Cours de Statistique
HEI 3 - 2014/2015

A. RIDARD

Table des matières

Chapitre 1

Modes d'échantillonnage et paramètres d'une population

Si la Statistique descriptive consiste en l'étude d'une population toute entière d'individus selon un ou plusieurs caractères, la Statistique inférentielle permet d'estimer ou de tester des caractéristiques d'une population de taille N à partir d'un échantillon de taille n. Avant de préciser ces caractéristiques, nous allons présenter différentes manières de prélever les échantillons.

1.1 Modes d'échantillonnage

1.1.1 Sondage aléatoire simple

Il est important de rappeler que chaque individu d'une population est caractérisé par un ou plusieurs caractères appelés aussi variables. On distingue deux types et quatre sous-types de variables.

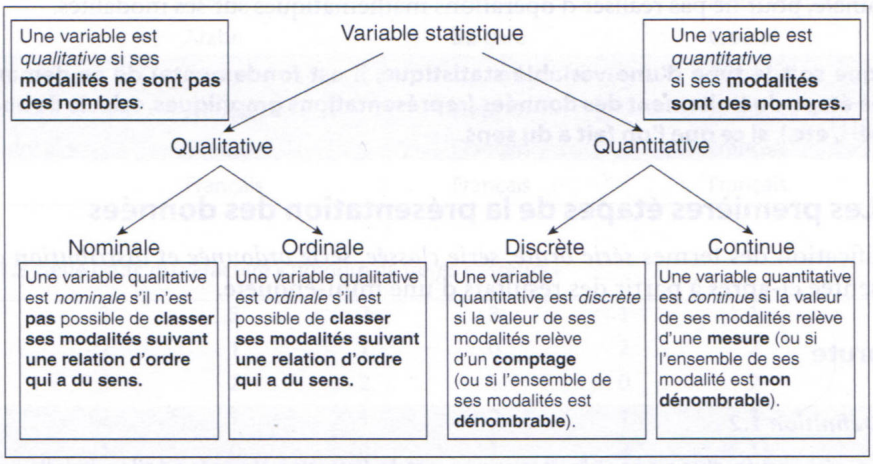


FIGURE 1.1 – Deux types et quatre sous-types de variables

Attention, une variable est une application (au sens mathématique du terme) qui, à chaque individu, associe une valeur (numérique ou non). Si, en Mathématiques, l'usage est plutôt d'appeler f, g ou h les applications, en Statistique celles-ci sont notées X, Y ou Z. Les minuscules x, y ou z représentent alors les réalisations (valeurs) de ces variables (applications). Autrement dit, si l'on note ω un individu de la population et X la variable étudiée, alors $X(\omega) = x$ signifie que le caractère X a pour valeur x pour l'individu ω. Si l'individu est choisi au hasard, la variable est dite aléatoire (va).

Un n-échantillon aléatoire est un n-uplet de variables aléatoires $(X_1, ..., X_n)$ qui, à un n-uplet d'individus choisis au hasard dans la population $(\omega_1, ..., \omega_n)$, associe le n-uplet de valeurs $(x_1, ..., x_n)$ où $x_i = X(\omega_i)$.

Remarque.
- On fera bien la différence entre majuscules et minuscules pour éviter toute confusion entre applications et valeurs.
- Une variable aléatoire possède une loi de probabilité qui régit son comportement. Si la variable est discrète, la loi est définie par un diagramme en bâtons qui, à chaque valeur possible, associe sa probabilité. Si la variable est continue, le diagramme est remplacé par une courbe de densité [1].

Un n-échantillon aléatoire $(X_1, ..., X_n)$ est dit simple si les va sont indépendantes et de même loi. Dans ce cas, la loi est celle de la va étudiée X. Cela se produit si les individus sont choisis au hasard :
- soit avec remise
- soit sans remise (ou simultanément) à condition que le taux de sondage $\frac{n}{N}$ soit inférieur à 10%.

Remarque. Le premier cas est théorique et le deuxième pratique.

1.1.2 Sondage en strates

Dans un sondage aléatoire simple, tous les échantillons d'une population de taille N sont possibles avec la même probabilité. On imagine que certains d'entre eux puissent s'avérer a priori indésirables.

Plus concrètement, dans l'étude du lancement d'un nouveau produit financier, on peut supposer des différences de comportement entre les "petits" et les "gros" clients de la banque. Il serait malencontreux que les hasards de l'échantillonnage conduisent à n'interroger que les clients appartenant à une seule de ces catégories, ou simplement que l'échantillon soit trop déséquilibré en faveur de l'une d'elles. S'il existe dans la base de sondage une information auxiliaire permettant de distinguer, a priori, les catégories de petits et gros clients, on aura tout à gagner à utiliser cette information pour répartir l'échantillon dans chaque souspopulation. C'est le principe de la stratification : découper la population en sous-ensembles appelés strates et réaliser un sondage aléatoire simple dans chacune d'elles.

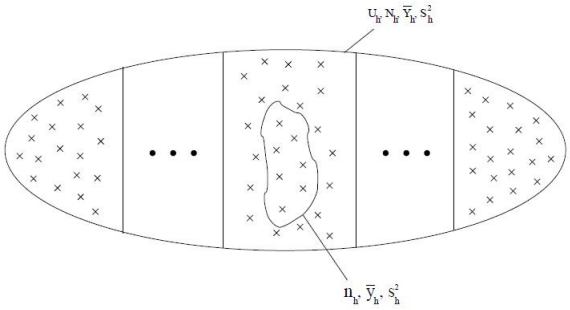


FIGURE 1.2 – Découpage en strates

Nous avons alors deux manières de choisir les n_h :
- allocation proportionnelle : $\forall h, \frac{n_h}{N_h} = \frac{n}{N}$
- allocation optimale [2] : $\forall h, \frac{n_h}{N_h} = n\frac{S_H}{\sum_h N_H S_H}$.

Exemple. Pour illustrer les différents modes d'échantillonnage, on pourra se reporter à l'annexe 1.

1. Plus précisément, la va X est (absolument) continue de densité f si pour tout intervalle I, on a :
$$P(X \in I) = \int_I f(x)\,dx$$

2. On cherche la répartition de l'échantillon qui maximise la précision (et donc qui minimise la variance). Pour cela, on va augmenter les effectifs échantillonnés dans les strates où la variabilité est grande et diminuer les effectifs échantillonnés dans les strates homogènes.

Dans toute la suite du cours, les échantillons aléatoires seront supposés simples.

1.2 Paramètres d'une population

1.2.1 Moyenne et variance d'une variable aléatoire

Soit X la va étudiée.

Nous avons déjà vu qu'une va possède une loi de probabilité régissant son comportement, elle possède aussi deux caractéristiques importantes :
– une caractéristique de tendance centrale, l'espérance ou la moyenne, définie par

$$m = E(X) = \begin{cases} \sum\limits_i x_i P(X = x_i) & \text{si } X \text{ est discrète} \\ \int_{\mathbf{R}} x f(x)\, dx & \text{si } X \text{ est continue de densité } f \end{cases}$$

– une caractéristique de dispersion, la variance, définie par :

$$\sigma^2 = V(X) = E\Big((X-m)^2\Big) = \begin{cases} \sum\limits_i (x_i - m)^2 P(X = x_i) & \text{si } X \text{ est discrète} \\ \int_{\mathbf{R}} (x-m)^2 f(x)\, dx & \text{si } X \text{ est continue de densité } f \end{cases}$$

$\sigma = \sqrt{V(X)}$ est l'écart-type de X.

Exercice. Montrer que $V(X) = E(X^2) - E(X)^2$

Remarque.
Si X est discrète, les formules sont à comparer avec les formules de statistiques descriptives rappelées dans le tableau suivant avec 3 types de données :

1. Les données brutes (individu par individu) de la forme :

$$x_1,\ x_2,\ \ldots,\ x_n$$

2. Les données regroupées dans le cas discret de la forme :

x_1	x_2	$\ldots$	x_p
n_1	n_2	$\ldots$	n_p

3. Les données regroupées par classe dans le cas continu de la forme :

$[e_1, e_2[$	$[e_2, e_3[$	$\ldots$	$[e_p, e_{p+1}[$
n_1	n_2	$\ldots$	n_p

En notant $n = \sum\limits_{i=1}^{p} n_i$ l'effectif total, $f_i = \frac{n_i}{n}$ la fréquence associée à n_i et $c_i = \frac{e_i + e_{i+1}}{2}$ le centre de la classe $[e_i, e_{i+1}[$, on a :

Type de données	1	2	3
Moyenne $\bar{x}$	$\frac{1}{n}\sum\limits_{i=1}^{n} x_i$	$\frac{1}{n}\sum\limits_{i=1}^{p} n_i x_i = \sum\limits_{i=1}^{p} f_i x_i$	$\frac{1}{n}\sum\limits_{i=1}^{p} n_i c_i = \sum\limits_{i=1}^{p} f_i c_i$
Variance s^2	$\frac{1}{n}\sum\limits_{i=1}^{n} (x_i - \bar{x})^2$	$\frac{1}{n}\sum\limits_{i=1}^{p} n_i (x_i - \bar{x})^2 = \sum\limits_{i=1}^{p} f_i (x_i - \bar{x})^2$	$\frac{1}{n}\sum\limits_{i=1}^{p} n_i (c_i - \bar{x})^2 = \sum\limits_{i=1}^{p} f_i (c_i - \bar{x})^2$

Exemple (Loi de Bernoulli)**.**
La loi de Bernoulli $\mathcal{B}(p)$ est définie par :

$$P(X = 1) = p \text{ et } P(X = 0) = q = 1 - p$$

Son espérance et sa variance sont :

$$E(X) = p \text{ et } V(X) = pq$$

Elle modélise une expérience aléatoire à deux issues possibles : le succès de probabilité p et l'échec de probabilté $q = 1 - p$.

Exemple (Loi exponentielle)**.**
La loi exponentielle $\mathcal{E}(\lambda)$ a pour densité :

$$f(x) = \begin{cases} \lambda e^{-\lambda x} \text{ sur } [0, +\infty[\\ 0 \text{ ailleurs} \end{cases}$$

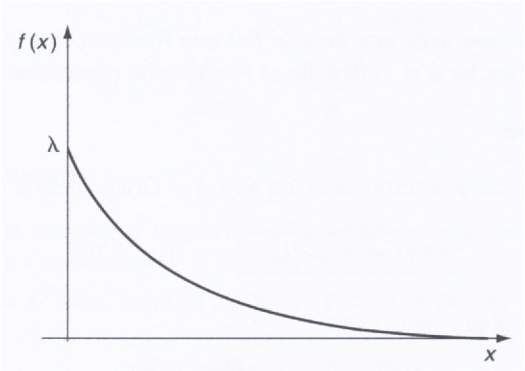


FIGURE 1.3 – Densité de la loi exponentielle

Son espérance et sa variance sont :

$$E(X) = \frac{1}{\lambda} \text{ et } V(X) = \frac{1}{\lambda^2}$$

Elle modélise la durée de vie de phénomènes sans vieillissement (comme les composants électroniques) car pour tout $s, t > 0$:

$$P(X \geq s + t | X \geq s) = P(X \geq t)$$

Pour démontrer ce résultat (et pas seulement), introduisons la notion de **fonction de répartition**.

Definition. *La fonction de répartition d'une va X est l'application F de* $\mathbf{R}$ *dans* $[0, 1]$ *définie par*

$$F(x) = P(X \leq x)$$

F est une fonction croissante et continue à droite telle que $F(-\infty) = 0$ et $F(+\infty) = 1$ (dans $\bar{\mathbf{R}}$).

La fonction de répartition d'une va discrète est une fonction en escalier :

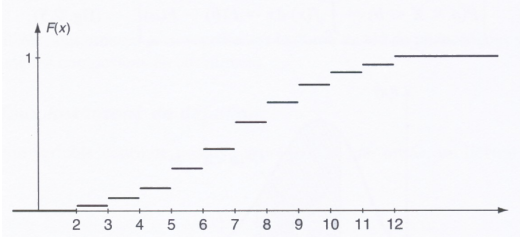


FIGURE 1.4 – Fonction de répartition d'une va discrète

Pour une va continue, cela ressemble à :

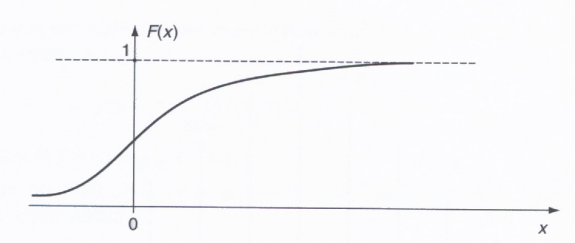


FIGURE 1.5 – Fonction de répartition d'une va continue

De plus, si X est continue de densité f, alors sa fonction de répartition F est dérivable et admet f pour dérivée. En effet,

$$F(x) = P(X \in]-\infty, x]) = \int_{-\infty}^{x} f(x)\,dx$$

Par conséquent, on a :

$$P(a < X \leq b) = \int_{a}^{b} f(x)\,dx = F(b) - F(a)$$

Remarque. Pour une variable continue :
– $P(X = x) = 0$ et $P(X \leq x) = P(X < x)$
– Une probabilité est une aire sous la courbe de sa densité

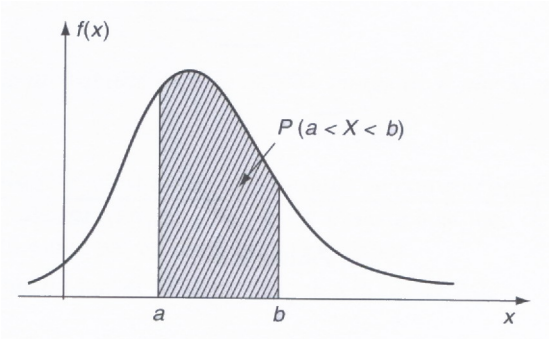


FIGURE 1.6 – Aire sous la courbe

1.2.2 Proportion

On s'intéresse tout simplement à la proportion p c'est à dire à la part des individus dans une population possédant un certain caractère.

Remarque. Cette proportion p est en fait la moyenne de la va de Bernouilli qui, à un individu, associe 1 s'il possède le caractère désiré et 0 sinon.

Chapitre 2

Estimation

2.1 Estimation ponctuelle et estimateur

L'estimation consiste à donner des valeurs approchées aux paramètres d'une population (m, σ^2, p) à l'aide d'un échantillon (aléatoire simple) de n observations issues de cette population.

2.1.1 Loi Forte des Grands Nombres et applications

Théorème (LFGN). *Si $(X_1, ..., X_n)$ est un échantillon de va indépendantes et de même loi d'espérance m, alors $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \simeq m$ dès que n est assez grand (≥ 30).*

La LFGN nous assure que la moyenne empirique $\bar{X}$ "est" une application constante égale à la moyenne théorique m dès que n est assez grand. Toute réalisation $\bar{x}$ de $\bar{X}$ est donc une estimation de m. On dit aussi que $\bar{X}$ est un estimateur de m.

On notera bien ici la différence entre l'estimateur $\bar{X}$ (en majuscule) qui est une va et l'estimation $\bar{x}$ (en minuscule) qui est une valeur. C'est la même différence qu'entre l'application f et la valeur $f(x)$ en Mathématiques !

De même, la LFGN nous assure que la variance empirique $S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ est un estimateur de la variance théorique σ^2.

En utilisant la remarque faite précédemment sur la proportion et la LFGN, on montre enfin que la fréquence empirique $F = \frac{1}{n}\sum_{i=1}^{n} X_i$ est un estimateur de la proportion p où les X_i sont des Bernouilli de paramètre p.

Attention, le même paramètre peut être estimé à l'aide d'estimateurs différents. Il convient donc de définir les qualités exigées d'un estimateur pour choisir "le meilleur".

2.1.2 Qualités d'un estimateur

Soit θ le paramètre à estimer et T un estimateur [1] .

La première qualité d'un estimateur est d'être **convergent** :

$$\lim_{n \to +\infty} T = \theta$$

1. Va qui est fonction des observations X_i et dont la loi dépend de θ(définition rigoureuse).

Deux estimateurs convergents ne convergent cependant pas nécessairement à la même vitesse, on parle alors de précision.

Supposons maintenant la loi de probabilité de T connue pour une valeur donnée de θ.

Biais

L'erreur d'estimation $T - \theta$ est une va qui se décompose de la manière suivante :

$$T - \theta = T - E(T) + E(T) - \theta$$

$T - E(T)$ représente les fluctuations aléatoires de T autour de sa valeur moyenne tandis que $E(T) - \theta$, appelé **biais**, correspond à une erreur systématique due au fait que T varie autour de sa valeur centrale $E(T)$ et non autour de θ.

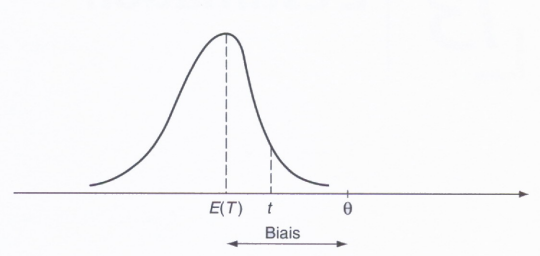


FIGURE 2.1 – Biais

Il est donc souhaitable d'utiliser des estimateurs **sans biais** vérifiant $E(T) = \theta$.

Ainsi, $\bar{X}$ est un estimateur sans biais de m.

Attention, S^2 est biaisé [2] pour σ^2. En effet, de $X_i - m = X_i - \bar{X} + \bar{X} - m$, on tire la décomposition :

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - m)^2 - (\bar{X} - m)^2$$

et donc le biais :

$$E(S^2) - \sigma^2 = \frac{n-1}{n}\sigma^2 - \sigma^2 = -\frac{\sigma^2}{n}$$

Pour ne pas sous-estimer σ^2, on préfèrera souvent la variance corrigée d'espérance σ^2 :

$$S^{*2} = \frac{n}{n-1}S^2$$

Attention, l'écart-type corrigé S^* reste biaisé pour σ mais asymptotiquement sans biais [3].

Précision

La précision de T est souvent mesurée par l'erreur quadratique moyenne $E\Big((T-\theta)^2\Big)$ qui se décompose [4] sous la forme :

$$E\Big((T-\theta)^2\Big) = V(T) + \Big(E(T) - \theta\Big)^2$$

De deux estimateurs sans biais, le plus précis est donc celui de variance minimale.

Ainsi, l'estimateur $D = \frac{1}{n}\sum_{i=1}^{n}(X_i - m)^2$ est meilleur que S^{*2} dès que m est connue.

Exemple. Pour illustrer les différents estimateurs de la variance, on pourra se reporter à l'annexe 1.

2. Mais asymptotiquement sans biais ie $\lim_{n\to+\infty}(E(S^2) - \sigma^2) = 0$
3. Il n'existe pas d'expression générale donnant $E(S^*)$ pour toute distribution.
4. Il s'agit de la décomposition biais-variance

2.2 Estimation par intervalle de confiance

Il est souvent plus intéressant de fournir un renseignement du type $a < \theta < b$ plutôt que $\hat{\theta} = c$.

2.2.1 Principe

Soit T un estimateur de θ (le meilleur possible) dont on connaît la loi de probabilité [5].

Etant donnée une valeur θ_0 de θ (sa vraie valeur par exemple), déterminer un intervalle de probabilité de niveau $1 - \alpha$ pour T revient à chercher deux réels $t_1 < t_2$ vérifiant :

$$P(t_1 < T < t_2 | \theta = \theta_0) = 1 - \alpha$$

On peut traduire cette méthode dans un plan (θ, T) où l'on trace $t_1(\theta)$ et $t_2(\theta)$:

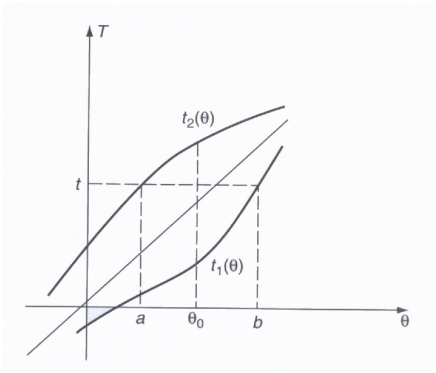


FIGURE 2.2 – Estimation par intervalles

On lit alors l'intervalle de probabilité selon la verticale issue de θ_0 et l'intervalle de confiance selon l'horizontale issue de t (réalisation de T).

Si l'on augmente le niveau de confiance $1 - \alpha$, les courbes s'écartent et donc l'intervalle grandit.
Si la taille de l'échantillon augmente, les courbes se rapprochent et donc l'intervalle diminue.

2.2.2 Loi normale

La loi normale $\mathcal{N}(m, \sigma)$ a pour densité :

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2\right)$$

Son espérance et sa variance sont :

$$E(X) = m \text{ et } V(X) = \sigma^2$$

Dans le calcul des probabilités, on utilise le changement de variable $U = \dfrac{X - m}{\sigma}$ pour se ramener à la loi normale centrée réduite $\mathcal{N}(0, 1)$ qui est tabulée (cf Annexes).

Pour démontrer ce résultat (et pas seulement), nous avons besoin des propriétés suivantes :
– une transformée affine de gaussienne est encore une gaussienne
– $E(aX + b) = aE(X) + b$
– $V(aX + b) = a^2 V(X)$

5. La loi peut, par exemple, être caractérisée par sa densité qui est fonction de θ

Enfin, si $X_1 \sim \mathcal{N}(m_1, \sigma_1)$ et $X_2 \sim \mathcal{N}(m_2, \sigma_2)$ sont indépendantes, alors

$$aX_1 + bX_2 \sim \mathcal{N}\left(am_1 + bm_2, \sqrt{a^2\sigma_1^2 + b^2\sigma_2^2}\right)$$

En effet, on a :
– une combinaison linéaire de deux gaussiennes indépendantes est encore une gausienne
– $E(aX + bY) = aE(X) + bE(Y)$
– Si X et Y sont indépendantes [6], alors $V(X + Y) = V(X) + V(Y)$

Exercice.
On note X la note d'un candidat, choisi au hasard parmi tous les candidats ayant passé un examen, et l'on suppose que $X \sim \mathcal{N}(7, 2)$.

1. Déterminer la proportion de candidats ayant obtenu au moins 10/20.
2. Déterminer le premier décile c'est à dire la note en dessous de laquelle se situent 10% des candidats.
3. Le but de cette question est de réajuster à l'aide d'une transformation affine $Y = aX + b$ (a et b étant des réels positifs) les notes de la promotion de sorte que :
   – 50% des candidats aient obtenu au moins 10/20
   – le premier décile soit égal à 7
   (a) Déterminer la loi de Y en fonction de a et b.
   (b) Déterminer un système de deux équations en a et b issu des deux conditions et conclure.

2.2.3 Moyenne

Supposons $X \sim \mathcal{N}(m, \sigma)$ et estimons m.

Quand σ est connu

$\bar{X}$ est le meilleur estimateur de m et $\bar{X} \sim \mathcal{N}(m, \frac{\sigma}{\sqrt{n}})$ ou encore $U = \frac{\sqrt{n}}{\sigma}(\bar{X} - m) \sim \mathcal{N}(0, 1)$.

L'intervalle de probabilité (à risques symétriques) de U au niveau $1 - \alpha$ est donc :

$$-u_{(1-\frac{\alpha}{2})} < U < u_{(1-\frac{\alpha}{2})}$$

où $u_{(1-\frac{\alpha}{2})}$ vérifie [7] $P(U < u_{(1-\frac{\alpha}{2})}) = 1 - \frac{\alpha}{2}$

L'intervalle de confiance est alors :

$$\bar{x} - u_{(1-\frac{\alpha}{2})}\frac{\sigma}{\sqrt{n}} < m < \bar{x} + u_{(1-\frac{\alpha}{2})}\frac{\sigma}{\sqrt{n}}$$

Quand σ est inconnu

On utilise ici le fait que $T = \frac{\sqrt{n}}{S^*}(\bar{X} - m) \sim \mathcal{T}_{n-1}$ [8].

L'intervalle de probabilité (à risques symétriques) de T au niveau $1 - \alpha$ est donc :

$$-t_{(1-\frac{\alpha}{2})} < T < t_{(1-\frac{\alpha}{2})}$$

où $t_{(1-\frac{\alpha}{2})}$ vérifie [9] $P(T_{n-1} < t_{(1-\frac{\alpha}{2})}) = 1 - \frac{\alpha}{2}$

6. il suffit en fait qu'elle soit non corrélées c'est à dire que $E(XY) = E(X)E(Y)$
7. L'utilisation de la table de la loi normale centrée réduite fournit par exemple $u_{(1-\frac{\alpha}{2})} = 1,96$ pour $1 - \alpha = 0,95$
8. $\mathcal{T}_n = \frac{U}{\sqrt{\frac{X}{n}}}$ avec $U \sim \mathcal{N}(0, 1)$ et $X \sim \chi^2_n$ indépendantes
9. On utilise ici la table de la loi de Student (cf Annexes)

L'intervalle de confiance est alors :

$$\bar{x} - t_{(1-\frac{\alpha}{2})}\frac{s^*}{\sqrt{n}} < m < \bar{x} + t_{(1-\frac{\alpha}{2})}\frac{s^*}{\sqrt{n}}$$

Quand l'échantillon n'est plus gaussien

On utilise le Théorème Central Limite :

Théorème (TCL)**.** *Si $(X_1, ..., X_n)$ est un échantillon de va indépendantes et de même loi d'espérance m et d'écart-type σ, alors $\frac{\sqrt{n}}{\sigma}(\bar{X} - m) \sim \mathcal{N}(0, 1)$ dès que n est assez grand (≥ 30).*

En effet, si l'échantillon n'est plus gaussien mais de grande taille ($n > 30$), le TCL (accompagné du théorème de Slutsky pour le cas où σ est inconnu) nous assure que les variables $\frac{\sqrt{n}}{\sigma}(\bar{X}-m)$ et $\frac{\sqrt{n}}{S^*}(\bar{X}-m)$ suivent approximativement une $\mathcal{N}(0, 1)$ et donc ...

Exercice.
Un artisan qui fabrique des objets de maroquinerie souhaite estimer le nombre moyen m de porte-cartes vendus quotidiennement. En notant ses ventes sur 36 jours, il obtient une moyenne de 120 et un écart-type corrigé de 17. Donner un intervalle de confiance pour m au seuil de 95% dans les deux cas suivants :

1. Si le nombre de porte-cartes est gaussien.
2. Sans l'hypothèse de normalité.

2.2.4 Variance

Supposons $X \sim \mathcal{N}(m, \sigma)$ et estimons σ^2.

Quand m est connue

$D = \frac{1}{n}\sum_{i=1}^{n}(X_i - m)^2$ est le meilleur estimateur de σ^2 et $\frac{nD}{\sigma^2} \sim \chi^2_n$ [10].

L'intervalle de probabilité de $\frac{nD}{\sigma^2}$ au niveau $1 - \alpha$ est donc :

$$k_{\frac{\alpha}{2}} < \frac{nD}{\sigma^2} < k_{(1-\frac{\alpha}{2})}$$

où $k_{\frac{\alpha}{2}}, k_{(1-\frac{\alpha}{2})}$ vérifient [11] $P(k_{\frac{\alpha}{2}} < \chi^2_n < k_{(1-\frac{\alpha}{2})}) = 1 - \alpha$

L'intervalle de confiance est alors :

$$\frac{nd}{k_{(1-\frac{\alpha}{2})}} < \sigma^2 < \frac{nd}{k_{\frac{\alpha}{2}}}$$

Quand m est inconnue

On utilise ici $S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ et le fait que $\frac{nS^2}{\sigma^2} \sim \chi^2_{n-1}$.

L'intervalle de probabilité de $\frac{nS^2}{\sigma^2}$ au niveau $1 - \alpha$ est donc :

$$k_{\frac{\alpha}{2}} < \frac{nS^2}{\sigma^2} < k_{(1-\frac{\alpha}{2})}$$

10. $\chi^2_n = \sum_{i=1}^{n} U_i^2$ avec les $U_i \sim \mathcal{N}(0, 1)$ indépendantes

11. On utilise ici la table de la loi du χ^2 (cf Annexes)

où $k_{\frac{\alpha}{2}}, k_{(1-\frac{\alpha}{2})}$ vérifient $P(k_{\frac{\alpha}{2}} < \chi^2_{n-1} < k_{(1-\frac{\alpha}{2})}) = 1-\alpha$

L'intervalle de confiance est alors :

$$\frac{ns^2}{k_{(1-\frac{\alpha}{2})}} < \sigma^2 < \frac{ns^2}{k_{\frac{\alpha}{2}}}$$

Nota Bene. *Ces intervalles ne sont valables que si X suit une loi normale.*

Exercice.
En mesurant la quantité d'alcool (gr/l) contenue dans 10 cidres doux du marché, on obtient :

$$5,42 - 5,55 - 5,61 - 5,91 - 5,93 - 6,15 - 6,20 - 6,79 - 7,07 - 7,37$$

Supposons que la quantité d'alcool suive une loi normale de moyenne μ et d'écart-type σ.

1. Déterminer l'intervalle de confiance, au centième près, pour μ au niveau 95% :
   (a) Si $\sigma = 0,6$ g/l.
   (b) Si σ est inconnu.
2. Déterminer l'intervalle de confiance, au centième près, pour σ^2 au niveau 95% :
   (a) Si $\mu = 6$ g/l.
   (b) Si μ est inconnue.

2.2.5 Proportion

Etant donné une population infinie (ou finie si le tirage s'effectue avec remise) où une proportion p des individus possède un certain caractère, il s'agit de trouver un intervalle de confiance pour p à partir de la fréquence f obtenue dans un n-échantillon.

Le nombre d'individus nF possédant le caractère étudié dans le n-échantillon suit une loi binomiale $\mathcal{B}(n,p)$ donc si n est grand, l'approximation d'une binomiale par une gaussienne fournit :

$$nF \sim \mathcal{N}\left(np, \sqrt{np(1-p)}\right)$$

ou encore

$$F \sim \mathcal{N}\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

L'intervalle de probabilité (à risques symétriques) de F au niveau $1-\alpha$ est donc :

$$p - u_{(1-\frac{\alpha}{2})}\sqrt{\frac{p(1-p)}{n}} < F < p + u_{(1-\frac{\alpha}{2})}\sqrt{\frac{p(1-p)}{n}}$$

où $u_{(1-\frac{\alpha}{2})}$ vérifie [12] $P(U < u_{(1-\frac{\alpha}{2})}) = 1 - \frac{\alpha}{2}$

L'intervalle de confiance est alors :

$$f - u_{(1-\frac{\alpha}{2})}\sqrt{\frac{f(1-f)}{n}} < p < f + u_{(1-\frac{\alpha}{2})}\sqrt{\frac{f(1-f)}{n}}$$

Exercice.
Un échantillon de 100 votants choisis au hasard parmi tous les votants d'une circonscription a montré que 55% d'entre eux étaient favorables à un certain candidat.

1. Estimer par intervalle de confiance la proportion de votants favorables à ce candidat au seuil de 95%.
2. Déterminer la taille de l'échantillon minimal pour assurer, au seuil de 95%, une incertitude n'excédant pas 2%.

12. L'utilisation de la table de la loi normale centrée réduite fournit par exemple $u_{(1-\frac{\alpha}{2})} = 1,96$ pour $1-\alpha = 0,95$

Chapitre 3

Tests statistiques

3.1 Introduction

3.1.1 Les faiseurs de pluie

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 $\mathcal{N}(600, 100)$. Des entrepreneurs, surnommés faiseurs de pluie, prétendaient pouvoir augmenter de 50 mm le niveau moyen de pluie, ceci par insémination des nuages au moyen d'iodure d'argent. Leur procédé fut mis à l'essai entre 1951 et 1959 et on releva les hauteurs de pluies suivantes :

Année	1951	1952	1953	1954	1955	1956	1957	1958	1959
mm	510	614	780	512	501	534	603	788	650

Que pouvait-on en conclure ? Deux hypothèses s'affrontaient : ou bien l'insémination était sans effet, ou bien elle augmentait réellement le niveau moyen de pluie de 50 mm. Si m désigne l'espérance mathématique de X, variable aléatoire égale au niveau annuel de pluie, ces hypothèses pouvaient se formaliser comme suit :

$$\begin{cases} H_0 : m = 600 \text{ mm} \\ H_1 : m = 650 \text{ mm} \end{cases}$$

Les agriculteurs, hésitant à opter pour le procédé forcément onéreux des faiseurs de pluie, tenaient à l'hypothèse H_0 et il fallait donc que l'expérience puisse les convaincre. Ils choisirent $\alpha = 0,05$ comme niveau de probabilité autrement dit ils étaient prêts à accepter H_1 si le résultat obtenu faisait partie d'une éventualité improbable qui n'avait que 5 chances sur 100 de se produire sous H_0.

Question : pouvons-nous rejeter l'hypothèse H_0 (au profit de H_1) ?

Puisqu'il s'agit de tester la valeur de m, il est naturel d'utiliser la moyenne empirique $\bar{X}$ des observations. En fait, on utilise une variable, appelée variable de décision, qui dépend du paramètre à tester m et dont la loi sous H_0 est tabulée :

$$T = \frac{\bar{X} - m_0}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0, 1) \text{ si } H_0 \text{ est vraie}$$

Si T est trop grand, supérieur à un seuil k qui n'a que 5 chances sur 100 d'être dépassé si H_0 est vraie [1], on optera pour H_1 avec une probabilité de se tromper égale à 0,05. Par contre, si $T < k$, on conservera H_0 faute de preuves suffisantes.

Ici, la table fournit :

$$k = 1,64$$

La règle de décision est donc :

1. Ce raisonnement probabiliste est à comparer avec le raisonnement par l'absurde sauf que le résultat impossible est ici remplacé par un résultat très peu probable, et la négation de l'hypothèse de départ par l'hypthèse alternative H_1

– Si $T > 1,64$, on rejette H_0 (et on accepte H_1)
– Si $T < 1,64$, on ne rejette pas H_0

L'ensemble d'événements $\{T > 1,64\}$ s'appelle la région critique (ou région de rejet de H_0). Son complémentaire $\{T < 1,64\}$ s'appelle la région d'acceptation de H_0.

Ici, les données relevées indiquent que $t = \frac{610,2-600}{100/3} = 0,306$ donc on ne rejette pas H_0.

Attention, on peut accepter H_0 à tort. En effet, on commet une erreur chaque fois que $\bar{X}$ prend une valeur inférieure à 655, mais $T = \dfrac{\bar{X} - 650}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0,1)$ si H_1 est vraie donc on commet une erreur avec une probabilité :

$$\beta = P\left(U < \frac{655 - 650}{100/3}\right) = P(U < 0,15) = 0,56$$

α (resp. β) s'appelle le risque de première (resp. deuxième) espèce.

Il convient enfin de remarquer le rôle particulier joué par H_0 : si la forme de la région critique $\{T > k\}$ est indiquée par la nature de H_1 ($650 > 600$), la valeur de k ne dépend que de H_0.

3.1.2 Quelques généralités

Un test est un mécanisme qui permet de trancher entre deux hypothèses au vu des résultats d'un échantillon.

En notant H_0 et H_1 ces deux hypothèses, dont une et une seule est vraie, les quatre cas possibles sont représentés dans le tableau suivant :

Décision \ Vérité	H_0	H_1
H_0	1-α	β
H_1	α	1-β

α et β désignent les probabilités d'erreur de première et deuxième espèce :
– α est la probabilité de rejeter H_0 à tort
– β est la probabilité de conserver H_0 à tort

Notons que ces erreurs correspondent à des risques différents. Ainsi, dans l'exemple des faiseurs de pluie, le risque de première espèce consiste à acheter un procédé d'insémination inefficace alors que le risque de deuxième espèce consiste à perdre une occasion d'augmenter le niveau de pluie et donc d'obtenir une récolte plus abondante.

Dans la pratique des tests statistiques, il est de règle de se fixer α ce qui fait jouer à H_0 un rôle prééminent :
– H_0 peut être une hypothèse solidement établie n'ayant jamais été contredite par l'expérience
– H_0 peut être une hypothèse de prudence (l'innocuité d'un vaccin, l'innocence d'une personne)
– H_0 peut être une hypothèse à laquelle on tient pour des raisons qui peuvent être subjectives
– H_0 peut être la seule hypothèse facile à formuler ($m = m_0$ contre $m \neq m_0$)

α étant fixé, β sera alors déterminé comme résultat d'un calcul (à condition que la loi de probabilité sous H_1 soit connue). Notons cependant que β varie dans le sens contraire de α. En effet, diminuer α conduit à une règle de décision plus stricte qui aboutit à n'abandonner H_0 que dans des cas rarissimes et donc à conserver H_0 bien souvent à tort ce qui revient à augmenter β ou encore à diminuer la puissance du test [2] $1 - \beta$.

2. La méthode de Neyman et Pearson permet de maximiser la puissance du test $1 - \beta$ pour une valeur donnée de α en choisissant la variable de décision et la région critique optimales

α étant fixé, il importe maintenant de choisir une variable de décision : variable dont la loi doit être connue sous H_0 et bien entendu différente sous H_1.

La région critique W est alors l'ensemble des valeurs de la variable de décision qui conduisent à écarter H_0 (au profit de H_1). Sa forme étant déterminée par la nature de H_1, sa détermination exacte se fait en écrivant que :

$$P(W|H_0) = \alpha$$

La région d'acceptation étant son complémentaire $\bar{W}$, on a :

$$P(\bar{W}|H_0) = 1 - \alpha$$

Pour résumer, voici la démarche d'un test :

1. Choix de H_0 et H_1
2. Détermination de la variable de décision et de sa loi sous H_0
3. Détermination de la forme de la région critique (selon H_1) et de ses bornes en fonction de α
4. Calcul de la valeur expérimentale de la variable de décision
5. Conclusion : rejet ou non de H_0

3.2 Tests de conformité

3.2.1 Moyenne

Supposons $X \sim \mathcal{N}(m, \sigma)$ et testons m.

Quand σ est connu

Sous $H_0 : m = m_0$, la variable de décision $T = \dfrac{\bar{X} - m_0}{\frac{\sigma}{\sqrt{n}}}$ suit une $\mathcal{N}(0, 1)$.

Ainsi, pour le test $\begin{cases} H_0 : m = m_0 \\ H_1 : m = m_1 \text{ avec } m_1 > m_0 \end{cases}$, la région critique est :

$$\{T > k\} \text{ où } k \text{ vérifie } P(U > k) = \alpha$$

ou encore

$$[u_{(1-\alpha)}; +\infty[$$

Remarque. Pour le test $\begin{cases} H_0 : m = m_0 \\ H_1 : m > m_0 \end{cases}$, la région critique est encore $[u_{(1-\alpha)}; +\infty[$.

Exercice. Montrer que :

1. Pour le test $\begin{cases} H_0 : m = m_0 \\ H_1 : m < m_0 \end{cases}$, la région critique est $]-\infty; -u_{(1-\alpha)}]$.
2. Pour le test $\begin{cases} H_0 : m = m_0 \\ H_1 : m \neq m_0 \end{cases}$, la région critique est $]-\infty; -u_{(1-\frac{\alpha}{2})}] \cup [u_{(1-\frac{\alpha}{2})}; +\infty[$.

Quand σ est inconnu

Sous $H_0 : m = m_0$, la variable de décision $T = \dfrac{\bar{X} - m_0}{\frac{S^*}{\sqrt{n}}}$ suit une $\mathcal{T}_{n-1}$.

Ainsi, pour le test $\begin{cases} H_0 : m = m_0 \\ H_1 : m \neq m_0 \end{cases}$, la région critique est :

$$\{|T| > k\} \text{ où } k \text{ vérifie } P(|\mathcal{T}_{n-1}| > k) = \alpha$$

ou encore

$$]-\infty; -t_{(1-\frac{\alpha}{2})}] \cup [t_{(1-\frac{\alpha}{2})}; +\infty[$$

Nota Bene. *Si l'échantillon n'est plus gaussien mais de grande taille ($n > 30$), le TCL (accompagné du théorème de Slutsky pour le cas où σ est inconnu) nous assure que T suit approximativement une $\mathcal{N}(0,1)$, et donc que les régions critiques sont les mêmes que dans le cas où l'échantillon est gaussien avec σ connu.*

Exercice.
Lors d'une enquête sur le temps de sommeil par nuit des enfants de 2 à 3 ans dans un département français, on a trouvé une moyenne de 10,2 heures dans un groupe de 40 enfants avec un écart type de 2,1 heures. En France, la moyenne du temps de sommeil par nuit est de 11,7 heures chez les enfants de cet âge.

1. La moyenne dans ce département est-elle significativement différente (test bilatéral) au seuil de 5% de la moyenne française ?
2. Est-elle significativement inférieure (test unilatéral) au seuil de 5% ?

3.2.2 Variance

Supposons $X \sim \mathcal{N}(m, \sigma)$ et testons σ.

Quand m est connue

Sous $H_0 : \sigma = \sigma_0$, la variable de décision $T = \dfrac{nD}{\sigma_0^2}$ suit un χ^2_n.

Ainsi, pour le test $\begin{cases} H_0 : \sigma = \sigma_0 \\ H_1 : \sigma > \sigma_0 \end{cases}$, la région critique est :

$$\{T > k\} \text{ où } k \text{ vérifie } P(\chi^2_n > k) = \alpha$$

ou encore

$$[k_{(1-\alpha)}; +\infty[$$

Quand m est inconnue

Sous $H_0 : \sigma = \sigma_0$, la variable de décision $T = \dfrac{nS^2}{\sigma_0^2} = \dfrac{(n-1)S^{*2}}{\sigma_0^2}$ suit un χ^2_{n-1}.

Ainsi, pour le test $\begin{cases} H_0 : \sigma = \sigma_0 \\ H_1 : \sigma < \sigma_0 \end{cases}$, la région critique est :

$$\{T < k\} \text{ où } k \text{ vérifie } P(\chi^2_{n-1} < k) = \alpha$$

ou encore

$$[0; k_\alpha]$$

Exercice.
Un agent immobilier prétend, lors d'une interview, que le prix moyen des transactions immobilières dans un quartier niçois est de 2400 euros du mètre carré avec un écart-type de 220 euros. Le journaliste chargé du dossier à paraître dans une revue spécialisée décide de vérifier ces affirmations à partir des 50 dernières transactions effectuées par quatre agences du quartier

1695	2202	2722	2534	2494	2648	2298	1997	2118	2767
2867	2391	2029	2121	2105	2565	2652	2497	2822	2713
2014	2350	2343	2398	2505	2630	2169	2661	2325	2031
2683	2328	2710	2417	2264	2299	2531	2423	2592	2577
2568	1992	2872	2603	2415	2072	2475	2089	2140	2720

On en tire : $\bar{x} = 2408.66$ et $s^* = 272.01$

En admettant que le prix de vente suive une loi normale, tester au seuil de 5% les affirmations de l'agent immobilier.

Nota Bene. *Si l'échantillon n'est plus gaussien mais de grande taille ($n > 30$), la variable de décision* $T = \dfrac{S^{*2} - \sigma_0^2}{\sqrt{\frac{2S^{*4}}{n-1}}}$ *suit approximativement une* $\mathcal{N}(0,1)$ *et donc ...*

3.2.3 Proportion

Sous $H_0 : p = p_0$, la variable de décision $T = \dfrac{F - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$ suit approximativement une $\mathcal{N}(0,1)$.

Exercice.
Sur un échantillon de 300 patients traités par un certain remède, 243 ont été guéris.
La proportion de guérison est-elle significativement supérieure à 75% au seuil de 5% ?

3.3 Tests de comparaison de deux échantillons indépendants

3.3.1 Moyennes

Soit deux échantillons **gaussiens** indépendants $X_1 \sim \mathcal{N}(m_1; \sigma_1)$ et $X_2 \sim \mathcal{N}(m_2; \sigma_2)$.

Alors, sous $H_0 : m_1 = m_2$:

– Si les variances sont connues, $T = \dfrac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$ suit une $\mathcal{N}(0,1)$.

– Si les variances sont inconnues et supposées égales [3], $T = \dfrac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}}}$ suit une $\mathcal{T}_{n_1+n_2-2}$ où

$$S_p^2 = \frac{(n_1-1)S_1^{*2} + (n_2-1)S_2^{*2}}{n_1+n_2-2} \text{ est la variance de pool [4]}$$

– Si les variances sont inconnues et supposées différentes, $T = \dfrac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^{*2}}{n_1} + \frac{S_2^{*2}}{n_2}}}$ suit une $\mathcal{T}_m$ où

$$m = \frac{1}{\frac{c^2}{n_1-1} + \frac{(1-c)^2}{n_2-1}} \text{ avec } c = \frac{\frac{S_1^2}{n_1-1}}{\frac{S_1^2}{n_1-1} + \frac{S_2^2}{n_2-1}}$$

Nota Bene. *Si les échantillons ne sont plus gaussiens mais de grandes tailles ($n > 30$), les variables de décision suivent toutes approximativement une* $\mathcal{N}(0,1)$ *et donc ...*

3.3.2 Variances

Soit deux échantillons **gaussiens** indépendants $X_1 \sim \mathcal{N}(m_1; \sigma_1)$ et $X_2 \sim \mathcal{N}(m_2; \sigma_2)$ où $S_1^* > S_2^*$.

Alors, sous $H_0 : \sigma_1^2/\sigma_2^2 = 1$:

– Si les espérances sont connues, $T = \dfrac{D_1}{D_2}$ suit une loi de Fisher [5] $\mathcal{F}(n_1, n_2)$.

– Si les espérances sont inconnues, $T = \dfrac{S_1^{*2}}{S_2^{*2}}$ suit une $\mathcal{F}(n_1 - 1, n_2 - 1)$.

3. Le test de comparaison des espérances doit donc être précédé par celui des variances

4. S_p^2, apppelée aussi variance combinée, n'est rien d'autre que la moyenne des variances corrigées des échantillons, pondérées par les tailles des échantillons diminuées de 1

5. $\mathcal{F}(n,p) = \dfrac{\chi_n^2/n}{\chi_p^2/p}$

Remarque. Le test étant ici $\begin{cases} H_0 : \sigma_1^2/\sigma_2^2 = 1 \\ H_1 : \sigma_1^2/\sigma_2^2 > 1 \end{cases}$, la région critique est $[f_{(1-\alpha)}; +\infty[$.

Exercice.
Les QI de 9 enfants d'un quartier d'une grande ville ont une moyenne de 107 avec un écart-type de 10. Les QI de 12 enfants d'un autre quartier ont une moyenne de 112 avec un écart-type de 9. On suppose que la variable aléatoire associée au QI suit une loi Normale.
Y a-t-il une différence significative au seuil de 5% entre les QI moyens des 2 quartiers ?

Nota Bene. *Si les échantillons ne sont plus gaussiens mais de grandes tailles ($n > 30$) et de distributions unimodales pas trop dissymétriques, on considère la variable de décision* $T = \dfrac{S_1^{*2} - S_2^{*2}}{\sqrt{\frac{2S_1^{*4}}{n_1-1} + \frac{2S_2^{*4}}{n_2-1}}}$ *qui, sous* $H_0 : \sigma_1^2 = \sigma_2^2$, *suit approximativement une* $\mathcal{N}(0,1)$ *et donc ...*

Exercice.
Des essais cliniques sont menés auprès de 137 patients atteints d'une maladie pulmonaire sans gravité afin de tester l'efficacité d'un traitement à la pulmotrycine.
Le protocole est le suivant :
- Des exercices respiratoires sont prescrits à 67 patients choisis au hasard ainsi qu'un placebo (groupe témoin).
- Les mêmes exercices respiratoires sont precrits aux 70 autres patients ainsi que de la pulmotricyne (groupe traité).
- Au bout de trois mois, l'amélioration de la capacité pulmonaire de chaque patient est mesurée sur une échelle de 0 (pas d'amélioration) à 10 (récupération totale).

Voici les résultats obtenus :

Amélioration	Groupe témoin (A)	Groupe traité (B)
0	2	0
1	8	0
2	4	3
3	7	0
4	14	10
5	9	14
6	5	13
7	4	17
8	7	10
9	2	3
10	5	0

On en tire : $\bar{x}_A = 4.78$, $s_A^2 = 7.37$, $\bar{x}_B = 6$ et $s_B^2 = 2.66$

L'amélioration moyenne du groupe traité est supérieure à celle du groupe témoin (de combien ?) mais cette différence doit-elle être attribuée aux bienfaits de la pulmotrycine ou aux fluctuations d'échantillonage (le même protocole sur d'autres individus n'aurait sans doute pas donné les mêmes résultats). Autrement dit, la supériorité observée entre les deux moyennes empiriques est-elle statistiquement significative au seuil de 2% ?

3.3.3 Proportions

Soit deux échantillons indépendants où F_1 et F_2 sont approximativement gaussiennes [6].

Alors, sous $H_0 : p_1 = p_2$, $T = \dfrac{F_1 - F_2}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}}$ suit une $\mathcal{N}(0,1)$ où $\hat{p} = \frac{n_1 f_1 + n_2 f_2}{n_1 + n_2}$ et donc ...

6. C'est le cas dès que les échantillons sont de grande taille

Exercice.
Lors des primaires d'une campagne présidentielle, des sympathisants d'un parti sont interrogés sur leur opinion à propos d'un candidat avant et après un débat télévisé. Avant le débat, 64% des 980 personnes interrogées déclarent avoir une opinion positive sur le candidat. Après le débat, cette proportion n'est plus que de 61% chez 1001 autres personnes interrogées.

Cette baisse est-elle significative au seuil de 5% ?

3.4 Test d'indépendance du chi 2

Soit deux variables X et Y et un n-échantillon fournissant les effectifs observés N_{ij} [7].

On considère alors la distance $D^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(N_{ij} - N_{ij}^t)^2}{N_{ij}^t}$ entre les effectifs observés N_{ij} et les effectifs théoriques d'indépendances $N_{ij}^t = \frac{N_{i.}N_{.j}}{n}$, qui ne saurait être trop grande sous l'hypothèse nulle d'indépendance.

Par ailleurs, D^2 suit approximativement un $\chi^2_{(I-1)(J-1)}$ sous H_0 dès que $n_{ij}^t \geq 5$ pour tout i, j.

Par conséquent, la région critique est $[k_{(1-\alpha)}; +\infty[$.

Exercice.
Voici la préférence de 80 hommes et 70 femmes pour un type de vin.

	Rouge	Rosé	Blanc
Homme	40	20	20
Femme	30	30	10

Les 150 couples observés permettent-ils de conclure que le type de vin préféré est indépendant du sexe ?

3.5 Test d'ajustement du chi 2

Un test d'ajustement a pour objectif de tester si une distribution observée est modélisable par une loi théorique discrète ou discrétisée, c'est à dire divisée en k classes de probabilités $p_1, p_2, \cdots, p_k$.

Soit donc un n-échantillon de cette loi théorique fournissant les effectifs $N_1, N_2, \cdots, N_k$ de chaque classe.

On considère alors la distance $D^2 = \sum_{i=1}^{k} \frac{(N_i - N_i^t)^2}{N_i^t}$ entre les effectifs observés N_i et les effectifs théoriques d'ajustement $N_i^t = np_i$, qui ne saurait être trop grande sous l'hypothèse nulle H_0 : l'ajustement est correct.

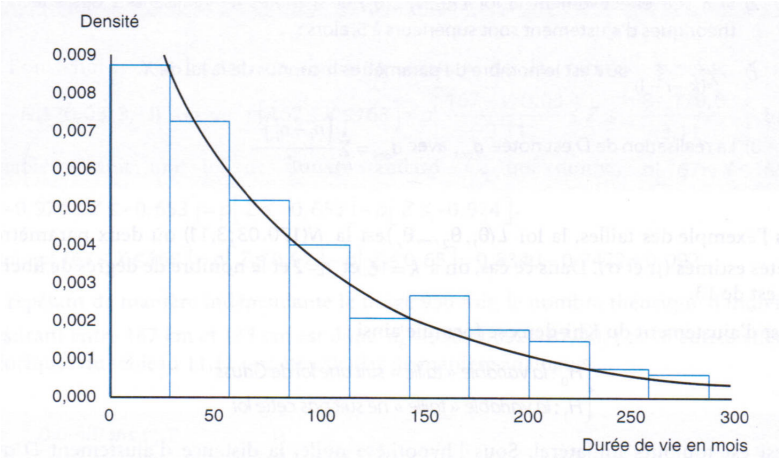
Par ailleurs, si la loi théorique possède l paramètres à estimer, D^2 suit approximativement [8] un χ^2_{k-1-l} sous H_0 dès que $n_i^t \geq 5$ pour tout i (un regroupement de classes permettra toujours de vérifier ses conditions).

Par conséquent, la région critique est $[k_{(1-\alpha)}; +\infty[$.

Exercice (Quel modèle pour la durée de vie de la TX100 ?)**.**
L'objectif est d'ajuster l'histogramme des durées de vie observées par la densité de probabilité d'une loi

7. Evidemment, N_{ij}désigne le nombre d'individus ayant la i^e valeur pour X et la j^e pour Y
8. En réalité, si les estimations ne sont pas celles du maximum de vraisemblance effectuées au moyen des k classes, la loi limite de D^2 n'est plus un χ^2 mais reste comprise entre un χ^2_{k-1} et un χ^2_{k-1-l}

théorique. On constate que l'histogramme est fortement dissymétrique avec un étalement à droite. Cette forme rappelle celle d'une loi exponentielle. Cet ajustement est-il correct ?



Le calcul de la distance d^2 d'ajustement est détaillé dans le tableau suivant.

	A	B	C	D	E	F	G	H
1	Classe	Inf	Sup	ni	n^t_i	d^2i		Centre
2	1	0	29	59	67,554282	1,0832138		14,5
3	2	29	58	49	47,968097	0,0221986		43,5
4	3	58	87	35	34,060585	0,0259097		72,5
5	4	87	116	27	24,185313	0,3275733		101,5
6	5	116	145	14	17,173204	0,5863334		130,5
7	6	145	174	18	12,194133	2,7642873		159,5
8	7	174	203	12	8,6586573	1,2894114		188,5
9	8	203	232	10	6,1482308	2,4130724		217,5
10	9	232	261	5	4,3656586	0,0921714		246,5
11	10	261	290	4	3,0999121	0,2613488		275,5
12				233		8,8655201		
13	Paramètre	0,0118067						
14						0,3537673		
15								
16	Classe	Inf	Sup	ni	n^t_i	d^2i		Centre
17	1	0	29	59	67,554282	1,0832138		14,5
18	2	29	58	49	47,968097	0,0221986		43,5
19	3	58	87	35	34,060585	0,0259097		72,5
20	4	87	116	27	24,185313	0,3275733		101,5
21	5	116	145	14	17,173204	0,5863334		130,5
22	6	145	174	18	12,194133	2,7642873		159,5
23	7	174	203	12	8,6586573	1,2894114		188,5
24	8	203	232	10	6,1482308	2,4130724		217,5
25	9+10	232	290	9	7,4655708	0,3153775		261
26				233		8,8273775		
27	Paramètre	0,0118067						
28						0,2652885		

Exercice (ex 1 de janvier 2013 pour un ajustement gaussien)**.**

Chapitre 4

Régression linéaire

En 1986, l'Organisation mondiale de la santé (OMS) a présenté les résultats d'une importante étude sur les facteurs de dépenses énergétiques des individus. Celle-ci indique que le métabolisme de base des individus dépend de leurs poids, taille, sexe, âge, état physiologique, régime alimentaire, activité physique et de l'absorption de certaines substances. Le tableau suivant indique seulement les valeurs du métabolisme de base et du poids de 20 individus.

Poids (kg)	Métabolisme (kcal/jour)	Poids (kg)	Métabolisme (kcal/jour)
70	1 819,80	47	1 255,15
75	1 758,08	60	1 371,25
75	1 660,76	47	1 243,04
92	1 836,48	60	1 289,36
73	1 729,03	45	1 314,94
94	2 129,72	53	1 310,58
75	1 884,94	51	1 390,70
70	1 690,12	55	1 401,50
88	1 932,28	58	1 392,22
48	1 095,32	55	1 415,75

La régression du métabolisme par le poids réalisée par Excel fournit :

Statistiques de la régression					
Coefficient de détermination multiple		0,934			
Coefficient de détermination R^2		**0,873**			
Coefficient de détermination R^2		0,866			
Erreur-type (*des erreurs*)		**104,5**			
Observations		20			
ANALYSE DE VARIANCE					
	Degré de liberté	***Somme des carrés***	***Moyenne des carrés***	***F***	***Valeur critique de F***
Régression	1	1 350 243,9	1 350 243,9	123,587 94	1,70964E-09
Résidus	18	196 656,6	10 925,4		
Total	19	1 546 900,5			
	Coefficients	***Erreur-type***	***Statistique t***	***Probabilité***	
Constante	426,037	103,42	4,119	0,000 643 901	
Poids	17,351	1,56	11,117	1,709 64E-09	

4.1 A partir de toute la population (Statistique descriptive)

L'objectif est de montrer comment et sous quelles conditions il est possible de modéliser une relation entre deux variables quantitatives par une équation du type $Y = f(X)$. La modélisation est effectuée en 3 étapes :

- On construit le nuage de points pour, d'une part, infirmer ou confirmer l'intuition de dépendance et, d'autre part, déterminer la forme du modèle (nature de f : linéaire, puissance, exponentielle, logistique)
- On construit le modèle en utilisant la méthode des moindres carrées ordinaires (MCO) s'il est linéaire (sinon, on effectue un changement de variables pour se ramener au cas linéaire).
- On mesure la qualité du modèle

4.1.1 Interpréter le nuage de points

Indépendance

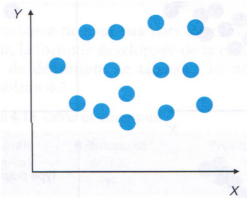


FIGURE 4.1 – Absence de lien entre X et Y

Dépendances de formes différentes

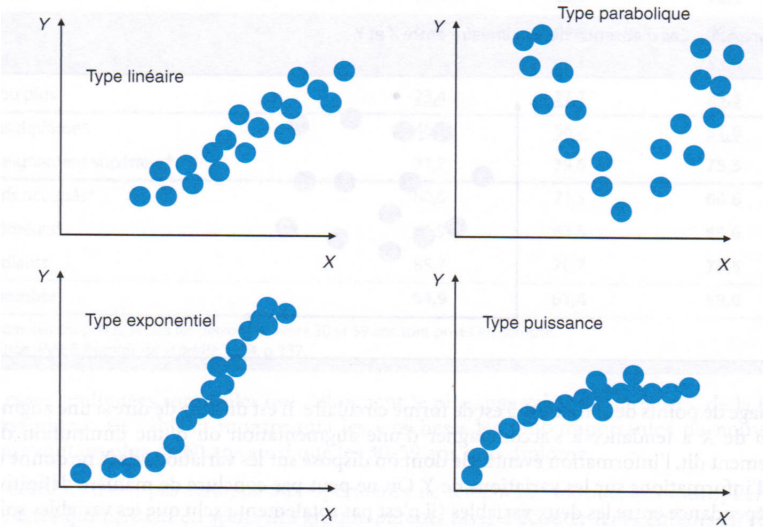


FIGURE 4.2 – Quatre cas de dépendance de formes différentes

Non corrélation

Definition (Covariance)**.** *Soit (X, Y) un couple de va quantitatives, de moyennes respectives m_X et m_Y, pour lequel N couples d'observations (x_i, y_i) ont été relevés. La covariance du couple (X, Y) est définie par :*

$$Cov(X, Y) = \frac{1}{N} \sum_{i=1}^{N} (x_i - m_X)(y_i - m_Y)$$

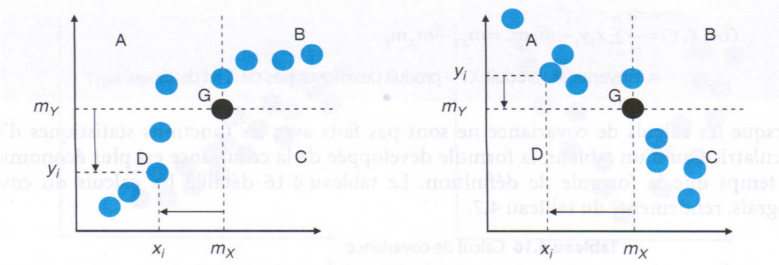


FIGURE 4.3 – Interprétation du signe de la covariance

Interprétation du signe de la covariance :
- La covariance est un indicateur de monotonie : si la covariance est positive (resp. négative), alors X et Y varient en général dans le même sens (resp. dans le sens contraire).
- Si la covariance est nulle ou presque nulle, alors il n'y a pas de tendance croissante ou décroissante et les variables sont dites non corrélées.
- Attention, la covariance n'est pas un indicateur d'indépendance.

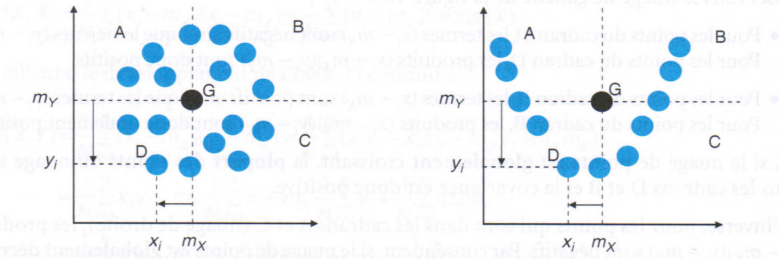


FIGURE 4.4 – Nuages à covariance nulle

Remarque. Deux variables indépendantes sont non corrélées mais la réciproque est fausse.

Definition (Coefficient de corrélation linéaire)**.** *Soit (X, Y) un couple de va quantitatives, d'écarts-types respectifs σ_X et σ_Y. Le coefficient de corrélation linéaire du couple (X, Y) est défini par :*

$$r = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}$$

4.1.2 Construire le modèle

Nous nous plaçons dans le cas où le nuage de points suggère une relation linéaire entre deux variables X et Y. Modéliser la relation consiste à chercher l'équation d'une droite qui ajuste au mieux le nuage de points.

Y est la variable expliquée (ou dépendante) et X la variable explicative (ou indépendante).

Une relation du type $y = ax + b$ définit une droite. Réaliser une régression linéaire de Y en X consiste à rechercher la meilleure droite d'ajustement, à condition de définir ce que l'on entend par "meilleure", c'est à dire à condition de choisir un critère d'optimisation.

En fait, on cherche a et b qui minimisent : $\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{N}(y_i - ax_i - b)^2 = f(a, b)$.

La résolution du système correspondant à l'annulation des deux dérivées partielles de f fournit :

Proposition. *La droite de régression de Y en X d'équation $\hat{y} = ax + b$ est telle que :*

– *La moyenne des valeurs ajustées de Y est égale à la moyenne des valeurs observées de Y :*

$$\frac{1}{N}\sum_{i=1}^{N}\hat{y}_i = \frac{1}{N}\sum_{i=1}^{N} y_i$$

ce qui revient à dire que la droite d'ajustement passe par le point moyen $G(m_X, m_Y)$ ou encore que la moyenne des résidus $e_i = y_i - \hat{y}_i$ est nulle.

– *La pente a vérifie :*

$$a = \frac{Cov(X,Y)}{Var(X)}$$

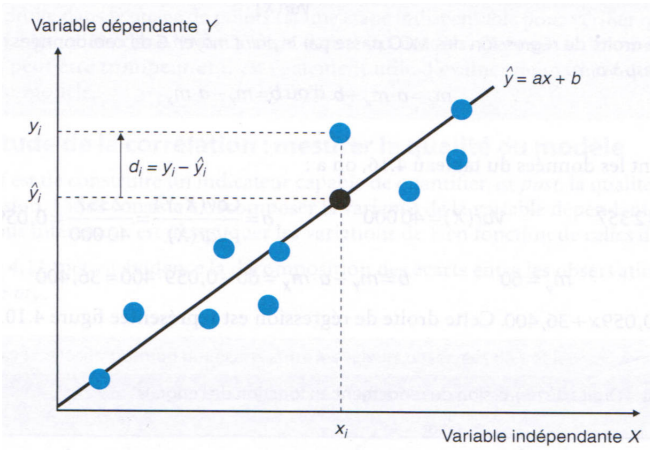


FIGURE 4.5 – Valeurs ajustées de Y et résidus

En fait, on a pour tout i : $y_i = ax_i + b + e_i = \hat{y}_i + e_i$ avec $\frac{1}{N}\sum_{i=1}^{N} e_i = 0$.
En moyenne Y se comporte donc comme $aX + b$ et pour un x_i donné, il n'y a qu'un y_i possible.

Remarque. De nombreux modèles non linéaires se ramènent au modèle linéaire :

– Le modèle puissance $Y = cX^d$ se ramène à

$$Y' = \ln c + dX' \text{ avec } Y' = \ln Y \text{ et } X' = \ln X$$

– Le modèle exponentielle $Y = ce^{dX}$ se ramène à

$$Y' = \ln c + dX \text{ avec } Y' = \ln Y$$

– Le modèle logistique $Y = \dfrac{e^{c+dX}}{1+e^{c+dX}}$ se ramène à

$$Y' = c + dX \text{ avec } Y' = \ln\frac{Y}{1-Y}$$

4.1.3 Mesurer la qualité du modèle

L'objectif est de construire un indicateur capable de quantifier, a posteriori, la qualité de la droite de régression. L'idée consiste à décomposer la variance de la variable expliquée Y.

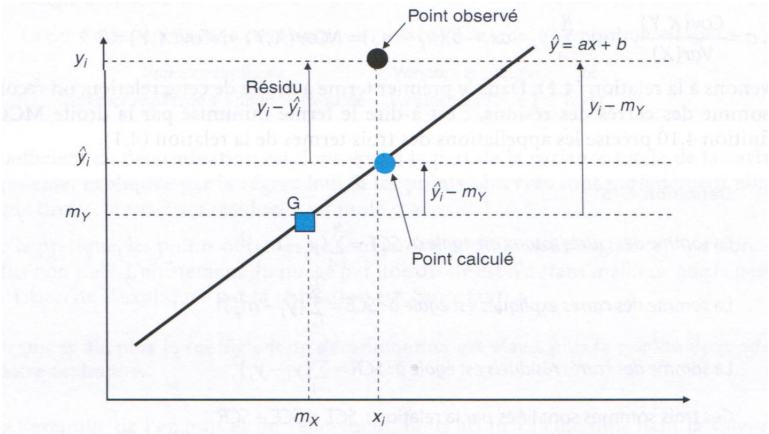


FIGURE 4.6 – Décomposition des écarts entre les valeurs observées de Y et leur moyenne

De la décomposition $y_i - m_Y = (y_i - \hat{y}_i) + (\hat{y}_i - m_Y)$, on tire après calculs [1] :

$$\begin{array}{ccccc} \sum_{i=1}^{N}(y_i - m_Y)^2 & = & \sum_{i=1}^{N}(y_i - \hat{y}_i)^2 & + & \sum_{i=1}^{N}(\hat{y}_i - m_Y)^2 \\ \text{Somme des Carrés Totaux (SCT)} & = & \text{Somme des Carrés Résiduels (SCR)} & + & \text{Somme des Carrés Expliqués (SCE)} \\ \text{Variations totales} & = & \text{Variations résiduelles} & + & \text{Variations expliquées par le modèle} \end{array}$$

Remarque. En divisant par N, il vient le théorème de décomposition des variances :

$$\begin{array}{ccccc} \frac{1}{N}\sum_{i=1}^{N}(y_i - m_Y)^2 & = & \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 & + & \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - m_Y)^2 \\ Var(Y) & = & \text{Variance Résiduelle} & + & \text{Variance Expliquée} \end{array}$$

Enfin, nous avons :

Proposition. *Le carré du coefficient de corrélation linéaire, appelé coefficient de détermination, vérifie :*

$$r^2 = \frac{SCE}{SCT}$$

Le coefficient de détermination est donc égal à la part des variations de Y expliquées par la régression. Si les points observés sont parfaitement alignés sur une droite, r^2 vaut 1. Plus le coefficient de détermination est élevé, plus la qualité du modèle linéaire est bonne.

4.2 A partir d'un échantillon (Statistique inférentielle)

4.2.1 Ce qui change

Précédemment, nous avons posées deux hypothèses simplificatrices :
- On a supposé que l'intégralité de la population était connue ce qui conduit à considérer les coefficients a et b de la régression comme étant les coefficients "réels" et non des estimations.
- On a fait comme si la variable dépendante Y était intégralement expliquée par la variable explicative X ce qui revient à supposer que pour un x_i donné, il existe un unique y_i.

Ici, le modèle de régression linéaire est analysé en relâchant ces deux hypothèses. Ainsi, le raisonnement tient compte à présent de l'incertitude à deux niveaux :
- Celle liée à l'échantillonnage ce qui donne pour tout i : $y_i = \alpha x_i + \beta + e_i$.
- Celle liée au fait que tous les facteurs explicatifs n'étant jamais pris en compte, il convient de rajouter au modèle un terme aléatoire d'erreur E_i ce qui revient à supposer que pour un x_i donné, il est possible d'obtenir différents y_i. On a donc pour tout i : $Y_i = \alpha x_i + \beta + E_i$.

1. Grâce au théorème de Pythagore

4.2.2 Hypothèses du modèle

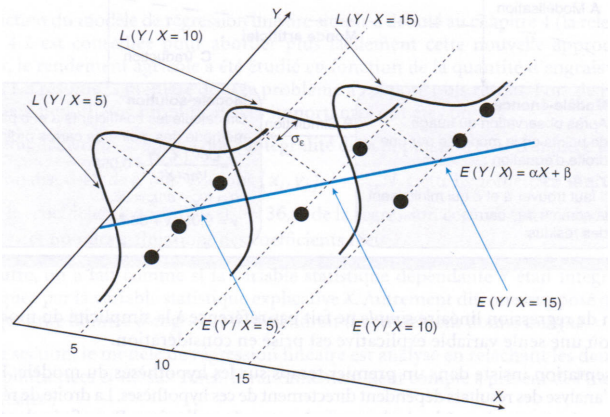


FIGURE 4.7 – Hypothèses de la régression linéaire simple

On suppose que pour chaque essai i de 1 à n, on ait : $Y_i = \alpha x_i + \beta + E_i$ où :
- E_i est une va appelée également terme d'erreur
- Y_i est la réponse aléatoire attendue pour l'essai i
- x_i est la réalisation du facteur explicatif X pour l'essai i
- n est la taille de l'échantillon

De plus, on suppose que les termes d'erreur E_i sont des va indépendantes et de même loi : $E_i \sim \mathcal{N}(0; \sigma_E)$.

Remarque.
- Les variances sont supposées égales (hypothèse d'homoscédasticité)
- L'hypothèse de normalité des E_i entraîne que les Y_i sont aussi des va gaussiennes : $Y_i \sim \mathcal{N}(\alpha x_i + \beta; \sigma_E)$

4.2.3 Estimation des coefficients du modèle

L'objectif est d'utiliser le modèle de régression $E(Y|X = x) = \alpha x + \beta$ pour faire de la prévision. On s'appuie pour cela sur le résultat suivant :

Proposition.
L'estimateur des MCO de α est :

$$A = \frac{\frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{\frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^2}$$

L'estimateur des MCO de β est :

$$B = \bar{Y} - A\bar{X}$$

L'estimateur des MCO du coefficient de détermination ρ^2 est :

$$R^2 = \frac{SCE}{SCT}$$

où $SCE = \sum_{i=1}^{N}(\hat{Y}_i - \bar{Y})^2$ *et* $SCT = \sum_{i=1}^{N}(Y_i - \bar{Y})^2$.

4.2.4 Tests de la nullité de la pente

Même si l'estimation de α est non nulle, il est nécessaire de savoir si sa différence avec 0 est significative [2].

Test de Student

Pour cela, on peut réaliser le test bilatéral $\begin{cases} H_0 : \alpha = 0 \\ H_1 : \alpha \neq 0 \end{cases}$ avec la variable de décision $\frac{A-\alpha}{S_a}$ qui suit $\mathcal{T}_{n-2}$ où $S_a^2 = \frac{S_E^2}{nS_X^2}$ est un estimateur de $Var(A)$.

Exercice.
En reprenant l'exemple du début de chapitre, estimer α par intervalle de confiance au seuil de 95%.

Remarque. La variable $\frac{B-\beta}{S_b}$ suit $\mathcal{T}_{n-2}$ où S_b^2 est un estimateur de $Var(B)$.
Comme dans l'exercice précédent, on peut estimer β par intervalle de confiance.

Test de Fisher

On peut aussi réaliser le test $\begin{cases} H_0 : \frac{SCE}{SCR/(n-2)} = 1 \\ H_1 : \frac{SCE}{SCR/(n-2)} > 1 \end{cases}$ qui revient à effectuer le test $\begin{cases} H_0 : \alpha = 0 \\ H_1 : \alpha \neq 0 \end{cases}$.

En effet, on montre que :
– Si $\alpha = 0$, alors $SCR/(n-2)$ et SCE sont deux estimateurs non biaisés de σ_E^2.
– Si $\alpha \neq 0$, alors $SCR/(n-2)$ reste sans biais mais SCE surestime σ_E^2.

La variable de décision étant $\frac{SCE}{SCR/(n-2)} = \frac{R^2}{(1-R^2)/(n-2)}$ qui suit $\mathcal{F}(1; n-2)$ sous H_0.

4.2.5 Intervalle de prévision

Notons $Y_0 = Y|(X = x_0)$ et $\hat{Y}_0 = \hat{Y}|(X = x_0)$ où x_0 est une valeur non observée de X.

On montre alors que

$$\frac{Y_0 - \hat{Y}_0}{S_E\sqrt{1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{ns_X^2}}} \sim \mathcal{T}_{n-2}$$

Par conséquent, l'intervalle de prévision pour y_0, de confiance [3] $1-\alpha$, est :

$$\left[\hat{y}_0 \pm t_{1-\frac{\alpha}{2}} s_E \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{ns_X^2}}\right]$$

Exercice.
En reprenant l'exemple du début de chapitre, déterminer l'intervalle de prévision du métabolisme pour un poids de 100 kg au seuil de 95%.

2. Si $\alpha = 0$, X et Y ne sont pas liées linéairement
3. Ne pas confondre le α précisant le niveau de confiance et la pente de la droite de régression

Chapitre 5

Analyse de variance

L'analyse de la variance utilise un vocabulaire spécifique : les variables qualitatives susceptibles d'influer sur la distribution de la variable quantitative étudiée sont appelées facteurs (ou facteurs de variabilité) et leurs modalités niveaux ou catégories.

5.1 Un facteur

Une importante entreprise agro-alimentaire cherche à optimiser le rendement de ses plantations de maïs. Trois variétés de maïs sont testées et plantées sur dix parcelles de deux hectares. Le tableau suivant indique les rendements obtenus. Le responsable de l'étude se demande si la variété de maïs a une influence sur le rendement. Pour identifier une éventuelle influence, le problème est simplifié : il s'agit de tester si le rendement moyen est différent selon la variété de maïs utilisée.

Variété 1	Variété 2	Variété 3
206	206	181
205	218	199
199	192	194
187	207	190
193	196	200
183	187	205
198	197	199
201	186	193
180	212	213
213	190	199
$\bar{x}_1 = 196,5$	$\bar{x}_2 = 199,1$	$\bar{x}_3 = 197,3$

L'ANOVA à un facteur réalisée par Excel fournit :

RAPPORT DÉTAILLÉ				
Groupes	Nombre[1] d'échantillons	Somme	Moyenne	Variance
Variété 1	10	1965	196,5	113,4
Variété 2	10	1991	199,1	122,1
Variété 3	10	1973	197,3	74,5

ANALYSE DE VARIANCE						
Source des variations	Somme des carrés	Degré de liberté	Moyenne des carrés	F[2]	Proba-bilité	Valeur critique pour F
Entre Groupes	35,467	2	17,733	0,172	0,84	3,35
A l'intérieur des groupes	2 789,500	27	103,315			
Total	2 824,967	29				

5.1.1 Hypothèses du modèle

L'analyse de la variance fait intervenir une variable quantitative mesurée sur plusieurs populations. Chaque population correspond à un niveau (une modalité) du facteur explicatif envisagé.

Les notations utilisées sont les suivantes :
- k va X_i (i allant de 1 à k), d'espérance m_i et d'écart-type σ_i. Les va X_i sont définies dans les mêmes termes mais sont mesurées sur k populations P_i.
- On tire un échantillon de taille n_i dans chaque population P_i. Ainsi, $(X_{i1}, ..., X_{in_i})$ est un n_i-échantillon issu de X_i.
- L'effectif total des échantillons est $n = \sum_{i=1}^{k} n_i$.

On teste $\begin{cases} H_0 : m_1 = ... = m_k \\ H_1 : \text{au moins deux des espérances sont différentes} \end{cases}$

Il convient ici de remarquer que si l'hypothèse nulle est retenue à l'issue du test, on considère que la variable qualitative (le facteur) définissant les populations n'a pas d'influence sur la variable quantitative.

Les va X_i sont supposées indépendantes et pour tout i, $X_i \sim \mathcal{N}(m_i; \sigma)$ où σ^2 est la variance commune des populations [1].

5.1.2 La méthode de l'ANOVA

La variance commune σ^2 joue un rôle fondamental car la méthode de l'ANOVA (ANalysis Of VAriance) utilise deux estimations de cette variance. Les deux estimateurs correspondants sont construits ci-après. Le premier repose sur la variabilité des observations à l'intérieur de chaque échantillon, le second mesure la variablité des moyennes entre les échantillons.

Estimation de la variance commune par la variance intra-échantillon

La variance intra-échantillon est définie par :

$$Var_{intra} = \frac{(n_1 - 1)S_1^{*2} + ... + (n_k - 1)S_k^{*2}}{(n_1 + ... + n_k) - k} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)^2}{n - k} = \frac{SCE}{n - k}$$

C'est un estimateur non biaisé de la variance commune σ^2 des populations.

Remarque. SCE désigne la Somme des Carrés des Erreurs

Estimation de la variance commune par la variance inter-échantillon

La variance inter-échantillon est définie par :

$$Var_{inter} = \frac{\sum_{i=1}^{k} n_i(\bar{X}_i - \bar{X})^2}{k - 1} = \frac{SCI_e}{k - 1}$$

où $\bar{X} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij}$ est la moyenne empirique générale.

Sous H_0, c'est un estimateur non biaisé de la variance commune σ^2 des populations.

1. la normalité et l'homoscédasticité (égalité des variances) peuvent être testées à l'aide d'un logiciel adapté

Variable de décision

Sous H_0, la variable de décision $\frac{SCI_e/(k-1)}{SCE/(n-k)}$ suit $\mathcal{F}(k-1; n-k)$ et le test est unilatéral à droite (cf. test de Fisher en régression linéaire).

Remarque. Si H_0 est rejetée, on pourra comparer les moyennes deux à deux [2].

Exercice.
En reprenant l'exemple, indiquer si la variété a un effet significatif sur le rendement au seuil de 5%.

5.2 Régression linéaire et analyse de variance à un facteur

5.2.1 Points communs

L'objectif de fond des deux analyses est d'identifier des facteurs explicatifs (et donc des leviers d'actions) de la variabilité d'une variable quantitative.

Par ailleurs, les deux approches reposent sur la décomposition de la somme des carrés totaux :
– Pour la régression, la décomposition $SCT = SCR + SCE$ est obtenue à l'aide de :

$$Y_i - \bar{Y} = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y})$$

– Pour l'ANOVA, la décomposition $SCT = SCE + SCI_e$ est obtenue à l'aide de :

$$X_{ij} - \bar{X} = (X_{ij} - \bar{X}_i) + (\bar{X}_i - \bar{X})$$

Les correspondances de notations entre la régression et l'ANOVA à un facteur sont consignées dans le tableau suivant :

Source de variation	Somme des carrés	Degrés de liberté	Variances	Variable de décision
Expliquée par la régression	*SCE* (notée SCI_e pour l'ANOVA)	1	*SCE*	$F_R = \frac{SCE}{SCR/n-2}$
Résidus ou erreurs	*SCR* (notée *SCE* pour l'ANOVA)	$n-2$	$\frac{SCR}{n-2}$	
Total	*SCT*	$n-1$	//	//

FIGURE 5.1 – Tableau ANOVA dans le cadre de la régression simple ($k = 2$)

5.2.2 Différences

Les différences de fond portent sur la nature des variables explicatives, ce qui a des répercussions sur la forme des résultats finaux.

Pour la régression, la variable explicative X est quantitative. Le fait de disposer de n couples d'observations à valeurs numériques (x_i, y_i) permet d'envisager la modélisation du lien (s'il existe) entre la variable expliquée Y et la variable explicative X par une fonction linéaire dont on cherche l'équation. Les hypothèses du modèle portent sur les termes d'erreur E_i. L'analyse de régression conduit à :
– confirmer ou infirmer l'hypothèse d'un effet de la variable explicative sur les variations de la variable expliquée
– donner, s'il y a lieu, l'équation de la fonction qui lie les deux variables
– tester a posteriori la qualité du modèle
– utiliser le modèle pour effectuer des prévisions

2. ou réaliser un test de comparaisons multiples de moyennes à l'aide d'un logiciel adapté

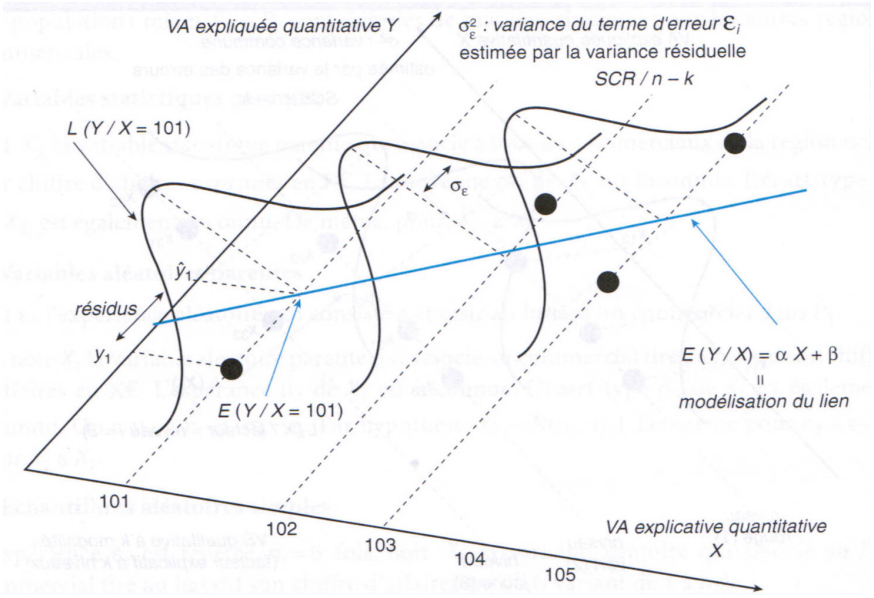


FIGURE 5.2 – Principales caractéristiques de la régression

Pour l'ANOVA, la variable explicative, appelée facteur, est qualitative. On dispose de n observations x_{ij} où l'indice i identifie le niveau (la modalité) du facteur explicatif. Le caractère qualitatif du facteur explicatif exclut toute tentative de modélisation par une fonction mathématique de l'influence éventuelle entre le facteur de variabilité et la variable espliquée X. Les hypothèses du modèle portent sur les va X_i. L'analyse de la variance conduit à confirmer ou infirmer l'hypothèse d'égalité des espérances m_i des va X_i. Autrement dit, à retenir ou non l'hyptohèse d'un impact statistiquement discernable du facteur explicatif sur les variations de la variable expliquée.

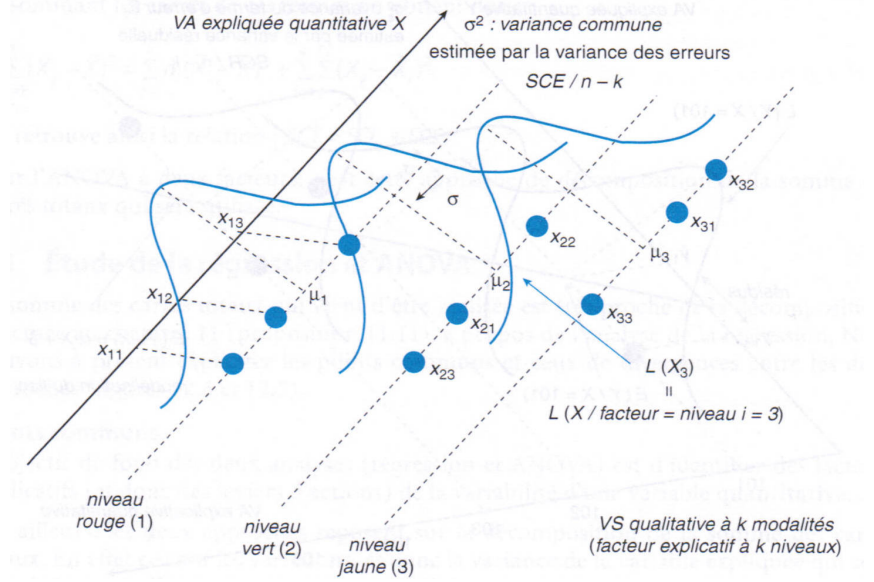


FIGURE 5.3 – Principales caractéristiques de l'ANOVA

5.3 Deux facteurs

5.3.1 Sans répétition d'expérience

Un exemple

Une chaîne d'hypermarchés vient d'installer, à titre d'essai, dans cinq magasins quelques caisses automatiques, c'est à dire des caisses où les clients enregistrent eux-mêmes le montant de leur achat. Trois méthodes sont à comparer : la caisse traditionnelle avec une caissière, la caisse automatique sans assistance et la caisse automatique avec l'assistance d'une hôtesse. Dans le tableau suivant, chaque observation est spécifique à une méthode et un magasin. La grandeur mesurée en secondes est le temps de passage aux caisses. La question est de savoir s'il est légitime d'étendre la pratique des caisses automatiques. Le seuil des tests est fixé à 5% et les temps de passage sont supposés gaussiens.

Magasin	Avec caissière (MA)	Sans caissière (MB)	Sans caissière avec assistance (MC)	Moyennes de catégories
A	190	226	200	$\bar{x}_{.1}=205,33$
B	195	250	170	$\bar{x}_{.2}=205,00$
C	188	212	180	$\bar{x}_{.3}=193,33$
D	156	170	190	$\bar{x}_{.4}=172,00$
E	164	220	200	$\bar{x}_{.5}=194,67$
Moyennes d'échantillons	$\bar{x}_{1.}=178,60$	$\bar{x}_{2.}=215,60$	$\bar{x}_{3.}=188,00$	$\bar{x}=194,07$

L'ANOVA à deux facteurs sans répétition d'expérience réalisée par Excel fournit :

ANALYSE DE VARIANCE						
Source des variations	Somme des carrés	Degré de liberté	Moyenne des carrés	F	Probabilité (*critique*)	Valeur critique pour *F*
Lignes (*magasin : inter-catégorie*)	2202,9	4	550,7	1,43	0,31	3,84
Colonnes (*méthode : inter-échantillon*)	3698,5	2	1849,3	4,79	0,04	4,46
Erreur (*SCE*)	3091,5	8	386,4			
Total (*SCT*)	8992,9	14				

La méthode

Soit k échantillons et $n = k \times h$ va X_{ij} issues de k populations présentant h catégories.

L'ANOVA à un facteur a été effectuée en décomposant la somme des carrés totaux en deux composantes. Nous allons ici la décomposer en trois.

De la décomposition $X_{ij} - \bar{X} = (\bar{X}_{i.} - \bar{X}) + (\bar{X}_{.j} - \bar{X}) + (X_{ij} - \bar{X}_{i.} - \bar{X}_{.j} + \bar{X})$, on tire :

$$\sum_{i=1}^{k}\sum_{j=1}^{h}(X_{ij} - \bar{X})^2 = h\sum_{i=1}^{k}(\bar{X}_{i.} - \bar{X})^2 + k\sum_{j=1}^{h}(\bar{X}_{.j} - \bar{X})^2 + \sum_{i=1}^{k}\sum_{j=1}^{h}(X_{ij} - \bar{X}_{i.} - \bar{X}_{.j} + \bar{X})^2$$

que l'on notera $SCT = SCI_e + SCI_c + SCE$.

Les sommes ayant pour degré de liberté, de gauche à droite : $n-1$, $k-1$, $h-1$ et $(k-1)(h-1)$.

Si les hypothèses de l'ANOVA sont réunies [3], alors :

3. les n va X_{ij} sont indépendantes, gaussiennes et de même variance

1. Pour la différence de moyennes inter-échantillon, la variable de décision est :

$$F_E = \frac{SCI_e/k-1}{SCE/(k-1)(h-1)} \sim \mathcal{F}(k-1,(k-1)(h-1))$$

2. Pour la différence de moyennes inter-catégorie, la variable de décision est :

$$F_C = \frac{SCI_c/h-1}{SCE/(k-1)(h-1)} \sim \mathcal{F}(h-1,(k-1)(h-1))$$

3. Pour chacun des deux tests unilatéraux à droite $\begin{cases} H_0 : \text{les espérances concernées sont toutes égales} \\ H_1 : \text{au moins deux des espérances sont différentes} \end{cases}$, la règle de décision au seuil de signification α est ...

Exercice.
En reprenant l'exemple :

1. Indiquer si le magasin (resp. la méthode) a un effet significatif sur le temps de passage aux caisses au seuil de 5%.
2. Est-il possible de privilégier la méthode A à la méthode C ?

5.3.2 Avec répétition d'expérience

Un exemple

Les serveurs des débits de boissons ont pour coutume d'effectuer des rotations entre la salle, la terrasse et le bar. Ces rotations de services ont lieu, suivant les établissements, à la journée ou de manière hebdomadaire. Cette tradition a pour but de ne pas désavantager les serveurs entre eux, car ceux-ci sont en général rémunérés à un taux fixe appliqué à leur chiffre d'affaires. Le patron d'un débit de boissons s'interroge sur l'opportunité de ces rotations. Il pense par ailleurs qu'un second facteur explicatif des différences de salaires de ses employés est lié à leur expérience. Pour étudier la pertinence de son intuition, il relève les données présentées dans le tableau suivant.
Le problème est de déterminer s'il existe des différences de chiffres d'affaires en tenant compte de deux facteurs (le "lieu de travail" et le "niveau d'expérience"). Pour détecter d'éventuelles interactions entre ces deux variables explicatives, deux mesures ont été relevées pour chaque échantillon et chaque catégorie.

	Salle	Terrasse	Bar	Moyenne par catégorie
sans sans	433 454	510 540	560 540	$\overline{x}_{\cdot 1\cdot} = 506,2$
faible faible	460 482	528 527	587 580	$\overline{x}_{\cdot 2\cdot} = 527,3$
moyenne moyenne	546 570	560 570	610 600	$\overline{x}_{\cdot 3\cdot} = 576,0$
assez élevée assez élevée	615 577	625 635	610 620	$\overline{x}_{\cdot 4\cdot} = 613,7$
élevée élevée	630 620	568 675	580 667	$\overline{x}_{\cdot 5\cdot} = 623,3$
Moyenne par échantillon	$\overline{x}_{1\cdot\cdot} = 538,7$	$\overline{x}_{2\cdot\cdot} = 573,8$	$\overline{x}_{3\cdot\cdot} = 595,4$	$\overline{\overline{x}} = 569,3$

L'ANOVA à deux facteurs avec répétition d'expérience réalisée par Excel fournit :

ANALYSE DE VARIANCE						
Source des variations	**Somme des carrés**	**Degré de liberté**	**Moyenne des carrés**	***F***	**Probabilité (*critique*)**	**Valeur critique pour *F* ($f_{0,95}$)**
Échantillon (*statut : inter-catégorie* SCI_c)	64 079,5	4	16 019,9	20,18	0,00	3,06
Colonnes (*région : inter-échantillon* SCI_e)	16 378,2	2	8 189,1	10,32	0,00	3,68
Intéraction (SCI_{ec})	12 430,1	8	1 553,8	1,96	0,12	2,64
À l'intérieur du groupe (*erreur SCE*)	11 906,5	15	793,8			
Total (*SCT*)	10 4794,3	29				

La méthode

Soit k échantillons et $n = k \times h \times g$ va X_{ijl} issues de k populations présentant h catégories et pour lesquelles g réalisations sont observées.

L'ANOVA à deux facteurs sans répétition a été effectuée en décomposant la somme des carrés totaux en trois composantes. Nous allons ici la décomposer en quatre.

De $X_{ijl} - \bar{X} = (\bar{X}_{i.} - \bar{X}) + (\bar{X}_{.j} - \bar{X}) + (\bar{X}_{ij} - \bar{X}_{i.} - \bar{X}_{.j} + \bar{X}) + (X_{ijl} - \bar{X}_{ij})$, on tire :

$$\sum_{i=1}^{k}\sum_{j=1}^{h}\sum_{l=1}^{g}(X_{ijl}-\bar{X})^2 = hg\sum_{i=1}^{k}(\bar{X}_{i.}-\bar{X})^2+kg\sum_{j=1}^{h}(\bar{X}_{.j}-\bar{X})^2+g\sum_{i=1}^{k}\sum_{j=1}^{h}(\bar{X}_{ij}-\bar{X}_{i.}-\bar{X}_{.j}+\bar{X})^2+\sum_{i=1}^{k}\sum_{j=1}^{h}\sum_{l=1}^{g}(X_{ijl}-\bar{X}_{ij})^2$$

que l'on notera $SCT = SCI_e + SCI_c + SCI_{ec} + SCE$.

Les sommes ayant pour degré de liberté, de gauche à droite : $n-1$, $k-1$, $h-1$, $(k-1)(h-1)$ et $kh(g-1)$.

Si les hypothèses de l'ANOVA sont réunies [4], alors :

1. Pour la différence de moyennes inter-échantillon, la variable de décision est :
$$F_E = \frac{SCI_e/k-1}{SCE/kh(g-1)} \sim \mathcal{F}(k-1, kh(g-1))$$
2. Pour la différence de moyennes inter-catégorie, la variable de décision est :
$$F_C = \frac{SCI_c/h-1}{SCE/kh(g-1)} \sim \mathcal{F}(h-1, kh(g-1))$$
3. Pour la différence de moyennes due aux interactions entre les échantillons et les catégories, la variable de décision est :
$$F_{EC} = \frac{SCI_{ec}/(k-1)(h-1)}{SCE/kh(g-1)} \sim \mathcal{F}((k-1)(h-1), kh(g-1))$$
4. Pour chacun des trois tests unilatéraux à droite $\begin{cases} H_0 : \text{les espérances concernées sont toutes égales} \\ H_1 : \text{au moins deux des espérances sont différentes} \end{cases}$, la règle de décision au seuil de signification α est ...

Exercice.
En reprenant l'exemple, indiquer si l'interaction entre le lieu de travail et le niveau d'expérience (resp. le lieu de travail puis le niveau d'expérience) a un effet significatif sur le chiffre d'affaire au seuil de 5%.

4. les n va $X_{ij.}$ sont indépendantes, gaussiennes et de même variance

Remarque. Si l'interaction a un effet significatif, on remplacera dans F_E (resp. F_C), $SCE/kh(g-1)$ par $SCI_{ec}/(k-1)(h-1)$ pour savoir si l'échantillon (resp. la catégorie) a un effet significatif en plus de celui de l'interaction déjà mis en évidence.

Exercice.
Le rendement d'une réaction chimique réalisée dans des conditions industrielles pourrait dépendre de la température et de la pression. Afin de s'en assurer, une étude est conduite en adoptant trois valeurs de température (facteur A) et deux valeurs de pression (facteur B). La grandeur mesurée est la quantité de produit chimique formée, exprimée en excès par rapport à la plus petite valeur obtenue (celle du couple 225 °C, 50 bars).
Que pouvons-nous conclure à partir des résultats suivants :

	225	250	275
50	0	28	65
	10	32	65
60	42	58	75
	38	62	75

Annexes

Annexe 1

Une étude réalisée sur les étudiants d'une promotion HEI2 décrits par leur sexe et leur moyenne de fin d'année dans le GR1 fournit les résultas suivants :

		GR 1			Moyenne		
Moyenne	M.	10,3684733		Population	10,6323641		
	Mlle	11,2846226					
Ecart-type	M.	2,02032477					
	Mlle	2,33281115		n/N	0,06		
		nHsH GR1			Population	Echantillon prop	Echantillon optim GR1
	M.	529,32509		M.	262	16	15
	Mlle	247,277982		Mlle	106	6	7
	Total	776,603072		Total	368	22	22
		Moyenne					
	EAS	9,38772727			Variance		
				Population M	4,08171217		
		Moyenne					
	ESPG	9,909375			EASG26		
	ESPF	11,48		Variance	2,73489364		
	ESP	10,3377273		Variance c	2,84428938		
		Moyenne			26	52	78
	ESOG	10,646		Variance c	2,84428938	5,416586425	4,330231211
	ESOF	11,2114286		d	2,80471774	5,313416717	4,28327036
	ESO	10,8259091					

On peut remarquer que : ...

Table 3

Loi Normale Centrée Réduite

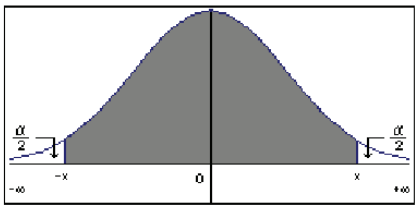
Fonction de répartition F(z)=P(Z<z)

Exemple : P(Z<1.96)= 0.97500 se trouve en ligne 1.9 et colonne 0.06

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,50000	0,50399	0,50798	0,51197	0,51595	0,51994	0,52392	0,52790	0,53188	0,53586
0,1	0,53983	0,54380	0,54776	0,55172	0,55567	0,55962	0,56356	0,56750	0,57142	0,57535
0,2	0,57926	0,58317	0,58706	0,59095	0,59484	0,59871	0,60257	0,60642	0,61026	0,61409
0,3	0,61791	0,62172	0,62552	0,62930	0,63307	0,63683	0,64058	0,64431	0,64803	0,65173
0,4	0,65542	0,65910	0,66276	0,66640	0,67003	0,67365	0,67724	0,68082	0,68439	0,68793
0,5	0,69146	0,69498	0,69847	0,70194	0,70540	0,70884	0,71226	0,71566	0,71904	0,72241
0,6	0,72575	0,72907	0,73237	0,73565	0,73891	0,74215	0,74537	0,74857	0,75175	0,75490
0,7	0,75804	0,76115	0,76424	0,76731	0,77035	0,77337	0,77637	0,77935	0,78231	0,78524
0,8	0,78815	0,79103	0,79389	0,79673	0,79955	0,80234	0,80511	0,80785	0,81057	0,81327
0,9	0,81594	0,81859	0,82121	0,82382	0,82639	0,82894	0,83147	0,83398	0,83646	0,83891
1,0	0,84135	0,84375	0,84614	0,84850	0,85083	0,85314	0,85543	0,85769	0,85993	0,86214
1,1	0,86433	0,86650	0,86864	0,87076	0,87286	0,87493	0,87698	0,87900	0,88100	0,88298
1,2	0,88493	0,88686	0,88877	0,89065	0,89251	0,89435	0,89617	0,89796	0,89973	0,90148
1,3	0,90320	0,90490	0,90658	0,90824	0,90988	0,91149	0,91309	0,91466	0,91621	0,91774
1,4	0,91924	0,92073	0,92220	0,92364	0,92507	0,92647	0,92786	0,92922	0,93056	0,93189
1,5	0,93319	0,93448	0,93575	0,93699	0,93822	0,93943	0,94062	0,94179	0,94295	0,94408
1,6	0,94520	0,94630	0,94738	0,94845	0,94950	0,95053	0,95154	0,95254	0,95352	0,95449
1,7	0,95544	0,95637	0,95728	0,95819	0,95907	0,95994	0,96080	0,96164	0,96246	0,96327
1,8	0,96407	0,96485	0,96562	0,96638	0,96712	0,96784	0,96856	0,96926	0,96995	0,97062
1,9	0,97128	0,97193	0,97257	0,97320	0,97381	0,97441	0,97500	0,97558	0,97615	0,97670
2,0	0,97725	0,97778	0,97831	0,97882	0,97933	0,97982	0,98030	0,98077	0,98124	0,98169
2,1	0,98214	0,98257	0,98300	0,98341	0,98382	0,98422	0,98461	0,98500	0,98537	0,98574
2,2	0,98610	0,98645	0,98679	0,98713	0,98745	0,98778	0,98809	0,98840	0,98870	0,98899
2,3	0,98928	0,98956	0,98983	0,99010	0,99036	0,99061	0,99086	0,99111	0,99134	0,99158
2,4	0,99180	0,99202	0,99224	0,99245	0,99266	0,99286	0,99305	0,99324	0,99343	0,99361
2,5	0,99379	0,99396	0,99413	0,99430	0,99446	0,99461	0,99477	0,99492	0,99506	0,99520
2,6	0,99534	0,99547	0,99560	0,99573	0,99585	0,99598	0,99609	0,99621	0,99632	0,99643
2,7	0,99653	0,99664	0,99674	0,99683	0,99693	0,99702	0,99711	0,99720	0,99728	0,99736
2,8	0,99744	0,99752	0,99760	0,99767	0,99774	0,99781	0,99788	0,99795	0,99801	0,99807
2,9	0,99813	0,99819	0,99825	0,99831	0,99836	0,99841	0,99846	0,99851	0,99856	0,99861
3,0	0,99865	0,99869	0,99874	0,99878	0,99882	0,99886	0,99889	0,99893	0,99897	0,99900
3,1	0,99903	0,99906	0,99910	0,99913	0,99916	0,99918	0,99921	0,99924	0,99926	0,99929
3,2	0,99931	0,99934	0,99936	0,99938	0,99940	0,99942	0,99944	0,99946	0,99948	0,99950
3,3	0,99952	0,99953	0,99955	0,99957	0,99958	0,99960	0,99961	0,99962	0,99964	0,99965
3,4	0,99966	0,99968	0,99969	0,99970	0,99971	0,99972	0,99973	0,99974	0,99975	0,99976

Table 4

Loi de Student



α	1	0,8	0,6	0,4	0,2	0,1	0,05	0,02	0,01	0,002	0,001
1 − α	0	0,2	0,4	0,6	0,8	0,9	0,95	0,98	0,99	0,998	0,999
ν = ddl											
1	0,0000	0,3249	0,7265	1,3764	3,0777	6,3137	12,706	31,821	63,656	318,29	636,58
2	0,0000	0,2887	0,6172	1,0607	1,8856	2,9200	4,3027	6,9645	9,9250	22,328	31,600
3	0,0000	0,2767	0,5844	0,9785	1,6377	2,3534	3,1824	4,5407	5,8408	10,214	12,924
4	0,0000	0,2707	0,5686	0,9410	1,5332	2,1318	2,7765	3,7469	4,6041	7,1729	8,6101
5	0,0000	0,2672	0,5594	0,9195	1,4759	2,0150	2,5706	3,3649	4,0321	5,8935	6,8685
6	0,0000	0,2648	0,5534	0,9057	1,4398	1,9432	2,4469	3,1427	3,7074	5,2075	5,9587
7	0,0000	0,2632	0,5491	0,8960	1,4149	1,8946	2,3646	2,9979	3,4995	4,7853	5,4081
8	0,0000	0,2619	0,5459	0,8889	1,3968	1,8595	2,3060	2,8965	3,3554	4,5008	5,0414
9	0,0000	0,2610	0,5435	0,8834	1,3830	1,8331	2,2622	2,8214	3,2498	4,2969	4,7809
10	0,0000	0,2602	0,5415	0,8791	1,3722	1,8125	2,2281	2,7638	3,1693	4,1437	4,5868
11	0,0000	0,2596	0,5399	0,8755	1,3634	1,7959	2,2010	2,7181	3,1058	4,0248	4,4369
12	0,0000	0,2590	0,5386	0,8726	1,3562	1,7823	2,1788	2,6810	3,0545	3,9296	4,3178
13	0,0000	0,2586	0,5375	0,8702	1,3502	1,7709	2,1604	2,6503	3,0123	3,8520	4,2209
14	0,0000	0,2582	0,5366	0,8681	1,3450	1,7613	2,1448	2,6245	2,9768	3,7874	4,1403
15	0,0000	0,2579	0,5357	0,8662	1,3406	1,7531	2,1315	2,6025	2,9467	3,7329	4,0728
16	0,0000	0,2576	0,5350	0,8647	1,3368	1,7459	2,1199	2,5835	2,9208	3,6861	4,0149
17	0,0000	0,2573	0,5344	0,8633	1,3334	1,7396	2,1098	2,5669	2,8982	3,6458	3,9651
18	0,0000	0,2571	0,5338	0,8620	1,3304	1,7341	2,1009	2,5524	2,8784	3,6105	3,9217
19	0,0000	0,2569	0,5333	0,8610	1,3277	1,7291	2,0930	2,5395	2,8609	3,5793	3,8833
20	0,0000	0,2567	0,5329	0,8600	1,3253	1,7247	2,0860	2,5280	2,8453	3,5518	3,8496
21	0,0000	0,2566	0,5325	0,8591	1,3232	1,7207	2,0796	2,5176	2,8314	3,5271	3,8193
22	0,0000	0,2564	0,5321	0,8583	1,3212	1,7171	2,0739	2,5083	2,8188	3,5050	3,7922
23	0,0000	0,2563	0,5317	0,8575	1,3195	1,7139	2,0687	2,4999	2,8073	3,4850	3,7676
24	0,0000	0,2562	0,5314	0,8569	1,3178	1,7109	2,0639	2,4922	2,7970	3,4668	3,7454
25	0,0000	0,2561	0,5312	0,8562	1,3163	1,7081	2,0595	2,4851	2,7874	3,4502	3,7251
26	0,0000	0,2560	0,5309	0,8557	1,3150	1,7056	2,0555	2,4786	2,7787	3,4350	3,7067
27	0,0000	0,2559	0,5306	0,8551	1,3137	1,7033	2,0518	2,4727	2,7707	3,4210	3,6895
28	0,0000	0,2558	0,5304	0,8546	1,3125	1,7011	2,0484	2,4671	2,7633	3,4082	3,6739
29	0,0000	0,2557	0,5302	0,8542	1,3114	1,6991	2,0452	2,4620	2,7564	3,3963	3,6595
30	0,0000	0,2556	0,5300	0,8538	1,3104	1,6973	2,0423	2,4573	2,7500	3,3852	3,6460
40	0,0000	0,2550	0,5286	0,8507	1,3031	1,6839	2,0211	2,4233	2,7045	3,3069	3,5510
50	0,0000	0,2547	0,5278	0,8489	1,2987	1,6759	2,0086	2,4033	2,6778	3,2614	3,4960
60	0,0000	0,2545	0,5272	0,8477	1,2958	1,6706	2,0003	2,3901	2,6603	3,2317	3,4602
70	0,0000	0,2543	0,5268	0,8468	1,2938	1,6669	1,9944	2,3808	2,6479	3,2108	3,4350
80	0,0000	0,2542	0,5265	0,8461	1,2922	1,6641	1,9901	2,3739	2,6387	3,1952	3,4164
90	0,0000	0,2541	0,5263	0,8456	1,2910	1,6620	1,9867	2,3685	2,6316	3,1832	3,4019
100	0,0000	0,2540	0,5261	0,8452	1,2901	1,6602	1,9840	2,3642	2,6259	3,1738	3,3905
200	0,0000	0,2537	0,5252	0,8434	1,2858	1,6525	1,9719	2,3451	2,6006	3,1315	3,3398
∞	0,0000	0,2533	0,5244	0,8416	1,2816	1,6449	1,9600	2,3263	2,5758	3,0903	3,2906

Table 5

Loi du χ^2

$$P(\chi_\nu^2 \geq \chi_{\nu,\alpha}^2) = \alpha$$

$1-\alpha$	0,001	0,005	0,01	0,025	0,05	0,1	0,5	0,9	0,95	0,975	0,99	0,995	0,999
α	0,999	0,995	0,99	0,975	0,95	0,9	0,5	0,1	0,05	0,025	0,01	0,005	0,001
ν = ddl													
1	0,00	0,00	0,00	0,00	0,00	0,02	0,45	2,71	3,84	5,02	6,63	7,88	10,83
2	0,00	0,01	0,02	0,05	0,10	0,21	1,39	4,61	5,99	7,38	9,21	10,60	13,82
3	0,02	0,07	0,11	0,22	0,35	0,58	2,37	6,25	7,81	9,35	11,34	12,84	16,27
4	0,09	0,21	0,30	0,48	0,71	1,06	3,36	7,78	9,49	11,14	13,28	14,86	18,47
5	0,21	0,41	0,55	0,83	1,15	1,61	4,35	9,24	11,07	12,83	15,09	16,75	20,51
6	0,38	0,68	0,87	1,24	1,64	2,20	5,35	10,64	12,59	14,45	16,81	18,55	22,46
7	0,60	0,99	1,24	1,69	2,17	2,83	6,35	12,02	14,07	16,01	18,48	20,28	24,32
8	0,86	1,34	1,65	2,18	2,73	3,49	7,34	13,36	15,51	17,53	20,09	21,95	26,12
9	1,15	1,73	2,09	2,70	3,33	4,17	8,34	14,68	16,92	19,02	21,67	23,59	27,88
10	1,48	2,16	2,56	3,25	3,94	4,87	9,34	15,99	18,31	20,48	23,21	25,19	29,59
11	1,83	2,60	3,05	3,82	4,57	5,58	10,34	17,28	19,68	21,92	24,73	26,76	31,26
12	2,21	3,07	3,57	4,40	5,23	6,30	11,34	18,55	21,03	23,34	26,22	28,30	32,91
13	2,62	3,57	4,11	5,01	5,89	7,04	12,34	19,81	22,36	24,74	27,69	29,82	34,53
14	3,04	4,07	4,66	5,63	6,57	7,79	13,34	21,06	23,68	26,12	29,14	31,32	36,12
15	3,48	4,60	5,23	6,26	7,26	8,55	14,34	22,31	25,00	27,49	30,58	32,80	37,70
16	3,94	5,14	5,81	6,91	7,96	9,31	15,34	23,54	26,30	28,85	32,00	34,27	39,25
17	4,42	5,70	6,41	7,56	8,67	10,09	16,34	24,77	27,59	30,19	33,41	35,72	40,79
18	4,90	6,26	7,01	8,23	9,39	10,86	17,34	25,99	28,87	31,53	34,81	37,16	42,31
19	5,41	6,84	7,63	8,91	10,12	11,65	18,34	27,20	30,14	32,85	36,19	38,58	43,82
20	5,92	7,43	8,26	9,59	10,85	12,44	19,34	28,41	31,41	34,17	37,57	40,00	45,31
21	6,45	8,03	8,90	10,28	11,59	13,24	20,34	29,62	32,67	35,48	38,93	41,40	46,80
22	6,98	8,64	9,54	10,98	12,34	14,04	21,34	30,81	33,92	36,78	40,29	42,80	48,27
23	7,53	9,26	10,20	11,69	13,09	14,85	22,34	32,01	35,17	38,08	41,64	44,18	49,73
24	8,08	9,89	10,86	12,40	13,85	15,66	23,34	33,20	36,42	39,36	42,98	45,56	51,18
25	8,65	10,52	11,52	13,12	14,61	16,47	24,34	34,38	37,65	40,65	44,31	46,93	52,62
26	9,22	11,16	12,20	13,84	15,38	17,29	25,34	35,56	38,89	41,92	45,64	48,29	54,05
27	9,80	11,81	12,88	14,57	16,15	18,11	26,34	36,74	40,11	43,19	46,96	49,65	55,48
28	10,39	12,46	13,56	15,31	16,93	18,94	27,34	37,92	41,34	44,46	48,28	50,99	56,89
29	10,99	13,12	14,26	16,05	17,71	19,77	28,34	39,09	42,56	45,72	49,59	52,34	58,30
30	11,59	13,79	14,95	16,79	18,49	20,60	29,34	40,26	43,77	46,98	50,89	53,67	59,70

Pour $\nu > 30$, La loi du χ^2 peut –être approximée par la loi normale $N(\nu, \sqrt{\nu})$

Table 6

Loi de Fisher F

$$P(F_{\nu_1,\nu_2} < \mathsf{f}_{\nu_1,\nu_2,\alpha}) = \alpha$$

$\alpha = 0{,}975$

ν_2 \ ν_1	1	2	3	4	5	6	7	8	9	10	15	20	30	50	100	200	500	•
1	648	800	864	900	922	937	948	957	963	969	985	993	1001	1008	1013	1016	1017	1018
2	38,5	39,0	39,2	39,3	39,3	39,4	39,4	39,4	39,4	39,4	39,4	39,4	39,5	39,5	39,5	39,5	39,5	39,5
3	17,4	16,0	15,4	15,1	14,9	14,7	14,6	14,5	14,5	14,4	14,3	14,2	14,1	14,0	14,0	13,9	13,9	13,9
4	12,2	10,6	9,98	9,60	9,36	9,20	9,07	8,98	8,90	8,84	8,66	8,56	8,46	8,38	8,32	8,29	8,27	8,26
5	10,0	8,43	7,76	7,39	7,15	6,98	6,85	6,76	6,68	6,62	6,43	6,33	6,23	6,14	6,08	6,05	6,03	6,02
6	8,81	7,26	6,60	6,23	5,99	5,82	5,70	5,60	5,52	5,46	5,27	5,17	5,07	4,98	4,92	4,88	4,86	4,85
7	8,07	6,54	5,89	5,52	5,29	5,12	4,99	4,90	4,82	4,76	4,57	4,47	4,36	4,28	4,21	4,18	4,16	4,14
8	7,57	6,06	5,42	5,05	4,82	4,65	4,53	4,43	4,36	4,30	4,10	4,00	3,89	3,81	3,74	3,70	3,68	3,67
9	7,21	5,71	5,08	4,72	4,48	4,32	4,20	4,10	4,03	3,96	3,77	3,67	3,56	3,47	3,40	3,37	3,35	3,33
10	6,94	5,46	4,83	4,47	4,24	4,07	3,95	3,85	3,78	3,72	3,52	3,42	3,31	3,22	3,15	3,12	3,09	3,08
11	6,72	5,26	4,63	4,28	4,04	3,88	3,76	3,66	3,59	3,53	3,33	3,23	3,12	3,03	2,96	2,92	2,90	2,88
12	6,55	5,10	4,47	4,12	3,89	3,73	3,61	3,51	3,44	3,37	3,18	3,07	2,96	2,87	2,80	2,76	2,74	2,72
13	6,41	4,97	4,35	4,00	3,77	3,60	3,48	3,39	3,31	3,25	3,05	2,95	2,84	2,74	2,67	2,63	2,61	2,60
14	6,30	4,86	4,24	3,89	3,66	3,50	3,38	3,29	3,21	3,15	2,95	2,84	2,73	2,64	2,56	2,53	2,50	2,49
15	6,20	4,76	4,15	3,80	3,58	3,41	3,29	3,20	3,12	3,06	2,86	2,76	2,64	2,55	2,47	2,44	2,41	2,40
16	6,12	4,69	4,08	3,73	3,50	3,34	3,22	3,12	3,05	2,99	2,79	2,68	2,57	2,47	2,40	2,36	2,33	2,32
17	6,04	4,62	4,01	3,66	3,44	3,28	3,16	3,06	2,98	2,92	2,72	2,62	2,50	2,41	2,33	2,29	2,26	2,25
18	5,98	4,56	3,95	3,61	3,38	3,22	3,10	3,01	2,93	2,87	2,67	2,56	2,44	2,35	2,27	2,23	2,20	2,19
19	5,92	4,51	3,90	3,56	3,33	3,17	3,05	2,96	2,88	2,82	2,62	2,51	2,39	2,30	2,22	2,18	2,15	2,13
20	5,87	4,46	3,86	3,51	3,29	3,13	3,01	2,91	2,84	2,77	2,57	2,46	2,35	2,25	2,17	2,13	2,10	2,09
22	5,79	4,38	3,78	3,44	3,22	3,05	2,93	2,84	2,76	2,70	2,50	2,39	2,27	2,17	2,09	2,05	2,02	2,00
24	5,72	4,32	3,72	3,38	3,15	2,99	2,87	2,78	2,70	2,64	2,44	2,33	2,21	2,11	2,02	1,98	1,95	1,94
26	5,66	4,27	3,67	3,33	3,10	2,94	2,82	2,73	2,65	2,59	2,39	2,28	2,16	2,05	1,97	1,92	1,90	1,88
28	5,61	4,22	3,63	3,29	3,06	2,90	2,78	2,69	2,61	2,55	2,34	2,23	2,11	2,01	1,92	1,88	1,85	1,83
30	5,57	4,18	3,59	3,25	3,03	2,87	2,75	2,65	2,57	2,51	2,31	2,20	2,07	1,97	1,88	1,84	1,81	1,79
40	5,42	4,05	3,46	3,13	2,90	2,74	2,62	2,53	2,45	2,39	2,18	2,07	1,94	1,83	1,74	1,69	1,66	1,64
50	5,34	3,98	3,39	3,06	2,83	2,67	2,55	2,46	2,38	2,32	2,11	1,99	1,87	1,75	1,66	1,60	1,57	1,55
60	5,29	3,93	3,34	3,01	2,79	2,63	2,51	2,41	2,33	2,27	2,06	1,94	1,82	1,70	1,60	1,54	1,51	1,48
80	5,22	3,86	3,28	2,95	2,73	2,57	2,45	2,36	2,28	2,21	2,00	1,88	1,75	1,63	1,53	1,47	1,43	1,40
100	5,18	3,83	3,25	2,92	2,70	2,54	2,42	2,32	2,24	2,18	1,97	1,85	1,71	1,59	1,48	1,42	1,38	1,35
200	5,10	3,76	3,18	2,85	2,63	2,47	2,35	2,26	2,18	2,11	1,90	1,78	1,64	1,51	1,39	1,32	1,27	1,23
500	5,05	3,72	3,14	2,81	2,59	2,43	2,31	2,22	2,14	2,07	1,86	1,74	1,60	1,46	1,34	1,25	1,19	1,14
•	5,02	3,69	3,12	2,79	2,57	2,41	2,29	2,19	2,11	2,05	1,83	1,71	1,57	1,43	1,30	1,21	1,13	1,00

Loi de Fisher F (suite)

$$P(F_{\nu_1,\nu_2} < \mathsf{f}_{\nu_1,\nu_2,\alpha}) = \alpha$$

$\alpha = 0{,}95$

ν_2 \ ν_1	1	2	3	4	5	6	7	8	9	10	15	20	30	50	100	200	500	•
1	161	200	216	225	230	234	237	239	241	242	246	248	250	252	253	254	254	254
2	18,5	19,0	19,2	19,2	19,3	19,3	19,4	19,4	19,4	19,4	19,4	19,4	19,5	19,5	19,5	19,5	19,5	19,5
3	10,1	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79	8,70	8,66	8,62	8,58	8,55	8,54	8,53	8,53
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,86	5,80	5,75	5,70	5,66	5,65	5,64	5,63
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74	4,62	4,56	4,50	4,44	4,41	4,39	4,37	4,37
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	3,94	3,87	3,81	3,75	3,71	3,69	3,68	3,67
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	3,51	3,44	3,38	3,32	3,27	3,25	3,24	3,23
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35	3,22	3,15	3,08	3,02	2,97	2,95	2,94	2,93
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14	3,01	2,94	2,86	2,80	2,76	2,73	2,72	2,71
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98	2,85	2,77	2,70	2,64	2,59	2,56	2,55	2,54
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85	2,72	2,65	2,57	2,51	2,46	2,43	2,42	2,40
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75	2,62	2,54	2,47	2,40	2,35	2,32	2,31	2,30
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67	2,53	2,46	2,38	2,31	2,26	2,23	2,22	2,21
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60	2,46	2,39	2,31	2,24	2,19	2,16	2,14	2,13
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54	2,40	2,33	2,25	2,18	2,12	2,10	2,08	2,07
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,35	2,28	2,19	2,12	2,07	2,04	2,02	2,01
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45	2,31	2,23	2,15	2,08	2,02	1,99	1,97	1,96
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41	2,27	2,19	2,11	2,04	1,98	1,95	1,93	1,92
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38	2,23	2,16	2,07	2,00	1,94	1,91	1,89	1,88
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35	2,20	2,12	2,04	1,97	1,91	1,88	1,86	1,84
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30	2,15	2,07	1,98	1,91	1,85	1,82	1,80	1,78
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25	2,11	2,03	1,94	1,86	1,80	1,77	1,75	1,73
26	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22	2,07	1,99	1,90	1,82	1,76	1,73	1,71	1,69
28	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24	2,19	2,04	1,96	1,87	1,79	1,73	1,69	1,67	1,65
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16	2,01	1,93	1,84	1,76	1,70	1,66	1,64	1,62
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08	1,92	1,84	1,74	1,66	1,59	1,55	1,53	1,51
50	4,03	3,18	2,79	2,56	2,40	2,29	2,20	2,13	2,07	2,03	1,87	1,78	1,69	1,60	1,52	1,48	1,46	1,44
60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99	1,84	1,75	1,65	1,56	1,48	1,44	1,41	1,39
80	3,96	3,11	2,72	2,49	2,33	2,21	2,13	2,06	2,00	1,95	1,79	1,70	1,60	1,51	1,43	1,38	1,35	1,32
100	3,94	3,09	2,70	2,46	2,31	2,19	2,10	2,03	1,97	1,93	1,77	1,68	1,57	1,48	1,39	1,34	1,31	1,28
200	3,89	3,04	2,65	2,42	2,26	2,14	2,06	1,98	1,93	1,88	1,72	1,62	1,52	1,41	1,32	1,26	1,22	1,19
500	3,86	3,01	2,62	2,39	2,23	2,12	2,03	1,96	1,90	1,85	1,69	1,59	1,48	1,38	1,28	1,21	1,16	1,11
•	3,84	3,00	2,60	2,37	2,21	2,10	2,01	1,94	1,88	1,83	1,67	1,57	1,46	1,35	1,24	1,17	1,11	1,00

Bibliographie

[SAP] G. SAPORTA, *Probabilités, analyse des données et Statistique*, TECHNIP, 2006
[TRI] B. TRIBOUT, *Statistique pour économistes et gestionnaires*, PEARSON, 2007