H_1) sous la forme $W = \left\{(x_1, x_2, ..., x_n) / D \geq \chi_\alpha^2\right\}$, la donnée de l'erreur de première espèce α permet immédiatement de déterminer le seuil χ_α^2 à partir de l'équation $Prob(\chi^2 \geq \chi_\alpha^2) = \alpha$ où χ^2 est la variable du chi- deux à $m-1$ degrés de libertés (cf. tables de valeurs annexées).

En bref, on rejette H_0 si $D_{calculé} \geq \chi_\alpha^2$ (décision H_0 retenue dans le cas contraire).

• Cependant, la convergence de la loi de D vers la loi du chi- deux, n'est admise que si les effectifs théoriques $N.p_i$ sont tous supérieurs à 5, cette convergence étant d'autant mieux satisfaite que ces effectifs $N.p_i$ sont grands. A cet égard, on **regroupera** donc les *classes d'effectifs inférieurs à 5* avec les classes adjacentes, *la valeur du nombre de degrés de libertés $\nu = m-1$ étant à apprécier une fois ces regroupements préalables effectués.*

• Enfin, si dans la détermination de la loi théorique, il est nécessaire d'estimer k paramètres, il convient de *diminuer d'autant le nombre de degrés de libertés*, égal en définitive à $\nu = (m-1) - k$.

Par exemple, l'approximation par la loi de POISSON, entraîne $k = 1$.

b) Le test de KOLMOGOROV

Préférable au test de chi- deux lorsque la distribution avec laquelle on teste l'ajustement est une *loi continue* (dont notamment la loi normale), ce test est basé sur la **comparaison des fonctions de répartition.**

Partant d'un échantillon indépendant $(X_1, X_2, ..., X_n)$ de taille n et de variable parente X (de fonction de répartition inconnue $F(x)$) et d'une loi donnée de fonction de répartition $F_0(x)$, on se propose de tester : $\begin{cases} H_0 : F(x) = F_0(x) \\ H_1 : F(x) \neq F_0(x) \end{cases}$.

Pour cela, les valeurs de l'échantillon étant *rangées par ordre croissant*, soient $X(i)$ les **statistiques d'ordre** ainsi obtenues, on compare $F_0(x)$ à la *fonction de répartition empirique* $\widehat{F}(x)$, fonction en escalier définie par :

$$\widehat{F} = \begin{cases} 0 \ \ si \ \ x \prec X(i) \\ i/n \ \ si \ \ X(i) \leq x \prec X(i+1) \\ 1 \ \ si \ \ X \geq X(n) \end{cases}$$

Cette comparaison a lieu par l'intermédiaire de la **distance de KOLMOGOROV** , soit $D(F_0,\widehat{F}) = \underset{x}{Sup}\left(\left|F_0(x)-\widehat{F}(x)\right|\right)$, dont s'agissant de données sous forme d'échantillon (voir applications pour des données réparties en classes), l'expression est:

$$D(F_0,\widehat{F}) = \underset{i=1,2,...,n}{Sup}\left\{\left|F_0(X(i))-\frac{i}{n}\right|,\left|F_0(X(i))-\frac{i-1}{n}\right|\right\}.$$

Sous l'hypothèse H_0 (soit $F(x)=F_0(x)$), la loi de $D(F_0,\widehat{F})$ ne dépend pas de F_0. En effet, suivant un résultat relevant du calcul des probabilités et à la base de la simulation stochastique, les variables $F_0(X(i))$ sont *uniformes* sur $[0,1]$. Plus précisément, on montre que pour n *assez grand*, $\Pr ob(\sqrt{n}.D(F_0,\widehat{F}) \geq t)$ est proche de $2.\sum_{k=1}^{k=+\infty}(-1)^{k+1}.\exp(-2.k^2.t^2)$, série qui converge très rapidement.

Des éléments ci-dessus, résulte la valeur du seuil $D_{\alpha,n}$ tel que $\Pr ob(D \geq D_{\alpha,n}) = \alpha$. En effet, en notant par t_α la valeur de $t \backslash 2.\sum_{k=1}^{k=+\infty}(-1)^{k+1}.\exp(-2.k^2.t^2) = \alpha$, il en résulte aisément $D_{\alpha,n} = \frac{t_\alpha}{\sqrt{n}}$ (t_α indépendant de n). Ce résultat, valable dès que n est assez grand ($n \geq 40$) est numériquement fourni par la *table des valeurs de la fonction* t_α de KOLMOGOROV (cf. annexes), table de laquelle ressortent pour les valeurs les plus courantes du risque de première espèce α, les données :

$D_{\alpha,n} \backslash \alpha$	0,20	0,15	0,10	0,05	0,01
	$1,07/\sqrt{n}$	$1,14/\sqrt{n}$	$1,22/\sqrt{n}$	$1,36/\sqrt{n}$	$1,63/\sqrt{n}$

• Pour le cas des petites valeurs de $n, (n \prec 40)$, la table fournit en fonction de n, la valeur du seuil $D_{\alpha,n}$ vérifiant $\Pr ob(D \geq D_{\alpha,n}) = \alpha$. Ces valeurs sont indiquées ici pour le test bilatéral $\begin{cases} H_0 : F = F_0 \\ H_1 : F \neq F_0 \end{cases}$ mais il existe aussi des tables dans le cas des tests unilatéraux $\begin{cases} H_0 : F = F_0 \\ H_1 : F \succ F_0 \end{cases}$ ou $\begin{cases} H_0 : F = F_0 \\ H_1 : F \prec F_0 \end{cases}$.

En conclusion, sur la base de la *région critique* $W = \left\{(x_1,x_2,...,x_n)/D(F_0,\widehat{F}) \geq D_{\alpha,n}\right\}$ on appliquera, s'agissant du **test de KOLMOGOROV**, la **règle de décision** :

- $D_{calculé} \geq D_{\alpha,n} \Rightarrow$ on décide H_1 , c'est-à-dire le *rejet de l'ajustement*;
- $D_{alculé} \prec D_{\alpha,n} \Rightarrow$ *on ne rejette pas l'ajustement.*

c) Le test de normalité de SHAPIRO et WILK

Applicable à une série d'observations indépendantes $(x_1,x_2,...,x_n)$ d'une variable *quantitative* dont on teste la **normalité** ou non et valable pour des *tailles n d'échantillons, relativement faibles* $(5 \leq n \leq 50)$, le test de SHAPIRO-WILK, dont il faut souligner la *puissance élevée*, est basé sur le **rapport** W **de deux estimations liées à la variance** dont provient l'échantillon :

- l'une fonction des *étendues partielles* $x_n - x_1, x_{n-1} - x_2, \ldots$, les données x_i ayant préalablement été classées par ordre croissant $(x_1 \leq x_2 \leq \ldots \leq x_n)$;
- l'autre égale à $(n-1).s^2 = \sum_{i=1}^{i=n} (x_i - \bar{x})^2$.

• Pratiquement, les étapes suivantes sont à mettre en œuvre :

Etape 1 : Classer les données x_i par *ordre croissant* ($\Rightarrow$ statistique d'ordre $x_1 \leq x_2 \leq \ldots \leq x_n$).

Etape 2 : Calculer $T_n = \sum_{i=1}^{i=n} (x_i - \bar{x})^2$ avec $\bar{x} = \frac{\sum_{i=1}^{i=n} x_i}{n}$.

Etape 3 : Calculer les *différences* $d_1 = x_n - x_1, d_2 = x_{n-1} - x_2, \ldots, d_i = x_{n-i+1} - x_i, \ldots$ Si n est *pair* on formera ainsi $\frac{n}{2}$ différences, et si n est *impair*, on se contentera de $\frac{n-1}{2}$ différences, l'observation médiane n'étant pas utilisée alors.

Etape 4 : On calcule la quantité $W = \frac{b^2}{T_n}$ où $b = \sum_{i=1}^{i=n} a_i.d_i$, les *coefficients* a_i étant fournis en fonction de n et de i $(5 \leq n \leq 50)$ par une *table de valeurs* jointe (cf. annexes).

Etape 5 : On compare W à un *seuil* W_α lu en fonction de α et de n dans la *table de* SHAPIRO-WILK jointe également en annexes. D'où, la **règle de décision** :

→ Si $W \succ W_\alpha$, la normalité de la distribution n'est pas rejetée ;

→ Si $W \leq W_\alpha$, l'hypothèse de normalité est rejetée.

d) La méthode graphique de la droite de HENRY (normalité)

D'une **grande simplicité**, cette *méthode graphique* part du résultat suivant lequel, sous l'hypothèse de la normalité de X, la variable $\xi = \frac{X - m}{\sigma}$ suit la loi normale, centrée, réduite de type $N(0,1)$.

Ainsi, pour tout x, a-t-on $\Pr ob(X \leq x) = \Pr ob(\xi \leq t = \frac{x - m}{\sigma}) = \Pi(t)$, $\Pi(t)$ étant la fonction de répartition de la variable ξ dont la table des valeurs bien connue est annexée).

En d'autres termes, et partant de la fonction de répartition empirique $\hat{F}(x)$ associée aux données $(x_1, x_2, \ldots, x_n)$ qu'on aura *préalablement classées par ordre croissant*, ici encore, si aux valeurs de $\hat{F}(x_i), (1 \leq i \leq n)$, on associe les nombres t_i solutions des équations $\hat{F}(x_i) = \Pi(t_i)$, la relation qui lie les t_i aux x_i est **linéaire** dès lors que la variable aléatoire X suit la loi normale (puisqu'en théorie, on doit avoir $t_i = \frac{x_i - m}{\sigma}$.

Il s'ensuit le mode opératoire suivant de la droite de HENRY :

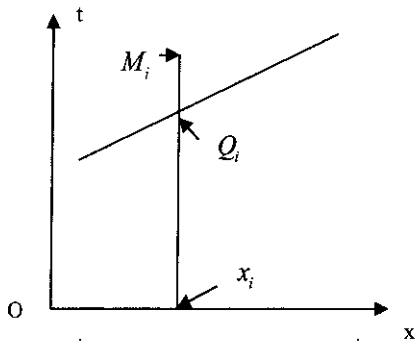
Etape 1 : Classer les données x_i par *ordre croissant* ($\Rightarrow$ statistique d'ordre $x_1 \leq x_2 \leq ... \leq x_n$) et exprimer, point par point, la fonction de répartition empirique $\widehat{F}(x)$ (cf. la définition donnée au paragraphe b) précédent). A noter cependant, que le plus souvent, les données auront préalablement été réparties en classes, ce qui revient à calculer les fréquences cumulées (cf. exemple en applications ci-après).

Etape 2 : Associer aux valeurs x_i, les nombres t_i vérifiant les relations $\widehat{F}(x_i) = \Pi(t_i)$.

Etape 3 : Dresser dans un repère (x,t) les points $M_i(x_i,t_i)$.

Etape 4 : Si ces points M_i sont sensiblement **alignés**, la **normalité** de X *est vérifiée* (ajustement rejeté dans le cas contraire). La droite en question, dite « *droite de HENRY* » qui sera tracée de façon à passer le plus près possible des n points $M_i(x_i,t_i)$ et dont *en théorie* l'équation est $t = \dfrac{x-m}{\sigma}$, coupe donc l'axe des x au point de coordonnées $(m,0)$.

On en déduit donc, à travers cette intersection avec l'axe des abscisses, une évaluation de $m = E(X)$. De même, la pente de droite qui est en théorie l'inverse de σ, fournit en conséquence une évaluation graphique de $\sigma(X)$.



Le test de KOLMOGOROV peut utilement appuyer les conclusions qu'inspire la droite de HENRY. Il suffit pour cela de repérer le point M_i le plus éloigné de la droite en question, S désignant la distance correspondante, soit $S = M_iQ_i$.

Si $S \geq D_{\alpha,n}$ ($D_{\alpha,n}$ étant le seuil critique du test de KOLMOGOROV dont la détermination a déjà été présentée au paragraphe b)), on rejette l'hypothèse de normalité (du moins au voisinage de la valeur x_i).

Dans le cas contraire, l'écart en question est jugé ne pas remettre en cause la conclusion de normalité.

3.2 Tests de comparaison entre échantillons indépendants

Quelques cas parmi les tests les plus connus sont présentés ci-après, une distinction étant opérée entre le cas de deux populations ($K = 2$) et celui d'un nombre supérieur de populations ($K > 2$). A noter que le test du chi- deux est également très présent pour les comparaisons entre deux échantillons notamment pour le cas de variables discrètes, comme cela pourra être constaté à travers les applications ci-après.

a) Le test de KOLMOGOROV-SMIRNOV pour l'identité de deux distributions

Variante du test de KOLMOGOROV précédent, le présent **test de KOLMOGOROV-SMIRNOV** a pour objet de tester $\begin{cases} H_0 : F_X = F_Y \\ H_1 : F_X \neq F_Y \end{cases}$ dans les conditions ci-après :

On part de deux échantillons indépendants de variables parentes X et Y et de tailles respectives n_X et n_Y, F_X et F_Y étant les fonctions de répartition de X et de Y. Considérant les *fonctions de répartition empiriques* $\widehat{F_X}$ et $\widehat{F_Y}$ définies dans les conditions déjà exposées au paragraphe 3.1.b) précédent, on utilise cette fois la **distance** :

$$D(\widehat{F_X},\widehat{F_Y}) = \underset{x\in R}{Sup}\left|\widehat{F_x}(x)-\widehat{F_y}(x)\right|$$

(resp. $D(\widehat{F_X},\widehat{F_Y}) = \underset{x\in R}{Sup}(\widehat{F_X}(x)-\widehat{F_Y}(x))$ ou $D(\widehat{F_X},\widehat{F_Y}) = \underset{x\in R}{Sup}(\widehat{F_Y}(x)-\widehat{F_X}(x))$ lorsqu'il s'agit des *tests unilatéraux* $\begin{cases}H_0 : F_X = F_Y\\ H_1 : F_X \succ F_Y\end{cases}$ et $\begin{cases}H_0 : F_X = F_Y\\ H_1 : F_X \prec F_Y\end{cases}$).

• Or, sous l'hypothèse H_0, et pour n_X et n_Y grands, pratiquement $n_X \geq 40, n_Y \geq 40$, on montre que $\Pr ob(\sqrt{\frac{n_X.n_Y}{n_X+n_Y}}.D(\widehat{F_X},\widehat{F_Y}) \geq t)$ converge également vers la série $2.\sum_{k=1}^{k=+\infty}(-1)^{k+1}.\exp(-2.k^2.t^2)$. Notant donc par D_{α,n_X,n_Y} le seuil vérifiant la relation $\Pr ob(D(\widehat{F_X},\widehat{F_Y}) \geq D_{\alpha,n_X,n_Y}$, il vient immédiatement $D_{\alpha,n_X,n_Y} = \sqrt{\frac{n_X+n_Y}{n_X.n_Y}}t_\alpha$ (t_α ayant pour ce qui est du test bilatéral, des valeurs comparables à celles déjà indiquées précédemment, c'est-à-dire :

D_{α,n_X,n_Y} \ α	0,20	0,15	0,10	0,05	0,01
	$1,07.\sqrt{\frac{n_X+n_Y}{n_X.n_Y}}$	$1,14.\sqrt{\frac{n_X+n_Y}{n_X.n_Y}}$	$1,22.\sqrt{\frac{n_X+n_Y}{n_X.n_Y}}$	$1,36.\sqrt{\frac{n_X+n_Y}{n_X.n_Y}}$	$1,63.\sqrt{\frac{n_X+n_Y}{n_X.n_Y}}$

Pour ce qui est des petits échantillons, les tables jointes en annexes fournissent pour diverses valeurs de n_X et de n_Y, et pour les cas $\alpha = 0,01$ et $\alpha = 0,05$, les valeurs du seuil $D_{\alpha,n_X,n_Y} / \Pr ob(D \geq D_{\alpha,n_X,n_Y}) = \alpha$.

• En conclusion et à partir de la *distance calculée* $D_{calculé}(\widehat{F_X},\widehat{F_Y})$, on conclura à H_1 si $D_{calculé}(\widehat{F_X},\widehat{F_Y}) \geq D_{\alpha,n_X,n_Y}$ et à H_0 autrement, tout cela au risque de première espèce α. Ici encore, des tables spécifiques traitent le cas unilatéral $\begin{cases}H_0 : F_X = F_Y\\ H_1 : F_X \succ F_Y\end{cases}$ (resp. $\begin{cases}H_0 : F_X = F_Y\\ H_1 : F_X \prec F_Y\end{cases}$), la distance à considérer étant $D(\widehat{F_X},\widehat{F_Y}) = \underset{x\in R}{Sup}(\widehat{F_X}(x)-\widehat{F_Y}(x))$ (resp. $D(\widehat{F_X},\widehat{F_Y}) = \underset{x\in R}{Sup}(\widehat{F_Y}(x)-\widehat{F_X}(x))$).

Les éléments susmentionnés sont illustrés à travers l'exemple ci-après dans lequel à partir de l'observation de la taille du domaine vital (en km²) pour neuf ours femelles (X) et six ours mâles (Y), on teste l'hypothèse H_0 : il n'y a pas de différence significative dans les tailles du domaine vital pour les ours mâles et femelles, contre l'hypothèse H_1 : la taille du domaine vital pour les ours mâles diffère significativement de celle du domaine vital pour les ours femelles.

Les domaines vitaux (en km²) ainsi observés pour les ours considérés sont les suivants :

Ours femelles (X)	37	72	60	49	18	50	102	49	20
Ours mâles (Y)	94	504	173	560	274	168			

La mise en œuvre du test conduit à classer par ordre croissant les valeurs de X et de Y et à mettre en regard les fonctions de répartition $\widehat{F_X}$ et $\widehat{F_Y}$, pour en déduire, point par point, l'expression des différences $\left|\widehat{F_X}-\widehat{F_Y}\right|$ et à fortiori de $Sup\left|\widehat{F_X}-\widehat{F_Y}\right|$. Dans le tableau ci-dessous N_X et N_Y désignent, pour chaque valeur x, le nombre de valeurs de X (resp. de Y) inférieures ou égales à x, les rapports $N_X/9$ et $N_Y/6$ étant en conséquence les valeurs de $\widehat{F_X}$ et de $\widehat{F_Y}$ au point x.

Valeurs x	N_X	N_Y	$\widehat{F_X}=N_X/9$	$\widehat{F_Y}=N_Y/6$	$\left\|\widehat{F_X}-\widehat{F_Y}\right\|$
18	1	0	0,111	0	0,111
20	2	0	0,222	0	0,222
37	3	0	0,333	0	0,333
49	5	0	0,555	0	0,555
50	6	0	0,666	0	0,666
60	7	0	0,777	0	0,777
72	8	0	0,888	0	0,888
94	8	1	0,888	0,166	0,722
102	9	1	1,000	0,166	0,833
168	9	2	1,000	0,333	0,666
173	9	3	1,000	0,500	0,500
274	9	4	1,000	0,666	0,333
504	9	5	1,000	0,833	0,166
560	9	6	1,000	1,000	0

La distance $D(\widehat{F_X},\widehat{F_Y})=\underset{x\in R}{Sup}\left|\widehat{F_X}(x)-\widehat{F_Y}(x)\right|$ a donc pour valeur 0,888 dans le cas proposé ici. S'agissant d'échantillons petits et d'un test bilatéral, les tables annexées fournissent pour $n_X=9, n_Y=6, \alpha=0,05$, la valeur $D_{0.05,9,6}$, du seuil D_{α,n_X,n_Y} / $\Pr ob(D\geq D_{\alpha,n_X,n_Y})=\alpha$.

L'inégalité $0,888\geq 0,722$ conduit donc ici à rejeter l'hypothèse H_0 puisque $D_{calculé}\geq D_{\alpha,n_X,n_Y}$.

b) Les tests de MANN-WHITNEY (« U test ») et WILCOXON pour l'identité de deux distributions

Plus *puissant* que le test de KOLMOGOROV-SMIRNOV, le test **de MANN-WHITNEY** est fondé non pas sur les valeurs observées $(x_1,x_2,...,x_{n_X})$ et $(y_1,y_2,...,y_{n_Y})$, mais sur leurs **rangs** dans un *classement commun.*

Plus précisément, ce test repose sur le fait que si les $X_i(1\leq i\leq n_X)$ et les $Y_j(1\leq j\leq n_Y)$ ont *même loi*, l'ensemble qui résulte de leur mélange est *homogène*, les probabilités $\Pr ob(X_i\prec Y_j),\forall(i,j)$ étant proches de $1/2$ lorsque les lois de X et de Y sont identiques.

Ainsi, notant par U le nombre de couples (i, j) pour lesquels $X_i \prec Y_j$ et considérant qu'il y a au total $n_X.n_Y$ couples (i, j) possibles, l'hypothèse $H_0 : F_X = F_Y$ conduit pour U à une valeur proche de $\frac{n_X.n_Y}{2}$. Le test en question qui porte sur $\begin{cases} H_0 : F_X = F_Y \\ H_1 : F_X \neq F_Y \end{cases}$ est donc construit autour de la *région critique* :

$$W = \left\{ (x_1, x_2, ..., x_{n_X}), (y_1, y_2, ..., y_{n_Y}) / \left| U - \frac{n_X.n_Y}{2} \right| \geq U_\alpha \right\},$$

U_α étant déterminé ici encore à partir de l'erreur de première espèce α.

Or, pour des *grands échantillons* $(n_X \geq 20, n_Y \geq 20)$, on montre que *la statistique* $\dfrac{U - \frac{n_X.n_Y}{2}}{\sqrt{\frac{n_X.n_Y.(n_X + n_Y + 1)}{12}}}$ **converge** vers la *loi normale centrée réduite* $N(0,1)$. Dès lors, notant par t_α le nombre vérifiant $Prob(|\xi| \geq t_\alpha) = \alpha$, il vient immédiatement l'expression du seuil critique $U_\alpha = t_\alpha \cdot \sqrt{\frac{n_X.n_Y.(n_X + n_Y + 1)}{12}}$.

La transposition du test précédent aux tests unilatéraux $\begin{cases} H_0 : F_X = F_Y \\ H_1 : F_X \succ F_Y \end{cases}$ ou $\begin{cases} H_0 : F_X = F_Y \\ H_1 : F_X \prec F_Y \end{cases}$ est immédiate. D'autre part, pour le cas des *petits échantillons*, on pourra recourir à des *tables de valeurs spécifiques*.

• Notamment, lorsque les échantillons sont de *grandes tailles*, la *comptabilisation du nombre total* U de couples (i, j) pour lesquels $X_i \prec Y_j$, peut être simplifiée en considérant *la somme* W *des rangs des valeurs* x_i *dans l'échantillon mélangé*, soit $(x_1, x_2, ..., x_{n_X}, y_1, y_2, ..., y_{n_Y})$, somme dont on montre qu'elle est reliée à U par la relation :

$$U = n_X.n_Y + \frac{n_X.(n_X + 1)}{2} - W.$$

Le **test de la somme des rangs** ainsi construit, qui est le **test de WILCOXON**, conduit immédiatement, à partir de la relation $Prob\left(\dfrac{\left| U - \frac{n_X.n_Y}{2} \right|}{\sqrt{\frac{n_X.n_Y.(n_X + n_Y + 1)}{12}}} \geq t_\alpha \right) = \alpha$ à la nouvelle équation $Prob\left(\dfrac{\left| W - \frac{n_X.(n_X + n_Y + 1)}{2} \right|}{\sqrt{\frac{n_X.n_Y.(n_X + n_Y + 1)}{12}}} \geq t_\alpha \right) = \alpha$, d'où, la *région critique* :

$$W = \left\{ (x_1, x_2, ..., x_{n_X}), (y_1, y_2, ..., y_{n_Y}) / \left| W - \frac{n_X.(n_X + n_Y + 1)}{2} \right| \geq W_\alpha \right\},$$

avec $W_\alpha = t_\alpha \cdot \sqrt{\frac{n_X.n_Y.(n_X + n_Y + 1)}{12}}$.

Ici encore, des tables de valeurs permettent de traiter le cas des petits échantillons.

• Enfin, pour les deux tests précédents qu'on regroupe fréquemment sous l'appellation commune du **test de WILCOXON-MANN-WHITNEY**, la *détermination des rangs* **en cas d'ex-aequo** pourra s'effectuer par *tirage au sort* ou de préférence en attribuant aux ex-aequo un **rang moyen** (cf. applications ci-après). Lorsque les valeurs identiques sont nombreuses, il sera fait usage d'une **formule de correction** *sur la variance*, formule dont les effets restent néanmoins souvent négligeables (du moins, pour un nombre limité d'ex-aequo). La *formule corrigée* en question vise, pour U ou W, à remplacer la variance $\sqrt{\frac{n_X.n_Y.(n_X+n_Y+1)}{12}}$ par la quantité $\sqrt{\frac{n_X.n_Y}{N.(N-1)}.\left[\frac{N^3-N}{12}-T\right]}$ avec $N=n_X+n_Y$ et $T=\sum_{g=1}^{g=N_g}\frac{t_g^3-t_g}{12}$, N_g nombre de valeurs différentes des rangs, formule dans laquelle pour une valeur g donnée du rang, t_g désigne le nombre d'ex-aequo constaté.

c) Le choix du test approprié

Il reste lié à la nature des différences que l'on souhaite mettre en évidence entre les deux échantillons à comparer.

S'il s'agit de mettre en évidence les différences qui concernent la tendance centrale entre les deux populations dont sont extraits les deux échantillons (médiane qui dans le cas de distributions symétriques est aussi la moyenne), on privilégiera les tests de MANN-WHITNEY-WILCOXON et KOLMOGOROV-SIRNOV, tout en sachant que le test de la médiane (test de MOOD) développé en applications ci-après forme également une méthode classique en ce sens.

S'il s'agit d'une comparaison plus large englobant tendance centrale et dispersion ou aplatissement, il conviendra de préférer le test de chi- deux ou le test bilatéral de KOLMOGOROV-SIRNOV.

d) Le test « H » de KRUSKAL-WALLIS pour l'identité de K distributions ($K \geq 2$)

Utilisant également la notion de **rang**, ce test constitue en quelques sorte, la *généralisation du test précédent de WILCOXON-MANN-WHITNEY*. On dispose cette fois, de K séries de valeurs $x_{i,j}$ d'effectifs respectifs $n_i, (1 \leq i \leq K, 1 \leq j \leq n_i)$, et on cherche à tester l'hypothèse de *l'identité des distributions* respectives de ces k échantillons, soit le test de l'hypothèse nulle $H_0 : F_{X_1} = F_{X_2} = ... = F_{X_K}$ contre l'hypothèse alternative $H_1 : \exists (i,j) / F_{X_i} \neq F_{X_j}$.

Après avoir *classé par ordre croissant* tous les $x_{i,j}$ dans un ensemble regroupé et leur avoir affecté leur **rang** correspondant dans ce classement, l'idée est de calculer, pour *chaque classe* $i, (1 \leq i \leq K)$, le *rang moyen des éléments qu'elle contient*, soit $\overline{R_i} = \frac{\sum_{j=1}^{j=n_i} R_{i,j}}{n_i}$ et de **comparer** *ces rangs moyens* $\overline{R_1}, \overline{R_2}, ..., \overline{R_K}$ au **rang moyen théorique** $\overline{R}$ qui, dans l'hypothèse H_0 de l'identité des distributions entre les K échantillons est égal $\overline{R} = \frac{\sum_{i=1}^{i=N} i}{N} = \frac{N+1}{2}$, avec $N = \sum_{i=1}^{i=K} n_i$.

On calcule alors, la **somme pondérée des distances** entre les $\overline{R_i}$ et la valeur $\overline{R}$, ou plutôt la *statistique* $H = \frac{12}{N.(N+1)} . \sum_{i=1}^{i=K} n_i . (\overline{R_i} - \overline{R})^2$ dont on montre qu'elle suit la **loi du chi-deux** à $K-1$ degrés de libertés. Dès lors, la détermination du seuil $\chi_\alpha^2 / \Pr ob(\chi^2 \geq \chi_\alpha^2) = \alpha$ et la comparaison de H à χ_α^2 permet de conclure quant au choix de H_0 ou de H_1.

• A noter la possibilité d'appliquer un *coefficient correctif sur la variance* lorsqu'il y a plusieurs **ex-aequo**, la formule consistant à remplacer H par $H' = \frac{H}{C}$, avec C égal à $1 - \frac{\sum_{g=1}^{g=N_g} (t_g^3 - t_g)}{N.(N^2-1)}$ (t_g et N_g étant définis ci-après).

Dans la formule précédente, N_g désigne le nombre de valeurs différentes des rangs dans le classement regroupé, formule dans laquelle pour une valeur g donnée du rang, t_g désigne le nombre d'ex-aequo constaté. Ici encore, pour le calcul de H, on remplacera ces rangs ex-aequo par leur valeur moyenne.

• Enfin, comme pour le test ANOVA, une méthode inspirée des **contrastes**, permet d'examiner, lorsque l'hypothèse H_1 est vraie, *quelles sont les distributions qui diffèrent*. La méthode consiste à comparer deux à deux les différentes distributions ce qui conduit à $\frac{K.(K-1)}{2}$ tests (K nombre de populations considérées).

Souhaitant conserver à *l'erreur de première espèce* α, la valeur totale qui correspond au test de KRUSKAL-WALLIS, cela induit pour chacun des tests de comparaison ci-dessus, *l'erreur de première espèce réduite* $\alpha' = \frac{\alpha}{K.(K-1)}$. Il s'ensuit, pour tout couple (i, j), un test se résumant, pour n *assez grand*, à la *région critique* :

$$\left|\overline{R_i} - \overline{R_j}\right| \geq t_\alpha . \sqrt{\frac{n.(n+1)}{12} . (\frac{1}{n_i} + \frac{1}{n_j})} \text{ avec } n = \sum_{i=1}^{i=K} n_i \text{ et } t_{\alpha'} \text{ vérifiant } \Pr ob(|\xi| \leq t_{\alpha'}) = \alpha' .$$

3.3 Tests de comparaison entre échantillons appariés

Pour rappel, dans les échantillons appariés, les séries de mesures effectuées suivant K modalités (dans le cas général) portent sur les mêmes individus, le cas $K = 2$ étant le plus courant bien évidemment.

Par exemple, des mesures biologiques avant et après un traitement, la double correction des copies, les rendements obtenus par des variétés cultivées sur différentes parcelles....

a) Le test des signes pour l'identité de deux distributions

Peu puissant, mais très simple, ce test qui s'applique au cas de deux échantillons, les données étant associées par *paires*, n'utilise qu'un *classement au niveau de chaque paire* et s'abstient donc de données quantitatives.

Pratiquement, après avoir *éliminé les paires* (x_i, y_i) *pour lesquelles la différence* $d_i = x_i - y_i$ *est nulle*, on comptabilise au sein des n paires restantes, le nombre d de *différences de signe positif* (**paires « + »**).

Or, manifestement sous l'hypothèse H_0 : $F_X = F_Y$, la variable D qui caractérise le nombre de paires « + » parmi les n paires restantes, suit la loi binomiale $B(n, 1/2)$.

On est donc ramené ici à un *test d'ajustement* (**test binomial**), dont le mode opératoire est facilité par la table de valeurs annexée (*table des valeurs critiques du test binomial*).

En effet, cette table fournit pour tout couple (d,n), la probabilité $\Pr ob(D \le d)$ pour le cas de la loi $B(n, 1/2)$.

Raisonnant par exemple, sur le test unilatéral $\begin{cases} H_0 : F_X = F_Y \\ H_1 : F_X \prec F_Y \end{cases}$, la *région critique* a pour forme $d \le d_\alpha$ où d désigne la valeur observée de D et d_α est tel que, relativement à la loi binomiale $D : B(n, 1/2)$, on ait la relation $\Pr ob(D \le d_\alpha) = \alpha$.

Cette comparaison de d à d_α est équivalente à la comparaison de $\Pr ob(D \le d)$ à α, la probabilité $\Pr ob(D \le d)$ étant lue dans la table susmentionnée. Concrètement, et pour le test considéré, on décidera H_1 si $\Pr ob(D \le d) \le \alpha$.

Dans le cas du *test bilatéral*, on se ramènera au test précédent avec l'erreur de première espèce $\frac{\alpha}{2}$, ce qui revient à *multiplier par deux* la valeur lue dans la table (on a cette fois, $2.\Pr ob(D \le d) \le \alpha$).

Par exemple, on se propose d'évaluer la justesse d'un appareil de mesure en comparant les hauteurs de sept arbres, calculées à partir dudit appareil puis de façon exacte (au sol, après qu'ils aient été abattus, par exemple). Les données sont les suivantes :

n° d'arbre (i)	1	2	3	4	5	6	7
hauteur en m suivant mesure par appareil (x_i)	20,4	25,4	25,6	25,6	26 ,6	28,6	30,5
hauteur en m suivant mesure exacte (y_i)	20,7	26,3	26,8	28,1	26,2	28,9	32,3

Il s'ensuit les différences :

n° d'arbre (i)	1	2	3	4	5	6	7
différence $d_i = x_i - y_i$	-0,3	-0,9	-1,2	-2,5	**+0,4**	-0,3	-1,8

On a donc ici, une seule différence positive, ce qui entraîne $d = 1$. Or, pour $n = 7$, la table annexée des valeurs critiques du test binomial fournit la probabilité $\Pr ob(D \le 1) = 0,062$, probabilité à doubler si on considère le test bilatéral.

Fixant l'erreur de première espèce α à 0,05, on n'a pas ici $\Pr ob(D \le 1) \le 0,05$. On ne peut donc pas rejeter l'hypothèse de l'exactitude de l'appareil de mesure.

- Pour le cas des *grands échantillons* $(n \ge 30)$, on pourra utiliser *l'approximation de la loi binomiale* $B(n, 1/2)$ *par la loi normale* de moyenne $\frac{n}{2}$ et d'écart- type $\frac{\sqrt{n}}{2}$ (théorème de MOIVRE-LAPLACE).

Ainsi $\Pr ob(D \le d) = \Pr ob(\xi \le \dfrac{d - \dfrac{n}{2}}{\dfrac{\sqrt{n}}{2}})$, ou $\Pr ob(D \le d) = \Pr ob(\xi \le \dfrac{d - \dfrac{n}{2} + 0,5}{\dfrac{\sqrt{n}}{2}})$ en utilisant la *correction de continuité* de YATES.

Pour rappel, la correction de continuité de YATES vise à corriger l'imprécision suivant laquelle la convergence d'une loi discrète vers une loi continue conduit à confondre inégalités larges et strictes puisque la loi limite qui est continue ne charge pas les points. Le principe en est d'assimiler $\Pr ob(X = x)$ à $\Pr ob(x - \frac{1}{2} \le X \le x + \frac{1}{2})$ ce qui, entre la variable discrète "X" et sa limite continue "X^*", conduit par exemple, aux relations $\Pr ob(X \le x) = \Pr ob(X^* \le x + 0,5), \Pr ob(X \ge x) = \Pr ob(X^* \ge x - 0,5)$....

b) Le test des rangs « signés » de WILCOXON pour l'identité de deux distributions

Plus puissante que le test précédent, la présente version du **test de WILCOXON**, également fondée sur la notion de **rang**, prend en compte l'*amplitude* des valeurs $d_i = x_i - y_i$, *une paire ayant un grand « d_i » se voyant donner plus de poids (à travers son rang) qu'une paire ayant un petit « d_i »*. Le mode opératoire en est le suivant :

→ on *calcule* les différences $d_i = x_i - y_i$;

→ on *classe* les d_i par valeurs croissantes ;

→ on « *signe* » les d_i par « + » ou « - » suivant le signe de $d_i = x_i - y_i$, les paires pour lesquelles $d_i = 0$ étant éliminées;

→ on *attribue* un rang aux $|d_i|$ en utilisant ici encore le principe des *rangs moyens* pour les ex-aequo ;

→ on *calcule* les sommes m et p égales respectivement aux sommes des rangs des d_i signés par « - » et des rangs signés par « + ». Il est immédiat à ce sujet, que $m + p = \sum_{k=1}^{k=n} k = \dfrac{n.(n+1)}{2}$.

Sous l'hypothèse H_0, on a pour les statistiques M et P associées aux sommes susmentionnées, $E(M) = E(P) = \dfrac{n.(n+1)}{4}$ et $\sigma_M^2 = \sigma_P^2 = \dfrac{n.(n+1).(2n+1)}{24}$. Admettant que pour n *assez grand* ($n \ge 30$), la loi de $\dfrac{P - \dfrac{n.(n+1)}{4}}{\sqrt{\dfrac{n.(n+1).(2n+1)}{24}}}$ converge vers la *loi normale centrée réduite* $N(0,1)$. On conclut aisément dès lors au *choix entre les hypothèses* $H_0 : F_X = F_Y$ et $H_1 : F_X \ne F_Y$ suivant le *positionnement* de $\left| \dfrac{P - \dfrac{n.(n+1)}{4}}{\sqrt{\dfrac{n.(n+1).(2n+1)}{24}}} \right|$ par rapport au nombre t_α vérifiant $\Pr ob(|\xi| \ge t_\alpha) = \alpha$ (pour le test unilatéral, il en est de même, mais sans la valeur absolue).

• Dans le cas des petits échantillons ($n \leq 30$), la table de WILCOXON (cf. annexes), fournit relativement à la statistique P et en fonction des valeurs de n, le seuil $T_\alpha / \Pr ob(P \geq T_\alpha) = \alpha$. Pour le test unilatéral à droite, on retiendra donc H_1 lorsque $P_{calculé} \geq T_\alpha$. Dans le cas du test bilatéral, on considérera les seuils $T_{1-\alpha/2}$ et $T_{\alpha/2}$ vérifiant respectivement $\Pr ob(P \leq T_{1-\alpha/2}) = 1 - \alpha/2$ et $\Pr ob(P \geq T_{\alpha/2}) = \alpha/2$.

• Enfin, dans le cas où il y a un grand nombre d'ex aequo parmi les valeurs $|d_i|$, on corrigera la variance à l'instar du test de MANN-WHITNEY-WILCOXON, la formule préconisée en ce sens, étant après correction :

$$VarP = \frac{n.(n+1).(2n+1)}{24} - \frac{1}{48}.\sum_{g=1}^{g=N_g} t_g.(1-t_g).(1+t_g)$$

(t_g nombre d'ex aequo pour chaque valeur distincte de $|d_i|$ considérée et N_g nombre total de ces valeurs distinctes).

c) Le test de MAC NEMAR pour l'identité de deux distributions suivant variables binaires

En fait, il s'agit de *comparer* **avant** et **après** un traitement donné, des *mesures* effectuées sur les mêmes individus et dont le renseignement fourni est de type **binaire** (bon, mauvais – content, mécontent....).

Désignant par π_1 et π_2 la proportion de la valeur « 1 » respectivement dans chacune des deux populations considérées, il s'agit ici de tester : $\begin{cases} H_0 : \pi_1 = \pi_2 \\ H_1 : \pi_1 \neq \pi_2 \end{cases}$. Le **tableau de contingence** ci-dessous, rassemble les données en mettant notamment en évidence, parmi les n éléments testés, les quatre situations possibles entre *avant* et *après* le traitement en question (on a $n = a+b+c+d$).

Avant\Après	0	1	Σ
0	a	b	$a+b$
1	c	d	$c+d$
Σ	$a+c$	$b+d$	$n=a+b+c+d$

Dans les conditions ci-dessus, l'égalité des proportions $\pi_1 = \frac{c+d}{n}$ et $\pi_2 = \frac{b+d}{n}$ conduit à la comparaison de b avec c, c'est-à-dire plus précisément, des échanges de réponses 0 → 1 et 1 → 0 entre *avant* et *après* le traitement, l'*invariance* de la proportion π de la valeur « 1 » qui correspond à l'hypothèse H_0, conduisant pour ces échanges , à l'effectif commun théorique $\frac{b+c}{2}$.

En effet, notant par c^* et b^* les effectifs théoriques qui correspondent respectivement aux transitions 1 → 0 et 0 → 1, lorsque H_0 est vraie, on a d'une part $c^* = b^* = y$, ce qui implique $a+2.y+d=n$ et d'autre part, s'agissant des effectifs observés, la relation $a+b+c+d=n$. Il s'ensuit aisément $2.y = b+c$, ce qui justifie la valeur commune susmentionnée $c^* = b^* = \frac{b+c}{2}$.

Le tableau ci-après, résume cette **comparaison** entre *effectifs absolus observés* et *effectifs absolus théoriques*, pour ces transitions 1 → 0 et 0 → 1.

Echanges de réponses	0 → 1	1 → 0
Fréquences absolues observées	b	c
Fréquences absolues théoriques	$\frac{b+c}{2}$	$\frac{b+c}{2}$

L'application du **test de chi- deux** relatif à la *signification ou non* de cette différence entre fréquences observées et théoriques **(test d'homogénéité)**, conduit immédiatement à la statistique $\chi^2 = \frac{(\frac{b+c}{2}-b)^2}{\frac{b+c}{2}} + \frac{(\frac{b+c}{2}-c)^2}{\frac{b+c}{2}} = \frac{(b-c)^2}{b+c}$, distance qui suit la *loi du chi –deux à un degré de liberté*. Il en résulte immédiatement au *risque de première espèce* α *donné*, la *région critique* caractérisée par $\frac{(b-c)^2}{b+c} \geq \chi_\alpha^2$ avec $\Pr ob(\chi^2 \geq \chi_\alpha^2) = \alpha$.

A noter, *pour les petits échantillons*, la recommandation d'appliquer la *correction de continuité* de YATES, la distance à considérer étant alors $\frac{(|b-c|-1)^2}{b+c}$.

d) Le test de COCHRAN pour l'identité de K ***distributions suivant variables binaires,*** $K \geq 2$

Version « binaire » du test de FRIEDMAN (cf. paragraphe suivant e)), il constitue une *extension* du test de MAC NEMAR. On considère ainsi un échantillon de taille n dans lequel une **variable binaire** est mesurée K fois à travers *divers traitements*, la question étant, pour les proportions π_i de la valeur « 1 » au sein de ces mesures pour chacun des traitements en question $(1 \leq i \leq K)$, de tester l'hypothèse $H_0 : \pi_1 = \pi_2 = ... = \pi_K$ contre l'hypothèse alternative $H_1 : \exists (i,j) / \pi_i \neq \pi_j$.

Les données (mesures binaires $x_{i,j}$) sont rassemblées dans le *tableau de contingence* ci-dessous :

Bloc (n° de l'échantillon)	Traitements « i », $1 \leq i \leq K$						Somme
	X_1	X_2	.	X_j	.	X_K	
1	$x_{1,1}$	$x_{1,2}$	.	$x_{1,j}$	.	$x_{1,K}$	L_1
2	$x_{2,1}$	$x_{2,2}$	.	$x_{2,j}$	.	$x_{2,K}$	L_2
.	.	.	.	.	.	.	.
i	$x_{i,1}$	$x_{i,2}$	.	$x_{i,j}$	.	$x_{i,K}$	L_i
.	.	.	.	.	.	.	.
n	$x_{n,1}$	$x_{n,2}$	.	$x_{n,j}$	.	$x_{n,K}$	L_n
Somme	C_1	C_2	.	C_j	.	C_K	S (*)

(*) $S = \sum_{j=1}^{j=K} C_j = \sum_{i=1}^{i=n} L_i$.

La **statistique de COCHRAN** est définie par :

$$Q = K.(K-1).\frac{\sum_{j=1}^{j=K}(C_j - \frac{S}{K})^2}{\sum_{i=1}^{i=n} L_i.(K-L_i)} = \frac{(K-1).\left[K.\sum_{j=1}^{j=K} C_j^2 - (\sum_{j=1}^{j=K} C_j)^2\right]}{K.\sum_{i=1}^{i=n} L_i - \sum_{i=1}^{i=n} L_i^2}.$$

Lorsque H_0 est vraie, il est montré que Q suit *asymptotiquement* la **loi du** χ^2 à $K-1$ degrés de libertés, approximation supposée correcte dès que $n \geq 4$ et $n.K \geq 24$. Le cas de l'identité des traitements correspond en théorie à $C_j = \frac{S}{K}, \forall j / (1 \leq j \leq K) \Rightarrow Q = 0$. Dès lors, la *région critique* du test qui traduit des écarts significatifs entre les C_j et le rapport théorique S/K est fournie immédiatement par la condition $Q \geq \chi_\alpha^2$ où χ_α^2 vérifie, relativement à la loi $\chi^2(K-1)$, la condition $\Pr ob(\chi^2(K-1) \geq \chi_\alpha^2) = \alpha$.

e) Le test de FRIEDMAN pour l'identité de K ***distributions,*** $K \geq 2$

Généralisant le test des signes et formant une *analyse de variance* (ANOVA) *non paramétrique*, le test propose à partir d'un échantillon de taille n soumis à K mesures à travers divers traitements $X_i, (1 \leq i \leq K)$, distincts, soient $x_{i,j} (1 \leq i \leq n, 1 \leq j \leq K)$ les données ainsi collectées, de tester l'hypothèse H_0 : « *les K traitements sont identiques* », contre l'hypothèse H_1 : « *il existe au moins deux traitements qui diffèrent* ».

Comme pour le test de KRUSKAL-WALLIS, la condition exigée ici des données est de pouvoir être classées (qu'elles soient *quantitatives* ou *ordinales*). Cette fois, du fait qu'il s'agit de *données appariées par lignes* « blocs », c'est pour chaque élément considéré de l'échantillon (c'est-à-dire par ligne), que les *rangs* sont déterminés, la somme qui en résulte étant constante pour toutes les lignes et égale à $\sum_{i=1}^{i=K} i = \frac{K.(K+1)}{2}$. Les données sont résumées ci-dessous :

Bloc (n° de l'échantillon)	Traitements « j », $1 \leq j \leq K$						Somme
	X_1	X_2	.	X_j	.	X_K	
1	$r_{1,1}$	$r_{1,2}$	.	$r_{1,j}$	.	$r_{1,K}$	$R_{1.} = K.(K+1)/2$
2	$r_{2,1}$	$r_{2,2}$	.	$r_{2,j}$	.	$r_{2,K}$	$R_{2.} = K.(K+1)/2$
.	.	.	.	.	.	.	.
i	$r_{i,1}$	$r_{i,2}$	.	$r_{i,j}$	.	$r_{i,K}$	$R_{i.} = K.(K+1)/2$
.	.	.	.	.	.	.	.
n	$r_{n,1}$	$r_{n,2}$	.	$r_{n,j}$	.	$r_{n,K}$	$R_{n.} = K.(K+1)/2$
Somme	$S_{.1}$	$S_{.2}$	.	$S_{.j}$	.	$S_{.K}$	

Notant par $\overline{R_{.j}}$ les moyennes $\frac{S_{.j}}{n}$ pour chacun des traitements « j » ci-dessus $(1 \leq j \leq K)$, la statistique choisie ici pour comparer les effets des K traitements en question, est la **statistique du test de FRIEDMAN** .

Cette statistique est définie par :

$$F = \frac{12.n}{K.(K+1)} . \sum_{j=1}^{j=K} (\overline{R_{.j}} - \overline{R})^2 \text{, avec } \overline{R} = \frac{\sum_{i=1}^{i=n} \sum_{j=1}^{j=K} r_{i,j}}{n.K} = \frac{n.K.(K+1)/2}{n.K} = \frac{K+1}{2}$$

Il en résulte $F = \frac{12.n}{K.(K+1)} . \sum_{j=1}^{j=K} (\overline{R_{.j}} - \frac{K+1}{2})^2$, ou encore après développement, l'expression $F = \frac{12}{n.K.(K+1)} . \sum_{j=1}^{j=K} S_{.j}^2 - 3.n.(K+1)$.

Or, on montre que, *pour n assez grand*, F qui exprime la *variabilité* entre les traitements ($F = 0$ si ceux-ci sont équivalents « hypothèse H_0 »), suit une **loi de chi- deux** à $K-1$ degrés de liberté, soit $\chi^2(K-1)$. Dès lors, la *région critique* du test de FRIEDMAN est générée par la valeur du seuil $\chi_\alpha^2 / \Pr ob(\chi^2(K-1) \geq \chi_\alpha^2) = \alpha$, région critique caractérisée par la relation $F_{calculé} \geq \chi_\alpha^2$.

- Pour le cas des *petits échantillons*, on déterminera le seuil $F_\alpha / \Pr ob(F \geq F_\alpha) = \alpha$ à partir de la *table des valeurs de la statistique de FRIEDMAN* jointe en annexes.

- D'autre part, lorsqu'il y a des **ex-aequo** à l'intérieur d'un bloc, on utilisera ici encore le principe du *rang moyen* avec en outre l'application d'un *facteur correctif* suivant la formule $F_{corrigé} = \frac{F}{C}$ avec $C = 1 - \frac{\sum_{i=1}^{i=n} \sum_{g=1}^{g=G_i} (t_{i,g}^3 - t_{i,g})}{n.(K^3 - K)}$, G_i désignant pour le bloc $i, (1 \leq i \leq n)$, le nombre de valeurs différentes et $t_{i,g}$ le nombre de répétitions de la valeur n° g $(1 \leq g \leq G_i)$.

- Enfin, comme pour le test de KRUSKAL-WALLIS, on peut lorsque l'hypothèse H_0 est rejetée, déterminer la *source des écarts entre les traitements*, ceci à travers les $\frac{K.(K-1)}{2}$ tests qui permettent de comparer, deux à deux, les traitements effectués. Considérant ainsi l'erreur de première espèce réduite $\alpha' = \frac{\alpha}{K.(K-1)}$ (pour conserver le risque global α sur l'ensemble des comparaisons), chaque test, formulé par exemple à partir d'une *comparaison sur les rangs moyens par traitement*, conduit à la région critique :

$$\left| \overline{R_{.j}} - \overline{R_{.i}} \right| \geq t_{\alpha'} \sqrt{\frac{K.(K+1)}{6.n}}, \; t_{\alpha'} \text{ vérifiant, pour } n \text{ assez grand, } \Pr ob(|\xi| \geq t_{\alpha'}) = \alpha'$$

3.4 Tests d'association

Le **test du r de PEARSON** antérieurement développé au paragraphe 2.3.d) du présent rappel de cours, constitue un exemple classique de test d'association pour *variables quantitatives* et pour des **liaisons linéaires**.

Mais, il demeure le cas des *variables qualitatives* (*ordinales ou nominales*) et qui plus est, la notion d'association et d'indépendance ne peut pas être restreinte aux liaisons linéaires. Il est rappelé en effet, qu'il existe des variables qui ont un faible coefficient de corrélation linéaire et qui cependant sont très liées par une relation monotone autre que linéaire.

Les statistiques du **rhô de SPEARMAN** et du **tau de KENDALL** offrent cette alternative utile au coefficient $r_{x,y}$ de PEARSON.

a) Le coefficient Rhô de corrélation des rangs de SPEARMAN pour la liaison entre variables quantitatives ou ordinales

D'une portée plus large que le coefficient $r_{x,y}$ de PEARSON puisque détectant les liaisons non linéaires et s'étendant aux données ordinales, le **coefficient de SPEARMAN** qui est basé sur les *rangs des valeurs des variables* et non les valeurs elles-mêmes, permet de s'affranchir de l'hypothèse contraignante de normalité à l'instar des tests non paramétriques.

Partant d'un échantillon de n valeurs $(x_i, y_i), 1 \le i \le n$, de deux variables aléatoires X et Y, et notant pour tout i, les **rangs** respectifs R_i et S_i des valeurs x_i et y_i au sein des *séries* respectives $(x_1, x_2, ..., x_n)$ et $(y_1, y_2, ..., y_n)$ *préalablement ordonnées* par ordre croissant, le coefficient de SPEARMAN est défini par $\rho = \dfrac{\text{cov}(R,S)}{\sqrt{Var(R).Var(S)}} \Rightarrow -1 \le \rho \le 1$.

En pratique, le coefficient de SPEARMAN est fourni par la formule :

$$\rho = 1 - \frac{6.\sum_{i=1}^{i=n}(R_i - S_i)^2}{n.(n^2-1)}.$$

En effet, il est immédiat que $\overline{R} = \overline{S} = \dfrac{\sum_{i=1}^{i=n} R_i}{n}$, soit en définitive $\overline{R} = \overline{S} = \dfrac{1}{n}.\sum_{i=1}^{i=n} i = \dfrac{n+1}{2}$.

Par ailleurs, $\sum_{i=1}^{i=n} R_i^2 = \sum_{i=1}^{i=n} S_i^2 = \sum_{i=1}^{i=n} i^2 = \dfrac{n.(n+1).(2n+1)}{6}$ (du moins lorsqu'il n'y a pas d'ex aequo). Il en résulte $\sum_{i=1}^{i=n}(R_i - \overline{R})^2 = \sum_{i=1}^{i=n}(S_i - \overline{S})^2 = \sum_{i=1}^{i=n} R_i^2 - n.\overline{R}^2 = \dfrac{n.(n^2-1)}{12}$.

Enfin, on a parallèlement, $\sum_{i=1}^{i=n}(R_i - S_i)^2 = \sum_{i=1}^{i=n} R_i^2 + \sum_{i=1}^{i=n} S_i^2 - 2.\sum_{i=1}^{i=n} R_i.S_i$, soit en développant,

$\sum_{i=1}^{i=n}(R_i - S_i)^2 = \dfrac{n.(n+1).(2n+1)}{3} - 2.\sum_{i=1}^{i=n} R_i.S_i$.

Finalement, $\rho = \dfrac{\dfrac{(n+1).(2n+1)}{6} - \sum_{i=1}^{i=n}\dfrac{(R_i - S_i)^2}{2n} - \dfrac{(n+1)^2}{4}}{\dfrac{n^3-n}{12}}$, ce qui après simplification établit l'expression annoncée $\rho = 1 - \dfrac{6.\sum_{i=1}^{i=n}(R_i - S_i)^2}{n.(n^2-1)}$.

• Or, à l'instar du test r de PEARSON présenté au paragraphe 2.3.d) antérieur, on montre que pour n *assez grand* $(n \ge 10)$, la variable $T = \rho.\sqrt{\dfrac{n-2}{1-\rho^2}}$ suit approximativement la **loi de STUDENT** à $\nu = n-2$ degrés de libertés.

Ramenant ainsi l'**indépendance** des variables X et Y à un *test bilatéral* dont l'hypothèse H_0 est la *nullité de l'espérance de* ρ et à fortiori de T, il vient immédiatement la région critique $|\rho| \geq \frac{t_\alpha}{\sqrt{n-2+t_\alpha^2}}$ où t_α vérifie $\Pr ob(|T| \geq t_\alpha) = \alpha$, T variable de STUDENT à $\nu = n-2$ degrés de libertés.

A noter également, la *convergence pour n grand*, de ρ vers **la loi normale** de variance $\frac{1}{n-1}$, ce qui relativement au test bilatéral portant sur la nullité ou non de l'espérance mathématique de ρ, conduit à partir de la variable gaussienne centrée réduite $\xi = \sqrt{n-1}.\rho$, à la région critique définie par $|\rho| \geq \frac{t_\alpha}{\sqrt{n-1}}$ où t_α vérifie $\Pr ob(|\xi| \geq t_\alpha) = \alpha$.

- Pour les petits échantillons, il conviendra d'utiliser une table de valeurs spécifique.

- Le test du coefficient de SPEARMAN peut aussi être utilisé pour évaluer si, au sein d'une échantillon $(x_1, x_2, ..., x_n)$ de n valeurs d'une variable aléatoire X, *lesdites valeurs* sont **indépendantes** ou non.

Pour cela, on *compare les rangs des* x_i dans l'échantillon (après classement par ordre croissant comme habituellement), soient R_i, à *l'ordre croissant naturel* $(1, 2, ..., n)$. le coefficient de corrélation entre les deux séries de rangs étant nul lorsque l'hypothèse d'indépendance est satisfaite.

On a immédiatement $\rho = 1 - \frac{6.\sum_{i=1}^{i=n}(R_i - i)^2}{n.(n^2-1)}$ et ici encore la règle de décision de H_1 (*région critique*) est caractérisée par $|\rho| \geq \pi = \frac{t_\alpha}{\sqrt{n-1}}$, où t_α vérifie $\Pr ob(|\xi| \geq t_\alpha) = \alpha$ (du moins pour les valeurs de n supérieures ou égales à 30).

- Enfin, lorsqu'il y a des **ex-aequo**, les calculs précédents des sommes $\sum_{i=1}^{i=n} R_i^2, \sum_{i=1}^{i=n} R_i.S_i$ sont *erronés*. On pourra alors utiliser un *coefficient corrigé* et plus compliqué défini par :

$$\rho = \frac{S_X + S_Y - \sum_{i=1}^{i=n}(R_i - S_i)^2}{2.\sqrt{S_X.S_Y}},$$

formule dans laquelle $S_X = \frac{n.(n^2-1) - \sum_{i=1}^{i=g}(t_i^3 - t_i)}{12}$ et $S_Y = \frac{n.(n^2-1) - \sum_{j=1}^{j=h}(t_j^3 - t_j)}{12}$, g étant le nombre de valeurs différentes au sein des rangs de X et t_i étant le nombre d'ex-aequo pour le rang $i, (1 \leq i \leq g)$. De même, h et t_j pour ce qui concerne les rangs de Y.

L'autre coefficient d'association qui est le coefficient tau de KENDALL présente relativement au traitement des valeurs ex-aequo, une simplicité plus grande comme on pourra le constater ci-après.

b) Le coefficient Tau de corrélation des rangs de KENDALL pour la liaison entre variables quantitatives ou ordinales

De même puissance que le test précédent, le **test de KENDALL** qui présente l'avantage d'une possibilité de généralisation à plus de deux variables, est basé sur le **nombre d'inversions** *constatées dans les classements des rangs entre les deux ensembles considérés*, la *mesure de la corrélation* étant définie ici par le coefficient de KENDALL égal, au signe près, à $\tau = 1 - 2.\dfrac{Nombre\ d'inversions}{Nombre\ d'associations\ possibles}$, expression qui, pour un échantillon de taille n, s'écrit aussi $\tau = 1 - 2.\dfrac{Nombre\ d'inversions}{\dfrac{n.(n-1)}{2}}$ (puisque le nombre d'associations deux à deux possibles est C_n^2).

Il est immédiat de constater que τ est aussi le rapport entre :

- la différence entre le nombre de paires de rangs en concordance, soit n_c, et le nombre de paires en discordance, soit n_d ;
- le nombre total de paires « non ordonnées » possibles, soit $C_n^2 = \dfrac{n.(n-1)}{2}$

La technique ci-dessous, illustrée sous forme d'un exemple, facilite le *dénombrement des concordances et discordances* en question, les notations utilisées étant celles du paragraphe précédent.

Quatre produits ont été notés de 0 à 20 par deux utilisateurs différents (X,Y), soient $(x_i, y_i), (1 \le i \le 4)$, les données ainsi recueillies :

	$i=1$	$i=2$	$i=3$	$i=4$
X	14	18	12	10
Y	16	11	18	14

Raisonnant sur les rangs (R_i, S_i) on a immédiatement :

	$i=1$	$i=2$	$i=3$	$i=4$
X	3	4	2	1
Y	3	1	4	2

On réordonne alors les rangs des échantillons de façon à ce que, par rapport à X, la présentation des rangs s'effectue suivant l'ordre naturel $(1,2,3,4)$:

	$i=4$	$i=3$	$i=1$	$i=2$
X	1	2	3	4
Y	2	4	3	1

Puis, pour toutes les comparaisons deux à deux possibles (en nombre égal à $\dfrac{n.(n-1)}{2}$), on affecte un *coefficient* « + » lorsque les préférences de X et Y sont en *concordances* et « - » dans le cas contraire, *concordance* signifiant « Y augmente quand X augmente ». Ainsi, pour le couple (4,3), X et Y privilégient tous les deux 4 à 3 et sont donc en concordances. Il en est autrement pour le couple (4,2) pour lequel X augmente et Y diminue.

Le tableau symétrique (4,4) ci-après (matrice (n,n) dans le cas général), résume ces comparaisons deux à deux.

	1	2	3	4
1		-	-	+
2			-	-
3				+
4				

On a donc en conclusion, pour l'exemple en question, $n_c = 2, n_d = 4 \Rightarrow \tau = \dfrac{n_c - n_d}{\dfrac{n.(n-1)}{2}} = -0,33$.

• A noter également, l'autre technique possible ci-dessous, suivant *procédure de marquage.*

Dans le tableau des rangs, on relie les classements comparables lorsqu'il y a inversion, c'est à dire lorsque les x_i et y_i n'ont pas le même rang dans leurs séries associées.

	$i=4$	$i=3$	$i=1$	$i=2$
X	1	2	3	4
Y	2	4	3	1

Le *nombre d'inversions* est égal au *nombre d'intersections des lignes*, soit 2 pour le cas présent. Il s'ensuit $\tau = 1 - \dfrac{2 \times 2}{6} = 0,33$, ce qui correspond au résultat précédent, au signe près.

• Or, on montre que pour n *assez grand* (en pratique dès $n \geq 10$), la *loi du tau de KENDALL* converge vers la **loi normale** de variance $\dfrac{2.(2n+5)}{9n.(n-1)}$, l'hypothèse H_0 se traduisant par ailleurs par la nullité de $E(\tau)$.

En effet, s'il y a *indépendance* entre X et Y, les nombres de *concordances* et de *discordances* sont *statistiquement équilibrés*, ce qui entraîne $E(\tau) = 0$. Finalement la *région critique* (rejet de l'indépendance de X et Y), est caractérisée par la condition $|\tau| \geq t_\alpha . \sqrt{\dfrac{2.(2n+5)}{9n.(n-1)}}$ où t_α vérifie $Prob(|\xi| \geq t_\alpha) = \alpha$.

• Pour le cas des *petits échantillons* ($n \leq 10$), il conviendra d'utiliser une *table de valeurs spécifique* (table de KENDALL).

• Enfin, lorsqu'il y a des **valeurs ex aequo**, on utilisera la procédure habituelle du *rang moyen* qui cependant, altère le calcul de τ.

On pourra recourir alors à la formule corrective $\tau = \dfrac{S}{\sqrt{\dfrac{n.(n-1)}{2} - T_X} . \sqrt{\dfrac{n.(n-1)}{2} - T_y}}$

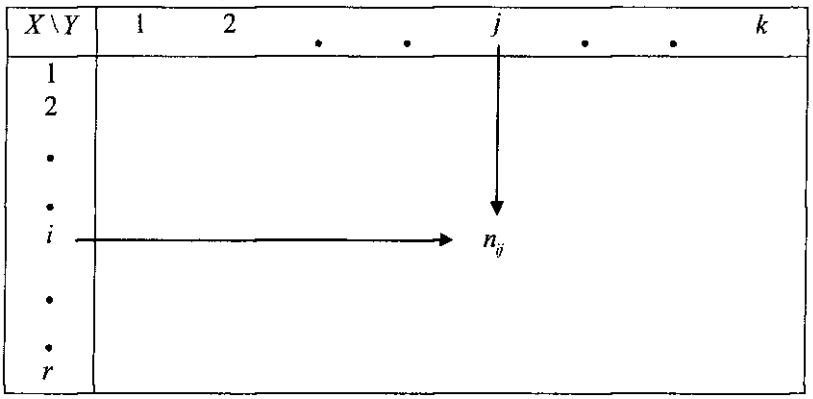
dans laquelle $T_X = \dfrac{\sum_{i=1}^{i=g_X} t_i.(t_i - 1)}{2}$ avec g_X désignant le nombre de valeurs distinctes dans la série des rangs des x_i et t_i le nombre de valeurs ex-aequo au rang $i, (1 \leq i \leq g_X)$.

De même, pour $T_Y = \dfrac{\sum_{j=1}^{j=g_Y} t_j.(t_j-1)}{2}$, avec g_Y désignant le nombre de valeurs distinctes dans la série des rangs des y_j et t_j le nombre de valeurs ex-aequo au rang $j, (1 \le j \le g_Y)$.

c) Le test de contingences du chi- deux pour la liaison entre variables qualitatives nominales

Les données relatives à X et Y sont supposées cette fois, être présentées sous forme d'un tableau de contingences respectivement à r et k classes, l'effectif correspondant à l'intersection des classes i et j étant noté n_{ij}, la question étant de tester :

$$\begin{cases} H_0 : X \text{ et } Y \text{ sont indépendantes} \\ H_1 : X \text{ et } Y \text{ sont liées} \end{cases}$$



Sous l'hypothèse H_0 d'*indépendance* de X et de Y, il est immédiat que la *probabilité conjointe* $p_{i,j} = \Pr ob(X = i, Y = j)$ est égale au *produit des probabilités* $\Pr ob(X = i).\Pr ob(Y = j)$, soit $p_{i.}.p_{.j}$.

Or ces différentes probabilités admettent pour estimateurs respectifs $\widehat{p_{i,j}} = \dfrac{n_{ij}}{n}, \widehat{p_{i.}} = \dfrac{n_{i.}}{n}, \widehat{p_{.j}} = \dfrac{n_{.j}}{n}$, avec $n = \sum_{i=1}^{i=r}\sum_{j=1}^{j=k} n_{ij} \; n_{i.} = \sum_{j=1}^{j=k} n_{ij} \; n_{.j} = \sum_{i=1}^{i=r} n_{ij}$. On montre alors que, sous l'hypothèse H_0, l'indicateur des *écarts entre la distribution observée* (les n_{ij}) *et la distribution théorique* qui correspond à l'*indépendance* (c'est-à-dire, les produits $n.\widehat{p_{i.}}.\widehat{p_{.j}}$) est fourni par la variable $V = \sum_{i=1}^{i=r}\sum_{j=1}^{j=k} \dfrac{(n_{ij} - n.\widehat{p_{i.}}.\widehat{p_{.j}})^2}{n.\widehat{p_{i.}}.\widehat{p_{.j}}}$ dont la distribution suit la **loi du chi- deux** à $\nu = (r-1).(k-1)$ degrés de libertés.

En effet, la définition de cet indicateur d'écart résulte immédiatement des éléments du paragraphe 3.1 précédent relatif au test d'adéquation de χ^2. La variable V suit, à cet égard, la loi du χ^2 à $\nu = (N-1) - p$ degrés de libertés, où en l'occurrence N qui désigne le nombre total de termes dans la somme des écarts considérée est égal à $r.k$, et où p qui représente le nombre de paramètres à estimer est égal à la somme des $r-1$ probabilités qui définissent la loi de X ($r-1$ car la $r^{ième}$ probabilité résulte de la condition $\sum_{i=1}^{i=r} p_{i.} = 1$) et des $k-1$ probabilités qui permettent de caractériser la loi de Y. En résumé, on a bien $\nu = r.k - (r-1) - (k-1) = (r-1).(k-1)$.

• Dès lors, il s'ensuit la *région critique* (zone de rejet de H_0), caractérisée par $V_{calculé} \geq \chi^2_\alpha$ où χ^2_α vérifie $\Pr ob(\chi^2 \geq \chi^2_\alpha) = \alpha$ (χ^2_α lu dans la *table des valeurs de la loi de* χ^2 annexée).

• Comme pour l'adéquation, les classes dont le nombre influe directement sur la valeur de v doivent comprendre un *degré de signification* suffisant. Concrètement, on pourra retenir entre autres, les conditions de COCHRAN à savoir $n_{ij} \succ 0, \forall (i,j)$, et pour 80% des cas, $n_{ij} \geq 5$.

B - Applications

1. Tests à un échantillon sous modèle ganssien

1.1 Test t de STUDENT et pluviométrie

Enoncé : Des relevés effectués pendant de nombreuses années ont permis d'établir que le niveau naturel des pluies dans la Beauce, en millimètres par an, suit une loi normale de moyenne 600.

On cherche à savoir si cette espérance mathématique est toujours la même. A cet effet, on dispose de mesures sur 9 années consécutives, les niveaux de précipitation ainsi relevés étant les suivants :

510	614	780	512	501	534	603	788	650

Que peut-on conclure au risque de première espèce $\alpha = 0,05$?

Solution : Il s'agit de tester $\begin{cases} H_0 : m = m_0 \\ H_1 : m \neq m_0 \end{cases}$ avec $m_0 = 600$, n petit, et σ inconnu, test classique dit « *t de STUDENT à un échantillon* ». Reprenant, le mode opératoire présenté en rappels de cours du présent chapitre (cf. paragraphe 2.3.a), la **région critique** du test a pour forme $W = \{(x_1, x_2, ..., x_n) / |\bar{x} - m| \geq \varepsilon \}$, $\bar{x}$ désignant la moyenne empirique des observations, soit $\bar{x} = \dfrac{\sum_{i=1}^{i=n} x_i}{n}$.

La donnée de l'erreur de première espèce α conduit à la relation qui permet d'expliciter ε et qui s'écrit $\Pr ob(|\bar{x} - m| \geq \varepsilon / m = m_0) = \alpha$. Se référant à la statistique $T = \dfrac{\overline{X} - m}{\hat{S}/\sqrt{n}}$ où $\hat{S}^2 = \dfrac{1}{n-1}.\left[\sum_{i=1}^{i=n} X_i^2 - n.\overline{X}^2\right]$, statistique dont il a été montré qu'elle suit la loi de STUDENT à $v = n-1$ degrés de libertés (cf. application 1.1 du chapitre I), on a immédiatement $\Pr ob(|T| \geq \dfrac{\varepsilon}{\hat{S}/\sqrt{n}}) = \alpha$.

La lecture dans la table des valeurs annexée du seuil $t_\alpha / \Pr ob(|T| \geq t_\alpha) = \alpha$ permet de déduire la valeur de ε égale à $t_\alpha . \hat{s}/\sqrt{n}$ et à fortiori, la **règle de décision** du test.

Ainsi décide-t-on H_1 si $\bar{x} \notin \left]m_0 - t_\alpha . \frac{\hat{s}}{\sqrt{n}}, m_0 + t_\alpha . \frac{\hat{s}}{\sqrt{n}}\right[$ et H_0 sinon. Numériquement, on obtient successivement $\bar{x} = \frac{\sum_{i=1}^{i=9} x_i}{9} = 610,22$ et $\hat{s}^2 = \frac{1}{9-1}.\left[\sum_{i=1}^{i=9} x_i^2 - 9.\bar{x}^2\right] \Rightarrow \hat{s} = 111,53$.

Par ailleurs, s'agissant du test bilatéral tel présentement, $\alpha = 0,05 \Rightarrow t_\alpha = 2,306$. D'où en définitive, la *région critique* $\bar{x} \notin]514,27 - 685,73[$.

Or, pour l'échantillon proposé, on a $\bar{x} = 610,22$. A l'évidence, on est loin de se trouver dans la zone de décision de H_1 (région critique) et il n'est donc pas permis de conclure à une éventuelle évolution de la pluviométrie moyenne ($m_0 = 600$).

1.2 Test de proportion et étude de marché

Enoncé : Le lancement d'un nouveau produit nécessitant une réorganisation complète d'un atelier (donc des dépenses importantes), une entreprise décide de faire une étude de marché sous forme de questionnaire dont on retient en particulier l'information suivante « la personne interrogée est ou n'est pas intéressée par le nouveau produit ».

Soit p la proportion (réelle mais inconnue) des personnes intéressées par le nouveau produit. L'entreprise juge qu'il est raisonnable de lancer ce nouveau produit si plus de 50% des personnes interrogées sont réellement intéressées. Elle décide donc, à partir des réponses obtenues sur un échantillon de taille $n = 100$, de faire le test $\begin{cases} H_0 : p = 0,5 \\ H_1 : p \succ 0,5 \end{cases}$.

1°) Préciser la signification des risques de 1ère et de 2ème espèce.

2°) Déterminer la région critique associée au test, au seuil de 1%, ceci pour l'échantillon de 100 personnes ainsi considéré.

3°) Que conclure si 58 des 100 personnes interrogées déclarent être intéressées par le nouveau produit ?

Solution : 1°) **L'erreur de première espèce** α =Prob(décider H_1 / H_0 vraie) décrit le risque de lancer à tort le nouveau produit. A contrario, **l'erreur de seconde espèce** β = Prob(décider H_0 / H_1 vraie) représente le risque associé au manque d'opportunité qui serait de ne pas lancer ce nouveau produit alors qu'en réalité l'intérêt est suffisant pour en justifier le développement.

2°) Le test proposé est un test **unilatéral** de la forme $\begin{cases} H_0 : p = p_0 \\ H_1 : p \succ p_0 \end{cases}$, avec $p_0 = 0,5$. Se référant aux rappels de cours du présent chapitre (cf. paragraphe 2.3.b), la **région critique** du test a pour forme $F_n \geq \pi$ où F_n désigne la fréquence $\frac{X}{n}$ des observations faites.

La taille élevée de n autorise ici le recours à la **loi normale** (cf. théorème de MOIVRE LAPLACE). Ainsi $\alpha = \Pr ob(F_n \geq \pi / p = p_0) = \Pr ob(\xi \geq \frac{\pi - p_0}{\sqrt{\frac{p_0.(1-p_0)}{n}}})$.

Par suite, $\pi = p_0 + t_\alpha \cdot \sqrt{\dfrac{p_0.(1-p_0)}{n}}$ où t_α vérifie $\Pr ob(\xi \geq t_\alpha) = \alpha$. Les données numériques $\alpha = 1\%, n = 100, p_0 = 0,5$ entraînent $t_\alpha = 2,33$ (cf. table de valeurs annexée de la fonction de répartition $\Pi(t) = \Pr ob(\xi \leq t)$), puis $\pi = 0,55$.

3°) L'enquête de satisfaction suivant laquelle 58 des 100 personnes interrogées sont intéressées par le nouveau produit suscite le lancement de celui-ci puisqu'on se trouve présentement dans la région critique (zone de décision de H_1) définie par $f_n \geq \pi = 0,55$, le risque d'erreur étant maîtrisé à 1% en la circonstance. Il n'y a donc pas à hésiter !.

1.3 Risques client et fournisseur

Enoncé : Un cahier des charges impose aux fabricants d'un conditionnement pharmaceutique une conservation égale à 7 degrés au bout de 80 heures. On admet que cette température de conservation à 80 heures, soit X, suit la loi normale $N(m,\sigma)$ dans laquelle on suppose que σ reste constant et égal à un degré.

Le client demande une expertise de la livraison qu'il reçoit, les enjeux en étant :

- H_0 : $m = m_0 = 7 \Rightarrow$ acceptation de la livraison ;
- $H_1 : m \neq m_0 \Rightarrow$ rejet de la livraison.

1°) On enregistre les meures de cette conservation à 80 heures, à travers un échantillon de $n = 40$ conditionnements réfrigérés, soient $(x_i), 1 \leq i \leq 40$. Désignant par $\bar{x} = \dfrac{\sum_{i=1}^{i=n} x_i}{n}$ la moyenne des observations effectuées, construire un intervalle de confiance dans lequel, sous l'hypothèse H_0, $\bar{x}$ est située dans plus de 99% des cas.

Qu'en conclure quant à la règle de décision du test considéré ?

2°) L'échantillon de 40 enregistrements susmentionné conduit à une température moyenne de conservation à 80 heures égale à 7,3 degrés. Quelle conclusion devra-t-on en tirer et suivant quel risque ?

3°) Un spécialiste du froid considère après études et essais que la température moyenne à 80 heures du conditionnement considéré n'est pas de $m = m_0 = 7$ degrés, mais de $m = m_1 = 7,2$ degrés (toujours cependant avec le même écart-type de un degré). Si cette affirmation était vraie, quelle serait alors la probabilité que la fabrication soit acceptée à tort par le client à la suite de l'application de la règle de décision établie au 1°) ?

4°) Que devient cette probabilité (risque de 2ème espèce) lorsqu'on fonde la décision d'acceptation ou de rejet de la livraison sur des échantillons de taille $n = 500$ et non plus $n = 40$?. Interpréter le résultat obtenu.

Solution : 1°) On a affaire dans ce problème, à un *test paramétrique bilatéral de conformité portant sur un modèle normal*, tels ceux décrits en rappels de cours du présent chapitre (cf. paragraphe 2.3.a).

Pour un tel test bilatéral, la caractérisation de la *région critique* équivaut à la détermination de *l'intervalle de confiance* au seuil $1-\alpha$.

En effet, sous l'hypothèse $H_0, \overline{X} = \dfrac{\sum_{i=1}^{i=n} X_i}{n}$ dont la loi est normale, soit $N(m, \dfrac{\sigma}{\sqrt{n}})$, a pour moyenne $m = m_0 = 7$ et pour écart-type $\sigma_{\overline{X}} = \dfrac{1}{\sqrt{40}} = 0{,}158$ (on suppose en effet, dans l'énoncé, $n = 40$ et $\sigma = 1$). Notant par t_α le seuil vérifiant $\Pr ob(-t_\alpha \le \xi \le t_\alpha) = 2.\Pi(t_\alpha) - 1 = 0{,}99$, il vient immédiatement, par lecture de la table annexée des valeurs de la fonction de répartition $\Pi(t) = \Pr ob(\xi \le t)$ de la variable normale centrée réduite, la valeur $t_\alpha = 2{,}575$.

Or, ξ, c'est aussi $\dfrac{\overline{X} - E(\overline{X})}{\sigma_{\overline{X}}} = \dfrac{\overline{X} - m}{\sigma/\sqrt{n}}$. Ainsi, sous l'hypothèse $H_0 : m = m_0 = 7$, a-t-on l'encadrement $-2{,}575 \le \dfrac{\overline{X} - 7}{0{,}158} \le 2{,}575 \Rightarrow 6{,}593 \le \overline{X} \le 7{,}407$.

Cet *intervalle de confiance* en fonction duquel $\Pr ob(|\overline{X} - 7| \ge 0{,}407) = 0{,}01$ correspond à l'intervalle défini par $\Pr ob(|\overline{X} - m_0| \ge \varepsilon) = \alpha$ et qui suivant les rappels de cours (cf. paragraphe 2.3) définit la *région critique du test bilatéral* $\begin{cases} H_0 : m = m_0 \\ H_1 : m \ne m_0 \end{cases}$.

2°) La valeur $\overline{x} = 7{,}3$ qui est observée est comprise dans la *région d'acceptation* de l'hypothèse H_0, soit $]6{,}593 - 7{,}407[$. *Il n'y a donc pas lieu ici de refuser la livraison.*

Par contre, si le risque α qui est la probabilité de refuser une livraison conforme, c'est à dire le **risque du fournisseur,** est **faible** puisque maîtrisé à 1%, il en est **autrement** du risque β encouru quand la livraison est acceptée et qui, étant égal à la probabilité d'accepter à tort une livraison non conforme, est le **risque du client.**

En fait, ce risque reste inconnu puisque fonction des valeurs de m sous l'hypothèse H_1 (cf. rappels de cours du présent chapitre, paragraphe 2.3.a, **courbe d'efficacité**).

3°) Dans l'hypothèse où on accepterait la fabrication et où on aurait en réalité $m = m_1 = 7{,}2$, on aurait $\beta = \Pr ob(6{,}593 \le \overline{X} \le 7{,}407 / m = 7{,}2)$, soit par passage à la variable normale centrée réduite $\xi = \dfrac{\overline{X} - 7{,}2}{\sigma_{\overline{X}}} = \dfrac{\overline{X} - 7{,}2}{0{,}158}$, un risque de 2ème espèce encouru finalement égal à $\beta = \Pr ob(\dfrac{6{,}593 - 7{,}2}{0{,}158} \le \xi \le \dfrac{7{,}407 - 7{,}2}{0{,}158}) = \Pi(1{,}31) - \Pi(-3{,}84)$, soit $\beta = \Pi(1{,}31) = 0{,}903$ (car $\Pi(-3{,}84)$ est quasiment négligeable).

On remarque donc ici toute **l'importance** du *risque pris par le client* dans les conditions de déroulement d'un tel test au contraire du *risque fournisseur* qui est très confortable.

4°) Pour **diminuer** *son risque*, le *client* doit disposer de **plus d'informations**. Ainsi, si $n = 500$, la nouvelle *région critique* du test, caractérisée par $\Pr ob(|\overline{X} - m_0| \ge \varepsilon) = \alpha$ où $\overline{X}$ suit la loi normale $N(m, \sigma/\sqrt{n})$ est donc définie par $|\overline{X} - m_0| \ge \varepsilon$, soit numériquement :

$\varepsilon = 2,575 \times \frac{1}{\sqrt{500}} = 0,115$. D'où la zone de rejet de H_0, à savoir, la *région critique*, $\bar{x} \notin]6,885 - 7,115[$.

Dans ces conditions, l'erreur de seconde espèce susmentionnée au 3°) pour la valeur de m égale à 7,2, vaut $\beta = Prob(\frac{6,885-7,2}{1/\sqrt{500}} \leq \xi \leq \frac{7,115-7,2}{1/\sqrt{500}}) = \Pi(-1,90) - \Pi(-7,05)$, soit $\beta = 1 - \Pi(1,90) = 2,9\%$ ($\Pi(-7,05)$ étant négligeable).

L'augmentation importante de n a permis ici de *ramener le risque du fournisseur à un niveau très faible, le risque du fournisseur restant bien évidemment invariant* et égal à 1%.

1.4 Test séquentiel de WALD portant sur une moyenne

Exposée par Abraham WALD en 1947, la théorie des tests séquentiels (ou progressifs) a eu un impact considérable sur le contrôle statistique des fabrications industrielles, ces méthodes permettant de diminuer jusqu'à 50% la taille des échantillons prélevés dans lesdits contrôles de qualité.

Enoncé : On considère une suite $(X_i), i \geq 1$ de variables aléatoires indépendantes équidistribuées de loi parente normale, soit $N(\theta, \sigma^2)$, la variance σ^2 étant supposée connue et θ étant un paramètre inconnu.

On souhaite tester $H_0 : \theta = \theta_0$ contre $H_1 : \theta = \theta_1$ où θ_0 et θ_1 sont donnés et vérifient $\theta_0 \prec \theta_1$.

1°) Notant par $f(x,\theta)$ la densité de probabilité de la variable parente de loi $N(\theta, \sigma^2)$ et posant $Z = \ln \frac{f(X,\theta_1)}{f(X,\theta_0)}$, calculer les espérances $E_{\theta_0}(Z)$ et $E_{\theta_1}(Z)$.

2°) On considère le test séquentiel du rapport de vraisemblance fondé sur l'observation de la suite $(X_i), i \geq 1$, et dont les erreurs de première et de seconde espèce sont supposées fixées et égales respectivement à α et β. On note par N le nombre aléatoire des observations effectuées.

Expliciter des valeurs approchées de $E_{\theta_0}(N)$ et $E_{\theta_1}(N)$.

3°) En vue d'essais métallurgiques, on a composé un alliage devant contenir 2,4% d'un élément donné. Pour vérifier cette préparation on effectue des dosages avec une méthode rapide et fiable, l'écart- type étant supposé connu et égal à $\sigma = 0,5$. Par ailleurs, l'hypothèse de normalité de la teneur en question est admise.

On se fixe :

- à la valeur 5%, le risque de première espèce α, qui est le risque de refuser un alliage dont la concentration serait bien égale à $\theta = \theta_0 = 2,4\%$;
- à la valeur 5%, le risque de deuxième espèce β, d'accepter un alliage qui contiendrait la teneur $\theta = \theta_1 = 2,8\%$ de l'élément considéré au lieu des 2,4% prévus initialement.

a) Expliciter numériquement la règle de décision du test séquentiel correspondant.

b) Les premières mesures effectuées conduisent à la série de valeurs ci-dessous :

n	1	2	3	4	5	6	7	8	9	10
x_n	2,875	2,280	2,390	2,265	2,400	3,150	1,400	2,090	2,403	2,403

Qu'en conclure et au bout de combien d'observations ?

c) Calculer les valeurs approchées de $E_{\theta_0}(N)$ et $E_{\theta_1}(N)$.

d) Quelle serait la taille n_0 de l'échantillon requise dans le cadre d'un test classique suivant la méthode de NEYMAN et PEARSON ?

e) Comparer $E_{\theta_0}(N)$ et $E_{\theta_1}(N)$ à n_0. Qu'en conclure ?

Solution : 1°) De la définition $f(X,\theta)=\frac{1}{\sigma.\sqrt{2\pi}}.\exp(-\frac{1}{2.\sigma^2}.(X-\theta)^2)$, on obtient aisément $\ln\frac{f(X,\theta_1)}{f(X,\theta_0)}=\frac{(\theta_1-\theta_0)}{\sigma^2}.\left[X-\frac{\theta_0+\theta_1}{2}\right]$.

Il en résulte, par linéarité de l'espérance mathématique, $E_\theta\left[\ln\frac{f(X,\theta_1)}{f(X,\theta_0)}\right]=\frac{\theta_1-\theta_0}{\sigma^2}.\left[E_\theta(X)-\frac{\theta_0+\theta_1}{2}\right]$, soit pour les cas particuliers, $\theta=\theta_0$ et $\theta=\theta_1$, les expressions $E_{\theta_0}\left[\ln\frac{f(X,\theta_1)}{f(X,\theta_0)}\right]=-\frac{(\theta_1-\theta_0)^2}{2.\sigma^2}$ et $E_{\theta_1}\left[\ln\frac{f(X,\theta_1)}{f(X,\theta_0)}\right]=\frac{(\theta_1-\theta_0)^2}{2.\sigma^2}$.

2°) Se référant aux rappels de cours du présent chapitre (cf. paragraphe 2.7) et considérant les *rapports de vraisemblance* $R_1,R_2,...,R_N,..$ où $R_N=\frac{f(x_1,x_2,...,x_N,\theta_1)}{f(x_1,x_2,...,x_N,\theta_0)}$, il est rappelé que le test de WALD conduit au processus de décision suivant :

- $R_N \prec \frac{\beta}{1-\alpha}$ on décide H_0 ;
- $R_N \succ \frac{1-\beta}{\alpha}$ on décide H_1 ;
- $\frac{\beta}{1-\alpha} \le R_N \le \frac{1-\beta}{\alpha}$ on procède à une observation supplémentaire, aucune décision n'étant prise.

Or, l'indépendance des observations en fonction desquelles la densité de probabilité du N - uplet est égale au produit des densités de probabilité entraîne $\ln R_N=\sum_{i=1}^{i=N}\ln\frac{f(X_i,\theta_1)}{f(X_i,\theta_0)}$ et, par passage à l'espérance mathématique et linéarité, $E_\theta[\ln R_N]=\sum_{i=1}^{i=N}E_\theta\left[\ln\frac{f(X_i,\theta_1)}{f(X_i,\theta_0)}\right]$.

En définitive, et compte tenu des résultats de la 1ère question, $E_\theta[\ln R_N]=\frac{N.(\theta_1-\theta_0)}{\sigma^2}.\left[E_\theta(X)-\frac{\theta_0+\theta_1}{2}\right]$.

En particulier, pour $\theta = \theta_0$ et $\theta = \theta_1$, on a respectivement les résultats $E_{\theta_0}\left[\ln R_N\right] = -N.\frac{(\theta_1-\theta_0)^2}{2.\sigma^2}$ et $E_{\theta_1}\left[\ln R_N\right] = N.\frac{(\theta_1-\theta_0)^2}{2.\sigma^2}$, expressions desquelles on déduit, entre autres, la relation $N = \frac{-2.\sigma^2}{(\theta_1-\theta_0)^2}.E_{\theta_0}\left[\ln R_N\right]$.

• Lorsque le nombre d'observations nécessaires pour conclure est égal à n, c'est à dire lorsque $N = n$, et que par ailleurs l'hypothèse $H_0 : \theta = \theta_0$ est satisfaite, on se trouve être nécessairement dans l'une des deux éventualités $R_n \prec \frac{\beta}{1-\alpha}$ ou $R_n \succ \frac{1-\beta}{\alpha}$, ceci respectivement avec les probabilités Prob(décider H_0 / H_0 vraie), soit $1-\alpha$, et Prob (décider H_1 / H_0 vraie), soit α.

Assimilant les valeurs de R_n aux bornes susmentionnées (ce qui est admissible car on peut supposer que les valeurs de R_n ne s'écartent que faiblement de ces bornes), on a donc pour la variable aléatoire $E_{\theta_0}\left[\ln R_N\right]$ de valeurs $\ln\frac{\beta}{1-\alpha}$ et $\ln\frac{1-\beta}{\alpha}$ et de probabilités associées $1-\alpha$ et α, l'espérance mathématique :

$$E_N\left[E_{\theta_0}(\ln R_N)\right] = (1-\alpha).\ln\frac{\beta}{1-\alpha} + \alpha.\ln\frac{1-\beta}{\alpha}$$

Du résultat ci-dessus et de la relation précédente $N = \frac{-2.\sigma^2}{(\theta_1-\theta_0)^2}.E_{\theta_0}\left[\ln R_N\right]$, on déduit en définitive, l'expression approchée $E_{\theta_0}(N) = \frac{-2.\sigma^2}{(\theta_1-\theta_0)^2}.\left[(1-\alpha).\ln\frac{\beta}{1-\alpha} + \alpha.\ln\frac{1-\beta}{\alpha}\right]$.

De même, $E_{\theta_1}(N) = \frac{2.\sigma^2}{(\theta_1-\theta_0)^2}.\left[\beta.\ln\frac{\beta}{1-\alpha} + (1-\beta).\ln\frac{1-\beta}{\alpha}\right]$.

3-a) Reprenant les résultats des rappels de cours (cf. paragraphe 2.7 du présent chapitre) relativement au test séquentiel portant sur une moyenne, la **règle de décision** à mettre en œuvre est la suivante :

- si $\sum_{i=1}^{i=n} x_i \prec n.q - h_1$ on décide H_0 ;
- si $\sum_{i=1}^{i=n} x_i \succ n.q - h_2$ on décide H_1 ;
- si $n.q - h_1 \le \sum_{i=1}^{i=n} x_i \le n.q + h_2$ on procède à une observation supplémentaire avant toute décision éventuelle, les quantités q, h_1, h_2 étant respectivement égales à $q = \frac{\theta_0+\theta_1}{2}, h_1 = \frac{\sigma^2}{\theta_1-\theta_0}.\ln\frac{1-\alpha}{\beta}$, et $h_2 = \frac{\sigma^2}{\theta_1-\theta_0}.\ln\frac{1-\beta}{\alpha}$.

Dans le cadre numérique proposé, $\alpha = \beta = 5\% - \sigma = 0,5 - \theta_0 = 2,4 - \theta_1 = 2,8$. La mise en œuvre du test conduit à tracer les droites $A(n)$ et $B(n)$ dont les équations sont explicitées ci-après.

- $A(n) = n.q - h_1$, soit numériquement, $A(n) = 2,6.n - 1,84$;

- $B(n) = n.q + h_2$, soit numériquement, $B(n) = 2,6.n + 1,84$.

Le *positionnement* des points représentatifs $(p, \sum_{i=1}^{p} x_i)$ entre ou à l'extérieur de ces droites permet de conclure à la *poursuite des observations* ou à une *décision.*

3-b) En fonction des mesures effectuées et de leurs résultats x_i, la somme $\sum_{i=1}^{i=n} x_i$ et les frontières $A(n)$ et $B(n)$ sont calculées pour les valeurs successives de n comme indiqué ci-dessous :

n	1	2	3	4	5	6	7	8	9	10
x_n	2,875	2,280	2,390	2,265	2,400	3,150	1,400	2,090	2,403	2,403
$\sum_{i=1}^{i=n} x_i$	2,875	5,155	7,545	9,810	12,210	15,360	16,760	18,850	21,253	23,656
$A(n)$	0,76	3,36	5,96	8,56	11,16	13,76	16,36	18,96	21,56	24,16
$B(n)$	4,44	7,04	9,64	12,24	14,84	17,44	20,04	22,64	25,24	27,84

On observe que pour les sept premières observations, on a $A(n) \leq \sum_{i=1}^{i=n} x_i \leq B(n)$. Par contre, à la 8ème observation, on a $\sum_{i=1}^{i=n} x_i \succ A(n)$ ce qui permet de conclure à l'hypothèse H_0 au bout d'un *nombre relativement faible d'observations.*

- A noter que le tracé des droites $A(n), B(n)$, et des points $M_n(n, \sum_{i=1}^{i=n} x_i)$ aurait fourni également une *solution graphique* tout aussi simple, le *test séquentiel de* WALD permettant de conclure dès qu'on sort de la zone d'incertitude qui est située entre les deux droites en question (cf. paragraphe 2.7 des rappels de cours).

3-c) Reprenant les expressions antérieures de $E_{\theta_0}(N), E_{\theta_1}(N)$, ainsi que celles de h_1 et de h_2, il est immédiat que :

- $E_{\theta_0}(N) = \frac{2.\sigma^2}{(\theta_1 - \theta_0)^2}.\left[(1-\alpha).\ln\frac{1-\alpha}{\beta} - \alpha.\ln\frac{1-\beta}{\alpha}\right] = \frac{2}{\theta_1 - \theta_0}.\left[(1-\alpha).h_1 - \alpha.h_2\right]$;

- $E_{\theta_1}(N) = \frac{2.\sigma^2}{(\theta_1 - \theta_0)^2}.\left[\beta.\ln\frac{\beta}{1-\alpha} + (1-\beta).\ln\frac{1-\beta}{\alpha}\right] = \frac{2}{\theta_1 - \theta_0}.\left[(1-\beta).h_2 - \beta.h_1\right]$.

On a donc, numériquement, $E_{\theta_0}(N) = E_{\theta_1}(N) = 8,28$ soit en arrondissant à l'entier le plus proche, la valeur commune 8.

- En fait, de façon générale, et pour le **test composite** $\begin{cases} H_0 : \theta \leq \theta_0 \\ H_1 : \theta \succ \theta_1 \end{cases}$, on montre que

$E_\theta(N) = \dfrac{\left[\ln\frac{\beta}{1-\alpha}\right].P(\theta) + \left[\ln\frac{1-\beta}{\alpha}\right].(1-P(\theta))}{E_\theta(Z)}$ avec $Z = \ln\dfrac{f(X,\theta_1)}{f(X,\theta_0)}$ et où $P(\theta)$ désigne la probabilité de conclure en faveur de θ.

La fonction précédente qui passe par les points $E_{\theta_0}(N)$ et $E_{\theta_1}(N)$ calculés ci-dessus, atteint son *maximum* au voisinage de la valeur $\theta = q = \frac{\theta_0 + \theta_1}{2}$, soit un *extrema* dont on peut montrer qu'il est égal à $E_q(N) = \frac{h_1 h_2}{\sigma^2}$. Pour le cas présent et numériquement, il en résulte $E_\theta(N) \leq E_q(N) = 13{,}54$.

La valeur de ce maximum reste donc *faible* en tout état de cause et tend à *privilégier le test séquentiel au test traditionnel*, comme le confirme le calcul de la question suivante.

3-d) Dans le cadre d'un **test classique** suivant la méthode de NEYMAN et PEARSON, la donnée des erreurs de première espèce et de seconde espèce conduit, pour le test considéré, aux relations :

- $\alpha = \text{Prob (décider } H_1 / H_0 \text{ vraie)} = Prob(\bar{x} \geq \pi / \theta = \theta_0)$;
- $\beta = \text{Prob (décider } H_1 / H_0 \text{ vraie)} = Prob(\bar{x} \prec \pi / \theta = \theta_1)$.

L'hypothèse de **normalité** qui a été admise, permet donc d'écrire relativement à la variable normale centrée réduite $\xi = \frac{\bar{x} - \theta}{\sigma / \sqrt{n}}$, les conditions $\alpha = Prob(\xi \geq \frac{\pi - \theta_0}{\sigma / \sqrt{n}})$ et $\beta = Prob(\xi \prec \frac{\pi - \theta_1}{\sigma / \sqrt{n}})$. Notant par $t_{1-\alpha}$ et t_β les nombres qui vérifient respectivement $1 - \alpha = Prob(\xi \prec t_{1-\alpha})$ et $\beta = Prob(\xi \leq t_\beta)$, il vient immédiatement $\begin{cases} \frac{\pi - \theta_0}{\sigma / \sqrt{n}} = t_{1-\alpha} \\ \frac{\pi - \theta_1}{\sigma / \sqrt{n}} = t_\beta \end{cases}$, d'où les valeurs de π et de n par résolution.

Concrètement, on a $t_{0,95} = 1{,}645$, $t_{0,05} = -1{,}645$, $\frac{\pi - 2{,}4}{\pi - 2{,}8} = -1 \Rightarrow \pi = 2{,}6$. D'où, $n = \frac{t_{0,95}^2 . \sigma^2}{(\pi - \theta_0)^2}$, soit numériquement, $n = 16{,}92$. Cette valeur qu'on arrondira par excès à $n_0 = 17$ constitue la **taille minimale de l'échantillon** requis pour contenir à la valeur 5% les risques α et β.

3-e) Les calculs précédents révèlent donc la nécessité de disposer au moins de 17 observations dans le *test classique* alors qu'en moyenne, le *test séquentiel* permet de conclure en 13 observations. En moyenne, *ce dernier test* est donc **préférable**.

• Il faut noter néanmoins qu'avec le test séquentiel, la valeur $E_\theta(N) = 13$, constitue un *nombre moyen aléatoire* et non une certitude. En d'autres termes, on peut-être amené à des situations dans lesquelles N *dépasse largement* $E_\theta(N)$, voire même n_0, ce qui ferait perdre tout intérêt au test séquentiel.

C'est pourquoi, **en pratique**, *on tronque le test séquentiel* lorsque N atteint 2,5 fois la valeur moyenne maximale $E_\theta(N)$, l'hypothèse H_0 étant acceptée ou refusée suivant qu'on se trouve en dessous ou au dessus de la droite qui est parallèle aux droites $A(n)$ et $B(n)$ et qui passent par l'origine (droite d'équation $y = n.q$).

1.5 Ajustements par la loi normale

Enoncé : Les données ci-dessous représentent le taux d'oxygénation d'un cours d'eau à 20° dans une certaine région. Soient $x_i, (1 \le i \le n)$, les données ainsi considérées de la variable aléatoire X en question et $f_i, (1 \le i \le n)$, les fréquences absolues associées.

x_i : taux d'oxygénation par jour	f_i : fréquence
$x_i \le 0,1$	12
$0,1 \prec x_i \le 0,15$	20
$0,15 \prec x_i \le 0,20$	23
$0,20 \prec x_i \le 0,25$	15
$x_i \succ 0,25$	13

On se propose de tester, au seuil $\alpha = 5\%$, la normalité ou non de cette distribution en utilisant successivement :

1°) le test de chi- deux ;

2°) le test de KOLMOGOROV ;

3°) la méthode graphique de la droite de HENRY.

Solution : 1°) **Préalablement** à la *mise en œuvre des différents test proposés*, il convient **d'estimer** les paramètres m et σ de la *loi normale* $N(m,\sigma)$ par laquelle on souhaite modéliser la variable X « *taux d'oxygénation* », ce qui conduit au tableau de calculs ci-dessous (les classes extrêmes sont assimilées en la circonstance aux classes $]0,05-0,1]$ et $]0,25-0,30]$ d'amplitude commune 0,05 et toutes les classes sont assimilées à leurs centres respectives, soient x_i^*.

Classe i	Centre x_i^*	Fréquence f_i	$f_i.x_i^*$	$f_i.x_i^{*2}$
$x_i \le 0,1$	0,075	12	0,9	0,0675
$0,1 \prec x_i \le 0,15$	1,125	20	2,5	0,3125
$0,15 \prec x_i \le 0,20$	0,175	23	4,025	0,7044
$0,20 \prec x_i \le 0,25$	0,225	15	3,375	0,7594
$x_i \succ 0,25$	0,275	13	3,575	0,9831
Σ		83	14,375	2,8269

Notant par N le nombre de données collectées (taille de l'échantillon égale à $\sum_{i=1}^{i=n} f_i$) et par n le nombre de classes formées dans le tableau précédent $(n=5)$, il s'ensuit, relativement aux estimations respectives de m et de σ^2, soient $\bar{x} = \frac{1}{N}.\sum_{i=1}^{i=n} f_i.x_i$ et $\hat{s}^2 = \frac{1}{N-1}.\left[\sum_{i=1}^{i=n} f_i.x_i^2 - N.\bar{x}^2\right]$, les valeurs $\bar{x} = 0,1732$ et $\hat{s} = 0,064$.

• Revenant au **test de chi- deux** qui est la première méthode suggérée ici, il convient de signaler que sa *mise en œuvre est plutôt à réserver aux variables discrètes*. Mais, *quand les données sont regroupées en classes* comme c'est le cas ici, le test est *applicable*, les classes extrêmes devant cependant rester ouvertes (ce qui est vérifié pour le cas proposé).

Cette application du test de χ^2 conduit au tableau de calculs ci-dessous :

Classe i	f_i	p_i(*)	$N.p_i$
$x_i \leq 0,1$	12	0,1271	10,55
$0,1 \prec x_i \leq 0,15$	20	0,2323	19,28
$0,15 \prec x_i \leq 0,20$	23	0,3034	25 ,18
$0,20 \prec x_i \leq 0,25$	15	0,2221	18,43
$x_i \succ 0,25$	13	0,1151	9,56
Σ	83	1,0000	83

(*) Pour ce qui est du calcul des p_i associés aux classes et ceci suivant la loi théorique $N(m=0,1732;\sigma=0,064)$, on a ,par exemple pour la classe $]0,10-0,15[$, $\Pr ob(0,1 \prec X \leq 0,15) = \Pr ob(\frac{0,1-0,1732}{0,064} \prec \xi \leq \frac{0,15-0,1732}{0,064})$, soit en utilisant les tables de valeurs annexée de la fonction $\Pi(t) = \Pr ob(\xi \leq t)$, le résultat $\Pr ob(0,1 \prec X \leq 0,15) = \Pi(-0,3616) - \Pi(-1,1414) = 0,2323$. Et ainsi de suite....

Par ailleurs, pour les classes extrêmes (qui doivent être considérées comme ouvertes dans le calcul d'ajustement), on a par exemple :

$$\Pr ob(X \leq 0,1) = \Pr ob(\xi \leq \frac{0,1-0,1732}{0,064} = -1,1414) = 0,1271.$$

Enfin, toutes les classes ayant une fréquence théorique $N.p_i$ supérieure à 5, *il n'a pas été nécessaire, pour l'exemple proposé, de procéder à des regroupements.*

• En conclusion, $\chi^2_{calculé} = \sum_{i=1}^{i=5} \frac{(f_i - N.p_i)^2}{N.p_i} = 2,30$, la statistique associée à la *distance de chi- deux* suivant en la circonstance la loi du χ^2 à $\nu = 5-1-2$ degrés de libertés (puisque l'ajustement par la loi normale va de pair avec l'estimation de deux paramètres et que le nombre de classes est par ailleurs de 5 dans l'exemple traité).

Raisonnant à partir d'une *erreur de première espèce* α égale à 5% et notant par χ^2_α le nombre vérifiant $\Pr ob(\chi^2 \geq \chi^2_\alpha) = \alpha$, il vient immédiatement, par lecture dans la table des valeurs annexée, $\chi^2_\alpha = 5,991$.

On constate que $\chi^2_{calculé} = 2,30$ est inférieur au seuil $\chi^2_\alpha = 5,991$ ce qui *n'autorise pas à rejeter* l'hypothèse d'ajustement par la loi normale. L'éloignement de $\chi^2_{calculé}$ par rapport au seuil critique χ^2_α laisse envisager avec confiance la validité de cet ajustement par la loi normale. Mais, il reste néanmoins impossible de calculer l'erreur de 2ème espèce (et à fortiori la puissance du test) en absence d'hypothèses plus précises quant au type de loi suivie par X sous l'hypothèse alternative H_1.

2°) Lorsque les données sont réparties en classes, le test de KOLMOGOROV qui conduit à *comparer fréquences cumulées* **théoriques**, soient $F_0(x_i)$, et *fréquences cumulées* **observées**, soient $\widehat{F}(x_i)$, s'explicite à travers la **distance** :

$$D(F_0, \widehat{F}) = \underset{i=1,2,...,k}{Sup} \left\{ \left| F_0(x_i) - \widehat{F}(x_{i-1}) \right|, \left| F_0(x_i) - \widehat{F}(x_i) \right| \right\}$$

(k désignant le nombre des classes et $x_i\,(1 \le i \le k)$, les valeurs de X pour chacune de ces classes.

Pour les données constatées, les valeurs des classes étant définies par les bornes supérieures 0,1 – 0,15 – 0,20..., on a par exemple, $\widehat{F}(0,1) = \frac{12}{83} = 0,1446$, $\widehat{F}(0,15) = \frac{32}{83}$....
D'autre part, $F_0(0,1) = p_0 = 0,1271$; $F_0(0,15) = p_0 + p_1 = 0,3594$ ....

- En résumé, on obtient le tableau de calculs ci-dessous :

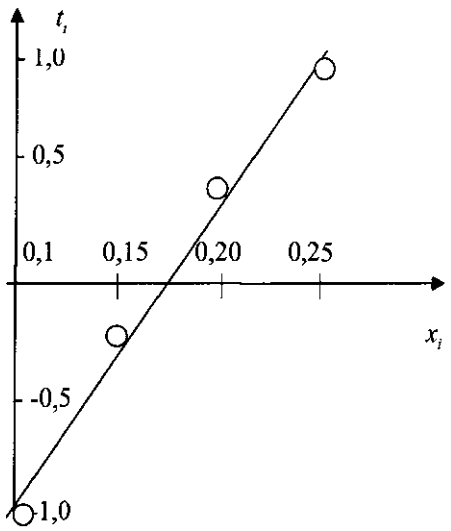
x_i	$\widehat{F}(x_i)$	$F_0(x_i)$	$\left\|F_0(x_i) - \widehat{F}(x_i)\right\|$	$\left\|F_0(x_i) - \widehat{F}(x_{i-1})\right\|$
0,1	0,1446	0,1271	0,01748	0,1271
0,15	0,3855	0,3594	0,02614	0,2148
0,20	0,6627	0,6628	0,00015	0,2772
0,25	0,8434	0,8849	0,04153	0,2222
$+\infty$	1	1	0	0,1566

Pour l'échantillon en question, la *distance de* KOLGOMOROV, soit $D(F_0, \widehat{F}) = \underset{i=1,2,...,5}{Sup} \left\{ \left| F_0(x_i) - \widehat{F}(x_{i-1}) \right|, \left| F_0(x_i) - \widehat{F}(x_i) \right| \right\}$ est donc égale numériquement à 0,2772.

Par ailleurs, la table de valeurs annexée (table de KOLMOGOROV pour un échantillon), fournit pour le test *bilatéral* et $n = 5$, $\alpha = 0,05$, la valeur D_α vérifiant $\Pr ob(D(F_0, \widehat{F}) \ge D_\alpha) = \alpha$. Il vient ainsi $D_\alpha = 0,565$.

La **région critique** du test étant caractérisée par $D_{calculé} \ge D_\alpha$ et les résultats numériques obtenus conduisant à $D_{calculé} = 0,2772 \prec D_\alpha = 0,565$, on conclut ici encore *qu'il n'y a pas lieu de rejeter l'hypothèse d'ajustement des données par la loi normale.*

3°) La **méthode graphique** de la *droite de* HENRY entraîne pour sa mise en œuvre et à partir des données précédentes, le tableau de calculs ci-après, les nombres t_i vérifiant les relations $\widehat{F}(x_i) = \Pr ob(\xi \le t_i), i = 1,2...,5$.



Valeurs x_i	$\widehat{F}(x_i)$	t_i
0,10	0,1446	-1,06
0 ,15	0,3855	-0,29
0,20	0,6627	+0,42
0,25	0,8434	+1,01
$+\infty$	1	$+\infty$

Le tracé de la droite de HENRY montre un alignement manifeste des points représentatifs $M_i(x_i,t_i)$ qui laisse envisager un ajustement par la loi normale.

Plus précisément, cette dernière a pour moyenne l'abscisse de l'intersection de la droite de HENRY (dont il est rappelé que l'équation est $t_i = \frac{x_i - m}{\sigma}$) avec l'axe des x_i, soit $m = 0{,}1725$.

Quant à l'écart- type, il est estimé par l'inverse de la pente de la droite de HENRY, soit $\sigma = 0{,}068$. Bien que le graphique soit très imprécis, les valeurs qu'on obtient ainsi sont très proches de celles obtenues par les calculs (cf. 1ère question).

2. Tests à un échantillon sous autres modèles

2.1 Test paramétrique pour le modèle de POISSON

Enoncé : On dispose d'un échantillon de n variables aléatoires indépendantes, soit $(X_1, X_2, ..., X_n)$ de loi parente X suivant une loi de POISSON de paramètre θ.

On souhaite tester, suivant l'erreur de première espèce $\alpha = 0{,}05$, l'hypothèse $H_0 : \theta = \theta_0 = 0{,}5$ contre l'hypothèse $H_1 : \theta = \theta_1 = 1$.

1°) A partir de la méthode de NEYMAN et PEARSON, montrer que la statistique de décision du test est la moyenne empirique $\overline{X}$, résultat dont on indiquera une justification.

2°) On suppose disposer d'un échantillon de taille $n = 10$. Déterminer la région critique du test au seuil $\alpha = 5\%$ et exprimer, par exemple, la conclusion à retenir lorsque les données dont on dispose conduisent à la somme $\sum_{i=1}^{i=10} x_i = 9$.

3°) Partant cette fois d'un échantillon de taille $n = 100$, on suppose avoir $\bar{x} = \frac{\sum_{i=1}^{i=100} x_i}{100} = 0{,}8$.
Quelle conclusion faut-il retenir (toujours suivant l'hypothèse $\alpha = 5\%$) ?

4°) Calculer la puissance du test pour chacune des valeurs précédentes de n.

5°) Quelle est la taille minimale de l'échantillon dont on doit disposer, soit n^*, pour satisfaire à la fois $\alpha \leq 0{,}05$ et $\beta \leq 0{,}10$?

Solution : 1°) Rappelant que $\Pr ob(X=x)=\frac{\theta^x.e^{-x}}{x!}$, le *rapport des fonctions de vraisemblance* $\frac{L(x_1,x_2,...,x_n,\theta_0)}{L(x_1,x_2,...,x_n,\theta_1)}$, soit $\frac{L_0}{L_1}$ suivant une transcription allégée, est égal à $\frac{\prod_{i=1}^{i=n}\theta_0^{x_i}.e^{-\theta_0}/x_i!}{\prod_{i=1}^{i=n}\theta_1^{x_i}.e^{-\theta_1}/x_i!}$.

Selon le *théorème de* NEYMAN et PEARSON, le test dont la *région critique* est définie par $W=\left\{(x_1,x_2,...,x_n)/\frac{L_0}{L_1}\le k\right\}$ est de **puissance maximale**.

Passant aux *log- vraisemblance*, la construction susmentionnée de W conduit donc à la relation, $(\ln\theta_0-\ln\theta_1).\sum_{i=1}^{i=n}x_i-(\theta_0-\theta_1)\le\ln k$, ce qui, compte tenu que $\theta_1\succ\theta_0$, entraîne $\sum_{i=1}^{i=n}x_i\ge\frac{1}{\ln\theta_1-\ln\theta_0}.\left[-\ln k+\theta_1-\theta_0\right]$. En résumé, la **région critique** est définie ici par une équation de la forme $\sum_{i=1}^{i=n}x_i\ge K'$ ou encore $\bar{x}\ge K$ (avec $\bar{x}=\frac{\sum_{i=1}^{i=n}x_i}{n}$ et $K=\frac{K'}{n}$).

• Le paramètre θ sur lequel porte le présent test est l'espérance mathématique de X, soit $E(X)$. Il est donc assez logique que la règle de décision correspondante ait pour *fonction discriminante*, la statistique $\overline{X}$ (comme cela a été montré directement ci-dessus), puisque $\overline{X}$ est *l'estimateur ponctuel* (E.M.V) de $E(X)$.

2°) $\alpha=\Pr ob(décider\ \ H_1/H_0\ \ vraie)$, soit $\alpha=\Pr ob(\sum_{i=1}^{i=n}x_i\ge K'/\theta=\theta_0)$. Mais, si tous les $X_i,(1\le i\le n)$, ont pour loi parente, la loi de POISSON de paramètre θ, leur somme $S=\sum_{i=1}^{i=n}X_i$ suit la loi de POISSON de paramètre $n.\theta$ (puisque la loi de POISSON est *stable pour l'addition* lorsque les variables X_i sont *indépendantes*).

Numériquement, il s'agit donc de trouver le seuil K' (K' entier), tel que $\Pr ob(S\ge K')=0,05$ où S suit la loi de POISSON de paramètre $n.\theta_0$, soit en l'occurrence la valeur 5. Utilisant à cet effet, des tables de valeurs portant sur la fonction de répartition $\Pr ob(X\le x)$ de ladite loi (cf. tables ou calculateurs en ligne), on constate que $\Pr ob(X\le x)$ dépasse la valeur $1-\alpha=95\%$, pour x compris entre 8 et 9.

Une valeur réelle obtenue par interpolation linéaire n'ayant pas de sens ici puisque les X_i sont des variables entières, on retiendra donc le *seuil critique* $K'=9$ qui conduit à une *erreur de première espèce* égale à $\alpha=1-0,9682=3,18\%$.

• Pour l'échantillon dont on dispose $\sum_{i=1}^{i=9}x_i=9$, c'est-à-dire la borne inférieure de la région critique dans laquelle on décide H_1. C'est donc cette dernière hypothèse $\theta=\theta_1=1$ qu'on retiendra.

3°) Cette fois, c'est par rapport à la loi de POISSON de paramètre $n.\theta_0 = 100 \times 0,5 = 50$ qu'il convient de déterminer le seuil K' (ou le seuil K si on raisonne sur $\bar{x}$). Toutefois, eu égard à la *grande valeur* de n, on peut admettre ici que le **théorème central limite** s'applique et qu'en conséquence, la moyenne empirique $\overline{X} = \frac{\sum_{i=1}^{i=n} X_i}{n}$ converge, sous l'hypothèse H_0, vers la **loi normale** de moyenne $\theta_0 = 0,5$ et de variance $Var(\overline{X}) = \frac{1}{n^2}.n.\theta_0 = \frac{\theta_0}{n}$ (numériquement, $Var(\overline{X}) = \frac{0,5}{100}$).

On a donc $\Pr ob(\overline{X} \geq K / \theta = \theta_0) = 0,05$, ce qui s'écrit aussi, suivant la variable normale centrée réduite ξ associée à $\overline{X}$, $\Pr ob(\xi \geq \frac{K - \theta_0}{\sqrt{\frac{\theta_0}{n}}}) = 0,05$. Ainsi, obtient-on $K = \theta_0 + t_\alpha \sqrt{\frac{\theta_0}{n}}$, où t_α désigne le nombre vérifiant $\Pr ob(\xi \geq t) = 0,05$.

Numériquement, il vient $t_\alpha = 1,645$ et $K = 0,616$. Or, pour l'échantillon considéré dans l'énoncé, la valeur observée de la moyenne empirique est $\bar{x} = 0,8$. Dans la mesure où $\bar{x} = 0,8 \geq K = 0,616$, on est conduit ici encore à *retenir le choix de l'hypothèse* H_1.

4°) Revenant au cas $n = 10$, l'erreur de 2ème espèce β est égale à la probabilité $\Pr ob(\textit{décider } H_0 / H_1 \textit{ vraie})$, soit $\beta = \Pr ob(S \prec K' / \theta = \theta_1 = 1)$, où S suit la loi de POISSON de paramètre $n.\theta_1 = 10$ et où $K' = 9$.

Par lecture dans une table de valeurs ou par calcul direct, il en résulte immédiatement $\beta = 0,333$ ce qui entraîne, pour la *puissance du test*, la valeur $\eta = 1 - \beta = 0,666$.

- D'autre part, lorsque $n = 100$ et *utilisant l'approximation par la loi normale*, on a $\beta = \Pr ob(\overline{X} \prec K / \theta = \theta_1)$, avec $K = 0,616$, soit $\beta = \Pr ob(\xi \prec \frac{0,616 - 1}{\sqrt{1/100}} = -3,84)$, c'est à dire moins de 10^{-4}. Autrement dit, *la puissance du test avoisine alors la valeur maximale unité.*

5°) On cherche la valeur minimale de n telle qu'on ait simultanément :

$$\begin{cases} \Pr ob(\overline{X} \geq K / \theta = \theta_0 = 0,5) \leq 0,05 \\ \Pr ob(\overline{X} \prec K / \theta = \theta_1 = 1) \leq 0,10 \end{cases}$$

Admettant la distribution normale de $\overline{X}$ (par convergence), il en résulte, par passage à la variable normale centrée réduite ξ de loi $N(0,1)$:

$$\begin{cases} \Pr ob(\xi \geq \frac{K - 0,5}{\sqrt{0,5/n}}) \leq 0,05 \\ \Pr ob(\xi \prec \frac{K - 1}{\sqrt{1/n}}) \leq 0,10 \end{cases}$$

A partir des nombres $t_{0,95}$ et $t_{0.10}$ vérifiant respectivement $Prob(\xi \leq t_{0,95}) = 0,95$ et $Prob(\xi \prec t_{0,10}) = 0,10$, soient $t_{0,95} = 1,645$ et $t_{0.10} = -1,28$, il découle aisément les relations $\frac{K-0,5}{\sqrt{0,5/n}} \geq 1,645$ et $\frac{K-1}{\sqrt{1/n}} \leq -1,28$. Partant du système $\begin{cases} \frac{K-0,5}{\sqrt{0,5/n}} = 1,645 \\ \frac{K-1}{\sqrt{1/n}} = -1,28 \end{cases}$, il vient immédiatement par quotient, $\frac{K-0,5}{\sqrt{0,5.(K-1)}} = -\frac{1,645}{1,28} \Rightarrow K = 0,738$ et, par substitution, $n = (-\frac{1,28}{K-1})^2 = 23,87$.

Arrondissant par excès à l'entier le plus proche, les inégalités susmentionnées, sont satisfaites dès que $n \geq n^* = 24$.

• La valeur n^* ainsi trouvée étant inférieure à 30, on peut douter de la validité de l'approximation par la loi normale. A défaut, c'est par **approximations successives** à partir de conditions ci-dessus qu'il faut procéder, à savoir, pour rappel, les conditions $Prob(\sum_{i=1}^{i=n} X_i \geq K'/\theta = \theta_0) = \alpha \leq 0,05$ et $Prob(\sum_{i=1}^{i=n} X_i \prec K'/\theta = \theta_1) = \beta \leq 0,10$, $\sum_{i=1}^{i=n} X_i$ suivant la loi de POISSON de paramètre $n.\theta$.

Les résultats obtenus ci-après en faisant varier n puis en déterminant le *seuil critique* K' correspondant à partir des tables de valeurs de la loi de POISSON ou mieux, à partir d'un calcul direct su EXCEL, montrent qu'en définitive, *l'optimum atteint* ($n^* = 23$) *est très proche de celui obtenu précédemment par approximation normale.*

n	20	21	22	23	24
K'	15	16	17	17	18
α	0,049	0,040	0,032	0,046	0,038
β	0,105	0,111	0,117	0,082	0,087

2.2 Test paramétrique pour le modèle de RAYLEIGH

La loi de RAYLEIGH est fréquemment rencontrée en théorie du signal pour décrire le bruit en sorties de certains récepteurs de transmission. Dans un cadre plus large qui est celui des processus stationnaires, cette loi modélise valablement la hauteur des vagues ou d'une crue (se référer pour cela à l'application 1.4 du chapitre II).

Enoncé : On considère dans ce problème, une variable aléatoire X de densité de probabilité $f_\lambda(x) = \frac{2x}{\lambda}.\exp(-\frac{x^2}{\lambda}).1_{[0,+\infty[}(x)$ où $\lambda \succ 0$ (modèle de RAYLEIGH).

1°) Déterminer l'estimateur du maximum de vraisemblance du paramètre λ, soit $\hat{\lambda}$.

2°) On considère le test $\begin{cases} H_0 : \lambda = \lambda_0 \\ H_1 : \lambda = \lambda_1 \end{cases}$ dans lequel $\lambda_1 \succ \lambda_0$. En utilisant le théorème de NEYMAN et PEARSON, montrer que le test optimal (U.P.P) pour ce problème, s'exprime en fonction de la statistique $\hat{\lambda}$. En expliciter la forme de la région critique.

3-a) Montrer que la variable $U = \dfrac{2.n.\hat{\lambda}}{\lambda}$ suit la loi du chi- deux à $2n$ degrés de liberté.

3-b) En déduire, la détermination de la région critique du test considéré.

3-c) On suppose $\alpha = 0,05$; $\lambda_0 = 4$; $\lambda_1 = 5$; et $n = 15$. Définir la région critique correspondante et évaluer la puissance du test qui en résulte.

4°) Supposant n grand et admettant la validité du théorème central limite, expliciter numériquement :

a) la région critique ;

b) la puissance du test ;

c) la taille minimale de l'échantillon à considérer, soit n^*, pour obtenir une puissance au moins égale à $1-\beta$;

d) Traiter les questions ci-dessus dans le cadre de l'application numérique $\alpha = 0,05; \beta = 0,10,\ n = 30; \lambda_0 = 4; \lambda_1 = 5$.

Solution : 1°) Se référant à l'application 1.4 du chapitre II, on admettra que *l'estimateur du maximum de vraisemblance* de λ, soit $\hat{\lambda}$ qui est la solution de l'équation $\dfrac{\partial L}{\partial \lambda} = 0$, est déterminé par la statistique $\hat{\lambda} = \dfrac{\sum_{i=1}^{i=n} X_i^2}{n}$.

2°) Suivant le théorème de NEYMAN et PEARSON, la *région critique* qui porte sur le rapport des vraisemblances est définie par $W = \left\{ (x_1, x_2, ..., x_n) / \dfrac{L(x_1, x_2, ..., x_n, \lambda_0)}{L(x_1, x_2, ..., x_n, \lambda_1)} \le k \right\}$, soit en utilisant les log – vraisemblances, la condition $\ln \dfrac{L_0}{L_1} \le \ln k$.

Or, $\ln L(x_1, x_2, ..., x_n, \lambda) = \sum_{i=1}^{i=n} \ln x_i - \dfrac{1}{\lambda} \cdot \sum_{i=1}^{i=n} x_i^2 + n.\ln 2 - n.\ln \lambda$. Il s'ensuit, pour ce qui est de la région critique, la définition $(\dfrac{1}{\lambda_1} - \dfrac{1}{\lambda_0}) . \sum_{i=1}^{i=n} x_i^2 - n.(\ln \lambda_0 - \ln \lambda_1) \le \ln k$, soit le résultat $\sum_{i=1}^{i=n} x_i^2 \ge \dfrac{1}{1/\lambda_0 - 1/\lambda_1} . \left[-\ln k + n.(\ln \lambda_1 - \ln \lambda_0 \right]$, puisque $\lambda_1 \succ \lambda_0$.

• En conclusion, la *région critique du test* a pour forme $\hat{\lambda} \ge K$ (puisque $\sum_{i=1}^{i=n} x_i^2 = n.\hat{\lambda}$). Le théorème de NEYMAN et PEARSON assure par ailleurs qu'il s'agit bien d'un **test de puissance maximum** (test U.M.P).

3-a) En premier lieu, on constate immédiatement que la variable aléatoire X^2 suit la *loi exponentielle* de paramètre $1/\lambda$. En effet, le changement de variable $y = x^2$ dans la densité de probabilité $f_\lambda(x)$ de X, implique en notant par $g(y)$ la densité de probabilité de Y, la relation $g(y).dy = \dfrac{2.\sqrt{y}}{\lambda} . \exp(-\dfrac{y}{\lambda}) . \dfrac{dy}{2.\sqrt{y}} = \dfrac{1}{\lambda} . \exp(-\dfrac{y}{\lambda}) . dy$.

Ainsi $g(y)=\frac{1}{\lambda}.\exp(-\frac{y}{\lambda}), y\in R^+$, ce qui est bien le résultat annoncé.

- Dans ces conditions, la *somme* $\sum_{i=1}^{i=n} X_i^2$ *des n variables aléatoires exponentielles indépendantes* suit la *loi Gamma n* de densité de probabilité $\frac{(1/\lambda)^n.e^{-x/\lambda}.x^{n-1}}{(n-1)!}$ (cf. rappels de cours du chapitre I, paragraphe 1.1).

De façon équivalente, $\frac{1}{\lambda}.\sum_{i=1}^{i=n} X_i^2$ suit la *loi Gamma n* de paramètre 1, à savoir de densité de probabilité $\frac{e^{-z}.z^{n-1}}{(n-1)!}$ (il suffit de procéder au changement de variable $Z=X/\lambda$ dans la densité de probabilité susmentionnée). Enfin, un nouveau changement de variable $U=2Z$ aboutit à la densité de probabilité $\frac{e^{-u}.u^{n-1}}{2^n.(n-1)!}$ qui correspond en définitive à la densité de probabilité de la loi $\chi^2(2n)$.

En effet, pour cette dernière, ou plutôt pour la loi $\chi^2(n)$, il est rappelé que la densité de probabilité en est définie par $\varphi(x)=\frac{x^{n/2-1}.e^{-x/2}}{2^{n/2}.\Gamma(n/2)}$ (cf. rappels de cours du chapitre I, paragraphe 1.1). On a donc bien établi que $\frac{2}{\lambda}.\sum_{i=1}^{i=n} X_i^2 = 2n.\frac{\hat{\lambda}}{\lambda}$ suit la loi $\chi^2(2n)$.

3-b) $\alpha=\Pr ob(décider\ \ H_1/H_0\ \ vraie)$, soit $\alpha=\Pr ob(\hat{\lambda}\geq K/\lambda=\lambda_0)$. Il en résulte en utilisant le résultat de la question précédente $\alpha=\Pr ob(2n.\frac{\hat{\lambda}}{\lambda}\geq 2n.\frac{K}{\lambda}/\lambda=\lambda_0)$ où $2n.\frac{\hat{\lambda}}{\lambda}$ suit la loi $\chi^2(2n)$. Déterminant à l'aide de la table de valeurs annexée, le seuil χ_α^2 vérifiant $\Pr ob(\chi^2(2n)\geq\chi_\alpha^2)=\alpha$, on en déduit immédiatement $K=\frac{\lambda_0.\chi_\alpha^2}{2n}$.

En résumé, si $\hat{\lambda}\geq\frac{\lambda_0.\chi_\alpha^2}{2n}$ on décide H_1, H_0 étant retenu sinon.

3-c) Pour $n=15$, soit $2n=30$, la table annexée fournit le seuil critique $\chi_\alpha^2=43,773$, soit $K=5,83$. Quant à *l'erreur de 2ème espèce* β, elle est définie par la probabilité $\Pr ob(décider\ \ H_0/H_1\ \ vraie)=\Pr ob(\hat{\lambda}\prec K/\lambda=\lambda_1)$, soit $\beta=\Pr ob(\chi^2(2n)\prec K.\frac{2n}{\lambda_1})$.

Numériquement, $\beta=\Pr ob(\chi^2(30)\prec 5,83\times\frac{30}{5}=34,98)$. Un calculateur serait nécessaire pour avoir la vraie valeur de β, mais le constat, par lecture dans la table annexée, de la valeur $\Pr ob(\chi^2(30)\geq 36,25)=0,2$ permet de conclure d'ores et déjà à la majoration $\beta\prec 1-0,2=80\%$ et, pour la puissance η du test, à la minoration $\eta\succ 20\%$.

• La valeur de $2n$ est comprise ici dans la *zone de validité de la convergence de la variable aléatoire* $\sqrt{2.\chi^2}-\sqrt{2.\nu-1}$ *vers la loi normale* $N(0,1)$. Dans ces conditions, on peut avoir un résultat plus précis pour β, à partir de cette convergence, c'est-à-dire en écrivant $\beta=\Pr ob(\chi^2 \prec 34,98)=\Pr ob(\sqrt{2.\chi^2}-\sqrt{2\times 30-1}=\sqrt{69,96}-\sqrt{59}=0,67)$.

Par lecture dans la table de valeurs de la loi normale, centrée, réduite (cf. annexes), soit $N(0,1)$, il vient $\beta=0,749$, ou encore une puissance de 25% (après arrondi).

• *Autre approximation possible*, celle de la *loi de chi- deux*, $\chi^2(n)$, vers la *loi normale* $\xi: N(m=n,\sigma^2=2n)$ (cf. rappels de cours du chapitre I, paragraphe 1.2), ce qui conduit en l'occurrence à $\beta=\Pr ob(\chi^2(30)\prec 34,98)=\Pr ob(\xi \prec \frac{34,98-30}{\sqrt{60}}=0,64)$=0,74, soit une puissance de 26% qui ici encore, est très proche des estimations précédentes.

4-a) Le *théorème central limite* appliqué à la somme $\sum_{i=1}^{i=n} X_i^2$ entraîne sa convergence vers la *loi normale* de moyenne $\sum_{i=1}^{i=n} E(X_i^2)$ et de variance $\sum_{i=1}^{i=n} Var(X_i^2)$. Or, comme cela a été établi dans l'application 1.4 du chapitre II, $E(X^2)=\int_0^{+\infty}\frac{2.x^3}{\lambda}.\exp(-\frac{x^2}{\lambda}).dx=\lambda$.

D'autre part, $Var(X^2)=E(X^4)-\left[E(X^2)\right]^2$ avec $E(X^4)=\int_0^{+\infty}\frac{2.x^5}{\lambda}.\exp(-\frac{x^2}{\lambda}).dx$.

Une intégration par parties suivant $U=x^4, dV=\frac{2x}{\lambda}.\exp(-\frac{x^2}{\lambda})\Rightarrow dU=4x^3.dx$ et $V=-\exp(-\frac{x^2}{\lambda})$ conduit à $E(X^4)=-\left[4.x^3.\exp(-\frac{x^2}{\lambda})\right]_0^{+\infty}+4.\int_0^{+\infty}x^3.\exp(-\frac{x^2}{\lambda}).dx$.

Le premier de ces deux termes étant nul, on a finalement, par rapprochement du second terme avec l'intégrale qui définit $E(X^2)$, le résultat $E(X^4)=2.\lambda^2$.

En résumé, $Var(X^2)=E(X^4)-\left[E(X^2)\right]^2=2.\lambda^2-\lambda^2=\lambda^2$. Ainsi $E(\sum_{i=1}^{i=n}X_i^2)=n.\lambda$ et $Var(\sum_{i=1}^{i=n}X_i^2)=n.\lambda^2$. Il faut donc retenir que la somme $\sum_{i=1}^{i=n}X_i^2$ converge vers la loi normale $N(m=n.\lambda,\sigma=\lambda.\sqrt{n})$, ce qui équivaut à la convergence de $\hat{\lambda}$ vers la *loi normale* $N(\lambda,\sigma=\frac{\lambda}{\sqrt{n}})$.

• Pour ce qui est de la **région critique** $\hat{\lambda}\geq K$, la donnée de l'erreur de première espèce α, conduit donc à la relation $\alpha=\Pr ob(\hat{\lambda}\geq K/\lambda=\lambda_0)$, soit par passage à la variable normale centrée réduite ξ de loi $N(0,1)$, $\alpha=\Pr ob(\xi\geq\frac{K-\lambda_0}{\lambda_0/\sqrt{n}})$. Désignant par t_α le nombre vérifiant $\alpha=\Pr ob(\xi\geq t_\alpha)$, on a donc $K=\lambda_0+t_\alpha.\frac{\lambda_0}{\sqrt{n}}$.

Relativement aux données numériques antérieures $\alpha = 0,05; \lambda_0 = 4; \lambda_1 = 5, n = 15$, le fait théoriquement contestable d'admettre cette convergence conduirait ici au *seuil critique* $K = 4 + 1,645 \times \frac{4}{\sqrt{15}} = 5,71$, ce qui n'est pas très éloigné, néanmoins de la valeur exacte trouvée précédemment, à savoir, 5,83. On peut donc compter sur le fait d'une *bonne précision* quand, à fortiori, $n = 30$.

4-b) $\beta = \Pr ob(\hat{\lambda} \prec K / \lambda = \lambda_1) \Rightarrow \beta = \Pr ob(\xi \prec \frac{K - \lambda_1}{\lambda_1 / \sqrt{n}})$. Avec les données antérieures, on trouve donc $\beta = \Pr ob(\xi \prec \frac{5,71 - 5}{5/\sqrt{15}} = 0,55)$, soit $\beta = 0,71$ (au lieu des 74% précédents).

4-c) On cherche n^* tel que l'on ait, simultanément $\begin{cases} \alpha = \Pr ob(\hat{\lambda} \geq K / \lambda = \lambda_0) \\ \beta \geq \Pr ob(\hat{\lambda} \prec K / \lambda = \lambda_1) \end{cases}$.

Partant d'une égalité, pour ce qui est de l'erreur de 2ème espèce β, et recourant à la variable normale centrée réduite $\xi : N(0,1)$, les équations susmentionnées s'écrivent :

$$\begin{cases} \Pr ob(\xi \geq \frac{K - \lambda_0}{\lambda_0 / \sqrt{n}}) = \alpha \\ \Pr ob(\xi \prec \frac{K - \lambda_1}{\lambda_1 / \sqrt{n}}) = \beta \end{cases}$$

Notant par t_α et t_β les nombres qui vérifient respectivement les conditions $\Pr ob(\xi \geq t_\alpha) = \alpha$ et $\Pr ob(\xi \prec t_\beta) = \beta$, on a donc $\begin{cases} \frac{K - \lambda_0}{\lambda_0 / \sqrt{n}} = t_\alpha \\ \frac{K - \lambda_1}{\lambda_1 / \sqrt{n}} = t_\beta \end{cases}$.

Par quotient, et après développement, il en résulte $K = \frac{\lambda_0 . \lambda_1 . (t_\beta - t_\alpha)}{t_\beta . \lambda_1 - t_\alpha . \lambda_0}$, puis par substitution, la valeur de n^* qui est aussi la valeur minimale cherchée lorsque pour β, on passe de l'égalité à une inégalité.

4-d) Les données $\alpha = 0,05; \beta = 0,10; \lambda_0 = 4; \lambda_1 = 5; n = 30$ conduisent pour chacune des questions précédentes à $K = 4 + 1,645 \times \frac{4}{\sqrt{30}} = 5,20$ et $\beta = \Pr ob(\xi \prec \frac{5,2 - 5,0}{5/\sqrt{30}}) = 0,59$, c'est-à-dire une puissance de 41%.

Enfin, si on souhaite, pour le test considéré, une puissance d'au moins 90% (ou encore une erreur de 2ème espèce β inférieure à 10%), on a, avec les notations antérieures, $t_\beta / \Pr ob(\xi \prec t_\beta) = 0,10 \Rightarrow t_\beta = -1,28$ (par lecture dans la table de valeurs annexée de la fonction de répartition $\Pi(t) = \Pr ob(\xi \leq t)$).

Ainsi $K = 4,507$ ce qui entraîne, par exemple, à partir de la relation $\dfrac{K-\lambda_0}{\lambda_0/\sqrt{n}} = t_\alpha$, la valeur cherchée $n^* = \left[\dfrac{\lambda_0 . t_\alpha}{K-\lambda_0}\right]^2$, soit numériquement, $n^* = 168,43$.

Par arrondi et par excès, il faut donc disposer d'une **échantillon de taille 169** pour obtenir ici un *test de puissance supérieure* à 90%, l'erreur de première espèce étant quant à elle égale à 5%.

2.3 Tests portant sur un modèle de revenus (PARETO)

Enoncé : On se propose dans cette application, de construire un test non paramétrique d'ajustement puis un test paramétrique de conformité autour de la loi de PARETO, loi dont la densité de probabilité est $f(x,\theta) = \dfrac{\theta}{x^{\theta+1}} . 1_{[1,+\infty[}(x)$ et dont il est pressenti qu'elle modélise le revenu annuel X d'un individu (exprimé en 10^4 euros).

PARTIE I

On a noté les revenus annuels de 5 personnes prises au hasard dans la population de référence, les données ainsi recueillies étant :

1,1	3,2	2,3	3,8	2,4

1°) Déterminer la fonction de répartition de la variable aléatoire X.

2°) Utilisant un test de KOLMOGOROV, peut-on admettre au niveau de signification de 5% ($\alpha = 0,05$), que ces données sont issues d'une loi de PARETO de paramètre $\theta = 1$?

PARTIE II

A partir d'un échantillon de taille n de variables aléatoires indépendantes de loi parente X de type PARETO susmentionné, soit $(X_1, X_2, ..., X_n)$, on se propose de tester $\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta = \theta_1 \end{cases}$ où θ_0 et θ_1 sont deux valeurs fixées vérifiant $\theta_0 \prec \theta_1$.

1°) A l'aide du théorème de NEYMAN et PEARSON, montrer que la statistique de décision du test est $Y = \sum_{i=1}^{i=n} \ln X_i$.

2°) Explicitant la loi suivie par Y, en déduire la règle de décision du test au niveau de signification α donné. Exprimer la puissance du test (application numérique : $n = 5; \theta_0 = 1 ; \theta_1 = 1,5; \alpha = 0,01$).

3°) On suppose ici que $n = 100$. Utilisant le théorème central limite, qu'obtient-on pour ce qui est des résultats précédents (seuil et puissance) ?

Solution : I-1°) $F_X(x) = \int_{-\infty}^{x} f(t,\theta).dt$, soit $F_Y(x) = 0$ si $x \leq 1$ et $F_X(x) = \int_1^x \dfrac{\theta.dt}{t^{\theta+1}}$ si $x \succ 1$, soit après intégration, $F_X(x) = 1 - x^{-\theta}$ $(x \succ 1)$.

I-2°) S'agissant d'une *loi continue* et utilisant l'expression précédente de la fonction de répartition, le **test d'ajustement** le plus approprié ici est le test de KOLMOGOROV qu'il est donc proposé de mettre en œuvre.

Le *classement par ordre croissant* des cinq valeurs observées, et la *comparaison de la fréquence cumulée empirique* $\widehat{F}(x_i) = 1/n$ *avec la fréquence théorique* $F_0(x_i) = 1 - x_i^{-\theta}$ (dans laquelle on suppose $\theta = 1$), conduit au tableau de calculs ci-dessous (cf. méthode décrite en rappels de cours du présent chapitre, paragraphe 3.1.b).

x_i	$\widehat{F}(x_i) = 1/n$	$F_0(x_i)$	$\lvert F_0(x_i) - 1/n \rvert$	$\lvert F_0(x_i) - 1/n - 1 \rvert$
1,1	0,2	0,0909	0,1091	0,0909
2,3	0,4	0,5652	0,1652	0,3652
2,4	0,6	0,5833	0,0166	0,1833
3,2	0,8	0,6875	0,1125	0,0875
3,8	1,0	0,7368	0,2632	0,0632

Ainsi pour l'échantillon en question, la **distance de KOLMOGOROV**, soit la statistique, $D(F_0, \widehat{F}) = Sup_{i=1,2,...,n}\left\{\left|F_0(X(i)) - \frac{i}{n}\right|, \left|F_0(X(i)) - \frac{i-1}{n}\right|\right\}$ prend-elle numériquement la valeur 0,3652.

• Par ailleurs, *la table de* KOLMOGOROV *pour un échantillon* (cf. annexes), fournit, pour $\alpha = 0,05; n = 5$; et en l'occurrence le *test bilatéral*, la valeur D_α telle que $\Pr ob(D(F_0, \widehat{F}) \geq D_\alpha) = \alpha$. Il vient ainsi $D_\alpha = 0,565$.

La **région critique** du test étant caractérisée par $D_{calculé} \geq D_\alpha$ et les résultats numériques obtenus précédemment entraînant $D_{calculé} = 0,3652 \prec D_\alpha = 0,565$, on conclut *qu'il n'y a pas lieu de rejeter l'hypothèse* H_0 *suivant laquelle les données proposées sont la réalisation d'une loi de* PARETO de paramètre $\theta = 1$.

II-1°) La *vraisemblance* $L(x_1, x_2, ..., x_n, \theta)$ associée à un échantillon de taille n, soit $(x_1, x_2, ..., x_n)$, est égale, pour la loi proposée, à $L(x_1, x_2, ..., x_n, \theta) = \frac{\theta^n}{\prod_{i=1}^{i=n} x_i^{\theta+1}} . \prod_{i=1}^{i=n} 1_{[1,+\infty[}(x_i)$.

La *région critique*, selon le théorème de NEYMAN et PEARSON, du **test paramétrique** proposé, est caractérisée suivant le *rapport des vraisemblances*, par la relation $\frac{L(x_1, x_2, ..., x_n, \theta_0)}{L(x_1, x_2, ..., x_n, \theta_1)} \leq k$.

En utilisant la *log- vraisemblance* $\ln L = n.\ln\theta - (\theta + 1).\sum_{i=1}^{i=n} \ln x_i$, avec $x_i \geq 1, (1 \leq i \leq n)$, la *région critique* est donc définie par $n.\ln\frac{\theta_0}{\theta_1} + (\theta_1 - \theta_0).\sum_{i=1}^{i=n} \ln x_i \leq \ln k$, c'est-à-dire une relation qui est de la forme $\sum_{i=1}^{i=n} \ln x_i \leq K$, avec $K = \frac{1}{\theta_1 - \theta_0}.\left[\ln k + n.\ln\frac{\theta_1}{\theta_0}\right]$.

On constate bien ainsi que la *statistique du test* est $Y = \sum_{i=1}^{i=n} \ln X_i$.

II-2°) Le *changement de variable* $Y_i = \ln X_i$ dans la *densité de probabilité élémentaire* $f(x_i).dx_i$ de la loi de PARETO conduit pour Y_i à des valeurs sur $[0,+\infty[$ et à la densité $g(y_i)$ vérifiant élémentairement $g(y_i).dy_i = f(x_i).dx_i$, avec $x_i = e^{y_i}$, soit $g(y_i).dy_i = \frac{\theta}{e^{y_i(\theta+1)}}.e^{y_i}.dy_i$.

Ainsi, les variables aléatoires Y_i dont la densité de probabilité s'écrit $g(y_i) = \theta.\exp(-\theta.y_i)$, suivent-elles la **loi exponentielle** de paramètre θ. Dans ces conditions et conformément aux propriétés rappelées au chapitre I (cf. rappels de cours, paragraphe 1.1), la *somme* $Y = \sum_{i=1}^{i=n} Y_i = \sum_{i=1}^{i=n} \ln X_i$ suit-elle la **loi Gamma- n** de paramètre θ, c'est-à-dire de densité de probabilité $\frac{\theta^n x^{n-1}.e^{-\theta.x}}{(n-1)!}$.

• Comme cela a déjà été mis en évidence dans l'application 2.2 précédente (modèle de RAYLEIGH), *la variable* $U = 2.\theta.Y$ dont, par nouveau changement de variable, la densité de probabilité est $\frac{e^{-\frac{u}{2}}.u^{n-1}}{2^n.(n-1)!}$ suit la **loi de chi- deux** à $2n$ degrés de liberté, soit $\chi^2(2n)$, résultat qui permet ainsi de déterminer aisément la règle de décision du test et sa puissance.

En effet, $\alpha = \Pr ob(Y = \sum_{i=1}^{i=n} \ln X_i \leq K / \theta = \theta_0) = \Pr ob(2.\theta.Y = \chi^2(2n) \prec 2.\theta.K / \theta = \theta_0)$.

Déterminant dans la table de valeurs annexée (cf. loi de χ^2), le seuil χ^2_α vérifiant $\Pr ob(\chi^2(2n) \prec \chi^2_\alpha) = \alpha$, il vient immédiatement $K = \frac{\chi^2_\alpha}{2.\theta_0}$.

On décidera donc H_1 si $\sum_{i=1}^{i=n} \ln x_i \leq \frac{\chi^2_\alpha}{2.\theta_0}$ et H_0 dans le cas contraire.

D'autre part, $\beta = \Pr ob(\sum_{i=1}^{i=n} \ln X_i \succ K / \theta = \theta_1)$, soit $\beta = \Pr ob(\chi^2(2n) \succ 2.\theta_1.\frac{\chi^2_\alpha}{2.\theta_0})$.

D'où par lecture dans la table des valeurs annexée de la loi de χ^2, la valeur de β et à fortiori de la *puissance* $\eta = 1 - \beta$ du test.

• L'application numérique $n = 5; \alpha = 0,01; \theta_0 = 1; \theta_1 = 1,5$ conduit pour la loi de χ^2 à $\nu = 2n$ degrés de libertés, soit $\nu = 10$, à $\chi^2_\alpha = 2,558$ (cf. tables de valeurs annexées). Il en résulte $K = 1,279$.

L'erreur de 2ème espèce β associée au test $\begin{cases} H_0 : \theta = \theta_0 = 1,0 \\ H_1 : \theta = \theta_1 = 1,5 \end{cases}$ est, quant à elle, égale à $\Pr ob(\chi^2(10) \succ 2 \times 1,5 \times 1,279 = 3,84)$. A défaut de disposer d'un calculateur, la table de valeurs annexée montre que $\Pr ob(\chi^2(10) \geq 0,975) = 3,25$ et $\Pr ob(\chi^2(10) \geq 0,95) = 3,94$. β est donc comprise entre 95% et 97,5%, la *puissance du test* étant pour sa part comprise entre 2,5% et 5%.

II-3°) Si $n=100$, on peut admettre en fonction du **théorème central limite** que la somme $\sum_{i=1}^{i=n} \ln X_i$ converge vers la *loi normale* de moyenne $\frac{n}{\theta}$ et de variance $\frac{n}{\theta^2}$, les variables aléatoires X_i étant, pour rappel, des variables exponentielles de paramètre θ et à fortiori d'espérance et de variance respectivement égales à $\frac{1}{\theta}$ et $\frac{1}{\theta^2}$.

La *région critique* caractérisée par $Y=\sum_{i=1}^{i=n} \ln X_i \leq K$ est donc déterminée, par passage à la *variable normale centrée réduite* ξ : $N(0,1)$, par $\alpha = \Pr ob(\xi \leq \frac{K - n/\theta_0}{\sqrt{n}/\theta_0})$, d'où en notant par t_α le nombre vérifiant $\alpha = \Pr ob(\xi \leq t_\alpha)$, le *seuil critique* $K = \frac{n}{\theta_0} + t_\alpha . \frac{\sqrt{n}}{\theta_0}$.

Numériquement, si $\alpha = 0,01$ et $n=100$, on a immédiatement $t_\alpha = -2,33$ et $K = 76,7$. Il s'ensuit $\beta = \Pr ob(\xi \geq \frac{K - n/\theta_1}{\sqrt{n}/\theta_1} = 1,505)$, soit environ $\beta = 0,066$.

La *puissance correspondante du test*, soit $\eta = 1-\beta$, avoisine donc 93,4%.

2.4 Test entre deux lois pour une étude de clientèle

Comme le montre la présente application, la méthode de NEYMAN et PEARSON peut être ntilisée an-delà du cadre de tests portant sur la valeur d'un paramètre.

Enoncé : Une agence de voyage souhaite cibler sa clientèle. Elle sait que les coordonnées du lieu d'habitation d'un client (X,Y) rapportées au lieu de naissance $(0,0)$ sont une information significative pour le goût de ce client. Elle distingue :

- la population 1 (hypothèse H_0) dont la loi de probabilité a pour densité :

$$f_1(x,y) = \frac{1}{2\pi} . \exp\left[-(\frac{x^2+y^2}{2}) \right] .dx.dy$$

- la population 2 (hypothèse H_1) dont la loi de probabilité a pour densité :

$$f_2(x,y) = \frac{1}{16} . 1_{[-2,+2]}(x) . 1_{[-2,+2]}(y) .dx.dy$$

L'agence souhaite tester l'hypothèse qu'un nouveau client vivant en (x,y) appartienne à la population 1 plutôt qu'à la population 2.

1°) Montrer que les variables aléatoires X et Y sont indépendantes quelle que soit l'hypothèse choisie.

2°) Exprimant le rapport des fonctions de vraisemblance sous les hypothèses H_0 et H_1, proposer à l'aide du théorème de NEYMAN et PEARSON un test de puissance maximale et de première espèce α dont on caractérisera la région critique W.

3°) A partir de la statistique appropriée, caractériser graphiquement la région critique de ce test dans R^2 (on supposera $\alpha = 5\%$).

4°) Pour un client dont la valeur du couple (X,Y) est $(1,9-1,9)$, qu'en conclure ?

5°) Relativement au test ci-dessus, calculer la valeur du risque de $2^{ème}$ espèce β et la puissance du test.

Solntion : 1°) Sous l'hypothèse H_0 et considérant la densité de probabilité $\varphi_1(x)$ de la variable aléatoire X, il est immédiat d'après la *loi des probabilités marginales*, que $\varphi_1(x) = \frac{1}{2\pi}.\exp(-\frac{x^2}{2}).\int_{-\infty}^{+\infty}\exp(-\frac{y^2}{2}).dy = \frac{1}{\sqrt{2\pi}}.\exp(-\frac{x^2}{2})$. De même, par symétrie, Y admet sur R, la densité de probabilité $\psi_1(y) = \frac{1}{\sqrt{2\pi}}.\exp(-\frac{y^2}{2})$, ce qui établit l'indépendance puisque $f_1(x,y) = \varphi_1(x).\psi_1(y)$.

De même, sous H_1, on a pour X, $\varphi_2(x) = \frac{1}{16}.1_{[-2,+2]}(x).\int_{-2}^{+2}dy = \frac{1}{4}.1_{[-2,+2]}(x)$. Par symétrie, on a aussi, pour Y, $\psi_2(y) = \frac{1}{4}.1_{[-2,+2]}(y)$ et à fortiori $f_2(x,y) = \varphi_2(x).\psi_2(y)$. ce qui établit l'indépendance.

2°) Selon le **théorème de NEYMAN et PEARSON**, le test qui, portant sur *le rapport de vraisemblance*, est défini par $W = \left\{(x_1,x_2,...,x_n)/\frac{L_0}{L_1} \le k\right\}$ est de *puissance maximale*. L_0 et L_1 désignant les fonctions de vraisemblance associées à l'échantillon en question, et ceci respectivement suivant les hypothèses H_0 et H_1.

En l'occurrence, dans le cadre d'un échantillon (x,y) de taille $n=2$, l'expression de la région critique susmentionnée s'écrit $W = \left\{(x,y)/\frac{f_1(x,y)}{f_2(x,y)} \le k\right\}$, la relation $\frac{f_1x,y)}{f_2(x,y)} \le k$, s'écrivant concrètement, $\frac{16}{2\pi}.\frac{\exp(-\frac{(x^2+y^2)}{2})}{1_{[-2,+2]}(x).1_{[-2,+2]}(y)} \le k$.

Lorsque $x \notin [-2,+2]$, $1_{[-2,+2]}(x).1_{[-2,+2]}(y) = 0$. Le rapport de vraisemblance $\frac{L_0}{L_1}$ est *infini*, ce qui conduit à retenir l'hypothèse H_0 puisque la vraisemblance associée à l'hypothèse H_1 est nulle, sinon lorsque $x \in [-2,+2]$, on a $1_{[-2,+2]}(x).1_{[-2,+2]}(y) = 1$, ce qui entraîne pour la région critique, la caractérisation $\frac{16}{2\pi}.\exp(-\frac{(x^2+y^2)}{2}) \le k$, ou encore, $x^2+y^2 \ge -2.\ln(\frac{2\pi.k}{16})$, c'est-à-dire une relation de la forme $x^2+y^2 \ge K$.

- En conclusion, $W = \left\{(x,y) \in R^2 \cap [-2,+2]\times[-2,+2]/x^2+y^2 \ge K\right\}$

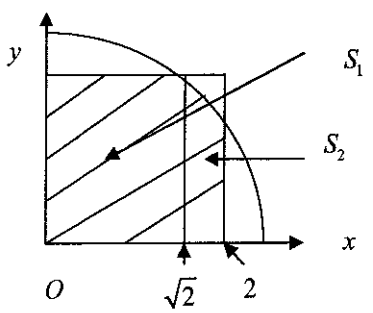
3°) Lorsqu'on est sous l'hypothèse H_0, les variables X et Y suivent *la loi normale, centrée, réduite, et sont indépendantes*. La somme X^2+Y^2 suit donc une **loi du chi- deux** à deux degrés de liberté (cf. rappels de cours du chapitre I, paragraphe 1.1).

L'erreur de première espèce α étant égale à 5%, on peut écrire $\Pr ob(W/H_0)=0,05 \Rightarrow \Pr ob(X^2+Y^2 \geq K)=0,05$. Désignant par K_α le nombre qui, pour la loi de chi- deux à deux degrés de liberté, soit $\chi^2(2)$, vérifie $\Pr ob(\chi^2(2) \geq K)=0,05$, il vient immédiatement, par lecture dans la table des valeurs annexée, $K_\alpha = 5,99$.

Ainsi, la *région critique* (zone de décision de H_1) est-elle définie finalement par l'ensemble $W=\left\{(x,y) \in [-2,+2]\times[-2,+2] / x^2+y^2 \geq 5,99\right\}$. Géométriquement, il s'agit de l'intersection entre l'extérieur du cercle de rayon 2,45 et l'intérieur du carré de côté 4 et de centre O.

4°) S'agissant du point de coordonnées $(x=1,9; y=1,9)$, il appartient au carré et est extérieur au cercle puisque $x^2+y^2=7,22 \succ 5,99$. En conclusion, pour le client en question, on décide H_1 pour ce qui est de sa population de rattachement.

5°) *L'erreur de 2ème espèce* β est égale à Prob(décider H_0/H_1 vraie), soit en explicitant, $\beta = \Pr ob(X^2+Y^2 \prec K / H_1 \ \textit{vraie})$ avec $K=5,99$. Lorsque H_1 est vraie, (X,Y) a pour densité de probabilité $\frac{1}{16}.1_{[-2,+2]}(x).1_{[-2,+2]}(y)$.



Ainsi $\beta = \iint\limits_{\substack{x^2+y^2 \prec 5,99 \\ -2 \leq x \leq 2 \\ -2 \leq y \leq 2}} \frac{dx.dy}{16}$, soit $\beta = 4.\frac{S}{16}$, avec $S = \iint\limits_{\substack{0 \leq x \leq 2 \\ 0 \leq y \leq 2 \\ x^2+y^2 \prec 5,99}} dx.dy$, c'est-à-dire *l'aire hachurée ci-contre.* Décomposant cette aire en *aires partielles* S_1 et S_2 définies ci-contre, on a $S_1 = 2 \times \sqrt{2} = 2,82$ et

$S_2 = (\int_{\sqrt{2}}^{2}(\int_{0}^{\sqrt{5,99-x^2}} dy).dx = \int_{\sqrt{2}}^{2}\sqrt{5,99-x^2}.dx$. Posant $x=\sqrt{5,99}.\cos t \Rightarrow t = \arccos(\frac{x}{\sqrt{5,99}})$

il s'ensuit, pour nouvelles bornes de l'intégrale précédente, $0,956 = ar\cos(\frac{\sqrt{2}}{\sqrt{5,99}})$ et

$0,614 = ar\cos(\frac{2}{\sqrt{5,99}})$, d'où $S_2 = -\int_{0,956}^{0,614} 5,99.\sin^2 t.dt = -\frac{5,99}{2}.\int_{0,956}^{0,614}(1-\cos 2t).dt$, soit

$S_2 = \frac{5,99}{2}.\left[t-\frac{\sin 2t}{2}\right]_{0,614}^{0,956} = 1,026$.

• En conclusion, $S = S_1 + S_2 = 3,846$ et $\beta = \frac{S}{4} = 0,962$. On remarque que la puissance du test qui est égale à $\eta = 1 - \beta = 3,8\%$ est particulièrement faible dans cet exemple, et ceci bien qu'il s'agit d'un test de puissance maximale. La petite valeur de la taille de l'échantillon $(n = 2)$ est une explication à la faiblesse de cette puissance.

2.5 Test séquentiel de WALD et contrôle de réception

L'un des aspects les plus courants du contrôle de qualité est celui de la réception opérée par le client, de la fabrication du fournisseur, le contrôle étant alors qualitatif (par attributs) encore que dans l'application 1.3 précédente, c'est un contrôle par mesures qui est effectué. Au-delà du plan d'échantillonnage simple, les plans d'échantillonnage multiples trouvent à travers la méthode progressive (séquentielle), une généralisation qui est la plus économique.

Enoncé : Il est proposé de développer quelques aspects du plan de contrôle par attributs qu'on limitera ici au suivi de la proportion p de pièces défectueuses $(0 \leq p \leq 1)$, chaque élément contrôlé étant ainsi classé conforme ou non conforme (un autre mode est celui du contrôle du nombre moyen de défauts par unité testée).

De façon générale, le client souhaite limiter à β le risque d'accepter une fabrication très mauvaise ($p \geq p_2$) et le fournisseur souhaite limiter à α le risque de se voir refuser une fabrication qui serait de bonne qualité ($p \leq p_1$, avec $p_1 \prec p_2$).

PARTIE I (plan simple)

Dans ce mode d'échantillonnage, on procède à un prélèvement de n pièces à l'issue duquel on décide du rejet ou non de la livraison en fonction du nombre de pièces défectueuses constatées.

I-1°) α et β étant fixées, écrire les équations permettant d'exprimer la taille de l'échantillon à prélever, soit n, et la règle de décision correspondante.

I-2°) On fixe α et β à respectivement 1% et 5%. Par ailleurs, on considère que $p_1 = 0,10$ et $p_2 = 0,30$. Expliciter la mise en œuvre du contrôle.

PARTIE II (plan progressif)

Le recours aux plans multiples, pallie au coût manifeste des plans simples, suivant lequel, n étant fixé, il faut un contrôle qui reste tout aussi important, que les premières pièces prélevées soient conformes, ou que les premiers prélèvements créent un doute manifeste.

Ainsi, l'échantillonnage double vise-t-il à créer une étape intermédiaire :

- on prélève tout d'abord n_1 pièces. Si le taux de non-conformité est inférieur à c_1, on accepte le lot. Si ce taux est supérieur à c_2, on refuse le lot ;
- dans la zone intermédiaire d'un taux de non-conformité compris entre c_1 et c_2, on procède alors au prélèvement d'un nouvel échantillon de taille n_2.

Imaginable à « triple détente » voire au-delà, ce dispositif est généralisé lorsque la taille de l'échantillon est aléatoire, ce qui est l'objet du test séquentiel de WALD.

II-1°) Expliciter à partir du rapport des vraisemblances, la règle de décision du test séquentiel de WALD dont on montrera qu'elle conduit, relativement au nombre cumulé de pièces non conformes, soit $X(n)$, et lorsque la taille aléatoire N de l'échantillon prélevé prend la valeur n, au mode opératoire :

- $X(n) \prec A(n) = -h_1 + s.n \Rightarrow$ on accepte la livraison ;
- $X(n) \succ B(n) = h_2 + s.n \Rightarrow$ on refuse la livraison ;
- $A(n) \leq X(n) \leq B(n) \Rightarrow$ on procède à un prélèvement supplémentaire.

(s, h_1, h_2) étant à déterminer en fonction de α, β, p_1, p_2.

II-2°) Déterminer les valeurs approchées de $E_{p_1}(N), E_{p_2}(N)$, et de $\underset{p}{Max}\, E_p(N)$.

II-3°) Expliciter numériquement les résultats précédents pour les valeurs numériques $\alpha = 1\%, \beta = 5\%, p_1 = 0{,}10, p_2 = 0{,}30$. Qu'en conclure ?

Solution : I-1°) Il s'agit de tester $\begin{cases} H_0 : p \leq p_1 \\ H_1 : p \geq p_2 \end{cases}$, **test composite** pour lequel les risques $\alpha(p)$ et $\beta(p)$ sont majorés par les risques α et β inhérents au test $\begin{cases} H_0 : p = p_1 \\ H_1 : p = p_2 \end{cases}$.

Raisonnant donc sur ces majorants, la donnée des *risques maximum* α et β conduit, X étant le nombre de pièces défectueuses au sein de l'échantillon de taille n, aux équations $\begin{cases} \alpha = \Pr ob(X \geq c / p = p_1) \\ \beta = \Pr ob(X \prec c / p = p_2) \end{cases}$.

X suivant la *loi binomiale* $B(n, p)$, il faut recourir à une *méthode exacte* (par approximations successives, par exemple), ou *aux approximations habituelles* par les lois de POISSON et GAUSS, ce dernier cas étant le plus courant et admis pour la suite.

I-2°) Introduisant la *variable normale centrée réduite* $\xi = \dfrac{X - n.p}{\sqrt{n.p.q}}$ et la *fonction de répartition* $\Pi(t) = \Pr ob(\xi \leq t)$, les équations précédentes s'écrivent :

$$\begin{cases} \alpha = \Pr ob(\xi \geq \dfrac{c - n.p_1}{\sqrt{n.p_1.(1-p_1)}}) \Rightarrow \Pi(\dfrac{c - n.p_1}{\sqrt{n.p_1.(1-p_1)}}) = 1 - \alpha \\ \beta = \Pr ob(\xi \prec \dfrac{c - n.p_2}{\sqrt{n.p_2.(1-p_2)}}) \Rightarrow \Pi(\dfrac{c - n.p_2}{\sqrt{n.p_2.(1-p_2)}}) = \beta \end{cases}$$

Notant par t_α le seuil vérifiant $\Pr ob(\xi \leq t_\alpha) = \alpha$, il vient $\dfrac{c - n.p_1}{\sqrt{n.p_1.(1-p_1)}} = t_{1-\alpha}$ et $\dfrac{c - n.p_2}{\sqrt{n.p_2.(1-p_2)}} = t_\beta$, d'où n et c par résolution du système d'équations en question.

Pour le cas présent, $\alpha = 1\%, \beta = 5\%, p_1 = 0{,}10, p_2 = 0{,}30 \Rightarrow t_{0,99} = 2{,}33$ et $t_{0,05} = -1{,}645$. Il s'ensuit, par quotient $\dfrac{c - n.p_1}{c - n.p_2} = \dfrac{t_{1-\alpha}}{t_\beta}.\dfrac{\sqrt{p_1.(1-p_1)}}{\sqrt{p_2.(1-p_2)}}$, soit numériquement $c = 0{,}196.n$.

Par *substitution* dans la première équation, par exemple, il vient $n = 52,76$, soit par arrondi, $n = 53$, et par suite $c = 10,35$, seuil qu'on arrondira à la valeur entière $c = 10$.

En conclusion, pour les risques α et β, maîtrisés respectivement à 1% et 5% ; et à partir d'échantillons de taille 53 et du nombre X de pièces défectueuses sur ledit échantillon, on décidera H_1 si $X \geq 10$ et H_0 sinon.

II-1°) Pour chacune des observations effectuées, la variable à considérer ici est la **variable de BERNOULLI** de loi :

- $X = 0 \Rightarrow$ pièce conforme $\Rightarrow$ probabilité associée $q = 1 - p$;
- $X = 1 \Rightarrow$ pièce non conforme $\Rightarrow$ probabilité associée p ;

loi qu'on peut caractériser par $\Pr ob(X = x) = p^x.(1-p)^{1-x}, x \in \{0,1\}$.

Explicitant le mode opératoire mentionné en rappels de cours (cf. paragraphe 2.7), on a successivement $R_n = \dfrac{\prod_{i=1}^{i=n} p_2^{x_i}.(1-p_2)^{1-x_i}}{\prod_{i=1}^{i=n} p_1^{x_i}.(1-p_1)^{1-x_i}} = \dfrac{p_2^{\sum_{i=1}^{i=n} x_i}.(1-p_2)^{n-\sum_{i=1}^{i=n} x_i}}{p_1^{\sum_{i=1}^{i=n} x_i}.(1-p_1)^{n-\sum_{i=1}^{i=n} x_i}}$, soit en passant au logarithme népérien et après développement et regroupement, l'expression :

$$\ln R_n = \left[\ln\frac{p_2}{p_1} + \ln\frac{1-p_1}{1-p_2}\right].\sum_{i=1}^{i=n} x_i - n.\ln\frac{1-p_1}{1-p_2}.$$

Il s'ensuit la **règle de décision** :

- **décider** H_1 si $\ln R_n \succ \ln\dfrac{1-\beta}{\alpha}$, ce qui entraîne la relation :

$$\sum_{i=1}^{i=n} x_i \succ \left[\frac{\ln\frac{1-p_1}{1-p_2}}{\ln\frac{p_2}{p_1} + \ln\frac{1-p_1}{1-p_2}}\right].n + \frac{\ln\frac{1-\beta}{\alpha}}{\ln\frac{p_2}{p_1} + \ln\frac{1-p_1}{1-p_2}} ;$$

- **décider** H_0 si $\ln R_n \prec \ln\dfrac{\beta}{1-\alpha}$, ce qui entraîne la relation :

$$\sum_{i=1}^{i=n} x_i \succ \left[\frac{\ln\frac{1-p_1}{1-p_2}}{\ln\frac{p_2}{p_1} + \ln\frac{1-p_1}{1-p_2}}\right].n - \frac{\ln\frac{1-\alpha}{\beta}}{\ln\frac{p_2}{p_1} + \ln\frac{1-p_1}{1-p_2}}$$

- la **mise en œuvre d'un prélèvement complémentaire** dans les autres cas.

• On est donc bien confronté aux bornes mentionnées dans l'énoncé avec :

$$h_1 = \frac{\ln\frac{1-\alpha}{\beta}}{\ln\frac{p_2}{p_1} + \ln\frac{1-p_1}{1-p_2}} ; h_2 = \frac{\ln\frac{1-\beta}{\alpha}}{\ln\frac{p_2}{p_1} + \ln\frac{1-p_1}{1-p_2}}, s = \frac{\ln\frac{1-p_1}{1-p_2}}{\ln\frac{p_2}{p_1} + \ln\frac{1-p_1}{1-p_2}}.$$

II-2°) Utilisant les résultats de l'application précédente 1.4, on est amené à calculer $E_p(Z)$

avec $Z=\dfrac{\ln f(X,p_2)}{\ln f(X,p_1)}$, $\dfrac{f(X,p_2)}{f(X,p_1)}=\dfrac{p_2^X.(1-p_2)^{1-X}}{p_1^X.(1-p_1)^{1-X}}\Rightarrow Z=\left[\ln\dfrac{p_2}{p_1}+\ln\dfrac{1-p_1}{1-p_2}\right].X+\ln\dfrac{1-p_2}{1-p_1}$

La variable de BERNOULLI ayant pour espérance mathématique la valeur p, on a ainsi pour les espérances $E_{p_1}(Z)$ et $E_{p_2}(Z)$ et par linéarité de l'espérance mathématique, les résultats :

$$E_{p_1}(Z)=\left[\ln\frac{p_2}{p_1}+\ln\frac{1-p_1}{1-p_2}\right].p_1+\ln\frac{1-p_2}{1-p_1},\ E_{p_1}(Z)=\left[\ln\frac{p_2}{p_1}+\ln\frac{1-p_1}{1-p_2}\right].p_2+\ln\frac{1-p_2}{1-p_1}.$$

De l'expression générale $E_\theta(N)=\dfrac{\left[\ln\dfrac{\beta}{1-\alpha}\right].P(\theta)+\left[\ln\dfrac{1-\beta}{\alpha}\right].(1-P(\theta))}{E_\theta(Z)}$ admise dans l'application antérieure 1.4, résulte pour $p=p_1$, par exemple :

$$E_{p_1}(N)=\frac{(1-\alpha).\ln\dfrac{\beta}{1-\alpha}+\alpha.\ln\dfrac{1-\beta}{\alpha}}{E_{p_1}(Z)}=\frac{-(1-\alpha).\ln\dfrac{1-\alpha}{\beta}+\alpha.\ln\dfrac{1-\beta}{\alpha}}{\left[\ln\dfrac{p_2}{p_1}+\ln\dfrac{1-p_1}{1-p_2}\right].\left[p_1-\dfrac{\ln\dfrac{1-p_1}{1-p_2}}{\ln\dfrac{p_2}{p_1}+\ln\dfrac{1-p_1}{1-p_2}}\right]}.$$

Ainsi $E_{p_1}(N)=\dfrac{(1-\alpha).h_1-\alpha.h_2}{s-p_1}$. De même et après calculs semblables, on obtient $E_{p_2}(N)=\dfrac{(1-\beta).h_2-\beta.h_1}{p_2-s}$.

- Quant à $\underset{p}{Max}\,E_p(N)$, on admettra qu'il s'agit d'un extrema atteint au voisinage de $p=s$ et dont la valeur est $E_s(N)=\dfrac{h_1.h_2}{\sigma^2}$ avec $\sigma=s.(1-s)$ (cf. résultat général mentionné dans l'application 1.4 antérieure, la variance ayant pour expression $p.(1-p)$ dans le cas de la loi de BERNOULLI. D'où, le résultat annoncé, $E_s(N)=\underset{p}{Max}\,E_p(N)=\dfrac{h_1.h_2}{s.(1-s)}$.

II-3°) L'application numérique $\alpha=1\%,\beta=5\%,p_1=0{,}10,p_2=0{,}30$ conduit, suivant les formules antérieures, à $h_1=2{,}21;h_2=3{,}37;s=0{,}25$, d'où $E_s(N)=39{,}65$, soit la valeur 40 par arrondi.

Ces calculs montrent ici encore, **l'économie** à attendre d'un *plan progressif* qui, en moyenne, exige 40 prélèvements, le *plan simple* exigeant quant à lui 53 prélèvements pour obtenir les mêmes risques (cf. 1ère question).

L'inconvénient en demeure cependant l'incertitude liée aux fluctuations de N autour de sa moyenne dont on a montré qu'elle était bornée supérieurement par 40 pour l'exemple traité ici. On pallie à cette difficulté en tronquant le test à $3.E_s(N)$ une décision étant alors prise suivant qu'on se trouve au dessus (H_1) ou au dessous (H_0) de la droite d'équation $y=s.n$.

2.6 Ajustement par une loi uniforme

Enoncé : Pour une étude de maladies cardio-vasculaires, l'enquête menée dans quatre villes à travers des échantillons de tailles respectives 200,400,300,100 révèle les nombres de malades suivants :

Ville i	1	2	3	4	Σ
f_i	22	38	25	16	100

Ces données montrent-elles une différence significative d'une ville à l'autre quant au nombre de personnes atteintes ?. On admettra un niveau de signification α égal à 5%.

Solution : *L'absence de différence significative entre villes* équivaut au *caractère uniforme* du nombre de malades sur chacune des villes en question, ce qui induit, en théorie, un nombre de malades proportionnel à la taille de l'échantillon considéré.

Ramenant donc le problème posé à un **test d'ajustement** dont la **distance de chi- deux** forme ici la méthode la plus appropriée, on est donc conduit au tableau de calculs ci-après, n étant le nombre de classes, soit $n=4$, et N la taille $\sum_{i=1}^{i=n} f_i$ de l'échantillon des malades considéré, soit $N=100$.

Villes i	Fréquences observées f_i	Fréquences théoriques $N.p_i$
1	22	20
2	38	40
3	25	30
4	15	10
Σ	100	100

Toutes les classes étant suffisamment significatives puisque $N.p_i \succ 5, i=1,2,3,4$, il n'y a pas de regroupement à opérer, la **distance de chi- deux calculée**, soit $\sum_{i=1}^{i=n} \frac{(f_i - N.p_i)^2}{N.p_i}$ étant numériquement égale à 3,63.

• Or, la statistique associée à la distance de chi- deux suit la loi de χ^2 à $\nu = n-1-k$ degrés de libertés où $n=4$ et $k=0$, aucun paramètre n'étant à estimer pour caractériser la loi uniforme par laquelle on modélise. Numériquement, $\nu=3$.

Notant par χ^2_α le seuil vérifiant $Prob(\chi^2(3) \geq \chi^2_\alpha) = \alpha$, la lecture dans la table de valeurs annexée, fournit pour $\alpha = 0,05$ et $\nu = 3$, le résultat $\chi^2_\alpha = 7,815$ *ce qui n'autorise pas à considérer la différence entre villes comme significative.*

2.7 Tests non paramétriques de conformité à une valeur standard

Non limités aux problèmes de comparaison entre échantillons tels ceux développés en rappels de cours, les tests non paramétriques forment, dans le cadre de la conformité à une valeur standard, une alternative aux tests paramétriques classiques (notamment STUDENT), dès qu'on n'entre plus dans le cadre d'un modèle gaussien ou d'échantillons de grande taille.

Enoncé : Plusieurs adaptations de tests mentionnés dans les rappels de cours du présent chapitre, sont proposées dans cette application.

On mesure les tailles de dix étudiants de sexe masculin, les données recensées étant les suivantes, en cm :

171	165	172	174	178	180	175	180	167	169

Admettant que la taille moyenne d'un adulte est supposée être égale à $m_0 = 170$ cm et notant par m l'espérance de taille d'un étudiant, on souhaite, à travers les diverses méthodes ci-après, déterminer si cette dernière moyenne m est supérieure ou non à l'espérance de taille générale, à savoir $m_0 = 170$, bref un test unilatéral $\begin{cases} H_0 : m = m_0 \\ H_1 : m \succ m_0 \end{cases}$ dont on fixera à 5% le niveau de signification.

1°) (test des signes) : Supposant la distribution des tailles symétrique, appliquer le test des signes aux différences $d_i = x_i - m_0$ (sous l'hypothèse H_0) et en déduire la décision à retenir.

2°) (test des rangs signés de WILCOXON) : Dans ce test plus puissant que le précédent et qui exige la symétrie de la distribution des données (par suite, la coïncidence de la moyenne et de la médiane), on applique la procédure exposée en rappels de cours (cf. paragraphe 3.3.b) aux différences $d_i = x_i - m_0$.

Reprenant les données de la précédente question, quelle conclusion résulte de la mise en œuvre du test en question, α étant toujours égal à 5% ?.

3°) (test paramétrique de STUDENT)

On admet que la taille d'un individu est modélisée par la loi normale. Toujours suivant les données susmentionnées et au niveau de signification $\alpha = 0,05$, quelle conclusion peut-on retenir quant au test $\begin{cases} H_0 : m = m_0 \\ H_1 : m \succ m_0 \end{cases}$?

Solution : 1°) Portant sur la **médiane** (mais aussi sur la moyenne, si on fait l'hypothèse supplémentaire de la symétrie de la distribution étudiée), le **test des signes** *ne requiert aucune hypothèse sur la distribution des données*. En contrepartie, il est d'une *puissance très faible* car faisant abstraction des valeurs prises (seuls les signes des différences $d_i = x_i - m_0$ sont utilisés). Il entraîne donc fort logiquement, une perte d'information importante.

Pour le cas proposé et comme cela est indiqué ci-dessus, la *symétrie de la distribution* ramène le *test de moyenne* $\begin{cases} H_0 : m = m_0 \\ H_1 : m \succ m_0 \end{cases}$ à un *test de médiane*. La médiane étant définie pour rappel, par le nombre c tel que $\Pr ob(X \succ c) = \Pr ob(X \prec c)$ (soit, la valeur commune $\frac{1}{2}$ lorsqu'il s'agit de variables continues), il en résulte que la variable D qui caractérise le nombre de différences $d_i = x_i - m_0$ positives ($x_i, 1 \le i \le n$, désignant les valeurs proposées de l'échantillon), suit la **loi binomiale** $B(n, \frac{1}{2})$ lorsque l'hypothèse $H_0 : m = m_0$ est vérifiée.

Concrètement, sept différences d_i positives sont constatées dans l'échantillon proposé, ce qui entraîne $D_{calculé} = 7$. Reprenant les résultats des rappels de cours relatifs au test des signes pour comparaison de deux distributions (cf. paragraphe 3.3.a), il est immédiat que la **région critique** du test a pour forme $D \geq D_\alpha$, D_α étant caractérisé par la relation

$\alpha = \Pr ob(décider \quad H_1 / H_0 \quad vraie) = \Pr ob(D \geq D_\alpha / m = m_0)$.

Un *calcul direct* ou le recours à la *table annexée* relative au *test binomial* montre que la première valeur de D telle que $\Pr ob(D \geq D_\alpha) \leq \alpha = 0,05$ est $D_\alpha = 8$. L'inégalité $D_{calculé} = 7 \leq D_\alpha = 8$ *ne permet donc pas d'écarter l'hypothèse* H_0, du moins pour le test unilatéral et en limitant à $\alpha = 5\%$ le risque de conclure à tort au rejet de H_0.

• A noter que si des valeurs de l'échantillon étaient égales à m_0, il conviendrait de les écarter préalablement au décompte des différences d_i positives.

2°) La mise en place du **test des rangs signés de WILCOXON** appliqué aux différences $d_i = x_i - 170$, conduit après *tri par ordre croissant* des $|d_i|$ aux *calculs des rangs et signes* ci-dessous :

i	1	10	3	9	4	7	2	5	6	8
x_i	171	169	172	167	174	175	-165	178	180	180
$d_i = x_i - 170$	1	-1	2	-3	4	5	-5	8	10	10
$\lvert d_i \rvert$	1	1	2	3	4	5	5	8	10	10
Rangs (*)	1,5	1,5	3	4	5	6,5	6,5	8	9,5	9,5
Signes	+	-	+	-	+	+	-	+	+	+

(*) Le principe utilisé est celui du rang moyen lorsqu'il y a des valeurs ex-aequo. Par exemple, les deux premières valeurs qui correspondent aux deux premières positions 1 et 2, ont pour rang moyen $(1+2)/2 = 1,5$.

• Retenant pour **fonction discriminante** du test, la statistique W^+ définie par *la somme des rangs signés positivement*, il vient pour l'échantillon étudié, l'expression numérique $W^+_{calculé} = 1,5+3+5+6,5+8+9,5+9,5 = 43$. Bien évidemment, on remarquera que $W^-_{calculé} = 1,5+4+6,5 = 12$ et que $W^+_{calculé} + W^-_{calculé} = 43+12 = \dfrac{n.(n+1)}{2}$, avec $n = 10$.

Sous l'hypothèse H_0, et reprenant les éléments fournis en rappels de cours (cf. paragraphe 3.3.b du présent chapitre), la statistique W^+ a pour moyenne $\dfrac{n.(n+1)}{4}$ (puisque $E(W^+) = E(W^-) = \dfrac{1}{2}.\sum_{k=1}^{k=n} k$) et pour variance $\dfrac{n.(n+1).(2n+1)}{24}$. A partir de $n = 15$, on peut envisager une *approximation* de $\dfrac{W^+ - \dfrac{n.(n+1)}{4}}{\sqrt{\dfrac{n.(n+1).(2n+1)}{24}}}$ par la **loi normale** *centrée réduite*, soit $\xi : N(0,1)$.

Il s'ensuit, relativement au *test unilatéral* qui est proposé ici, la région critique caractérisée par $W^+ \geq W_\alpha$ où W_α vérifie $\alpha = \Pr ob(W \geq W_\alpha / m = m_0)$, la décision à retenir étant H_1 si $W_{calculé} \geq W_\alpha$ et H_0 sinon.

Pour ce qui est des petites valeurs de n, ce qui n'autorise pas l'approximation par la loi normale, on pourra utiliser des *tables de valeurs* telle celle annexée, qui fournissent la valeur du seuil W_α en fonction de n et de α. Ainsi, pour $n = 10$ et $p = 1 - \alpha = 0,95$ lit-on $W_\alpha = 44$.

- Ici encore et bien qu'il s'agisse d'un test plus puissant que le test des signes, la relation $W_{calculé} = 43 \prec W_\alpha = 44$ *ne permet pas de rejeter l'hypothèse* H_0.

- A noter que, comme précédemment, il faudrait éliminer les différences $d_i = x_i - 170$ qui sont nulles, avant l'affectation des rangs et le calcul de W^+.

- Enfin, et bien que n soit petit et inférieur à 15, l'approximation de W^+ par la loi normale conduirait au seuil $W_\alpha = \frac{n.(n+1)}{4} + 1,645.\sqrt{\frac{n.(n+1)(2n+1)}{24}}$ (puisque pour $\alpha = 0,05$, $\Pr ob(\xi \geq t_\alpha) = 0,05 \Rightarrow t_\alpha = 1,645$), soit numériquement, la valeur $W_\alpha = 43,64$. C'est très proche de la valeur précédente fournie par la table ad hoc, et cela légitime largement l'usage de l'approximation normale, même pour les petites valeurs de n.

3°) Si on admet le résultat bien connu suivant lequel la distribution du poids aléatoire d'un individu s'apparente à une **loi normale** (*théorie des causes élémentaires et théorème central limite*), on entre alors dans le domaine d'application du **test** t **de STUDENT**, *test paramétrique* qui est *le plus puissant* des tests considérés ici.

La variance étant inconnue et devant être estimée par $\hat{S}^2$, on est ainsi conduit à une région critique de la forme $\bar{x} \geq \pi$ (puisqu'en effet, pour le cas dont il s'agit, à savoir $\begin{cases} H_0 : m = m_0 \\ H_1 : m \succ m_0 \end{cases}$, le test est unilatéral), la fonction pivotale $T = \frac{\overline{X} - m}{\hat{S}/\sqrt{n}}$ suivant par ailleurs, la loi de STUDENT à $\nu = n - 1$ degrés de liberté.

Utilisant la variable auxiliaire $Y = X - 170$ pour simplifier les calculs de $\bar{x}$ et de $\hat{S}$, ces derniers sont rassemblés ci-dessous :

x_i	171	165	172	174	178	180	175	180	167	169	171
$y_i = x_i - 170$	1	-5	2	4	8	10	5	10	-3	-1	31
y_i^2	1	25	4	16	64	100	25	100	9	1	945

$\bar{y} = 3,1$

$\bar{x} = \bar{y} + 170$

$\widehat{S_Y}^2 = \frac{1}{10-1}.\left[\sum_{i=1}^{i=10} y_i^2 - 10.\bar{y}^2\right] = (5,26)^2$

$\widehat{S_X} = \widehat{S_Y} = 5,26$

Explicitant la région critique $\bar{x} \geq \pi$ où π vérifie $\alpha = \Pr ob(\bar{x} \geq \pi / m = m_0)$, il vient la relation $\alpha = \Pr ob(T \geq \frac{\pi - m_0}{\hat{S}/\sqrt{n}})$, soit $\pi = m_0 + t_\alpha . \hat{S}/\sqrt{n}$, t_α vérifiant $\Pr ob(T \geq t_\alpha) = \alpha$.

Numériquement, la lecture dans la table des valeurs de la loi de STUDENT (cf. annexes) du nombre $t_{0,05}$ tel que $Prob(T \geq t_{0,05}) = 0,05$, conduit pour un nombre de degré de liberté $\nu = 10-1 = 9$, à la valeur $t_{0,05} = 1,833$. D'où le seuil $\pi = 173,05$.

• Or, $\bar{x} = 173,1$, valeur qui cette fois, est *supérieure au seuil critique* π, *ce qui conduit à retenir l'hypothèse* H_1 **au contraire des tests précédents**. Ce résultat souligne la puissance plus élevée du test paramétrique qui est plus apte, à risque égal, à discerner l'hypothèse H_1 de l'hypothèse H_0

3. Tests à deux échantillons sous modèle gaussien

3.1 Un exemple utilisant les tests de STUDENT et de FISHER SNEDECOR

Enoncé : Pour comparer l'influence de deux régimes alimentaires A et B sur le développement des doryphores, un entomologiste a mesuré, lors de la mue imaginale, le poids d'insectes élevés dans les conditions A pour les uns, et dans les conditions B pour les autres. Il a obtenu les résultats suivants :

Régime A (9 insectes mâles) :

100	94	119	111	113	84	102	107	99

Régime B (8 insectes mâles) :

107	115	99	111	114	127	145	140

Le poids d'un insecte choisi au hasard dans un élevage est une variable aléatoire que l'on désignera par X dans le cas A et par Y dans le cas B. On admet que X et Y suivent une loi normale.

1°) Montrer qu'il n'y a pas lieu de penser que X et Y aient des variances différentes. Pour la suite, on notera par σ^2 la valeur commune de ces deux variances.

2°) Exprimer une estimation de σ^2, soit $\hat{S}^2$, dont on calculera espérance mathématique et variance. En déduire les propriétés de $\hat{S}^2$.

3°) Estimer σ^2 par intervalle de confiance au seuil 95%.

4°) Montrer que le régime B est plus favorable au développement des insectes que le régime A.

Solution : 1°) Notant par $N(m_A, \sigma_A)$ et $N(m_B, \sigma_B)$ les lois respectives des poids des insectes suivant les régimes respectifs A et B, le test proposé ici est **bilatéral**, d'énoncé $\begin{cases} H_0 : \sigma_A^2 = \sigma_B^2 \\ H_1 : \sigma_A^2 \neq \sigma_B^2 \end{cases}$.

La mise en œuvre du **test** ad hoc **de FISHER SNEDECOR** (cf. rappels des cours du présent chapitre, paragraphe 2.4.a), conduit à la **région critique** :

$$W = \left\{ (x_1, x_2, ..., x_{n_A}), (y_1, y_2, ..., y_{n_B}) / F = \frac{\widehat{S_A}^2}{\widehat{S_B}^2} \notin]\pi_1, \pi_2[\right\}$$

Dans la formule antérieure, il est précisé que $\widehat{S_A}^2$ et $\widehat{S_B}^2$ désignent respectivement les *estimateurs ponctuels* (variances empiriques corrigées) de σ_A^2 et σ_B^2, variances dont les expressions sont $\widehat{S_A}^2 = \frac{1}{n_A - 1}.\sum_{i=1}^{i=n_A}(X_{i,A} - \overline{X_A})^2$ et $\widehat{S_B}^2 = \frac{1}{n_B - 1}.\sum_{i=1}^{i=n_A}(X_{i,B} - \overline{X_B})^2$.

Or, la statistique $\frac{\widehat{S_A}^2 / \sigma_A^2}{\widehat{S_B}^2 / \sigma_B^2}$ suivant la loi de FISHER SNEDECOR à $\nu_A = n_A - 1$ et $\nu_B = n_B - 1$ degrés de liberté, il en résulte immédiatement, sous l'hypothèse H_0, les expressions des seuils π_1 et π_2 vérifiant $Prob(F \le \pi_1) = \Pr ob(F \ge \pi_2) = \frac{\alpha}{2}$.

• Numériquement, les données étant $n_A = 9, n_B = 8$ et les valeurs $x_{i,A}$ et $x_{i,B}$ étant celles présentées dans l'énoncé, la mise en œuvre des formules précédentes conduit aux *estimations* :

$$\overline{x_A} = \frac{\sum_{i=1}^{i=n_A} x_{i,A}}{n_A} = 103,22 \qquad \widehat{s_A}^2 = \frac{1}{n_A - 1}.\left[\sum_{i=1}^{i=n_A} x_{i,n_A}^2 - n_A.\overline{x_A}^2\right] = 112,94$$

$$\overline{x_B} = \frac{\sum_{i=1}^{i=n_A} x_{i,B}}{n_B} = 119,75 \qquad \widehat{s_B}^2 = \frac{1}{n_B - 1}.\left[\sum_{i=1}^{i=n_A} x_{i,n_B}^2 - n_B.\overline{x_B}^2\right] = 260,79$$

D'où $F_{calculé} = \widehat{s_A}^2 / \widehat{s_B}^2 = 0,433$.

• Pour la loi $F(\nu_A = n_A - 1 = 8, \nu_B = n_B - 1 = 7)$ dont la *table des valeurs de la fonction de répartition* est annexée, et pour un *niveau de test* (erreur de première espèce α) égal à 5%, on a immédiatement $\Pr ob(F \ge \pi_2) = \alpha/2 = 0,025 \Rightarrow \Pr ob(F \prec \pi_2) = 0,975$, le seuil π_2 lu dans la table étant égal ,quant à lui, à $\pi_2 = 4,90$.

D'autre part, $Prob(F \le \pi_1) = 0,025$ équivaut à $\Pr ob(\frac{1}{F} \ge \frac{1}{\pi_1}) = 0,025$, soit par complémentarité, $\Pr ob(\frac{1}{F} \prec \frac{1}{\pi_1}) = 0,975$. La variable $\frac{1}{F}$ suivant la loi de FISHER SNEDECOR, $F(7,8)$ (et non plus $F(8,7)$ comme c'est le cas pour F), il s'ensuit immédiatement $\frac{1}{\pi_1} = 4,53 \Rightarrow \pi_1 = 0,22$.

• Pour les échantillons considérés, on a $\pi_1 = 0,22 \prec F_{calculé} = \frac{\widehat{s_A}^2}{\widehat{s_B}^2} = 0,433 \prec \pi_2 = 4,90$.

On est manifestement dans la zone d'acceptation de H_0.

2°) Se référant à l'application 1.4 du chapitre I, la somme $(n_A-1).\dfrac{\widehat{S_A}^2}{\sigma_A^2}+(n_B-1).\dfrac{\widehat{S_B}^2}{\sigma_B^2}$ suit une loi de chi-deux à n_A+n_B-2 degrés de liberté, *l'estimateur commun* $\hat{S}^2$ qui en résulte, sous l'hypothèse $\sigma_A^2=\sigma_B^2=\sigma^2$, vérifiant :

$$\frac{(n_A-1).S_A^2+(n_B-1).S_B^2}{\sigma^2}=(n_A+n_B-2).\frac{\hat{S}^2}{\sigma^2}\Rightarrow\hat{S}^2=\frac{(n_A-1).\widehat{S_A}^2+(n_B-1).\widehat{S_B}^2}{(n_A+n_B-2)}.$$

Numériquement, il en ressort, pour les échantillons en question, *l'estimation* $\hat{s}^2=\dfrac{8\times112,94+7\times260,79}{9+8-2}=181,94$ de σ^2.

• Revenant au cas général de l'estimateur d'une variance et se reportant plus précisément à l'application 1.2 du chapitre I, il est rappelé que $E(\hat{S}^2)=\sigma^2$ et $Var(\hat{S}^2)=\dfrac{2.\sigma^4}{n-1}$ (du moins, pour le cas d'un modèle normal).

Pour l'estimateur $\hat{S}^2=\dfrac{(n_A-1).\widehat{S_A}^2+(n_B-1).\widehat{S_B}^2}{(n_A+n_B-2)}$ et posant $\sigma_A^2=\sigma_B^2=\sigma^2$, on a donc immédiatement, par linéarité :

$$E(\hat{S}^2)=\frac{(n_A-1).E(\widehat{S_A}^2)+(n_B-1).E(\widehat{S_B}^2)}{(n_A+n_B-2)}=\frac{1}{(n_A+n_B-2)}.\left[(n_A-1)+(n_B-1)\right].\sigma^2=\sigma^2$$

Ainsi, $\hat{S}^2$ est-il un estimateur **sans biais**.

D'autre part, par *pseudo linéarité*, on a :

$$Var(\hat{S}^2)=\frac{1}{(n_A+n_B-2)^2}.\left[(n_A-1)^2.Var(\widehat{S_A}^2)+(n_B-1)^2.Var(\widehat{S_B}^2)\right],$$

soit compte tenu des expressions $Var(\widehat{S_A}^2)=\dfrac{2.\sigma_A^2}{n_A-1}, Var(\widehat{S_B}^2)=\dfrac{2\sigma_B^2}{n_B-1}$, et de *l'homoscédasticité* $\sigma_A^2=\sigma_B^2=\sigma^2$, le résultat :

$$Var(\hat{S}^2)=\frac{1}{(n_A+n_B-2)^2}.\left[(n_A-1)^2.\frac{2\sigma^4}{n_A-1}+(n_B-1)^2.\frac{2\sigma^4}{n_B-1}\right]=\frac{2\sigma^4}{n_A+n_B-2}.$$

Ainsi, $\hat{S}^2$ est-il un estimateur **convergent** puisque $Var(\hat{S}^2)$ tend vers 0 lorsque n_A et n_B sont grands.

3°) S'inspirant des rappels de cours du chapitre II (cf. paragraphe 4.4) et partant du résultat précédent suivant lequel $(n_A+n_B-2).\dfrac{\hat{S}^2}{\sigma^2}$ suit la loi du chi-deux à $\nu=n_A+n_B-2$ degrés de liberté, la recherche de l'intervalle de confiance $[a,b]$ satisfaisant à la condition $Prob(a\le\sigma^2\le b)=1-\alpha$, conduit au résultat :

$$\frac{(n_A+n_B-2).\hat{S}^2}{t_A}\le\sigma^2\le\frac{(n_A+n_B-2).\hat{S}^2}{t_B},\ t_A \text{ et } t_B \text{ définis ci-après.}$$

Pour la loi $\chi^2(n_A+n_B-2)$, les seuils t_A et t_B précédents vérifient respectivement les relations $\Pr ob(\chi^2(n_A+n_B-2)\succ t_A)=\alpha/2$ et $\Pr ob(\chi^2(n_A+n_B-2)\prec t_B)=\alpha/2$.

Numériquement et pour $1-\alpha=95\%$, on a successivement, par lecture dans la table de valeurs annexée (distribution de χ^2), $\Pr ob(\chi^2(15)\succ t_A)=0,025\Rightarrow t_A=27,488$ et $\Pr ob(\chi^2(15)\prec t_B)=0,025\Rightarrow \Pr ob(\chi^2(15)\succ t_B)=0,975\Rightarrow t_B=6,262$.

Ainsi obtient-on pour σ^2, **l'intervalle de confiance** $15\times\frac{181,94}{27,488}\le\sigma^2\le 15\times\frac{181,94}{6,262}$, soit $99,28\le\sigma^2\le 435,81$.

4°) On teste dans cette question l'hypothèse $H_0: m_A=m_B$ contre l'hypothèse $H_1: m_A\prec m_B$. La réponse est le **test paramétrique t de STUDENT à deux échantillons**, les échantillons étant ici de petites tailles et l'hypothèse d'*homoscédasticité* étant admise par ailleurs ($\sigma_A^2=\sigma_B^2$).

Dans ces conditions, la *statistique « discriminante »* étant $\overline{X_A}-\overline{X_B}$, la **région critique** est caractérisée par $\overline{X_A}-\overline{X_B}\le t_\alpha\cdot\sqrt{\frac{1}{n_A}+\frac{1}{n_B}}\hat{S}$ (où t_α vérifie $\Pr ob(T\le t_\alpha)=\alpha$, T étant par ailleurs, la variable de STUDENT à $\nu=n_A+n_B-2$ degrés de liberté (cf. paragraphe 2.4.b des rappels de cours du présent chapitre).

Par lecture dans la table de valeurs annexée et pour $\alpha=0,05$ (cf. table de distribution de STUDENT), on obtient aisément pour $\nu=15$, le seuil $t_\alpha=-1,753$ et à fortiori, la **régle de décision** suivant laquelle :

→ on décide H_1 si $\overline{x_A}-\overline{x_B}\le -1,753\times\sqrt{\frac{1}{9}+\frac{1}{8}}\times\sqrt{181,94}=-11,49$;

→ on décide H_0 sinon.

Or, sur les échantillons dont on dispose, on a $\overline{x_A}-\overline{x_B}=-16,53$. C'est donc la *décision* H_1, à savoir la conclusion suivant laquelle *le régime B est plus favorable au développement que le régime A* qu'il faut retenir ici.

3.2 Comparaison de moyennes sur échantillons appariés

Enoncé : Une société de location de voitures met en place une expérience afin de décider si deux types de pneus sont différents ou non. Onze voitures sont conduites sur un parcours précis avec des pneus de type A. Ceux-ci sont alors remplacés par des pneus de type B et les voitures sont de nouveau conduites sur le même parcours.

Les consommations de carburant en km/litre des voitures en question, pour chacun des deux types de pneus A et B, soient X_A et X_B, sont indiquées dans le tableau ci-dessous, l'hypothèse de leur distribution gaussienne étant par ailleurs admise.

Consom/Voiture	1	2	3	4	5	6	7	8	9	10	11
X_A	4,2	4,7	6,6	7,0	6,7	4,5	5,7	6	7,4	4,9	6,1
X_B	4,1	4,9	6,2	6,9	6,8	4,4	5,7	5,8	6,9	4,9	6,0

Au niveau de signification 5%, quelle conclusion peut-on en retenir ?

Solution : *L'hypothèse de normalité* permet d'envisager ici un **test paramétrique de STUDENT** (qu'on choisira *bilatéral*), tel celui exposé en rappels de cours du présent chapitre (cf. paragraphe 2.5).

Utilisant la variable $D = X_A - X_B$, ce test se ramène, pour ladite variable, à tester la nullité ou non de la moyenne $m_A - m_B$ et utilise la statistique $\dfrac{\overline{D} - (m_A - m_B)}{\widehat{S}_D / \sqrt{n}}$ dont la loi est de type STUDENT à $\nu = n - 1$ degrés de liberté.

Les calculs qui en résultent sont les suivants :

- $\overline{d} = \dfrac{\sum_{i=1}^{i=11} d_i}{11} = \dfrac{1}{11}.[(4,2-4,1)+(4,7-4,9)+...+(6,1-6,0)] = 0,109$;
- $\widehat{s_D}^2 = \dfrac{1}{11-1}.\left[\sum_{i=1}^{i=11} d_i^2 - 11.\overline{d}^2\right] = 0,0409 \Rightarrow \widehat{s_D} = 0,202$.

La région critique est définie par $\overline{d} \notin]-\pi, +\pi[$, où π vérifie :

$$\Pr ob(D \notin]-\pi, +\pi[/ m_A - m_B = 0) = \alpha$$

Il en résulte la relation $\Pr ob(|T| = \left|\dfrac{\overline{D}}{\widehat{S}_D/\sqrt{n}}\right| \geq \dfrac{\pi}{\widehat{S}_D/\sqrt{n}}) = \alpha$, où T désigne la loi de STUDENT à $\nu = n-1$ degrés de liberté. La lecture dans la table de valeurs annexée (cf. table de STUDENT, test bilatéral), fournit pour t_α vérifiant $\Pr ob(|T| \geq t_\alpha) = \alpha$ et pour $\alpha = 0,05$ et $n = 11$, la valeur $t_\alpha = 2,2281$.

• Ainsi $\pi = 0,136$. Or, par rapport aux données collectées, $\overline{d}$ qui est égal à 0,109 reste donc inférieur au seuil critique susmentionné $\pi = 0,136$. *On ne peut donc pas, dans ces conditions, retenir l'hypothèse d'une différence significative entre ls pneus utilisés* (du moins au seuil $\alpha = 0,05$ ni même d'ailleurs, au seuil $\alpha = 0,10$).

3.3 Comparaison de variances entre deux types de solutions aqueuses

Enoncé : On se propose d'illustrer deux tests paramétriques sous modèle normal pour comparer des variances entre deux populations. Plus précisément, on s'intéresse à une méthode qui doit servir à l'analyse de deux catégories de solutions. Dans l'une la concentration de l'élément dosé est de 50mg/l et dans l'autre, elle est deux fois plus grande.

Pour savoir si la reproductabilité de la méthode est la même dans les deux cas, on l'applique à deux séries de 10 dosages concernant respectivement les deux catégories de solution susmentionnées, c'est-à-dire à 10 solutions dont la concentration de l'élément considéré a pour valeur 50mg/l et 10 solutions dont la concentration de l'élément considéré a pour valeur 100mg/l.

La méthode conduit, pour ces dosages, aux valeurs présentées ci-après.

Solutions A à 50mg/l	52	49	49	52	50	52	48	47	53	51
Solutions B à 100mg/l	102	98	97	101	103	104	95	100	99	97

L'hypothèse H_0 à contrôler est $\sigma_A^2 = \sigma_B^2$ et on suppose que la normalité des résultats fournis par la méthode d'analyse en question a été vérifiée lors de sa conception.

1°) Au niveau de signification $\alpha = 0,05$, quelle conclusion peut-on retenir ici ?

2°) Quelle est la valeur du risque de 2ème espèce β lorsque, sous l'hypothèse H_1, on a en réalité $\lambda = \frac{\sigma_A}{\sigma_B} = \frac{1}{2}$? Qu'en est-il lorsque $\lambda = \frac{1}{3}$? Interpréter ces résultats.

3°) Considérant le cas $\lambda = \frac{1}{2}$, quelle est la taille minimale commune $(n = n_A = n_B)$ que doivent avoir les deux échantillons prélevés pour que le risque β soit ramené de la valeur obtenue au 2°) à la valeur 10% ?

4°) Que deviennent les calculs antérieurs lorsque les concentrations de l'élément dosé dans les solutions analysées ne sont plus exactement connues ?

5°) On utilise cette fois les étendues et non pas les variances pour caractériser la dispersion des résultats fournis par la méthode, respectivement sur les solutions à 50mg/l et les solutions à 100mg/l.

A partir de la fonction discriminante définie par $F' = \frac{(x_A)_{\max} - (x_A)_{\min}}{(x_B)_{\max} - (x_B)_{\min}} = \frac{w_A}{w_B}$, qu'en conclure quant au test bilatéral $\begin{cases} H_0 : \sigma_A^2 = \sigma_B^2 \\ H_1 : \sigma_A^2 \neq \sigma_B^2 \end{cases}$?.

On admettra que, pour $P = 0,975$ et $n_A = n_B = 10$, le nombre F_P qui vérifie la relation $\Pr ob(F' \prec F_P) = P$ est égal à 2,11 (sous l'hypothèse $H_0 : \sigma_A^2 = \sigma_B^2$).

Solution : 1°) Le test à mettre en œuvre est à priori de type **bilatéral** $\begin{cases} H_0 : \sigma_A^2 = \sigma_B^2 \\ H_1 : \sigma_A^2 \neq \sigma_B^2 \end{cases}$.
Minatenant, si on suppose que la variance ne peut qu'augmenter avec la concentration, ou si, à l'examen des résultats expérimentaux, il s'avère peu vraisemblable que $\sigma_A^2 \succ \sigma_B^2$, on peut opter pour un *test unilatéral* $\begin{cases} H_0 : \sigma_A^2 = \sigma_B^2 \\ H_1 : \sigma_A^2 \prec \sigma_B^2 \end{cases}$ *dont la puissance est plus élevée.*

D'autre part, l'hypothèse de normalité des données autorise le recours au **test paramétrique de FISHER SNEDECOR**, l'hypothèse à considérer ici étant celle où *les moyennes m_A et m_B des distributions comparées sont connues* (respectivement égale à $50mg/l$ et $100mg/l$).

Se référant aux rappels de cours du présent chapitre (cf. paragraphe 2.4.a), la *fonction discriminante du test*, est dans les conditions susmentionnées, le rapport S_A^2 / S_B^2 où

$S_A^2 = \frac{1}{n_A} \cdot \sum_{i=1}^{i=n_A} (X_{i,A} - m_A)^2$ et $S_B^2 = \frac{1}{n_B} \cdot \sum_{i=1}^{i=n_B} (X_{i,B} - m_B)^2$.

Numériquement, $S_A^2 = \frac{1}{10}.\left[(52-50)^2 + ... + (51-50)^2\right] = 3,7$. De même, on a $S_B^2 = \frac{1}{10}.\left[(102-100)^2 + ... + (97-100)^2\right] = 7,8$.

Or la statistique $F = \frac{S_A^2 / \sigma_A^2}{S_B^2 / \sigma_B^2}$ suit, pour m_A et m_B connues, la loi de FISHER SNEDECOR à $\nu_1 = n_A$ et $\nu_2 = n_B$ degrés de liberté. Sous l'hypothèse $H_0 : \sigma_A^2 = \sigma_B^2$, c'est aussi la loi suivie par la *fonction discriminante* S_A^2 / S_B^2, la **région critique** du *test bilatéral* s'écrivant :

$$W = \left\{(x_{1,A}, x_{2,A}, ..., x_{n_A,A}), (x_{1,B}, x_{2,B}, ..., x_{n_B,B}) / F = \frac{S_A^2}{S_B^2} \notin]\pi_1, \pi_2[\right\}.$$

Utilisant la *symétrie des risques*, on détermine π_1 et π_2 à partir des conditions respectives $\Pr ob(F \leq \pi_1) = \alpha/2$ et $\Pr ob(F \geq \pi_2) = \alpha/2$.

Ainsi, pour $\alpha = 0,05; \nu_1 = 10; \nu_2 = 10$, la consultation de la table de valeurs annexée (cf. fonction de répartition de la loi de FISHER SNEDECOR), conduit-elle, pour ce qui est du seuil π_2 vérifiant $\Pr ob(F \prec \pi_2) = 1 - \alpha/2 = 0,975$, à la valeur $\pi_2 = 3,72$. D'autre part, $\Pr ob(F \leq \pi_1) = 0,025 \Leftrightarrow \Pr ob(\frac{1}{F} \geq \frac{1}{\pi_1}) = 0,025$, soit $\Pr ob(\frac{1}{F} \prec \frac{1}{\pi_1}) = 0,975$.

Lorsque F suit la loi de FISHER SNEDECOR à (ν_1, ν_2) degrés de liberté, il est immédiat que $\frac{1}{F}$ suit la loi de FISHER SNEDECOR à (ν_2, ν_1) degrés de liberté, soit en l'occurrence $(10,10)$. Pour le cas présent, on a donc $\frac{1}{\pi_1} = 3,72 \Rightarrow \pi_1 = 0,269$.

• En définitive, et si on considère le test bilatéral, la *zone d'acceptation de l'hypothèse* H_0 (c'est-à-dire le complémentaire de la région critique) est caractérisée par la condition $0,269 \prec \frac{S_A^2}{S_B^2} \prec 3,72$. Or, pour les données dont on dispose, on a $\frac{S_A^2}{S_B^2} = \frac{3,7}{7,8} = 0,47$.

Manifestement, *l'hypothèse* H_0 *doit donc être retenue ici.*

2°) *L'erreur de 2ème espèce* est définie par la probabilité Prob(décider H_0 / H_1 vraie), c'est-à-dire $\Pr ob(0,269 \prec \frac{S_A^2}{S_B^2} \prec 3,72 / H_1 vraie)$. Or, la statistique $F = \frac{S_A^2 / \sigma_A^2}{S_B^2 / \sigma_B^2}$ suit la loi de FISHER SNEDECOR à n_A et n_B degrés de liberté.

Sous l'hypothèse $H_1 : \sigma_A^2 \neq \sigma_B^2$ et plus précisément si on suppose $\lambda^2 = \frac{\sigma_A^2}{\sigma_B^2} = \frac{1}{4}$, on a donc $\frac{1}{\lambda^2}.\frac{S_A^2}{S_B^2} = 4.\frac{S_A^2}{S_B^2}$ de loi $F(n_A = 10, n_B = 10)$.

En définitive, $\beta = \Pr ob(0,269 \prec \frac{S_A^2}{S_B^2} \prec 3,72) = \Pr ob(1,076 \prec 4.\frac{S_A^2}{S_B^2} \prec 14,88)$. Recourant à un *calculateur des tables de valeurs* de la loi de FISHER SNEDECOR, plus précisément la loi $F(10,10)$, (sur EXCEL par exemple, tel le site suivant l'accès google→ tables numériques en ligne→ www.geai.univ-brest.fr), on a immédiatement l'évaluation approchée $\beta \simeq \Pr ob(F \succ 1,076) = 0,45$ ($\Pr ob(F \prec 14,88)$ étant très proche de 1).

- On notera que la même valeur de β est obtenue lorsque c'est σ_A qui est le double de σ_B ($\lambda = 2$). En effet, on a alors $\beta = \Pr ob(\frac{0,269}{4} \prec \frac{1}{4}.\frac{S_A^2}{S_B^2} \prec \frac{3,72}{4})$, soit en passant aux inverses, $\beta = \Pr ob(\frac{4}{3,72} = 1,076 \prec 4.\frac{S_B^2}{S_A^2} \prec \frac{4}{0,269} = 14,88) = 0,45$ (puisque $4.\frac{S_B^2}{S_A^2}$ suit la loi $F(n_B, n_A)$ qui, numériquement, est aussi la loi $F(10,10)$.

On retiendra donc que lorsque l'une des variances est quatre fois plus grande que l'autre, il y a 45% de chances que le test effectué ne permette pas de mettre cet écart en évidence ce qui est considérable.

A titre complémentaire, si cet écart passe du stade 4 au stade 9 (soit $\lambda = \frac{1}{3}$), on a $\beta = \Pr ob(2,421 \prec 9.\frac{S_A^2}{S_B^2} \prec 33,48) = \Pr ob(F \succ 2,421)$. Suivant le calculateur cité plus haut, on trouve cette fois, $\beta = 9\%$, c'est-à-dire un risque heureusement plus faible !.

3°) Dans cette question, c'est sur la taille de l'échantillon que l'on souhaite agir, λ étant constant. On veut ramener ainsi, pour $\lambda = \frac{1}{2}$, la probabilité $\beta = \Pr ob(4.\pi_1 \prec 4.\frac{S_A^2}{S_B^2} \prec 4.\pi_2)$ à moins de 10%.

A cet effet, il est proposé de procéder par *approximations successives* à l'aide du calculateur en ligne susmentionné, en se méfiant toutefois du changement des valeurs des seuils π_1 et π_2 lorsqu'on modifie n_A et n_B. Il s'ensuit, α étant toujours égal à 5% et pour le test bilatéral, les résultats suivants :

$n = n_A = n_B$	π_1	π_2	$\beta = \Pr ob(F \succ 4.\pi_1)$
10	0,27	3,72	0,45
20	0,41	2,46	0,14
25	0,45	2,23	0,08
23	0,43	2,31	0,10

Bref, la valeur la plus proche de l'objectif recherché $\beta = 10\%$, est ici $n = n_A = n_B = 23$.

4°) Lorsque les *concentrations ne sont pas connues* exactement, on les estimera par les moyennes $\overline{x_A} = \frac{1}{n_A}.\sum_{i=1}^{i=n_A} x_{i,A}$ et $\overline{x_B} = \frac{1}{n_B}.\sum_{i=1}^{i=n_B} x_{i,B}$, ce qui numériquement, conduit aux estimations $\overline{x_A} = 50,3$ et $\overline{x_B} = 99,6$.

Toujours suivant les rappels de cours, ce sont les statistiques $\widehat{S_A}^2 = \frac{1}{n_A - 1}.\sum_{i=1}^{i=n_A}(X_{i,A} - \overline{X_A})^2$ et $\widehat{S_B}^2 = \frac{1}{n_B - 1}.\sum_{i=1}^{i=n_B}(X_{i,B} - \overline{X_B})^2$ qu'il faut considérer ici, la *fonction discriminante* étant désormais $F = \frac{\widehat{S_A}^2}{\widehat{S_B}^2}$. Numériquement, il vient les estimations $\widehat{s_A}^2 = \frac{1}{n_A - 1}.\sum_{i=1}^{i=n_A}(x_{i,A} - \overline{x_A})^2 = 4,011$ et $\widehat{s_B}^2 = \frac{1}{n_B - 1}.\sum_{i=1}^{i=n_B}(x_{i,B} - \overline{x_B})^2 = 8,49$.

La statistique à considérer alors est $F = \frac{\widehat{S_A}^2}{\widehat{S_B}^2}$ étant entendu que $\frac{\widehat{S_A}^2 / \sigma_A^2}{\widehat{S_B}^2 / \sigma_B^2}$ suit la loi de FISHER SNEDECOR à $\nu_A = n_A - 1$ et $\nu_B = n_B - 1$ degrés de liberté.

Sous l'hypothèse H_0 : $\sigma_A^2 = \sigma_B^2$ et pour la loi $F(9,9)$, on a par lecture dans la table annexée ou utilisation du calculateur cité précédemment, $\Pr ob(F \prec \pi_1) = 0,025$, ce qui entraîne $\pi_1 = 0,248$, et $\Pr ob(F \prec \pi_2) = 0,975$, d'où $\pi_2 = \frac{1}{\pi_1} = 4,03$. Relativement à la *zone d'acceptation de* H_0 qui est désormais $0,25 \prec \frac{\widehat{S_A}^2}{\widehat{S_B}^2} \prec 4,03$, on obtient, pour ce qui est des valeurs observées, $F_{calculé} = \frac{\widehat{s_A}^2}{\widehat{s_B}^2} = \frac{4,01}{8,49} = 0,48$, ce qui est une valeur très proche de celle obtenue au 1°) pour m_A et m_B connues et qui conduit ici encore à la *conclusion d'acceptation de* H_0 puisque $0,25 \prec F_{calculé} \prec 4,03$.

On remarquera à ce sujet, qu'entre les deux hypothèses « *moyennes connues* » et « *moyennes inconnues* », les règles de décision $]0,27 - 3,72[$ et $]0,25 - 4,03[$ sont assez proches. Quant à β, sous l'hypothèse $\lambda^2 = \frac{\sigma_A^2}{\sigma_B^2} = \frac{1}{4}$, on trouve présentement l'expression $\beta = \Pr ob(1 \prec \frac{\widehat{S_A}^2}{\widehat{S_B}^2} \prec 16,12)$, quantité approchée par $\Pr ob(F \succ 1) = 50\%$. Lorsque $\lambda = \frac{1}{3}$, cette erreur $\beta = \Pr ob(2,25 \prec \frac{\widehat{S_A}^2}{\widehat{S_B}^2} \prec 36,27)$ prend la valeur 12%.

5°) Décrivant les *dispersions*, *non plus par les écarts-types* mais par les **étendues**, w_A et w_B, on a pour ces dernières et à partir des données proposées, $w_A = 53 - 47 = 6$ et $w_B = 104 - 95 = 9$. Ainsi, la fonction discriminante $\frac{w_A}{w_B}$ a-t-elle pour valeur $\frac{w_A}{w_B} = 0,67$.

La **région critique** du *test bilatéral* a pour forme $F' = \frac{w_A}{w_B} \notin]\pi_1, \pi_2[$ où $\Pr ob(F \prec \pi_1) = 0,025$ et $\Pr ob(F \succ \pi_2) = 0,025$.

A partir de la tabulation des valeurs de la statistique $F' = \frac{w_A}{w_B}$ (sous l'hypothèse H_0) (cf. extraits fournis dans l'énoncé), on a, pour $n_A = n_B = 10$, $\Pr ob(F' \prec 2,11) = 0,975$.

On a donc $\pi_2 = 2,11$. D'autre part, $\Pr ob(F' \prec \pi_1) = 0,025 \Leftrightarrow \Pr ob(\frac{1}{F'} \succ \frac{1}{\pi_1}) = 0,025$, soit $\Pr ob(\frac{1}{F'} \prec \frac{1}{\pi_1}) = 0,975$. Ainsi, $\pi_1 = \frac{1}{\pi_2} = \frac{1}{2,11} = 0,47$.

- En conclusion, la *zone d'acceptation de* H_0 étant donc $0,47 \prec \frac{w_A}{w_B} \prec 2,11$ pour le test bilatéral $\begin{cases} H_0 : \sigma_A^2 = \sigma_B^2 \\ H_1 : \sigma_A^2 \neq \sigma_B^2 \end{cases}$, on est conduit ici à retenir ladite hypothèse puisque, suivant les données proposées $\frac{w_A}{w_B} = 0,47$, *conclusion corroborant les résultats du test de* FISHER SNEDECOR.

3.4 Comparaison de proportions

Enoncé : Deux groupes A et B sont formés chacun de 100 malades. Un sérum est administré au groupe A, mais pas au groupe B. A part cela, les traitements sont identiques. Les résultats montrent que dans les groupes A et B, respectivement 75 et 65 personnes guérissent.

1°) Tester l'hypothèse selon laquelle le sérum est efficace aux seuils respectifs 1%, 5% , et 10%.

2°) On travaille dans cette question sur des échantillons de plus grande taille, soit $n = 300$, les nombres de malades recensés étant respectivement de 225 et 195. Tester l'hypothèse d'efficacité du sérum au seuil $\alpha = 1\%$. Qu'en conclure ?

Solution : 1°) Notant respectivement par p_A et p_B les proportions des guérisons parmi les groupes « *traité* » (A) et « *non traité* » (B), le test de l'hypothèse d'efficacité du sérum est de type **unilatéral** : $\begin{cases} H_0 : p_A = p_B \\ H_1 : p_A \succ p_B \end{cases}$.

Suivant les résultats des rappels de cours du présent chapitre (cf . paragraphe 2.4.c), la *fonction discriminante* est la différence des *fréquences empiriques* $F_A - F_B = \frac{X_A}{n_A} - \frac{X_B}{n_B}$, dont, lorsque n_A et n_B sont suffisamment grands, la loi converge vers la **loi normale** $N(p_A - p_B, \sqrt{\frac{p_A.(1-p_A)}{n_A} + \frac{p_B.(1-p_B)}{n_B}})$.

En effet X_A et X_B qui désignent les effectifs des guérisons parmi, respectivement, les malades *traités* et les malades *non traités*, suivent les *lois binomiales* $B(n_A, p_A)$ et $B(n_B, p_B)$ dont le théorème de MOIVRE-LAPLACE assure la convergence vers les *lois normales* $N(p_A, \sqrt{n_A.p_A.(1-p_A)})$ et $N(p_B, \sqrt{n_B.p_B.(1-p_B)})$.

Des résultats immédiats $E(F_A - F_B) = E(F_A) - E(F_B) = p_A - p_B$ (linéarité de l'espérance) et $Var(F_A - F_B) = Var(F_A) + Var(F_B) = \frac{1}{n_A^2}.Var(X_A) + \frac{1}{n_B^2}.Var(X_B)$, découle la loi limite de $F_A - F_B$ explicitée ci-dessus.

• Il s'ensuit, relativement au test unilatéral proposé, la **région critique** $F_A - F_B \geq \pi$, la donnée de l'erreur de première espèce α conduisant à la relation :

$$\alpha = \Pr ob(F_A - F_B \geq \pi / p_A = p_B).$$

Finalement, en introduisant la *variable normale centrée réduite* associée à $F_A - F_B$, soit $\xi = \frac{F_A - F_B - (p_A - p_B)}{\sqrt{\frac{p_A.(1-p_A)}{n_A} + \frac{p_B.(1-p_B)}{n_B}}}$, on a, sous l'hypothèse $H_0 : p_A = p_B = p$, l'équation

$$\Pr ob(\xi \geq \frac{\pi}{\sqrt{p.(1-p)}.\sqrt{\frac{1}{n_A} + \frac{1}{n_B}}}) = \alpha.$$

Notant par t_α la valeur vérifiant $\Pr ob(\xi \geq t_\alpha) = \alpha$ (t_α lu dans la table de valeurs annexée de la fonction de répartition $\Pi(t) = \Pr ob(\xi \leq t)$, il vient immédiatement, le **seuil critique**, $\pi = t_\alpha.\sqrt{p.(1-p)}.\sqrt{\frac{1}{n_A} + \frac{1}{n_B}}$. Dans cette expression p demeure toutefois *inconnu*. On estimera cette valeur commune de p_A et p_B par l'estimateur $\hat{p} = \frac{X_A + X_B}{n_A + n_B}$.

• Numériquement, une lecture dans la table de valeurs susmentionnée, fournit pour les diverses erreurs de première espèce $\alpha = 1\%, \alpha = 5\%, \alpha = 10\%$, les valeurs correspondantes t_α ci-dessous :

α	1%	5%	10%
t_α	2,33	1,65	1,28

D'autre part, sous l'hypothèse H_0, la valeur commune $p = p_A = p_B$ et dont l'estimateur est $\hat{p} = \frac{X_A + X_B}{n_A + n_B}$ a pour estimation $\frac{75 + 65}{200} = 0,70$. Enfin, s'agissant des données observées, on a pour $F_A - F_B$, la valeur calculée $\frac{75}{100} - \frac{65}{100} = 0,10$.

En définitive, pour chacune des trois erreurs de première espèce susmentionnées, la décision à retenir est la la suivante :

	Erreur de première espèce α		
	1%	5%	10%
Région critique	$[0,150;+\infty[$	$[0,107;+\infty[$	$[0,083;+\infty[$
Décision retenue	H_0	H_0	H_1

Comme on pouvait s'y attendre, la conclusion dépend du risque qu'on est prêt à accepter et qui, pour ce qui est de α, correspond au risque de continuer à appliquer un traitement qui n'est pas efficace. Ainsi, si au niveau de signification 10%, l'efficacité du traitement peut être retenue, il en est autrement pour les niveaux inférieurs 1% et 5%.

2°) En *augmentant la taille de l'échantillon on accroît la fiabilité de la décision.* Ainsi, les nouvelles données $n_A = n_B = 300, \alpha = 1\%, X_A = 225, X_B = 195$, conduisent-elles à $t_\alpha = 2,33$; $p = 0,70$ et $\pi = 0,087$.

La *valeur calculée* de $F_A - F_B = \frac{X_A}{n_A} - \frac{X_B}{n_B}$, étant égale à $\frac{225}{300} - \frac{195}{300} = 0,10$, soit une *valeur supérieure au seuil critique* $\pi = 0,087$, c'est donc l'hypothèse H_1, qui au niveau de signification 1%, doit être retenue cette fois.

En d'autres termes, l'augmentation de la taille de l'échantillon conduit, pour α bloqué, à diminuer le risque de 2ème espèce et à fortiori à augmenter la puissance du test.

3.5 Tables de contingences (2,2) et échantillons indépendants

Le test d'indépendance de chi-deux permet entre autres de comparer une proportion entre deux échantillons et il équivaut, lorsque n est grand, au test paramétrique correspondant qui fait appel à la loi normale.

Enoncé : On reprend dans cette application, les données de l'application 3.4 précédente, le test de l'efficacité du sérum considéré ayant conduit à ne pas se prononcer du moins pour le test unilatéral et un niveau de signification égal à 5%.

Les données sont rappelées à travers le tableau (2,2) ci-dessous :

	Guéris	Non guéris	Total
Groupe A (sérum)	75	25	100
Groupe B (sans sérum)	65	35	100
Total	140	60	200

La question est celle de l'efficacité ou non du sérum, c'est-à-dire d'une différence ou non entre les deux groupes, ce qui est équivalent à affirmer que la guérison est indépendante de l'administration du sérum.

1°) Quelles conclusions faut-il retenir du test de chi-deux quant au problème considéré (on comparera à cet égard, ce que fournit l'application ou non de la correction de continuité de YATES) ?

2°) On considère sous leur forme générale, les tables de contingence (2,2) définies par :

	I	II	Total
Groupe I	A	B	$L_1 = A + B$
Groupe II	C	D	$L_2 = C + D$
Total	$T_1 = A + C$	$T_2 = B + D$	$N = A + B + C + D$

2-a) Montrer que la distance de chi-deux liée au test d'indépendance entre les données est égale à $\chi^2 = \dfrac{N.(A.D-B.C)^2}{(A+B).(B+C).(C+D).(A+C)}$ ou, avec la correction de continuité de YATES, $\chi^2_{corrigé} = \dfrac{N.(|A.D-B.C|^2 - N/2)}{(A+B).(B+C).(C+D).(A+C)}$.

2-b) Retrouver ainsi les résultats de la 1ère question.

3-a) Montrer que, s'agissant d'un test de comparaison entre deux proportions, le test de chi-deux est équivalent, lorsque N est grand, au test paramétrique sous modèle gaussien.

3-b) Retrouver ainsi, numériquement, les liens entre les calculs de la présente application et de l'application précédente 3.4.

Solution : 1°) *Tester l'indépendance entre la guérison et l'administration du sérum considéré*, c'est *rapprocher* les *effectifs constatés* des *effectifs théoriques* qui résultent de la relation d'indépendance suivant laquelle la *loi du couple est égale au produit des lois.*

Or comme le montre le tableau ci-dessous, les *lois marginales* peuvent être exprimées aisément ici et à fortiori les *effectifs théoriques sous l'hypothèse d'indépendance* (H_0).

Ces lois marginales sont :

Probabilités	Guéris	Non guéris	Total
Groupe A	0,375	0,125	0,5
Groupe B	0,325	0,175	0,5
Total	0,7	0,3	1

Pour les *effectifs théoriques*, on a par exemple et pour la paire « *guéris-groupe A* », une probabilité théorique égale au produit $0,5\times0,7$, c'est-à-dire un effectif théorique égal à $200\times0,35=70$. Plus généralement, il s'ensuit le tableau suivant :

	Guéris	Non guéris	Total
Groupe A	70	30	100
Groupe B	70	30	100
Total	140	60	200

Dans ces conditions, la distance de chi-deux est égale à :

$$\chi^2_{calculé} = \frac{(75-70)^2}{70} + \frac{(25-30)^2}{30} + \frac{(65-70)^2}{70} + \frac{(35-30)^2}{30}, \text{ soit } \chi^2_{calculé} = 2,38.$$

Mais, comme cela a déjà été mentionné en rappels de cours (cf. paragraphe 3.4.c) et dans les applications antérieures, χ^2 suit la **loi du chi-deux** à $\nu=(r-1).(k-1)$ degrés de liberté, soit pour $r=k=2$, la valeur $\nu=1$.

Raisonnant sur le **test unilatéral** $\begin{cases} H_0 : p_A = p_B \\ H_1 : p_A \succ p_B \end{cases}$, tel celui traité dans l'application 3.4 précédente, une lecture dans la table de valeurs annexée (cf. valeurs critiques du χ^2), fournit immédiatement, au niveau de signification $\alpha=0,05$, la valeur du **seuil critique** vérifiant $Prob(\chi^2 \geq \chi^2_\alpha) = \alpha$, soit $\chi^2_\alpha = 2,71$.

Le seuil précédent passe à la valeur $\chi^2_\alpha = 3,84$ lorsqu'on considère, au lieu du test unilatéral, le *test bilatéral* $\begin{cases} H_0 : p_A = p_B \\ H_1 : p_A \neq p_B \end{cases}$.

• En conclusion, l'inégalité $\chi^2_{calculé} = 2,38 \prec \chi^2_\alpha = 2,71$ permet de *conclure à l'impossibilité de rejeter* H_0, décision qui reste également à retenir si on applique à la distance de chi-deux , la *correction de continuité* de YATES, dont on sait qu'elle est plus conservatrice. En effet, avec cette dernière :

$$\chi^2_{calculé} = \frac{(|75-70|-0,5)^2}{70} + \frac{(|25-30|-0,5)^2}{30} + \frac{(|65-70|-0,5)^2}{70} + \frac{(|35-30|-0,5)^2}{30} = 1,93$$

soit, une *valeur plus faible* que celle obtenue sans correction.

2-a) Pour chacune des paires du tableau (2,2) proposé et en fonction des lois marginales, les *effectifs théoriques* sont immédiatement les suivants :

	I	II	Total
Groupe I	$L_1.T_1/N$	$L_1.T_2/N$	L_1
Groupe II	$L_2.T_1/N$	$L_2.T_2/N$	L_2
Total	T_1	T_2	N

Ainsi, la **distance de chi-deux** est-elle égale, sous sa forme générale, à :

$$\chi^2_{calculé} = \frac{(A - L_1.T_1/N)^2}{L_1.T_1/N} + \frac{(B - L_1.T_2/N)^2}{L_1.T_2/N} + \frac{(C - L_2.T_1/N)^2}{L_2.T_1/N} + \frac{(D - L_2.T_2/N)^2}{L_2.T_2/N}$$

Or, développant $A - L_1.T_1/N$, il vient $A - \frac{(A+B).(A+C)}{(A+B+C+D)} = \frac{A.D - B.C}{N}$, avec $N = A+B+C+D$. De même, pour les trois autres termes, au signe près tel par exemple, $B - L_1.T_2/N = B - \frac{(A+B).(B+D)}{A+B+C+D} = \frac{B.C - A.D}{N}$. Bref, après mise en facteurs, on obtient :

$$\chi^2_{calculé} = \frac{(A.D-B.C)^2}{N^2}.\left[\frac{N}{L_1.T_1} + \frac{N}{L_1.T_2} + \frac{N}{L_2.T_1} + \frac{N}{L_2.T_2}\right], \text{ soit;}$$

$$\chi^2_{calculé} = \frac{(A.D-B.C)^2}{N.L_1.L_2.T_1.T_2}.\left[L_1.T_1 + L_1.T_2 + L_2.T_1 + L_2.T_2\right] = \frac{1}{N}.(A.D-B.C)^2.\frac{1}{L_1.L_2.T_1.T_2}.(L_1+L_2).(T_1+T_2).$$

Comme $L_1 + L_2 = T_1 + T_2 = N$, on a donc après simplifications, le résultat attendu, à savoir $\chi^2_{calculé} = \frac{N.(A.D-B.C)^2}{L_1.T_1.L_2.T_2}$.

• Si on applique la correction de continuité de YATES, chacun des termes est remplacé, à l'instar de $(A - L_1.T_1/N)^2$ par $\frac{(|A - L_1.T_1/N| - 0,5)^2}{L_1.T_1/N} = (\frac{|A.D-B.C|}{N} - 0,5)^2$, c'est-à-dire, $(\frac{|A.D-B.C| - 0,5.N}{N})^2$.

Par analogie avec les calculs antérieurs, le résultat $\chi^2_{calculé} = \dfrac{N.(|A.D - B.C| - 0,5.N)^2}{L_1.T_1.L_2.T_2}$ est manifeste.

2-b) L'application des formules précédentes, conduit numériquement, aux valeurs $\chi^2_{calculé} = \dfrac{200.(75\times35 - 65\times25)^2}{60\times140\times100\times100} = 2,38$ et, suivant correction de continuité, à la valeur, $\chi^2_{calculé} = \dfrac{200.(|75\times35 - 65\times25| - 100)^2}{60\times140\times100\times100} = 1,93$. *Les résultats antérieurs sont ainsi retrouvés.*

3-a) Procédant par analogie entre le présent test d'indépendance et le test paramétrique de comparaison de proportions traité dans l'application antérieure 3.4, on a en désignant par F_A et F_B les taux de guérison constatés pour chacun des groupes «*traité*» et « *non traité* » et en notant par n_A et n_B les tailles des effectifs desdits groupes, $A = F_A.L_1$, $B = (1-F_A).L_1$, $C = F_B.L_2$, et $D = (1-F_B).L_2$, avec en outre $L_1 = n_A$ et $L_2 = n_B$.

D'autre part, sous l'hypothèse H_0 de l'égalité des probabilités de guérison entre les deux groupes ($p_A = p_B = p$), on a $T_1 = p.N$ et $T_2 = (1-p).N$. Ainsi, compte tenu de l'expression de $\chi^2_{calculé}$ obtenue dans la question antérieure a-t-on :

$$\chi^2_{calculé} = N.\frac{\left[F_A.(1-F_B).L_1.L_2 - F_B.(1-F_A).L_1.L_2\right]^2}{L_1.L_2.T_1.T_2} = \frac{N.L_1.L_2.(F_A-F_B)^2}{T_1.T_2} = \frac{L_1.L_2.(F_A-F_B)^2}{N.p.(1-p)}.$$

Mais, $N = L_1 + L_2 \Rightarrow \dfrac{L_1.L_2}{N} = \dfrac{1}{1/L_1 + 1/L_2}$. On obtient donc en définitive, en tenant compte des relations $F_1 = n_A$ et $F_2 = n_B$, l'expression $\chi^2_{calculé} = \dfrac{(F_A-F_B)^2}{p.(1-p).(1/n_A + 1/n_B)}$.

Ce résultat coïncide avec le *carré de la statistique* $\dfrac{F_A - F_B}{\sqrt{p.(1-p)}.\sqrt{1/n_A + 1/n_B}}$ qui est *celle utilisée dans la mise en œuvre du test paramétrique de comparaison* entre deux proportions.

3-b) Ainsi, pour le test paramétrique de l'application 3.4 antérieure, on avait $F_A - F_B = 0,10; p = 0,70, n_A = n_B = 100 \Rightarrow \dfrac{F_A - F_B}{\sqrt{p.(1-p}.\sqrt{1/n_A + 1/n_B}} = 1,543$. Or, le *carré* de cet indicateur, vaut $2,38$, et c'est justement *la valeur fournie par la distance de chi-deux* dans la 1ère question.

- Plus précisément, **les deux tests** (*chi-deux et test paramérique sous loi normale*) *conduisent rigoureusement aux mêmes seuils* du moins dès que le théorème de MOIVRE LAPLACE est applicable (n est assez grand).

On remarque en effet qui si V suit la *loi du chi-deux à un degré de liberté*, la variable $Z = \sqrt{V}$ suit alors *la loi normale centrée réduite* $N(0,1)$.

En effet, par définition V a pour densité de probabilité $g(v)=\frac{1}{\sqrt{2v}.\Gamma(1/2)}.e^{-v/2}.1_{v\geq 0}(v)$ (cf. rappels de cours du chapitre I, paragraphe 1.1), $\Gamma(1/2)$ étant égal à $(-1/2)!=\sqrt{\pi}$. On a donc, élémentairement, $g(v).dv=\frac{1}{\sqrt{2.\pi.v}}.e^{-v/2}.dv,(\ v\geq 0\ )$.

Désignant par $f(z)$ la densité de probabilité de la variable aléatoire $Z=\sqrt{V}$, il faut se méfier toutefois ici que l'image inverse d'un ouvert $]0,a[$ est ici $]-\sqrt{a},0[\cup]0,+\sqrt{a}[$ et non pas $]0,\sqrt{a}[$. Autrement dit, le *théorème de la mesure image* fournit pour le cas présent, le résultat $f(\sqrt{v})+f(-\sqrt{v})$, ce qui revient à diviser par deux le résultat obtenu suite au changement de variable dans la densité éléméntaire susmentionnée.

Finalement, posant $v=z^2$ et après avoir appliqué la division ci-dessus, Z a pour densité de probabilité élémentaire $f(z).dz=\frac{1}{2}.\frac{2z}{z.\sqrt{2\pi}}.e^{-\frac{z^2}{2}}.dz\Rightarrow f(z)=\frac{1}{\sqrt{2\pi}}.e^{-\frac{z^2}{2}}$. Il s'agit bien de la densité de probabilité de la variable normale centrée réduite $N(0,1)$.

• Concrètement, et pour le **test unilatéral**, le **seuil critique** au niveau de signification $\alpha=0,05$ fourni par la **loi du chi-deux** est $\chi^2_\alpha=2,71$ (cf. 1ère question). Or, cette valeur est aussi $(1,645)^2$, 1,645 étant pour la **loi normale** $\xi:N(0,1)$, le seuil critique unilatéral t_α / $\Pr ob(\xi\geq t_\alpha)=0,05$. On retrouve donc ainsi la relation annoncée $Z=\sqrt{V}$.

De même, pour le **test bilatéral**, le test de chi-deux conduit au seuil $\chi^2_\alpha=3,84$ (cf. 1ère question). Quant à la loi normale, elle entraîne pour $\alpha=0,05$, la valeur $t_\alpha=1,96$, t_α vérifiant $\Pr ob(|\xi|\geq t_\alpha)=0,05$. Or, on a précisément $(1,96)^2=3,84$.

3.6 Corrélation entre taille et poids (coefficient « r » de BRAVAIS PEARSON)

Portant sur des variables quantitatives, le coefficient de corrélation linéaire de PEARSON est couramment utilisé. Mais en tant qu'indicateur de l'écart entre les données observées et les données théoriques qui correspondent à la droite de régression, il n'a de sens que pour les relations linéaires.

En outre, les distributions d'échantillonnage dudit coefficient conduisent à supposer que les variables étudiées sont distribuées normalement et même de loi conjointe binormale, ce qui est le cas en l'occurrence pour le couple (taille, poids) étudié ci-après, les observations étant, par ailleurs, supposées indépendantes (pas d'autocorrélation).

Dans ces conditions, le test de nullité du coefficient de corrélation équivant à un test paramétrique d'indépendance.

Enoncé : Un questionnaire effectué en début d'année sur une population de 128 étudiants permet d'établir le tableau de contingence ci-dessous où P et T désignent respectivement le poids en kg et la taille en cm.

Les données sont rassemblées dans le tableau présenté ci-après, ceci après avoir été regroupées par classes, les effectifs observés correspondants étant portés dans ledit tableau.

Poids P (en kg)	Taille T (en cm)				
	$[150-160[$	$[160-170[$	$[170-180[$	$[180-190[$	$[190-200[$
$[40-50[$	9	5	1	0	0
$[50-60[$	4	34	9	0	0
$[60-70[$	0	5	20	10	0
$[70-80[$	0	3	9	12	0
$[80-90[$	0	0	0	6	1

1°) Calculer le coefficient de corrélation linéaire entre P et T.

2°) Que dire, au niveau de signification $\alpha = 5\%$, de l'indépendance ou non des variables aléatoires P et T ?

3°) Expliciter un intervalle de confiance du coefficient de corrélation linéaire en question, ceci au seuil 95%.

4°) On souhaite tester le degré de cette corrélation à travers le test unilatéral $\begin{cases} H_0 : r = 0,7 \\ H_1 : r \succ 0,7 \end{cases}$

Qu'en conclure au niveau de signification 5% pour les données proposées ci-dessus.

Solution : 1°) *Ramenant chaque classe à sa valeur centrale*, il est proposé en outre de passer par les **variables auxiliaires** $U = \dfrac{P-65}{10}$ et $V = \dfrac{T-175}{10}$ pour alléger les calculs.

Il en résulte le tableau « *simplifié* » (U,V) ci-dessous :

$U \backslash V$	-2	-1	0	1	2
-2	9	5	1	0	0
-1	4	34	9	0	0
0	0	5	20	10	0
1	0	3	9	12	0
2	0	0	0	6	1

Notant par $f_{i.}, f_{.j}, f_{ij}$ les efectifs respectivement associés aux valeurs u_i, v_j, et (u_i, v_j), et par N l'effectif total $\sum_i \sum_j f_{ij}$, on obtient les *estimations* suivantes :

Paramètre	Estimation	Valeur
$E(U)$	$\bar{u} = \dfrac{1}{N} \cdot \sum_i u_i . f_{i.}$	–0,305
$E(V)$	$\bar{v} = \dfrac{1}{N} \cdot \sum_i v_j . f_{.j}$	-0,336
$Var(U)$	$\dfrac{1}{N-1} \cdot \left[\sum_i u_i^2 . f_{i.} - N.\bar{u}^2 \right]$	1,158
$Var(V)$	$\dfrac{1}{N-1} \cdot \left[\sum_j v_j^2 . f_{.j} - N.\bar{v}^2 \right]$	0,918
$Cov(U,V)$	$\dfrac{1}{N-1} \cdot \left[\sum_i \sum_j u_i . v_j . f_{ij} - N.\bar{u}.\bar{v} \right]$	0,787

Or il est rappelé que pour tout couple $(a.X+b,c.Y+d)$ où X et Y sont aléatoires et a,b,c,d réels, on a d'une part $Var(a.X+b)=a^2.Var(X), Var(c.Y+d)=c^2.Var(Y)$, et d'autre part $Cov(a.X+b,c.Y+d)=a.c.Cov(X,Y)$. Ainsi, obtient-on pour ce qui est du *coefficient de corrélation linéaire* :

$$r_{P,T}=\frac{\text{cov}(P,T)}{\sqrt{Var(P).Var(T)}}=\frac{100.\text{cov}(U,V)}{\sqrt{100.Var(U).100.Var(V)}}=r_{U,V}$$

Numériquement, on a $r_{P,T}=0,76$ ($r_{P,T}=0,7629$ très précisément).

2°) Compte tenu de l'hypothèse de **binormalité** du couple (P,T), *le test de l'indépendance* de ces deux variables se ramène ici au *test de nullité ou non* de leur coefficient de corrélation linéaire, soit $\begin{cases}H_0 : r=0\\ H_1 : r\neq 0\end{cases}$.

Mais sous l'hypothèse H_0, la variable $T=\frac{r.\sqrt{n-2}}{\sqrt{1-r^2}}$ suit la **loi de STUDENT** à $\nu=n-2$ degrés de liberté, loi qui lorsque n *est grand*, s'apparente à la **loi normale** (cf. rappels de cours du présent chapitre, paragraphe 2.3.d).

Pour le cas présent, $n=128$ et $\alpha=0,05$ ce qui, pour un **test bilatéral**, entraîne la valeur $t_\alpha=1,96$ (en réalité $t_\alpha=1,979$ si on ne recoure pas à l'approximation normale), et le **seuil critique** $\pi=\frac{t_\alpha}{\sqrt{n-2+t_\alpha^2}}=0,1736$ (pour rappel, t_α est solution de l'équation $Prob(|T|\geq t_\alpha)=\alpha$). En conclusion, on rejette l'hypothèse d'indépendance H_0 si, pour le coefficient $r_{P,T}$ calculé à partir des données, on a $|r_{P,T}|\geq 0,174$ ce qui est le cas présentement puisque $r_{P,T}=0,76\geq 0,174$.

- Une autre méthode est celle de la **transformation de FISHER**, soit $z=Argthr$, ou encore, $z=\frac{1}{2}.\ln\frac{1+r}{1-r}$ (cf. rappels de cours). On montre que sous l'hypothèse nulle $H_0 : r=0$, z converge vers la *loi normale* de moyenne 0 et de variance $\frac{1}{n-3}$.

Ramenant le test en question à celui de la *nullité ou non* de la moyenne $E(z)$, on est conduit ainsi à la **région critique** $|z|\geq\frac{t_\alpha}{\sqrt{n-3}}$ avec $t_\alpha=1,96$, soit numériquement $|z|\geq 0,1735$. C'est un *résultat très proche du seuil critique obtenu par la méthode antérieure.*

3°) La *transformée de FISHER* permet également d'obtenir aisément un **intervalle de confiance** pour r. A partir de la valeur calculée $r_{P,T}$ (estimation ponctuelle r_0 de r), on a, à partir de l'estimation ponctuelle $z_0=\frac{1}{2}.\ln\frac{1+r_0}{1-r_0}$, *l'estimation par intervalle de confiance* $\left[z_0-\frac{t_\alpha}{\sqrt{n-3}}, z_0+\frac{t_\alpha}{\sqrt{n-3}}\right]$ de z. Par transformation inverse, résulte l'encadrement cherché pour r.

Concrètement, le nombre $t_\alpha / \Pr ob(|\xi| \le t_\alpha) = \alpha$ est égal à $t_\alpha = 1,96$ lorsque $\alpha = 0,95$. Par ailleurs, $r_0 = 0,76 \Rightarrow z_0 = 0,996$. Il s'ensuit, pour z, l'intervalle de confiance $[0,821 - 1,171]$. Suivant la *transformation de FISHER inverse*, $r = thz$, on obtient pour r l'intervalle de confiance $[0,676 - 0,825]$.

4°) Cette fois, le test est **unilatéral** de type $\begin{cases} r = r_0 = 0,7 \\ r \succ r_0 = 0,7 \end{cases}$. Utilisant de nouveau la **méthode de la transformation de FISHER**, on est confronté relativement à z, à un test $\begin{cases} H_0 : z = z_0 \\ H_1 : z \succ z_0 \end{cases}$ où $z_0 = \frac{1}{2}.\ln\frac{1+r_0}{1-r_0}$, soit numériquement $z_0 = 0,867$.

Se référant au *test unilatéral de conformité d'une moyenne* (cf. paragraphe 2.3.a), la **région critique** du test a pour expression $z \ge z_0 + t_\alpha . \frac{1}{\sqrt{n-3}}$ où t_α vérifie $\Pr ob(\xi \ge t_\alpha) = \alpha$, soit $t_\alpha = 1,645$ lorsque $\alpha = 0,05$.

Ainsi a-t-on pour z, la région critique $z \ge \pi = 1,014$. Par transformation inverse $r = thz$, *la région critique du test qui porte sur* r, est donc $r \ge th(1,014) = 0,768$. Or, sur la base des données fournies $r = 0,763$. On ne peut donc pas en l'occurrence retenir l'hypothèse $r \succ 0,7$ puisqu'on n'a pas $r_{calculé} \ge \pi = 0,768$.

4. Tests à deux échantillons sous autres modèles

4.1 Test paramétrique de comparaison de moyennes sous modèle exponentiel

Enoncé : On considère deux variables aléatoires exponentielles, soient X et Y, de paramètres respectifs λ et μ. On se propose de construire un test paramétrique de comparaison entre les moyennes $1/\lambda$ et $1/\mu$ en se ramenant à des distributions d'échantillonnage connues.

1°) Montrer que la variable aléatoire $T = 2.\lambda.X$ suit une loi de chi-deux à deux degrés de liberté.

2°) $(X_1, X_2, ..., X_n)$ étant un échantillon de n variables aléatoires indépendantes de loi parente X, caractériser la loi de la statistique $Z_n = 2.\lambda.(X_1 + X_2 + ... + X_n)$.

3°) Soit $(Y_1, Y_2, ..., Y_p)$ un autre échantillon de p variables aléatoires indépendantes de loi parente Y. Caractériser la loi de la variable $U(n,p) = \frac{p.\lambda.(X_1 + X_2 + ... + X_n)}{n.\mu.(Y_1 + Y_2 + ... + Y_p)}$.

4°) Construire à partir des résultats précédents, un test paramétrique de comparaison entre $H_0 : \lambda = \mu$ et $H_1 : \lambda \ne \mu$.

5°) Appliquer le test précédent aux deux séries de données :

X	0,633	0,161	0,100	1,197	0,152	0,182	0,418	0.192	0,029	0,885	0,278	0,008

Y	0,361	0,085	0,293	0,080	0,077	0,020	0,036	0,095	0,197	0,206

Solution : 1°) Le changement de variable $t = 2.\lambda.x$ dans la densité de probabilité élémentaire de la variable exponentielle X, soit $f(x).dx = \lambda.e^{-\lambda.x}.dx$, conduit immédiatement pour la variable T, à la densité de probabilité élémentaire $h(t).dt = \lambda.e^{-\frac{t}{2}}.\frac{dt}{2.\lambda} = \frac{e^{-t}}{2}.dt$. Or $h(t) = \frac{e^{-\frac{t}{2}}}{2}$ coïncide avec la densité de probabilité de la loi $\chi^2(2)$ (cf. rappels de cours du chapitre I, paragraphe 1.1).

2°) Suivant les résultats de l'application 1.1 du chapitre I, la *somme des variables de chi-deux indépendantes*, soient $\chi^2(n_1)$ et $\chi^2(n_2)$, est elle-même une *loi de type chi-deux* à $n_1 + n_2$ degrés de liberté, soit $\chi^2(n_1 + n_2)$.

Il en résulte donc immédiatement, pour ce qui est de la statistique $Z = \sum_{i=1}^{i=n}(2.\lambda.X_i)$, une loi de type chi-deux à $\sum_{i=1}^{i=n} 2 = 2n$ degrés de liberté.

3°) Toujours d'après les rappels de cours du chapitre I (paragraphe 1.1), lorsque X et Y suivent respectivement les lois $\chi^2(n)$ et $\chi^2(p)$, la *statistique quotient* $F = \frac{X/n}{Y/p}$ suit la **loi de FISHER SNEDECOR** à n et p degrés de liberté, soit $F(n,p)$. Pour le cas présent, $2.\lambda.\sum_{i=1}^{i=n} X_i$ et $2.\mu.\sum_{i=1}^{i=p} Y_i$ suivent respectivement les lois $\chi^2(2n)$ et $\chi^2(2p)$. Il en ressort, pour le quotient $U(n,p) = \frac{2.\lambda.(X_1 + X_2 + ... + X_n)/2n}{2.\mu.(Y_1 + Y_2 + ... + Y_p)/2p}$, la loi de FISHER SNEDECOR, $F(2n, 2p)$.

Après simplifications, $U(n,p)$ s'écrit $\frac{p.\lambda.\sum_{i=1}^{i=n} X_i}{n.\mu.\sum_{i=1}^{i=p} Y_i}$, ou encore par introduction des moyennes empiriques, $U(n,p) = \frac{\lambda.\overline{X}}{\mu.\overline{Y}}$.

4°) Relativement au test $\begin{cases} H_0 : \lambda = \mu \\ H_1 : \lambda \neq \mu \end{cases}$, le recours fort logique à la statistique $\frac{\overline{X}}{\overline{Y}}$ et au rapport $U(n,p)$, conduit pour le cas d'un *test bilatéral*, à **la région critique** caractérisée par $W = \left\{ (x_1, x_2, ..., x_n), (y_1, y_2, ..., y_p) / \frac{\overline{x}}{\overline{y}} \notin]\pi_1, \pi_2[\right\}$.

Mais, *sous l'hypothèse* H_0, la statistique $\frac{\overline{X}}{\overline{Y}}$ qui coïncide avec le rapport $U(n,p)$ (puisque $\lambda = \mu$), suit la loi $F(2n, 2p)$.

Notant par F_1 et F_2 les seuils vérifiant respectivement $\Pr ob(F \le F_1) = \alpha/2$ et $\Pr ob(F \ge F_2) = \alpha/2$, il vient immédiatement, au *niveau de signification* α, la **règle de décision** suivante :

- si $\frac{\bar{x}}{\bar{y}} \notin]F_1, F_2[$, on décide H_1 ;
- si $\frac{\bar{x}}{\bar{y}} \in]F_1, F_2[$, on décide H_0.

5°) Pour les données proposées, $n = 12, p = 10, \bar{x} = 0,353, \bar{y} = 0,145$. D'autre part, par lecture dans la table de valeurs correspondante annexée et pour la loi $F(24,20)$, on a lorsque $\alpha = 5\%$, $\Pr ob(F \ge F_2) = 0,025 \Rightarrow \Pr ob(F \prec F_2) = 0,975 \Rightarrow F_2 = 2,41$.

Par ailleurs, $\Pr ob(F \le F_1) = 0,025 \Rightarrow \Pr ob(\frac{1}{F} \ge \frac{1}{F_1}) = 0,025$. Comme la variable $\frac{1}{F}$ suit manifestement la loi $F(20,24)$, il s'ensuit après lecture dans la table de valeurs susmentionnée $\frac{1}{F_1} = 2,33$ soit $F_1 = 0,429$.

Pour l'exemple proposé, on a $\frac{\bar{x}}{\bar{y}} = 2,43$, *valeur qui est située à l'extérieur de l'intervalle* $]0,429 - 2,410[$. C'est donc le **rejet de l'hypothèse** $H_0 : \lambda = \mu$ qu'on est conduit à retenir ici.

4.2 Comparaison du test de WILCOXON avec le test paramétrique du rapport des vraisemblances pour deux échantillons exponentiels

Enoncé : On considère deux échantillons (X_1, X_2) et (Y_1, Y_2, Y_3) indépendants, de tailles $n_X = 2$ et $n_Y = 3$ et de lois parentes de type exponentiel et de paramètres respectifs λ (pour ce qui est de X) et μ (pour ce qui est de Y). On souhaite tester l'hypothèse $H_0 : \lambda = \mu$ contre l'hypothèse $H_1 : \lambda \succ \mu$, ceci au niveau de signification $\alpha = 10\%$.

1°) « *Construction d'un test paramétrique* »

1-a) Reprenant les résultats de l'application précédente 4.1 relative à la construction d'un test paramétrique de comparaison de moyennes sous modèle exponentiel, montrer que la région critique du test est $\frac{\bar{Y}}{\bar{X}} \ge F_{0.90,6,4}$ où $F_{0.90,6,4}$ est le $90^{\text{ème}}$ percentile de la fonction de répartition de la variable de FISHER SNEDECOR à $\nu_1 = 6$ et $\nu_2 = 4$ degrés de liberté.

1-b) Exprimer la puissance du test suivant l'hypothèse alternative $H_1 : \lambda = a.\mu$, pour $a \succ 1$.

1-c) Exprimer numériquement cette puissance pour diverses valeurs de $a \ge 1$, variant de 1 à 16.

2°) « *Test de MANN-WHITNEY-WILCOXON* »

2-a) Expliciter la région de rejet de l'hypothèse $H_0 : \lambda = \mu$, lorsque pour traiter le test unilatéral précédent, on utilise la statistique de WILCOXON.

2-b) Montrer que si Z_1, Z_2, Z_3, Z_4 sont des variables aléatoires exponentielles indépendantes de paramètres respectifs $\lambda_1, \lambda_2, \lambda_3, \lambda_4$, on a :

$$\Pr ob(Z_1 \prec Z_2 \prec Z_3 \prec Z_4) = \frac{\lambda_1}{\lambda_1+\lambda_2+\lambda_3+\lambda_4} \cdot \frac{\lambda_2}{\lambda_2+\lambda_3+\lambda_4} \cdot \frac{\lambda_3}{\lambda_3+\lambda_4}.$$

En déduire, de façon plus générale, la relation :

$$\Pr ob(Z_1 \prec Z_2 \prec \ldots \prec Z_l) = \frac{\lambda_1}{\lambda_1+\lambda_2+\ldots+\lambda_l} \cdot \frac{\lambda_2}{\lambda_2+\lambda_3+\ldots+\lambda_l} \ldots \frac{\lambda_{l-1}}{\lambda_{l-1}+\lambda_l}$$

2-c) Expliciter en fonction de a la puissance du test de WILCOXON pour le test considéré dans ce problème, c'est-à-dire à partir des échantillons (X_1, X_2) et (Y_1, Y_2, Y_3).

2-d) Qu'en conclure ?

Solution : 1°) Lorsque les variables Z_i sont *exponentielles et indépendantes* de paramètre θ, la somme $2.\theta.\sum_{i=1}^{i=n} Z_i$ suit la *loi du chi- deux* à $2n$ degrés de liberté, soit $\chi^2(2n)$. Ainsi, les statistiques $2.\lambda.\sum_{i=1}^{i=2} X_i$ et $2.\mu.\sum_{i=1}^{i=3} Y_i$ suivent-elles pour le présent cas, les lois respectives $\chi^2(4)$ et $\chi^2(6)$.

Dans ces conditions, la statistique $\dfrac{2.\lambda.\sum_{i=1}^{i=2} X_i / 4}{2.\mu.\sum_{i=1}^{i=3} Y_i / 6}$ suit la **loi de FISHER SNEDECOR** à 4 et 6 degrés de liberté, soit $F(4,6)$. Cette statistique, s'écrit aussi $\dfrac{\lambda.\overline{X}}{\mu.\overline{Y}}$ en introduisant les moyennes empiriques (cf. application 4.1 précédente).

- Pour le **test unilatéral** $\begin{cases} H_0 : \lambda = \mu \\ H_1 : \lambda \succ \mu \end{cases}$, il faut néanmoins se méfier que $E(\overline{X}) = \dfrac{1}{\lambda}$ et $E(\overline{Y}) = \dfrac{1}{\mu}$. *En raison de ces relations inversement proportionnelles*, on *rejette l'hypothèse* H_0 lorsque $\dfrac{\overline{Y}}{\overline{X}}$ *est grand* (et non pas $\dfrac{\overline{X}}{\overline{Y}}$!). Une autre façon de constater ce résultat est de comparer pour $\lambda \succ \mu$, les fonctions de répartition des variables aléatoires X et Y, soient $F(x) = 1 - e^{-\lambda.x}$ et $G(x) = 1 - e^{-\mu.x}$. On montre aisément que $G(x) \succ F(x)$, $\forall x \in R^+$ (on dit que G est **stochastiquement plus grand** que F).

En définitive, la **région critique** du test unilatéral considéré, est donc, ici, $\dfrac{\overline{Y}}{\overline{X}} \geq F_\alpha$ où F_α vérifie, sous l'hypothèse nulle $H_0 : \lambda = \mu$, l'équation $\Pr ob(\dfrac{1}{F} = \dfrac{\overline{Y}}{\overline{X}} \geq F_\alpha) = 0{,}10$. Mais, quand F suit la loi $F(4,6)$, la variable $\dfrac{1}{F}$ suit la loi $F(6,4)$. Ainsi, la règle de décision du test est-elle fixée par l'équation $\Pr ob(F(6,4) = \dfrac{\overline{Y}}{\overline{X}} \geq F_\alpha) = 0{,}10$.

F_α qui, pour la loi $F(6,4)$ est caractérisé par la relation $\Pr ob(F_{6,4} \prec F_\alpha) = 0{,}90$, en constitue donc le 90ème percentile, soit $F_{0.90,6,4}$, ce qui montre le résultat annoncé.

1-b) La **puissance** η du test est égale à $1-\beta$ où $\beta = \Pr ob(\frac{\overline{Y}}{\overline{X}} \prec F_\alpha / \lambda = a.\mu)$. Or, la statistique $\frac{2.\lambda.\overline{X}}{2.\mu.\overline{Y}}$ suit la loi de FISHER SNEDECOR, $F(4,6)$, statistique qui, sous l'hypothèse H_1 est aussi égale à $a.\frac{\overline{X}}{\overline{Y}}$ (puisque $\lambda = a.\mu$).

En conclusion, $\eta = 1-\beta = \Pr ob(\frac{\overline{Y}}{\overline{X}} \succeq F_\alpha / \lambda = a.\mu)$, soit en inversant les deux membres de l'inégalité et en multipliant de part et d'autre par a, l'expression ci-après de la puissance, $\eta = \Pr ob(a.\frac{\overline{X}}{\overline{Y}} = F(4,6) \leq \frac{a}{F_\alpha})$, avec $F_\alpha = F_{0.90,6,4}$.

1-c) Le recours à un *calculateur en ligne* des valeurs de la loi de FISHER SNEDECOR (cf. site geai.univ-brest.fr) conduit pour $\alpha = 0{,}10; \nu_1 = 6; \nu_2 = 4$ au **seuil critique** $F_{0.90,6,4} = 4{,}01$.

Pour diverses valeurs de a allant de 1 à 16 et en procédant par *approximations successives*, le calculateur fournit, pour η, les valeurs approchées ci-dessous :

a	1	2	3	4	5	6	7	8
$a/F_{0.90,6,4}$	0,25	0,50	0,75	1,00	1,25	1,50	1,75	2,00
η	0,10	0,26	0,41	0,53	0,62	0,69	0,74	0,78
a	9	10	11	12	13	14	15	16
$a/F_{0.90,6,4}$	2,25	2,50	2,75	3,00	3,25	3,50	3,75	4,00
η	0,82	0,85	0,87	0,89	0,90	0,92	0,93	0,94

2-a) La **statistique *W* de WILCOXON** pour le problème posé $\begin{cases} H_0 : \lambda = \mu \\ H_1 : \lambda \succ \mu \end{cases}$ à partir d'échantillons de tailles respectives $n_X = 2$ et $n_Y = 3$, est égale à la *somme des rangs* des X_i dans l'échantillon (X_i, Y_i) *regroupé et trié par ordre croissant.*

La **région critique** du test est définie par $W \geq W_\alpha$ où le seuil critique, W_α, qui vérifie, sous l'hypothèse H_0, $\Pr ob(W \geq W_\alpha) = \alpha$, peut être calculé directement (cf. application 4.5 du présent chapitre), soit à l'aide de tables de valeurs, soit par calculateur en ligne.

2-b) La densité de probabilité élémentaire du n-uplet (Z_1, Z_2, Z_3, Z_4), avec $n = 4$, s'écrit immédiatement $\lambda_1.\lambda_2.\lambda_3.\lambda_4.e^{-\lambda_1.z_1}.e^{-\lambda_2.z_2}.e^{-\lambda_2.z_3}.e^{-\lambda_4.z_4}.dz_1.dz_2.dz_3.dz_4$. Le changement de variables $U_1 = Z_1, U_2 = Z_2 - Z_1, U_3 = Z_3 - Z_2, U_4 = Z_4 - Z_3$ conduit à la nouvelle densité de probabilité : $\lambda_1.\lambda_2.\lambda_3.\lambda_4.e^{-\lambda_1.u_1}.e^{-\lambda_2.(u_1+u_2)}.e^{-\lambda_2.(u_1+u_2+u_3)}.e^{-\lambda_4.(u_1+u_2+u_3+u_4)}.du_1.du_2.du_3.du_4$, cette transformation ayant manifestement la valeur 1 comme *jacobien*.

Dans ces conditions, la probabilité $\Pr ob(Z_1 \prec Z_2 \prec Z_3 \prec Z_4)$ qui est aussi égale à $\Pr ob(U_1 \succ 0, U_2 \succ 0, U_3 \succ 0, U_4 \succ 0)$ est donc fournie par l'intégrale :

$$\lambda_1.\lambda_2.\lambda_3.\lambda_4.\int\int\int\int_{R^{+4}} e^{-(\lambda_1+\lambda_2+\lambda_3+\lambda_4).u_1}.e^{-(\lambda_2+\lambda_3+\lambda_4).u_2}.e^{-(\lambda_3+\lambda_4).u_3}.e^{-\lambda_4.u_4}.du_1.du_2.du_3.du_4 \ ;$$

intégrale dont le théorème de FUBINI permet d'écrire l'expression sous la forme du produit des intégrales simples :

$$\lambda_1.\lambda_2.\lambda_3.\lambda_4.(\int_0^{+\infty} e^{-(\lambda_1+\lambda_2+\lambda_3+\lambda_4).u_1}.du_1).(\int_0^{+\infty} e^{-(\lambda_2+\lambda_3+\lambda_4).u_2}.du_2).(\int_0^{+\infty} e^{-(\lambda_3+\lambda_4).u_3}.du_3).(\int_0^{+\infty} e^{-\lambda_4.u_4}.du_4).$$

Il en résulte immédiatement, après calcul des intégrales simples en question, le résultat cherché $\Pr ob(Z_1 \prec Z_2 \prec Z_3 \prec Z_4) = \frac{\lambda_1}{\lambda_1+\lambda_2+\lambda_3+\lambda_4}.\frac{\lambda_2}{\lambda_2+\lambda_3+\lambda_4}.\frac{\lambda_3}{\lambda_3+\lambda_4}$.

- De façon plus générale et par exemple, par récurrence, on obtient la relation $\Pr ob(Z_1 \prec Z_2 \prec ... \prec Z_l) = \frac{\lambda_1}{\lambda_1+\lambda_2+...+\lambda_l}.\frac{\lambda_2}{\lambda_2+\lambda_3+...+\lambda_l}....\frac{\lambda_{l-1}}{\lambda_{l-1}+\lambda_l}$.

2-c) Pour le test *unilatéral* proposé $\begin{cases} H_0 : \lambda = \mu \\ H_1 : \lambda \succ \mu \end{cases}$, le **test non paramétrique de MANN-WHITNEY-WILCOXON** fait appel à la *somme des rangs des valeurs* X_i dans *l'échantillon* $(X_1, X_2, Y_1, Y_2, Y_3)$ *classé par ordre croissant*, soit W, la statistique ainsi définie.

La **région critique** du test a pour forme $W \leq W_\alpha$ où W_α est caractérisé à partir de l'erreur de première espèce α, par l'équation $\Pr ob(W \leq W_\alpha / \lambda = \mu) = \alpha$. Il ne faut pas oublier en effet, le résultat précédent (cf. 1-a) suivant lequel $\lambda \succ \mu$ entraîne pour les fonctions de répartition $F(x)$ et $G(y)$, l'inégalité $F(x) \prec G(y)$.

Dans le cas de *petits échantillons*, un *calcul direct* ou *l'usage des tables de valeurs*, ou *l'usage d'un calculateur sur le net*, permet la détermination du seuil critique W_α. Ainsi, suivant calculateur et pour les données $\alpha = 10\%, n_X = 2, n_Y = 3$ a-t-on, par exemple, $W_\alpha = 3$.

- Quant à la **puissance du test**, c'est $\eta = 1 - \beta$ avec $\beta = \Pr ob(W \succ W_\alpha / \lambda = a.\mu)$, puisque sous l'hypothèse $H_1 : \lambda = a.\mu$ avec $a \succ 1$. On a donc $\eta = \Pr ob(W \leq W_\alpha / \lambda = a.\mu)$.

Les cas de permutations possibles entre X_1, X_2, Y_1, Y_2, Y_3 pour lesquels la somme des rangs des X_i est inférieure ou égale à $W_\alpha = 3$, correspondent aux permutations pour lesquelles les deux premiers rangs sont occupés par les X_i et le reste par les Y_i, soient des « *5-uplets* » vérifiant, par exemple, $X_1 \prec X_2 \prec Y_1 \prec Y_2 \prec Y_3$.

Ces cas sont en nombre égal à $2!.3! = 12$, la valeur commune des probabilités étant déterminée, d'après la question précédente appliquée à $\Pr ob(X_1 \prec X_2 \prec Y_1 \prec Y_2 \prec Y_3)$, par l'expression $\frac{\lambda}{\lambda+\lambda+\mu+\mu+\mu}.\frac{\lambda}{\lambda+\mu+\mu+\mu}.\frac{\mu}{\mu+\mu+\mu}.\frac{\mu}{\mu+\mu}$. Or, sous l'hypothèse H_1, $\lambda = a.\mu$, ce qui pour la probabilité en question induit immédiatement le résultat ci-après.

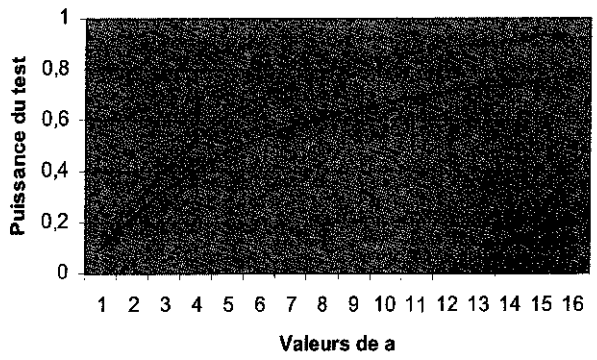
On a $Prob(X_1 \prec X_2 \prec Y_1 \prec Y_2 \prec Y_3) = \frac{a.\mu}{(2.a+3).\mu} \cdot \frac{a.\mu}{(a+3).\mu} \cdot \frac{\mu}{3.\mu} \cdot \frac{\mu}{2.\mu}$, soit après réduction, $Prob(X_1 \prec X_2 \prec Y_1 \prec Y_2 \prec Y_3) = \frac{a^2}{6.(a+3).(2.a+3)}$.

Ainsi, la **puissance du test** qui est égale à 12. $Prob(X_1 \prec X_2 \prec Y_1 \prec Y_2 \prec Y_3)$ a-t-elle pour expression $\eta = \frac{2.a^2}{(a+3).(2.a+3)}$.

2-d) Le calcul de la puissance η pour diverses valeurs de a variant entre 1 et 16, permet de dresser un **tableau comparatif** entre le *test paramétrique de la méthode du rapport de vraisemblance de* NEYMAN *et* PEARSON et *le test non paramétrique de* WILCOXON.

a	Puissance η_1 selon le test paramétrique	Puissance η_2 selon le test de WILCOXON	Différence $\eta_1 - \eta_2$
1	0,10	0,10	0
2	0,26	0,23	0,03
3	0,41	0,33	0,08
4	0,53	0,41	0,12
5	0,62	0,48	0,14
6	0,69	0,53	0,16
7	0,74	0,58	0,16
8	0,78	0,61	0,17
9	0,82	0,64	0,18
10	0,85	0,67	0,18
11	0,87	0,69	0,18
12	0,89	0,71	0,18
13	0,90	0,73	0,17
14	0,92	0,74	0,18
15	0,93	0,76	0,17
16	0,94	0,77	0,17

Les calculs précédents mettent en évidence la **supériorité** de la puissance *du test paramétrique* sur celle du *test non paramétrique* de MANN-WHITNEY-WILCOXON, ce que montre clairement les représentations graphiques ci-dessous (*test paramétrique en pointillé et test non paramétrique en trait plein*) :



4.3 Au sujet du traitement des ex- aequo dans les tests de rangs

Enoncé : Les données ci-dessous établies à partir de deux échantillons indépendants extraits respectivement d'une population A de patients suivant un traitement, et d'une population B de patients non traités (administration d'un placebo), conduisent pour un indicateur de santé considéré aux séries de valeurs ci-dessous :

Malades traités A	19	11	14	17	23	11	15	19	11	8
Malades non traités B	23	19	17	19	23	12	11	14	19	8

1°) Mettre en œuvre le test non paramétrique de MANN-WHITHNEY-WILCOXON au niveau de signification $\alpha = 5\%$ pour se prononcer quant à la différence significative ou non, d'effets entre le médicament étudié et le placebo (on raisonnera par test bilatéral).

2°) La conclusion précédente est-elle modifiée si on applique à la variance de la statistique W de WILCOXON, la correction inhérente aux valeurs ex-aequo.

Solution : 1°) Selon la méthode habituelle (cf. paragraphe 3.2.b du présent chapitre), le classement par ordre croissant des données suivant un échantillon obtenu par regroupement, conduit au tableau ci-dessous, la particularité *de cette application* étant cependant un *grand nombre de valeurs ex-aequo.*

Valeurs	8	8	11	11	11	11	12	14	14	15
Population	A	B	A	A	A	B	B	A	B	A
Rang (*)	1,5	1,5	4,5	4,5	4,5	4,5	7	8,5	8,5	10
Valeurs	17	17	19	19	19	19	19	23	23	23
Population	A	B	B	B	B	A	A	A	B	B
Rang (*)	11,5	11,5	15	15	15	15	15	19	19	19

(*) Le **principe du rang moyen** est préférable à un *tirage au sort* pour fixer les rangs des ex-aequo. Ainsi, pour les deux premières valeurs égales à 8, on prendra comme valeur du rang, la moyenne $(1+2)/2 = 1,5$. De même, pour la valeur commune 11, le rang moyen correspondant est $(3+4+5+6)/4 = 4,5$ et ainsi de suite...

On a donc en définitive, en considérant la somme des rangs des valeurs qui correspondent à la population A dans l'échantillon regroupé ainsi formé, le résultat $W_{calculé} = 1,5+4,5+...+19 = 94$. L'usage d'un *calculateur* (cf. tables numériques en lignes www.geai.univ.brest.fr) voire tout simplement *l'approximation gaussienne*, conduit, pour le *niveau de signification* $\alpha = 5\%$, à la **région critique** $W \notin]79,131[$.

L'hypothèse d'une *différence d'effets* par rapport à l'indicateur mesuré, quant aux malades traités et ceux qui ne le sont pas, *ne peut donc pas être retenue ici* puisque $W_{calculé} = 94 \in]79,131[$, c'est-à-dire qu'on se *situe à l'extérieur de la région critique.*

- A noter que l'usage de **l'approximation normale** de la statistique W par la loi $N(\frac{n_A.(n_A+n_B+1)}{2}, \sigma^2 = \frac{n_A.n_B}{12}.(n_A+n_B+1))$ conduit sensiblement aux mêmes résultats.

En effet, on a numériquement $E(W) = 105, \sigma(W) = 13,23$. A partir de l'erreur de première espèce $\alpha = 5\%$, on a $Prob(|W-105| \geq W_\alpha) = 0,05$.

Suivant la variable normale centrée réduite $\xi = \frac{W-105}{13,23}$ de loi $N(0,1)$, la relation précédente s'écrit $\Pr ob(|\xi| \geq \frac{W_\alpha}{13,23}) = 0,05$. A partir de la table usuelle annexée des valeurs de la fonction de répartition $\Pi(t) = \Pr ob(\xi \prec t)$, l'équation $\Pr ob(|\xi| \geq t_\alpha) = 0,05$ qui s'écrit immédiatement $2.\Pi(t_\alpha) - 1 = 0,95$, entraîne $t_\alpha = 1,96$. Finalement, le seuil critique W_α est égal à $1.96 \times 13,23 = 25,93$, la **région critique** s'écrivant donc, après arrondi, sous la forme $W \notin]79,131[$. *C'est quasiment le même résultat que celui obtenu précédemment par calcul exact.*

2°) Se référant aux rappels de cours, la **correction pour valeurs ex-aequo** applicable à la variance $Var(W)$ s'écrit $Var(W) = \frac{n_A n_B}{N.(N-1)} . \left[\frac{N^3 - N}{12} - T \right]$ avec $N = n_A + n_B$ et $T = \sum_{g=1}^{g=N_g} \frac{t_g^3 - t_g}{12}$, N_g étant le nombre de valeurs différentes des rangs et , pour un rang g donné, t_g le nombre de valeurs ex-aequo.

Numériquement, l'application des formules ci-dessus, s'écrit :

$$T = \frac{2.(4-1)}{12} + \frac{4.(16-1)}{12} + \frac{2.(4-1)}{12} + \frac{2.(4-1)}{12} + \frac{5.(25-1)}{12} + \frac{3.(9-1)}{12} = 18,5 .$$

On a donc $\sigma(W) = \sqrt{\frac{100}{20 \times 19} . \left[\frac{20 \times (400-1)}{12} - 18,5 \right]} = 13,04$.

• Bien que le nombre de valeurs ex-aequo soit particulièrement important ici, *il n'y a pas de différence sensible* entre la valeur de l'écart- type *corrigé* soit 13,04, et l'écart- type *non corrigé* qui est 13,23. En tout état de cause, on n'est pas conduit ici à remettre en question la conclusion précédente qui est celle de l'hypothèse H_0.

4.4 Etude de tendance à l'aide d'échantillons indépendants puis appariés

Enoncé : Les apiculteurs du Texas s'inquiètent de la progression des abeilles africaines plus agressives mais moins productives que les abeilles domestiques. Les pouvoirs publics sont prêts à donner des fonds pour combattre ce phénomène si on peut démontrer que la proportion d'abeilles africaines a augmenté de façon significative ces dernières années.

Les données sont récoltées à l'aide de pièges répartis sur le territoire texan. Des spécialistes identifient les abeilles capturées ce qui permet d'associer à chaque piège la proportion d'abeilles africaines observées.

Deux séries de données ont été ainsi obtenues, l'une en 1980 et l'autre en 1990, quant au pourcentage d'abeilles africaines observées.

Piège	1	2	3	4	5	6	7	8	9	10
% en 1980	0,330	0,146	0,518	0,339	0,693	0,249	0,438	0,695	0,135	0,388
% en 1990	0,360	0,177	0,524	0,447	0,140	0,392	0,534	0,263	0,157	0,566

1°) En l'absence de toute information sur la manière dont on a réparti les pièges en 1980 et en 1990, que peut-on conclure au niveau de signification 5% quant à l'augmentation ou non de la proportion des abeilles africaines ?

2°) En fait, les pièges ont été localisés aux mêmes endroits en 1980 et en 1990. Dans ces conditions que peut-on conclure en appliquant le test des signes au niveau de signification $\alpha = 5\%$.

Solution : 1°) Le test proposé consiste à choisir entre l'hypothèse nulle H_0 « *les lois des pourcentages des abeilles africaines en 1980 (X) et en 1990 (Y) sont les mêmes* », et l'hypothèse alternative H_1 « *la proportion des abeilles africaines a augmenté entre 1980 et 1990* ».

L'absence de toute hypothèse pour la nature de la distribution suivie par X et Y conduit à se tourner vers des **tests non paramétriques** dont le plus puissant est le **test de WILCOXON**. Equivalent au test de MANN et WHITNEY, mais de mise en œuvre plus commode, ce test qui est basé sur les rangs conduit (cf. rappels de cours du présent chapitre, paragraphe 3.2.b) aux étapes suivantes :

→ *Classer par ordre croissant,* l'échantillon regroupé des x_i et des y_j et *noter les rangs correspondants* (les valeurs ex-aequo étant traitées par un rang moyen ou avec une formule de correction).

Rang	Valeur	Variable	Rang	Valeur	Variable
1	0,135	X	11	0,388	X
2	0,140	Y	12	0,392	Y
3	0,146	X	13	0,438	X
4	0,157	Y	14	0,447	Y
5	0,177	Y	15	0,518	X
6	0,249	X	16	0,524	Y
7	0,263	Y	17	0,534	Y
8	0,330	X	18	0,566	Y
9	0,339	X	19	0,693	X
10	0,360	Y	20	0,695	X

→ *Calculer la somme des rangs des valeurs* x_i dans l'échantillon mélangé, soit $W_{calculé} = 105$.

→ *Expliciter la région critique* pour le *test unilatéral* proposé. Or, lorsque $F_X \succ F_Y$, la somme des rangs associés à X est plus faible que celle des rangs associés à Y, d'où une région critique de la forme $W \leq W_\alpha$.

→ *Déterminer* W_α à partir de la donnée de l'erreur de première espèce α. Or, la faible taille des échantillons conduit à utiliser la table spécifique annexée (cf. « *valeurs critiques pour le test unilatéral de WILCOXON-MANN-WHITNEY* »), ce qui, pour $n_1 = 10; n_2 = 10; \alpha = 5\%$, entraîne le seuil $W_\alpha = 82$.

L'inégalité $W_{calculé} = 105 \succ W_\alpha = 82$ *ne permet donc pas ici de retenir l'hypothèse* H_1.

- A noter que la convergence vers la loi normale conduit au seuil dont l'expression est $W_\alpha = 105 - 1,645 \times \sqrt{\frac{10 \times 10 \times 21}{12}} = 83,2$, résultat qui, en dépit des faibles valeurs de n_X et de n_Y est peu éloigné de ce que fournit la table susmentionnée.

• En fait, la méthode de recensement des données utilisée ici comporte deux faiblesses. D'une part, la grande dispersion des valeurs favorise leur interclassement et brouille le discernement entre les hypothèses H_0 et H_1. D'autre part, il faut s'attendre à ce que les proportions d'abeilles africaines soient liées aux emplacements des pièges et aux nombres d'insectes qui y sont capturés. Or à cet égard, il n'existe aucune assurance d'équirépartition ni en 1980 (X), ni en 1990 (Y).

La méthode suivant **échantillons appariés** pour laquelle les lieux où on porte l'attention en 1980 (X) et en 1990 (Y) sont les mêmes, est donc à cet égard plus appropriée.

2°) Les écarts $d_i = x_i - y_i$ sont représentés ci-dessous :

Piège	1	2	3	4	5	6	7	8	9	10
$d_i = x_i - y_i$	-0,030	-0,031	-0,006	-0,108	+0,553	-0,143	-0,096	+0,432	-0,022	-0,178
Signe	-	-	-	-	+	-	-	+	-	-

Le **test des signes** qui porte sur le nombre des signes « - » et « + » consiste à tester l'hypothèse selon laquelle ces différences appartiennent à une distribution de médiane nulle ($\Pr ob(X \prec Y) = \Pr ob(X \succ Y)$), ou non (test unilatéral à gauche, $M \prec 0$, M désignant la médiane de la variable $X - Y$).

Sous l'hypothèse H_0, la loi du nombre D de signes « d_i » positifs suit la **loi binomiale** $B(10, 1/2)$. La **région critique** relative au test considéré est caractérisée quant à elle par l'ensemble $D \leq d_\alpha$ où $\Pr ob(D \leq d_\alpha) = \alpha$. Mais, $D_{calculé}$ étant égal à 2 pour l'échantillon considéré, on a, sous l'hypothèse H_0, $\Pr ob(D \leq 2) = \frac{1}{2^{10}}.(C_{10}^0 + C_{10}^1 + C_{10}^2)$, soit la valeur numérique 0,055 (puisque sous l'hypothèse H_0, $\Pr ob(D = d) = \frac{1}{2^{10}} C_{10}^d$).

Bref, numériquement, on obtient $\Pr ob(D \leq 2) = 0,055$ (résultat qui est aussi celui que fournit la table annexée (cf. « *table des valeurs critiques de la loi binomiale* »).

• En définitive, la définition de la région critique $D \leq d_\alpha$ étant équivalente à la condition $\Pr ob(D \leq d) \leq \alpha$, on a pour le cas présent, $\Pr ob(D \leq d_{calculé}) = 0,055 \succ 0,05$. *On ne peut donc pas, en l'occurrence, retenir l'hypothèse de rejet* H_1. Par contre, il est manifeste que la zone de rejet de H_0 est proche et d'ailleurs, si on modifie à la hausse, le niveau α de signification du test, par la valeur $\alpha = 5,5\%$, c'est H_1 qu'il faudrait retenir.

4.5 Evaluation de l'efficacité d'un traitement suivant plusieurs tests non paramétriques

Enoncé : Un médecin souhaite vérifier l'efficacité d'un traitement dont il pense qu'il peut prolonger la vie de malades ayant déjà eu un infarctus. Il choisit pour cela dix malades comparables à tous points de vue, en prend cinq au hasard à qui il applique le traitement, les cinq autres étant des témoins non traités à qui on administre un placebo.

Les résultats concernant la durée de survie (exprimée en années) des malades étudiés sont présentés ci-après :

Malades traités (X)	6,5	4,2	17,8	7,9	13,2
Malades non traités (Y)	6,7	0,4	2,9	1,2	5,6

PARTIE I (test de la médiane de MOOD)

Datant de 1950, ce « test des signes » consiste à partir de la médiane générale calculée par regroupement des deux échantillons, soit M, à tester l'indépendance entre deux caractères à deux classes, en l'occurrence, malades traités et non traités, et, survie à plus de M années et à moins de M années. A cet effet, on procédera comme suit :

1°) Calculer M.

2°) Compléter le tableau des effectifs ci-dessous :

	Durée de vie $\leq M$	Durée de vie $\succ M$	Σ
Malades traités			
Malades non traités			
Σ			

3°) calculer les probabilités d'obtenir un écart entre les deux groupes supérieur ou égal à la configuration précédente (méthode exacte de FISHER).

4°) En déduire la règle de décision du test.

PARTIE II (test de WILCOXON-MANN-WHITNEY)

Au contraire du test antérieur qui entraîne une perte d'information notable puisque ignorant les valeurs (X) et (Y) en les remplaçant par un tableau qui se résume à totaliser les effectifs de celles qui sont inférieures à la médiane et de celles qui lui sont supérieures, le test classique de WILCOXON est plus puissant puisqu'il s'accompagne d'une meilleure prise en compte de ces valeurs à travers leurs rangs.

1°) Toujours au seuil $\alpha = 0,05$, expliciter la règle de décision inhérente à un tel test et conclure quant à l'efficacité ou non du traitement étudié.

2°) Retrouver ce résultat par une méthode exacte (sans l'usage des tables de valeurs).

PARTIE III (test de KOLMOGOROV-SMIRNOV)

1°) Suivant la version du test de KOLMOGOROV applicable à la comparaison de deux échantillons indépendants, exprimer la règle de décision correspondante et la conclusion à retenir pour ce qui est des données proposées.

2°) Que conclure de l'ensemble des résultats antérieurs ?

Solution : I-1°) La **médiane** est le nombre qui, pour une *variable continue*, vérifie la relation $\Pr ob(X \leq x) = \frac{1}{2}$. Pour une *variable discrète*, c'est $M = x_{\left(\frac{n+1}{2}\right)}$ si n est impair et $M = \frac{1}{2}.(x_{\frac{n}{2}} + x_{\left(\frac{n}{2}+1\right)})$ si n est pair, les données ayant été préalablement classées par ordre croissant. Ainsi, pour le cas présent et à partir de la distribution regroupée ci-dessous, a-t-on, $n = 10 \Rightarrow M = \frac{1}{2}.(5,6 + 6,5) = 6,05$.

0,4	1,2	2,9	4,2	5,6	6,5	6,7	7,9	13,2	17,8
Y	Y	Y	X	Y	X	Y	X	X	X

I-2°) Arrondissant la valeur de M à 6, les données précédentes conduisent immédiatement, pour le tableau proposé dans l'énoncé, aux renseignements ci-dessous :

	Durée de vie ≤ 6	Durée de vie $\succ 6$	Σ
Malades traités	1	4	5
Malades non traités	4	1	5
Σ	5	5	10

I-3°) Enumérant toutes les situations pour lesquelles l'écart, entre les deux groupes, dans la répartition des deux modalités considérées est supérieur ou égal à la configuration susmentionnée, on a deux formes possibles pour le cas proposé à savoir $\begin{matrix}1 & 4\\ 4 & 1\end{matrix}$ et $\begin{matrix}0 & 5\\ 5 & 0\end{matrix}$.

Groupes	Modalités		Σ
	≤ 6	$\succ 6$	
T	A	B	L_1
N.T	C	D	L_2
Σ	T_1	T_2	N

De façon générale, et pour le schéma ci-contre, le calcul de la probabilité associée sous l'hypothèse nulle H_0, équivaut au fait d'obtenir A éléments de type « T » et C éléments de type « $N.T$ » lors d'un tirage sans replacement dans une population qui comprend L_1 éléments de type « T » et L_2 éléments de type « $N.T$ ».

Bref, un modèle décrit par la **loi hypergéométrique** et en fonction duquel la probabilité cherchée est $\dfrac{C_{L_1}^{A}.C_{L_2}^{C}}{C_N^{T_1}} = \dfrac{L_1!L_2!T_1!T_2!}{A!B!C!D!N!}$. Numériquement, on obtient ainsi pour la configuration $\begin{matrix}0 & 5\\ 5 & 0\end{matrix}$ la probabilité $\dfrac{5!5!}{10!} = 0,00397$ et pour la configuration $\begin{matrix}1 & 4\\ 4 & 1\end{matrix}$, le résultat $\dfrac{5!5!5!5!}{1!4!1!4!10!} = 0,0992$. Au total, la probabilité d'obtenir, sous H_0, un écart dans la répartition des deux modalités supérieur ou égal à la configuration observée est donc égal à la somme $0,0992 + 0,00397 = 0,103$.

I-4°) La probabilité calculée précédemment représente la probabilité de rejeter à tort H_0, c'est-à-dire la probabilité d'obtenir, entre les deux groupes, un écart dans la répartition des deux modalités considérées, supérieur ou égal à celui qui a été observé. Si on souhaite limiter ce risque à 5%, le résultat obtenu pour cette probabilité, à savoir 0,103, ne permet pas de rejeter l'hypothèse H_0 puisqu'on a $0,103 \succ 0,05$.

• Lorsque les effectifs des groupes dépassent 15 unités, on ramènera le problème en question à un *test de* χ^2 dont la transcription de la distance correspondante, pour le schéma général précédent, conduit à l'indicateur :

$$\chi^2_{calculé} = \frac{N.(A.D - B.C)^2}{(A+C).(B+D).(A+B).(C+D)}$$

(χ^2 suivant par ailleurs, la loi du chi- deux à un degré de liberté, soit $\chi^2(1)$. A cet égard, on se reportera aux rappels de cours du présent chapitre (cf. paragraphe 3.4.c), pour constater que le nombre de degrés de liberté, lorsqu'il s'agit de tableaux (2,2) est bien $\nu = (r-1).(k-1)$, soit $\nu = (2-1).(2-1) = 1$).

Finalement et pour le test de chi- deux en question, on rejette l'hypothèse H_0 si $\chi^2_{calculé} \geq \chi^2_\alpha$, le seuil critique χ^2_α étant déterminé par la relation $\Pr ob(\chi^2(1) \geq \chi^2_\alpha) = \alpha$.

A noter enfin, l'usage possible de la *correction de continuité* de YATES pour le cas des petits échantillons, ceci suivant *l'indicateur corrigé* :

$$\chi^2_{calculé} = \frac{N.(|A.D - B.C| - N/2)^2}{(A+B).(B+D).(A+C).(C+D)}$$

II-1°) Reprenant les résultats présentés en rappels de cours du présent chapitre (cf. paragraphe 3.2.b), le **test de MANN-WHITNEY-WILCOXON** pour deux échantillons indépendants conduit successivement à :

→ *Classer par ordre croissant*, l'échantillon mélangé (X,Y) et *affecter les rangs correspondants* (principe du rang moyen en cas d'égalité). Il s'ensuit pour les données proposées, le tableau :

Valeurs	0,4	1,2	2,9	4,2	5,6	6,5	6,7	7,9	13,2	17,8
Variables	*Y*	*Y*	*Y*	*X*	*Y*	*X*	*Y*	*X*	*X*	*X*
Rangs	1	2	3	4	5	6	7	8	9	10

→ *Calculer la somme des rangs associés aux valeurs de* X, soit $W_{calculé} = 37$.

→ *Lire la valeur du seuil critique* $W_\alpha / \Pr ob(W \geq W_\alpha) = \alpha$ (cf. table annexée des valeurs critiques pour le test unilatéral de WILCOXON-MANN-WHITNEY), soit, pour $N_1 = 5$, $N_2 = 5$, et $\alpha = 0,05$, la valeur $W_\alpha = 36$.

→ *Comparer* W_α *et* $W_{calculé}$. En l'occurrence, la relation $W_{calculé} = 37 \succ W_\alpha = 36$ conduit à rejeter l'hypothèse H_0 au contraire du test précédent de la médiane, différence de résultat qui souligne la meilleure sensibilité du présent test de WILCOXON-MANN-WHITNEY.

II-2°) Afin de calculer exactement la probabilité $\Pr ob(W \geq 37)$, énumérons toutes les répartitions de rang (relatives à X) pour lesquelles $\sum_{i=1}^{i=5} R_i \geq 37, 1 \leq R_i \leq 10$. Il en résulte immédiatement, les 7 cas possibles :

$(6,7,8,9,10) \Rightarrow W = 40$
$(5,7,8,9,10) \Rightarrow W = 39$
$(5,6,8,9,10) \Rightarrow W = 38$
$(4,7,8,9,10) \Rightarrow W = 38$
$(5,6,7,9,10) \Rightarrow W = 37$
$(4,6,8,9,10) \Rightarrow W = 37$
$(3,7,8,9,10) \Rightarrow W = 37$

Or, parmi les 10 ! permutations possibles des rangs des 10 valeurs de l'échantillon, chacun des cas ci-dessus est assimilable à un partitionnement au sein duquel les cinq valeurs de rangs liées à X (respectivement, les cinq valeurs de rangs liées à Y), sont bloquées, le nombre de permutations qui, sous cette configuration, sont ainsi possibles, étant égal à 5 !.5 !.

Ainsi, chacun des sept cas ci-dessus, intervient-il sous forme non ordonnée, avec la probabilité $\frac{5!5!}{10!}$, les sept cas étant par ailleurs équiprobables.

En définitive, on a $\Pr ob(W \geq 37) = 7 \times \frac{5!5!}{10!} = 0,026$.

Au niveau de signification $\alpha = 5\%$, *on rejette donc l'hypothèse* H_0 et on conclut donc à l'efficacité du traitement.

• Le même raisonnement étendu des cinq cas supplémentaires pour lesquels $\sum_{i=1}^{i=5} R_i = 36$ conduit à l'évaluation $\Pr ob(W \geq 36) = 0,045$. Plus encore, allant jusqu'à la valeur 35, on a $\Pr ob(W \geq 35) = 0,06$. A partir de ces deux résultats, on retrouve donc la valeur par arrondi du seuil critique fourni par la table de FISHER, à savoir $W_\alpha = 36$.

III-1°) Si on aborde la question posée de l'efficacité ou non traitement sur un plan **unilatéral**, la **distance de KOLMOGOROV** à considérer est caractérisée par $D(\widehat{F_X}, \widehat{F_Y}) = \underset{x \in R}{Sup}(\widehat{F_Y(x)} - \widehat{F_X(x)})$ (du moins dans le sens où on teste la supériorité des valeurs de X par rapport à celles de Y).

La mise en œuvre du test de KOLMOGOROV (cf. rappels de cours, paragraphe 3.2.a du présent chapitre) conduit successivement à :

→ *Classer par ordre croissant* les valeurs de X et de Y et *mettre en regard les fonctions de répartition* $\widehat{F_X}$ et $\widehat{F_Y}$ ainsi que les différences $\widehat{F_Y} - \widehat{F_X}$, données desquelles on déduit la *distance* $D(\widehat{F_X}, \widehat{F_Y}) = \underset{x \in R}{Sup}(\widehat{F_Y(x)} - \widehat{F_X(x)})$. Cette étape est décrite ci-dessous, N_X et N_Y désignant les effectifs cumulés associés aux valeurs de X et de Y :

Valeurs	N_X	N_Y	$\widehat{F_X} = N_X/5$	$\widehat{F_Y} = N_Y/5$	$\widehat{F_Y} - \widehat{F_X}$
0,4	0	1	0	0,2	0,2
1,2	0	2	0	0,4	0,4
2,9	0	3	0	0,6	0,6
4,2	1	3	0,2	0,6	0,4
5,6	1	4	0,2	0,8	0,6
6,5	2	4	0,4	0,8	0,4
6,7	2	5	0,4	1,0	0,6
7,9	3	5	0,6	1,0	0,4
13,2	4	5	0,8	1,0	0,2
17,8	5	5	1,0	1,0	0

Ainsi, la valeur calculée de la statistique de KOLMOGOROV-SMIRNOV est-elle égale ici à $D(\widehat{F_X}, \widehat{F_Y}) = 0,6$.

→ La *région critique* ayant pour forme $D(\widehat{F_X}, \widehat{F_Y}) \geq D_\alpha$ et les valeurs de la distribution de KOLMOGOROV-SMIRNOV étant tabulées, *déterminer, pour l'erreur de première espèce* α *donnée, la valeur du seuil critique* D_α. A cet effet, l'utilisation de la table annexée qui porte sur la statistique $K_D = n.D$ (où $n = n_X = n_Y$) (cf. table « *valeurs critiques de* K_D *pour le test de KOLMOGOROV-SMIRNOV à deux échantillons petits et de même taille* »), conduit pour $\alpha = 0,05$ et $n_X = n_Y = 5$, *version unilatérale*, à la valeur $K_\alpha = 4$.

→ *Comparer* $K_{calculé}$ à K_α, ce qui pour le cas traité ici, s'écrit $K_{calculé} = 3 \prec K_\alpha = 4$.

→ On se trouve donc être ici à l'extérieur de la région critique $K \geq 4$, *ce qui ne permet pas de retenir l'hypothèse de l'efficacité du traitement.*

III-2°) Des trois tests précédents, c'est donc le test de WILCOXON-MANN-WHITNEY qui est le plus puissant et qui, en l'occurrence est le seul à pouvoir, à risque de première espèce constant, conclure au rejet de l'hypothèse H_0, c'est-à-dire à l'efficacité du traitement.

4.6 Etude d'impact suivant le test de MAC NEMAR

Ce test qui s'applique à un ensemble de sujets mesurés de façon ordinale ou nominale à deux instants séparés par un certain traitement (au sens large) est particulièrement adapté à l'analyse du changement des sujets en question sur un certain point, entre l'avant et l'après (une formation, une lecture, une visite, un constat, un traitement médical...).

Enoncé : Trente sujets sont soumis à un test avant (« pré test ») et après (« post test ») une formation, les résultats possibles, en étant « réussite » ou « échec ». Paris ceux-ci, ils sont :

- 3 sujets à avoir réussi les deux tests avant et après formation ;
- 21 sujets à avoir échoué au premier test (pré test) ;
- 12 sujets à avoir réussi au post test.

1°) Considérant les deux étapes « avant » et « après » et les états correspondants aux résultats d'un test (« état 0 » si échec et « état 1 » si réussite), établir la matrice des probabilités de transition entre les étapes en question.

2°) Testant l'invariance de la proportion de réussite (ou d'échec) entre avant et après la formation, en déduire à l'aide de la distance de chi- deux, la conclusion à retenir quant à l'efficacité ou non de la formation, ceci au niveau de signification $\alpha = 5\%$.

Solution : 1°) La **table de contingences** dressée ci-dessous quant aux nombres de réussite aux tests « *avant* » et « *après* » la formation peut aisément être complétée par les données portées en caractère gras et italique.

Avant\Après	Echec	Réussite	Total
Echec	***12***	***9***	21
Réussite	***6***	3	***9***
Total	***18***	12	30

Ainsi, par complémentarité, il y a 9 réussites au pré test et donc 6 sujets ayant réussi au pré test puis échoué au post test. De même, par complémentarité, il y a 18 échecs au total pour le post test et donc 12 cas d'échecs répétés aux pré et post tests. Enfin, par complémentarité à 30, il y a 9 sujets qui ayant échoué au pré test réussissent le post test.

2°) Bien que proche du chi- deux, le **test de MAC NEMAR** se distingue du test d'homogénéité pour tables de contingences (2,2) tel celui développé dans l'application 3.5 antérieure, par le fait que ce sont les mêmes personnes qui sont interrogées ici avant et après la formation et que les informations recueillies (avant et après) ne sont donc pas indépendantes. Comme les échantillons sont appariés, les effectifs des couples (échec, échec) et (réussite, réussite) n'apportent aucune information sur l'écart entre « *l'avant* » et « *l'après* » formation, ces écarts étant décrits par les flux *Echec* → *Réussite* et *Réussite* → *Echec*.

Avant\Après	0	1
0	*a*	*b*
1	*c*	*d*

Plus précisément et reprenant les notations ci-contre et les résultats des rappels de cours (cf. paragraphe 3.3.c), le test de l'hypothèse H_0 conduit à comparer les effectifs des échanges 0→ 1 et 1→ 0, soient b et c, à l'effectif théorique commun $(b+c)/2$ (flux égaux).

Le calcul de la *distance de chi- deux* sous la réserve habituelle d'un effectif théorique supérieur à 5 ($\frac{b+c}{2} \succ 5$) permet de conclure par comparaison au seuil χ^2_α vérifiant $\Pr ob(\chi^2 \geq \chi^2_\alpha) = \alpha$, le nombre de degrés de liberté étant égal à 1 en l'occurrence.

Or, cette distance, c'est $\frac{(b-\frac{b+c}{2})^2}{\frac{(b+c)}{2}} + \frac{(c-\frac{b+c}{2})^2}{\frac{(b+c)}{2}} = \frac{(b-c)^2}{b+c}$ ou encore, *avec la correction de continuité* de YATES, $\frac{(\left|b-\frac{b+c}{2}\right|-0,5)^2}{\frac{(b+c)}{2}} + \frac{(\left|c-\frac{b+c}{2}\right|-0,5)^2}{\frac{(b+c)}{2}} = \frac{(|b-c|-1)^2}{b+c}$.

Numériquement, $b=9, c=6 \Rightarrow \frac{b+c}{2} \succ 5$, ce qui légitime l'usage du test de chi- deux.

Il en ressort $\chi^2_{calculé} = \frac{(9-6)^2}{15} = 0,6$ qu'il faut comparer, au niveau de signification $\alpha = 5\%$, au seuil χ^2_α vérifiant $\Pr ob(\chi^2(1) \geq \chi^2_\alpha) = 0,05$, soit $\chi^2_\alpha = 3,84$ (cf. table de valeurs annexée).

La relation $\chi^2_{calculé} \prec \chi^2_\alpha$ conduit ici à *ne pas rejeter l'hypothèse H_0 d'une formation sans effet.*

• C'est d'autant plus vrai si, tenant compte de la petite taille des effectifs considérés, on applique la correction de continuité de YATES et la distance correspondante $\frac{(|b-c|-1)^2}{b+c}$ puisqu'on obtient alors pour $\chi^2_{calculé}$ la valeur 0,26 (et non plus 0,6 !).

• A noter que $\sqrt{\chi^2(1)}$ suivant la loi normale centrée réduite $N(0,1)$, le test se ramène à utiliser la statistique $\frac{|b-c|}{\sqrt{b+c}}$ et à comparer le résultat à $t_\alpha / \Pr ob(|\xi| \geq t_\alpha) = 0,05 \Rightarrow t_\alpha = 1,96$ (qui est aussi la racine carrée de $\chi^2_\alpha = 3,84$).

4.7 Coefficient de contingence

Applicable aux tables de contingence (k,k), le coefficient de contingence constitue une mesure d'association simple dont la valeur augmente avec le degré de dépendance des classifications effectuées. Cette notion peu usitée est illustrée ci-après pour le cas courant de tables $(2,2)$.

Enoncé : Considérant une table de données (k,k) pour lesquelles le test d'indépendance conduit à la distance de chi- deux, soit χ^2, le nombre des observations effectuées étant par ailleurs égal à n, on définit le coefficient de contingence par la formule $C = \sqrt{\frac{\chi^2}{\chi^2+n}}$.

1°) Evaluer la valeur maximale de C pour le cas particulier $k=2$.

2°) Le tableau ci-dessous, présente les relations entre la couleur des yeux et des cheveux pour un échantillon de 200 femmes.

		Couleur des cheveux		Total
		Blonds	Non blonds	
Couleur des yeux	Bleus	49	25	74
	Non bleus	30	96	126
Total		79	121	200

Calculer le coefficient de contingence. Que peut-on en conclure par comparaison avec le coefficient maximal obtenu à la 1ère question ?

Solution : 1°) La *valeur maximale du coefficient de contingence* C est atteinte lorsque les deux *classifications* étudiées sont *totalement dépendantes* ou associées, tous ceux qui ont les cheveux blonds, par exemple, ayant les yeux bleus, et par complémentarité, tous ceux ayant les cheveux non blonds n'ayant pas les yeux bleus. Il s'ensuit ainsi, dans cette situation extrême, un tableau de données de la forme ci-dessous :

	Modalités		Total
	I	II	
Groupe A	a	0	a
Groupe B	0	b	b
Total	a	b	$a+b$

Suivant la formule synthétique déjà démontrée dans l'application 3.5 antérieure, la **distance de chi- deux calculée**, sous l'hypothèse de l'indépendance entre les deux classifications, s'écrit (sans correction de continuité) sous la forme :

$$\chi^2_{calculé} = \frac{n.(a.b-0)^2}{a.b.a.b} = n$$

Ainsi, C est-il égal en la circonstance à $C_{\max} = \sqrt{\frac{n}{n+n}} = \frac{\sqrt{2}}{2} = 0,707$. On remarque ainsi, qu'au contraire du *coefficient de corrélation linéaire* de PEARSON dont la valeur absolue reste comprise entre 0 et 1, le *coefficient de contingence* qui est tout autre, a un champ de valeurs plus large qui est fonction de la taille de la table de contingence considérée.

A cet égard, on peut montrer qu'on a, plus généralement, pour une table (k,k), le résultat $C \leq C_{\max} = \sqrt{\frac{k-1}{k}}$.

2°) Pour les données proposées, on a immédiatement :

$$\chi^2_{calculé} = \frac{200\times(49\times96-25\times30)^2}{79\times121\times74\times126} = 35,08$$

Pour ce qui est du coefficient de contingence, $C = \sqrt{\frac{\chi^2}{\chi^2+n}}$, on a donc numériquement, $C = \sqrt{\frac{35,08}{35,08+200}} = 0,386$.

• Le même calcul, avec correction de continuité de YATES, conduirait à un résultat assez proche, puisqu'on aurait $\chi^2_{calculé} = \dfrac{200\times(|49\times 96-25\times 30|-100)^2}{79\times 121\times 74\times 126} = 33{,}33$, d'où pour C, la valeur 0,378.

Cette valeur qui est légèrement supérieure au milieu de l'intervalle $[C_{min}=0-C_{max}=0{,}707]$ traduit un degré d'association notable. Elle est d'ailleurs, largement supérieure à la valeur de C qui correspond au seuil critique $\chi^2_\alpha = 3{,}84$ à partir duquel on rejette l'hypothèse d'indépendance H_0, soit $C = 0{,}137$.

4.8 Alternative au « r » de PEARSON, le coefficient « τ » de KENDALL

D'une puissance légèrement inférieure au test paramétrique du r de PEARSON lorsque les conditions de validité de ce dernier sont vérifiées (notamment le caractère de binormalité de la distribution du couple), le test du tau de KENDALL, dont le champ d'application est celui des variables ordinales (donc plus large que le cadre des variables quantitatives), s'impose dès qu'on s'écarte des conditions susmentionnées du r de PEARSON (petits échantillons, distribution inconnue, modèle non linéaire...).

Enoncé : On mesure (en cm) la longueur (X) et la largeur (Y) de dix fleurs de l'espèce « iris setosa », les résultats obtenus en étant les suivants :

Fleur	1	2	3	4	5	6	7	8	9	10
Longueur	4,5	5,1	4,6	4,4	5,2	4,8	5,5	5,0	5,6	5,3
Largeur	3,0	3,6	3,4	2,9	3,5	3,1	4,2	3,2	3,7	3,8

1°) Pour les données susmentionnées, calculer le coefficient de corrélation linéaire de PEARSON et le coefficient tau de KENDALL.

2°) On observe une fleur supplémentaire dont les dimensions sont très distinctes des données précédentes puisque de longueur 4,9 et de largeur 1,0.

Comment varient les coefficients « r » de PEARSON et « τ » de KENDALL, compte tenu de ce nouvel élément ?. Qu'en conclure ?

3°) A partir du test de KENDALL, existe-t-il une liaison significative entre les deux grandeurs mesurées (X) et (Y), ceci au niveau de signification $\alpha = 0{,}05$?

Solution : 1°) Notant par $(x_i, y_i), 1 \le i \le 10$, les données proposées, le **coefficient de corrélation linéaire de PEARSON**, s'écrit $r_{X,Y} = \dfrac{\sum_{i=1}^{i=10} x_i.y_i - 10.\bar{x}.\bar{y}}{\sqrt{\sum_{i=1}^{i=10} x_i^2 - 10.\bar{x}^2}.\sqrt{\sum_{i=1}^{i=10} y_i^2 - 10.\bar{y}^2}}$,

soit numériquement $r_{X,Y} = 0{,}86$.

• D'autre part, la mise en œuvre du calcul du **coefficient « τ » des rangs de KENDALL** conduit à *dénombrer le nombre d'inversions* constatées dans les classements des rangs entre les deux ensembles de valeurs considérés, soit N, le coefficient τ étant égal à $\tau = -1 + 2.\dfrac{N}{\dfrac{n.(n-1)}{2}}$, n désignant le nombre de couples (x_i, y_i) (cf. rappels de cours ; paragraphe 3.4.b).

Le mode opératoire ci-dessous facilite le calcul de N :

→ Pour les n couples (x_i, y_i) *expliciter* au sein de chacune des séries (X) et (Y) *les rangs correspondants* :

Fleur	1	2	3	4	5	6	7	8	9	10
Rang (X)	2	6	3	1	7	4	9	5	10	8
Rang (Y)	2	7	5	1	6	3	10	4	8	9

A noter qu'il n'y a pas d'ex aequo ici, mais dans le cas contraire, on aurait appliqué la *méthode du rang moyen*, avec pour le calcul de τ, l'utilisation d'une formule corrective (cf. rappels de cours).

→ *Réordonner les rangs* de façon à ce que, par rapport à (X), la présentation de ceux-ci s'effectue dans l'ordre croissant naturel, soit :

Fleur	4	1	3	6	8	2	5	10	7	9
Rang (X)	1	2	3	4	5	6	7	8	9	10
Rang (Y)	1	2	5	3	4	7	6	9	10	8

→ *Pour toutes les comparaisons deux à deux possibles*, entre fleurs, comparaisons qu'on aura réduit de moitié par symétrie (elles sont donc au nombre totale de $C_n^2 = \frac{n.(n-1)}{2}$, soit numériquement 45), *affecter* un coefficient « + » pour les préférences de X et de Y qui sont en concordances et un coefficient « - » dans le cas contraire.

Pour faciliter cette énumération et le report des résultats, on pourra s'appuyer sur le tableau symétrique (10,10) ci-dessous :

	1	2	3	4	5	6	7	8	9	10
1		+	+	+	+	+	+	+	+	+
2			+	+	-	+	+	+	+	+
3				+	+	-	+	-	+	+
4					+	+	+	+	+	+
5						+	+	+	+	+
6							+	+	+	+
7								+	-	+
8									+	+
9										-
10										

Concrètement, pour remplir ce tableau, on identifie à partir de l'énumération des couples de fleurs $(i, j), j \succ i$, ceux pour lesquels le sens des variations de X et de Y coïncident (marquage par un « + ») et à contrario, les discordances (marquage par un « - »), le classement des rangs X par ordre croissant facilitant la mise en œuvre de cette procédure.

Ainsi, (1,2), (1,3), (1,4), (1,5), (1,6), (1,7), (1,8), (1,9), et (1,10) sont-ils tous en concordances parce que de rangs évoluant tous deux dans le même sens (croissant, pour les couples en question), donc marqués positivement. De même pour (2,3) et (2,4). Par contre, il en est autrement pour (2,5) pour lequel le rang (X) augmente de 6 à 7 alors que le rang (Y) diminue de 7 à 6, couple qu'on marquera ainsi par « - ». Et ainsi de suite, pour les autres paires….

→ Notant par S la différence entre le nombre de paires marquées par « + » et celles qui sont marquées par « - », *en déduire la valeur du coefficient tau* de KENDALL, soit $\tau = \dfrac{S}{\dfrac{n.(n-1)}{2}}$. Pour le cas présent, on obtient $\tau = \dfrac{40-5}{45} = 0,78$, valeur assez proche de la valeur unité obtenue lorsque toutes les paires sont en accord et qui confirme donc une liaison probable comme l'indiquait déjà le coefficient « r » de PEARSON.

• Autre méthode possible, la *procédure de marquage* présentée en rappels de cours (cf. paragraphe 3.4.b) et dont la mise en œuvre conduit à relier les rangs « X » et « Y » identiques et à compter le nombre d'intersection des liaisons ainsi obtenues, pour identifier le nombre de paires en divergence.

Il en résulte le tableau ci-dessous duquel on recense 5 paires divergentes :

Fleur	4	1	3	6	8	2	5	10	7	9
Rang (X)	1	2	3	4	5	6	7	8	9	10
Rang (Y)	1	2	5	3	4	7	6	9	10	8

Ainsi retrouve-t-on le résultat précédent puisque le nombre total de paires étant égal à $C_n^2 = \dfrac{n.(n-1)}{2}$, soit 45, on a 45-5=40 paires convergentes et de ce fait $\tau = \dfrac{40-5}{45} = 0,78$.

2°) Avec le couple $(4,9-1,0)$ portant sur une $11^{ème}$ fleur ajoutée à l'échantillon des dix fleurs précédent, la *nouvelle valeur du coefficient de corrélation linéaire* de PEARSON est de 0,46. Autrement dit, la valeur «aberrante» qui a été ainsi rajoutée modifie considérablement le coefficient « r » et en souligne touts sa *sensibilité.*

• C'est *différent* pour le coefficient « τ » de KENDALL. En effet, suivant la procédure de marquage des rangs identiques et des intersections des lignes qui les relient, procédure qui a été explicitée ci-dessus et qui est la plus simple à mettre en œuvre, le nouveau tableau des rangs qui en résulte conduit à 7 paires divergentes de par les étapes suivantes :

→ *Tableau des rangs*

Fleur	1	2	3	4	5	6	7	8	9	10	11
Rang (X)	2	7	3	1	8	4	10	6	11	9	5
Rang (Y)	3	8	6	2	7	4	11	5	9	10	1

→ *Tableau des rangs classés par ordre croissant suivant (X) et marquage*

Fleur	4	1	3	6	11	8	2	5	10	7	9
Rang (X)	1	2	3	4	5	6	7	8	9	10	11
Rang (Y)	2	3	6	4	1	5	8	7	10	11	9

On a donc $\tau = \dfrac{48-7}{55} = 0,74$, c'est-à-dire une *valeur très proche* de celle trouvée précédemment sans ajout de la donnée « aberrante », et ceci au contraire du coefficient « r » de PEARSON dont on a constaté qu'il était très sensible.

3°) Le **test de l'indépendance** entre X et Y se ramène à vérifier si les nombre de paires discordantes et concordantes sont égales, c'est-à-dire au test $\begin{cases} H_0 : \tau = 0 \\ H_1 : \tau \neq 0 \end{cases}$ (cf. rappels de cours).

Le cas des petits échantillons ($n \leq 10$), tel dans le présent exercice, oblige à recourir à une *table de valeurs spécifique*, à défaut de pouvoir utiliser la convergence vers la loi normale. Un *extrait de ces tables* est reproduit ci-dessous, la *valeur lue* en fonction du niveau de signification fixé (1% ou 5%) correspondant à la *différence entre les nombres de paires en concordance et de paires en discordance.*

N	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
$\alpha = 0,05$	8	11	13	16	18	21	23	26	28	33	35	38	42	45	49	52
$\alpha = 0,01$	10	13	17	20	24	27	31	36	40	43	49	52	58	63	67	72

Ainsi, pour $N = 10$ et $\alpha = 5\%$ lit-on une valeur de seuil égale à 21. Or, pour l'échantillon considéré (celui à dix valeurs), la valeur observée quant à cette différence entre les nombres des paires concordantes et discordantes est égale à 40-5=35 (cf. 1ère question).

C'est donc l'hypothèse H_1 qu'on peut retenir ici puisque $D_{calculé} \succ D_\alpha$, c'est-à-dire l'hypothèse d'une liaison entre les grandeurs X et Y.

- Utilisant *l'approximation normale* (cf. rappels de cours), le **seuil critique** obtenu serait égal à $1,96 \times \frac{n.(n-1)}{2} \times \sqrt{\frac{2.(2n+5)}{9n.(n-1)}}$, soit numériquement la valeur 21,91. C'est une *valeur très proche* de celle fournie par la table de KENDALL et ceci valide donc le large usage que l'on peut faire de l'approximation normale.

4.9 Coefficient Rhô de SPEARMAN

Egalement basé sur la notion de rang, le coefficient Rhô de SPEARMAN est d'une portée et d'une efficacité semblables à celles du coefficient Tau de KENDALL.

Enoncé : Le tableau ci-dessous indique la mortalité annuelle moyenne (X) pour les hommes âgés de 45 ans à 64 ans de 1958 à 1964 et la concentration (Y) en ions calcium de l'eau potable, pour trente villes d'Angleterre et du Pays de Galles.

Ville	Mortalité pour 100000 habitants	Calcium (ppm)	Ville	Mortalité pour 100000 habitants	Calcium (ppm)
Newcastle	1,702	44	Southampton	1,369	68
Northampton	1,309	59	Southend	1,257	50
Norwich	1,259	133	Southport	1,587	75
Nottingham	1,427	27	Southshields	1,713	71
Oldham	1,724	6	Stockport	1,557	13
Oxford	1,175	107	Stoke	1,640	57
Plymouth	1,486	5	Sunderland	1,709	71
Portsmouth	1,456	90	Wallasey	1,625	20
Preston	1,696	6	Walsall	1,527	60
Reading	1,236	101	West Bromwich	1,627	53
Rochdale	1,711	13	West Ham	1,486	122
Rotherdam	1,444	14	Wolverhampton	1,485	81
St Helens	1,591	49	York	1,378	71
Salford	1,987	8	Cardiff	1,519	21
Sheffield	1,495	14	Newport	1,581	14

1°) Calculer pour les données en question, la valeur du coefficient Rhô de SPEARMAN.

2°) Peut-on conclure au niveau de signification 5% à un lien entre les facteurs (X) et (Y) étudiés ?

3°) Comment est modifiée la valeur du coefficient Rhô si on applique les formules de correction pour ex-aequo ?

Solution : 1°) Le calcul des **rangs** R_i et S_i qui correspondent aux données (X_i,Y_i) associées à la ville $i, 1 \le i \le 30$, conduit au tableau ci-dessous, le principe du *rang moyen* ayant été appliqué pour le cas des valeurs ex-aequo.

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
R_i	25	5	4	8	29	1	12,5	10	24	2	27	9	20	30	14
S_i	13	18	30	12	2,5	28	1	26	2,5	27	5,5	8	14	4	8
i	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
R_i	6	3	19	28	17	23	26	21	16	22	12,5	11	7	15	18
S_i	20	15	24	22	5,5	17	22	10	19	16	29	25	22	11	8

Il en résulte pour le coefficient Rhô, soit $\rho = 1 - \dfrac{6.\sum_{i=1}^{i=n}(R_i - S_i)^2}{n.(n^2-1)}$ (cf. rappels de cours du présent chapitre, paragraphe 3.4.a), la valeur numérique $\rho = -0,487$.

2°) n étant assez grand ici puisque supérieur à 10, l'approximation de la statistique $T = \rho.\sqrt{\dfrac{n-2}{1-\rho^2}}$ par la **loi de STUDENT** à $\nu = n-2$ degrés de liberté, est autorisée ici (cf. rappels de cours du présent chapitre, paragraphe 3.4.a).

Le test d'indépendance $\begin{cases} H_0 : \rho = 0 \\ H_1 : \rho \neq 0 \end{cases}$ conduit donc à la **région critique** définie par $|\rho| \ge \pi = \dfrac{t_\alpha}{\sqrt{n-2+t_\alpha^2}}$ où t_α vérifie $\Pr ob(|T| \ge t_\alpha) = \alpha$. Numériquement, et par lecture dans la table de STUDENT annexée, on a pour $\alpha = 0,05$ et $\nu = 28$, $t_\alpha = 2,05$ (*test bilatéral*). En définitive, le seuil critique est $\pi = 0,36$.

La relation $|\rho_{calculé}| = 0,487 \succ \pi = 0,36$ permet de conclure au rejet de l'hypothèse H_0, c'est-à-dire à l'existence d'une *association significative* entre X et Y.

• A noter également, pour le calcul de π et pour n grand, l'autre approximation de ρ fournie, sous l'hypothèse nulle H_0, par la **loi normale centrée** de variance $\dfrac{1}{n-1}$, approximation de laquelle résulte le seuil critique $\pi = \dfrac{t_\alpha}{\sqrt{n-1}}$ où t_α vérifie la relation $\Pr ob(|\xi| \ge t_\alpha) = \alpha$. Pour $\alpha = 0,05$, on obtient numériquement $\pi = 0,364$, valeur très proche du seuil précédent fourni par la loi de STUDENT.

3°) Utilisant les formules de correction des rappels de cours, on a pour X, une valeur avec deux ex-aequo. Ainsi $S_X = \frac{30\times(30^2-1)-(2^3-2)}{12} = 2247$. De même, on a pour Y, deux valeurs ex-aequo au rang 2,5, deux valeurs ex-aequo au rang 5,5, trois ex-aequo au rang 8, trois ex-aequo au rang 22.

Il s'ensuit $S_Y = \frac{30\times(30^2-1)-\left[(2^3-2)+(2^3-2)+(3^3-3)+(3^3-3)\right]}{12} = 2242,5$.

Finalement, $\rho = \frac{S_X + S_Y - \sum_{i=1}^{i=n}(R_i - S_i)^2}{2.\sqrt{S_X.S_Y}} = -0,488$. On constate qu'il s'agit d'une valeur quasiment identique au résultat sans correction obtenu précédemment, ce qui montre le faible effet de cette dernière, du moins quand il y a peu d'ex-aequo.

5. Tests à plus de deux échantillons

5.1 Analyse de variance (test ANOVA de FISHER)

Enoncé : Le tableau ci-dessous représente le nombre de km/l parcourus par des automobiles de même modèle utilisant cinq marques différentes de carburants.

Marque A	12	15	14	11	15
Marque B	14	12	15		
Marque C	11	12	10	14	
Marque D	15	18	16	17	14
Marque E	10	12	14	12	

On suppose que l'hypothèse de normalité des consommations en carburants est satisfaite.

1°) Vérifier l'homoscédasticité desdites consommations entre les cinq marques de carburants considérées, et ceci au niveau de signification 5%.

2°) Tester si, aux seuils respectifs $\alpha = 5\%$, puis $\alpha = 1\%$, il y a une différence significative ou non entre au moins deux des marques de carburants en question.

3°) Analysant plus finement les écarts, répondre aux questions suivantes :

a) Les consommations pour les marques A et B sont-elles distinctes ?

b) La marque D diffère-t-elle des autres marques ?

Solution : 1°) Dès lors que l'hypothèse de **normalité** est admise, c'est le **test de BARTLETT** qui est probablement le plus puissant pour tester l'homoscédasticité des consommations entre marques et c'est donc sa mise en œuvre qui est proposée ici, selon les modalités et les notations explicitées en rappels de cours (cf. paragraphe 2.6.b).

On se trouve confronté ainsi à la comparaison de K échantillons d'effectifs respectifs $n_i (1 \le i \le K)$, soit numériquement $K = 5, n_1 = 5, n_2 = 3, n_3 = 4, n_4 = 5, n_5 = 4$, les consommations étant notées quant à elles, par $x_{i,j}, (1 \le i \le K, 1 \le j \le n_i)$.

Dès lors, les étapes à suivre sont :

→ Calculer les *variances corrigées* $S_i^2 = \frac{1}{n_i - 1}.\left[\sum_{j=1}^{j=n_i}(x_{i,j} - \overline{x_i})^2\right]$, soit, pour chacune des populations (marques de carburants), les résultats ci-dessous :

Statistique	Marque				
	A	B	C	D	E
$\overline{x_i} = \frac{\sum_{j=1}^{j=n_i} x_{i,j}}{n_i}$	13,40	13,67	11,75	16	12
$\sum_{j=1}^{j=n_i}(x_{i,j} - \overline{x_i})^2$	13,20	4,67	8,75	10	8
S_i^2	3,30	2,33	2,92	2,50	2,67

→ Calculer la *somme des écarts « intraclasses »*, $S_W = \sum_{i=1}^{i=K}\sum_{j=1}^{j=n_i}(x_{i,j} - \overline{x_i})^2$, soit numériquement $S_W = 44,62$.

→ *Expliciter la statistique de* BARTLETT, soit $B = \frac{(N-K).\ln(\frac{S_W}{N-K}) - \sum_{i=1}^{i=K}(n_i - 1)\ln S_i^2}{1 + \frac{1}{3.(K-1)}.\left[\sum_{i=1}^{i=K}\frac{1}{n_i - 1} - \frac{1}{N-K}\right]}$

avec $N = \sum_{i=1}^{i=K} n_i$, le résultat numérique en résultant étant $B_{calculé} = \frac{16,41 - 16,29}{1 + 0,13} = 0,105$.

→ *Sous l'hypothèse d'homoscédasticité* $H_0 : \sigma_i^2 = \sigma_j^2, \forall (i,j) / i \neq j$, la statistique B suit sensiblement la **loi du chi- deux** à $\nu = K - 1$ degrés de liberté (soit $\nu = 4$, présentement). Il en résulte, au niveau de signification $\alpha = 5\%$, le **seuil critique** $\chi_\alpha^2 = 9,488$ (χ_α^2 vérifiant $Prob(\chi^2 \geq \chi_\alpha^2) = \alpha$, et lu dans la table des valeurs annexée -> cf. loi du chi- deux).

- Autant dire ici que l'homoscédasticité est largement vérifiée puisque $B_{calculé} = 0,105 \prec \chi_\alpha^2 = 9,49$.

2°) Le **test ANOVA de FISHER** appliqué au cas proposé et dont l'utilisation est validée par la satisfaction des hypothèses de *normalité* et d'*homoscédasticité* conduit aux étapes ci-dessous (cf. rappels de cours, paragraphe 3.2.a) :

→ Calculer les sommes des écarts « *intraclasses* », $S_W = \sum_{i=1}^{i=K}\sum_{j=1}^{j=n_i}(x_{i,j} - \overline{x_i})^2$, et des écarts « *interclasses* », $S_B = \sum_{i=1}^{i=K} n_i.(\overline{x_i} - \overline{x})^2$, les valeurs numériques correspondantes étant respectivement, pour les données proposées, $S_W = 44,62$ (cf. 1ère question), et $S_B = 54,22$.

→ *Former la valeur calculée* de la **statistique de FISHER**, $F = \frac{S_B / K-1}{S_W / N-K}$.

Numériquement, $F_{calculé} = \dfrac{54,22/4}{44,62/16} = 4,86$.

→ Relativement à la loi de F qui est la **loi de FISHER SNEDECOR** à $\nu_1 = K-1$ et $\nu_2 = N-K$ degrés de liberté, *déterminer au niveau de signification* α, le **seuil critique** F_α à partir de l'équation $\Pr ob(F \geq F_\alpha) = \alpha$. Suivant lecture dans la table des valeurs annexée (cf. « *fonction de répartition de la loi de FISHER SNEDECOR* », il vient pour $(\nu_1 = 4, \nu_2 = 16, \alpha = 5\%)$ puis $(\nu_1 = 4, \nu_2 = 16, \alpha = 1\%)$, les seuils respectifs $\chi^2_\alpha = 3,01$ et $\chi^2_\alpha = 4,77$.

• Dans les deux cas $\alpha = 5\%$ et $\alpha = 1\%$, on a $F_{calculé} = 4,86 \succ \chi^2_\alpha$, ce qui conduit à retenir l'hypothèse H_1 (*différence significative entre au moins deux des cinq marques de carburants considérées*) et ceci avec un risque d'erreur très faible (c'est-à-dire moins de 1%).

3-a) Utilisant la **méthode des contrastes de SCHEFFE**, la comparaison entre les marques A et B suggère de recourir à la combinaison linéaire $C = m_1 - m_2$. Pour cette dernière et suivant la formule mentionnée en rappels de cours (cf. paragraphe 2.6.a), le contraste calculé, soit $\widehat{C}$, est égal à $\sum_{i=1}^{i=K} C_i.\overline{x_i}$, avec $C_1 = 1, C_2 = -1, C_3 = C_4 = C_5 = 0$. Ainsi, a-t-on $\widehat{C} = \overline{x_1} - \overline{x_2}$ =-0,27.

Par ailleurs, le seuil π à partir duquel le contraste est jugé significatif a pour expression $\pi = \sqrt{(K-1).F_\alpha . \dfrac{S_W}{N-K} . \sum_{i=1}^{i=K} \dfrac{C_i^2}{n_i}}$, expression dans laquelle F_α vérifie, pour la loi de FISHER SNEDECOR, $F(K-1, N-K)$, la relation $\Pr ob(F \geq F_\alpha) = \alpha$.

Numériquement, pour $\alpha = 5\%$, et relativement à la loi $F(4,16)$, il vient successivement $F_\alpha = 3,01$ et $\pi = \sqrt{4 \times 3,01 \times \dfrac{44,62}{16} \times (\dfrac{1}{5} + \dfrac{1}{3})} = 4,23$.

La relation $\left|\widehat{C}\right| = 0,27 \prec 4,23$ montre qu'il n'y a pas de différence significative (et de loin !) entre les marques de carburants A et B au plan de la consommation en carburants.

• Le même exercice appliqué à la comparaison entre la marque D avec la marque E, conduirait de même à $\widehat{C} = 4$ et $\pi = 3,88$, c'est-à-dire, cette fois, à la conclusion de l'hypothèse alternative H_1, à savoir une différence significative de consommation entre les deux marques considérées).

3-b) Enfin, on peut *utiliser différemment les contrastes*, comme, par exemple, pour comparer une marque à l'ensemble des autres. A cet égard, et considérant la marque D dont ion cherche à savoir si elle se différencie des autres marques, on pourra considérer, la combinaison linéaire $m_1 + m_2 + m_3 + m_5 - 4.m_4 = 0$, à laquelle le contraste calculé associé est $\widehat{C} = 4.\overline{x_4} - \overline{x_1} - \overline{x_2} - \overline{x_3} - \overline{x_5}$, soit numériquement $\widehat{C} = 13,18$. Par ailleurs, on a pour π, la valeur $\pi = \sqrt{4 \times 3,01 \times \dfrac{44,62}{16} \times (\dfrac{1}{5} + \dfrac{1}{3} + \dfrac{1}{4} + \dfrac{16}{5} + \dfrac{1}{4})} = 11,92$.

Ici encore, la relation $\widehat{C} = 13,18 \succ \pi$ conduit à retenir l'hypothèse d'une différence significative entre la marque D et les autres marques, au plan de la consommation en carburants.

5.2 Test de KRUSKAL-WALLIS

Au contraire du test ANOVA de FISHER avec lequel il présente beaucoup de similitudes, le test de KRUSKAL-WALLIS est un test non paramétrique qui porte sur les rangs et non sur les valeurs des observations. Il offre l'avantage de ne pas exiger la condition de normalité des distributions ni l'hypothèse d'homoscédasticité, l'indépendance des observations restant par ailleurs une hypothèse nécessaire ici. A défaut de cette dernière, c'est le test de FRIEDMAN (pour échantillons appariés) qu'il faudrait utiliser.

Enoncé : Un psychologne qui gère un foyer pour délinquants juvéniles cherche à montrer qu'il parvient effectivement à réduire la délinquance. Il suit à cet effet 10 jeunes dans son établissement, 10 jeunes délinquants qui vivent chez leurs parents, et 10 jeunes délinquants qui vivent dans une famille adoptive. Il compte les jours d'absentéisme scolaire pour chaque jeune en question, les données ainsi obtenues étant résumées ci-dessous :

Parents	15	18	19	14	5	8	12	13	7	13
Adoptés	16	14	20	22	19	5	17	18	12	18
Foyer	10	13	14	11	7	3	4	18	2	5

1°) Il y a-t-il des différences significatives entre les trois séries de données ci-dessus. au niveau de signification $\alpha = 5\%$?

2°) Si oui, analyser ces différences deux à deux.

Solution : 1°) Se référant aux rappels de cours du présent chapitre (cf. paragraphe 3.2.d), la mise en œuvre du **test de KRUSKAL-WALLIS** pour échantillons indépendants, conduit, à partir des rangs calculés dans un échantillon commun obtenu par regroupement, aux étapes suivantes, les données étant, avec les notations du cours, $K = 3, n_i = 10 (1 \le i \le K), N = 30, \alpha = 5\%$:

→ *Calculer, à partir de l'échantillon regroupé des valeurs* $x_{i,j}$ *classées par ordre croissant, les rangs associés* R_{ij}, le *principe du rang moyen* étant appliqué pour les valeurs ex-aequo. Les calculs par tris successifs conduit sur tableur EXCEL conduisent, pour les rangs R_{ij} en question, aux valeurs :

Parents	5	7,5	9	12,5	14	14	18	20	24,5	27,5
Adoptés	5	12,5	18	21	22	24,5	24,5	27,5	29	30
Foyer	1	2	3	5	7,5	10	11	14	18	24,5

→ Pour chaque classe $i / 1 \le i \le K$, *calculer le rang moyen* $\overline{R_i} = \dfrac{\sum_{j=1}^{j=n_i} R_{ij}}{n_i}$, soit numériquement et à partir des résultats ci-dessus, $\overline{R_1} = 15,2$ - $\overline{R_2} = 21,4$ - $\overline{R_3} = 9,6$.

→ *Calculer le* **rang moyen théorique** $\overline{R}$ qui, sous l'hypothèse H_0 de l'identité entre les K distributions est égal à $\overline{R} = \dfrac{\sum_{i=1}^{i=N} i}{N} = \dfrac{N+1}{2}$, soit numériquement $\overline{R} = 15,5$.

→ *Former la statistique* $H=\frac{12}{N.(N+1)}.\sum_{i=1}^{i=K} n_i.(\overline{R_i}-\overline{R})^2$, qui décrit la somme des écarts entre les $\overline{R_i}$ et $\overline{R}$, soit numériquement $H_{calculé}=8,99$.

→ Compte tenu des valeurs de rangs ex-aequo, *calculer le* **coefficient correctif** $C=1-\frac{\sum_{g=1}^{g=N_g}(t_g^3-t_g)}{N.(N^2-1)}$, où N_g désigne le nombre de valeurs différentes de rangs dans l'échantillon regroupé (ici, $N_g=18$), et t_g le nombre de valeurs ex-aequo pour la valeur $g/1\le g\le N_g$. Ainsi $C=1-\frac{1}{30\times(900-1)}.\left[3\times(3^3-1)+3\times(2^3-1)+(4^3-1)\right]=0,987$.

→ En déduire la *valeur corrigée* $H'=\frac{H}{C}$ de H, soit $H'=9,11$.

→ H suivant la **loi de chi- deux** à $K-1$ degré de liberté et χ_α^2 vérifiant la relation $Prob(\chi^2\ge\chi_\alpha^2)=\alpha$ pour la loi considérée, *en déduire*, par comparaison de H' à χ_α^2, *la règle de décision du test*, la **région critique** en étant caractérisée par $H'\ge\chi_\alpha^2$. Pour le cas présent ($K=3,\alpha=0,05$), on obtient par lecture dans la table de valeurs annexée (cf. loi de χ^2), $\chi_\alpha^2=5,99$.

• Assurément, la relation $H'=9,11\succ\chi_\alpha^2=5,99$ conduit à retenir l'hypothèse H_1, c'est-à-dire une réduction de la délinquance entre les enfants qui vivent dans le foyer considéré et ceux qui vivent chez leurs parents ou dans une famille adoptive.

2°) Menant une analyse plus détaillée des différences deux à deux et se référant aux rappels de cours, paragraphe 3.2.d, l'opposition, par exemple, entre «*foyer*» et «*parents*» conduit à comparer $\left|\overline{R_1}-\overline{R_3}\right|=\left|15,2-9,6\right|=5,6$ à $\pi_{1,3}=t_{\alpha'}.\sqrt{\frac{n.(n+1)}{12}.(\frac{1}{n_1}+\frac{1}{n_3})}$ où $n_1=n_3=10, n=30, K=3, \alpha=0,05, \alpha'=\frac{\alpha}{K.(K-1)}=0,083$. Quant à $t_{\alpha'}$, il vérifie relativement à la loi normale centrée réduite, ξ, $Prob(|\xi|\ge t_{\alpha'})=\alpha'$.

En définitive, et par lecture dans la table de valeurs annexée portant sur la loi normale centrée réduite, on a $t_{\alpha'}=2,39$, ce qui entraîne $\pi_{1,3}=9,17$. La relation $\left|\overline{R_1}-\overline{R_3}\right|=5,6\prec\pi_{1,3}=9,17$ ne permet pas de conclure à une différence significative de la délinquance entre les groupes 1 et 3.

• Il en est tout autrement, par contre, si on compare les groupes 2 et 3, soient «*adoptés*» et «*foyer*». Le seuil π reste inchangé ($\pi_{1,3}=\pi_{2,3}$) puisque les effectifs $n_i(1\le i\le 3)$, sont tous égaux à 10, et la différence $\left|\overline{R_2}-\overline{R_3}\right|$ a pour valeur $\left|21,4-9,6\right|=11,8$.

On a donc, cette fois, $\left|\overline{R_2}-\overline{R_3}\right|=11,8\succ\pi_{2,3}=9,17$, ce qui permet de conclure à une différence significative entre les deux groupes en question.

5.3 Test de la médiane généralisée

Portant sur des variables quantitatives, ce test qui s'applique à K groupes d'où sont extraits K échantillons de tailles égales ou non, vise à se prononcer sur l'égalité ou non des médianes entre lesdits groupes, formant ainsi une généralisation du test de la médiane de MOOD. Son mode opératoire est décrit dans le cadre de l'illustration proposée ci-dessous.

Enoncé : Un chercheur dans un centre de santé publique veut étudier l'influence du degré d'instruction de la mère sur le soin avec lequel elle assure la surveillance médicale de son enfant. A cet effet, il considère le niveau maximum de culture atteint par la mère et ceci à travers son diplôme le plus élevé, et il associe à cela, par ailleurs, le nombre de visites médicales de contrôle effectuées pour l'enfant durant ses deux premières années.

L'étude a été faite à partir de 44 mères tirées au sort parmi les naissances enregistrées dans une maternité durant une période donnée, six niveaux d'instruction étant considérés. Les résultats obtenus sont :

Ecole élémentaire	Collège	Lycée (bac)	1er cycle universitaire	Licence	Maîtrise
4	2	2	9	2	2
3	4	0	4	4	6
0	1	4	2	5	
7	6	3	3	2	
1	3	8			
2	0	0			
0	2	5			
3	5	2			
5	1	1			
1	2	7			
	1	6			
		5			
		1			

S'inspirant de l'application 4.5 du présent chapitre et de l'exercice 16 ci-après (cf. exercices complémentaires), il est proposé, pour tester l'hypothèse H_0 : « *il n'y a pas de différence significative entre les nombres de visites de contrôle en fonction du degré d'instruction de la mère* » et l'hypothèse contraire H_1, de répondre successivement aux questions suivantes :

1°) Calculer la médiane générale M du nombre de visites de contrôle pour l'ensemble des personnes étudiées.

2°) Dresser un tableau de contingences $(2,6)$ portant d'une part, sur le nombre de mères dont la fréquence des visites de contrôle pour leur enfant dépasse (resp. est inférieur à) M, et d'autre part, sur les niveaux d'études.

3°) Mettre en regard les effectifs O_{ij} observés ci-dessus, et les effectifs théoriques E_{ij} qui résultent de l'hypothèse H_0 d'égalité des médianes entre les groupes.

4°) Calculer la distance de chi- deux décrivant les écarts entre les O_{ij} et les E_{ij}, soit $\chi^2_{calculé} = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$, ceci après avoir effectué les regroupements utiles éventuels.

5°) Identifiant la loi de χ^2 précédente (nombre de degrés de liberté), conclure quant à la décision à retenir ici, le niveau de signification choisi étant fixé à 5%).

Solution : 1°) Pour rappel, *les données ayant préalablement été regroupées et classées par ordre croissant*, soient $x_{(i)}, 1 \le i \le 44$, la **médiane** est définie lorsque n est pair (c'est le cas ici), par $M = \frac{1}{2}.\left[x_{(n/2)} + x_{(n+1/2)}\right]$. Numériquement, on obtient pour les données proposées, la valeur $M = \frac{x_{(22)} + x_{(23)}}{2} = 2,5$.

2°) Pour chacun des niveaux de diplômes considérés, les **effectifs observés** des mères dont les visites de consultation de pédiatrie sont supérieurs à M (resp. inférieurs), sont rassemblés dans le tableau de contingences ci-après, soient O_{ij}.

	Niveaux de diplômes					
	Elémentaire	Collège	bac	1er cycle	licence	Maîtrise
Nombre de mères dont la fréquence des visites est supérieure à M	5 ***5***	4 ***5,5***	7 ***6,5***	3 ***2***	2 ***2***	1 ***1***
Nombre de mères dont la fréquence des visites est inférieure à M	5 ***5***	7 ***5,5***	6 ***6,5***	1 ***2***	2 ***2***	1 ***1***

3°) Parallèlement, sous l'hypothèse $H_0 : M_1 = M_2 = ... = M_6 = M$, les **effectifs théoriques,** dont le calcul est immédiat, sont portés ci-dessus en italique et caractères gras, soient E_{ij}.

4°) La *condition de validité du test de chi- deux* suivant laquelle sont seules significatives *les classes d'effectifs théoriques supérieurs à* 5 (cf. rappels de cours, paragraphe 3.1.a), n'est pas satisfaite ici pour les trois niveaux de diplômes « *1er cycle* », « *licence* », et « *maîtrise* ». C'est donc un **regroupement** sous l'appellation « *études universitaires* » qui est suggéré, le nouveau tableau de calculs qui en résulte étant :

	Niveaux de diplômes			
	élémentaire	collège	bac	université
Nombre de mères dont la fréquence des visites est supérieure à M	5 ***5***	4 ***5,5***	7 ***6,5***	6 ***5***
Nombre de mères dont la fréquence des visites est inférieure à M	5 ***5***	7 ***5,5***	6 ***6,5***	4 ***5***

4°) La **distance** de χ^2 entre les *effectifs observés* O_{ij} et les *effectifs théoriques* E_{ij}, a pour expression $\chi^2 = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$, ce qui conduit, pour les données ci-dessus, à la valeur calculée $\chi^2_{calculé} = \frac{(5-5)^2}{5} + \frac{(5-5)^2}{5} + \frac{(4-5,5)^2}{5,5} + ... + \frac{(4-5)^2}{5} = 1,295$.

5°) Pour le cas d'un tableau de contingences (r,k), la variable de χ^2 correspondante présente un nombre de degrés de liberté égale à $\nu = (r-1).(k-1)$ (cf. paragraphe 3.4.c des rappels de cours). Ainsi, numériquement, $\nu = (2-1)\times(4-1) = 3$.

Or, au niveau de signification $\alpha = 5\%$, le seuil χ^2_α qui vérifie $Prob(\chi^2 \geq \chi^2_\alpha) = \alpha$ est égal, pour $\nu = 3$, à $\chi^2_\alpha = 7{,}815$ (cf. lecture dans la table annexée relative à la loi du chi- deux). La relation $\chi^2_{calculé} = 1{,}295 \prec \chi^2_\alpha = 7{,}815$ conduit à ne pas pouvoir rejeter l'hypothèse H_0 pour l'exemple proposé.

5.4 Test de FRIEDMAN appliqué à un problème d'ergonomie

Egalement non paramétrique et portant sur les rangs, le test de FRIEDMAN qui est un test d'identité ou non des distributions desquelles on a extrait, pour chacune, un échantillon correspondant, se différencie du test de KRUSKAL-WALLIS par le fait que les données sont appariées par blocs (c'est le même élément de l'échantillon qu'on soumet aux différents traitements étudiés, ce qui suppose l'égalité des tailles des échantillons pour les K populations considérées).

Enoncé : Un ergonome désire étudier la forme la plus économique pour un orifice dans lequel des ouvriers doivent faire passer une fiche. Ils comparent ainsi cinq formes d'orifices de moins en moins évasés. Grâce à un appareillage avec cellule photo- électrique, il mesure en millièmes de seconde, le temps mis par un ouvrier pour mettre la fiche en position dans l'orifice.

Chaque sujet effectue plusieurs essais et l'ergonome note pour chacun, le temps médian. Les résultats ainsi obtenus pour 7 sujets sont :

Sujet	Forme 1	Forme 2	Forme 3	Forme 4	Forme 5
1	244	417	178	195	452
2	235	307	225	346	613
3	308	290	257	427	438
4	343	305	290	215	534
5	254	263	252	340	469
6	251	291	417	263	445
7	333	414	414	276	441

Etudier à l'aide du test de FRIEDMAN si les médianes qui correspondent aux cinq formes d'orifices considérées sont égales ou non.

Solution : Se référant aux rappels de cours du présent chapitre (cf. paragraphe 3.3.e), la mise en œuvre du **test de FRIEDMAN** conduit aux étapes suivantes :

→ *Déterminer par ligne* (élément $i, 1 \leq i \leq n$, de l'échantillon), *les rangs* R_{ij} qui correspondent aux diverses valeurs observées x_{ij} pour les traitements étudiés ($1 \leq i \leq K$), le principe du rang moyen étant utilisé en cas de valeurs ex-aequo.

Pour les données proposées ($n = 7, K = 5$), il s'ensuit :

Sujet	Forme 1	Forme 2	Forme 3	Forme 4	Forme 5
1	3	4	1	2	5
2	2	3	1	4	5
3	3	2	1	4	5
4	4	3	2	1	5
5	2	3	1	4	5
6	1	3	4	2	5
7	2	3,5	3,5	1	5

→ *Calculer, pour chacun des traitements « j », le rang moyen* $\overline{R_j} = \frac{\sum_{i=1}^{i=n} R_{ij}}{n}$, soit numériquement :

j	1	2	3	4	5
$\overline{R_j}$	2,43	3,07	1,93	2,57	5

→ *Former la* **statistique de FRIEDMAN** définie par $F = \frac{12.n}{K.(K+1)} . \sum_{j=1}^{j=K} (\overline{R_j} - \overline{R})^2$, avec $\overline{R} = \frac{K+1}{2}$, soit numériquement et pour $K = 5$, la valeur $F_{calculé} = 15,86$.

→ *Appliquer le facteur correctif* $C = 1 - \frac{\sum_{i=1}^{i=n} \sum_{g=1}^{g=G_i} (t_{i,g}^3 - t_{i,g})}{n.(K^3 - K)}$ pour les valeurs ex-aequo, ce qui se résume ici, à un seul terme, $C = 1 - \frac{(2^3 - 2)}{7.(5^3 - 5)} = 0,993$. Ainsi après correction, la valeur calculée à considérer est-elle $F'_{calculé} = \frac{F}{C} = 15,97$.

→ Admettant (pour n assez grand), que F suit, sous l'hypothèse nulle H_0, la **loi du chi- deux** à $K-1$ degrés de liberté, déterminer au niveau de signification α (on prendra ici $\alpha = 5\%$), le seuil χ_α^2 vérifiant $\Pr ob(\chi^2 \geq \chi_\alpha^2) = \alpha$. Par lecture dans la table de valeurs annexée et pour $\alpha = 5\%, \nu = K - 1 = 4$, on a immédiatement $\chi_\alpha^2 = 9,49$.

• La **région critique** du test étant caractérisée par $F \geq \chi_\alpha^2$, c'est l'hypothèse H_1 qu'on peut assurément retenir ici puisque $F_{calculé} = 15,97 \succ 9,49$; c'est-à-dire la conclusion d'une *différence significative* entre au moins deux des cinq valeurs médianes qui correspondent aux formes d'orifices étudiées.

• A noter que, même au seuil $\alpha = 1\%$, c'est l'hypothèse H_1 qui continue à prévaloir puisqu'on a alors $\chi_\alpha^2 = 13,27$ et donc toujours, $F_{calculé} = 15,97 \succ \chi_\alpha^2 = 13,27$.

• Enfin, comme pour le test de KRUSKAL-WALLIS (cf. application 5.2), on peut analyser plus précisément la nature des différences mises en évidence par le choix de H_1. Ainsi considérant l'erreur de première espèce réduite $\alpha' = \frac{\alpha}{K.(K-1)}$ (cf. rappels de cours, paragraphe 3.3.e), la confrontation des formes 3 et 5, conduit à comparer $|\overline{R_3} - \overline{R_5}| = 3,07$ à $\pi = t_{\alpha'} \sqrt{\frac{K.(K+1)}{6.n}}$ où $t_{\alpha'}$ vérifie, relativement à la variable normale centrée réduite $\xi : N(0,1)$, la relation $\Pr ob(|\xi| \geq t_{\alpha'}) = \alpha'$.

Il en résulte, successivement, $\alpha' = 0,0025$, puis $t_{\alpha'} = 2,81$, et $\pi = 2,01$. La relation $|\overline{R_3} - \overline{R_5}| = 3,07 \succ \pi = 2,01$ conduit à retenir la conclusion d'une différence significative entre les formes 3 et 5.

Plus généralement, les différences $\left|\overline{R_i} - \overline{R_j}\right|$ sont calculées ci-dessous :

Forme	1	2	3	4	5
1		0,64	0,50	0,14	2,57
2			1,14	0,50	1,93
3				0,64	3,07
4					2,43
5					

Les différences deux à deux significatives (supérieures au seuil critique $\pi = 2,01$) sont donc $(1,5),(3,5)$ et $(4,5)$.

5.5 Comparaisons sur échantillons liés et données binaires (test de COCHRAN)

Enoncé : Une interview est conduite auprès de 18 personnes, la question posée étant «entre deux médicaments donnés A et B, lequel des deux médicaments préférez-vous pour supprimer vos maux de tête ? ». Les réponses sont ainsi de type binaire, et peuvent par exemple, être codée, par 1 s'il s'agit du médicament A, et par 0 si le choix est le médicament B.

Pour chaque personne, trois interviews sont conduites dans des circonstances distinctes :

- interview 1→ la question est posée avant une campagne publicitaire portant sur le médicament A ;
- interview 2→ la question est posée après ladite campagne publicitaire ;
- interview 3→ la question est posée après un accident ayant mis en évidence les risques potentiels du médicament A.

Les données sont résumées dans le tableau ci-après :

Personne	Interview 1	Interview 2	Interview 3
1	0	0	0
2	1	1	0
3	0	1	0
4	0	0	0
5	1	0	0
6	1	1	0
7	1	1	0
8	0	1	0
9	1	0	0
10	0	0	0
11	1	1	1
12	1	1	1
13	1	1	0
14	1	1	0
15	1	1	0
16	1	1	1
17	1	1	0
18	1	1	0

La préférence des utilisateurs est-elle impactée ou non par les circonstances considérées précédemment ?

Solution : S'agissant de K groupes appariés ($K=3$) puisque portant sur les mêmes personnes d'un échantillon de taille n ($n=18$), les données étant par ailleurs binaires, c'est le **test non paramétrique de COCHRAN** qui est suggéré ici pour répondre à la question posée, test dont la mise en œuvre est décrite en rappels de cours du présent chapitre (cf. paragraphe 3.3.d).

Le calcul des *sommes en lignes* ($L_i, 1 \le i \le n$) *et en colonnes* ($C_j, 1 \le j \le K$), des données collectées conduit au tableau ci-dessous :

Personne	Groupe 1	Groupe 2	Groupe 3	Σ (L_i)
1	0	0	0	0
2	1	1	0	2
3	0	1	0	1
4	0	0	0	0
5	1	0	0	1
6	1	1	0	2
7	1	1	0	2
8	0	1	0	1
9	1	0	0	1
10	0	0	0	0
11	1	1	1	3
12	1	1	1	3
13	1	1	0	2
14	1	1	0	2
15	1	1	0	2
16	1	1	1	3
17	1	1	0	2
18	1	1	0	2
Σ (C_j)	13	13	3	29

La **statistique de COCHRAN** est définie par $Q = K.(K-1).\dfrac{\sum_{j=1}^{j=K}(C_j - S/K)^2}{\sum_{i=1}^{i=n} L_i.(K-L_i)}$, avec $S = \sum_{i=1}^{i=n} L_i = \sum_{j=1}^{j=K} C_j$. On obtient donc aisément, après calculs, la valeur calculée à partir des données ci-dessus, soit $Q_{calculé} = 16,67$.

Or, lorsque l'hypothèse H_0 est vraie et que les conditions $n \ge 4$ et $n.K \ge 24$ sont remplies (c'est le cas ici puisque $n=18$ et $n.K=54$), il a été montré que Q suivait *asymptotiquement*, la **loi du chi- deux** à $K-1$ degrés de liberté. Dès lors, notant par χ^2_α le seuil vérifiant, pour la loi $\chi^2(K-1)$ en question, la relation $Prob(\chi^2 \ge \chi^2_\alpha) = \alpha$, la **région critique** du test est caractérisée immédiatement par $Q \ge \chi^2_\alpha$.

Pour le niveau de signification classique de 5%, on lit dans la table de valeurs annexée de la loi de chi- deux à $\nu = 3-1 = 2$ degrés de liberté, $\chi^2_\alpha = 5,99$.

La relation $Q_{calculé} = 16,67 \succ \chi^2_\alpha = 5,99$ conduit assurément à retenir ici l'hypothèse H_1 qui est celle d'une différence significative entre au moins deux des trois groupes considérés, c'est-à-dire un impact manifeste des circonstances sur la préférence des utilisateurs pour le médicament A ou le médicament B.

C'est même vrai pour $\alpha = 1\%$ puisque, pour ce dernier niveau de signification, on a $\chi^2_\alpha = 9,21$.

C - Exercices complémentaires

1. Un radar actif de surveillance aérienne a des caractéristiques telles qu'une éventuelle cible réfléchit 20 impulsions lors d'un balayage. A l'aide d'un traitement adapté, ces n impulsions réfléchies, en cas de présence de la cible, fournissent un vecteur d'observations $(z_i), 1 \le i \le n$, avec :

- $H_0 : z_i = b_i$, en l'absence de cible ;
- $H_1 : z_i = A + b_i$, en présence de cible ;

les b_i désignant des variables aléatoires gaussiennes $N(0,\sigma^2)$, indépendantes et modélisant les divers bruits.

1°) A l'aide de la méthode de NEYMAN et PEARSON, construire le test entre les deux hypothèses H_0 et H_1, et exprimer la règle de décision dans les conditions numériques $A = 0,7; \alpha = 10^{-4}$, et $\sigma = 0,6$.

2°) Pour le test considéré, exprimer la valeur de l'erreur de 2ème espèce, β. Interpréter le résultat obtenu.

Solution : 1°) On raisonne ici à partir d'échantillons $(z_1, z_2, ..., z_n)$ de taille n (numériquement $n = 20$), les variables *indépendantes* en question, soient Z_i, ayant pour *loi parente*, la variable gaussienne centrée $N(0,\sigma^2)$ lorsque l'hypothèse H_0 est vraie, et la variable gaussienne de moyenne A et de variance σ^2, soit $N(A,\sigma^2)$, lorsque c'est l'hypothèse H_1 qui est vérifiée.

Il est immédiat en effet, que si $Z_i = A + B_i$ où B_i est de loi $N(0,\sigma^2)$, Z_i est gaussienne avec en outre, $E(Z_i) = A + E(B_i) = A$ et $Var(Z_i) = Var(B_i) = \sigma^2$.

Dans ces conditions, le test proposé, se ramène, relativement à la moyenne m de la loi « *parente* », $Z : N(m,\sigma^2)$, à tester $H_0 : m = 0$ contre $H_1 : m = A$.

D'après les éléments des rappels de cours (cf. paragraphe 2.3.a), la **région critique** a pour forme $\bar{z} \ge K$, avec $\alpha = \text{Prob}\, H_1 / H_0$ vraie), soit $\alpha = \Pr ob(\bar{z} \ge K / m = 0)$.

Considérant la variable normale centrée réduite associée à $\bar{z}$, soit $\xi = \frac{\bar{z}-0}{\sigma/\sqrt{n}}$, il vient immédiatement la relation, $\Pr ob(\xi \geq \frac{K}{\sigma/\sqrt{n}}) = \alpha$, soit en désignant par t_α, le nombre qui vérifie $\Pr ob(\xi \geq t_\alpha) = \alpha$, l'expression $K = t_\alpha . \frac{\sigma}{\sqrt{n}}$.

Numériquement, on a $\alpha = 10^{-4}; t_\alpha = 3,62; \sigma = 0,6; n = 20$, d'où $K = 0,48$. Concrètement, α qui est la probabilité de détecter la présence d'une cible sachant qu'il n'y en a pas, est donc la fréquence théorique des fausses détections.

2°) L'erreur β est définie par Prob(décider H_0 / H_1 vraie), soit $\Pr ob(\bar{z} \prec 0,48 / m = A)$. Utilisant la variable ξ, normale, centrée, réduite, associée à $\bar{z}$, soit $\xi = \frac{\bar{z}-A}{\sigma/\sqrt{n}}$, il en résulte $\beta = \Pr ob(\xi \prec \frac{0,48-0,7}{0,6/\sqrt{20}} \approx -1,64) = 0,051$. Concrètement, β représente ici la fréquence théorique des détections manquées.

2. L'écart- type de la teneur d'un composant dans un médicament est de 8 milligrammes. Un nouveau procédé de fabrication vise à diminuer cet écart-type. Pour 10 mesures de la teneur en question sur des unités fabriquées suivant le nouveau procédé, on obtient (en mg) :

726 – 725 – 722 – 727 – 718 – 723 – 731 – 719 – 724 - 726

On suppose que les mesures sont des variables aléatoires normales, identiquement distribuées, et indépendantes.

Le but recherché est-il atteint ? (on raisonnera ici à partir d'une erreur de première espèce fixée à la valeur 10%).

Solution : 1°) A partir de l'échantillon de taille $n = 10$ susmentionné, il s'agit de tester $\begin{cases} H_0 : \sigma = \sigma_0 \\ H_1 : \sigma \prec \sigma_0 \end{cases}$, avec $\sigma_0 = 8$. L'hypothèse de *normalité* permet ici de se placer dans les conditions du **test de conformité** exposé en rappels de cours du présent chapitre (cf. paragraphe 2.3.c).

La moyenne m de cette teneur X étant inconnue, la **région critique** du test a pour forme $W = \left\{ (x_1, x_2, ..., x_n) / \hat{S}^2 \leq \pi / \sigma = \sigma_0 \right\}$, avec $\hat{S}^2 = \frac{1}{n-1} . \sum_{i=1}^{i=n} (x_i - \bar{x})^2 = \frac{1}{n-1} . \left[\sum_{i=1}^{i=n} x_i^2 - n.\bar{x}^2 \right]$.

La variable $V = \frac{(n-1).\hat{S}^2}{\sigma^2}$ suivant alors la loi du chi- deux à $n-1$ degrés de liberté, soit $\chi^2(n-1)$, il s'ensuit de la donnée de l'erreur de première espèce α :

$$\alpha = \Pr ob(\hat{S}^2 \leq \pi / \sigma = \sigma_0) = \Pr ob(V \leq \frac{(n-1).\pi}{\sigma_0^2})$$

Notant par χ_α^2 le nombre vérifiant $\Pr ob(V \leq \chi_\alpha^2) = \alpha$, il en résulte $\pi = \chi_\alpha^2 . \frac{\sigma_0^2}{n-1}$.

Numériquement, un calcul sur EXCEL conduit aux résultats $\bar{x} = 724,1; \hat{S}^2 = 14,76$; $\chi^2_\alpha = 4,168$; et $\pi = 29,64$. On retient que $\hat{S}^2 = 14,76 \prec \pi = 29,64$, ce qui, avec un risque d'erreur inférieur à 10%, permet de conclure à l'efficacité du nouveau procédé.

3. On considère un échantillon de n valeurs indépendantes d'une variable aléatoire X, soit $(x_1, x_2, ..., x_n)$ à partir duquel on souhaite tester, relativement à la loi $P(X)$ suivie par X, les hypothèses :

- H_0 : $P(X)$ est la loi uniforme sur $[0,1]$, c'est-à-dire de densité de probabilité $f_0(x) = 1_{[0,1]}(x)$;

- H_1 : $P(X)$ est la loi Bêta, $B(2,1)$, c'est-à-dire de densité de probabilité $f_1(x) = 2x.1_{[0,1]}(x)$.

1°) Exprimant le rapport des vraisemblances, proposer à l'aide du théorème de NEYMAN et PEARSON, un test de puissance maximale dont on caractérisera la région critique W.

2°) A partir de la statistique appropriée dont on déterminera la loi sous l'hypothèse H_0, en déduire suivant l'erreur de première espèce α donnée, la détermination de la règle de décision du test en question.

Solution : 1°) Sous l'hypothèse H_0, la *fonction de vraisemblance* $L_0(x_1, x_2, ..., x_n)$ associée à l'échantillon $(x_1, x_2, ..., x_n)$ est égale au produit des densités de probabilité, soit en l'occurrence.

$$L_0(x_1, x_2, ..., x_n) = \prod_{i=1}^{i=n} 1_{[0,1]}(x_i).$$

De même, sous l'hypothèse H_1 a-t-on $L_1(x_1, x_2, ..., x_n) = 2^n . \prod_{i=1}^{i=n} 1_{[0,1]}(x_i).x_i$.

• En définitive, le **rapport de vraisemblance** $\frac{L_0}{L_1}$ est égal à $2^n . \prod_{i=1}^{i=n} x_i$. La **région critique** du test, soit, $W = \left\{ (x_1, x_2, ..., x_n) / \frac{L_0}{L_1} \le k \right\}$, a donc pour forme $W = \left\{ (x_1, x_2, ..., x_n) / \frac{1}{2^n . \prod_{i=1}^{i=n} x_i} \le k \right\}$, ce qui en introduisant les log- vraisemblances, s'écrit :

$$W = \left\{ (x_1, x_2, ..., x_n) / -n.\ln 2 - \sum_{i=1}^{i=n} \ln x_i \le \ln k \right\}.$$

En résumé, cette région critique a pour forme $-\sum_{i=1}^{i=n} \ln x_i \le K$, avec $K = \ln k + n.\ln 2$.

2°) Sous l'hypothèse H_0, les variables X_i sont uniformes sur $[0,1]$. Posant $Y_i = -\ln X_i, 1 \le i \le n$, ce qui entraîne $X_i = e^{-Y_i}$, il est immédiat que les variables Y_i ont pour densités de probabilité, les fonctions $g(y_i) / g(y_i).dy_i = e^{-y_i}.dy_i$ où $y_i \in [0, +\infty[$ (ceci d'après le *théorème de la mesure image*). Bref, les variables aléatoires Y_i $(1 \le i \le n)$, sont de type **exponentiel** (de paramètre 1).

Revenant à la région critique W et à la statistique $-\sum_{i=1}^{i=n} \ln x_i$, cette dernière suit la **loi Gamma- *n*** de paramètre 1 puisque somme de n variables aléatoires exponentielles indépendantes.

La détermination de la règle de décision, à partir de l'erreur de première espèce α, conduit pour la loi Q en question à déterminer le *quantile d'ordre* α, c'est-à-dire le nombre $Q_\alpha / prob(Q \le Q_\alpha) = \alpha$ (soit par tables de valeurs, soit par calcul, ce qui est plus compliqué). Il s'ensuit, la **règle de décision** à savoir, le choix de H_1 si $Q_{calculé} \le Q_\alpha$ et le choix de l'hypothèse nulle H_0 dans le cas contraire (il est rappelé que $Q_{calculé} = -\sum_{i=1}^{i=n} \ln x_i$).

L'avantage du test du rapport de vraisemblance est qu'il garantit la meilleure puissance possible d'après le théorème de NEYMAN et PEARSON.

4. On considère n variables aléatoires $(X_1, X_2, ..., X_n)$ indépendantes et équidistribuées de loi parente exponentielle, c'est-à-dire de densité de probabilité définie par $f_\theta(x) = \frac{1}{\theta}.\exp(-\frac{x}{\theta}).1_{[0,+\infty[}(x)$.

Etant donnée une valeur θ_0 fixée de θ, on souhaite tester $\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta \succ \theta_0 \end{cases}$.

1°) Suivant le théorème de NEYMAN et PEARSON, donner la forme de la région critique correspondante.

2°) Expliciter cette région critique, au niveau α (erreur de première espèce).

3°) Qu'en est-il lorsque n est grand ?

4°) Application numérique : $n = 10, \theta_0 = 1200, \alpha = 0{,}05$.

Solution : 1°) La *fonction de vraisemblance* associée à un n - échantillon est égale, pour la loi considérée, à $L(X_1, X_2, ..., X_n, \theta) = \prod_{i=1}^{i=n} f_\theta(X_i) = \frac{1}{\theta^n}.\exp(-\frac{\sum_{i=1}^{i=n} X_i}{\theta})$, avec $X_i \ge 0, 1 \le i \le n$.

Le **théorème de NEYMAN et PEARSON** qui porte sur le *rapport des vraisemblances* conduit donc à la **région critique** définie par $W = \left\{ (x_1, x_2, ..., x_n) / \frac{L(x_1, x_2, ..., x_n, \theta_0)}{L(x_1, x_2, ..., x_n, \theta_1)} \le k \right\}$, soit par passage à la log- vraisemblance la relation $n.\ln\frac{\theta_1}{\theta_0} - (\frac{1}{\theta_0} - \frac{1}{\theta_1}).\sum_{i=1}^{i=n} x_i \le \ln k$.

Dans le cadre du test proposé, les valeurs θ_1, sous l'hypothèse H_1, sont supérieures à θ_0. La condition précédente s'écrit en conséquence, $-n.\ln\frac{\theta_1}{\theta_0} + (\frac{1}{\theta_0} - \frac{1}{\theta_1}).\sum_{i=1}^{i=n} x_i \ge -\ln k$, soit une relation de la forme $\sum_{i=1}^{i=n} x_i \ge K$, avec $K = \left[-\ln k + n.\ln\frac{\theta_1}{\theta_0} \right].\frac{\theta_0.\theta_1}{\theta_1 - \theta_0}$.

2°) Pour cette question, il est proposé de se reporter aux résultats de l'application 2.2 du présent chapitre relative au modèle de RAYLEIGH.

D'après lesdits résultats, la somme $Z = \frac{\sum_{i=1}^{i=n} X_i}{\theta}$ suit la *loi Gamma – n* de densité de probabilité $\frac{e^{-z}.z^{n-1}}{(n-1)!}$, la variable 2.Z étant quant à elle, une *variable de* χ^2 *à* $2n$ *degrés de liberté*. En définitive, la **région critique** du test $\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta \succ \theta_0 \end{cases}$, définie à partir de l'erreur de 1ère espèce et l'équation $\Pr ob(\sum_{i=1}^{i=n} X_i \geq K / \theta = \theta_0)$ est déterminée par la relation :

$$\Pr ob(\frac{2}{\theta}.\sum_{i=1}^{i=n} X_i = \chi^2(2n) \geq \frac{2.K}{\theta} / \theta = \theta_0) = \alpha .$$

Notant par χ_α^2 le nombre qui vérifie $\Pr ob(\chi^2(2n) \geq \chi_\alpha^2) = \alpha$, il en résulte le seuil critique $K = \frac{\theta_0.\chi_\alpha^2}{2}$. Ainsi, si relativement aux données proposées, la somme calculée $\sum_{i=1}^{i=n} x_i$ est supérieure ou égale à K, on décide H_1, l'hypothèse nulle H_0 étant retenue dans le cas contraire.

• H_1 ayant la forme d'une *hypothèse multiple*, il convient de remarquer que l'erreur de 2ème espèce β (et à fortiori, la *puissance du test*), sont des fonctions de θ, dont on peut cependant exprimer un minorant.

En effet, $\beta = \Pr ob(\sum_{i=1}^{i=n} x_i \prec K / \theta \succ \theta_0)$ où, cette fois, c'est $2.\frac{\sum_{i=1}^{i=n} X_i}{\theta}$ qui suit la loi $\chi^2(2n)$.

Il en résulte $\beta = \Pr ob(\chi^2(2n) \prec \frac{2.K}{\theta})$. θ étant minoré par θ_0, β est donc majoré par $\beta_0 = \Pr ob(\chi^2(2n) \prec \frac{2.K}{\theta_0})$, d'où, pour ce qui est de la puissance $\eta = 1 - \beta$ du test considéré, le *minorant* $1 - \beta_0$.

3°) Lorsque n est *grand* $(n \geq 30)$, on pourra, au-delà de la convergence de la loi de χ^2 vers la loi normale, admettre le *théorème central limite* et le caractère **normal** de la somme $\sum_{i=1}^{i=n} X_i$ dont il est immédiat qu'elle a pour moyenne $\sum_{i=1}^{i=n} E(X_i) = n.\theta$ et pour variance $\sum_{i=1}^{i=n} Var(X_i) = n.\theta^2$.

Le **seuil critique** du test, caractérisé par $\Pr ob(\sum_{i=1}^{i=n} x_i \geq K / \theta = \theta_0) = \alpha$, soit par passage à la variable normale centrée réduite ξ associée, $\Pr ob(\xi \geq \frac{K - n.\theta_0}{\theta_0.\sqrt{n}}) = \alpha$, a donc pour expression $K = n.\theta_0 + t_\alpha.\theta_0.\sqrt{n}$ où t_α vérifie $\Pr ob(\xi \geq t_\alpha) = \alpha$.

4°) Supposant $n = 10, \theta_0 = 1200, \alpha = 5\%$, on a successivement $\chi_\alpha^2 = 31,41$ (lecture dans table annexée relative à la loi du χ^2 à 20 degrés de liberté), et $K = 18846$.

Ainsi, pour l'exemple considéré, l'hypothèse H_0 doit-elle être rejetée lorsque la somme $\sum_{i=1}^{i=n} x_i$ des observations effectuées est supérieure à 18.846 (acceptation de H_0 dans le cas contraire).

Quant au minorant de la puissance du test, c'est $1-\beta_0$, avec $\beta_0 = \Pr ob(\chi^2(2n) \prec \frac{2.K}{\theta_0} = 31,41) = 1-\alpha = 0,95$. On a donc $\eta \geq 0,05$.

5. Englobant la loi de RAYLEIGH, la famille des lois de WEIBULL, caractérisée par la densité de probabilité $f(x,\theta,\alpha) = \alpha.\theta.e^{-\theta.x^\alpha}.x^{\alpha-1}.1_{[0,+\infty[}(x)$, trouve des applications intéressantes, notamment en théorie de la fiabilité, la souplesse offerte par la multiplicité des paramètres, permettant de recouvrir une grande diversité de lois courantes (loi exponentielle lorsque $\alpha = 1$, loi de RAYLEIGH lorsque $\alpha = 2,\ldots$).

On considère dans ce problème le cas $\alpha = 3$, c'est-à-dire une variable aléatoire X dont la densité de probabilité est $f(x,\theta) = \frac{3}{\theta}.x^2.\exp(-\frac{x^3}{\theta}).1_{[0,+\infty[}(x), \theta \succ 0$.

A l'aide d'un échantillon $(X_1, X_2, ..., X_n)$ de loi parente X, on veut tester $\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta = \theta_1 \end{cases}$, avec $\theta_1 \succ \theta_0$.

1°) A l'aide du théorème de NEYMAN et PEARSON, exprimer la statistique T_n du test le plus puissant et expliciter la région critique dudit test.

2°) Après avoir vérifié que $Y = \frac{2}{\theta}.X^3$ suit une loi de chi- deux à 2 degrés de liberté, exprimer le seuil critique du test en fonction de l'erreur de première espèce α, supposée connue.

3°) On suppose que $\theta_0 = \frac{1}{2}, \theta_1 = 1, n = 10, \alpha = 5\%$, et $\sum_{i=1}^{i=10} x_i^3 = 18$. Qu'en conclure ? Quelle est la puissance du test ainsi construit ?

4°) Pour tester si une variable X possède la densité de probabilité définie par $f(x,\theta) = \frac{3}{\theta}.x^2.\exp(-\frac{x^3}{\theta}).1_{[0,+\infty[}(x), \theta \succ 0$, on considère un échantillon de n variables aléatoires indépendantes de loi parente X, soit $(X_1, X_2, ..., X_n)$, et on teste si la loi de $Y = \frac{2}{\theta}.X^3$ est une loi de chi-deux à deux degrés de liberté, soit $\chi^2(2)$.

En d'autres termes, on considère le test d'ajustement :

- H_0 : Y suit la loi $\chi^2(2)$;
- H_1 : Y ne suit pas la loi $\chi^2(2)$.

On ne dispose cependant que de cinq valeurs de la variable X, auxquelles correspondent les cinq valeurs de Y : 1,1 - 2,3 - 5,4 - 7,8 - 10,2.

Effectuer un test de KOLMOGOROV avec les risques de première espèce, $\alpha = 5\%$, puis $\alpha = 1\%$. Qu'en conclure au vu des données ci-dessus ?

Solution : 1°) La fonction de vraisemblance $L(X_1, X_2, ..., X_n, \theta)$ associée à un échantillon $(X_1, X_2, ..., X_n)$ de variable parente X, a pour expression le produit des densités de probabilités soit $(\frac{3}{\theta})^n.(\prod_{i=1}^{i=n} X_i^2).\exp(-\sum_{i=1}^{i=n}\frac{X_i^3}{\theta}).\prod_{i=1}^{i=n} 1_{[0,+\infty[}(X_i)$. Il en résulte, en passant à la log-vraisemblance, le résultat $\ln L = n.\ln 3 - n.\ln\theta + \sum_{i=1}^{i=n} 2.\ln X_i - \frac{1}{\theta}.\sum_{i=1}^{i=n} X_i^3$, où $X_i \geq 0, \forall i / 1 \leq i \leq n$.

La **région critique** du test $\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta = \theta_1 \end{cases}$, avec $\theta_1 \succ \theta_0$, qui porte sur le *rapport des vraisemblances*, soit $\frac{L(X_1, X_2, ..., X_n, \theta_0)}{L(X_1, X_2, ..., X_n, \theta_1)} \leq k$, a donc pour caractérisation à partir des log-vraisemblances $\sum_{i=1}^{i=n} X_i^3 \geq \frac{1}{\left[\frac{1}{\theta_0} - \frac{1}{\theta_1}\right]}.\left[-\ln k - n.\ln\frac{\theta_0}{\theta_1}\right]$. Bref, on obtient un test dont la *statistique discriminante* est $T_n = \sum_{i=1}^{i=n} X_i^3$, la région critique étant définie à partir des données $x_i (1 \leq i \leq n)$ collectées, par la relation $\Pr ob(\sum_{i=1}^{i=n} x_i^3 \geq K / \theta = \theta_0) = \alpha$.

Il s'agit bien, en outre, du test le plus puissant, d'après le théorème de NEYMAN et PEARSON.

2°) Le changement de variable $Y = \frac{2}{\theta}.X^3$ (ce qui implique $X = (\frac{\theta.Y}{2})^{1/3}$), entraîne, pour ce qui est de la densité de probabilité de Y, soit $g(y)$, la relation $g(y).dy = f(x).dx$, avec $f(x) = \frac{3}{\theta}.(\frac{\theta.y}{2})^{2/3}.\exp(-\frac{y}{2})$ et $dx = \frac{1}{3}.\frac{\theta}{2}.(\frac{\theta.y}{2})^{-2/3}.dy$. En définitive, $g(y) = \frac{1}{2}.\exp(-\frac{y}{2})$, densité de probabilité de la **loi exponentielle** de paramètre $1/2$ dont il est aisé de constater à partir des rappels de cours du chapitre I (cf. paragraphe 1.1), qu'elle coïncide avec la densité de la **loi du chi-deux** à *deux degrés de liberté*, soit $\chi^2(2)$.

• Or il est rappelé que si V_1 et V_2 sont indépendantes et de lois respectives $\chi^2(n_1)$ et $\chi^2(n_2)$, la somme $V_1 + V_2$ suit la loi $\chi^2(n_1 + n_2)$ (cf. application 1.1 du chapitre I). Dès lors, la statistique $\sum_{i=1}^{i=n}(\frac{2.X_i^3}{\theta})$ suit la loi de chi-deux à $\nu = 2n$ degrés de liberté.

Reprenant donc la relation $\Pr ob(\sum_{i=1}^{i=n} X_i^3 \geq K / \theta = \theta_0) = \alpha$, on a en fonction du résultat ci-dessus $\Pr ob(\chi^2(2n) = \frac{2}{\theta}.\sum_{i=1}^{i=n} X_i^3 \geq \frac{2.K}{\theta} / \theta = \theta_0) = \alpha$. Le report à la table de valeurs annexée relative à la loi $\chi^2(2n)$ fournit aisément le nombre $t_\alpha / \Pr ob(\chi^2(2n) \geq t_\alpha) = \alpha$, ce qui entraîne, pour le **seuil critique**, la valeur $K = \frac{\theta_0.t_\alpha}{2}$. Ainsi décide-t-on H_1 si, relativement aux données $x_i (1 \leq i \leq n)$ proposées, $\sum_{i=1}^{i=n} x_i^3 \geq \frac{\theta_0.t_\alpha}{2}$ et H_0 dans le cas contraire.

3°) L'application numérique proposée conduit à $t_\alpha = 31,41$ et $K = 7,85$. La relation $\sum_{i=1}^{i=10} x_i^3 = 18 \succ K = 7,85$ autorise donc le rejet de l'hypothèse $H_0 : \theta = \theta_0 = \frac{1}{2}$, au seuil $\alpha = 5\%$.

Quant à β =Prob(décider H_0 / H_1 vraie), c'est $\Pr ob(\chi^2(2n) \prec \frac{2.K}{\theta_1} = 15,70)$.

Utilisant un calculateur en ligne de la fonction de répartition de la loi de χ^2, on a $\Pr ob(\chi^2(20) \succ 15,70) = 0,735$, ce qui montre que $\beta = 26,5\%$. Le test proposé ici est donc puissant puisque $\eta = 1 - \beta = 73,5\%$.

6. On met au point une méthode de dosage de l'élément de base d'un alliage, la méthode étant considérée comme acceptable si l'écart-type ne dépasse pas la valeur $\sigma_0 = 0,2$ (exprimée en pourcentage).

On admet les risques suivants :

- $\alpha = 5\%$, pour ce qui est du risque de refuser la méthode bien que $\sigma \le 0,2$;

- $\beta = 5\%$, pour ce qui est du risque d'accepter la méthode d'analyse alors que $\sigma \ge 0,3$.

Par ailleurs, on suppose d'une part que la distribution de la teneur en alliage pour l'élément de base considéré est de type « loi normale » et d'autre part que la moyenne m de cette distribution est constante et égale à la valeur 95,25%.

1°) Expliciter la règle de décision relative à un tel test $\begin{cases} H_0 : \theta \le \theta_0 = 0,2 \\ H_1 : \theta \ge \theta_1 = 0,3 \end{cases}$ suivant la méthode séquentielle de WALD.

2°) Calculer les espérances $E_{\theta_0}(N)$ et $E_{\theta_1}(N)$, N étant la taille aléatoire de l'échantillon prélevé.

3°) Les premières mesures effectuées conduisent à la série de données ci-dessous :

n	1	2	3	4	5	6	7	8	9	10
x_n	95,31	95,82	95,03	95,29	95,17	94,68	95,35	95,05	94,97	95,68

Que peut-on en conclure ?

4°) Quelle serait la taille minimale requise pour contenir α et β à la valeur 5%, dans le cadre d'un test classique suivant la méthode de NEYMAN et PEARSON ?.

Qu'en conclure, en définitive, quant à la comparaison entre test classique et test séquentiel ?

Solution : 1°) Sous l'hypothèse de **normalité** de la distribution étudiée, la *fonction de vraisemblance* associée à un échantillon de taille n, a pour expression :

$$L(x_1, x_2, ..., x_n, \sigma) = (\frac{1}{\sigma.\sqrt{2\pi}})^n . \exp(-\frac{1}{2.\sigma^2} . \sum_{i=1}^{i=n} (x_i - m)^2)$$

Il en résulte, pour le *rapport des vraisemblances* $R_n = \frac{L(x_1, x_2, ..., x_n, \sigma_1)}{L(x_1, x_2, ..., x_n, \sigma_0)}$, ou plutôt son logarithme népérien, l'expression ci-après.

$$\ln R_n = -n.\ln\sigma_1 + n.\ln\sigma_0 + (\frac{1}{2.\sigma_0^2} - \frac{1}{2.\sigma_1^2}).\sum_{i=1}^{i=n}(x_i - m)^2 .$$

Ramenant le test proposé $\begin{cases} H_0 : \sigma \le \sigma_0 \\ H_1 : \sigma \ge \sigma_1 \end{cases}$, avec $\sigma_1 \succ \sigma_0$ au test séquentiel entre deux hypothèses simples $\begin{cases} H_0 : \sigma = \sigma_0 \\ H_1 : \sigma = \sigma_1 \end{cases}$ décrit en rappels de cours, paragraphe 2.7, la *région d'acceptation de l'hypothèse* H_1 caractérisée par $\ln R_n \succ \ln\frac{1-\beta}{\alpha}$, conduit à la relation :

$$\sum_{i=1}^{i=n}(x_i - m)^2 \succ n.\ln(\frac{\sigma_1^2}{\sigma_0^2}) + \frac{2}{1/\sigma_0^2 - 1/\sigma_1^2}.\ln\frac{1-\beta}{\alpha},$$

relation qui est de la forme $\sum_{i=1}^{i=n}(x_i - m)^2 \succ B(n) = k.n + g_2$, avec $k = \frac{\ln(\frac{\sigma_1^2}{\sigma_0^2})}{1/\sigma_0^2 - 1/\sigma_1^2}$ et

$g_2 = \frac{2}{1/\sigma_0^2 - 1/\sigma_1^2}.\ln(\frac{1-\beta}{\alpha})$.

De même, la *région d'acceptation de l'hypothèse* H_0, qui s'écrit $\ln R_n \prec \ln\frac{\beta}{1-\alpha}$ conduit à la relation $\sum_{i=1}^{i=n}(x_i - m)^2 \prec \frac{n.\ln(\frac{\sigma_1^2}{\sigma_0^2})}{1/\sigma_0^2 - 1/\sigma_1^2} - \frac{2}{1/\sigma_0^2 - 1/\sigma_1^2}.\ln(\frac{1-\alpha}{\beta})$, relation de la forme $\sum_{i=1}^{i=n}(x_i - m)^2 \prec A(n) = k.n - g_1$, avec $g_1 = \frac{2}{1/\sigma_0^2 - 1/\sigma_1^2}.\ln(\frac{1-\alpha}{\beta})$.

On a donc en définitive, le mode opératoire suivant :

- $\sum_{i=1}^{i=n}(x_i - m)^2 \prec A(n) = k.n - g_1 \Rightarrow$ on décide $H_0 : \sigma \le \sigma_0$;
- $\sum_{i=1}^{i=n}(x_i - m)^2 \succ B(n) = k.n + g_2 \Rightarrow$ on décide $H_1 : \sigma \ge \sigma_1$;
- $A(n) \le \sum_{i=1}^{i=n}(x_i - m)^2 \le B(n) \Rightarrow$ on procède à une observation supplémentaire avant toute décision éventuelle.

Numériquement, les valeurs $\alpha = \beta = 5\%, \sigma_0 = 0,2$, et $\sigma_1 = 0,3$, entraînent à partir des formules précédentes $k = 0,0584$; $g_1 = 0,424; g_2 = 0,424$.

• On suppose que m est *connue* ici, mais dans le cas contraire, il suffirait de remplacer m par son estimateur ponctuel $\bar{x}$, les relations ci-dessus étant cependant à corriger légèrement, à savoir remplacer n par $n-1$. Ainsi, obtiendrait-on, sous cette hypothèse, le mode opératoire :

- $\sum_{i=1}^{i=n}(x_i - \bar{x})^2 \prec A(n) = k.(n-1) - g_1 \Rightarrow$ on décide $H_0 : \sigma \le \sigma_0$;

- $\sum_{i=1}^{i=n}(x_i-\bar{x})^2 \succ B(n)=k.(n-1)+g_2 \Rightarrow$ on décide $H_1:\sigma \geq \sigma_1$;

- $A(n) \prec \sum_{i=1}^{i=n}(x_i-\bar{x})^2 \prec B(n) \Rightarrow$ on procède à une observation supplémentaire avant toute décision éventuelle.

2°) S'inspirant des applications 1.4 et 2.5 du présent chapitre, on a tout d'abord, pour σ fixé à la valeur $\sigma=\sigma_0$, $E_{\sigma_0}\left[\frac{\ln f(X_i,\sigma_1)}{\ln f(X_i,\sigma_0)}\right]=E_{\sigma_0}\left[(\frac{1}{2.\sigma_0^2}-\frac{1}{2.\sigma_1^1}).(X_i-m)^2-\ln\frac{\sigma_1}{\sigma_0}\right]$, soit par linéarité, $E_{\sigma_0}\left[\ln\frac{f(X_i,\sigma_1)}{f(X_i,\sigma_0)}\right]=(\frac{1}{2.\sigma_0^2}-\frac{1}{2.\sigma_1^2}).\sigma_0^2-\ln\frac{\sigma_1}{\sigma_0}$.

Puis, pour ce qui est de $R_N=\prod_{i=1}^{i=N}\frac{f(X_i,\sigma_1)}{f(X_i,\sigma_0)}$, on a $E_{\sigma_0}[\ln R_N]=\sum_{i=1}^{i=N}E_{\sigma_0}\left[\ln\frac{f(X_i,\sigma_1)}{f(X_i,\sigma_0)}\right]$, soit $E_{\sigma_0}[\ln R_N]=N.\sigma_0^2.(\frac{1}{2.\sigma_0^2}-\frac{1}{2.\sigma_1^2})-N.\ln\frac{\sigma_1}{\sigma_0}$, d'où $N=\frac{E_{\sigma_0}[\ln R_N]}{\sigma_0^2.(\frac{1}{2.\sigma_0^2}-\frac{1}{2.\sigma_1^2})-\ln\frac{\sigma_1}{\sigma_0}}$ où encore, $N=\frac{E_{\sigma_0}[\ln R_N]}{\frac{1}{2}.(\frac{1}{\sigma_0^2}-\frac{1}{\sigma_1^2}).(\sigma_0^2-k)}$.

Mais de façon approchée, $E_N\left[E_{\sigma_0}[\ln R_N]\right]=\alpha.\ln\frac{1-\beta}{\alpha}+(1-\alpha).\ln\frac{\beta}{1-\alpha}$.

D'où en définitive et après développement, l'approximation de $E_{\sigma_0}(N)$ par l'expression $\frac{-\alpha.g_2+(1-\alpha).g_1}{k-\sigma_0^2}$.

• De même, $E_{\sigma_1}[\ln R_N]=N.\sigma_1^2.(\frac{1}{2.\sigma_0^2}-\frac{1}{2.\sigma_1^2})-N.\ln\frac{\sigma_1}{\sigma_0}$, $N=\frac{E_{\sigma_1}[\ln R_N]}{\sigma_1^2.(\frac{1}{2.\sigma_0^2}-\frac{1}{2.\sigma_1^2})-\ln\frac{\sigma_1}{\sigma_0}}$ ou encore $N=\frac{E_{\sigma_1}[\ln R_N]}{\frac{1}{2}.(\frac{1}{\sigma_0^2}-\frac{1}{\sigma_1^2}).(\sigma_1^2-k)}$.

Explicitant $E_N\left[E_{\sigma_1}[\ln R_N]\right]$ par $\beta.\ln\frac{\beta}{1-\alpha}+(1-\beta).\ln\frac{1-\beta}{\alpha}$, on obtient donc finalement l'approximation de $E_{\sigma_1}(N)$ par l'expression $\frac{-\beta.g_1+(1-\beta).g_2}{\sigma_1^2-k}$.

• Numériquement, on obtient $E_{\sigma_0}(N)=20,75$, soit après arrondi, la valeur entière $E_{\sigma_0}(N)=21$. De même, $E_{\sigma_1}(N)=12,08$, soit la valeur entière approchée $E_{\sigma_1}(N)=12$.

Se référant à l'expression générale de $E_\sigma(N)$ mentionnée dans l'application 2.5 du présent chapitre, on peut montrer que $E_\sigma(N)$ atteint son maximum pour $\sigma=k$, la valeur correspondante obtenue étant $\frac{g_1.g_2}{2.k^2}$, c'est-à-dire, numériquement et pour le cas présent, $n^*=26,36$, valeur qu'on arrondira à 26.

• Enfin, lorsque m est inconnue, tous les résultats précédents sont à augmenter d'une unité.

3°) L'application du mode opératoire susmentionné à la série d'observations proposée dans l'énoncé conduit au tableau de calculs ci-dessous :

n	1	2	3	4	5	6	7	8	9	10
x_n	95,31	95,82	95,03	95,29	95,17	94,68	95,35	95,05	94,97	95,68
$\sum_{i=1}^{i=n}(x_i-m)^2$	0,0036	0,3285	0,3769	0,3785	0,3849	0,7098	0,7198	0,7598	0,8382	1,0231
$A(n)$	-0,3656	-0,3072	-0,2488	-0,1904	-0,1320	-0,0730	-0,0152	0,0432	0,1016	0,1600
$B(n)$	0,4298	0,4357	0,4415	0,4474	0,4532	0,4590	0,4649	0,4707	0,4766	0,4824

Pour cet exemple, on constate que dès $n=6$, on entre dans la région critique puisque $\sum_{i=1}^{i=n}(x_i-m)^2 \succ B(n)$. On est donc amené à interrompre le test à la 6ème observation et à décider H_1, la méthode d'analyse contrôlée envisagée ici ne convenant manifestement pas puisque générant un écart-type trop grand.

4°) Dans le cadre d'un test classique et se référant aux rappels de cours (paragraphe 2.3.c), la fixation des erreurs de première et de seconde espèce conduit aux relations :

$$\begin{cases}\alpha = \Pr ob(S^2 \geq \pi / \sigma = \sigma_0) \\ \beta = \Pr ob(S^2 \prec \pi / \sigma = \sigma_1)\end{cases} \text{ où } S^2 = \sum_{i=1}^{i=n}(x_i-m)^2$$

(du moins si on suppose m connue).

Il en résulte, relativement à la statistique $V = \dfrac{n.S^2}{\sigma^2}$ dont il est rappelé qu'elle suit la loi du chi-deux à n degrés de liberté, soit $\chi^2(n)$, les relations $\begin{cases}\alpha = \Pr ob(V \geq \dfrac{n.\pi}{\sigma_0^2}) \\ \beta = \Pr ob(V \prec \dfrac{n.\pi}{\sigma_1^2})\end{cases}$.

L'erreur de première espèce α étant bloquée à 5%, la recherche par approximations successives de la valeur minimale de n pour laquelle l'erreur de 2ème espèce, β, est contenue à 5%, conduit au tableau ci-dessous, la lecture des nombres $t / Prob(\chi^2(n) \geq t) = P$ étant effectuée à partir de tables ou d'un calculateur en ligne.

n	χ^2_α	π	$\dfrac{n.\pi}{\sigma_1^2}$	β (en %)
20	31,41	0,063	13,96	18
30	43,77	0,059	19,45	7
40	55,76	0,056	24,78	3
35	49,80	0,057	22,13	4
33	47,40	0, 057	21,07	6
34	48,60	0,057	21,60	5

On retient donc la valeur minimale $n_0 = 34$ qui est nettement supérieure aux estimations $E_{\sigma_0}(N) = 21$ et $E_{\sigma_1}(N) = 12$, voire en tout état de cause, $\underset{\sigma}{Max}\, E_\sigma(N) = 26$, ce qui, ici encore, illustre l'intérêt du test séquentiel par rapport au test de NEYMAN et PEARSON.

7. Lors du championnat de France de ligue 1 de football (saison 2004-2005), on a relevé le nombre moyen de buts marqués par équipe et par match lors de l'ensemble des rencontres qui ont eu lieu pendant les 38 journées du championnat.

En tout, 824 buts ont été marqués lors des 380 matchs disputés. Le tableau suivant fournit la distribution du nombre X des buts marqués par équipe au cours d'un match.

Buts marqués par équipe et par match (x_i)	0	1	2	3	4	5	6	7	8	Total
Equipes ayant marqué ce nombre de buts (f_i)	268	266	152	53	13	7	0	0	1	760

Tester au seuil $\alpha = 5\%$, le caractère poissonnien de la distribution du nombre de buts marqués par une équipe au cours d'un match.

Solution : C'est le **test non paramétrique de** χ^2 qui correspond le mieux ici au problème de *l'ajustement* ou non de la distribution empirique proposée par la **loi de POISSON** caractérisée par $\Pr ob(X = x) = e^{-\lambda}.\frac{\lambda^x}{x!}$. Se référant à l'application 1.2 du chapitre II, c'est par $\hat{\lambda} = \frac{\sum_{i=1}^{i=n} f_i.x_i}{N}$ (avec $N = \sum_{i=1}^{i=n} f_i$), qu'on obtient le meilleur estimateur ponctuel de λ, à savoir numériquement $\hat{\lambda} = 1,084$. Rapprochant les *fréquences absolues observées* f_i avec les *fréquences absolues théoriques* $N.p_i$ où $p_i = \frac{\hat{\lambda}^{x_i}.e^{-\hat{\lambda}}}{x_i!}$, on obtient le tableau de calculs ci-dessous :

Classe	Effectif observé	Effectif théorique
0	268	257,01
1	266	278,65
2	152	151,06
3	53	54,59
4	13	14,80
5	7	3,21
6	0	0,58
7	0	0,09
8	1	0,01

Toutefois, il faut *regrouper* les cinq dernières classes pour vérifier la condition nécessaire $N.p_i \succ 5$ (cf. rappels de cours du présent chapitre, paragraphe 3.1.a). On a donc en résumé :

Classe	f_i	$N.p_i$	$(f_i - N.p_i)^2 / N.p_i$
0	268	257,01	0,470
1	266	278,65	0,574
2	152	151,06	0,006
3	53	54,59	0,046
4	21	18,69	0,286
Σ	760	759,99	1,383

Dans la mesure où le nombre de paramètres estimés pour réaliser l'ajustement est $k = 1$ et où il y a 5 classes après regroupement, la loi de la **distance de chi-deux**, $D = \sum_{i=1}^{i=n} \frac{(f_i - N.p_i)^2}{N.p_i}$ est donc la loi de χ^2 à $\nu = 5-1-1 = 3$ degrés de liberté.

Par lecture dans la table de valeurs annexée et pour $\alpha = 5\%$, le seuil χ^2_α qui vérifie $\Pr ob(\chi^2(3) \geq \chi^2_\alpha) = 0,05$ est égal à 7,815.

On a donc $\chi^2_{calculé} = 1,383 \prec \chi^2_\alpha = 7,815$ ce qui signifie qu'il n'y a pas lieu ici de rejeter l'hypothèse suivant laquelle le modèle de POISSON décrit le nombre de buts marqués par match et par équipe.

8. « *Qualité d'un générateur de nombres aléatoires uniformes* ».

On souhaite vérifier la qualité du générateur de nombres aléatoires d'une calculatrice scientifique. Pour cela, on procède à 250 tirages dans l'ensemble $\{0,1,2,...,9\}$ et on obtient les résultats suivants :

x	0	1	2	3	4	5	6	7	8	9
$N(x)$	28	32	23	26	23	31	18	19	19	31

Tester si le générateur produit des entiers uniformément répartis sur l'ensemble des entiers $\{0,1,2,3,4,5,6,7,8,9\}$, le risque choisi étant de 5%.

Solution : Préférable au test de χ^2, encore qu'ici les tailles des classes sont suffisamment significatives pour éviter des regroupements, le **test de KOLMOGOROV** est suggéré dans cet exercice.

Les données étant regroupées en classes, la mise en œuvre du test conduit à calculer la **distance de KOMOGOROV**, soit :

$$D(\widehat{F}, F_0) = Sup\left\{\left|\widehat{F}(X(i)) - F_0(X(i))\right|, \left|\widehat{F}(X(i-1)) - F_0(X(i))\right|\right\}$$

Il en résulte le tableau de calculs :

x_i	0	1	2	3	4	5	6	7	8	9
$\widehat{F}(x_i)$	0,112	0,240	0,332	0,436	0,528	0,652	0,724	0,800	0,876	1
$F_0(x_i)$	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1
$\widehat{F}(x_i) - F_0(x_i)$	0,012	0,040	0,032	0,036	0,028	0,052	0,024	0	-0,024	0
$\widehat{F}(x_{i-1}) - F_0(x_i)$	0,012	0,088	0,060	0,068	0,064	0,072	0,048	0,076	0,1	0,124

Le maximum constaté, qui constitue la distance $D_{calculé}(\widehat{F}, F_0)$ est immédiatement égal à 0,124. Si pour le test considéré $\begin{cases} H_0 : F = F_0 \\ H_1 : F \neq F_0 \end{cases}$ on choisit un risque de première espèce $\alpha = 5\%$, l'utilisation de la *table annexée* des valeurs de KOLMOGOROV pour le *test à un échantillon* conduit, relativement au seuil $D_{\alpha,n} / \Pr ob(D(F_0, \widehat{F}) \geq D_{\alpha,n}) = \alpha$, à la valeur $D_{\alpha,n} = 0,410$.

La relation $D_{calculé} = 0,124 \prec D_{\alpha,n} = 0,410$ ne permet pas de conclure ici au rejet de l'ajustement de la distribution proposée par la loi uniforme, ce qui, en d'autres termes, valide la qualité du générateur.

- A titre indicatif, le **test d'ajustement de** χ^2 conduirait aux calculs ci-dessous :

x_i	0	1	2	3	4	5	6	7	8	9	Σ
$N_{cumulé}$ (1)	28	60	83	109	132	163	181	200	219	250	-
$250.F_0(x_i)$ (2)	25	50	100	125	150	175	200	225	250	250	
$\frac{[(1)-(2)]^2}{(2)}$	0,36	2	0,86	0,81	0,39	1,13	0,21	0	0,16	0	5,91

Ainsi a-t-on $\chi^2_{calculé} = 5,91$. Or la *distance de chi-deux* suit ici la loi de χ^2 à $\nu = n-1-k$ degrés de liberté, soit $\nu = 10-1-0 = 9$, la lecture du seuil $\chi^2_\alpha / \Pr ob(\chi^2(9) \geq \chi^2_\alpha) = \alpha$, conduisant pour $\alpha = 0,05$, à la valeur $\chi^2_\alpha = 16,91$.

On est donc manifestement dans la zone d'acceptation de l'hypothèse nulle H_0 puique $\chi^2_{calculé} = 5,91 \prec \chi^2_\alpha = 16,91$, ce qui confirme le résultat précédent fourni par le test de KOLMOGOROV.

9. On dispose d'un échantillon réduit de 10 valeurs indépendantes dont, au risque $\alpha = 5\%$, on souhaite tester la normalité ou non, ces données étant les suivantes :

27,2 - 29,7 - 30,0 - 28,0 - 27,9 - 29,9 - 28,3 - 30,9 - 30,2 - 30,4

Choisissant un test approprié, qu'en conclure quant à cette hypothèse de normalité ?

Solution : La faible valeur de la taille de l'échantillon conduit ici à privilégier le **test non paramétrique de SHAPIRO-WILK** ou le **test de KOLMOGOROV**.

Appliquant le mode opératoire du test de SHAPIRO-WILK présenté en rappels de cours de ce chapitre (cf. paragraphe 3.1.c), on est amené à :

Etape 1 → Classer les données x_i par ordre croissant, soit :

27,2	27,9	28,0	28,3	29,7	29,9	30,0	30,2	30,4	30,9

Etape 2 → Calculer $T_n = \sum_{i=1}^{i=n} (x_i - \bar{x})^2$, soit numériquement $T_n = 14,625$.

Etape 3→ n étant pair, calculer les différences $d_i = x_{n-i+1} - x_i$, soient $d_1 = x_{10} - x_1 = 3,7$ - $d_2 = x_9 - x_2 = 2,5$ - $d_3 = x_8 - x_3 = 2,2$ - $d_4 = x_7 - x_4 = 1,7$ - $d_5 = x_6 - x_5 = 0,2$.

Etape 4 → Lire dans la table de SHAPIRO-WILK (cf. annexes), les valeurs des coefficients a_i, pour $n = 10$, soient $a_1 = 0,5739$ - $a_2 = 0,3291$ - $a_3 = 0,2141$ - $a_4 = 0,1224$ - $a_5 = 0,0399$.

Etape 5 → Calculer $b = \sum_{i=1}^{i=5} a_i.d_i$, soit $b = 3,633$, puis la quantité $W = \frac{b^2}{T_n}$, soit numériquement, $W = 0,903$.

Etape 6 → Lire dans la table de SHAPIRO-WILK annexée, le seuil $W_\alpha / \Pr ob(W \leq W_\alpha) = \alpha$, soit numériquement et pour $\alpha = 5\%, n = 10$, la valeur $W_\alpha = 0,842$.

La région critique du test étant caractérisée par $W \leq W_\alpha$, on se trouve être ici dans le cas où $W_{calculé} = 0,903 \succ W_\alpha = 0,842$. Dans ces conditions, on ne peut donc pas rejeter l'hypothèse nulle qui correspond à la normalité de la distribution proposée.

10. On a demandé à 30 personnes d'indiquer, parmi 5 chocolats noirs du marché (excellence, amère, amazonie, pâtissier, supérieur) leur chocolat préféré.

Le tableau suivant fournit le nombre de personnes ayant préféré chacun de ces chocolats :

Chocolat	Nombre de personnes ayant préféré ce chocolat
Excellence	7
Amère	5
Amazonie	5
Pâtissier	4
Supérieur	9
Total	30

1°) Peut-on considérer que chacun des chocolats est préféré par la même proportion de personnes ?

2°) Refaire le même exercice en multipliant les effectifs observés par 10.

3°) Interpréter les résultats obtenus.

Solution : 1°) S'agissant d'un ajustement à partir de **données qualitatives**, c'est le **test de chi-deux** qui s'impose ici. Relativement aux hypothèses H_0 : « *La distribution est uniforme* » et H_1 : « *La distribution n'est pas uniforme* », le calcul de la distance correspondante (cf. paragraphe 3.1.a des rappels de cours) conduit au tableau ci-après :

Classe	Effectifs observés N_i	Effectifs théoriques $N.p_i$	$\frac{(N_i - N.p_i)^2}{N.p_i}$
Excellence	7	6	0,166
Amère	5	6	0,166
Amazonie	5	6	0,166
Pâtissier	4	6	0,666
Supérieur	9	6	1,5
Total	30	30	2,667

Toutes les classes ayant un effectif théorique supérieur à 5, il n'y a pas de regroupement à opérer ici, la valeur calculée de la *distance de chi- deux* $D = \sum_{i=1}^{i=m} \frac{(N_i - N.p_i)^2}{N.p_i}$, étant égale à $D_{calculé} = 2,667$. D'autre part, D suit la **loi du chi- deux** à $\nu = m - 1 - k$ degrés de liberté, soit numériquement $\nu = 5 - 1 - 0 = 4$.

Fixant par ailleurs à 5%, la valeur du risque de 1ère espèce $\alpha = \Pr ob($ décider H_1 / H_0 vraie), il en résulte par lecture dans la table annexée des valeurs de loi de chi- deux, le seuil $\chi^2_{0,05} / \Pr ob(\chi^2(4) \geq \chi^2_{0,05}) = 0,05$, soit $\chi^2_{0,05} = 9,488$. On en conclut en conséquence à une *indifférence* du goût des consommateurs pour les chocolats proposés puisque $\chi^2_{calculé} = 2,67$ est inférieur à $\chi^2_{0,05} = 9,488$.

2°) Si on multiplie les effectifs précédents par 10, l'indicateur d'écart $\sum_{i=1}^{i=m} \frac{(N_i - N.p_i)^2}{N.p_i}$ est multiplié par 10 et on a donc $\chi^2_{calculé} = 26,67$. **Cette fois,** *l'hypothèse d'uniformité est largement rejetée.*

3°) Cet exercice et notamment la discordance des conclusions des questions précédentes, souligne la relative faiblesse des la puissance du test de χ^2, c'est-à-dire l'incapacité à distinguer les hypothèses H_0 et H_1 lorsque la taille de l'échantillon est faible.

Ainsi, la valeur élevée de l'erreur de 2ème espèce β, pour n faible (telle dans la 1ère question), conduit-elle à accepter largement l'hypothèse H_0 alors qu'il en est autrement dès que l'information possédée devient plus importante (cf. 2ème question).

Le regroupement des classes pour lesquelles $N.p_i \prec 5$ (ce n'est pas le cas néanmoins ici), est aussi une source importante de perte d'information qui montre les **limites** du test de χ^2 lorsque les effectifs sont faibles.

11. On veut vérifier si la consommation de beurre est liée au fait d'être breton. Une enquête auprès de 30 utilisateurs a donné les résultats résumés dans le tableau ci-dessous :

	Breton	Non Breton	Total
Consomme du beurre salé	10	5	15
Consomme du beurre doux	4	11	15
Total	14	16	30

Le type de beurre consommé est-il indépendant de la région du consommateur ?

Solution : La question qui est posée équivaut à **tester l'indépendance** ou non des *variables qualitatives nominales* « *consommer du beurre salé ou non* » et « *être breton ou non* », un test de chi-deux étant approprié ici, selon les conditions exposées en rappels de cours du présent chapitre (cf. paragraphe 3,4,c).

Par sommations, respectivement en lignes et en colonnes, on obtient immédiatement les estimateurs des probabilités associées aux états considérés à savoir $\widehat{p_{i.}}$ et $\widehat{p_{.j}}$. Au tableau des **fréquences absolues observées** n_{ij} présentées dans l'énoncé, il faut superposer le tableau des **fréquences théoriques** suivant lesquelles, sous l'hypothèse nulle H_0 d'indépendance entre les caractères étudiés, les estimations $\widehat{p_{ij}}$ des probabilités associées aux couples (i, j) sont égales aux produits $\widehat{p_{i.}} \cdot \widehat{p_{.j}}$.

Relativement à ces dernières, les fréquences absolues théoriques s'obtiennent en multipliant par le nombre total d'observations, soit n ($n = 30$ dans l'exemple proposé), les probabilités $\widehat{p_{i.}} \cdot \widehat{p_{.j}}$. On obtient donc le tableau ci-dessous :

	Breton	Non Breton
Consomme du beurre salé	7	8
Consomme du beurre doux	7	8

Par exemple, pour le couple « beurre salé » - « être breton », on a $\widehat{p_{i.}} = \frac{15}{30} = \frac{1}{2}$ et $\widehat{p_{.j}} = \frac{14}{30}$, d'où en définitive, $n.\ \widehat{p_{i.}} \cdot \widehat{p_{.j}} = 7$, et ainsi de suite....

Conformément aux résultats des rappels de cours, la statistique $V = \sum_{i=1}^{i=r}\sum_{j=1}^{j=k} \frac{(n_{ij} - n.\widehat{p_{i.}}.\widehat{p_{.j}})^2}{n.\widehat{p_{i.}}.\widehat{p_{.j}}}$ suit la loi du χ^2 à $\nu = (r-1).(k-1)$ degrés de liberté, soit $\nu = (2-1).(2-1) = 1$, pour le cas présent. Quant à la valeur calculée de V sur la base des données proposées et des résultats précédents, c'est immédiatement $V_{calculé} = \frac{(10-7)^2}{7} + \frac{(5-8)^2}{8} + \frac{(4-7)^2}{7} + \frac{(11-8)^2}{8} = 4,82$.

Considérant le seuil $\chi^2_\alpha / \Pr ob(V \geq \chi^2_\alpha) = \alpha$ et une erreur de première espèce α fixée à 5%, l'usage de la table de valeurs annexée portant sur la loi du chi- deux à un degré de liberté, entraîne $\chi^2_\alpha = 3,84$.

L'inégalité $\chi^2_{calculé} = 4,82 \succ \chi^2_\alpha = 3,84$ permet de conclure, au risque 5%, au rejet de l'hypothèse suivant laquelle le fait d'être breton ou non n'a pas d'influence sur le type de beurre consommé (salé ou doux).

• A noter cependant que cette conclusion est **remise en cause** si on applique la **correction de continuité de YATES** (plus conservatrice pour H_0), la *distance de chi-deux corrigée* étant $V'_{calculé} = \frac{(|10-7|-0,5)^2}{7} + ... + \frac{(|11-8|-0,5)^2}{8} = 3,35$.

On n'a plus $\chi^2_{calculé} \succ \chi^2_\alpha = 3,84$. Par contre, au niveau de signification $\alpha = 10\%$, ou pour le test **unilatéral** selon un niveau de signification $\alpha = 5\%$, la conclusion de rejet demeure puisqu'on a alors $\chi^2_\alpha = 2,71$.

12. On désire tester si un médicament a une influence sur le comportement psychomoteur. On choisit au hasard, 20 sujets qu'on répartit aléatoirement en deux groupes : le groupe témoin et le groupe expérimental. On leur fait subir la même expérience psychomotrice. On a administré auparavant le médicament aux sujets du groupe expérimental et un placebo au groupe témoin. Les résultats obtenus sont les suivants :

Groupe témoin	166	167	169	170	174	173	172	170	166	173
Groupe expérimental	167	162	165	168	162	160	164	158	165	169

On suppose que, dans chaque groupe, les résultats sont distribués selon une loi gaussienne, que la variance est la même pour les deux groupes, et que les performances des sujets sont indépendantes.

Tester au niveau de signification $\alpha = 5\%$, l'hypothèse selon laquelle le médicament n'a aucun effet sur le comportement psychomoteur.

Solution : Entre les deux échantillons, le test proposé est un **test *t* de STUDENT** de type **bilatéral**, tel celui exposé en rappels de cours du présent chapitre (cf. paragraphe 2.4.b), soit $\begin{cases} H_0 : m_X = m_Y \\ H_1 : m_X \neq m_Y \end{cases}$. Les hypothèses de *normalité* et d'égalité des variances (*homoscédasticité*), permettent de conclure à l'expression du **seuil critique** :

$$\pi = \frac{t_\alpha}{\sqrt{n_X + n_Y - 2}} . \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}} . \sqrt{(n_X - 1).\widehat{S_X}^2 + (n_Y - 1).\widehat{S_Y}^2}$$

où t_α vérifie $\Pr ob(|T| \geq t_\alpha) = \alpha$, T suivant par ailleurs, la **loi de STUDENT** à $\nu = n_X + n_Y - 2$ degrés de liberté.

A partir de la formule $\hat{S}^2 = \frac{1}{n-1}.\left[\sum_{i=1}^{i=n} x_i^2 - n.\bar{x}^2\right]$ et des données numériques proposées, l'utilisation du tableur EXCEL conduit aux résultats suivants :

$$\widehat{S_X}^2 = 8,89 \text{ - } \widehat{S_Y}^2 = 12,44 \text{ - } \nu = 18 \text{ - } t_\alpha = 2,10 \text{ - } \pi = 3,067 \text{ - } \bar{x} = 170 \text{ - } \bar{y} = 164.$$

Or $|\bar{x} - \bar{y}| = 170 - 164 = 6 \succ \pi = 3,067$. On en conclut donc, au niveau de signification 5%, au rejet de l'hypothèse suivant laquelle le médicament n'aurait aucun effet sur le comportement psychomoteur.

13. On considère la statistique W de la somme des rangs de WILCOXON-MANN-WHITNEY dans le cas où $n_X = 4$ et $n_Y = 3$.

1°) Enumérant toutes les valeurs possibles de W et les probabilités correspondantes, en déduire, sous l'hypothèse H_0, la distribution de W dont on dessinera le graphe.

2°) Vérifier empiriquement que $E(W) = \frac{n_X.(n_X + n_Y + 1)}{2}$. Vérifier de même qu'on a $Var(W) = \frac{n_X.n_Y.(n_X + n_Y + 1)}{12}$, et retrouver ainsi les résultats présentés en rappels de cours du chapitre III (cf. paragraphe 3.2.b).

Solution : 1°) Le nombre de combinaisons possibles entre les rangs des 3 valeurs de X et ceux des 4 valeurs de Y est immdiatement égal à $C_7^3 = 35$. Désignant par R_i les **rangs** *des valeurs de* X *dans l'échantillon regroupé* $(X,Y), 1 \le R_i \le 7$, les 35 combinaisons possibles entraînent pour R_i, les valeurs rassemblées ci-après, la somme $W = \sum_{i=1}^{i=4} R_i$ étant aussi indiquée pour chacun des éléments ainsi énumérés.

Par exemple, dans l'échantillon (X,Y) regroupé et trié par ordre croissant, le cas où les quatre premières valeurs rencontrées portent sur X conduit à la série des rangs $(1,2,3,4)$ de somme $W = 10$, et ainsi de suite....

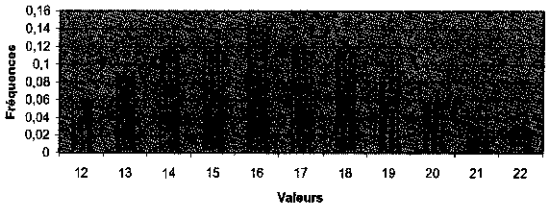
On obtient en définitive, le tableau ci-dessous, dont la construction pourrait également s'effectuer suivant une *méthode arborescente* telle en *théorie des graphes* (Recherche Opérationnelle).

Combinaison				W	Combinaison				W	Combinaison				W
1	2	3	4	10	1	3	4	7	15	1	4	6	7	18
1	2	3	5	11	1	3	5	6	15	1	5	6	7	19
1	2	3	6	12	1	3	5	7	16	2	4	5	6	17
1	2	3	7	13	1	3	6	7	17	2	4	5	7	18
1	2	4	5	12	2	3	4	5	14	2	4	6	7	19
1	2	4	6	13	2	3	4	6	15	2	5	6	7	20
1	2	4	7	14	2	3	4	7	16	3	4	5	6	18
1	2	5	6	14	2	3	5	6	16	3	4	5	7	19
1	2	5	7	15	2	3	5	7	17	3	4	6	7	20
1	2	6	7	16	2	3	6	7	18	3	5	6	7	21
1	3	4	5	13	1	4	5	6	16	4	5	6	7	22
1	3	4	6	14	1	4	5	7	17					

Sous l'hypothèse H_0, tous les cas sont **équiprobables** de probabilité commune $\frac{4!.3!}{7!} = 0,028571$. A partir du tableau de la page antérieure et dénombrant pour chaque valeur w_i de W, le nombre de cas favorables n_i, on en déduit aisément la loi de probabilité ci-dessous :

w_i	10	11	12	13	14	15	16	17	18	19	20	21	22	Σ
n_i	1	1	2	3	4	4	5	4	4	3	2	1	1	35
p_i(*)	0,028	0,028	0,057	0,086	0,114	0,114	0,143	0,114	0,114	0,086	0,057	0,028	0,028	1

(*) $p_i = \Pr ob(W = w_i)$.



L'histogramme correspondant a donc l'allure ci-contre et montre que même pour les faibles valeurs des tailles des échantillons , la *convergence* de W vers la **loi normale** est assurément *très rapide*.

2°) Par calculs par exemple sur EXCEL, on a immédiatement $E(W) = \sum_i w_i.p_i = 16$. C'est justement, la valeur fournie par la formule $E(W) = \frac{n_X.(n_X + n_Y + 1)}{2}$, soit présentement $E(W) = \frac{4 \times 8}{2} = 16$.

De même, la variance $Var(W) = \sum_i (w_i - E(W))^2.p_i$ est égale numériquement à $Var(W) = 8$. Or ici encore, c'est la valeur fournie par la formule littérale générale, à savoir, $Var(W) = \frac{n_X.n_Y.(n_X + n_Y + 1)}{12}$, soit numériquement $Var(W) = \frac{4 \times 3 \times 8}{12} = 8$.

L'exemple ainsi présenté, illustre les mécanismes de la démonstration qui, de façon générale, conduit aux résultats présentés en rappels de cours, quant au test de MANN-WHITNEY-WILCOXON.

14. Un même logiciel a été vendu à deux sociétés, soit en huit exemplaires à la société A et en 10 exemplaires à la société B. On a relevé ci-dessous le nombre d'utilisation de chaque exemplaire sur la même période de temps.

Société A	110	82	121	47	103	78	97	143		
Société B	92	101	38	71	52	108	65	64	88	111

Peut-on conclure que le logiciel est utilisé de façon similaire dans les deux sociétés ?

Pour répondre à cette question, on utilisera sucessivement :

1°) le test de WILCOXON-MANN-WHITNEY.

2°) Le test de KOLMOGOROV-SMIRNOV.

Solution : 1°) La mise en œuvre du **test non paramétrique de MANN-WHITNEY-WILCOXON**, pour le problème de comparaison proposé $(n_A = 8, n_B = 10)$ conduit à regrouper les valeurs relatives aux deux sociétés A et B dans un même échantillon qu'on triera par ordre croissant, la somme des rangs des valeurs liées à A (échantillon le plus petit), étant calculée ensuite (cf. rappels de cours du présent chapitre, paragraphe 3.2.b).

On obtient ainsi la suite :

Valeur	38	47	52	64	65	71	78	82	88
Société	B	A	B	B	B	B	A	A	B
Rang	1	2	3	4	5	6	7	8	9
Valeur	92	97	101	103	108	110	111	121	143
Société	B	A	B	A	B	A	B	A	A
Rang	10	11	12	13	14	15	16	17	18

La somme $W_{calculé}$ des rangs liés aux données qui portent sur la société A est égale à $W_{calculé} = 2+7+8+11+13+15+17+18 = 91$. Dans le contexte, c'est le test *bilatéral* qu'il faut considérer ici. Le recours à un *calculateur en ligne* (cf. geai.univ-brest.fr) conduit, par exemple, pour $\alpha = 10\%$, à la **région critique** définie par $W \notin [57,95]$. On peut aussi utiliser des tables de valeurs telles celles annexées et qui donnent les *valeurs critiques du test unilatéral de* WILCOXON-MANN-WHITNEY). Décomposant le test bilatéral traité présentement en deux tests unilatéraux $\begin{cases} H_0 : F_A = F_B \\ H_1 : F_A \prec F_B \end{cases}$ et $\begin{cases} H_0 : F_A = F_B \\ H_1 : F_A \succ F_B \end{cases}$, chacun d'erreur de première espèce égale à $\alpha' = \alpha/2 = 5\%$, il s'ensuit immédiatement les régions critiques respectives $W \leq 56$ et $W \geq 96$, d'où, pour le test bilatéral obtenu par union et au niveau de signification $\alpha = 10\%$, la région critique $W \notin [57,95]$ déjà obtenue ci-dessus.

• Or, pour les données proposées, $W_{calculé} = 91 \in [57,95]$. On est donc amené ici à ne pas rejeter l'hypothèse nulle H_0, c'est-à-dire à ne pas se prononcer sur une différence significative d'utilisation du logiciel entre les deux sociétés considérées A et B.

• Bien que n_A et n_B soient relativement faibles, l'approximation de W par la **loi normale** est déjà largement admissible, ce que tendait à montrer l'exercice 13 précédent. La statistique W de WILCOXON a pour moyenne $\frac{n_A.(n_A+n_B+1)}{2}$ et pour variance $\frac{n_A.n_B.(n_A+n_B+1)}{12}$, soit numériquement $E(W) = 76$ et $Var(W) = 126,66 \Rightarrow \sigma(W) = 11,25$.

La **région critique** est caractérisée quant à elle par $\Pr ob(|W-76| \geq W_\alpha) = \alpha$ (avec $\alpha = 10\%$). Passant à la variable normale, centrée, réduite $\xi = \frac{W-76}{11,25}$, on a donc la relation $\Pr ob(|\xi| \geq \frac{W_\alpha}{11,25}) = 0,10$. Or, la lecture dans la table annexée des valeurs de la fonction de répartition $\Pi(t) = \Pr ob(\xi \leq t)$, conduit à déterminer $t_\alpha / \Pr ob(|\xi| \geq t_\alpha) = 2.[1-\Pi(t_\alpha)] = 0,10$, soit immédiatement $t_\alpha = 1,645$.

Finalement, $\frac{W_\alpha}{11,25} = 1,645 \Rightarrow W_\alpha = 18,51$. En définitive, la **région critique** du test caractérisée par $W \notin]76-W_\alpha, 76+W_\alpha[$, s'écrit donc $W \notin]76-18,51, 76+18,51[$.

En arrondissant les bornes précédentes, on obtient donc la région critique $W \notin]57,95[$, ce qui est très proche du résultat obtenu précédemment et qui souligne la validité de l'approximation par la loi normale, même lorsqu'il s'agit de petits échantillons.

2°) La mise en œuvre du **test de KOLMOGOROV-SMIRNOV** conduit, quant à elle, à calculer la **distance** $D(\widehat{F_A},\widehat{F_B}) = \underset{x\in R}{Sup}\left|\widehat{F_A}(x)-\widehat{F_B}(x)\right|$ et pour cela à dresser le *tableau comparatif des fonctions de répartition empiriques* comme indiqué ci-après :

Valeurs	N_A	N_B	$\widehat{F_A}=\frac{N_A}{8}$	$\widehat{F_B}=\frac{N_B}{10}$	$\left\|\widehat{F_A}-\widehat{F_B}\right\|$
38	0	1	0	0,1	0,1
47	1	1	0,125	0,1	0,025
52	1	2	0,125	0,2	0,075
64	1	3	0,125	0,3	0,175
65	1	4	0,125	0,4	0,275
71	1	5	0,125	0,5	**0,375**
78	2	5	0,250	0,5	0,250
82	3	5	0,375	0,5	0,125
88	3	6	0,375	0,6	0,225
92	3	7	0,375	0,7	0,325
97	4	7	0,5	0,7	0,2
101	4	8	0,5	0,8	0,3
103	5	8	0,625	0,8	0,175
108	5	9	0,625	0,9	0,275
110	6	9	0,75	0,9	0,15
111	6	10	0,75	1	0,25
121	7	10	0,875	1	0,125
143	8	10	1	1	0

On a donc $D(\widehat{F_A},\widehat{F_B}) = 0,375$ compte tenu des données numériques susmentionnées.

- Comme cela est indiqué en rappels de cours (cf. paragraphe 3.2 du présent chapitre), la statistique du test qui est $\sqrt{\frac{n_A.n_B}{n_A+n_B}}.\underset{x\in R}{Sup}\left|\widehat{F_A}(x)-\widehat{F_B}(x)\right|$, suit, sous l'hypothèse $H_0 : F_A = F_B$, une loi dont les valeurs peuvent être tabulées, la probabilité $Prob(\sqrt{\frac{n_A.n_B}{n_A+n_B}}.D(\widehat{F_A},\widehat{F_B}) \geq t)$ étant convergente vers la somme $2.\sum_{k=1}^{k=+\infty}(-1)^{k+1}.\exp(-2.k^2.t^2)$.

Pour n_A et n_B assez grands et pour $\alpha = 10\%$, le seuil $1,22.\sqrt{\frac{n_A+n_B}{n_A.n_B}}$ peut ainsi être retenu à défaut de disposer des tables de valeurs ad hoc. Numériquement, cette dernière approximation conduit au seuil $1,22.\sqrt{\frac{8+10}{8\times 10}} = 0,57$. Bien que n_A et n_B seoient plutôt faibles, cette approximation demeure cependant acceptable, un calcul exact conduit à partir de la somme de la série susmentionnée conduisant en fait à un résultat très proche.

Finalement, $D_{calculé}(\widehat{F_A},\widehat{F_B}) = 0,375 \prec D_\alpha = 0,57$. On ne peut donc pas rejeter ici encore l'hypothèse H_0, à savoir une utilisation similaire du logiciel par les deux sociétés A et B.

15. 120 patients atteints d'une même maladie ont été traités par deux méthodes différentes. Parmi les 70 qui ont reçu le traitement A, 22 ont guéri et parmi les 50 qui ont reçu le traitement B, 25 ont guéri.

Le taux de guérison est-il meilleur pour le traitement B que pour le traitement A, ceci au niveau de signification $\alpha = 5\%$?

On appliquera successivement à cet effet, un test paramétrique et un test non paramétrique.

Solution : 1°) S'agissant du **test paramétrique de comparaison entre deux proportions**, on se reportera aux rappels de cours du présent chapitre (paragraphe 2.4.c),dont une illustration est fournie à travers l'application 3.4 précédente.

Le test proposé ici est **unilatéral** $\begin{cases} H_0 : p_A = p_B \\ H_1 : p_A \prec p_B \end{cases}$, les données en étant $n_A = 70, n_B = 50$, $X_A = 22, X_B = 25, \alpha = 5\%$. Le test a pour **région critique** $F_A - F_B \prec \pi$ où le seuil critique π est défini par $\pi = t_\alpha \sqrt{p.(1-p).(\frac{1}{n_A} + \frac{1}{n_B})}$ et t_α vérifie $\Pr ob(\xi \leq t_\alpha) = \alpha$, p étant estimé par ailleurs, sous l'hypothèse $H_0 : p_A = p_B$, par la statistique $\frac{X_A + X_B}{n_A + n_B}$. A noter que l'usage de la *loi normale* est autorisé par le fait que n_A et n_B sont suffisamment grands.

Ainsi, numériquement, $p = 0,392; t_\alpha = -1,645; \pi = -0,149; F_A - F_B = \frac{22}{70} - \frac{25}{50} = -0,185$.

Dans la mesure où $F_A - F_B = -0,185 \prec \pi = -0,149$, c'est donc l'hypothèse de rejet de H_0 qu'il faut retenir, c'est-à-dire la conclusion d'une différence entre les deux traitements étudiés.

2°) S'agissant d'un **test non paramétrique**, il faut se référer à l'application 4.5 du présent chapitre illustrant plusieurs tests non paramétriques pour comparer l'efficacité d'un traitement, *la taille suffisamment grande des effectifs* autorisant l'utilisation d'un **test de chi-deux** (à défaut, il faudrait recourir à la méthode exacte de FISHER).

Dans ce test, on va comparer les *effectifs observés* des malades guéris et non gnéris pour chacun des traitements en question, aux *effectifs théoriques* qui résulteraient d'une efficacité égale des deux traitements, c'est-à-dire selon la proportion commune $p = p_A = p_B = 0,392$, ou plus exactement par règles de trois au prorata des effectifs (par exemple, $70 \times \frac{47}{120} = 27,4; 50 \times \frac{47}{120} = 19,6; \ldots$).

En fait, on applique la relation d'indépendance suivant laquelle la loi du couple est égale au produit des lois marginales. Ainsi, pour la paire (guéris, traitementA) a-t-on le produit $\frac{47}{120} \times \frac{70}{120}$, ce qui en fréquence absolue fournit la valeur $120 \times \frac{47}{120} \times \frac{70}{120}$ indiquée ci-dessus. Et ainsi de suite...

Le tableau présenté ci-après rassemble les résultats obtenus quant aux effectifs observés O_{ij} et théoriques E_{ij}, la **distance de chi-deux** à considérer pour représenter l'écart entre l'observé et le théorique étant définie par la statistique $\sum_{i=1}^{i=2} \sum_{j=1}^{j=2} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$.

	Traitement A	Traitement B	Total
Guéris	22 *(27,4)*	25 *(19,6)*	47
Non guéris	48 *(42,6)*	25 *(30,4)*	73
Total	70	50	120

Dans ce tableau, les effectifs théoriques ont été portés entre parenthèses et en caractères italiques gras).

• La *distance de chi-deux* ainsi obtenue pour les données considérées et à partir de la formule précédemment rappelée est égale à $\chi^2_{calculé} = 4,20$, résultat qu'on peut obtenir également plus directement par la formule $\chi^2_{calculé} = \dfrac{n.(a.d-b.c)^2}{(a+c).(b+d).(a+b).(c+d)}$ applicable aux *tables de contingences* $(2,2)$ de la forme :

	Traitement A	Traitement B	Total
Guéris	a	b	$a+b$
Non guéris	c	d	$c+d$
Total	$a+c$	$b+d$	$n=a+b+c+d$

En fait, pour ces tables $(2,2)$ et comme cela a déjà été indiqué antérieurement, on rencontre fréquemment une *transcription plus conservatrice* de l'hypothèse H_0 et intégrant la **correction de continuité de YATES**, à savoir $\chi^2 = \sum_{i=1}^{i=2}\sum_{j=1}^{j=2} \dfrac{(|O_{ij}-E_{ij}|-0,5)^2}{E_{ij}}$, ou encore suivant la forme synthétique susmentionnée, $\chi^2 = \dfrac{n.(|a.d-b.c|-\frac{n}{2})^2}{(a+b).(c+d).(a+c).(b+d)}$.

Numériquement, et pour les données précédentes, $\chi^2_{calculé} = 3,45$.

• Revenant à la statistique χ^2, elle est caractérisée s'agissant d'un tableau $(2,2)$ par un nombre de degrés de liberté $\nu = (r-1).(k-1)$ où $r=2, k=2$, ce qui entraîne en définitive $\nu = 1$. Se référant aux tables de valeurs annexées relatives à la loi de χ^2, le seuil de rejet de l'hypothèse H_0, soit $\chi^2_\alpha / \Pr ob(\chi^2 \geq \chi^2_\alpha) = \alpha$, est égal à 2,71 pour le cas du test unilatéral et lorsque $\alpha = 5\%$.

Même dans l'hypothèse la plus conservatrice qui est celle d'un calcul de la distance de chi-deux avec correction de continuité de YATES, la relation $\chi^2_{calculé} = 3,45 > \chi^2_\alpha = 2,71$ conduit ici à retenir que les écarts entre distribution observée et distribution théorique obtenue sous l'hypothèse nulle H_0, sont suffisamment significatifs pour conclure au rejet de H_0, c'est-à-dire concrètement, à l'efficacité supérieure du traitement B par rapport au traitement A. cette conclusion, corrobore celle obtenue précedemment à travers le test paramétrique.

C'est d'ailleurs, un résultat logique puisque comme cela a été montré dans l'application 3.5 du présent chapitre, les deux tests sont équivalents du moins lorsque la convergence vers la loi normale est applicable.

16. On considère deux groupes de livres, les uns portant sur la médecine, les autres portant sur l'histoire. On cherche à savoir s'ils sont de même épaisseur ou non, les données résultant d' échantillons de tailles respectives 15 et 13, et étant :

Livres de médecine	29	39	60	78	82	112	125	170	192	224	263	275	276	286	756
Livres d'histoire	126	142	156	228	245	246	370	419	433	454	478	503	369		

A l'aide du test de la médiane (test de MOOD), que conclure au niveau de signification $\alpha = 5\%$?

Solution : 1°) Reprenant la procédure déjà exposée dans l'application 4.5 du présent chapitre (« *illustration de plusieurs tests non paramétriques pour évaluer l'efficacité d'un traitement* »), le **test de MOOD** conduit dans un premier temps, à calculer la valeur de la **médiane,** soit M, sur l'échantillon obtenu par regroupement des données relatives aux deux populations concernées.

Pour calculer cette valeur de M, il faut, après concaténation des échantillons, classer les données par ordre croissant, M étant égale à $x_{(\frac{n+1}{2})}$ si n est impair, et à $\frac{1}{2}.\left[x_{(\frac{n}{2})} + x_{(\frac{n+1}{2})}\right]$ si n est pair. Pour le cas considéré et après tri croiss ant, $n = 15+13 = 28$ est pair, et on a $M = \frac{1}{2}.(x_{(14)} + x_{(15)}) = \frac{1}{2}.(228+245) = 236,5$.

On dresse alors, une **table de contingences** $(2,2)$, d'une part relativement aux deux types d'ouvrages considérés, et d'autre part en fonction de la position par rapport à la médiane M. Le tableau qui en résulte est le suivant :

	Nombre de pages $\leq 236,5$	Nombre de pages $\succ 236,5$	Total
Ouvrages de médecine	10	5	15
Ouvrages d'histoire	4	9	13
Total	14	14	28

• La taille des classes, toutes supérieures ou égales à 5, autorise l'usage de la *distance de chi-deux* pour répondre de façon bilatérale à l'acceptation ou non de l'hypothèse nulle H_0 : « *épaisseur identique pour les deux types d'ouvrages* », tout en précisant ici que c'est autour de la *tendance centrale que forme la médiane* que cette épaisseur est appréciée.

Le test de chi-deux conduit à mettre en regard les *observations constatées* 10 – 5 – 4 – 9 et les *observations théoriques* sous l'hypothèse H_0, ces dernières étant obtenues par règle de trois, soient 7,5 – 7,5 – 6,5 – 6,5, données résumées ci-dessous :

	Nombre de pages $\leq 236,5$	Nombre de pages $\succ 236,5$	Total
Ouvrages de médecine	10 *(7,5)*	5 *(7,5)*	15
Ouvrages d'histoire	4 *(6,5)*	9 *(6,5)*	13
Total	14	14	28

Il en résulte la distance de chi-deux calculée $\chi^2_{calculé} = \frac{(10-7,5)^2}{7,5} + ... + \frac{(9-6,5)^2}{6,5} = 3,59$ ou plutôt, **avec** la correction de continuité de YATES, $\chi^2_{calculé} = 2,30$ (car pour l'échantillon considéré n n'est pas grand). C'est d'ailleurs, le même résultat qu'on retrouve à partir de la formule $\chi^2_{calculé} = \frac{n.(|a.d - b.c| - \frac{n}{2})^2}{(a+b).(c+d).(a+c).(b+d)}$ (cf. rappels de cours et exercice 15 précédent).

• Or, χ^2 suit la loi du chi-deux à un degré de liberté (c'est en effet, $(r-1).(k-1)$ avec $r=2, k=2$). La consultation des tables annexes (cf. tables de valeurs du chi-deux) conduit, dans ces conditions et pour le test bilatéral de niveau de signification $\alpha = 5\%$, au seuil $\chi^2_\alpha = 3,84$.

Suivant la variante « *conservatrice* » de la distance de chi-deux (celle qui tient compte de la correction de continuité de YATES), on ne peut donc pas conclure à une différence significative entre les deux types d'ouvrages puisque $\chi^2_{calculé} = 2,30 \prec \chi^2_\alpha = 3,84$.

• A défaut de recourir à la distance de chi-deux, la **méthode exacte de FISHER** déjà développée dans l'application 4.5 du présent chapitre, permet également de conclure et d'ailleurs son utilisation est incontournable lorsque les effectifs des classes sont petits ($n \prec 20$).

17. Relativement à leur choix pour un candidat donné à une élection, 19 personnes qui participent à une réunion de quartier avec ledit candidat sont sondées avant et après ce débat, les résultats en étant indiqués ci-dessous :

Electeur	1	2	3	4	5	6	7	8	9	10
Intention « avant »	OUI	OUI	OUI	NON	NON	OUI	NON	NON	OUI	OUI
Intention « après »	NON	OUI	NON	NON	NON	OUI	OUI	NON	NON	OUI
Electeur	11	12	13	14	15	16	17	18	19	
Intention « avant »	OUI	OUI	OUI	NON	OUI	OUI	NON	NON	OUI	
Intention « après »	NON	NON	NON	NON	NON	OUI	NON	OUI	NON	

Au niveau de signification $\alpha = 0,05$, la réunion en question a-t-elle ou non un impact sur les intentions de vote des participants ?

Solution : 1°) Plutôt que le *test des signes*, c'est le **test de MAC NEMAR** qui semble le plus approprié ici encore que les effectifs sont tous justes suffisants. Se référant aux rappels de cours (cf.paragraphe 3.3.c) et à l'application 4.6, le mode opératoire en consiste tout d'abord à retranscrire les données sous forme d'une *table de contingences* $(2,2)$ mettant en lumière les variations d'effectifs entre « *avant* » et « *après* » le débat considéré. Le résultat en est :

AVANT	APRES		Σ
	OUI	NON	
OUI	4	8	12
NON	2	5	7
Σ	6	13	19

AVANT	APRES		Σ
	OUI	NON	
OUI	A	B	$A+B$
NON	C	D	$C+D$
Σ	$A+C$	$B+D$	N

Avec la notation générale qui est celle des rappels de cours (cf. paragraphe 3.3.c), la comparaison des *effectifs observés* des transitions OUI → NON, soit B et NON → OUI soit C, à *l'effectif théorique* $\frac{B+C}{2}$, conduit à la *distance de chi-deux*, $\chi^2_{calculé} = \frac{(B-C)^2}{B+C}$, ou plutôt, avec la *correction de continuité* de YATES, la *distance corrigée* $\chi^2_{calculé} = \frac{(|B-C|-1)^2}{B+C}$. D'autre part, toutes les classes considérées sont suffisamment significatives pour autoriser l'usage du test de χ^2 (puisque la condition $N.p_i \geq 5$ dont l'écriture est ici $\frac{B+C}{2} = \frac{8+2}{2} = 5 \geq 5$ est satisfaite).

En définitive, on a numériquement $\chi^2_{calculé} = 3,6$ ou encore, avec la correction de continuité, $\chi^2_{calculé} = 2,5$. Or, pour le **test bilatéral** et la loi du chi-deux à $\nu = 1$ degré de liberté, le **seuil critique** $\chi^2_\alpha / \Pr ob(\chi^2(1) \geq \chi^2_\alpha) = 0,05$ est égal à $\chi^2_\alpha = 3,84$ (cf. tables de valeurs annexées). Il est donc manifeste qu'on ne peut pas conclure ici à un impact de la réunion de quartier sur les intentions de vote, du moins au seuil $\alpha = 5\%$, puisque $\chi^2_{calculé} = 3,6$ (et à fortiori $\chi^2_{calculé} = 2,5$ lorsqu'on applique la correction de continuité) ne sont pas supérieurs au seuil critique $\chi^2_\alpha = 3,84$.

• Quant au **test des signes**, dont les principes sont détaillés en rappels de cours (cf. paragraphe 3.3.a), il conduit après avoir éliminé les quatre paires OUI→OUI et les cinq paires NON→NON, à considérer le nombre de paires NON→OUI parmi les dix paires restantes, nombre dont sous l'hypothèse H_0 « *réunion de quartier sans impact sur le vote des participants* », la loi est **binomiale**, soit $B(n = 10, p = \frac{1}{2})$.

Or, pour ladite variable aléatoire, soit D, et sa valeur observée, soit $d = 2$, on a par définition, $\Pr ob(D \leq 2) = (C^0_{10} + C^1_{10} + C^2_{10}).\frac{1}{2^{10}}$, soit par consultation de la table de valeurs annexée (cf. « *valeurs critiques du test binomial* »), $\Pr ob(D \leq 2) = 0,055$. Comparant cette probabilité à $\frac{\alpha}{2} = 0,025$ (0,025 et non 0,05 puisque c'est le test bilatéral qui est considéré ici), on en conclut à l'hypothèse H_0 puisqu'on n'a pas $\Pr ob(D \leq 2) \leq 0,025$.

Cela rejoint les conclusions du test de MAC NEMAR.

18. On cherche à savoir si la concordance des résultats de deux tests légèrement différents de la ventilation pulmonaire est aussi bonne chez les malades qui présentent une insuffisance respiratoire que chez les témoins non malades.

Dans un groupe témoin de 103 non malades, le coefficient de corrélation entre les résultats des deux tests est égal à 0,776.

Dans un groupe de 36 malades, les résultats ont été les suivants :

- variance des résultats du premier test : 25 ;
- variance des résultats du deuxième test : 100 ;
- covariance des résultats des deux tests : 30.

1°) Calculer le coefficient de corrélation linéaire dans le groupe de 36 malades.

2°) Comparer les deux coefficients de corrélation en question, au risque $\alpha = 5\%$.

Solution : 1°) De la définition du *coefficient de corrélation linéaire* de BRAVAIS PEARSON, soit $r_{X,Y} = \frac{\text{cov}(X,Y)}{\sqrt{Var(X).Var(Y)}}$, il résulte numériquement, pour les données proposées et pour le groupe des malades, $r_{X,Y} = \frac{30}{\sqrt{25 \times 100}} = 0,6$.

2°) Notant par r_1 et r_2 les coefficients de corrélation qui correspondent respectivement aux « *non malades* » et aux « *malades* » et par n_1 et n_2 les tailles des échantillons dans chacune des populations considérées, la question posée est celle du **test de comparaison bilatéral** $\begin{cases} H_0 : r_1 = r_2 \\ H_1 : r_1 \neq r_2 \end{cases}$.

Or comme cela est indiqué en rappels de cours (cf. paragraphe 2.3.d), pour n asez grand, la **transformée de FISHER**, $z=\frac{1}{2}.\ln\frac{1+r}{1-r}$ suit la **loi normale** de moyenne $\frac{1}{2}.\ln\frac{1+r_{calculé}}{1-r_{calculé}}$ et de variance $\frac{1}{n-3}$. Ainsi, pour le groupe des 103 non malades, obtient-on la loi normale z_1 de moyenne $\frac{1}{2}.\ln\frac{1+0,776}{1-0,776}=1,035$ et d'écart-type $\frac{1}{\sqrt{n_1-3}}=\frac{1}{10}$. De même, pour le groupe des 36 malades, la loi normale z_2 correspondante a pour moyenne $\frac{1}{2}.\ln\frac{1+0,6}{1-0,6}=0,693$ et pour variance $\frac{1}{\sqrt{n_2-3}}=\frac{1}{\sqrt{33}}=0,174$.

En définitive, considérant la variable z_1-z_2 dont la loi est **normale** (puisque différence de deux variables normales indépendantes), plus précisément de moyenne $E(z_1)-E(z_2)$ et de variance $\frac{1}{n_1-3}+\frac{1}{n_2-3}$, la comparaison de r_1 et de r_2 équivaut à tester la nullité ou non de la moyenne de la variable z_1-z_2.

Se référant aux rappels de cours (cf. paragraphe 2.4.b), la *fonction discriminante* à considérer étant z_1-z_2, la **région critique** a pour forme $|z_1-z_2|\geq\pi$, avec $\Pr ob(|z_1-z_2|\geq\pi/H_0)=\alpha$. Or, sous l'hypothèse H_0, $\frac{z_1-z_2}{\sqrt{\frac{1}{n_1-3}+\frac{1}{n_2-3}}}$ suit la *loi normale centrée réduite* $N(0,1)$. Notant par t_α le seuil vérifiant $\Pr ob(|\xi|\geq t_\alpha)=\alpha$, soit $t_\alpha=1,96$ lorsque $\alpha=5\%$, on a donc $\pi=t_\alpha.\sqrt{\frac{1}{n_1-3}+\frac{1}{n_2-3}}$, soit numériquement $\pi=1,96.\sqrt{\frac{1}{100}+\frac{1}{33}}=0,393$.

- Or, pour les observations effectuées, $|r_{1,calculé}-r_{2,calculé}|=|1,035-0,693|=0,342$. La relation $O,342\prec\pi=0,393$ ne permet pas de rejeter l'hypothèse H_0 d'égalité entre les deux coefficients de corrélation à travers les données recueillies ici.

19. La relation entre l'autoritarisme des étudiants et leur conformisme social est étudiée à travers deux tests dont les résultats sont présentés ci-dessous à travers un nombre total de points. L'échantillon considéré est de 12 étudiants.

Etudiant	1	2	3	4	5	6	7	8	9	10	11	12
Evaluation suivant test de l'autoritarisme (X)	82	98	87	40	116	113	111	83	85	126	106	117
Evaluation suivant test de conformisme (Y)	42	46	39	37	65	88	86	56	62	92	54	81

1°) Calculer le coefficient tau de KENDALL.

2°) Peut-on considérer, au niveau de signification $\alpha=5\%$, qu'il existe une association entre l'autoritarisme et le conformisme social des étudiants ?

Solution : 1°) La transcription des données proposées en termes de rangs conduit au tableau ci-dessous :

Etudiant	1	2	3	4	5	6	7	8	9	10	11	12
Rang de X	2	6	5	1	10	9	8	3	4	12	7	11
Rang de Y	3	4	2	1	8	11	10	6	7	12	5	9

Ordonnant par *ordre croissant* les rangs de X, il en résulte le nouveau tableau :

Etudiant	4	1	8	9	3	2	11	7	6	5	12	10
Rang de X	1	2	3	4	5	6	7	8	9	10	11	12
Rang de Y	1	3	6	7	2	4	5	10	11	8	9	12

La **méthode graphique des intersections** (cf. rappels de cours et application 4.8 du présent chapitre), conduit à 11 intersections « *paires discordantes* » et à fortiori à 66 - 11=55 « *paires concordantes* », le nombre total d'associations deux à deux étant $C_n^2 = \frac{n.(n-1)}{2}$, soit 66 (puisque $n = 12$).

En définitive, $\tau = \frac{nombre\ de\ concordances - nombre\ de\ discordances}{C_n^2}$ est égal à $\frac{55-11}{66} = 0,66$.

La valeur absolue $|\tau|$ variant entre 0 (*indépendance*) et 1 (*liaison totale*), la valeur obtenue ici laisse supposer qu'il y a bien une association entre les deux variables étudiées, ce que va confirmer la question suivante.

2°) Le **test d'indépendance** entre X et Y se ramène au **test bilatéral** *de nullité ou non* de τ, test dont suivant les rappels de cours (cf. paragraphe 3.4.b), la **région critique** est caractérisée par $|\tau| \geq t_\alpha . \sqrt{\frac{2.(2n+5)}{9n.(n-1)}}$ où t_α vérifie $Prob(|\xi| \geq t_\alpha) = \alpha$. A noter que la condition $n \geq 10$ qui valide la convergence vers la loi normale, est satisfaite ici.

Numériquement, $\alpha = 5\% \Rightarrow t_\alpha = 1,96$, d'où le seuil critique $\pi = 0,43$. Or relativement aux données dont on dispose, $|\tau_{calculé}| = 0,66 \succ \pi = 0,43$. C'est donc l'hypothèse H_1 qu'on peut retenir ici, c'est-à-dire la confirmation de l'association entre X et Y.

20. Un chercheur a soumis quatre groupes de cinq élèves à un apprentissage de résolution de problèmes mathématiques. Chaque groupe apprend avec une méthode pédagogique propre :

- méthode A : uniquement verbale ;
- méthode B : méthode écrite ;
- méthode C : méthode suivant un schéma annoté ;
- méthode D : méthode suivant plusieurs schémas annotés.

L'apprentissage dure une heure pour chaque groupe et le même contenu est présent. Deux jours, après l'apprentissage, les sujets sont soumis à un test de raisonnement mathématique dont l'évaluation des résultats donne lieu à une note étalonnée entre 0 et 35 (plus la note est élevée, meilleur est le résultat).

Les notes ainsi obtenues sont :

Groupe expérimental			
A	B	C	D
6	14	22	23
13	10	11	19
16	14	19	25
14	19	19	24
14	25	23	25

On admettra par ailleurs, les hypothèses de normalité de la distribution des notes et l'indépendance des observations effectuées. De même, on suppose satisfaite l'homogénéité des variances entre groupes (homoscédasticité). Enfin, on chosira pour toutes les questions posées ci-après, un niveau de signification α fixé à 5%.

1°) Les quatre méthodes considérées ici sont-elles équivalentes ?

2°) La méthode verbale diffère-t-elle des autres méthodes ?

3°) La méthode écrite diffère-t-elle des méthodes avec schémas (un ou plusieurs) ?

4°) Le nombre de schémas a-t-il une influence décelable sur la performance ?

Solution : 1°) La mise en euvre du **test « ANOVA » de FISHER** (test paramétrique dont la validité est assurée du fait de la satisfaction des conditions d'indépendance, de normalité, et d'homoscédasticité, conduit aux calculs ci-après (cf. rappels de cours, paragraphe 2.6.a).

En premier lieu, il faut calculer les sommes des écarts « *intraclasses* », soit la somme $S_W = \sum_{i=1}^{i=K}\sum_{j=1}^{j=n_i}(x_{ij}-\overline{x_i})^2$, et des écarts « *interclasses* », soit la somme $S_B = \sum_{i=1}^{i=K} n_i.(x_i-\overline{x})^2$, les valeurs de K et des $n_i (1 \le i \le 4)$, étant respectivement $K=4, n_1=n_2=n_3=n_4=5$, pour le cas étudié. Il s'ensuit le tableau des calculs intermédiaires ci-dessous :

Statistique	Groupe				Σ
	A	B	C	D	
$\overline{x_i} = \frac{\sum_{j=1}^{j=n_i} x_{ij}}{n_i}$	12,6	16,4	18,8	23,2	
$\sum_{j=1}^{j=n_i}(x_{ij}-\overline{x_i})^2$	59,2	133,2	88,8	24,8	306
$S_i^2 = \frac{\sum_{j=1}^{j=n_i}(x_{ij}-\overline{x_i})^2}{n_i-1}$	14,8	33,3	22,2	6,2	
$n_i.(x_i-\overline{x})^2$	132,6	9,1	5,5	148,5	

D'où, $\overline{x} = \frac{\sum_{i=1}^{i=K} n_i.\overline{x_i}}{N}$, avec $N = \sum_{i=1}^{i=K} n_i$, soit $\overline{x} = 17,75$. De même, $S_W = 306$ et $S_B = 295,75$.

Le test $\begin{cases} H_0 : m_i = m_j, \forall (i,j) \\ H_1 : \exists (i,j) / m_i \neq m_j \end{cases}$ conduit à la *fonction discriminante* $F = \dfrac{S_B / K-1}{S_W / N-K}$, soit numériquement et pour les données proposées $(K=4, N=20)$, $F_{calculé} = 5,15$.

• Or, sous l'hypothèse H_0, F suit la **loi de FISHER SNEDECOR** à $(\nu_1 = 3, \nu_2 = 16)$ degrés de liberté, soit $F(3,16)$, le seuil F_α qui, pour la loi en question, vérifie $Prob(F \geq F_\alpha) = \alpha$, étant immédiatement égal à 3,24 lorsque $\alpha = 5\%$ (cf. lecture dans table de valeurs annexée).

C'est donc l'hypothèse H_1 qui est celle d'une différence significative pour au moins deux des groupes considérés qu'il faut retenir ici puisque $F_{calculé} = 5,15 \succ F_\alpha = 3,24$.

2°) Pour tester si la *méthode verbale* est **différente** des autres méthodes, on va utiliser la **méthode des contrastes de SCHEFFE** (cf. application 5.1 de ce chapitre) avec la combinaison linéaire $3.m_1 - m_2 - m_3 - m_4 = 0$. Il s'ensuit le *contraste calculé*, $\widehat{C} = 3.\overline{x_1} - \overline{x_2} - \overline{x_3} - \overline{x_4} = -20,6$.

Se référant aux rappels de cours (cf. paragraphe 2.6.a), la **région critique** du test (zone de décision de H_1) est caractérisée par $\left|\widehat{C}\right| \geq \pi = \sqrt{(K-1).F_\alpha . \dfrac{S_W}{N-K} . \sum_{i=1}^{i=K} \dfrac{C_i^2}{n_i}}$ où F_α vérifie, pour la loi de FISHER SNEDECOR, $F(K-1, N-K)$, la relation $\Pr ob(F \geq F_\alpha) = \alpha$.

Pour $\alpha = 0,05 \Rightarrow F_\alpha = 3,24$, on a ainsi $\pi = \sqrt{3 \times 3,24 \times \dfrac{306}{16} \times (\dfrac{9}{5} + \dfrac{1}{5} + \dfrac{1}{5} + \dfrac{1}{5})} = 21,12$. On ne peut donc pas conclure à un contraste significatif entre la *méthode verbale* et les autres méthodes, du moins au seuil $\alpha = 5\%$, puisque $\left|\widehat{C}\right| = 20,6 \prec \pi = 21,12$.

• On n'est pas loin cependant de la région critique. En effet, pour $\left|\widehat{C}\right| = 20,6$, on a $F_{observé} = 3,08$. Or, $\Pr ob(F \geq 3,08) = 0,058$. Autrement dit, au niveau de signification $\alpha = 6\%$, c'est H_1 (et non plus H_0) qu'on retiendrait.

3°) Dans cette question, il faut comparer la méthode B à l'ensemble des deux méthodes C et D, problème auquel la *méthode des contrastes* répond par me choix de la combinaison linéaire $2.m_2 - m_3 - m_4$. Il en résulte, le contraste calculé $\widehat{C} = 2.\overline{x_2} - \overline{x_3} - \overline{x_4} = -9,2$. D'autre part, on a cette fois $\pi = \sqrt{3 \times 3,24 \times \dfrac{306}{16} \times (\dfrac{4}{5} + \dfrac{1}{5} + \dfrac{1}{5})} = 14,94$.

La relation $\left|\widehat{C}\right| = 9,2 \prec \pi = 14,94$ ne permet pas de se prononcer ici sur une différence significative entre la *méthode verbale* et les *méthodes avec schémas* (un ou plusieurs).

4°) Enfin, tester si le nombre de schémas influe ou non sur la performance, c'est étudier l'existence ou non d'une différence significative entre les méthodes C et D, ce qui, toujours avec la *méthode des contrastes*, conduit à la combinaison linéaire $m_3 - m_4 = 0$ et au contraste calculé, $\widehat{C} = \overline{x_3} - \overline{x_4} = -4,5$. Il en résulte, le seuil critique $\pi = \sqrt{3 \times 3,24 \times \dfrac{306}{16} \times (\dfrac{1}{5} + \dfrac{1}{5})} = 8,62$.

Ici encore, la relation $\left|\widehat{C}\right| = 4,5 \prec \pi = 8,62$ ne permet pas de rejeter l'hypothèse H_0 et de conclure à une différence significative entre les méthodes C et D.

21. On considère trois traitements appliqués respectivement à des groupes de 5, 5, et 4 malades, l'évaluation des effets desdits traitements suivant une cotation qui évolue entre 0 et 100, donnant les résultats ci-après :

Traitement A	11	12	14	48	4
Traitement B	4	5	3	8	4
Traitement C	3	1	0	8	

1°) Les trois traitements ont-ils des effets différents au niveau de signification $\alpha = 5\%$?

2°) Qu'en est-il au niveau de signification $\alpha = 1\%$?

Solution : 1°) Supposant l'indépendance des mesures relatives aux malades considérés, l'absence d'indication sur la normalité ou non, suggère d'utiliser ici le **test non paramétrique de KRUSKAL-WALLIS** (cf. rappels de cours, paragraphe 3.2.d, et application 5.2 du présent chapitre).

Après regroupement des données eu un seul échantillon, classement par ordre croissant et affectation du rang selon le *principe du rang moyen* en cas de valeurs ex-aequo, il résulte pour les valeurs susmentionnées de l'énoncé, les rangs suivants :

Traitement A	6	11	12	13	14
Traitement B	3,5	6	6	8	9,5
Traitement C	1	2	3,5	9,5	

Il s'ensuit les *rangs moyens* $\overline{R_1} = 11,2; \overline{R_2} = 6,6; \overline{R_3} = 4,0$, le *rang moyen théorique* $\overline{R} = \dfrac{\sum_{i=1}^{i=N} i}{N} = \dfrac{N+1}{2}$ étant égal, quant à lui, à $\dfrac{14+1}{2} = 7,5$.

• La statistique $H = \dfrac{12}{N.(N+1)} . \sum_{i=1}^{i=K} n_i . (\overline{R_i} - \overline{R})^2$ décrit la somme pondérée des distances entre les $\overline{R_i}$ et le rang moyen théorique $\overline{R} = 7,5$, sa valeur calculée à partir des données de l'échantillon étant immédiatement :

$$H_{calculé} = \frac{12}{14 \times 15} . \left[5 \times (11,2 - 7,5)^2 + 5 \times (6,6 - 7,5)^2 + 4 \times (4,0 - 7,5)^2 \right] = 6,94 .$$

Compte tenu des trois valeurs ex-aequo, le coefficient correctif à appliquer est défini par $C = 1 - \dfrac{1}{14 \times (14^2 - 1)} . \left[(3^3 - 3) + (2^3 - 2) + (2^3 - 2) \right]$, soit numériquement, $C = 0,9868$ (cf. rappels de cours, paragraphe 3.2.d). Ainsi, après correction, la statistique qni décrit les écarts entre les rangs $\overline{R_i}$ et le rang moyen théorique $\overline{R}$, est-elle, pour les données étudiées, $H'_{calculé} = \dfrac{H}{C} = 7,03$.

• Admettant que H suit, sous l'hypothèse H_0, la **loi du chi-deux** à $K-1$ degrés de liberté, le seuil χ^2_α vérifiant $\Pr ob(H \geq \chi^2_\alpha) = \alpha$ est immédiatement, pour $\alpha = 5\%$, $\chi^2_\alpha = 5,99$ (cf. lecture dans table des valeurs de la loi $\chi^2(2)$ annexée).

Toutefois, lorsque les échantillons considérés ont, pour certains, une taille inférieure à la valeur 5, les conditions d'application de la statistique de KRUSKAL-WALLIS ne sont plus remplies et il est recommandé alors de se reporter à des tables spécifiques.

De telles tables consultables en ligne, fournissent ainsi pour $n_1 = 5, n_2 = 5, n_3 = 4$, les seuils 5,64 et 7,80 respectivement pour $\alpha = 5\%$ et $\alpha = 1\%$.

En tout état de cause, pour $\alpha = 5\%$, $H'_{calculé} \approx 7,03$ est supérieur aux deux seuils précédents (suivant χ^2, soit 5,99, et suivant table spécifique, soit 5,64), ce qui conduit à retenir la conclusion de différences significatives entre au moins deux des trois traitements considérés.

• Une analyse de ces différences, deux à deux, telle celle menée dans l'application 5.2 permettrait des investigations plus précises quant à la nature de la ou des différences en question.

2°) Pour $\alpha = 1\%$, la table de χ^2 fournit lorsque $\nu = K - 1 = 2$ degrés de liberté, le seuil $\chi^2_\alpha = 9,21$, seuil dont il est rappelé (cf. 1ère question), qu'il est de 7,80 lorsqu'on se réfère à la table de valeurs spécifiques applicable aux petits échantillons. Au demeurant, la relation $H'_{calculé} = 7,03 \prec 7,80$ ne permet plus ici de conclure à l'hypothèse H_1.

• C'est un résultat somme toute logique, que la diminution du risque de rejeter à tort l'hypothèse nulle H_0, entraîne une attitude plus conservatrice sur ladite hypothèse nulle et conduise ainsi à remettre en cause la décision de rejet retenue lorsque $\alpha = 5\%$.

22. On demande à trois mélomanes d'une revue d'écouter six versions différentes d'une symphonie de BEETHOVEN et pour chacune de classer suivant l'ordre préférentiel croissant variant de 1 à 6, l'organisation des six plans sonores (qui ressort de la disposition spatiale des instruments et à fortiori du chef d'orchestre).

Les trois séries indépendantes de rangs ainsi obtenus compte tenu des classements des trois mélomanes A,B,C contactés, sont rassemblées ci-dessous :

	a	B	c	d	e	f
A	1	6	3	2	5	4
B	1	5	6	4	2	3
C	6	3	2	5	4	1

Les six versions sont-elles appréciées ou non de la même façon ?

Solution : Ce sont les mêmes éléments « i » $(1 \le i \le 3)$ à qui on applique les différents traitements « j » $(1 \le j \le 6)$, formant donc ainsi des **données appariées**. Ces données étant par ailleurs de type **ordinal**, c'est finalement le **test non paramétrique de FRIEDMAN** qu'il convient d'appliquer ici dans les conditions décrites en rappels de cours du présent chapitre (cf. paragraphe 3.3.e) et dans l'application 5.3.

Il en résulte, pour chacune des versions a,b,c,d,e,f testées, les rangs moyens ci-dessous :

Rang moyen\Version	1	2	3	4	5	6
$\overline{R_i}$	2,67	4,67	3,67	3,67	3,67	2,67

Le *rang moyen théorique* étant égal à $\frac{K+1}{2} = 3,5$, il en résulte, pour la statistique de FRIEDMAN, $F = \frac{12.n}{K.(K+1)} . \sum_{j=1}^{j=K} (\overline{R_j} - \overline{R})^2$, la valeur numérique :

$$F_{calculé} = \frac{12 \times 3}{6 \times (6+1)} . \left[(2,67 - 3,5)^2 + ... + (2,67 - 3,5)^2 \right] = 2,43.$$

Or admettant que F suit la **loi de chi-deux** à $\nu = K - 1$ degrés de liberté (soit $\nu = 2$, pour le cas étudié), une lecture dans la table de valeurs annexée, fournit, pour le niveau de signification $\alpha = 5\%$, la valeur du seuil $\chi^2_\alpha / \Pr ob(\chi^2 \geq \chi^2_\alpha) = \alpha$. On lit $\chi^2_\alpha = 5,99$.

La **région critique** du test étant caractérisée par $F \geq \chi^2_\alpha$, la relation $F_{calculé} = 2,43 \prec \chi^2_\alpha = 5,99$ qui résulte des données proposées, ne permet pas en l'occurrence de rejeter l'hypothèse nulle H_0. En d'autres termes, on n'a pas mis en évidence de différence significative d'interprétation (au sens de l'organisation des plans sonores) entre les différentes versions de la symphonie considérée.

• A noter ici encore, l'existence de tables spécifiques du test de FRIEDMAN pour traiter le cas des petits échantillons.

23. Lors d'une étude sur la nature et la conséquence de la stratification sociale dans une petite ville du centre-ouest des Etats-Unis d'Amérique, HOLLINGSHEAD montra que les membres de cette communauté se répartissaient eux-mêmes en quatre classes sociales.

Son étude était centrée sur les corrélats de cette stratification parmi les jeunes. L'une de ses prédictions était que les adolescents des différentes classes sociales s'engageaient dans différentes voies d'étude (général, commercial, préparation à l'université), au lycée de la ville. Cette hypothèse fut testée en identifiant l'appartenance sociale de 390 lycéens et en déterminant leur choix scolaire.

L'hypothèse nulle correspond au fait que la proportion de lycéens inscrits dans chacune des trois filières considérées est la même pour chaque classe sociale. Pour l'hypothèse alternative H_1, la proportion de lycéens inscrits dans chaque filière diffère suivant les classes sociales.

Les données obtenues sont les suivantes :

Filière	Classe				Total
	I	II	III	IV	
Général	11	75	107	14	207
Commercial	1	31	60	10	102
Prépa.Université	23	40	16	2	81
Total	35	146	183	26	390

Qu'en conclure au niveau de signification $\alpha = 5\%$ quant à l'indépendance ou non entre la classe sociale et les voies d'études choisies ?

Solution : Il est proposé de recourir à un **test de contingeuces de chi-deux** (cf. paragraphe 3.4.c des rappels de cours du présent chapitre), les *effectifs théoriques* qui correspondent à l'hypothèse H_0 suivant laquelle la proportion des lycéens inscrits dans chacune des filières est la même pour toutes les classes sociales, étant calculés par règles de trois en fonction des formules habituelles sur l'indépendance des variables aléatoires.

Ainsi, en considérant les totaux en lignes (loi marginale de la variable « *filière* »), et les totaux en colonnes (loi marginale de la variable « *classe sociale* »), a-t-on, par exemple, pour le couple « *choisir la filière générale* » **et** « *appartenir à la classe sociale I* » ; une probabilité égale au produit des lois, soit le produit de Prob(« *choisir la filière générale* »)$=\frac{207}{390}$ par la probabilité Prob(« *appartenir à la classe sociale* I »)$=\frac{35}{290}$.

En résumé, la probabilité en question portant sur le couple « *choisir la filière générale* » **et** « *appartenir à la classe sociale I* », est égale à $\frac{207}{390}\times\frac{35}{390}$.

Mais, pour comparaison avec les *effectifs observés*, ce sont en fait, les effectifs absolus, qu'il faut considérer ici, soit $390\times\frac{207}{390}\times\frac{35}{390}=18,57$. Et ainsi de suite, pour tous les autres couples possibles, les résultats obtenus étant portés dans le tableau ci-dessous en caractères italiques gras.

Filière	Classe				Total
	I	II	III	IV	
Général	11 ***18,57***	75 ***77,49***	107 ***97,13***	14 ***13,80***	207
Commercial	1 ***9,15***	31 ***38,18***	60 ***47,86***	10 ***6,80***	102
Prépa.Université	23 ***7,27***	40 ***30,32***	16 ***38,01***	2 ***5,40***	81
Total	35	146	183	26	390

L'écart entre les *effectifs observés* O_{ij} et les *effectifs théoriques* E_{ij} est mesuré par ailleurs, par la **distance de chi-deux** définie par $D=\sum_{i=1}^{i=r}\sum_{j=1}^{j=k}\frac{(O_{ij}-E_{ij})^2}{O_{ij}}$ (numériquement $r=3,k=4$). A partir des données ci-dessus, il s'ensuit numériquement, la valeur calculée $D_{calculé}=69,39$.

Or, sous l'hypothèse nulle H_0, la statistique D suit la loi de chi-deux à $\nu=(r-1).(k-1)$ degrés de liberté, soit numériquement $\nu=(3-1)\times(4-1)=6$ degrés de liberté. Notant par χ^2_α, le seuil qui, pour la loi considérée, vérifie $\Pr ob(\chi^2\geq\chi^2_\alpha)=\alpha$, il est immédiat par lecture dans la table de valeurs annexée, que $\chi^2_\alpha=12,59$ lorsque $\alpha=5\%$ (de même, si $\alpha=1\%$, on a $\chi^2_\alpha=16,81$).

• Or, pour les données considérées, $D_{calculé}=69,39$ est largement supérieur au seuil critique χ^2_α (même lorsque $\alpha=1\%$). C'est donc, sans conteste, le rejet de l'hypothèse H_0 qu'il faut retenir dans cet exercice, c'est-à-dire la conclusion de l'inscription des lycéens différenciée en fonction de leur classe sociale.

CHAPITRE IV

REGRESSION

A - Rappels de cours

1. Régression linéaire simple

1.1 Le modèle

L'un des schémas les plus courants de l'ajustement est d'expliciter le lien de dépendance entre une grandeur mesurée x (supposée connue) et une grandeur Y dont le caractère aléatoire est imputable à un ensemble de facteurs imprévisibles et non mesurés, le modèle de régression en résultant ayant pour forme $Y = f(x) + \varepsilon$ où :

- Y désigne la variable aléatoire « **expliquée** » dite encore « **observée** » ;
- x désigne la variable « **explicative** » dite encore « **prédicteur** » ;
- ε désigne l'erreur aléatoire de prédiction sur Y dite encore « **résidu** ».

A noter que les raisonnements ci-dessus se généralisent aisément au cas où x est la réalisation d'une variable aléatoire X.

• Ainsi lorsqu'on dispose d'un échantillon de données de taille n, soit $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$, a-t-on un modèle qui s'écrit $Y_i = f(x_i) + \varepsilon_i, \forall i \in \{1, 2, ..., n\}$, les variables Y_i étant supposées *indépendantes* en général. On admettra également le plus souvent que les variables $\varepsilon_i (1 \le i \le n)$, sont *centrées* ($E(\varepsilon_i) = 0, \forall i$), et *de même loi* de variance σ^2.

L'hypothèse $E(\varepsilon_i) = 0$ traduit le fait que les facteurs autres que le prédicteur x, ont des effets qui se compensent. Quant à l'équirépartition des lois des ε_i, elle signifie que les expériences i $(1 \le i \le n)$ ont toutes été réalisées dans les mêmes conditions.

• Le **modèle de régression linéaire simple** par lequel $f(x)$ est *linéaire* est **fondamental**, sa définition en étant $Y_i = a.x_i + b + \varepsilon_i$, $\forall i \in \{1, 2, ..., n\}$.

A noter que c'est la linéarite de f par rapport aux coefficients $a, b, ..$ et non au prédicteur qui est déterminant pour justifier la linéarité ou non du modèle, les exemples ci-dessous formant ainsi des modèles linéaires à quelques adaptations près explicitées en applications du présent chapitre :

- $Y_i = a.x_i^2 + b.x_i + c + \varepsilon_i$;
- $Y_i = a.\ln x_i + b + \varepsilon_i$;
- $Y_i = a + b/x_i + \varepsilon_i$;

...

• Enfin, *classique* est le **modèle linéaire gaussien** dans lequel on suppose la *normalité* des résidus $\varepsilon_i, i \in \{1,2,...,n\}$ (donc de loi commune $N(0,\sigma^2)$), ne serait-ce qu'en raison du *théorème central-limite* et de la théorie des « *causes élémentaires* ».

1.2 Estimation des paramètres

La meilleure représentation affine des Y_i en fonction des $x_i (1 \le i \le n)$ est fournie par la **droite des moindres carrés** d'équation $y = \widehat{a_n}.x + \widehat{b_n}$ où $\widehat{a_n} = \dfrac{\sum_{i=1}^{i=n}(x_i - \overline{x_n}).(y_i - \overline{y_n})}{\sum_{i=1}^{i=n}(x_i - \overline{x_n})^2}$, soit également $\widehat{a_n} = \dfrac{\sum_{i=1}^{i=n} x_i.y_i - n.\overline{x_n}.\overline{y_n}}{\sum_{i=1}^{i=n} x_i^2 - n.\overline{x_n}^2}$, et $\widehat{b_n} = \overline{y_n} - \widehat{a_n}.\overline{x_n}$.

Tout d'abord, il est implicite dans l'énoncé ci-dessus que $\overline{x_n} = \dfrac{\sum_{i=1}^{i=n} x_i}{n}$ et $\overline{y_n} = \dfrac{\sum_{i=1}^{i=n} y_i}{n}$.

Au demeurant, la question posée ici est de *minimiser l'écart entre les prévisions* $f(x_i) = a.x_i + b$ *et les valeurs observées* y_i, l'indicateur qui en résulte étant défini par la somme $U(a,b) = \sum_{i=1}^{i=n}(y_i - a.x_i - b)^2$, ceci suivant la théorie des variables aléatoires dont il est rappelé qu'elle est construite au sens des *moindres carrés* (espace L^2 de fonctions de carré intégrable).

De l'écriture de la nullité des dérivées partielles $\dfrac{\partial U}{\partial a}$ et $\dfrac{\partial U}{\partial b}$ résulte le système d'équations :

$$\begin{cases} a.\sum_{i=1}^{i=n} x_i + n.b = \sum_{i=1}^{i=n} y_i \\ a.\sum_{i=1}^{i=n} x_i^2 + b.\sum_{i=1}^{i=n} x_i = \sum_{i=1}^{i=n} x_i.y_i \end{cases}$$

On en déduit les valeurs $\widehat{a_n}$ et $\widehat{b_n}$ annoncées.

On remarquera à cet égard, que ces valeurs ne sont que les traductions empiriques de l'équation de la *droite de régression* $Z = \dfrac{\mathrm{cov}(X,Y)}{Var(X)}.[X - E(X)] + E(Y)$ qui lie deux variables aléatoires X et Y.

1.3 Erreur moyenne

Cette erreur qui est représentée par la **norme** $\|\varepsilon\|^2$ est définie, au sens **quadratique** moyen (norme habituelle en la circonstance), par la *moyenne de la somme des carrés des écarts* entre les valeurs observées y_i et les valeurs théoriques fournies par le modèle, soient $\widehat{y_i} = \widehat{a_n}.x_i + \widehat{b_n}$. Ainsi $\|\varepsilon\|^2 = \dfrac{1}{n}.\sum_{i=1}^{i=n}(y_i - \widehat{a_n}.x_i - \widehat{b_n})^2$.

• On a $\|\varepsilon\|^2 = s_y^2.(1-r_{xy}^2)$ où s_y^2 et r_{xy}^2 représentent respectivement la *variance empirique* de Y et le *coefficient de corrélation empirique* de X et Y définis respectivement par $s_y^2 = \frac{1}{n}.\sum_{i=1}^{i=n}(y_i - \overline{y})^2$ et $r_{xy}^2 = \frac{\sum_{i=1}^{i=n} x_i.y_i - n.\overline{x_n}.\overline{y_n}}{(\sum_{i=1}^{i=n} x_i^2 - n.\overline{x_n}^2).(\sum_{i=1}^{i=n} y_i^2 - n.\overline{y_n}^2)}$.

Notant respectivement par c_{xy}, s_x^2, et s_y^2, la covariance et les variances empiriques (non corrigées) liées à X et Y, on a immédiatement $y_i = \frac{c_{xy}}{s_x^2}.\left[x_i - \overline{x_n}\right] + \overline{y_n}$.

Il s'ensuit $\|\varepsilon\|^2 = \frac{1}{n}.\sum_{i=1}^{i=n}\left[y_i - \frac{c_{xy}}{s_x^2}.(x_i - \overline{x_n}) - \overline{y_n}\right]^2$, soit en développant, l'expression :

$$\|\varepsilon\|^2 = \frac{1}{n}.\left[\sum_{i=1}^{i=n}(y_i - \overline{y_n})^2 + \frac{c_{xy}^2}{s_x^4}.\sum_{i=1}^{i=n}(x_i - \overline{x_n})^2 - 2.\frac{c_{xy}}{s_x^2}.\sum_{i=1}^{i=n}(x_i - \overline{x_n}).(y_i - \overline{y_n})\right].$$

Finalement, $\|\varepsilon\|^2 = s_y^2 + \frac{c_{xy}^2}{s_x^4}.s_x^2 - 2.\frac{c_{xy}^2}{s_x^2}$, soit $\|\varepsilon\|^2 = s_y^2 - \frac{c_{xy}^2}{s_x^2} = s_y^2.\left[1 - r_{xy}^2\right]$, ce qui établit le résultat annoncé.

1.4 Interprétation du coefficient de corrélation « empirique »

Semblable au coefficient de corrélation $\rho_{XY} = \frac{\mathrm{cov}(X,Y)}{\sqrt{VarX.VarY}}$, le *coefficient de corrélation empirique* r_{xY} décrit la **dépendance linéaire** ou non entre x et Y :

- $r_{xy}^2 = 1$ équivaut à $\|\varepsilon\|^2 = 0$, c'est-à-dire l'alignement des points (x_i, y_i) suivant une droite de pente positive si $r_{xy} = +1$ et de pente négative si $r_{xy} = -1$.
- $r_{xy}^2 = 0$ lorsque Y et x sont indépendantes. Réciproquement, la nullité de r_{xy} n'entraîne pas l'indépendance, mais elle signifie seulement qu'il n'y a pas dépendance linéaire entre x et Y.

1.5 Coefficient de détermination (coefficient de corrélation généralisé) et analyse de la variance

Le *coefficient de détermination* R^2 est défini par le *rapport* entre **l'écart expliqué** $\sum_{i=1}^{i=n}(\widehat{y_i} - \overline{y_n})^2$, c'est-à-dire *l'écart du au modèle* (c'est un écart maîtrisé), et **l'écart total** $\sum_{i=1}^{i=n}(y_i - \overline{y_n})^2$, la différence entre ces deux écarts, soit $\sum_{i=1}^{i=n}(y_i - \widehat{y_i})^2$ formant par ailleurs, *l'écart inexpliqué* (écart aléatoire) dit encore « **écart résiduel** ».

• Le coefficient de détermination R^2 varie entre 0 et 1 et pour le cas de la régression linéaire simple, on a $R^2 = r_{xy}^2$.

De $\|\varepsilon\|^2 = s_y^2.(1-r_{xy}^2)$ (cf. paragraphe 1.4 précédent), on déduit immédiatement l'expression $r_{xy}^2 = 1 - \frac{\|\varepsilon\|^2}{s_y^2}$.

Ainsi $r_{xy}^2 = 1 - \dfrac{\sum_{i=1}^{i=n}(y_i - \widehat{y_i})^2}{\sum_{i=1}^{i=n}(y_i - \overline{y_n})^2}$. Or, reprenant *l'équation de l'analyse de la variance* (cf. rappels de cours du chapitre III, paragraphe 2.6), on a immédiatement :

$$\sum_{i=1}^{i=n}(y_i - \overline{y_n})^2 = \sum_{i=1}^{i=n}(y_i - \widehat{y_i} + \widehat{y_i} - \overline{y_n})^2 = \sum_{i=1}^{i=n}(y_i - \widehat{y_i})^2 + \sum_{i=1}^{i=n}(\widehat{y_i} - \overline{y_n})^2 .$$

En effet, le double produit $2.\sum_{i=1}^{i=n}(y_i - \widehat{y_i}).(\widehat{y_i} - \overline{y_n})$ dont le développement conduit à l'expression $2.\sum_{i=1}^{i=n}(y_i - a.x_i - b).(a.x_i + b - \overline{y_n})$ est nul de par les équations des moindres carrés $\begin{cases} a.\sum_{i=1}^{i=n} x_i + n.b = \sum_{i=1}^{i=n} y_i \\ a.\sum_{i=1}^{i=n} x_i^2 + b.\sum_{i=1}^{i=n} x_i = \sum_{i=1}^{i=n} x_i.y_i \end{cases}$.

Plus précisément, on a en explicitant le développement susmentionné, la décomposition $(b - \overline{y_n}).\sum_{i=1}^{i=n}(y_i - a.x_i - b) + a.\sum_{i=1}^{i=n}(x_i.y_i - a.x_i^2 - b.x_i)$, chacune des sommes étant nulle compte tenu du système d'équations antérieur.

Revenant à $r_{xy}^2 = 1 - \dfrac{\sum_{i=1}^{i=n}(y_i - \widehat{y_i})^2}{\sum_{i=1}^{i=n}(y_i - \overline{y_n})^2} = \dfrac{\sum_{i=1}^{i=n}(y_i - \overline{y_n})^2 - \sum_{i=1}^{i=n}(y_i - \widehat{y_i})^2}{\sum_{i=1}^{i=n}(y_i - \overline{y_n})^2}$, c'est donc également suite au résultat obtenu ci-dessus, $r_{xy}^2 = \dfrac{\sum_{i=1}^{i=n}(\widehat{y_i} - \overline{y_n})^2}{\sum_{i=1}^{i=n}(y_i - \overline{y_n})^2}$, soit le coefficient de détermination R^2 caractérisé par le rapport $\dfrac{\text{écart expliqué}}{\text{écart total}}$. Ceci achève la démonstration du résultat annoncé.

- Le coefficient de détermination R^2 offre l'avantage de sa clarté et d'être *généralisable aux régressions non linéaires et multiples*. Il représente *l'évaluation de la contribution du modèle de régression considéré à la valeur de la variable dépendante* (variable expliquée). Plus les écarts expliqués s'approchent de l'écart total, plus performant est le modèle considéré, R^2 approchant alors la valeur unité.

- Enfin, **l'analyse de la variance** est applicable aux modèles de régression simple et même multiple. La statistique $F = (n-2).\dfrac{R^2}{1-R^2}$ suit la **loi de FISHER SNEDECOR** à $\nu_1 = 1, \nu_2 = n-2$ degrés de liberté, le test sur une régression significative se résumant à $\begin{cases} H_0 : a = 0 \\ H_1 : a \neq 0 \end{cases}$ et ayant pour *région critique* $F \geq F_\alpha$, F_α caractérisé par $\Pr ob(F \geq F_\alpha) = \alpha$.

L'équation de l'analyse de la variance, $\sum_{i=1}^{i=n}(y_i-\overline{y_n})^2=\sum_{i=1}^{i=n}(y_i-\widehat{y_i})^2+\sum_{i=1}^{i=n}(\widehat{y_i}-\overline{y_n})^2$ s'écrit sous la forme SCT=SCE + SCR où **SCT** est la somme des *écarts totaux*, **SCE** celle des *écarts résiduels*, et **SCR** celle des *écarts expliqués*.

Reprenant les résultats des rappels de cours du chapitre III relatifs à l'ANOVA (cf. paragraphe 2.6), $\frac{SCR}{2-1}$ et $\frac{SCE}{n-2}$ forment des estimateurs sans biais des variances expliquées et non expliquées de Y.

Sous l'hypothèse $H_0 : a=0$, le rapport $F=\frac{SCE/(2-1)}{SCR/(n-2)}=(n-2).\frac{R^2}{1-R^2}$ suit la loi de **FISHER SNEDECOR** à $(1,n-2)$ degrés de liberté. Pour une régression significative (hypothèse H_1), la part de la variance de Y qui est expliquée est supérieure à la part aléatoire due à l'erreur, F étant alors grand. Il s'ensuit la région critique $F \geq F_\alpha$ tel que cela a été annoncé.

• En fait, le *test de nullité du coefficient de corrélation linéaire* (test du r de PEARSON), le *test de la nullité du coefficient* a, et le *test ANOVA* sont équivalents.

1.6 Propriétés des estimateurs des coefficients de la droite de régression

Passant des réalisations aux variables aléatoires et donc aux estimateurs $\widehat{A_n}$ et $\widehat{B_n}$ de a et b, on a (en supposant par ailleurs X connue « $X=x$ »), $\widehat{A_n}=\frac{\text{cov}(x,Y)}{s_x^2}$ et $\widehat{B_n}=\overline{Y_n}-\frac{\text{cov}(x,Y)}{s_x^2}.\overline{x_n}$.

• $\widehat{A_n}$ et $\widehat{B_n}$ sont les *estimateurs sans biais et de variance minimum* de a et b (**théorème de GAUSS – MARKOV**).

Tout d'abord, supposer les valeurs de X connues ne signifie pas X constante, auquel cas on aurait d'ailleurs, $s_x^2=0$. Ceci dit, l'estimateur $\widehat{A_n}$ est bien sans biais.

En effet, $E(\widehat{A_n})=E\left[\frac{\text{cov}(x,Y)}{s_x^2}\right]=\frac{1}{s_x^2}.E\left[\frac{1}{n}.\sum_{i=1}^{i=n}x_i.Y_i-\overline{x_n}.\overline{Y_n}\right]$, soit par linéarité de l'espérance mathématique et tenant compte que les valeurs de X, soient x_i, sont connues, le résultat $E(\widehat{A_n})=\frac{1}{s_x^2}.\left[\frac{1}{n}.\sum_{i=1}^{i=n}x_i.E(Y_i)-\overline{x_n}.E(\overline{Y_n})\right]$. Or $E(Y_i)=a.x_i+b$ (car $E(x_i)=x_i$).

De même, $E(\overline{Y_n})=\frac{1}{n}.\sum_{i=1}^{i=n}E(Y_i)=a.\overline{x_n}+b$. D'où, finalement :

$$E(\widehat{A_n})=\frac{1}{s_x^2}.\left[\frac{1}{n}.\sum_{i=1}^{i=n}(a.x_i^2+b.x_i)-a.\overline{x_n}^2-b.\overline{x_n}\right]=\frac{1}{s_x^2}.\left[a.\sum_{i=1}^{i=n}(x_i^2-\overline{x_n}^2)+b.\overline{x_n}-b.\overline{x_n}\right]=a.\frac{s_x^2}{s_x^2},$$

soit $E(\widehat{A_n})=a$.

D'autre part, $E(\widehat{B_n})=E(\overline{Y_n})-\overline{x_n}.E(\widehat{A_n})$, soit $E(\widehat{B_n})=a.\overline{x_n}+b-a.\overline{x_n}=b$. Ceci achève la justification du résultat annoncé suivant lequel $\widehat{A_n}$ et $\widehat{B_n}$ sont **sans biais**.

On peut montrer de même, mais c'est plus compliqué, que d'une part $Var(\widehat{A}_n) = \frac{\sigma^2}{n.s_x^2}$ et d'autre part, $Var(\widehat{B}_n) = \frac{\sigma^2}{n}.\left[1 + \frac{\overline{x_n}}{s_x^2}\right]$, σ^2 étant, pour rappel, la variance commune aux ε_i.

Il est manifeste dès lors, que $\widehat{A}_n$ et $\widehat{B}_n$ sont des **estimateurs convergents** puisque $\lim_{n\to+\infty} Var(\widehat{A}_n) = \lim_{n\to+\infty} Var(\widehat{B}_n) = 0$.

Quant au théorème de GAUSS-MARKOV qui souligne la qualité encore plus forte des estimateurs des moindres carrés (à savoir, être de variance minimum), sa démonstration n'est pas incluse ici.

- $\widehat{\sigma_n}^2 = \frac{n}{n-2}.s_Y^2.\left[1 - r_{xY}^2\right] = \frac{1}{n-2}.\sum_{i=1}^{i=n}(Y_i - \widehat{A}_n.x_i - \widehat{B}_n)^2$ est un **estimateur sans biais** de σ^2.

On a $Var\varepsilon_i = \sigma^2 = Var(Y_i - a.x_i - b)$. On peut admettre que les résidus sont naturellement estimés par les *résidus empiriques* $\widehat{\varepsilon}_i = Y_i - \widehat{A}_n.x_i - \widehat{B}_n$ et prendre pour estimation de σ^2, la *variance empirique des résidus empiriques*, soit $s_{\widehat{\varepsilon}}^2 = \frac{1}{n}.\sum_{i=1}^{i=n}\widehat{\varepsilon}_i^2 - \overline{\widehat{\varepsilon}}^2$, avec $\overline{\widehat{\varepsilon}} = \frac{\sum_{i=1}^{i=n}\widehat{\varepsilon}_i}{n}$.

Il en résulte $s_{\widehat{\varepsilon}}^2 = \frac{1}{n}.\sum_{i=1}^{i=n}(Y_i - \widehat{A}_n.x_i - \widehat{B}_n)^2 - \overline{\widehat{\varepsilon}}^2$ avec $\overline{\widehat{\varepsilon}} = \frac{1}{n}.\left[\sum_{i=1}^{i=n}Y_i - \widehat{A}_n.\sum_{i=1}^{i=n}x_i - n.\widehat{B}_n\right]$.

Or, on a montré précédemment que $\overline{Y_n} = \widehat{A}_n.\overline{x_n} + \widehat{B}_n \Rightarrow \overline{\widehat{\varepsilon}} = 0$. Finalement, la variance empirique des résidus se résume à $s_{\widehat{\varepsilon}}^2 = \frac{1}{n}.\sum_{i=1}^{i=n}(Y_i - \widehat{A}_n.x_i - \widehat{B}_n)^2$, soit $s_Y^2.\left[1 - r_{xY}^2\right]$ compte tenu de l'évaluation de l'erreur moyenne entre observations et prévisions explicitée dans le paragraphe 1.3 précédent.

Toutefois, pour *débiaiser* cette variance calculée, le coefficient multiplicateur $\frac{n}{n-2}$ (et non pas $\frac{n}{n-1}$, car il y a ici deux échantillons) est nécessaire, d'où le résultat annoncé.

1.7 Intervalles de confiance et tests dans le cas du modèle linéaire gaussien

Si on suppose en outre la **normalité des résidus** $\varepsilon_i : N(0,\sigma^2)$, on a alors un ensemble de propriétés qu'on admettra et qui permettent la mise en œuvre de tests et d'intervalles de confiance sur $\widehat{A}_n, \widehat{B}_n$, et $\widehat{\sigma_n}^2$ comme cela est illustré dans les applications ci-après.

- $\widehat{A}_n$ suit la **loi normale** $N(a, \frac{\sigma^2}{n.s_x^2})$.
- $\widehat{B}_n$ suit la **loi normale** $N(b, \frac{\sigma^2}{n}.\left[1 + \frac{\overline{x_n}}{s_x^2}\right])$.
- Lorsque σ^2 est inconnu (σ^2 estimé par $\widehat{\sigma_n}^2 = \frac{n}{n-2}.s_Y^2.\left[1 - r_{xY}^2\right]$, $\frac{\widehat{A}_n - a}{\widehat{\sigma_{\widehat{A}_n}}}$ suit la **loi de STUDENT** à $\nu = n-2$ degrés de liberté.

- De même, $\frac{\widehat{B_n}-b}{\widehat{\sigma_{\widehat{B_n}}}}$ suit également la **loi de STUDENT** à $\nu=n-2$ degrés de liberté.
- σ_n^2 est indépendante de $\widehat{A_n}$ et $\widehat{B_n}$.
- $Cov(\widehat{A_n},\widehat{B_n})=-\frac{\sigma^2.\overline{x_n}}{n.s_x^2}$. Ainsi $\widehat{A_n}$ et $\widehat{B_n}$ ne sont-elles **pas indépendantes**.
- $\widehat{A_n},\widehat{B_n}$, et $\widehat{\sigma_n^2}$ sont les **estimateurs sans biais** de a,b, et σ^2, $\widehat{A_n},\widehat{B_n}$ et $\frac{n-2}{n}.\widehat{\sigma_n^2}$ formant par ailleurs les **estimateurs du maximum de vraisemblance** desdits paramètres a,b, et σ^2.
- Lorsque $X=x_0$, **l'intervalle de confiance de la prédiction** $Y_0=a.x_0+b$ a pour forme $\widehat{a_n}.x_0+\widehat{b_n}\pm t_\alpha.\widehat{\sigma_n}.\sqrt{1+\frac{1}{n}+\frac{(x_0-\overline{x_n})^2}{n.s_x^2}}$ où t_α vérifie, relativement à la loi de STUDENT à $\nu=n-2$ degrés de liberté, $Prob(|T|\le t_\alpha)=\alpha$ (si σ est connu, c'est la table de la loi normale $N(0,1)$ qu'on utilisera).
- Enfin, lorsque $X=x_0$, **l'intervalle de confiance de la moyenne des prédictions**, $\overline{Y_n}=\frac{\sum_{i=1}^{i=n}Y_{0,i}}{n}$ a pour forme $\widehat{a_n}.x_0+\widehat{b_n}\pm t_\alpha.\widehat{\sigma_n}.\sqrt{\frac{1}{n}+\frac{(x_0-\overline{x_n})^2}{n.s_x^2}}$ où t_α vérifie les conditions qui sont celles de l'alinéa précédent.

2. Régression linéaire multiple

2.1 Le modèle

Géneralisant de deux à p variables explicatives (prédicteurs) le modèle de régression linéaire simple, le **modèle de régression linéaire multiple** conduit à partir d'un échantillon de données de taille n, soit $(y_i,x_{i1},x_{i2},...,x_{ip}), i\in\{1,2,...,n\}$, à un modèle qui, formulé en termes de variables aléatoires s'écrit :

$Y_i=a_0+a_1.X_{i1}+a_2.X_{i2}+...+a_p.X_{ip}+\varepsilon_i, i\in\{1,2,...,n\}$, les X_i étant connues ou non.

2.2 Estimateurs des moindres carrés

Ici encore à partir du modèle $y_i=a_0+a_1.x_{i1}+a_2.x_{i2}+...+a_p.x_{ip}+\varepsilon_i, i\in\{1,2,...,n\}$, il s'agit de trouver les coefficients $\widehat{a_0},\widehat{a_1},...,\widehat{a_p}$ qui vérifient la relation $U(\widehat{a_0},\widehat{a_1},...,\widehat{a_p})=\underset{a_0,a_1,...,a_p}{Min}\sum_{i=1}^{i=n}(y_i-a_0-a_1.x_{i1}-...-a_p.x_{ip})^2$, ce qui revient à résoudre le système de $(p+1)$ équations à $(p+1)$ inconnues formé par les équations aux dérivées partielles $\frac{\partial U}{\partial a_k}=0, k\in\{1,2,...,n\}$.

• La *meilleure représentation affine* des Y_i en fonction des *variables explicatives* $x_k, (i \in \{1,2,...,n\}, k \in \{1,2,...,p\})$, est fournie par l'équation $y = \widehat{a_0} + \widehat{a_1}.x_1 + ... + \widehat{a_p}.x_p$ où les coefficients $\widehat{a_0}, \widehat{a_1}, ..., \widehat{a_p}$ sont solutions du système d'équations :

$$\begin{cases} \widehat{a_1}.s_{x_1}^2 + \widehat{a_2}.c_{x_1,x_2} + ... + \widehat{a_p}.c_{x_1,x_p} = c_{x_1,Y} \\ \widehat{a_1}.c_{x_2,x_1} + \widehat{a_2}.s_{x_2}^2 + ... + \widehat{a_p}.c_{x_2,x_p} = c_{x_2,Y} \\ \quad ... \\ \widehat{a_1}.c_{x_p,x_1} + \widehat{a_2}.c_{x_p,x_2} + ... + \widehat{a_p}.s_{x_p}^2 = c_{x_p,Y} \end{cases}$$

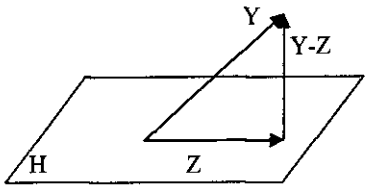
avec en outre, $\widehat{a_0} = \overline{y} - \widehat{a_1}.\overline{x_1} - \widehat{a_2}.\overline{x_2} - ... - \widehat{a_p}.\overline{x_p}$.

Il est précisé que dans le système ci-dessus, $s_{x_i}^2$ et c_{x_i,x_j} désignent les *variances et covariances empiriques* des x_i, x_j (i et j variant entre 1 et p).

La méthode algébrique telle celle employée pour la régression simple conduit aisément au résultat annoncé. Toutefois, pour varier la présentation et alléger les écritures, le recours à la théorie des probabilités et à une **méthode géométrique** est approprié ici.

On va chercher tout d'abord la *meilleure représentation affine* de Y par les p variables aléatoires $X_1, X_2, ..., X_p$, soit $Z = \beta_0 + \beta_1.X_1 + ... + \beta_p.X_p$.

A cet effet, et considérant le sous-espace vectoriel engendré par la variable constante 1 et les variables $X_i (1 \le i \le p)$, soit $H = Vect(1, X_1, ..., X_p)$, il est rappelé que le vecteur Z de H qui approche le mieux possible Y est fourni par la *projection orthogonale* de Y sur H.



En d'autres termes, $Y - Z \perp H$, résultat qu'on peut transcrire en écrivant que $Y - Z$ est orthogonal à chaque générateur de H, soient les $(p+1)$ équations :

$$\begin{array}{l} Y - Z \perp 1 \\ Y - Z \perp X_1 \\ \quad ... \\ Y - Z \perp X_p \end{array}$$

L'espace des variables aléatoires étant muni, classiquement, du produit scalaire $\prec X, Y \succ = E(X.Y)$, il s'ensuit des relations d'orthogonalité susmentionnées, le système :

$$\begin{cases} E\left[(Y - \beta_0 - \beta_1.X_1 - ... - \beta_p.X_p).1\right] = 0 \\ E\left[(Y - \beta_0 - \beta_1.X_1 - ... - \beta_p.X_p).X_1\right] = 0 \\ \quad ... \\ E\left[(Y - \beta_0 - \beta_1.X_1 - ... - \beta_p.X_p).X_p\right] = 0 \end{cases}$$

soit après développement suivant linéarité de l'espérance mathématique, le système d'équations présenté en page suivante :

$$\begin{cases} \beta_0 + \beta_1.E(X_1) + ... + \beta_p.E(X_p) = E(Y) \\ \beta_0.E(X_1) + \beta_1.E(X_1^2) + ... + \beta_p.E(X_1.X_p) = E(X_1.Y) \\ \qquad\qquad ... \\ \beta_0.E(X_p) + \beta_1.E(X_1.X_p) + ... + \beta_p.E(X_p^2) = E(X_p.Y) \end{cases}$$

Moyennant quoi, chercher la *meilleure représentation affine* de Y en fonction des X_i, c'est aussi chercher, en raisonnant sur les **variables centrées,** la *meilleure représentation linéaire* de $Y - E(Y)$ par les $X_i - E(X_i)$. Posant $Z^* = \alpha_0 + \alpha_1.\left[X_1 - E(X_1)\right] + ... + \alpha_p.\left[X_p - E(X_p)\right]$ et appliquant le système précédent au contexte des variables centrées considérées, il vient immédiatement $\alpha_0 = 0$ et :

$$\begin{cases} \alpha_1.Var(X_1) + \alpha_2.\text{cov}(X_1, X_2) + ... + \alpha_p.\text{cov}(X_1, X_p) = \text{cov}(X_1, Y) \\ \alpha_1.\text{cov}(X_1, X_2) + \alpha_2.Var(X_2) + ... + \alpha_p.\text{cov}(X_2, X_p) = \text{cov}(X_2, Y) \\ \qquad\qquad ... \\ \alpha_1.\text{cov}(X_1, X_p) + \alpha_2.\text{cov}(X_2, X_p) + ... + \alpha_p.Var(X_p) = \text{cov}(X_p, Y) \end{cases}$$

Remplaçant ensuite la matrice de variance, covariance par la matrice des variances et covariances empiriques, on aboutit immédiatement au résultat annoncé quant aux coefficients $\widehat{a_0}, \widehat{a_1}, ..., \widehat{a_p}$.

2.3 Etude des coefficients et analyse de la variance

Généralisant les résultats antérieurs de la régression linéaire simple, au modèle $y_i = a_0 + a_1.x_{i1} + a_2.x_{i2} + ... + a_p.x_{ip} + \varepsilon_i$, on montre, dans le cas de **modèles gaussiens** (ε_i de loi normale $N(0,\sigma^2)$) et en notant par $\widehat{A}_{j,n}$ les estimateurs des coefficients a_j, les résultats ci-dessous :

- $\dfrac{\widehat{A}_{j,n} - a_j}{\widehat{\sigma}_{\widehat{A}_{j,n}}}$ suit la **loi de STUDENT** à $n - p - 1$ degrés de liberté.
- $\dfrac{\widehat{A}_{j,n} - a_j}{\sigma_{\widehat{A}_{j,n}}}$ suit la **loi normale** $N(0,1)$.
- $(n-p-1).\dfrac{\widehat{\sigma}^2_{\widehat{A}_{j,n}}}{\sigma^2_{\widehat{A}_{j,n}}}$ suit la **loi du chi-deux** à $n - p - 1$ degrés de liberté.

Il faut néanmoins noter que le calcul des variances estimées $\widehat{\sigma}^2_{\widehat{A}_{j,n}}$ est un peu délicat et exige de recourir à une *approche matricielle* ou au *facteur d'inflation de la variance* (« *variance inflation factor* ») comme cela est développé dans les application 2.1 et 2.2 ci-après.

• On trouvera en page suivante, un tableau récapitulatif relatif à l'analyse de la variance et qui pour les divers facteurs de variations résume lois et degrés de liberté aux fins de mettre en œuvre les tests classiques permettant d'evaluer le caractère approprié ou non du modèle linéaire multiple.

Source de variation	Somme des carrés	Degrés de liberté	Estimateurs
Expliquée	$SCE=\sum_{i=1}^{i=n}(\widehat{y_i}-\overline{y})^2$	p	$SCE/(p-1)$
Résiduelle	$SCR=\sum_{i=1}^{i=n}(y_i-\widehat{y_i})^2$	$n-p-1$	$SCR/(n-p-1)$
Totale	$SCT=\sum_{i=1}^{i=n}(y_i-\overline{y})^2$	$n-1$	

Le *coefficient de détermination* R^2 étant toujours égal au rapport SCE/SCT, tester s'il y a ou non un modèle de régression linéaire multiple entre Y et les x_i, se résume à tester l'hypothèse $H_0 : a_1 = a_2 = ... = a_p = 0$ contre l'hypothèse alternative H_1 suivant laquelle «*l'un au moins des coefficients* a_i *n'est pas nul*».

La statistique associée $F = \dfrac{R^2/p}{(1-R^2)/(n-p-1)}$ suit la **loi de FISHER SNEDECOR** de type $F(p, n-p-1)$, la *région critique* du test ayant pour forme $F \geq F_\alpha$, F_α étant déterminé par la donnée de l'erreur de 1ère espèce α.

• A noter enfin, les limites suivant lesquelles le coefficient de détermination augmente mécaniquement (et non en fonction du caractère représentatif de la régression) lorsque le nombre de variables significatives augmente. Pour compenser cet effet, on uniformise le référentiel en **ajustant** R^2 suivant la formule :

$$R^2_{ajusté} = 1 - \frac{SCR/(n-p-1)}{SCT/(n-1)} = 1 - \frac{n-1}{n-p-1}.(1-R^2).$$

B - Applications

1. Modèles à une variable explicative

1.1 Autonr de la droite de régression

Dans cet exemple, dans lequel les hypothèses du modèle linéaire gaussien sont supposées satisfaites (résidus de loi normale $N(0,\sigma^2)$), les formules d'estimation et de test présentées en rappels de cours (cf. paragraphes 1.5 et 1.7) sont illustrées aux fins de prédictions et inférence statistique.

Enoncé : Une entreprise de maintenance d'ascenseurs mesure le nombre annuel Y de pannes, d'un certain type de portes coulissantes, en fonction du nombre X d'heures de maintenance par an sur ce type d'installation. On observe les résultats ci-dessous :

Heures	Pannes	Heures	Pannes	Heures	Pannes
9	2	12	3	12	2
3	3	15	3	24	1
3	4	6	3	21	2
18	2	6	2	21	2

PARTIE I - MODELISATION

1°) Expliciter la droite de régression de Y en X, soit $\hat{Y} = \hat{a}.X + \hat{b}$.

2°) Calculer le coefficient de détermination R^2 ainsi qu'une estimation de la variance résiduelle. Qu'en conclure quant à la pertinence de la droite de régressions comme prédicteur de la réponse ?

3°) Estimer la covariance entre le coefficient du modèle linéaire considéré $Y = a.X + b$.

PARTIE II - PREDICTIONS

On suppose dans cette partie que la valeur de la variable explicative X est fixée à $x_0 = 10$ heures, plusieurs prédictions étant proposées en ce point.

1°) Construire au seuil de confiance 95%, un intervalle de confiance pour la valeur moyenne de Y.

2°) Toujours au même seuil et au même point, construire un intervalle de confiance pour la prédiction Y.

PARTIE III – INFERENCE STATISTIQUE

1°) L'influence du nombre d'heures de maintenance sur le nombre de pannes observées est-elle significative au niveau $\alpha = 5\%$?

2°) Peut-on considérer que, pour une maintenance nulle, le nombre de pannes est égal à 5 ?

Solution : I-1°) Se reportant aux résultats des rappels de cours du présent chapitre (cf. paragraphe 1.2), la **droite de régression** de Y en X a pour équation $\hat{Y} = \hat{a}.X + \hat{b}$ où $\hat{a} = \dfrac{\sum_{i=1}^{i=n}(x_i - \bar{x}).(y_i - \bar{y})}{\sum_{i=1}^{i=n}(x_i - \bar{x})^2}$ et $\hat{b} = \bar{y} - \hat{a}.\bar{x}$, les données (x_i, y_i) étant par ailleurs, les suivantes :

Heures x_i	9	3	3	18	12	15	6	6	12	24	21	21
Pannes y_i	2	3	4	2	3	3	3	2	2	1	2	2

Il s'ensuit immédiatement, en appliquant les formules susmentionnées, $\bar{x} = 12,5$, $\bar{y} = 2,42$, $\hat{a} = -0,0753$, et $\hat{b} = 3,3579$, d'où la droite de régression cherchée d'équation $\hat{y} = -0,075.x + 3,358$.

I-2°) Le **coefficient de détermination** R^2 est égal au *rapport* entre la *somme des écarts expliqués* (dus au modèle), soit $\sum_{i=1}^{i=n}(\widehat{y_i} - \bar{y})^2$ et la *somme des écarts totaux* (de Y), soit $\sum_{i=1}^{i=n}(y_i - \bar{y})^2$. Calculant ainsi les valeurs $\widehat{y_i}$ à partir de l'équation de régression susmentionnée, soient $\widehat{y_i} = -0,075.x_i + 3,358$, il vient numériquement $\sum_{i=1}^{i=n}(\widehat{y_i} - \bar{y})^2 = 3,35$, $\sum_{i=1}^{i=n}(y_i - \bar{y})^2 = 6,92$, d'où $R^2 = 0,48$.

• Le **coefficient de corrélation linéaire** est, quant à lui, égal à $r_{xy} = \frac{\sum_{i=1}^{i=n}(x_i-\bar{x}).(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{i=n}(x_i-\bar{x})^2}.\sqrt{\sum_{i=1}^{i=n}(y_i-\bar{y})^2}} = -0,696$. On pourra vérifier à cette occasion, qu'on a bien $R^2 = r_{xy}^2$.

• Par ailleurs, la somme des écarts résiduels $\sum_{i=1}^{i=n}(y_i-\hat{y}_i)^2$ est égale à 3,57. On retrouve ainsi *l'équation de l'analyse de la variance* suivant laquelle la somme totale des écarts (soit, 6,92) est égale à la somme des écarts expliqués (soit 3,35) et des écarts résiduels (soit 3,57). Pour l'exemple proposé, on est loin de pouvoir émettre des certitudes quant à la pertinence de la droite de régression comme prédicteur de la réponse, la part inexpliquée des écarts restant importante ici.

I-3°) On a d'après les rappels de cours (cf. paragraphe 1.7), $cov(\hat{a},\hat{b}) = -\frac{\sigma^2.\bar{x}}{n.s_x^2}$. Estimant σ^2 par l'estimateur sans biais $\hat{\sigma}^2 = \frac{1}{n-2}.\sum_{i=1}^{i=n}(y_i-\hat{y}_i)^2$, soit $\hat{\sigma}^2 = 0,357$, il vient numériquement $cov(\hat{a},\hat{b}) = -0,007$.

II-1°) Pour la moyenne de $Y/X = x_0$, les rappels de cours (cf. paragraphe 1.7) enseignent que l'intervalle de confiance a pour forme $\hat{a}.x_0 + \hat{b} \pm t_\alpha.\hat{\sigma}.\sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{n.s_x^2}}$, $\hat{\sigma}^2$ étant estimé par la formule susmentionnée (cf. question I-3°) ou par $\frac{n}{n-2}.s_y^2.\left[1-r_{xy}^2\right]$, avec $s_y^2 = \frac{1}{n}.\sum_{i=1}^{i=n}(y_i-\bar{y})^2$, ce qui est équivalent.

Il en résulte, au seuil de confiance $\alpha = 95\%$ et pour les données antérieures, $t_\alpha = 2,2281$ (cf. lecture dans table annexée de la loi de STUDENT à $\nu = 10$ degrés de liberté, test bilatéral).

Par ailleurs, $\hat{\sigma}^2 = \frac{1}{n-2}.\left[\sum_{i=1}^{i=n}(y_i-\bar{y})^2 - \frac{(\sum_{i=1}^{i=n}(x_i-\bar{x}).(y_i-\bar{y}))^2}{\sum_{i=1}^{i=n}(x_i-\bar{x})^2}\right]$, soit numériquement, $\hat{\sigma}^2 = \frac{1}{10}.\left[6,917 - \frac{1980,25}{591}\right] = 0,357$, résultat dont on constatera, comme cela avait été annoncé ci-dessus, qu'il coïncide avec l'estimation de la variance résiduelle définie par $\hat{\sigma}^2 = \frac{1}{n-2}.\sum_{i=1}^{i=n}(y_i-\hat{y}_i)^2$.

• Finalement, au point $x_0 = 10$, on a $\hat{a}.x_0 + \hat{b} = 2,60$, d'où, pour $E\left[Y/x = x_0\right]$, l'intervalle de confiance $2,60 \pm 2,2281 \times \sqrt{0,357 \times (\frac{1}{12} + \frac{6,25}{591})}$, soit $\left[2,19-3,01\right]$.

II-2°) De même, pour la prédiction $Y/x=x_0$, on a (cf. rappels de cours, paragraphe 1.7), l'intervalle $\hat{a}.x_0+\hat{b}\pm t_\alpha.\hat{\sigma}.\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{n.s_x^2}}$, soit $2,60\pm 2,2281.\sqrt{0,357\times(1+\frac{1}{12}+\frac{6,25}{591})}$, c'est-à-dire $[1,21-4,00]$. Cette fois, l'encadrement est plus *imprécis*.

• On remarque, qui plus est, que la qualité de l'estimation est d'autant moins bonne que $x_0-\bar{x}$ est grand, c'est-à-dire que le point considéré est éloigné du centre de la zone de modélisation.

III-1°) Tester si l'influence du nombre des heures de maintenance est significative sur le nombre de pannes observées, c'est tester (au niveau de signification α), si $\begin{cases} a=0 \\ a\neq 0 \end{cases}$.

La statistique $\frac{\hat{a}-a}{\hat{\sigma}_{\hat{a}}}$ suivant la **loi de STUDENT** à $\nu=n-2$ degrés de liberté, le test a pour **région critique** $|T|=\left|\frac{\hat{a}}{\sqrt{\hat{\sigma}^2/n.s_x^2}}\right|\geq t_\alpha$, puisque, sous l'hypothèse H_0, $a=0$. Ainsi, a-t-on $|\hat{a}|\geq \pi=t_\alpha.\sqrt{\hat{\sigma}^2/n.s_x^2}$. Or, pour $\alpha=5\%$ et pour le test bilatéral, $t_\alpha=2,2281\Rightarrow \pi=0,054$. La relation $|\hat{a}|=0,075\succ \pi=0,054$ permet de conclure au caratère significatif de l'influence de x sur Y.

• Comme cela a été signalé en rappels de cours (cf. paragraphe 1.5), le **test de STUDENT** est *équivalent* au **test F de l'analyse de la variance** (ANOVA). Ainsi, $F=(n-2).\frac{R^2}{1-R^2}$ suit la **loi de FISHER SNEDECOR** à $\nu_1=1$ et $\nu_2=n-2$ degrés de liberté.

La région critique ayant pour forme $F\geq F_\alpha / \Pr ob(F\geq F_\alpha)=\alpha$ (cf. rappels de cours), il résulte d'une lecture dans la table de valeurs annexée (table de FISHER SNEDECOR) et pour $\alpha=0,05$, le seuil $F_\alpha=6,94$. Or $F_{calculé}=9,39\succ F_\alpha=6,94$. C'est donc l'hypothèse $H_1: F\succ 1$ qu'on retient ici ce qui est conforme au résultat antérieur fourni par le test de STUDENT.

III-2°) Dans cette question, c'est un test sur le coefficient b qui est proposé, à savoir $\begin{cases} H_0: b=5 \\ H_1: b\neq 5 \end{cases}$. Or, la statistique $\frac{\hat{b}-b}{\hat{\sigma}_{\hat{b}}}$ suit la **loi de STUDENT** à $\nu=n-2$ degrés de liberté, $\hat{\sigma}_{\hat{b}}^2$ étant égal à $\frac{\hat{\sigma}^2}{n}.\left[1+\frac{\bar{x}}{s_x^2}\right]$. On a donc cette fois, une **région critique** caractérisée par $|\hat{b}-5|\geq \pi=t_\alpha.\sqrt{\frac{\hat{\sigma}^2}{n}.\left[1+\frac{\bar{x}}{s_x^2}\right]}$ avec t_α vérifiant $\Pr ob(|T|\geq t_\alpha)=\alpha\Rightarrow t_\alpha=2,2281$ lorsque $\alpha=5\%$. Il s'ensuit numériquement $\pi=0,43$.

Or $\left|\hat{b}_{calculé} - 5\right| = |3,36 - 5| = 1,64 \succ 0,43$. C'est donc le rejet de l'hypothèse H_0 suivant laquelle le nombre de pannes est égal à 5 lorsqu'on effectue une maintenance nulle, qui est à retenir ici.

1.2 Parabole des moindres carrés et distance de freinage

Comme cela a été indiqué en rappels de cours (cf. paragraphe 1.1), un ensemble important de régressions non linéaires dont entre autres les régressions polynomiales et de FOURIER ressortent néanmoins du modèle linéaire, l'élément déterminant étant ici ladite linéarité par rapport aux coefficients du modèle et non par rapport au prédicteur. Les équations de la régression sont établies ci-après dans le cas le plus général.

Enoncé : PARTIE I – EQUATIONS DE LA REGRESSION

Etant donnés un prédicteur x et p fonctions données de x, soient $\varphi_1(x), \varphi_2(x), ..., \varphi_p(x)$, on recherche pour la variable expliquée Y, la meilleure représentation affine $\hat{y} = \widehat{a_0} + \widehat{a_1}.\varphi_1(x) + ... + \widehat{a_p}.\varphi_p(x)$, les données étant fournies quant à elles, par un échantillon de taille n, soit $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$.

Ecrire le système de $(p+1)$ équations à $(p+1)$ inconnues qui permet la détermination des coefficients $\widehat{a_0}, \widehat{a_1}, ..., \widehat{a_p}$.

PARTIE II- EXEMPLE DE L'AJUSTEMENT PARABOLIQUE

On considère ici, relativement au modèle développé en 1ère partie, le cas particulier $p = 2, \varphi_1(x) = x, \varphi_2(x) = x^2$.

Ecrire dans ces conditions, les équations de la régression.

PARTIE III – DISTANCE DE FREINAGE

Le tableau ci-dessous indique les distances de freinage d_i (en m) d'un véhicule roulant à diverses vitesses v_i (en km/h).

Vitesse v_i	20	30	40	50	60	70
Distance de freinage d_i	54	90	138	206	292	396

Il est admis par ailleurs, que la distance d est modélisée en fonction de v suivant l'équation $d_i = a_0 + a_1.v_i + a_2.v_i^2 + \varepsilon_i$.

1°) Calculer les coefficients $\widehat{a_0}, \widehat{a_1}$, et $\widehat{a_2}$ de la meilleure représentation parabolique $\hat{d} = \widehat{a_0} + \widehat{a_1}.v + \widehat{a_2}.v^2$ de d par les v_i.

2°) Calculer le coefficient de détermination. Qu'en conclure ?

Solution : I°) La *meilleure représentation affine* des y_i par les fonctions de la forme $a_0 + a_1.\varphi_1(x_i) + ... + a_p.\varphi_p(x_i)$ est caractérisée par les coefficients $\widehat{a_0}, \widehat{a_1}, ..., \widehat{a_p}$ qui minimisent la distance $\left\|y - \hat{y}\right\|$ dont, au sens des moindres carrés, l'expression est fournie par $U(a_0, a_1, ..., a_p) = \sum_{i=1}^{i=n} \left[y_i - a_0 - a_1.\varphi_1(x_i) - ... - a_p.\varphi_p(x_i) \right]^2$.

La *nullité des dérivées partielles* $\frac{\partial U}{\partial a_k} = 0, k \in \{1,2,...,p\}$, conduit aux coefficients $\widehat{a_k}$ cherchés, ce qu'on peut obtenir de façon encore plus rapide par la voie **géométriqne** à l'instar des développements des rappels de cours (cf. paragraphe 2.1).

En effet, considérant le *sous-espace vectoriel* $H = Vect(1, \varphi_1(x), \varphi_2(x), ..., \varphi_p(x))$ et le *produit scalaire* $\prec x, y \succ = \sum_{i=1}^{i=n} x_i.y_i$, la *meilleure représentation* de Y par les vecteurs de H est fournie par la **projection orthogonale** $\hat{Y}$ de Y sur H, c'est-à-dire $\hat{Y}$ vérifiant $Y - \hat{Y} \perp H$. Au regard, de la *condition nécessaire et suffisante d'orthogonalité* de $Y - \hat{Y}$ avec H, soit $Y - \hat{Y} \perp 1, Y - \hat{Y} \perp \varphi_1(x), ..., Y - \hat{Y} \perp \varphi_p(x)$, et compte tenu du produit scalaire susmentionné, il en résulte les équations :

$$\begin{cases} \sum_{i=1}^{i=n} \left[y_i - \widehat{a_0} - \widehat{a_1}.\varphi_1(x_i) - ... - \widehat{a_p}.\varphi_p(x_i) \right].1 = 0 \\ \sum_{i=1}^{i=n} \left[y_i - \widehat{a_0} - \widehat{a_1}.\varphi_1(x_i) - ... - \widehat{a_p}.\varphi_p(x_i) \right].\varphi_1(x_i) = 0 \\ ..... \\ \sum_{i=1}^{i=n} \left[y_i - \widehat{a_0} - \widehat{a_1}.\varphi_1(x_i) - ... - \widehat{a_p}.\varphi_p(x_i) \right].\varphi_p(x_i) = 0 \end{cases}$$

d'où le système d'équations :

$$\begin{cases} n.\widehat{a_0} + \widehat{a_1}.\sum_{i=1}^{i=n} \varphi_1(x_i) + ... \widehat{a_p}.\sum_{i=1}^{i=n} \varphi_p(x_i) = \sum_{i=1}^{i=n} y_i \\ \widehat{a_0}.\sum_{i=1}^{i=n} \varphi_1(x_i) + \widehat{a_1}.\sum_{i=1}^{i=n} \varphi_1^2(x_i) + ... + \widehat{a_p}.\sum_{i=1}^{i=n} \varphi_1(x_i).\varphi_p(x_i) = \sum_{i=1}^{i=n} \varphi_1(x_i).y_i \\ ... \\ \widehat{a_0}.\sum_{i=1}^{i=n} \varphi_p(x_i) + \widehat{a_1}.\sum_{i=1}^{i=n} \varphi_1(x_i).\varphi_p(x_i) + ... + \widehat{a_p}.\sum_{i=1}^{i=n} \varphi_p^2(x_i) = \sum_{i=1}^{i=n} \varphi_p(x_i).y_i \end{cases}$$

II- 1°) Les équations de la régression dans le cas où $p = 2$ et où $\varphi_1(x) = x, \varphi_2(x) = x^2$, s'écrivent, avec les notations $d_i = a_0 + a_1.v_i + a_2.v_i^2 + \varepsilon_i$:

$$\begin{cases} n.\widehat{a_0} + \widehat{a_1}.\sum_{i=1}^{i=n} v_i + \widehat{a_2}.\sum_{i=1}^{i=n} v_i^2 = \sum_{i=1}^{i=n} d_i \\ \widehat{a_0}.\sum_{i=1}^{i=n} v_i + \widehat{a_1}.\sum_{i=1}^{i=n} v_i^2 + \widehat{a_2}.\sum_{i=1}^{i=n} v_i^3 = \sum_{i=1}^{i=n} d_i.v_i \\ ... \\ \widehat{a_0}.\sum_{i=1}^{i=n} v_i^2 + \widehat{a_1}.\sum_{i=1}^{i=n} v_i^3 + \widehat{a_2}.\sum_{i=1}^{i=n} v_i^4 = \sum_{i=1}^{i=n} d_i.v_i^2 \end{cases}$$

Il en résulte numériquement, et pour les données proposées, le système d'équations explicité ci-après.

$$\begin{cases} 6.\widehat{a_0} + 270.\widehat{a_1} + 13900.\widehat{a_2} = 1176 \\ 270.\widehat{a_0} + 13900.\widehat{a_1} + 783000.\widehat{a_2} = 64840 \\ 13900.\widehat{a_0} + 783000.\widehat{a_1} + 46750000.\widehat{a_2} = 3830000 \end{cases}$$

Par résolution, on obtient $\widehat{a_0} = 41,771429, \widehat{a_1} = -1,095714, \widehat{a_2} = 0,087857$, d'où, après arrondis, la représentation parabolique $\hat{d} = 41,771 - 1,096.v + 0,088.v^2$ (il convient cependant de se méfier des arrondis dont la présence de puissances de la variable v augmente considérablement l'impact).

II- 2°) Pour les diverses valeurs de v considérées, les valeurs d_i sont comparées ci-dessous aux prédictions $\widehat{d_i}$ du modèle.

v_i	20	30	40	50	60	70
d_i	54	90	138	206	292	396
$\widehat{d_i}$	55,00	87,97	138,51	206,63	292,31	395,57

L'écart résiduel $\sum_{i=1}^{i=6}(d_i - \widehat{d_i})^2$ est donc égal à 6,057. Quant à **l'écart total**, c'est $\sum_{i=1}^{i=6}(d_i - \overline{d})^2 = 84080$. Enfin, **l'écart expliqué** (du au modèle) est égal à $\sum_{i=1}^{i=6}(\widehat{d_i} - \overline{d})^2 = 84073,936$.

On retrouve **l'éqnation de la variance** suivant laquelle l'écart total est égal à la somme de l'écart résiduel et de l'écart expliqué.

• Dans ce cadre, le **coefficient de détermination** R^2 est égal au rapport entre l'écart expliqué et l'écart total, soit $R^2 = \dfrac{84073,9}{84080} \approx 0,9999$, valeur qui avoisine 1 pour le cas présent et qui traduit toute la pertinence du modèle obtenu.

1.3 Equations non linéaires se ramenant au modèle linéaire (exemple de l'équation des gazs parfaits)

Enoncé : Le tableau ci-dessous, représente des valeurs expérimentales P_i de la pression d'un gaz donné (en pascals), en fonction de divers volumes V_i (en m^3).

Volume V_i	54,3	61,8	72,4	88,7	118,6	194,0
Pression P_i	61,2	49,5	37,6	28,4	19,2	10,1

D'après les lois de la thermodynamique, on admet que P et V sont reliés par un modèle de type $PV^\gamma = C$, où γ et C sont des constantes spécifiques au gaz considéré.

1°) Déterminer les meilleures estimations $\hat{\gamma}$ et $\widehat{C}$ de γ et de C et en déduire la meilleure approximation de P en fonction de V.

2°) Calculer P pour $V = 100m^3$.

Solntion : 1°) Il est immédiat que $P.V^{\gamma} = C$ s'écrit aussi $\ln P + \gamma.\ln V = \ln C$, soit la forme d'un **modèle linéaire** entre $\ln P$ et $\ln V$ (plus précisément $\ln P = -\gamma.\ln V + \ln C$).

Posant $x = \ln V$ et $Y = \ln P$, la meilleure approximation affine des y_i par les x_i est fournie par la **droite de régression** d'équation $\hat{y} = \hat{a}.x + \hat{b}$ où $\hat{a} = \dfrac{\sum_{i=1}^{i=n}(x_i - \bar{x}).(y_i - \bar{y})}{\sum_{i=1}^{i=n}(x_i - \bar{x})^2}$ et $\hat{b} = \bar{y} - \hat{a}.\bar{x}$. La mise en œuvre de ces formules, conduit au tableau de calculs ci-dessous :

V_i	54,3	61,8	72,4	88,7	118,6	194,0
P_i	61,2	49,5	37,6	28,4	19,2	10,1
$x_i = \ln V_i$	3,994	4,124	4,282	4,485	4,776	5,268
$y_i = \ln P_i$	4,114	3,902	3,627	3,346	2,955	2,313

Il s'ensuit $\bar{x} = 4,488$; $\bar{y} = 3,376$; $\sum_{i=1}^{i=6}(x_i - \bar{x})^2 = 1,109$; $\sum_{i=1}^{i=6}(x_i - \bar{x}).(y_i - \bar{y}) = -1,558$, $\hat{a} = -1,404$; $\hat{b} = 9,679$, ce qui entraîne en définitive, $\hat{y} = -1,404.x + 9,679$.

- Par rapprochement avec l'équation $y = -\gamma.x + \ln C$, il vient $\hat{\gamma} = 1,404$ et $\hat{C} = \exp(9,679) = 15971$. D'où, entre P et V, le modèle $P.V^{1,404} = 15971$.

2°) Lorsque $V = 100$, $\ln P = 3,213 \Rightarrow P = 24,83N/m^2$.

1.4 Modèle de régression (taille, poids)

Les équations de la régression demeurent applicables lorsque la variable explicative est elle-même aléatoire. Ci-dessous, un exemple inspiré de la statistique descriptive.

Enoncé : On a relevé ci-dessous l'âge x et le poids y de $n = 492$ enfants.

Poids (en kg)	Age (en années)					Total
	6	8	10	12	14	
11 à 15	1					1
15 à 19	9	1				10
19 à 23	16	13		1		30
23 à 27	8	32	14	1		55
27 à 31	5	17	26	10		58
31 à 35		6	33	28	10	77
35 à 39		1	21	26	13	61
39 à 43			3	31	31	65
43 à 47			1	17	41	59
47 à 51				10	30	40
51 à 55				1	18	19
55 à 59				2	6	8
59 à 63					8	8
63 à 67					1	1
Total	39	70	98	127	158	492

1°) Calculer le coefficient de corrélation linéaire entre l'âge et le poids.

2°) Expliciter la droite de régression de y en x.

Solution : 1°) On part cette fois d'un *tableau de contingences* dans lequel les classes qui caractérisent les valeurs de y peuvent être ramenées à leurs centres, les fréquences absolues observées relevées pour chaque couple (x_i, y_j) étant notées n_{ij}.

Dans ces conditions, les formules qui conduisent aux *variances, covariance, coefficient de corrélation linéaire,..*, **empiriques**, s'écrivent respectivement :

$$\bar{x} = \frac{\sum_i \sum_j x_i . n_{ij}}{N} \qquad s_x^2 = \frac{\sum_i \sum_j (x_i - \bar{x})^2 . n_{ij}}{N}$$

$$\bar{y} = \frac{\sum_i \sum_j y_j . n_{ij}}{N} \qquad s_y^2 = \frac{\sum_i \sum_j (y_j - \bar{y})^2 . n_{ij}}{N}$$

$$\text{(avec } N = \sum_i \sum_j n_{ij} \text{)}$$

$$c_{xy} = \frac{\sum_i \sum_j (x_i - \bar{x}).(y_j - \bar{y}).n_{ij}}{N}$$

Numériquement, et à partir d'un calcul effectué sur tableur, on obtient $\bar{x} = 11,20$; $\bar{y} = 36,51$; $\sum_i \sum_j (x_i - \bar{x})^2 . n_{ij} = 3232,48$; et $\sum_i \sum_j (y_j - \bar{y})^2 . n_{ij} = 49034,33$. Enfin, on a $\sum_i \sum_j (x_i - \bar{x}).(y_j - \bar{y}).n_{ij} = 10247,80$.

Il en résulte immédiatement $r_{xy} = \frac{10247,80}{\sqrt{3232,48 \times 49034,33}} = 0,814$. Ce coefficient est suffisamment proche de 1 pour justifier un *ajustement suivant un modèle linéaire*.

2°) Sous cette hypothèse, la droite de régression de y en x a pour équation $\hat{y} = \hat{a}.x + \hat{b}$ où, par transcription des formules habituelles aux données suivant tableau de contingences, tel le présent problème, on a $\hat{a} = \frac{\sum_i \sum_j (x_i - \bar{x}).(y_j - \bar{y}).n_{ij}}{\sum_i \sum_j (x_i - \bar{x})^2 . n_{ij}}$ et $\hat{b} = \bar{y} - \hat{a}.\bar{x}$.

Il vient numériquement, $\hat{a} = 3,17$ et $\hat{b} = 1,007$, d'où le modèle $\hat{y} = 3,17.x + 1,007$.

2. Modèles à plusieurs variables explicatives

2.1 Illnstration autour d'un modèle à deux variables explicatives

Enoncé : Le taux de mortalité infantile X, le taux d'analphabétisme Y, et le P.I.B par habitant Z, ont été étudiés dans neuf pays.

On suppose par ailleurs que les conditions de validité du modèle gaussien sont réunies ici.

Les données qui datent de 1984 sont rassemblées dans le tableau qui est présenté ci-après.

Pays	Numéro d'ordre	Taux de mortalité infantile en 1/1000	Taux d'analphabétisme en 1/100	P.I.B par habitant
Afrique du Sud	1	89,0	50,0	2680,0
Algérie	2	114,0	58,5	2266,0
Arabie Saoudite	3	111,0	75,4	10827,0
Argentine	4	44,0	5,3	2264,0
Australie	5	10,4	0,0	9938,0
Bahrein	6	57,0	20,9	8960,0
Brésil	7	75,0	23,9	1853,0
Cameroun	8	106,0	55,1	939,0
Canada	9	10,0	0,9	9857,0

1°) Considérant le modèle $Z = a + b.X + c.Y$, expliciter les estimateurs $\hat{a}, \hat{b}$, et $\hat{c}$ des paramètres a, b, et c, la prédiction de Z en résultant ayant pour forme $\hat{Z} = \hat{a} + \hat{b}.X + \hat{c}.Y$.

2°) Calculer le coefficient de détermination. Qu'en conclure quant à la validité du modèle ?

3°) Pour le Mexique, le taux de mortalité infantile est 54 (pour mille), le taux d'analphabétisme étant quant à lui de 17,3%. Evaluer l'estimation ponctuelle du P.I.B par habitant qui résulte de la régression linéaire multiple.

4°) On considère le test $H_0 : a = b = c$ contre H_1 : Il existe au moins un coefficient a, b, c non nul. Qu'en conclure au niveau de signification $\alpha \approx 5\%$?

5°) Considérant cette fois, le test $\begin{cases} H_0 : b = 0 \\ H_1 : b \neq 0 \end{cases}$ (toujours au niveau de signification $\alpha = 5\%$), indiquer si le taux de mortalité infantile contribue (ou non) de manière significative au sein de la régression.

Solution : 1°) Se reportant aux rappels de cours, la *meilleure représentation affine* de $Z - E(Z)$ par $X - E(X)$ et $Y - E(Y)$ est caractérisée par l'équation $\hat{Z} - E(Z) = \widehat{\alpha_1}.[X - E(X)] + \widehat{\alpha_2}.[Y - E(Y)]$ où $\widehat{\alpha_1}$ et $\widehat{\alpha_2}$ sont solutions du système :

$$\begin{cases} \alpha_1.Var(X) + \alpha_2.\text{cov}(X,Y) = \text{cov}(X,Z) \\ \alpha_1.\text{cov}(X,Y) + \alpha_2.Var(Y) = \text{cov}(Y,Z) \end{cases}$$

Remplaçant les moments ci-dessus par les *moyennes*, *variances*, et *covariances empiriques*, on en déduit comme suit, les estimateurs $\hat{a}, \hat{b}$, et $\hat{c}$ des coefficients de la régression linéaire multiple $Z = a + b.X + c.Y + \varepsilon$. On a ainsi :

$$\bar{x} = 68,49 \qquad \sum_{i=1}^{i=9}(x_i - \bar{x})^2 = 13275,61 \qquad \sum_{i=1}^{i=9}(x_i - \bar{x}).(y_i - \bar{y}) = 8693,22$$

$$\bar{y} = 32,22 \qquad \sum_{i=1}^{i=9}(y_i - \bar{y})^2 = 6335,90 \qquad \sum_{i=1}^{i=9}(z_i - \bar{z}).(y_i - \bar{y}) = -210629,77$$

$$\bar{z} = 5509,33 \qquad \sum_{i=1}^{i=9}(z_i - \bar{z})^2 = 142013220 \qquad \sum_{i=1}^{i=9}(x_i - \bar{x}).(z_i - \bar{z}) = -646541,87$$

Il en résulte, le système d'équations :

$$\begin{cases} 13275,61.\alpha_1 + 8693,22.\alpha_2 = -646541,87 \\ 8693,22.\alpha_1 + 6335,90.\alpha_2 = -210629,77 \end{cases}$$

Par résolution, $\widehat{\alpha_1} = -265,24$ et $\widehat{\alpha_2} = 330,68$, d'où, en conclusion, la relation $\widehat{Z} - 5509,33 = -265,24.[X - 68,49] + 330,68.[Y - 32,22]$, qui s'écrit encore :

$$Z = -265,24.X + 330,68.Y + 13020,10 .$$

2°) Le **coefficient de détermination** R^2 est défini par le *rapport* entre la *variance de Z expliquée par le modèle*, soit $\sum_{i=1}^{i=9}(\widehat{z_i} - \overline{z})^2$, et la *variance totale de Z*, soit $\sum_{i=1}^{i=9}(z_i - \overline{z})^2$.

Le tableau ci-dessous indique les valeurs $\widehat{z_i}$ pour les différentes données (x_i, y_i) :

x_i	y_i	$\widehat{z_i}$
89,0	50,0	5947,75
114,0	58,5	2127,51
111,0	75,4	8511,80
44,0	5,3	3102,07
10,4	0,0	10261,58
57,0	20,9	4812,60
75,0	23,9	1030,28
106,0	55,1	3125,12
10,0	0,9	10665,30

On obtient immédiatement $\sum_{i=1}^{i=9}(\widehat{z_i} - \overline{z})^2 = 101838313$ et $\sum_{i=1}^{i=9}(z_i - \overline{z})^2 = 142013220$, d'où $R^2 = 0,7171$.

• Le modèle est à priori *intéressant* puisque la part de la variablité de Z traduite par ce dernier dépasse ainsi 70% (l'idéal étant, pour rappel, la valeur 100%).

3°) Considérant les données $x = 54; y = 17,3$, la **prédiction** correspondante de Z fournie par le modèle obtenu au 1°) est $\widehat{Z} = -265,24 \times 54 + 330,68 \times 17,3 + 13020,10$, soit la valeur 4417,86.

4°) Le **test de signification globale** $H_0 : b = c = 0$ contre $H_1 : \exists$ au moins un coefficient b, c non nul, permet de vérifier s'il est possible d'effectuer ou non une prédiction meilleure que la simple constante, c'est-à-dire en d'autres termes, s'il existe au moins une variable qui apporte de l'information sur Z.

Se référant aux rappels de cours (cf. paragraphe 2.2), la statistique $F = \dfrac{R^2 / p}{(1-R^2) / n-p-1}$ suit **la loi de FISHER SNEDECOR** à $\nu_1 = p, \nu_2 = n - p - 1$ degrés de liberté, soit la loi $F(p, n-p-1)$. Sous l'hypothèse H_1, la statistique F qui est le rapport entre la *variance expliquée* et la *variance totale* prend des *grandes valeurs*, la **région critique** étant donc caractérisée par $F \geq F_\alpha$ où, sous l'hypothèse H_0, F_α vérifie $\Pr ob(F \geq F_\alpha) = \alpha$.

Par lecture dans la table de valeurs annexée de la loi $F(\nu_1 = 2, \nu_2 = 9-2-1 = 6)$ et pour le niveau de signification $\alpha = 5\%$, on a $F_\alpha = 5,14$. Or $F_{calculé} = \dfrac{R^2/2}{(1-R^2)/6} = 7,60$.

C'est donc l'hypothèse H_1 qu'il faut retenir ici puisque $F_{calculé} = 7,60 \succ F_\alpha = 5,14$, la pertinence du modèle linéaire considéré étant ainsi confirmée.

5°) Cette fois, on cherche à déterminer si la *variable explicative* X contribue ou non de manière significative à la régression, ce qui se résume au **test bilatéral** $\begin{cases} H_0 : b = 0 \\ H_1 : b \neq 0 \end{cases}$.

Or, suivant les rappels de cours (cf. paragraphe 2.3), la variable $\dfrac{\widehat{B} - b}{\sigma_{\widehat{B}}}$ suit la **loi de STUDENT** à $n-p-1$ degrés de liberté. *La difficulté réside dans l'évaluation de* $\widehat{\sigma}_{\widehat{B}}$ et il est vrai à cet égard que la formulation matricielle telle celle exposée dans l'application 2.2 ci-après est nettement plus commode.

• Dans le cadre général du **modèle linéaire multiple** caractérisé par l'équation $Y = \beta_0 + \beta_1.X_1 + ... + \beta_j.X_j + ... + \beta_p.X_p + \varepsilon$, on montre que $\widehat{\sigma}^2_{\widehat{\beta}_j} = VIF(\widehat{\beta}_j).\dfrac{\widehat{\sigma}^2}{\sum_{i=1}^{i=n}(x_{ji} - \overline{x}_j)^2}$

où $\widehat{\sigma}^2$ désigne *l'écart total résiduel* estimé par $\dfrac{1}{n-p-1}.\sum_{i=1}^{i=n}(y_i - \widehat{y}_i)^2$ et *VIF*, le **facteur d'inflation de la variance** défini par $\dfrac{1}{1-R_j^2}$, R_j^2 étant, quant à lui, le *coefficient de détermination de la régression de* X_j *en fonction des autres variables explicatives*, soient $X_1, X_2, ..., X_{j-1}, X_{j+1}, ..., X_p$.

• En résumé, et pour l'exemple proposé, la procédure ci-dessus conduit, pas à pas, à :

→ Calculer $\widehat{\sigma}^2 = \dfrac{1}{n-3}.\sum_{i=1}^{i=9}(z_i - \widehat{z}_i)^2 = 6695818$.

→ Expliciter la meilleure régression de la variable choisie, soit X, en fonction des autres variables explicatives, soit Y. Il vient immédiatement, en utilisant les formules habituelles, $\widehat{X} = \dfrac{\sum_{i=1}^{i=n}(x_i - \overline{x}).(y_i - \overline{y})}{\sum_{i=1}^{i=n}(y_i - \overline{y})^2}.\left[y - \overline{y}\right] + \overline{x}$, soit numériquement, l'équation

$\widehat{X} = \dfrac{8693,22}{6335,90}.\left[y - 32,22\right] + 68,49$.

→ En déduire le coefficient de détermination R_X^2 égal en la circonstance au carré du coefficient de corrélation linéaire $\dfrac{\text{cov}(X,Y)}{\sqrt{Var(X).Var(Y)}}$, soit $\left[\dfrac{8693,22}{\sqrt{13275,61 \times 6335,90}}\right]^2 = 0,898$.

→ En déduire :

$$VIF(\hat{B}) = \frac{1}{1-R_X^2} = 9,848 \text{ puis } \hat{\sigma}_{\hat{B}}^2 = VIF(\hat{B}).\frac{\hat{\sigma}^2}{\sum_{i=1}^{i=9}(x_i - \bar{x})^2} = 9,848 \times \frac{6695818}{13275,61} = 4967,24.$$

- La suite est habituelle. La **région critique** a pour forme $\left|\hat{B}\right| \geq \pi$. Or, sous l'hypothèse $H_0 : b = 0$, la statistique $\frac{\hat{B}}{\hat{\sigma}_{\hat{B}}}$ suit la **loi de STUDENT** à $n-p-1$ degrés de liberté. Désignant par t_α le nombre qui, au niveau de signification α donné, satisfait la relation $Prob(|T| \geq t_\alpha) = \alpha$, on a immédiatement $\pi = t_\alpha.\hat{\sigma}_{\hat{B}}$.

Numériquement, $\alpha = 5\%$, et $n = 9, p = 2$, entraînent suivant lecture dans la *table de* STUDENT *annexée*, $t_\alpha = 2,447$. Ainsi $\pi = 172,46$. Or $\left|\hat{B}_{calculé}\right| = 265,24$ suivant la régression explicitée à la 1ère question.

La relation $\left|\hat{B}_{calculé}\right| = 265,24 \succ \pi = 172,46$ conduit à retenir l'hypothèse H_1 suivant laquelle le taux de mortalité infantile contribue bien de manière significative à la prédiction du P.I.B. En d'autres termes, on ne saurait résumer ici, le modèle à une équation de la forme $Z = m + n.Y$ (suppression du facteur X).

2.2 Matrices et régression linéaire multiple

L'écriture matricielle facilite la lecture et la manipulation des modèles de régression multiple, ses principaux résultats en étant développés ci-dessous.

Enoncé : On considère le modèle de régression multiple à p variables (tel celui exposé en rappels de cours) et caractérisé par les équations :

$$Y_i = a_0 + a_1.x_{1i} + a_2.x_{2i} + ... + a_p.x_{pi} + \varepsilon_i \text{ avec } i = 1,2,...,n.$$

$\hat{a_0}, \hat{a_1}, ..., \hat{a_p}$ étant les estimateurs des moindres carrés des a_i, on forme les matrices suivantes :

$$\hat{A}_{(p+1,1)} = \begin{pmatrix} \hat{a_0} \\ \hat{a_1} \\ . \\ \hat{a_p} \end{pmatrix}, \; X_{(n,p+1)} = \begin{pmatrix} 1 & x_{11} & ... & x_{p1} \\ 1 & x_{12} & ... & x_{p2} \\ . & . & . & . \\ 1 & x_{1n} & ... & x_{pn} \end{pmatrix}, \; Y_{(n,1)} = \begin{pmatrix} y_1 \\ y_2 \\ . \\ y_n \end{pmatrix}, \; \varepsilon_{(n,1)} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ \varepsilon_n \end{pmatrix}, \; A_{(p+1,1)} = \begin{pmatrix} a_0 \\ a_1 \\ . \\ a_p \end{pmatrix}$$

On note par ailleurs par M^T la transposée de toute matrice M.

1°) Vérifier que $\hat{A} = (X^T.X)^{-1}.X^T.Y$ et en déduire que $\hat{A} = A + (X^T.X)^{-1}.X^T.\varepsilon$.

2°) Montrer que $\hat{A}$ est sans biais.

3°) Suivant les hypothèses de nullité des espérances $E(\varepsilon_i)$, d'homoscédasticité (c'est-à-dire, $E(\varepsilon_1^2) = E(\varepsilon_2^2) = ... E(\varepsilon_n^2) = \sigma^2$), et de non corrélation des erreurs ($E(\varepsilon_i.\varepsilon_j) = 0$ si $i \neq j$), montrer que $E\left[\varepsilon.\varepsilon^T\right] = \sigma^2.I_n$ où I_n désigne la matrice identique d'ordre n.

4°) Notant par $\Omega_{(p+1,p+1)}$ la matrice des variances et covariances de $\hat{A}$, montrer que $\Omega = \sigma^2.(X^T.X)^{-1}$.

5°) On souhaite étudier la variation du taux d'hémoglobine dans le sang au cours d'une opération chirurgicale en fonction de la durée de l'opération en question et du volume de sang perdu à cette occasion. On dispose des résultats suivants dans lesquels y_i représente la valeur observée en pourcentage de la variation du taux d'hémoglobine, x_{i1} est la durée de l'opération, et x_{i2} est le volume de sang perdu en litres.

y_i	-1,70	-4,61	-5,82	-1,17	-4,23	-3,31	+0,42	-2,98
x_{i1}	1,75	1,33	1,43	1,86	1,81	1,66	1,60	2,00
x_{i2}	0,52	0,59	0,61	0,50	0,54	0,49	0,27	0,47

On suppose que y_i est une réalisation d'une variable aléatoire Y_i de loi normale $N(a+b.x_{i1}+c.x_{i2},\sigma^2)$.

a) Estimer les paramètres inconnus a,b,c, et σ^2.

b) Tester l'hypothèse suivant laquelle la variation du taux d'hémoglobine ne dépend ni de la durée de l'opération, ni du volume de sang perdu.

c) Tester l'hypothèse suivant laquelle la variation du taux d'héméglobine ne dépend pas de la durée de l'opération.

Solution : 1°) Reprenant la théorie générale (cf. rappels de cours, paragraphe 2.1) appliquée à la *régression affine* de Y en fonction des $X_k, k \in \{1,2,...,p\}$, et notamment *l'approche géométrique*, il est immédiat que la *meilleure représentation affine* cherchée $\hat{Y} = \widehat{a_0} + \widehat{a_1}.X_1 + ... + \widehat{a_p}.X_p$ de Y par les X_k, est caractérisée par la **projection orthogonale** de Y sur le *sous-espace vectoriel* $H = Vect(1, X_1, X_2, ..., X_p)$, l'espace considéré étant, cette fois, muni du produit scalaire $\prec X,Y \succ = \sum_{i=1}^{i=n} x_i.y_i$.

Ainsi a-t-on $Y - \hat{Y} \perp H$, ce qui est vérifié, **si et seulement si**, on a simultanément $Y - \hat{Y} \perp 1, Y - \hat{Y} \perp X_1, ..., Y - \hat{Y} \perp X_p$, ce qui conduit au système d'équations :

$$\begin{cases} \sum_{i=1}^{i=n} (y_i - a_0 - a_1.x_{1i} - ... - a_p.x_{pi}).1 = 0 \\ \sum_{i=1}^{i=n} (y_i - a_0 - a_1.x_{1i} - ... - a_p.x_{pi}).x_{1i} = 0 \\ \qquad ... \\ \sum_{i=1}^{i=n} (y_i - a_0 - a_1.x_{1i} - ... - a_p.x_{pi}).x_{pi} = 0 \end{cases}$$

Soit, le système à $p+1$ équations et à $p+1$ inconnues, explicité en page suivante et dont il est proposé ensuite, d'effectuer la transcription matricielle qui conduit au résultat proposé dans l'énoncé.

$$\begin{cases} n.a_0 + a_1.\sum_{i=1}^{i=n} x_{1i} + ... + a_p.\sum_{i=1}^{i=n} x_{pi} = \sum_{i=1}^{i=n} y_i \\ a_0.\sum_{i=1}^{i=n} x_{1i} + a_1.\sum_{i=1}^{i=n} x_{1i}^2 + ... + a_p.\sum_{i=1}^{i=n} x_{1i}.x_{pi} = \sum_{i=1}^{i=n} x_{1i}.y_i \\ \quad ... \\ a_0.\sum_{i=1}^{i=n} x_{pi} + a_1.\sum_{i=1}^{i=n} x_{pi}.x_{1i} + ... + a_p.\sum_{i=1}^{i=n} x_{pi}^2 = \sum_{i=1}^{i=n} x_{pi}.y_i \end{cases}$$

• **Matriciellement**, on a, d'une part, $X^T_{(p+1,n)}.Y_{(n,1)} = \begin{pmatrix} 1 & 1 & ... & 1 \\ x_{11} & x_{12} & ... & x_{1n} \\ \vdots & & & \\ x_{p1} & x_{p2} & ... & x_{pn} \end{pmatrix}.\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$, soit la matrice unicolonne $\begin{pmatrix} \sum_{i=1}^{i=n} y_i \\ \sum_{i=1}^{i=n} x_{1i}.y_i \\ \vdots \\ \sum_{i=1}^{i=n} x_{pi}.y_i \end{pmatrix}$.

D'autre part, $X^T_{(p+1,n)}.X_{(n,p+1)} = \begin{pmatrix} 1 & 1 & ... & 1 \\ x_{11} & x_{12} & ... & x_{1n} \\ \vdots & \vdots & \vdots & \vdots \\ x_{p1} & x_{p2} & ... & x_{pn} \end{pmatrix}.\begin{pmatrix} 1 & x_{11} & ... & x_{p1} \\ 1 & x_{12} & ... & x_{p2} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n} & ... & x_{pn} \end{pmatrix}$, soit la matrice carrée symétrique $\begin{pmatrix} n & \sum_{i=1}^{i=n} x_{1i} & ... & \sum_{i=1}^{i=n} x_{pi} \\ \sum_{i=1}^{i=n} x_{1i} & \sum_{i=1}^{i=n} x_{1i}^2 & ... & \sum_{i=1}^{i=n} x_{1i}.x_{pi} \\ \vdots & \vdots & \vdots & \vdots \\ \sum_{i=1}^{i=n} x_{pi} & \sum_{i=1}^{i=n} x_{1i}.x_{pi} & ... & \sum_{i=1}^{i=n} x_{pi}^2 \end{pmatrix}$.

En définitive, par rapprochement avec le système d'équations explicité ci-dessus, il est immédiat que $\hat{A} = (X^T.X)^{-1}.X^T.Y$.

• D'autre part, remplaçant dans la relation ci-dessus, Y, par son expression manifeste $Y = X.A + \varepsilon$, il vient $\hat{A} = (X^T.X)^{-1}.(X^T.X).A + (X^T.X)^{-1}.X^T.\varepsilon$, d'où, la relation proposée dans l'énoncé $\hat{A} = A + (X^T.X)^{-1}.X^T.\varepsilon$.

2°) Par linéarité de l'espérance mathématique, $E(\hat{A}) = E(A) + E\left[(X^T.X)^{-1}.X^T.\varepsilon\right]$, soit $E(\hat{A}) = E(A) + (X^T.X)^{-1}.X^T.E(\varepsilon)$. Or, $E(A) = A$ (car les a_i ne sont pas aléatoires) et par ailleurs $E(\varepsilon) = 0$ (car les ε_i sont centrées).

Finalement $\hat{A}$ est sans biais puisque $E(\hat{A}) = A$.

3°) En premier lieu, on a $\varepsilon_{(n,1)}.\varepsilon_{(1,n)}^T = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.\begin{pmatrix} \varepsilon_1 & \varepsilon_2 & \dots & \varepsilon_n \end{pmatrix} = \begin{pmatrix} \varepsilon_1^2 & \varepsilon_1.\varepsilon_2 & \dots & \varepsilon_1.\varepsilon_n \\ \varepsilon_2.\varepsilon_1 & \varepsilon_2^2 & \dots & \varepsilon_2.\varepsilon_n \\ \vdots & \vdots & \vdots & \vdots \\ \varepsilon_n.\varepsilon_1 & \varepsilon_n.\varepsilon_2 & \dots & \varepsilon_n^2 \end{pmatrix}$.

Passant à l'espérance mathématique, on a d'une part $E(\varepsilon_i.\varepsilon_j) = 0$ si $i \neq j$ (hypothèse de *non corrélation des erreurs*) et d'autre part $E(\varepsilon_i^2) = Var(\varepsilon_i) + E(\varepsilon_i)^2 = Var(\varepsilon_i) = \sigma^2$ (en effet, par hypothèse, $E(\varepsilon_i) = 0, \forall i$, et par ailleurs, toutes les variances $Var(\varepsilon_i)$ sont égale à σ^2 suivant l'hypothèse d'*homoscédasticité*).

En résumé, $E\left[\varepsilon.\varepsilon^T\right] = \begin{pmatrix} \sigma^2 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma^2 \end{pmatrix} = \sigma^2.I_n$ où I_n désigne la *matrice identique d'ordre n*.

4°) La **matrice des variances et covariances** de $\widehat{A}$, soit Ω (matrice carrée symétrique d'ordre $p+1$), est égale, par définition, à $\begin{pmatrix} Var(\widehat{a_0}) & Cov(\widehat{a_0},\widehat{a_1}) & \dots & Cov(\widehat{a_0},\widehat{a_p}) \\ Cov(\widehat{a_1},\widehat{a_0}) & Var(\widehat{a_1}) & \dots & Cov(\widehat{a_1},\widehat{a_p}) \\ \vdots & \vdots & \vdots & \vdots \\ Cov(\widehat{a_p},\widehat{a_0}) & Cov(\widehat{a_p},\widehat{a_1}) & \dots & Var(\widehat{a_p}) \end{pmatrix}$.

Or, pour tout $i \in \{1,2,..,p\}$, on a $Var(\widehat{a_i}) = E\left[(\widehat{a_i} - E(\widehat{a_i}))^2\right] = E\left[(\widehat{a_i} - a_i).(\widehat{a_i} - a_i)\right]$, puisque suivant les résultats de la 2ème question, les estimateurs $\widehat{a_i}$ sont sans biais ($E(\widehat{a_i}) = a_i$). D'autre part, $Cov(\widehat{a_i},\widehat{a_j}) = E\left[(\widehat{a_i} - E(\widehat{a_i})).(\widehat{a_j} - E(\widehat{a_j}))\right]$, soit s'agissant de variables centrées, $Cov(\widehat{a_i},\widehat{a_j}) = E\left[(\widehat{a_i} - a_i).(\widehat{a_j} - a_j)\right]$.

- S'inspirant des questions antérieures, on constate que :

$\left[\widehat{A} - A\right].\left[\widehat{A} - A\right]^T = \begin{pmatrix} \widehat{a_0} - a_0 \\ \widehat{a_1} - a_1 \\ \vdots \\ \widehat{a_p} - a_p \end{pmatrix}.\begin{pmatrix} \widehat{a_0} - a_0 & \widehat{a_1} - a_1 & \dots & \widehat{a_p} - a_p \end{pmatrix}$, c'est-à-dire la matrice carrée $\begin{pmatrix} (\widehat{a_0} - a_0)^2 & (\widehat{a_0} - a_0).(\widehat{a_1} - a_1) & \dots & (\widehat{a_0} - a_0).(\widehat{a_p} - a_p) \\ (\widehat{a_1} - a_1).(\widehat{a_0} - a_0) & (\widehat{a_1} - a_1)^2 & \dots & (\widehat{a_1} - a_1).(\widehat{a_p} - a_p) \\ \vdots & \vdots & \vdots & \vdots \\ (\widehat{a_p} - a_p).(\widehat{a_0} - a_0) & (\widehat{a_p} - a_p).(\widehat{a_1} - a_1) & \dots & (\widehat{a_p} - a_p)^2 \end{pmatrix}$.

Passant à l'espérance mathématique, il est immédiat que $E\left\{\left[\widehat{A} - A\right].\left[\widehat{A} - A\right]^T\right\} = \Omega$.

- Or, $\widehat{A} - A = (X^T.X)^{-1}.X^T.\varepsilon$. De même, $\left[\widehat{A} - A\right]^T = \varepsilon^T.X.\left[(X^T.X)^{-1}\right]^T = \varepsilon^T.X.(X^T.X)^{-1}$ (en effet, $\left[(X^T.X)^{-1}\right]^T = (X^T.X)^{-1}$ puisque la matrice $(X^T.X)^{-1}$ est symétrique).

Finalement, $\Omega = E\left\{(X^T.X)^{-1}.X^T.\varepsilon.\varepsilon^T.X.(X^T.X)^{-1}\right\}$. Par linéarité de l'espérance mathématique, il en résulte $\Omega = (X^T.X)^{-1}.X^T.E\left(\varepsilon.\varepsilon^T\right).X.(X^T.X)^{-1}$. De la relation $E(\varepsilon.\varepsilon^T) = \sigma^2.I_n$ établie à la 3ème question, on obtient en définitive, le résultat $\Omega = \sigma^2.(X^T.X)^{-1}.X^T.I_n.X.(X^T.X)^{-1}$, qui s'écrit encore $\Omega = \sigma^2.(X^T.X)^{-1}.X^T X.(X^T.X)^{-1}$, soit après simplifications, $\Omega = \sigma^2.(X^T.X)^{-1}$. C'est le résultat cherché !.

5-a) La transposition des formulations matricielles précédentes à l'exemple proposé, conduit, en considérant les matrices $\hat{A} = \begin{pmatrix} \hat{a} \\ \hat{b} \\ \hat{c} \end{pmatrix}$, $X_{8,3)} = \begin{pmatrix} 1 & 1,75 & 0,52 \\ 1 & 1,33 & 0,59 \\ 1 & 1,43 & 0,61 \\ 1 & 1,86 & 0,50 \\ 1 & 1,81 & 0,54 \\ 1 & 1,66 & 0,49 \\ 1 & 1,60 & 0,27 \\ 1 & 2,00 & 0,47 \end{pmatrix}$, aux résultats $X^T.X = \begin{pmatrix} 8 & 13,44 & 3,99 \\ 13,44 & 22,93 & 6,66 \\ 3,99 & 6,66 & 2,06 \end{pmatrix}$. Par ailleurs, $X^T.Y = \begin{pmatrix} -23,4 \\ -38,04 \\ -12,93 \end{pmatrix}$.

L'inversion de la matrice (3,3), $X^T.X$ fait appel aux étapes ci-après :

→ *Calculer le déterminant de ladite matrice*, soit, après calculs, la valeur 0,197.

→ *Remplacer chacun des coefficients de la matrice $X^T.X$ par son cofacteur* (pour rappel, le cofacteur du coefficient situé à la $i^{ème}$ ligne et à la $j^{ème}$ colonne, est le déterminant de la matrice réduite obtenue en rayant la ligne i et la colonne j, le coefficient $(-1)^{i+j}$ étant par ailleurs à appliquer audit déterminant.

Ainsi, le cofacteur de 8 (élément à la 1ère ligne et à la 1ère colonne) est le déterminant $(-1)^{1+1} \times \begin{vmatrix} 22,93 & 6,66 \\ 6,66 & 2,06 \end{vmatrix} = 3,02$. De même, le cofacteur du coefficient 13,44 (à la 2ème ligne et la 1ère colonne), est égal à $(-1)^{2+1} \times \begin{vmatrix} 13,44 & 3,99 \\ 6,66 & 2,06 \end{vmatrix} = -1,20$. Et ainsi de suite, jusqu'à obtenir la matrice $H = \begin{pmatrix} 3,02 & -1,20 & -1,97 \\ -1,20 & 0,61 & 0,35 \\ -1,97 & 0,35 & 2,79 \end{pmatrix}$.

→ *Transposer la matrice ci-dessus*, qui pour le cas présent d'une matrice symétrique $(H^T = H)$ conduit à $\frac{1}{\det(X^T.X)}.H^T = \begin{pmatrix} 15,32 & -6,07 & -10,02 \\ -6,02 & 3,09 & 1,76 \\ -10,02 & 1,76 & 14,15 \end{pmatrix} \Rightarrow \hat{A} = \begin{pmatrix} 2,02 \\ 1,69 \\ -15,62 \end{pmatrix}$.

D'où, l'approximation cherchée $\hat{y} = 2,02 + 1,69.x_1 - 15,62.x_2$.

• Quant à σ^2, son estimation est fournie par $\frac{1}{n-p-1}.\sum_{i=1}^{i=n}(y_i-\widehat{y_i})^2$, soit pour le cas présent, $\widehat{\sigma}^2=\frac{1}{8-3}.\sum_{i=1}^{i=8}(y_i-\widehat{y_i})^2$, avec les valeurs :

y_i	-1,70	-4,61	-5,82	-1,17	-4,23	-3,31	0,42	-2,98
$\widehat{y_i}$	-3,14	-4,94	-5,09	-2,64	-3,95	-2,82	0,51	-1,93

En conclusion, $\widehat{\sigma}^2=\frac{6,995}{5}=1,399$.

5-b) Comme cela a déjà été présenté en rappels de cours (cf. paragraphe 2.2) et dans l'application 2.1, le **test de signification globale** $H_0 : b=c=0$ contre H_1 : »*Il existe au moins b ou c non nul*», permet de vérifier s'il existe une prédiction de y par x_1 et x_2 autre que la simple constante, la statistique utilisée étant $F=\frac{R^2/p}{(1-R^2)/n-p-1}$, F suivant la **loi de FISHER SNEDECOR** à p et $n-p-1$ degrés de liberté et R^2 désignant le *coefficient de détermination* $\frac{\sum_{i=1}^{i=n}(\widehat{y_i}-\overline{y})^2}{\sum_{i=1}^{i=n}(y_i-\overline{y})^2}$, soit numériquement $R^2=\frac{21,849}{28,844}=0,757$.

La **région critique** du test a pour forme $F\geq F_\alpha$, F_α vérifiant $Prob(F\geq F_\alpha)=\alpha$. Ainsi, pour $n=8$, $p=2$, et $\alpha=5\%$, a-t-on, suivant lecture dans la table de valeurs annexée (cf. loi de SNEDECOR), $F_\alpha=5,79$. Or, $F_{calculé}=\frac{R^2/2}{(1-R^2)/5}$ avec $R^2=0,757$, soit $F_{calculé}=7,81$.

La relation $F_{calculé}=7,81\succ F_\alpha=5,79$ conduit donc à retenir ici l'hypothèse H_1, c'est-à-dire la conclusion suivant laquelle la variation d'hémoglobine dépend d'au moins l'un des deux facteurs considérés (durée de l'opération, volume de sang perdu).

5-c) Cette fois, c'est le test $\begin{cases}H_0 : b=0\\H_1 : b\neq 0\end{cases}$ qui est proposé, la variable $\frac{\widehat{B}-b}{\widehat{\sigma}_{\widehat{B}}}$ suivant la **loi de STUDENT** à $n-p-1$ degrés de liberté (cf. rappels de cours, paragraphe 2.2).

Comme cela a été mis en évidence dans l'application 2.1, l'évaluation de $\widehat{\sigma}_{\widehat{B}}$ requiert des calculs un peu complexes, le résultat étant $\widehat{\sigma}^2_{\widehat{B}}=VIF(\widehat{B}).\frac{\widehat{\sigma}^2}{\sum_{i=1}^{i=n}(x_{1i}-\overline{x_1})^2}$. Reprenant, la procédure antérieurement exposée (cf. application 2.1), le coefficient de corrélation linéaire de X_1 et X_2 est égal à $\rho_{X_1,X_2}=\frac{\sum_{i=1}^{i=8}x_{1i}.x_{2i}-8.\overline{x_1}.\overline{x_2}}{\sqrt{(\sum_{i=1}^{i=8}x_{1i}^2-8.\overline{x_1}^2).(\sum_{i=1}^{i=8}x_{2i}^2-8.\overline{x_2}^2)}}$.

Par lecture des coefficients de la matrice $X^T.X$, il s'ensuit $\rho_{X_1,X_2} = \dfrac{6,66-8\times1,68\times0,499}{\sqrt{(22,93-8\times1,68^2).(2,06-8\times0,499^2)}} = -0,267$.

Ainsi, $VIF(\hat{B}) = \dfrac{1}{1-\rho^2_{X_1,X_2}} = 1,076$ et $\hat{\sigma}^2_{\hat{B}} = \dfrac{1,076\times1,399}{0,3484} = 4,32$.

• Mais, c'est là un **résultat simplificateur important**, $\hat{\sigma}^2_{\hat{B}}$ (comme également $\hat{\sigma}^2_{\hat{C}}$) est fourni directement par les *valeurs des coefficients de la diagonale* de la **matrice des variances et covariances** de $\hat{A}$ définie dans la 4ème question précédente, par l'équation $\Omega = \sigma^2.(X^T.X)^{-1}$.

On a en effet, d'après le résultat de la question 5-a) relatif à $(X^T.X)^{-1}$ et compte tenu de l'estimation $\hat{\sigma}^2 = 1,399$, $\Omega = 1,399\times\begin{pmatrix} 15,32 & -6,07 & -10,02 \\ -6,07 & 3,09 & 1,76 \\ -10,02 & 1,76 & 14,15 \end{pmatrix}$ soit, après produit,

$\Omega = \begin{pmatrix} 21,43 & -8,49 & -14,01 \\ -8,49 & 4,32 & 2,47 \\ -14,01 & 2,47 & 19,79 \end{pmatrix}$.

La variance estimée de $\hat{B}$ qui est lue à l'intersection de la 2ème ligne et de la 2ème colonne correspond bien au résultat du calcul antérieur à partir du facteur d'inflation de la variance, ce qui établit le résultat anoncé.

• En définitive, et pour conclure, la région critique du test a pour forme $|\hat{B}| \geq \pi$. Or, sous l'hypothèse $H_0 : b = 0$, la statistique $\dfrac{\hat{B}}{\hat{\sigma}_{\hat{B}}}$ suit la **loi de STUDENT** à $n-p-1$ degrés de liberté. Ainsi, si t_α désigne le seuil vérifiant $Prob(|T| \geq t_\alpha) = \alpha$, on a $\pi = t_\alpha.\hat{\sigma}_{\hat{B}}$.

Numériquement, $\alpha = 5\%, \nu = 8-2-1 = 5 \Rightarrow t_\alpha = 2,57$ (suivant lecture dans la table de STUDENT annexée). D'où, $\pi = 5,34$. Mais, $|\hat{B}_{calculé}| = 1,69 \prec \pi = 5,34$. On est donc amené cette fois, à ne pas rejeter l'hypothèse H_0 suivant laquelle la variation du taux d'hémoglobine ne dépend pas de la durée de l'opération.

C - Exercices complémentaires

1. Dans le cadre de travaux de recherche portant sur la durée de la saison de végétation en montagne, des stations météorologiques sont installées à différentes altitudes. La température moyenne ainsi que l'altitude (en mètres) de chaque saison sont relevées, les données en résultant étant présentées ci-dessous :

Altitude	1040	1230	1500	1600	1740	1950	2200	2530	2800	3100
Température	7,4	6	4,5	3,8	2,9	1,9	1,0	-1,2	-1,5	-4,5

A partir de l'altitude d'un lieu, on cherche à évaluer sa température moyenne sans avoir à implanter une nouvelle station et on opte pour un modèle de régression linéaire gaussien $y = a.x + b + \varepsilon$.

1°) Calculer les estimations de σ^2 et des coefficients de la droite de régression $\hat{y} = \hat{a}.x + \hat{b}$.

2°) Effectuer un test de pertinence permettant de vérifier que le coefficient a est non nul au risque de 5%.

3°) Sachant qu'une certaine plante ne survit qu'à une température moyenne supérieure à -6°, est-il raisonnable de penser qu'on ne trouvera pas cette plante à une altitude de 3500 mètres ?

Solution : 1°) Notant respectivement par x_i et y_i les résultats « *altitude* » et « *température* », il est rappelé que la **droite de régression** de y en x a pour équation $\hat{y} = \hat{a}.x + \hat{b}$ où $\hat{a} = \dfrac{\sum_{i=1}^{i=n}(x_i - \bar{x}).(y_i - \bar{y})}{\sum_{i=1}^{i=n}(x_i - \bar{x})^2}$ et $\hat{b} = \bar{y} - \hat{a}.\bar{x}$. Il en résulte numériquement :

$$\bar{x} = 1969 \qquad \sum_{i=1}^{i=10}(x_i - \bar{x})^2 = 4,155890$$

$$\bar{y} = 2,03 \qquad \sum_{i=1}^{i=10}(y_i - \bar{y})^2 = 121,201$$

$$\hat{a} = -0,005366 \qquad \sum_{i=1}^{i=10}(x_i - \bar{x}).(y_i - \bar{y}) = -22299,7$$

$$\hat{b} = 12,595272$$

d'où la droite de régression $\hat{y} = -0,00537.x + 12,59527$.

- Quant à σ^2 et se référant à l'application 5.1, son estimation est :

$$\hat{\sigma}^2 = \frac{1}{n-2}.\left[\sum_{i=1}^{i=n}(y_i - \bar{y})^2 - \frac{(\sum_{i=1}^{i=n}(x_i - \bar{x}).(y_i - \bar{y}))^2}{\sum_{i=1}^{i=n}(x_i - \bar{x})^2}\right]$$

soit, numériquement, $\hat{\sigma}^2 = \frac{1}{8} \times 1,545 = 0,193$.

- Il est à noter ici que la *somme des écarts totaux*, soit $\sum_{i=1}^{i=n}(y_i - \bar{y})^2 = 121,201$ et la *somme des écarts expliqués par le modèle*, soit $\sum_{i=1}^{i=n}(\hat{y}_i - \bar{y})^2 = 119,656$ sont très proches, leur différence formant *l'écart résiduel* $\sum_{i=1}^{i=n}(y_i - \hat{y}_i)^2 = 1,545$ déjà calculé ci-dessus. Ainsi, le **coefficient de détermination** $R^2 = \dfrac{\sum_{i=1}^{i=n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{i=n}(y_i - \bar{y})^2}$, soit $R^2 = 0,987$, est-il très proche de 1, de même le **coefficient de corrélation linéaire empirique** $r_{xy} = \sqrt{R^2} = 0,993$.

2°) Le test proposé est **bilatéral** de type $\begin{cases} H_0 : a = 0 \\ H_1 : a \neq 0 \end{cases}$. Or $\dfrac{\hat{a} - a}{\hat{\sigma} \Big/ \sqrt{\sum_{i=1}^{i=n} (x_i - \bar{x})^2}}$ suit la **loi de STUDENT** à $\nu = n - 2$ degrés de liberté (cf. rappels de cours, paragraphe 1.7). La **région critique** ayant pour forme $|\hat{a}| \geq \pi$, il s'ensuit, de par la donnée de l'erreur de première espèce $\alpha = Prob(|\hat{a}| \geq \pi / a = 0)$, la relation :

$$Prob(|T| = \frac{|\hat{a}|}{\hat{\sigma} \Big/ \sqrt{\sum_{i=1}^{i=n} (x_i - \bar{x})^2}} \geq \frac{\pi}{\hat{\sigma} \Big/ \sqrt{\sum_{i=1}^{i=n} (x_i - \bar{x})^2}}) = \alpha .$$

Notant par t_α le nombre vérifiant $Prob(|T| \geq t_\alpha) = \alpha$, il vient $\pi = t_\alpha . \dfrac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{i=n} (x_i - \bar{x})^2}}$.

Numériquement, et pour $n = 10, \alpha = 5\%$, on a successivement $t_\alpha = 2,306$; $\pi = 0,000496$. Or, $|\hat{a}|_{calculé} = 0,00536 \succ \pi = 0,00049$. C'est donc assurément, l'hypothèse $H_1 : a \neq 0$ qu'il faut retenir ici ce que confirme la *très forte valeur du coefficient de corrélation linéaire empirique* entre X et Y.

3°) Pour traiter cette question, on va expliciter **l'intervalle de prédiction** de la température moyenne à une altitude de 3500 mètres et vérifier si cet intervalle contient ou non des valeurs inférieures à -6°, ce qui conduirait alors à écarter l'hypothèse de survivance, à cette altitude, de la plante considérée. Le seuil choisi est fixé ici encore à 5%.

Suivant les rappels de cours (cf. paragraphe 1.7), la prédiction à la valeur x_0 a pour encadrement au seuil de confiance $1 - \alpha$, l'intervalle :

$$\hat{a}.x_0 + \hat{b} \pm t_\alpha . \hat{\sigma} . \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{n.s_x^2}}$$

t_α satisfaisant pour la loi de STUDENT, $T(n-2)$, la relation $Prob(|T(n-2)| \geq t_\alpha) = \alpha$. On obtient ainsi, numériquement, l'encadrement :

$$-6,18 \pm 2,306 \times 0,4393 \times \sqrt{1 + \frac{1}{10} + \frac{(3500 - 1969)^2}{4155890}}$$

soit, l'intervalle $[-7,47; -4,87]$.

• Assurément, on se trouve être dans une zone où la survivance de la plante en question n'est pas assurée.

2. On veut prédire la hauteur H d'un arbre en fonction de son diamètre D. Pour obtenir une régression linéaire , on effectue un changement de variables en posant $Y = \ln H$ et $X = \ln D$. On obtient les cinq mesures :

x_i	-1,61	-1,20	-0,97	-0,51	-0,42
y_i	2,22	2,27	2,38	2,60	2,65

1°) Calculer le coefficient de corrélation linéaire empirique entre X et Y.

2°) Expliciter l'équation de la droite de régression de Y en X.

3°) Tester la signification de cette régression au seuil 5%.

4°) Evaluer la hauteur prévue d'un arbre dont le diamètre est 0,70 m.

5°) Expliciter un intervalle de confiance de niveau 95% pour la prédiction de la hauteur d'un arbre de diamètre 0,70 m.

Solution : 1°) Le **coefficient de corrélation linéaire empirique** est défini par la formule $r_{xy} = \dfrac{\sum_{i=1}^{i=n}(x_i-\bar{x}).(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{i=n}(x_i-\bar{x})^2}.\sqrt{\sum_{i=1}^{i=n}(y_i-\bar{y})^2}}$, soit numériquement, $\bar{x}=-0,942$; $\bar{y}=2,424$; $\sum_{i=1}^{i=n}(x_i-\bar{x})^2=0,97268$; $\sum_{i=1}^{i=n}(y_i-\bar{y})^2=0,14932$; $\sum_{i=1}^{i=n}(x_i-\bar{x}).(y_i-\bar{y})=0,37124$, ce qui entraîne $r_{xy}=0,974$. Bref, un coefficient qui laisse présumer une corrélation linéaire étroite entre X et Y.

2°) La **droite de régression** de Y en x a pour équation $\hat{Y}=\hat{a}.x+\hat{b}$ où $\hat{a}=\dfrac{\sum_{i=1}^{i=n}(x_i-\bar{x}).(y_i-\bar{y})}{\sum_{i=1}^{i=n}(x_i-\bar{x})^2}$ et $\hat{b}=\bar{y}-\hat{a}.\bar{x}$. On a donc numériquement, $\hat{a}=0,382$ et $\hat{b}=2,783$, d'où l'équation cherchée $\hat{y}=0,382.x+2,783$.

3°) On peut indifféremment tester ici $H_0 : a=0$ contre $H_1 : a\neq 0$ ou utiliser *l'analyse de la variance* qui porte sur la comparaison entre l'estimateur des écarts dus au modèle (*écarts expliqués*) et ceux qui restent inexpliqués (*écarts résiduels*). Se reportant aux rappels de cours (cf. paragraphe 1.5), on a d'une part, pour *estimateur des écarts expliqués*, la statistique $SCR/(2-1)$ avec $SCR=\sum_{i=1}^{i=n}(\hat{y}_i-\bar{y})^2$, soit numériquement 0,14169, et pour *estimateur des écarts résiduels*, la statistique $SCE/(n-2)$ avec $SCE=\sum_{i=1}^{i=n}(y_i-\hat{y}_i)^2$, soit numériquement 0,00254, la somme $SCE+SCR$ étant par ailleurs égale à *l'écart total* $SCT=\sum_{i=1}^{i=n}(y_i-\bar{y})^2=0,14932$.

Lorsque $H_0 : a=0$ est varie, le rapport $F=\dfrac{SCR}{SCE/(n-2)}$ suit la **loi de FISHER SNEDECOR** de type $F(1,n-2)$. Dans le cas d'une régression significative, F est grand, la **région critique** ayant donc pour forme $F\geq F_\alpha$.

Par lecture dans la table de valeurs annexée, le seuil F_α qui, au niveau de signification $\alpha=5\%$, vérifie $Prob(F\geq F_\alpha)=\alpha$ est égal à $F_\alpha=10,13$ $(\nu_1=1,\nu_2=3)$. Or, $F_{calculé}=55,71$. C'est donc l'hypothèse H_1 de signification de la régression qu'on retient ici puisque $F_{calculé}=55,71\succ F_\alpha=10,13$.

4°) Pour un arbre dont le diamètre est $d^* = 0,70$, la prévision y^* associée à $x^* = \ln d^* = -0,357$ est égale à $\hat{a}.x^* + \hat{b}$, soit numériquement $y^* = 2,647$. Il en résulte pour la hauteur h^* de l'arbre (qui vérifie la relation $y^* = \ln h^*$) la valeur prévue par le modèle, soit $h^* = 14,117$ m.

5°) Plus précisément, l'intervalle de prédiction de y pour $x = x^* = -0,357$ est caractérisé par $\hat{a}.x^* + \hat{b} \pm t_\alpha.\hat{\sigma}.\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{i=n}(x_i - \bar{x})^2}}$, t_α vérifiant relativement à la **loi de STUDENT**, $T(n-2)$, la relation $\Pr ob(|T| \geq t_\alpha) = \alpha$.

On a ainsi, $\hat{a}.x^* + \hat{b} = 2,647$ (cf. 4ème question), $t_\alpha = 3,1824$ (cf. lecture dans table des valeurs annexée), $\hat{\sigma}^2 = \frac{SCE}{(n-2)} = 0,00254$ (cf. 3ème question), d'où, après calculs, l'encadrement $[2,447 - 2,847]$. Pour la hauteur $d = \exp(y)$ et au niveau de confiance 95% ($\alpha = 5\%$), on a donc comme encadrement le résultat $[11,56 - 17,24]$, en mètres.

3. On étudie la croissance d'une plante à partir d'un instant considéré comme instant initial. On effectue des mesures du diamètre de la tige principale et on obtient les résultats ci-dessous :

Temps t en semaines	0	2	6	10	14
Diamètre d_i en cms	0,4	1,2	5,4	6,4	7,8

1°) Posant $u_i = \ln(\frac{8}{d_i} - 1)$ montrer qu'il existe un modèle linéaire significatif entre les t_i et les u_i, modèle qu'on explicitera et au sujet duquel on calculera le coefficient de corrélation linéaire.

2°) En déduire que, pour la plante considérée, le diamètre de sa tige principale est donné par une relation de la forme $d(t) = \frac{8}{1 + C.e^{-\alpha.t}}$ dans laquelle on estimera C et α.

Solution : 1°) Le calcul des u_i conduit au tableau ci-dessous :

t_i	0	2	6	10	14
u_i	2,944	1,735	-0,7309	-1,3863	-3,6636

D'ores et déjà, en reportant les points $M_i(t_i, u_i)$ dans un repère (t,u), on remarque qu'ils sont sensiblement alignés, d'où la mise en œuvre de la régression linéaire proposée. Par rapport à cette dernière et en fonction des rappels de cours (cf. paragraphe 1.2), la meilleure représentation affine de u par t a pour forme, la droite de régression $\hat{u} = \hat{a}.t + \hat{b}$ où $\hat{a} = \frac{\sum_{i=1}^{i=n}(u_i - \bar{u}).(t_i - \bar{t})}{\sum_{i=1}^{i=n}(t_i - \bar{t})^2}$ et $\hat{b} = \bar{u} - \hat{a}.\bar{t}$ (bien évidemment $\bar{u} = \frac{\sum_{i=1}^{i=n} u_i}{n}$ et $\bar{t} = \frac{\sum_{i=1}^{i=n} t_i}{n}$).

Numériquement, on obtient ainsi et successivement, $\bar{u} = -0,2203$; $\bar{t} = 6,4$; $\sum_{i=1}^{i=5}(u_i - \bar{u}).(t_i - \bar{t}) = -59,0180$; $\sum_{i=1}^{i=5}(t_i - \bar{t})^2 = 131,2$; $\hat{a} = -0,4498$; $\hat{b} = 2,6586$, soit, après arrondis, le modèle $\hat{u} = -0,45.t + 2,66$.

• Le **coefficient de corrélation linéaire** est caractérisé *empiriquement*, par la formule $r_{tu} = \dfrac{\sum_{i=1}^{i=n}(u_i - \bar{u}).(t_i - \bar{t})}{\sqrt{\sum_{i=1}^{i=n}(u_i - \bar{u})^2}.\sqrt{\sum_{i=1}^{i=n}(t_i - \bar{t})^2}}$, soit numériquement, $r_{tu} = \dfrac{-59,0180}{\sqrt{27,3135 \times 131,2}} = -0,9859$ (arrondi à 0,99). Il est donc très proche de 1, ce qui valide le choix du modèle retenu.

2°) L'analogie entre la relation $d = \dfrac{8}{1 + C.e^{-\alpha.t}}$ et le modèle linéaire $\ln u = a.t + b$, conduit immédiatement, en posant $u = \ln(\dfrac{8}{d} - 1)$ à $1 + C.e^{-\alpha.t} = \dfrac{8}{d}$, soit $u = C.e^{-\alpha.t}$, ce qui entraîne $\ln u = -\alpha.t + \ln C$.

Par comparaison, on a donc $\alpha = -a, C = e^b$. Il en résulte pour α et C, les estimations $\hat{\alpha}$ et $\hat{C}$ égales respectivement à $-(-0,45) = 0,45$ et $\exp(2,66) = 14,276$. D'où, après arrondis, le modèle $d = \dfrac{8}{1 + 14,3.e^{-0,45.t}}$.

4. On souhaite modéliser l'investissement I_i (exprimé en millions d'euros) d'une firme i par $I_i = b_0 + b_1.V_i + b_2.K_i + b_3.S_i + \varepsilon_i, i \in \{1, 2, ..., n\}$, où V_i et K_i représentent respectivement les ventes de l'entreprise et le stock de capital exprimés en millions d'euros, S_i une variable indicatrice égale à 1 si la firme est privée et à 0 si la firme est publique, et ε_i le résidu aléatoire dont on supposera que la loi est de type normale, soit $N(0, \sigma^2)$.

A partir de 34 observations, la méthode des moindres carrés a conduit aux résultats :

$$\hat{I}_i = 10 + 0,2.V_i + 0,02.K_i + 0,3.S_i$$

avec pour matrice de variances – covariances des coefficients estimés, la matrice

$$\Omega = \begin{pmatrix} 4,0 & -0,16 & 0,10 & -0,20 \\ -0,16 & 0,01 & -0,002 & 0,005 \\ 0,10 & -0,002 & 0,0001 & -0,003 \\ -0,20 & 0,005 & -0,003 & 0,01 \end{pmatrix} \text{ et } \hat{b} = (\hat{b_0}, \hat{b_1}, \hat{b_2}, \hat{b_3})^T.$$

1°) Prévoir le montant de l'investissement pour une firme privée ayant un stock de capital de 1 milliards d'euros et des ventes de 100 millions d'euros.

2°) Tester au seuil $\alpha = 5\%$, l'hypothèse $H_0 : b_1 = 0,3$ contre $H_1 : b_1 \neq 0,3$.

3°) Tester au seuil $\alpha = 5\%$, l'hypothèse $H_0 : b_2 = 0,03$ contre $H_1 : b_2 \prec 0,03$.

4°) Tester au seuil $\alpha = 5\%$, l'hypothèse $H_0 : b_1 = b_3$ contre $H_1 : b_1 \neq b_3$.

Solution : 1°) Pour un stock de capital de 1 milliards d'euros ($K_i = 1000$), des ventes de 100 millions d'euros ($V_i = 100$), et une entreprise privée ($S_i = 1$), le modèle proposé fournit la prédiction $I_i = 50,3$ millions d'euros.

2°) Suivant les résultats rappelés précédemment (cf. rappels de cours, paragraphe 2.3), les statistiques $\frac{\widehat{B_1} - b_1}{\widehat{\sigma}_{\widehat{B_1}}}, \frac{\widehat{B_2} - b_2}{\widehat{\sigma}_{\widehat{B_2}}}, \frac{\widehat{B_3} - b_3}{\widehat{\sigma}_{\widehat{B_3}}}$ associées aux coefficients du **modèle linéaire multiple** $I = b_0 + b_1.V + b_2.K + b_3.S + \varepsilon$ snivent la **loi de STUDENT** à $n - p - 1$ degrés de liberté ($n = 34, p = 3$), *loi proche de la loi normale* par convergence puisque n avoisine la valeur 30.

D'autre part, comme cela a été mis en évidence dans l'application 2.2 du présent chapitre, les estimations $\widehat{\sigma}^2_{\widehat{B_1}}, \widehat{\sigma}^2_{\widehat{B_2}}, \widehat{\sigma}^2_{\widehat{B_3}}$, sont fournies par les *valeurs des coefficients de la diagonale de la matrice des variances et covariances des coefficients estimés*, matrice $\Omega = \sigma^2.(X^T.X)^{-1}$ fournie ici dans l'énoncé.

- Pour le *test bilatéral* $\begin{cases} H_0 : b_1 = 0,3 \\ H_1 : b_1 \neq 0,3 \end{cases}$, on a par lecture dans la matrice Ω, $\widehat{\sigma}^2_{\widehat{B_1}} = 0,01$.

Des applications antérieures 2.1 et 2.2 du présent chapitre et pour le test bilatéral en question, résulte la **région critique** $\left|\widehat{B_1} - 0,3\right| \geq t_\alpha.\widehat{\sigma}_{\widehat{B_1}}$ (puisque sous l'hypothèse $H_0 : b_1 = 0,3$), t_α étant égal quant à lui à 2,04 lorsque $\alpha = 5\%$ (cf. lecture dans table des valeurs de la **loi de STUDENT** annexée).

Finalement, $\left|\widehat{B_1} - 0,3\right|_{calculé} = 0,1 \prec 0,204$, ce qui ne permet pas de rejeter l'hypothèse nulle $H_0 : b_1 = 0,3$.

3°) S'agissant cette fois du *test unilatéral* $\begin{cases} b_2 = 0,03 \\ b_2 \prec 0,03 \end{cases}$, on a successivement $\widehat{\sigma}^2_{\widehat{B_2}} = 0,0001$ (cf. lecture dans la matrice Ω du coefficient à la 3ème ligne et à la 3ème colonne). La **région critique** a pour forme $\widehat{B_2} \leq \pi = 0,03 + t_\alpha.\widehat{\sigma}_{\widehat{B_2}}$, avec t_α vérifiant $\Pr ob(T \leq t_\alpha) = \alpha$.

Lorsque $\alpha = 5\%$ et pour le test unilatéral considéré, la consultation de la table de valeurs annexée de la **loi de STUDENT**, conduit immédiatement à $t_\alpha = -1,6973$. En définitive, $B_{2_{calculé}} = 0,02 \succ \pi = 0,013$ ce qui, ici encore, conduit à retenir l'hypothèse nulle $H_0 : b_2 = 0,03$.

4°) $\widehat{B_1}$ et $\widehat{B_3}$ suivent respectivement les **lois de STUDENT** de moyennes b_1 et b_3 et de variances $\widehat{\sigma}^2_{\widehat{B_1}} = 0,01$ et $\widehat{\sigma}^2_{\widehat{B_3}} = 0,01$. Se reportant aux éléments relatifs aux **tests de comparaison entre moyennes** (cf. rappels de cours du chapitre III, paragraphe 2.4.b), et recourant à *l'approximation normale* de la loi de STUDENT pour simplifier les choses, $\frac{\widehat{B_1} - \widehat{B_3} - (b_1 - b_3)}{\widehat{\sigma}_{\widehat{B_1} - \widehat{B_3}}}$ suit **la loi normale** $N(0,1)$. Il s'ensuit la **région critique** $\left|\widehat{B_1} - \widehat{B_3}\right| \geq \pi = t_\alpha.\widehat{\sigma}_{\widehat{B_1} - \widehat{B_3}}$ où t_α vérifie $\Pr ob(|\xi| \geq t_\alpha) = \alpha$, ξ désignant la loi normale $N(0,1)$. Pour $\alpha = 5\%$, $t_\alpha = 1,96$.

D'autre part, $\widehat{\sigma}^2_{\widehat{B_1} - \widehat{B_3}} = \widehat{\sigma}^2_{\widehat{B_1}} + \widehat{\sigma}^2_{\widehat{B_3}} - 2.\mathrm{cov}(\widehat{B_1}, \widehat{B_3}) = 0,01 + 0,01 - 2 \times 0,005 = 0,01$. Finalement, $\pi = 1,96 \times 0,1 = 0,196$. Or, $\left|\widehat{B_1} - \widehat{B_3}\right|_{calculé} 0,1 \prec \pi = 0,196$. C'est donc le résultat $H_0 : b_1 = b_3$ qu'il faut retenir ici.

ANNEXES

LOI NORMALE CENTREE REDUITE

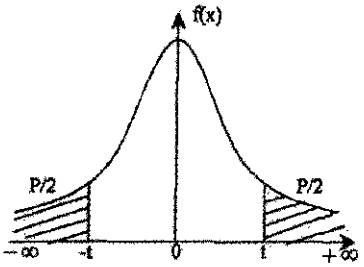
Cette table indique, pour certaines valeurs de t, la valeur de la surface hachurée, c'est-à-dire la valeur de :

$$\Pi(t) = \Pr ob(\xi \le t) = \frac{1}{\sqrt{2\pi}} \cdot \int_{-\infty}^{t} e^{-\frac{\xi^2}{2}} d\xi$$

t	0	1	2	3	4	5	6	7	8	9
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3,0	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990
3,1	0,9990	0,9991	0,9991	0,9991	0,9992	0,9992	0,9992	0,9992	0,9993	0,9993
3,2	0,9993	0,9993	0,9994	0,9994	0,9994	0,9994	0,9994	0,9995	0,9995	0,9995
3,3	0,9995	0,9995	0,9995	0,9996	0,9996	0,9996	0,9996	0,9996	0,9996	0,9997
3,4	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9998
3,5	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998
3,6	0,9998	0,9998	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999
3,7	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999
3,8	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999
3,9	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000

LOI DE STUDENT

Valeurs de t ayant la probabilité P d'être dépassées en valeur absolue.



V \P	90%	80%	70%	60%	50%	40%	30%	20%	10%	5%	1%
1	0,1584	0,3249	0,5095	0,7265	1,0000	1,3764	1,9626	3,0777	6,3137	12,7062	63,6559
2	0,1421	0,2887	0,4447	0,6172	0,8165	1,0607	1,3862	1,8856	2,9200	4,3027	9,9250
3	0,1366	0,2767	0,4242	0,5844	0,7649	0,9785	1,2498	1,6377	2,3534	3,1824	5,8408
4	0,1338	0,2707	0,4142	0,5686	0,7407	0,9410	1,1896	1,5332	2,1318	2,7765	4,6041
5	0,1322	0,2672	0,4082	0,5594	0,7267	0,9195	1,1558	1,4759	2,0150	2,5706	4,0321
6	0,1311	0,2648	0,4043	0,5534	0,7176	0,9057	1,1342	1,4398	1,9432	2,4469	3,7074
7	0,1303	0,2632	0,4015	0,5491	0,7111	0,8960	1,1192	1,4149	1,8946	2,3646	3,4995
8	0,1297	0,2619	0,3995	0,5459	0,7064	0,8889	1,1081	1,3968	1,8595	2,3060	3,3554
9	0,1293	0,2610	0,3979	0,5435	0,7027	0,8834	1,0997	1,3830	1,8331	2,2622	3,2498
10	0,1289	0,2602	0,3966	0,5415	0,6998	0,8791	1,0931	1,3722	1,8125	2,2281	3,1693
11	0,1286	0,2596	0,3956	0,5399	0,6974	0,8755	1,0877	1,3634	1,7959	2,2010	3,1058
12	0,1283	0,2590	0,3947	0,5386	0,6955	0,8726	1,0832	1,3562	1,7823	2,1788	3,0545
13	0,1281	0,2586	0,3940	0,5375	0,6938	0,8702	1,0795	1,3502	1,7709	2,1604	3,0123
14	0,1280	0,2582	0,3933	0,5366	0,6924	0,8681	1,0763	1,3450	1,7613	2,1448	2,9768
15	0,1278	0,2579	0,3928	0,5357	0,6912	0,8662	1,0735	1,3406	1,7531	2,1315	2,9467
16	0,1277	0,2576	0,3923	0,5350	0,6901	0,8647	1,0711	1,3368	1,7459	2,1199	2,9208
17	0,1276	0,2573	0,3918	0,5344	0,6892	0,8633	1,0690	1,3334	1,7396	2,1098	2,8982
18	0,1274	0,2571	0,3915	0,5338	0,6884	0,8620	1,0672	1,3304	1,7341	2,1009	2,8784
19	0,1274	0,2569	0,3912	0,5333	0,6876	0,8610	1,0655	1,3277	1,7291	2,0930	2,8609
20	0,1273	0,2567	0,3909	0,5329	0,6870	0,8600	1,0640	1,3253	1,7247	2,0860	2,8453
21	0,1272	0,2566	0,3906	0,5325	0,6864	0,8591	1,0627	1,3232	1,7207	2,0796	2,8314
22	0,1271	0,2564	0,3904	0,5321	0,6858	0,8583	1,0614	1,3212	1,7171	2,0739	2,8188
23	0,1271	0,2563	0,3902	0,5317	0,6853	0,8575	1,0603	1,3195	1,7139	2,0687	2,8073
24	0,1270	0,2562	0,3900	0,5314	0,6848	0,8569	1,0593	1,3178	1,7109	2,0639	2,7970
25	0,1269	0,2561	0,3898	0,5312	0,6844	0,8562	1,0584	1,3163	1,7081	2,0595	2,7874
26	0,1269	0,2560	0,3896	0,5309	0,6840	0,8557	1,0575	1,3150	1,7056	2,0555	2,7787
27	0,1268	0,2559	0,3894	0,5306	0,6837	0,8551	1,0567	1,3137	1,7033	2,0518	2,7707
28	0,1268	0,2558	0,3893	0,5304	0,6834	0,8546	1,0560	1,3125	1,7011	2,0484	2,7633
29	0,1268	0,2557	0,3892	0,5302	0,6830	0,8542	1,0553	1,3114	1,6991	2,0452	2,7564
30	0,1267	0,2556	0,3890	0,5300	0,6828	0,8538	1,0547	1,3104	1,6973	2,0423	2,7500
40	0,1265	0,2550	0,3881	0,5286	0,6807	0,8507	1,0500	1,3031	1,6839	2,0211	2,7045
50	0,1263	0,2547	0,3875	0,5278	0,6794	0,8489	1,0473	1,2987	1,6759	2,0086	2,6778
60	0,1262	0,2545	0,3872	0,5272	0,6786	0,8477	1,0455	1,2958	1,6706	2,0003	2,6603
80	0,1261	0,2542	0,3867	0,5265	0,6776	0,8461	1,0432	1,2922	1,6641	1,9901	2,6387
100	0,1260	0,2540	0,3864	0,5261	0,6770	0,8452	1,0418	1,2901	1,6602	1,9840	2,6259
120	0,1259	0,2539	0,3862	0,5258	0,6765	0,8446	1,0409	1,2886	1,6576	1,9799	2,6174
200	0,1258	0,2537	0,3859	0,5252	0,6757	0,8434	1,0391	1,2858	1,6525	1,9719	2,6006
∞	0,1257	0,2533	0,3853	0,5244	0,6745	0,8416	1,0364	1,2816	1,6449	1,9600	2,5758

LOI DU CHI-DEUX DE PEARSON

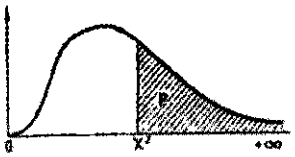
ν désigne le nombre de degrés de liberté.

Pour ν compris entre 30 et 100, on admettra que $\sqrt{2.\chi^2} - \sqrt{2.\nu - 1}$ est approximativement distribué selon la loi normale centrée réduite.

Pour ν supérieur à 100, on admettra que $(\chi^2 - \nu)/\sqrt{2.\nu}$ est approximativement distribuée selon la loi normale centrée réduite.

DISTRIBUTION DE χ^2 (Loi de K. Pearson)

Valeur de χ^2 ayant la probabilité P dêtre dépassée.



ν	0,9	0,8	0,7	0,5	0,3	0,2	0,1	0,05	0,02	0,01
1	0,0158	0,0642	0,148	0,455	1,074	1,642	2,706	3,841	5,412	6,635
2	0,211	0,446	0,713	1,386	2,408	3,219	4,605	5,991	7,824	9,210
3	0,584	1,005	1,424	2,366	3,665	4,642	6,251	7,815	9,837	11,345
4	1,064	1,649	2,195	3,357	4,878	5,989	7,779	9,488	11,668	13,277
5	1,610	2,343	3,000	4,351	6,064	7,289	9,236	11,070	13,388	15,086
6	2,204	3,070	3,828	5,348	7,231	8,558	10,645	12,592	15,033	16,812
7	2,833	3,822	4,671	6,346	8,383	9,803	12,017	14,067	16,622	18,475
8	3,490	4,594	5,527	7,344	9,524	11,030	13,362	15,507	18,168	20,090
9	4,168	5,380	6,393	8,343	10,656	12,242	14,684	16,919	19,679	21,666
10	4,865	6,179	7,267	9,342	11,781	13,442	15,987	18,307	21,161	23,209
11	5,578	6,989	8,148	10,341	12,899	14,631	17,275	19,675	22,618	24,725
12	6,304	7,807	9,034	11,340	14,011	15,812	18,549	21,026	24,054	26,217
13	7,041	8,634	9,926	12,340	15,119	16,985	19,812	22,362	25,471	27,688
14	7,790	9,467	10,821	13,339	16,222	18,151	21,064	23,685	26,873	29,141
15	8,547	10,307	11,721	14,339	17,322	19,311	22,307	24,996	28,259	30,578
16	9,312	11,152	12,624	15,338	18,418	20,465	23,542	26,296	29,633	32,000
17	10,085	12,002	13,531	16,338	19,511	21,615	24,769	27,587	30,995	33,409
18	10,865	12,857	14,440	17,338	20,601	22,760	25,989	28,869	32,346	34,805
19	11,651	13,716	15,352	18,338	21,689	23,900	27,204	30,144	33,687	36,191
20	12,443	14,578	16,266	19,337	22,775	25,038	28,412	31,410	35,020	37,566
21	13,240	15,445	17,182	20,337	23,858	26,171	29,615	32,671	36,343	38,932
22	14,041	16,314	18,101	21,337	24,939	27,301	30,813	33,924	37,659	40,289
23	14,848	17,187	19,021	22,337	26,018	28,429	32,007	35,172	38,968	41,638
24	15,659	18,062	19,943	23,337	27,096	29,553	33,196	36,415	40,270	42,980
25	16,473	18,940	20,867	24,337	28,172	30,675	34,382	37,652	41,566	44,314
26	17,292	19,820	21,792	25,336	29,246	31,795	35,563	38,885	42,856	45,642
27	18,114	20,703	22,719	26,336	30,319	32,912	36,741	40,113	44,140	46,963
28	18,939	21,588	23,647	27,336	31,391	34,027	37,916	41,337	45,419	48,278
29	19,768	22,475	24,577	28,336	32,461	35,139	39,087	42,557	46,693	49,588
30	20,599	23,364	25,508	29,336	33,530	36,250	40,256	43,773	47,962	50,892

LOI DE FISHER SNEDECOR

La table fournit pour $\alpha = 5\%$ et pour la variable de FISHER SNEDECOR à ν_1 et ν_2 degrés de liberté, soit $F(\nu_1,\nu_2)$, le seuil F_α qui a la probabilité α d'être dépassé.

ν_2 / ν_1	1	2	3	4	5	6	7	8	9	10	12	20	30	40	80	100
1	161 45	199.50	215.71	224 58	230.16	233.99	236.77	238 88	240.54	241 88	243.90	248.02	250.10	251.14	252.72	253 04
2	18 51	19.00	19.16	19 25	19.30	19.33	19.35	19.37	19 38	19.40	19.41	19.45	19.46	19.47	19.48	19.49
3	10 13	9 55	9.28	9 12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.66	8 62	8.59	8.56	8.55
4	7 71	6 94	6.59	6.39	6 26	6.16	6.09	6.04	6 00	5.96	5.91	5 80	5.75	5.72	5.67	5 66
5	6 61	5.79	5.41	5.19	5 05	4.95	4.88	4.82	4.77	4.74	4.68	4.56	4.50	4.46	4.41	4.41
6	5 99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.87	3.81	3.77	3.72	3.71
7	5.59	4.74	4.35	4.12	3.97	3 87	3.79	3.73	3.68	3.64	3.57	3.44	3.38	3.34	3.29	3.27
8	5 32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.15	3.08	3.04	2.99	2 97
9	5 12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	2.94	2.86	2.83	2.77	2.76
10	4.96	4 10	3.71	3 48	3 33	3 22	3.14	3 07	3.02	2.98	2.91	2 77	2.70	2.66	2.60	2.59
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2 95	2 90	2.85	2.79	2.65	2.57	2.53	2.47	2 46
12	4.75	3.89	3.49	3.26	3 11	3.00	2.91	2.85	2.80	2.75	2.69	2 54	2.47	2.43	2.36	2.35
13	4.67	3 81	3.41	3.18	3 03	2.92	2.83	2.77	2.71	2.67	2.60	2.46	2.38	2.34	2.27	2.26
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.39	2.31	2.27	2.20	2.19
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.33	2.25	2.20	2.14	2 12
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2 59	2 54	2.49	2.42	2 28	2 19	2.15	2.08	2 07
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2 23	2.15	2.10	2.03	2.02
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2 46	2.41	2 34	2 19	2.11	2.06	1.99	1 98
19	4 38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.16	2.07	2.03	1 96	1 94
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.12	2.04	1 99	1.92	1 91
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.10	2.01	1 96	1.89	1.88
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.07	1.98	1.94	1.86	1 85
23	4 28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.05	1.96	1.91	1.84	1 82
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.03	1.94	1 89	1.82	1 80
25	4.24	3 39	2.99	2 76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.01	1 92	1.87	1.80	1.78
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2 32	2.27	2.22	2.15	1.99	1.90	1.85	1.78	1.76
27	4 21	3.35	2.96	2.73	2 57	2.46	2.37	2 31	2 25	2 20	2.13	1.97	1 88	1.84	1 76	1.74
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	1.96	1.87	1.82	1.74	1 73
29	4.18	3.33	2.93	2.70	2.55	2 43	2 35	2 28	2.22	2.18	2 10	1.94	1.85	1.81	1.73	1.71
30	4 17	3.32	2.92	2.69	2 53	2.42	2 33	2.27	2.21	2.16	2.09	1.93	1 84	1.79	1.71	1 70
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2 12	2.08	2.00	1 84	1.74	1.69	1.61	1 59
60	4.00	3 15	2.76	2 53	2.37	2 25	2.17	2.10	2.04	1.99	1 92	1 75	1 65	1 59	1.50	1.48
80	3.96	3.11	2.72	2.49	2.33	2.21	2 13	2.06	2.00	1.95	1.88	1.70	1.60	1.54	1.45	1.43
100	3.94	3.09	2.70	2.46	2 31	2.19	2.10	2.03	1 97	1.93	1.85	1 68	1.57	1.52	1.41	1 39
200	3.89	3.04	2.65	2 42	2.26	2.14	2.06	1.98	1 93	1.88	1.80	1 62	1.52	1.46	1.35	1.32

LOI DE FISHER SNEDECOR

La table fournit pour $\alpha = 2{,}5\%$ et pour la variable de FISHER SNEDECOR à ν_1 et ν_2 degrés de liberté, soit $F(\nu_1,\nu_2)$, le seuil F_α qui a la probabilité α d'être dépassé.

ν_2 \ ν_1	1	2	3	4	5	6	7	8	9	10	12	20	30	40	80	100
1	647.79	799.48	864.15	899.60	921.83	937.11	948.20	956.64	963.28	968.63	976.72	993.08	1001.40	1005.60	1011.91	1013.16
2	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39	39.40	39.41	39.45	39.46	39.47	39.49	39.49
3	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47	14.42	14.34	14.17	14.08	14.04	13.97	13.96
4	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84	8.75	8.56	8.46	8.41	8.33	8.32
5	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62	6.52	6.33	6.23	6.18	6.10	6.08
6	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46	5.37	5.17	5.07	5.01	4.93	4.92
7	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76	4.67	4.47	4.36	4.31	4.23	4.21
8	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30	4.20	4.00	3.89	3.84	3.76	3.74
9	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96	3.87	3.67	3.56	3.51	3.42	3.40
10	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72	3.62	3.42	3.31	3.26	3.17	3.15
11	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59	3.53	3.43	3.23	3.12	3.06	2.97	2.96
12	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37	3.28	3.07	2.96	2.91	2.82	2.80
13	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31	3.25	3.15	2.95	2.84	2.78	2.69	2.67
14	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21	3.15	3.05	2.84	2.73	2.67	2.58	2.56
15	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06	2.96	2.76	2.64	2.59	2.49	2.47
16	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	3.05	2.99	2.89	2.68	2.57	2.51	2.42	2.40
17	6.04	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.98	2.92	2.82	2.62	2.50	2.44	2.35	2.33
18	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93	2.87	2.77	2.56	2.44	2.38	2.29	2.27
19	5.92	4.51	3.90	3.56	3.33	3.17	3.05	2.96	2.88	2.82	2.72	2.51	2.39	2.33	2.24	2.22
20	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77	2.68	2.46	2.35	2.29	2.19	2.17
21	5.83	4.42	3.82	3.48	3.25	3.09	2.97	2.87	2.80	2.73	2.64	2.42	2.31	2.25	2.15	2.13
22	5.79	4.38	3.78	3.44	3.22	3.05	2.93	2.84	2.76	2.70	2.60	2.39	2.27	2.21	2.11	2.09
23	5.75	4.35	3.75	3.41	3.18	3.02	2.90	2.81	2.73	2.67	2.57	2.36	2.24	2.18	2.08	2.06
24	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70	2.64	2.54	2.33	2.21	2.15	2.05	2.02
25	5.69	4.29	3.69	3.35	3.13	2.97	2.85	2.75	2.68	2.61	2.51	2.30	2.18	2.12	2.02	2.00
26	5.66	4.27	3.67	3.33	3.10	2.94	2.82	2.73	2.65	2.59	2.49	2.28	2.16	2.09	1.99	1.97
27	5.63	4.24	3.65	3.31	3.08	2.92	2.80	2.71	2.63	2.57	2.47	2.25	2.13	2.07	1.97	1.94
28	5.61	4.22	3.63	3.29	3.06	2.90	2.78	2.69	2.61	2.55	2.45	2.23	2.11	2.05	1.94	1.92
29	5.59	4.20	3.61	3.27	3.04	2.88	2.76	2.67	2.59	2.53	2.43	2.21	2.09	2.03	1.92	1.90
30	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	2.51	2.41	2.20	2.07	2.01	1.90	1.88
40	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45	2.39	2.29	2.07	1.94	1.88	1.76	1.74
60	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33	2.27	2.17	1.94	1.82	1.74	1.63	1.60
80	5.22	3.86	3.28	2.95	2.73	2.57	2.45	2.35	2.28	2.21	2.11	1.88	1.75	1.68	1.55	1.53
100	5.18	3.83	3.25	2.92	2.70	2.54	2.42	2.32	2.24	2.18	2.08	1.85	1.71	1.64	1.51	1.48
200	5.10	3.76	3.18	2.85	2.63	2.47	2.35	2.26	2.18	2.11	2.01	1.78	1.64	1.56	1.42	1.39

TEST DE SHAPIRO WILK

Les tables ci-dessous, fournissent pour la première d'entre elles les valeurs des coefficients a_i de SHAPIRO WILK en fonction de la taille n de l'échantillon et du rang i considéré, et pour la table inférieure, les valeurs du seuil critique W_α en fonction de l'erreur de première espèce α et de n.

i	n=5	n=10	n=15	n=20	n=25	n=30	n=35	n=40	n=45	n=50
1	0.6646	0.5739	0.5150	0.4734	0.4450	0.4254	0.4096	0.3964	0.3850	0.3751
2	0.2413	0.3291	0.3306	0.3211	0.3069	0.2944	0.2836	0.2737	0.2651	0.2574
3	0.0000	0.2141	0.2495	0.2565	0.2543	0.2487	0.2427	0.2368	0.2313	0.2260
4		0.1224	0.1878	0.2085	0.2148	0.2148	0.2127	0.2098	0.2065	0.2032
5		0.0399	0.1353	0.1686	0.1822	0.1870	0.1883	0.1878	0.1865	0.1847
6			0.0880	0.1334	0.1539	0.1630	0.1673	0.1691	0.1695	0.1691
7			0.0433	0.1013	0.1283	0.1415	0.1487	0.1526	0.1545	0.1554
8			0.0000	0.0711	0.1046	0.1219	0.1317	0.1376	0.1410	0.1430
9				0.0422	0.0823	0.1036	0.1160	0.1237	0.1286	0.1317
10				0.0140	0.0610	0.0862	0.1013	0.1108	0.1170	0.1212
11					0.0403	0.0697	0.0873	0.0986	0.1062	0.1113
12					0.0200	0.0537	0.0739	0.0870	0.0959	0.1020
13					0.0000	0.0381	0.0610	0.0759	0.0860	0.0932
14						0.0227	0.0484	0.0651	0.0765	0.0846
15						0.0076	0.0361	0.0546	0.0673	0.0764
16							0.0239	0.0444	0.0584	0.0685
17							0.0119	0.0343	0.0497	0.0608
18							0.0000	0.0244	0.0412	0.0532
19								0.0146	0.0328	0.0459
20								0.0049	0.0245	0.0386
21									0.0163	0.0314
22									0.0081	0.0244
23									0.0000	0.0174
24										0.0104
25										0.0035

α	n=5	n=10	n=15	n=20	n=25	n=30	n=35	n=40	n=45	n=50
0.01	0.686	0.781	0.835	0.868	0.888	0.900	0.910	0.919	0.926	0.930
0.05	0.762	0.842	0.881	0.905	0.918	0.927	0.934	0.940	0.945	0.947
0.10	0.806	0.869	0.901	0.920	0.931	0.939	0.944	0.949	0.953	0.955

TEST BINOMIAL

La table fournit relativement à la variable binomiale X : $B(n, 1/2)$ et pour les petites valeurs de n, la valeur de la probabilité $\Pr ob(X \leq x)$ (il est précisé que 031 se lit 0,031 et que « * » équivaut approximativement à la valeur 1,0.

$x \backslash n$	0	1	2	3	4	5	6	7	8	9	10	11
5	031	188	500	812	969	*						
6	016	109	344	656	891	984	*					
7	008	062	227	500	773	938	992	*				
8	004	035	145	363	637	855	965	996	*			
9	002	020	090	254	500	746	910	980	998	*		
10	001	011	055	172	377	623	828	945	989	999	*	
11	000	006	033	113	274	500	726	887	967	994	*	*
12	000	003	019	073	194	387	613	806	927	981	997	*

TEST DE WILCOXON MANN WHITNEY (échantillons indépendants)

Pour des échantillons de tailles respectives n_1 et n_2 (avec $n_1 \leq n_2$), la table ci-dessous, fournit, relativement au test de comparaison de WILCOXON MANN WHITNEY, les valeurs critiques W_α et W'_α qui correspondent aux tests unilatéraux respectivement à gauche (région critique : $W \leq W_\alpha$) et à droite (région critique : $W \geq W_\alpha$), les valeurs considérées de α, erreur de première espèce, étant de 1% et 5%.

N1	1				2				3				4				5				6				7				8			
	5%		1%		5%		1%		5%		1%		5%		1%		5%		1%		5%		1%		5%		1%		5%		1%	
N2	WS	W'S	WS	W'S	WS	W'S	WS	W'S	WS	W'S	WS	W'S	WS	W'S	WS	W'S	WS	W'S	WS	W'S	WS	W'S	WS	W'S	WS	W'S	WS	W'S	WS	W'S	WS	W'S
3	--	--	--	--	--	--	--	--	6	15	--	--																				
4	--	--	--	--	--	--	--	--	6	18	--	--	11	25	--	--																
5	--	--	--	--	3	13	--	--	7	20	--	--	12	28	10	30	19	36	16	39												
6	--	--	--	--	3	15	--	--	8	22	--	--	13	31	11	33	20	40	17	43	28	50	24	54								
7	--	--	--	--	3	17	--	--	8	25	6	27	14	34	11	37	21	44	18	47	29	55	25	59	39	66	34	71				
8	--	--	--	--	4	18	--	--	8	27	6	30	15	37	12	40	23	47	19	51	31	59	27	63	41	71	35	77	51	85	45	91
9	--	--	--	--	4	20	--	--	10	29	7	32	16	40	13	43	24	51	20	55	33	63	28	68	43	76	37	82	54	90	47	97
10	--	--	--	--	4	22	--	--	10	32	7	35	17	43	13	47	26	54	21	59	35	67	29	73	45	81	39	87	56	96	49	103
11	--	--	--	--	4	24	--	--	11	34	7	38	18	46	14	50	27	58	22	63	37	71	30	78	47	86	40	93	59	101	51	109
12	--	--	--	--	5	25	--	--	11	37	8	40	19	49	15	53	28	62	23	67	38	76	32	82	49	91	42	98	62	106	53	115
13	--	--	--	--	5	27	3	29	12	39	8	43	20	52	15	57	30	65	24	71	40	80	33	87	52	95	44	103	64	112	56	120
14	--	--	--	--	6	28	3	31	13	41	8	46	21	55	16	60	31	69	26	75	42	84	34	92	54	100	45	109	67	117	58	126
15	--	--	--	--	6	30	3	33	13	44	9	48	22	58	17	63	33	72	26	79	44	88	36	96	58	105	47	114	69	123	60	132
16	--	--	--	--	6	32	3	35	14	46	9	51	24	60	17	67	34	76	27	82	46	92	37	101	58	110	49	119	72	128	62	138
17	--	--	--	--	6	34	3	37	15	48	10	53	25	63	18	70	35	80	28	87	47	97	39	105	61	114	51	124	75	133	64	144
18	--	--	--	--	7	35	3	39	15	51	10	56	26	66	19	73	37	83	29	91	49	101	40	110	63	119	52	130	77	139	66	150
19	1	20	--	--	7	37	4	40	16	53	10	59	27	69	19	77	38	87	30	95	51	105	41	115	65	124	54	135	80	144	68	156
20	1	21	--	--	7	39	4	42	17	55	11	61	28	72	20	80	40	90	31	99	53	109	43	119	67	129	56	140	83	149	70	162
21	1	22	--	--	8	40	4	44	17	58	11	64	29	75	21	83	41	94	32	103	55	113	44	124	69	134	58	145	85	155	72	168
22	1	23	--	--	8	42	4	46	18	60	12	66	30	78	21	87	43	97	33	107	57	117	45	129	72	138	59	151	88	160	74	174
23	1	24	--	--	8	44	4	48	19	62	12	69	31	81	22	90	44	101	24	121	58	122	47	133	74	143	61	156	90	166	76	180
24	1	25	--	--	9	45	4	50	19	65	12	72	32	84	23	93	45	105	25	125	60	126	48	138	76	148	63	161	93	171	78	186
25	1	26	--	--	9	47	4	52	20	67	13	74	33	87	23	97	47	106	26	129	62	130	50	142	78	153	64	167	96	178	81	191

N1	9				10				11				12				13				14				15			
	5%		1%		5%		1%		5%		1%		5%		1%		5%		1%		5%		1%		5%		1%	
N2	WS	W'S	WS	W'S	WS	W'S	WS	W'S	WS	W'S	WS	W'S	WS	W'S	WS	W'S	WS	W'S	WS	W'S	WS	W'S	WS	W'S	WS	W'S	WS	W'S
9	65	105	59	112																								
10	69	111	61	119	82	128	74	136																				
11	72	117	63	126	86	134	77	143	100	153	91	162																
12	75	123	66	132	89	141	78	151	104	160	94	170	120	180	109	191												
13	78	129	68	139	92	148	82	158	108	167	97	178	125	187	113	199	142	209	130	221								
14	81	135	71	145	95	154	85	166	112	174	100	186	129	195	116	208	147	217	134	230	155	240	152	254				
15	84	141	73	152	99	161	88	172	116	181	103	194	133	203	120	216	152	225	138	239	171	249	156	264	192	273	176	288
16	87	147	76	158	103	167	91	178	120	188	107	201	138	210	124	224	158	234	142	248	178	258	161	273	197	283	181	299
17	90	153	78	165	106	174	93	187	123	196	110	209	142	216	127	233	161	242	146	257	182	266	165	283	203	292	186	309
18	93	159	81	171	110	180	96	194	127	208	113	217	146	226	131	241	166	250	150	266	187	275	170	292	208	302	190	320
19	96	165	83	178	113	187	99	201	131	210	116	225	150	234	134	250	171	292	154	275	192	284	174	302	214	311	195	330
20	99	171	85	185	117	193	102	208	135	217	119	233	155	241	138	258	175	267	158	284	197	293	178	312	220	320	200	340
21	102	177	88	191	120	200	105	215	139	224	123	240	159	249	142	266	180	275	162	293	202	302	183	321	225	330	205	350
22	105	183	90	198	126	207	108	222	143	231	126	248	163	257	145	275	185	289	166	302	207	311	187	331	231	339	210	360
23	108	189	93	204	127	213	110	230	147	238	129	256	168	264	149	283	189	292	170	311	212	320	192	340	235	349	214	371
24	111	195	95	211	130	220	113	237	151	245	132	264	172	272	153	291	194	300	174	320	218	328	196	350	242	358	219	381
25	114	201	99	217	134	226	116	244	155	252	136	271	178	280	156	300	199	308	178	329	223	337	200	360	248	367	224	391

TEST DES RANGS SIGNES DE WILCOXON (échantillons appariés)

La table ci-dessous, fournit pour la statistique P égale à la somme des rangs signés positivement, le seuil T_α qui, en fonction de n et de α, vérifie $Prob(P \geq T_\alpha) = \alpha$.

n	5	6	7	8	9	10	11	12	13	14	15	16
$\alpha = 2{,}5\%$	15	20	25	32	39	46	55	64	73	83	94	106
$\alpha = 5\%$	14	18	24	30	36	44	52	60	69	79	89	100
n	17	18	19	20	21	22	23	24	25	26	27	28
$\alpha = 2{,}5\%$	118	130	143	157	172	187	202	218	235	252	270	289
$\alpha = 5\%$	111	123	136	149	163	177	192	208	224	240	258	275
n	29	30	31	32	33	34	35	36	37	38	39	40
$\alpha = 2{,}5\%$	308	327	348	368	390	412	434	457	481	505	530	555
$\alpha = 5\%$	294	313	332	352	373	394	416	438	461	484	508	533

TEST DE KOLMOGOROV (pour un échantillon)

Relativement au test bilatéral et notant par D la statistique de KOLMOGOROV, la table fournit en fonction de la taille n de l'échantillon et de l'erreur de première espèce choisie, soit α, la valeur du seuil $D_{\alpha,n}$ vérifiant $\Pr ob(D \geq D_{\alpha,n}) = \alpha$.

Taille de l'échantillon	Risque de première espèce (α)				
	0,20	0,15	0,10	0,05	0,01
1	0,900	0,925	0,950	0,975	0,995
2	0,684	0,726	0,776	0,842	0,929
3	0,565	0,597	0,642	0,708	0,828
4	0,494	0,525	0,564	0,624	0,733
5	0,446	0,474	0,510	0,565	0,669
6	0,410	0,436	0,470	0,521	0,618
7	0,381	0,405	0,438	0,486	0,577
8	0,358	0,381	0,411	0,457	0,543
9	0,339	0,360	0,388	0,432	0,514
10	0,322	0,342	0,368	0,410	0,490
11	0,307	0,326	0,352	0,391	0,468
12	0,295	0,313	0,338	0,375	0,450
13	0,284	0,302	0,325	0,361	0,433
14	0,274	0,292	0,314	0,349	0,418
15	0,266	0,283	0,304	0,338	0,404
16	0,258	0,274	0,295	0,328	0,392
17	0,250	0,266	0,286	0,318	0,381
18	0,244	0,259	0,278	0,309	0,371
19	0,237	0,252	0,272	0,301	0,363
20	0,231	0,246	0,264	0,294	0,356
25	0,21	0,22	0,24	0,27	0,32
30	0,19	0,20	0,22	0,24	0,29
35	0,18	0,19	0,21	0,23	0,27
supérieur à 35	$1,07/\sqrt{n}$	$1,14/\sqrt{n}$	$1,22/\sqrt{n}$	$1,36/\sqrt{n}$	$1,63/\sqrt{n}$

TEST DE FRIEDMAN

Dans le cadre de la comparaison de K traitements à partir d'un échantillon de taille n (comparaison de K échantillons appariés), la table des valeurs ci-dessous fournit pour la statistique de FRIEDMAN et en fonction de n, K, et α (l'erreur de première espèce choisie), le seuil F_α vérifiant $\Pr ob(F \geq F_\alpha) = \alpha$.

	$k=3$		$k=4$		$k=5$		$k=6$	
n	0,05	0,01	0,05	0,01	0,05	0,01	0,05	0,01
2	-	-	6,000	-	7,600	8,000	9,143	9,714
3	6,000	-	7,400	9,000	8,533	10,13	9,857	11,76
4	6,500	8,000	7,800	9,600	8,800	11,20	10,29	12,71
5	6,400	8,400	7,800	9,960	8,960	11,68	10,49	13,23
6	7,000	9,000	7,600	10,20	9,067	11,87	10,57	13,62
7	7,143	8,857	7,800	10,54	9,143	12,11		
8	6,250	9,000	7,650	10,50	9,200	12,30		
9	6,222	9,556	7,667	10,73	9,244	12,44		
10	6,200	9,600	7,680	10,68				
11	6,545	9,455	7,691	10,75				
12	6,500	9,500	7,700	10,80				
13	6,615	9,385	7,800	10,85				
14	6,143	9,143	7,714	10,89				

TEST DE KOLMOGOROV-SMIRNOV (pour deux échantillons)

Pour le ***test bilatéral*** et dans le cas de petits échantillons (n_X et n_Y inférieurs à 15), la table ci-dessous fournit, relativement à la statistique D de KOLMOGOROV-SMIRNOV, la valeur du seuil D_{α,n_X,n_Y} vérifiant $\Pr ob(D \geq D_{\alpha,n}) = \alpha$, ceci en fonction de n_X et n_Y, et pour les valeurs 1% et 5% de l'erreur de première espèce α .Ce qui concerne le seuil 5% est à lire au dessus de la diagonale, les valeurs qui correspondent au seuil 1% étant quant à elles situées en dessous de ladite diagonale.

$n_X \setminus n_Y$

Attention !, lire, par exemple, 0,778 pour la valeur « 778 » et 1 pour la valeur 1000.

	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
5		800	800	750	778	800	709	717	692	657	733	800	647	667	642	650
6	1000		714	708	722	667	652	667	667	643	533	625	667	667	614	600
7	1000	857		714	667	657	623	631	615	643	590	571	571	571	571	564
8	875	833	857		639	600	602	625	596	571	558	625	566	611	539	550
9	889	883	778	764		589	596	583	556	556	556	542	536	556	520	517
10	900	800	757	750	700		545	550	569	529	533	525	524	511	495	550
11	818	818	766	727	707	700		545	524	532	509	506	497	490	488	486
12	833	833	714	708	694	667	652		519	512	517	500	490	500	474	483
13	800	769	714	692	667	646	636	608		489	492	486	475	470	462	462
14	800	762	786	679	667	643	623	619	571		467	473	467	460	455	450
15	800	767	714	675	667	667	618	600	590	586		475	455	456	446	450
16	800	750	688	688	653	625	602	604	582	563	554		456	444	437	437
17	800	716	706	647	647	624	588	583	576	563	557	526		435	437	429
18	788	778	690	653	667	600	596	583	585	556	544	535	536		415	422
19	747	728	684	645	626	595	584	570	559	556	533	526	514	515		421
20	800	733	664	650	617	650	577	583	550	543	533	525	515	506	492	

Une table semblable est présentée ci-dessous, pour le ***test unilatéral***.

	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
5		800	714	675	667	700	636	600	615	600	667	600	588	578	589	600
6	1000		667	625	611	600	576	667	590	571	567	563	649	511	561	550
7	857	833		607	571	571	571	547	549	571	533	526	513	516	519	514
8	875	833	750		556	550	545	500	519	517	500	563	500	500	487	500
9	800	778	746	750		556	525	528	504	500	511	479	484	500	468	467
10	800	733	714	700	678		518	500	492	486	500	475	465	456	447	500
11	800	742	714	693	636	627		485	469	474	461	455	455	444	440	436
12	800	750	690	667	639	617	583		455	464	467	458	441	444	434	433
13	769	692	692	644	624	600	601	590		428	446	438	434	423	421	415
14	729	714	714	643	635	600	584	560	560		438	428	420	413	414	407
15	800	700	667	625	622	600	576	567	574	529		421	412	411	400	417
16	738	688	652	688	604	588	568	562	538	536	500		401	403	395	400
17	741	667	647	625	601	582	556	549	534	525	514	511		386	390	383
18	722	722	659	611	611	578	545	556	526	519	511	493	490		389	378
19	737	675	647	612	579	547	545	535	526	508	498	497	489	468		379
20	750	667	650	625	578	600	536	533	519	507	500	488	479	472	450	

BIBLIOGRAPHIE

BAR-HEN Avner – Cours de statistique, université d'Aix-marseille (2007)

Commissariat à l'Energie Atomique – Statistique appliquée à l'exploitation des mesures, éditions MASSON (1978)

DELMAS Jean-François – Introduction au calcul des probabilités et à la statistique « exercices » (2008)

DUPUIS Jérôme – Cours de statistique inférentielle, université Paul Sabatier de Toulouse (2007)

GABBUD Christian – Association AGLA (graduées et gradués de l'Université de Lausanne et sciences actuarielles), bulletin n°30 (2006)

GAUDOIN Olivier – Méthodes statistiques pour l'ingénieur , ENSIMAG-INP de Grenoble (2007)

GENEST Christian – Tests d'indépendance, université Laval à Québec (2008)

FROMONT Magalie – Eléments de la théorie des tests, licence MASS (2006)

LACÔTE Guillaume – Estimation et introduction aux tests, ENSAE (2003)

LEGENDRE Pierre – Liaisons entre deux variables, université de Montréal (2007)

MOUCHIROUD Dominique – Outils pour la biologie, DEUG (2003)

MURRAY et SPIEGEL – Probabilités et statistiques, éditions SCHAUM (1981)

N'GUYEN Jean-Michel – Comparaison de proportions, faculté de médecine de Nantes (2008)

PALM Rudy – Utilisation du bootstrap pour les problèmes statistiques liés à l'estimation des paramètres (2002)

PARREINS Georges – Techniques statistiques, éditions DUNOD Technique (1974)

PHAN Thérèse et ROWENCZYK Jean-Pierre – Statistiques et probabilités, éditions DUNOD (2007)

PHILIPPE Anne – Exercices de statistique, université de Nantes (2006)

PHILIPPE Paulette – Exercices de statistique, université Balise pascal à Clermont-Ferrand (2008)

RAKOTOMALALA – Comparaison de populations, tests non paramétriques, université Louis Lumière Lyon 2 (2008)

SOLIANI Lamberto – Tests statistiques non paramétriques, éditions APRILE (2008)

TUFFERLY Stéphane – Data mining et statistique décisionnelle (intelligence des données), éditions TECHNIP (2007)

VINATIER Stéphane – Licence de biologie, faculté des sciences et techniques de Limoges (2007)

WILQUET Gaston – Techniques de la physique expérimentale (2007)

INDEX ALPHABETIQUE

M

N

O

P

Q

R

S

U

V

W

Y

Dans la COLLECTION TECHNOSUP :

Niveau **A : Approche** Niveau **B : Bases** Niveau **C : Compléments**

calcul scientifique :
- *VBA pour Excel. Bibliothèque mathématique avec applications pratiques* 600 p. (C) — D. ROUX
- Faire des maths avec *Mathematica*. Initiation, thèmes d'étude 160 p. (B) — N. VERDIER
- Eléments d'analyse. Calcul différentiel et intégral 240 p. (A) — B. RADI, A. EL HAMI
- Mathématiques des sciences appliquées 216 p. (B) — Ph. GOLDNER
- Suites et séries. Cours et exercices corrigés 192 p. (B) — A-J. RIQUET
- Analyse harmonique. Cours et exercices 192 p. (B) — B. ROSSETTO
- Calcul scientifique avec *Matlab* 288 p. (C) — J. KOKO
- Modélisation et analyse des systèmes linéaires 224 p. (C) — J.F. MASSIEU, Ph. DORLÉANS
- Tenseurs, variations et milieux continus 288 p. (C) — J.-F. GANGHOFFER

mesure :
- Mesure physique et instrumentation 192 p. (A) — D. BARCHIESI
- Capteurs électrochimiques. Fonctionnement, utilisation, conception 288 p. (C) — C. GONDRAN, P. FABRY
- Systèmes d'acquisition de données 240 p. (B) — E. ETIEN
- Traitement des mesures. Interprétation, modélisation 384 p. (B) — R. JOURNEAUX
- Traitement statistique du signal 216 p. (C) — M. BARRET
- Analyse spectrale Cours Supélec 192 p. (C) — G. FLEURY

probabilités :
- Probabilités pour modéliser et décider 256 p. (A) — N. SAVY
- Calcul des probabilités 224 p. (B) — J-P. BOULAY
- Modélisation probabiliste pour l'ingénieur 312 p. (C) — A. SMOLARZ

statistiques :
- Statistique *sans mathématique* 224 p. (A) — J. BADIA, R. BASTIDA, J.R. HAÏT
- Assimiler et utiliser les statistiques 288 p. (A) — L. PIBOULEAU
- Statistiques et expérimentation en biologie 192 p. (A) — J-Cl. LABERCHE
- Statistique mathématique 360 p. (B) — J-P. BOULAY

mécanique quantique ::
- *La mécanique quantique et ses applications Cours Supélec* 224 p. (C) — A. et M.-F. CHARLIER

ondes, corpuscules :
- Ondes et matière 352 p. (C) — D. BARCHIESI, M. LAMY DE LA CHAPELLE
- Physique pour l'électronique. Corpuscule, onde, état quantique, structures 320 p. ((B) — A. et D. DEVILLE
- Eléments de propagation électromagnétique 160 p. (B) — Ph. ROSNET
- *Techniques micro-ondes. Dispositifs passifs et tubes micro-ondes* Cours Supélec 320 p. (C) — M. HÉLIER
- Optoélectronique. Composants photoniques et fibres optiques Cours Supélec 320 p. (C) — Z. TOFFANO
- Le radar. Théorie et pratique Cours Supélec 160 p. (C) — J.-M. COLIN

optique :
- Optique physique. Interférences, diffraction, holographie. Cours et exercices 192 p. (A) — Fl. WEIL
- *Optique moderne. Polarisation, lasers, fibres optiques.* Cours et exercices 224 p. (B) — Fl. WEIL
- Exercices corrigés d'optique. Optique instrumentale. Optique de Fourier 192 p. (B) — J. SURREL

acoustique :
- Biophysique de l'environnement sonore 192 p. (B) — C. GELIS
- Acoustique générale 352 p. (C) — C. POTEL, M. BRUNEAU

géophysique :
- Murmures ionosphériques. Techniques de réception sous le seuil de 100 kHz 216 p. (C) — J-J. DELCOURT

géologie :
- Pétrologie sédimentaire 264 p. (B) — F. BOULVAIN

environnement :
- Les traitements de l'eau. Cours et problèmes résolus 256 p. (B) — Cl. CARDOT
- Techniques appliquées au traitement de l'eau 256 p. (B) — (coord.) Cl. CARDOT
- Pollution atmosphérique. Causes, conséquences, solutions, perspectives 224 p. (B) — P. MASCLET
- La lutte biologique. Application aux arthropodes et adventices. 320 p. (B) — B. PINTUREAU
- Biodégradation des matériaux. (288 p. (C) — A. CORNET, F. FEEUGEAS, B. TRIBOULET
- Gestion des déchets. Réglementation, organisation, mise en œuvre 224 p. (A) — *Th. ROGAUME*
- Traitement des déchets. Valorisation, élimination. 288 p. (A) — A. ADDOU

chimie :
- Révisions et autoévaluation en chimie structurale. 240 p. (A) — C. WATERLOT
- Comprendre la chimie organique 224 p. (A) — A. LASSALLE, D. ROBERT
- Assimiler la chimie organique 192 p. (A) — A. LASSALLE, D. ROBERT
- La chimie en IUT et BTS. Cours et exercices résolus 224 p. (A) — F. VAISSIAUX
- Chimie des solutions. Résumés de cours et exercices corrigés 224 p. (B) — P.-L. FABRE

génie chimique :
- Thermodynamique et cinétique chimique. Résumés de cours et exercices 224 p. (B) — P.-L. FABRE
- Cinétique et catalyse hétérogènes 320 p. (C) — B. GILOT, R. GUIRAUD
- Réactions et réacteurs chimiques 288 p. (B) — M. GUISNET, S. LAFORGE, D. COUTON
- Génie chimique. Les opérations unitaires 312 p. (C) — D. MORVAN

analyse physico-chimique :
- Les techniques de laboratoire 160 p. (A) — E. BOURGUET, C. AUGÉ
- Spectrométrie de masse 320 p. (C) — G. DUGUAY
- Séparation et analyse des biomolécules. Méthodes physico-chimiques 256 p. (B) — J.-P. SINE

mécanique :
- Mécanique générale. Cours, exercices et problèmes corrigés 288 p. (B) — Cl. CHÈZE, H. LANGE

- *Actions mécaniques. Statique. Inertie. De la théorie aux applications* 224 p. (A) — Cl. CHÈZE, F. BRONSARD
- Vibrations des structures 224 p. (B) — G. VENIZELOS

milieux continus :
- Mécanique des milieux déformables 288 p. (B) — M. FOURAR, Cl. CHÈZE
- *Théorie microscopique des liquides 480 p.* (C) — J.-L. BRETONNET
- Les écoulements de fluides newtoniens 384 p. (C) — J.-N. GENCE

génie mécanique
- Détermination des éléments de machines 360 p. — A. BOURDON, L. MANIN, D. PLAY
- *Conception et construction des moteurs alternatifs 288 p. (C)* — Ph. ARQUÈS
- Transmissions mécaniques de puissance. Boites de vitesses automatiques 288 p. (C) — Ph. ARQUÈS
- Ingénierie des turbomachines. Cours et exercices résolus 288 p. (C) — M. PLUVIOSE

structures :
- *Dimensionnement des structures. Résistance des matériaux 224 p. (B)* — D. CHÈZE
- Mécanique des structures. Du calcul analytique au calcul matriciel 288 p. (B) — J.-Ch. CRAVEUR, Cl. CHÈZE
- Vibrations des structures pour l'ingénieur et le technicien 264 p. (C) — B. COMBES
- Comprendre les éléments finis. Structures 288 p. (C) — A. CHATEAUNEUF

***thermodynamique* :**
- La thermodynamique des principes aux applications 312 p. (C) — Cl. CHEZE, P. BAUER
- Réactions thermiques en phase gazeuse. Modélisation 288 p. (C) — G.-M. CÔME
- La méthode modale en thermique 320 p. (C) — G. LEFEBVRE
- Transferts thermiques, application à l'habitat. Méthode nodale 224 p. (C) — H. CORTES, J. BLOT
- Combustion. Inflammation, combustion, pollution, applications 320 p. (C) — Ph. ARQUÈS

énergétique :
- Moteurs alternatifs à combustion interne. De la théorie à la compétition 288 p. (B) — Ph. ARQUÈS
- Machines à fluides. Principes et fonctionnement 288 p. (C) — M. PLUVIOSE
- Conversion d'énergie par turbomachines 288 p. (C) — M. PLUVIOSE
- Propulseurs aéronautiques et spatiaux 288 p. (C) — P. BAUER
- Piles à combustible. Principes, modélisation, applications 192 p. (B) — B. BLUNIER, A. MIRAOUI
- Énergie solaire. Calculs et optimisation 320 p. (B) — J. BERNARD
- Énergie nucléaire 1. De la théorie aux applications 256 p. (B) — J. BERNARD
- Énergie nucléaire 2. Les réacteurs nucléaires électrogènes 288 p. (B) — J. BERNARD

science des matériaux :
- Introduction à la cristallochimie. Solides cristallisés et empilements compacts 160 p. (A) — D. RIOU
- Propriétés et comportements des matériaux 320 p. (B) — A. CORNET, F. HLAWKA
- Métallurgie mécanique 320 p. (B) — A. CORNET, F. HLAWKA
- Fatigue des structures. Endurance, rupture, critères, contrôle, durabilité 320 p. (C) — G. HENAFF, F. MOREL
- Endommagement interfacial des métaux 256 p. (C) — G. SAINDRENAN, R. LE GALL, F. CHRISTIEN
- Cycle de vie des surfaces industrielles (312 p. (C) — F. HLAWKA, A. CORNET

génie civil :
- Béton armé. Application de l'eurocode 2 224 p. (B) — R. NICOT
- Analyse et dimensionnement sismiques 224 p. (B) — P. LESTUZZI

productique :
- Méthodes, productique et qualité 224 p. (B) — J.-M. CHATELET
- Maintenance industrielle 288p. (B) — J.M. AUBERVILLE
- Analyse et maintenance des automatismes industriels 192 p. (B) — A. REILLER
- TGAO La technologie de groupe 288 p. (C) — A. NADIF

génie industriel :
- Maîtriser la conduite de projet 192 p. (A) — C. ALONSO
- Management de projet technique 192 p. (B) — C. CAZAUBON, G. GRAMACIA, G. MASSARD
- Organisation et génie de production 224 p. (B) — F. LAMBERSEND
- *Méthode d'aide à la décision. Approche théorique et études de cas* 192 p. (B) — R. LABBE

électricité générale :
- Vade-mecum d'électrotechnique 312 p. (A) — C. LE TRIONNAIRE, J.-P. PICHENY
- Circuits électriques, Régimes continu, sinusoïdal et impulsionnel 192 p. (A) — J.-P. BANCAREL
- *Les lois de l'électricité 288 p.* (A) — M. PIOU
- Annales d'électrotechnique BTS Maintenance 256 p. (A) — D. VINCENT, N. ORTEGA
- Annales d'électrotechnique BTS Maintenance 1997/2008 224 p. (A) — D. VINCENT
- T.P. d'électrotechnique par simulation avec PSIMDEMO 224 p. (A) — F. LEPLUS

***machines électriques* :**
- Moteurs à courant alternatif 288 p. (A) — D. JACOB
- Le moteur asynchrone. Régimes statique et dynamique 160 p. (C) — L. MUTREL
- Machines à courant alternatif 240 p. (B) — D. NAMANE
- Modélisation et commande des moteurs triphasés 256 p. (C) — G. STURTZER, E. SMIGIEL
- Électrotechnique. Machines et réseaux Cours Supélec 256 p. (C) — J.-P. FANTON
- Machines électriques. Théorie et Mise en oeuvre Cours Supélec 256 p. (C) — Ph. BARRET
- Machines électriques tournantes 384 p. (C) — B. LAPORTE

électronique de puissance :
- Électronique de puissance, Principes, fonctionnement, dimensionnement 256 p. (A) — D. JACOB
- Les redresseurs, Redresseurs à diodes, à thyristors et mixtes 336 p. (B) — J. MIGNARD, C. PIN

électronique :
- Les fondamentaux en électronique 224 p. (A) — P. ROCHETTE
- Des clés pour l'électronique. Travaux dirigés illustrés par simulation 160 p. (A) — B. GIRAULT
- Les oscillateurs en électronique. Cours et exercices corrigés 160 p. (B) — G. COUTURIER
- L'outil graphique en électronique et automatique 224 p. (B) — J. BAILLOU, G. CHAUVAT, C. PEJOT

- Modulation d'amplitude. Cours et exercices 352 p. (C) — F. BIQUARD
- Amplificateurs fondamentaux et opérationnels. Cours et exercices corrigés 352 p. (B) — A. LANTZ
- Electronique radiofréquence Cours Supélec 256 p. (C) — A. PACAUD
- Electronique analogique rapide 216 p. (C) — A. FABRE
- Circuits spécialisés de l'électronique actuelle 320 p. (C) — A. et D. DEVILLE

semi-conducteurs :
- Composants à semiconducteurs Cours Supélec 256 p. (C) — O. BONNAUD, A. PACAUD
- Technologie mirco-électronique. Du silicium aux circuits intégrés 160 p. (B) — O. BONNAUD
- Détecteurs à semi-conducteurs, Principes et matériaux 224 p. (C) — J-P. PONPON
- Les semiconducteurs de puissance Cours Supélec 256 p. (C) — P. ALOÏSI

électronique numérique :
- Électronique numérique. Fonctionnement, modélisation, circuits intégrés 320 p. (B) — A. et D. DEVILLE
- Circuits intégrés numériques. Du transistor au microprocesseur 224 p. (A) — A.-Riadh BABA-ALI

traitement du signal :
- Signaux et systèmes continus et échantillonnés 192 p. (B) — M. VILLAIN
- Signaux et systèmes linéaires Cours Supélec 192 p. (B) — A. PACAUD
- Traitement du signal analogique. Cours 224 p. (A) — T. NEFFATI
- Traitement du signal analogique. Exercices et problèmes résolus 224 p. (B) — T. NEFFATI
- Ingénierie du signal. Théorie et pratique 224 p. (B) — Ph. COURMONTAGNE
- Analyse et traitement du signal. Approches pour l'ingénieur 320 p. (B) — Ph. GAILLARD, R. LENGELLE
- Théorie et pratique du signal déterministe et aléatoire, continu et discret 384 p. (B) — J.-P. TANGUY

filtrage numérique :
- Débuter en traitement numérique du signal 224 p. (A) — J.-N. MARTIN
- Analyse et contrôle numériques du signal 192 p. (B) — Ph. DESTUYNDER, F. SANTI
- Traitement numérique du signal. Théorie et applications 256 p. (C) — K. KPALMA, V. HAESE-COAT

automatique :
- Ce qu'il faut savoir sur les automatismes, Fiches-résumés 256 p. (A) — P. GRARE, I. KACEM
- Systèmes asservis linéaires 224 p. (B) — M. VILLAIN
- Asservissements linéaires continus 288 p. (B) — P. ROUSSEAU
- Commande analogique et numérique des systèmes 384 p. (B) — R. KOYS
- Systèmes et asservissements continus 320 p. (C) — É. OSTERTAG
- Régulation PID en génie électrique. Étude de cas 256 p. (C) — D. JACOB
- Problèmes résolus d'automatique 288 p. (B) — Ch. BURGAT
- Problèmes résolus d'automatique par thèmes et par types d'application 256 p. (B) — R. HUSSON

robotique :
- Traité de robotique. 1- Les architectures, conception, modélisation, équations 432 p. (C) — C. BOP
- Commande numérique des systèmes Cours Supélec 256 p. (C) — E. GODOY, É. OSTERTAG
- Commande et estimation multivariables 288 p. (C) — É. OSTERTAG
- Commande automatique des systèmes linéaires, utilisation de MATLAB 256 p. (C) — V. MINZU, B. LANG
- Commande et diagnostic des systèmes dynamiques 320 p. (C) — R. TOSCANO
- Ingénierie de la commande des systèmes 256 p. (C) — A. CROSNIER, G. ABBA, B. JOUVENCEL, R. ZAPATA

informatique industrielle :
- Bit après bit. Computers 1 320 p. (B) — J.-J. MERCIER
- Séquence après séquence. Computers 2 288 p. (B) — J.-J. MERCIER
- Instruction après instruction. Computers 3 320 p. (C) — J.-J. MERCIER
- Circuits logiques programmables. Mémoires, PLD, CPLD, FPGA 256 p. (B) — A. NKETSA
- Du binaire au processeur 320 p. (BC) — E. MESNARD
- Concevoir son microprocesseur 256 p. (B) — J.-Ch. BUISSON
- Logique combinatoire et séquentielle. Système, méthodes, réalisations 320 p. (C) — Cl. BRIE
- Architecture des systèmes sur puce 320 p. (C) — A. ATTOUI

image :
- Traitement de l'image et de la vidéo, avec exercices pratiques en Matlab et C++ 240 p. (C) — R. BELAROUSSI

communication, réseaux :
- Lignes de transmission 224 p. (B) — R. MEYS
- Transmission de l'information 192 p. (B) — Ph. FRAISSE, D. MARTY-DESSUS, R. PROTIÈRE
- Architectures des réseaux et télécommunications 192 p. (B) — P. LORENZ
- Réseaux Intranet et Internet, Architecture et mise en œuvre 336 p. (B) — J. PHILIP
- Codes correcteurs. Principes et exemples 192 p. (C) — J. BADRIKIAN

bases de données :
- Bases de données. Implémentation avec Access 256 p. (B) — J. AUBERT
- Conception méthodique des bases de données. Un guide de bonne pratique 224 p. (A) — G. BUENO
- Gradualité et imprécision dans les bases de données 320 p. (B) — P. BOSC, L. LIÉTARD, O. PIVERT, D. ROCACHER

bureautique :
- Excel pour l'ingénieur 320 p. (B) — Ph. BELLAN

processus temps réel :
- Approche du temps réel industriel 160 p. (A) — J.-M. DE GEETER
- Gestion des processus industriels temps réel 224 p. (B) — J.-J. MONTOIS

programmation :
- Algorithmes fondamentaux et langage C 320 p. (B) — J.-L. IMBERT
- Le langage C par l'exemple 320 p. (A) — Ph. ROBINET
- Du procédural à l'objet : les langages C et C++ 352 p. (B) — J. PHILIPP
- Belle programmation et langage C Cours Supélec 192 p. (C) — Y. NOYELLE
- Programmation avec le langage Python 336 p. (C) — X. DUPRE
- Compilations des langages de programmation 192 p. (C) — M. GAUTIER

génie logiciel :
- Méthode orientée objet intégrale MACAO 320 p. (B) — J.-B. CRAMPES

- La conception orientée objet, évidence ou fatalité 160 p. (B) — J.-L. CAVARERO, R. LECAT
- Conception des systèmes d'information. Méthodes et techniques 320 p. (B) — P. ANDRÉ, A. VAILLY
- Spécification des logiciels. Deux exemples : Z et UML 320 p. (C) — P. ANDRÉ, A. VAILLY
- Exercices corrigés de conception logicielle. Modélisation par la pratique 320 p. (B) — P. ANDRÉ, A. VAILLY
- Exercices corrigés d'UML. Passeport pour une maîtrise de la notation 320 p. (C) — P. ANDRÉ, A. VAILLY
- Exercices corrigés en langage Z. Spécifications formelles par l'exemple 256 p. (C) — P. ANDRÉ, A. VAILLY

ergonomie :
- Interfaces graphiques ergonomiques. Conception, modélisation 192 p. (B) — J.-B. CRAMPES
- Analyse des tâches en ergonomie 160 p. (A) — M. MOSCATO

logistique :
- Logistique interne 160 p. (A) — L. AMODEO, F. YALAOUI

sécurité :
- Sécurité des ouvrages. Risques. Géotechnique 320 p. (C) — J.-L. FAVRE
- Risques et sécurité 224 p. (C) — J.-F. GUYONNET

économie :
- Gestion financière. Analyse et politique financières de l'entreprise 256 p. (B) — A. RIVET
- Méthodes mathématiques pour les finances 384 p. (C) — J.-Ph. ARGAUD, O. DUBOIS
- Les marchés à terme agricoles 256 p. (B) — N. HABERT
- La Bourse et les produits boursiers 320 p. (B) — D. ARNOULD

législation :
- La réglementation du travail 160 p. (A) — Ph. MALINGREY
- Le travail salarié 256 p. (A) — P. IRIART
- Connaître et comprendre le droit. Principes et cas pratiques 256 p. (B) — C. GABET

éthique :
- Science, technologie et éthique 288 p. (B) — S. LAVELLE

Achevé d'imprimer sur rotative par l'Imprimerie Darantiere à Dijon-Quetigny en juin 2010
Dépôt légal : juin 2010 — N° d'impression : 10-0829
Imprimé en France

La collection TECHNOSUP dirigée par Claude Chèze est une sélection d'ouvrages dans toutes les disciplines, pour les filières technologiques des enseignements supérieurs.

Niveau A	Approche (éléments, résumés ou travaux dirigés)	IUT - BTS - 1er cycle
Niveau B	Bases (cours avec exercices et problèmes résolus)	IUP - Licence
Niveau C	Compléments (approfondissement, spécialisation)	Écoles d'ingénieurs, Master

L'ouvrage : niveau B (IUP - Licence)

S'appuyant sur le calcul des probabilités, dont les techniques usuelles ont été développées dans un précédent ouvrage de l'auteur, ce manuel présente les principales méthodes utilisées en statistique mathématique, à travers des rappels de cours et plus d'une centaine de problèmes corrigés qui les illustrent au moyen de cas concrets.

L'ouvrage s'adresse à un large public qui est celui des écoles d'ingénieurs et des I.U.T, mais aussi des écoles de commerce et des universités dans des spécialités aussi diverses que l'ingénierie, la médecine, la biologie, l'agriculture, la gestion, l'économie... Il traite une grande variété de modèles et donne un riche aperçu des techniques statistiques autour des sujets classiques, échantillonnage, estimation, décision, et régression. Il trouve ainsi un juste compromis entre la rigueur mathématique et la pratique effective.

La présentation est particulièrement fournie quant à l'estimation et la décision statistique où les tests non paramétriques trouvent une place importante.

L'auteur :
Jean-Pierre Boulay, Ingénieur de la Ville de Paris, enseigne le calcul des probabilités et la statistique à l'École Spéciale des Travaux Publics à Paris, où il assure également un cours de recherche opérationnelle.

Illustration de couverture : Dessin de Léonard de Vinci.

www.editions-ellipses.fr