UNIVERSITE MOHAMED V – AGDAL
Faculté des Sciences Juridiques, Economiques et Sociales
Filière des Sciences Economiques et Gestion

Semestre : IV

Sections : A, B et C

Module : Méthodes Quantitatives III

Matière : STATISTIQUE III

Session : printemps été 2011

Responsable de la matière : Adil ELMARHOUM

RAPPELS STATISTIQUES II

NOTION DE VARIABLES ALEATOIRES

I. DEFINITION

Une variable aléatoire X est une variable associée à une expérience ou à un groupe d'expériences aléatoires et servant à caractériser le résultat de cette expérience ou de ce groupe d'expériences.

On distingue les variables aléatoires discontinues ou discrètes et les variables aléatoires continues.

II. VARIABLE ALEATOIRE DISCONTINUE

2.1. Définition

Une variable aléatoire est discrète si elle varie de façon discontinue, la variable ne peut prendre que des valeurs entières.

Exemple :

- Soit X la variable aléatoire qui caractérise le résultat de l'expérience aléatoire "jet d'un dé homogène".

X est une variable aléatoire discrète, elle peut prendre les valeurs entières 1, 2, 3, 4, 5, et 6.

- Soit X la variable aléatoire qui caractérise le nombre de garçons dans une famille de quatre enfants.

X est une variable aléatoire discrète, elle peut prendre les valeurs entières 0, 1, 2, 3, et 4.

2.2. Distribution de probabilité

À chacune des valeurs x que peut prendre une variable aléatoire X, correspond une probabilité p(x), c'est la probabilité que la variable aléatoire X prenne la valeur x :

$$p(x) = p(X = x)$$

L'ensemble des valeurs admissibles x et des probabilités correspondantes p(x) constitue une distribution de probabilité discontinue. La relation entre x et p(x) est appelée loi de probabilité.

Pour toutes les distributions de probabilités dont les valeurs x correspondent à des événements complémentaires, le total des probabilités est égal à 1.

$$\sum p(x) = 1$$

La distribution cumulée des probabilités est appelée fonction de répartition :

$$F(x) = p(X \leq x) = \sum^{x} p(x)$$

$$0 \leq F(x) \leq 1$$

Exemple :

Soit X la variable aléatoire qui caractérise le résultat de l'expérience aléatoire "jet d'un dé homogène".

X est une variable aléatoire discrète, elle peut prendre les valeurs entières 1, 2, 3, 4, 5, et 6 avec la probabilité constante 1/6.

Distribution de probabilité de X

x	p(x)	F(x)
1	1/6	1/6
2	1/6	2/6
3	1/6	3/6
4	1/6	4/6
5	1/6	5/6
6	1/6	6/6
Total	1	

III. VARIABLE ALEATOIRE CONTINUE

Une variable aléatoire est continue si elle prend n'importe quelle valeur réelle appartenant à un intervalle donné.

Exemple :

Le poids est une variable aléatoire continue.
La taille est une variable aléatoire continue.

Un intervalle continu contient une infinité de valeurs. La probabilité d'obtenir exactement un résultat donné est généralement nulle, bien que ce résultat ne soit pas strictement impossible.

$$p(X = x) \approx 0$$

La notion de distribution de probabilité n'a donc plus de sens dans le cas continu. Par contre la fonction de répartition conserve toute sa signification.

Pour une variable aléatoire continue, on calcule la probabilité d'observer une valeur comprise dans un intervalle donné [x ; x+Δx].

$$p(x \le X \le x+\Delta x) = p(X \le x+\Delta x) - p(X \le x) = F(x+\Delta x) - F(x)$$

Cette probabilité tend vers p(x) quand Δx tend vers 0.

$$\lim_{\Delta x \to 0} p(x \le X \le x+\Delta x) = \lim_{\Delta x \to 0} F(x+\Delta x) - F(x)$$

$$\lim_{\Delta x \to 0} \frac{F(x+\Delta x) - F(x)}{\Delta x} = \lim_{\Delta x \to 0} \frac{\Delta F}{\Delta x} = \frac{dF}{dx} = F'(x) = f(x)$$

La fonction f(x), dérivée de la fonction de répartition F(x), est appelée fonction de densité de probabilité.

L'ensemble des valeurs admissibles pour une variable aléatoire continue et la fonction de densité de probabilité correspondante définissent une distribution de probabilité théorique continue.

Le produit f(x)dx est appelé élément de probabilité, c'est l'équivalent de la probabilité p(x) pour une variable aléatoire discontinue.

Pour une variable aléatoire continue, le cumul de la fonction de densité de probabilité est égal à 1 :

$$\int_{-\infty}^{+\infty} f(x)dx = 1$$

$$F(x) = \int_{-\infty}^{x} f(x)dx$$

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

Exemple :

Soit une variable aléatoire continue X définie par la fonction de densité de probabilité :

$$f(x) = \begin{cases} k & \text{si} \quad 0 \le x \le 1 \\ 0 & \text{sinon} \end{cases}$$

Pour déterminer la constante k, il faut :

$$\int_{-\infty}^{+\infty} f(x)dx = 1$$

$$\int_{0}^{1} k \times dx = 1$$

$$k \times x \Big]_{0}^{1} = 1$$

$$k = 1$$

$$f(x) = \begin{cases} 1 & \text{si} \quad 0 \le x \le 1 \\ 0 & \text{sinon} \end{cases}$$

On en déduit par intégration la fonction de répartition F(x) :

Si x < 0 :

$$F(x) = \int_{-\infty}^{x} f(x)dx = \int_{-\infty}^{0} 0 \times dx = 0$$

Si 0 ≤ x ≤ 1 :

$$F(x) = \int_{-\infty}^{x} f(x)dx = \int_{-\infty}^{0} 0 \times dx + \int_{0}^{x} 1 \times dx = x$$

Si x > 1 :

$$F(x) = \int_{-\infty}^{x} f(x)dx = \int_{-\infty}^{0} 0 \times dx + \int_{0}^{1} 1 \times dx + \int_{1}^{x} 0 \times dx = 1$$

$$F(x) = \begin{cases} 0 & \text{si} \quad x < 0 \\ x & \text{si} \quad 0 \le x \le 1 \\ 1 & \text{si} \quad x > 1 \end{cases}$$

CARACTERISTIQUES D'UNE VARIABLE ALEATOIRE

I. ESPERANCE MATHEMATIQUE

1.1. Définition

On appelle espérance mathématique la valeur moyenne de la variable, elle remplace la moyenne arithmétique dans le cas d'une variable statistique.

Cas discret : $E(X) = \sum x \times p(x)$

Cas continu : $E(X) = \int_{-\infty}^{+\infty} x \times f(x)dx$

Exemple :

- Soit X la variable aléatoire qui caractérise le nombre de garçons dans une famille de quatre enfants.

Distribution de probabilité de X

x	p(x)	F(x)
0	0,0625	0,0625
1	0,2500	0,3125
2	0,3750	0,6875
3	0,2500	0,9375
4	0,0625	1
Total	1	

$$E(X) = \sum x \times p(x) = 0 \times 0{,}0625 + 1 \times 0{,}25 + 2 \times 0{,}375 + 3 \times 0{,}25 + 4 \times 0{,}0625$$

$$E(X) = 2$$

Dans une famille de quatre enfants on doit s'attendre à avoir deux garçons.

Exemple :

Soit une variable aléatoire continue X définie par la fonction de densité de probabilité :

$$f(x) = \begin{cases} 1 & \text{si} \quad 0 \le x \le 1 \\ 0 & \text{sinon} \end{cases}$$

$$E(X) = \int_0^1 x \times dx = \frac{x^2}{2}\Big]_0^1 = \frac{1}{2}$$

1.2. Propriétés

- L'espérance d'une fonction d'une variable X est :

Cas discret : $E(g(X)) = \sum g(x) \times p(x)$

Cas continu : $E(g(X)) = \int_{-\infty}^{+\infty} g(x) \times f(x)dx$

Exemple :

Cas discret : $E(X^2) = \sum x^2 \times p(x)$

Cas continu : $E(X^2) = \int_{-\infty}^{+\infty} x^2 \times f(x)dx$

- L'espérance d'une constante est la constante : E(a) = a
- L'espérance d'une transformation linéaire est la transformation linéaire de l'espérance :

$$E(ax+b) = \sum (ax+b) \times p(x) = \sum axp(x) + \sum bp(x)$$

$$E(ax+b) = a\sum xp(x) + b\sum p(x)$$

$$E(ax+b) = aE(X) + b$$

- L'espérance d'une somme est la somme des espérances :

$$E(X + Y) = E(X) + E(Y)$$

- L'espérance d'une différence est la différence des espérances :

$$E(X - Y) = E(X) - E(Y)$$

- L'espérance d'un produit est le produit des espérances si les variables sont indépendantes :

$$E(X \times Y) = E(X) \times E(Y)$$

II. VARIANCE ET ECART-TYPE

2.1. Définition

Comme pour la moyenne, la variance d'une variable aléatoire conserve la même définition que la variance d'une variable statistique. C'est l'espérance mathématique des carrés des écarts par rapport à l'espérance.

- Cas discret : $V(X) = E[(X - E(X))^2] = \sum (x - E(X))^2 \times p(x)$

- Cas continu : $V(X) = E[(X - E(X))^2] = \int_{-\infty}^{+\infty} (x - E(X))^2 \times f(x)dx$

L'écart type est égal à la racine carrée de la variance :

$$\sigma = \sqrt{V(X)}$$

La variance est calculée à partir de la formule développée suivante :

$$V(X) = E[(X - E(X))^2] = E[X^2 - 2XE(X) + E(X)^2]$$

$$V(X) = E(X^2) - 2\,E(X)\,E(X) + E(X)^2$$

$$V(X) = E(X^2) - E(X)^2$$

La variance est donc égale à la différence entre l'espérance mathématique des carrés et le carré de l'espérance mathématique.

Exemple :

- Soit X la variable aléatoire qui caractérise le nombre de garçons dans une famille de quatre enfants.

Distribution de probabilité de X

x	p(x)	F(x)
0	0,0625	0,0625
1	0,2500	0,3125
2	0,3750	0,6875
3	0,2500	0,9375
4	0,0625	1
Total	1	

$$E(X) = \sum x \times p(x) = 0 \times 0{,}0625 + 1 \times 0{,}25 + 2 \times 0{,}375 + 3 \times 0{,}25 + 4 \times 0{,}0625 = 2$$

$$E(X^2) = \sum x^2 \times p(x) = 0^2 \times 0{,}0625 + 1^2 \times 0{,}25 + 2^2 \times 0{,}375 + 3^2 \times 0{,}25 + 4^2 \times 0{,}0625 = 5$$

$V(X) = E(X^2) - E(X)^2 = 5 - 2^2 = 1$

écart type est la racine carrée de 1 :

$$\sigma = \sqrt{1} = 1$$

Exemple :

Soit une variable aléatoire continue X définie par la fonction de densité de probabilité :

$$f(x) = \begin{cases} 1 & \text{si} \quad 0 \le x \le 1 \\ 0 & \text{sinon} \end{cases}$$

$$E(X) = \int_0^1 x \times dx = \frac{x^2}{2}\Big]_0^1 = \frac{1}{2}$$

$$E(X^2) = \int_0^1 x^2 \times dx = \frac{x3}{3}\Big]_0^1 = \frac{1}{3}$$

$$V(X) = E(X^2) - E(X)^2 = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}$$

$$\sigma = \frac{1}{\sqrt{12}}$$

2.2. Propriétés

- La variance d'une constante est nulle : V(a) = 0
- La variance d'une transformation linéaire est :

$$V(aX + b) = E[((aX + b) - E(aX + b))^2]$$

$$V(aX + b) = E[(aX + b - aE(X) - b)^2]$$

$$V(aX + b) = E[a^2(X - E(X))^2]$$

$$V(aX + b) = a^2V(X)$$

- La variance d'une somme est la somme des variances si les variables sont indépendantes :

$$V(X + Y) = E[((X + Y) - E(X+Y))^2]$$

$$V(X + Y) = E[(X + Y - E(X) - E(Y))^2]$$

$$V(X + Y) = E[((X-E(X)) + (Y-E(Y)))^2]$$

$$V(X + Y) = E[(X-E(X))^2 + 2\,(X-E(X))\,(Y-E(Y)) + (Y-E(Y))^2]$$

$$V(X + Y) = E[(X-E(X))^2] + 2\,E[(X-E(X))\,(Y-E(Y))] + E[(Y-E(Y))^2]$$

Si X et Y sont indépendantes, on peut écrire :

$$E[(X-E(X))\,(Y-E(Y))] = E(X-E(X))\,E(Y-E(Y)) = 0$$

$$V(X + Y) = E[(X-E(X))^2] + E[(Y-E(Y))^2]$$

$$V(X + Y) = V(X) + V(Y)$$

- La variance d'une différence est la somme des variances si les variables sont indépendantes :

$$V(X - Y) = E[((X - Y) - E(X-Y))^2]$$

$$V(X - Y) = E[(X - Y - E(X) + E(Y))^2]$$

$$V(X - Y) = E[((X-E(X)) - (Y-E(Y)))^2]$$

$$V(X - Y) = E[(X-E(X))^2 - 2\,(X-E(X))\,(Y-E(Y)) + (Y-E(Y))^2]$$

$$V(X - Y) = E[(X-E(X))^2] - 2\,E[(X-E(X))\,(Y-E(Y))] + E[(Y-E(Y))^2]$$

Si X et Y sont indépendantes, on peut écrire :

$$E[(X-E(X))\,(Y-E(Y))] = E(X-E(X))\,E(Y-E(Y)) = 0$$

$$V(X - Y) = E[(X-E(X))^2] + E[(Y-E(Y))^2]$$

$$V(X - Y) = V(X) + V(Y)$$

- Variable centrée réduite

Une variable aléatoire est dite centrée si son espérance mathématique est nulle, elle est dite réduite si son écart-type est égal à 1.

Toute variable aléatoire peut être transformée en une variable centrée réduite par le changement de variable $\frac{X - E(X)}{\sigma}$.

III. CONVERGENCE EN PROBABILITE

On dit qu'une variable aléatoire Xn converge en probabilité vers une constante a si :

$$\forall\ \varepsilon > 0, \qquad \lim_{n\to\infty} P(|X_n - a| > \varepsilon) = 0$$

Ceci signifie que l'écart entre le paramètre calculé à partir de l'échantillon et la vraie valeur du paramètre de la population est très faible quand la taille de l'échantillon est grande. Cet écart peut être mesuré par la variance. Ainsi on parle de convergence en probabilité si :

$$\lim_{n\to\infty} V(X_n) = 0$$

Exemple 1 :

Soit Xn une variable aléatoire qui désigne le nombre de succès obtenus lors de n prélèvements dans une population finie de taille N et dont la proportion de succès est p.

Désignons par $F_n = \frac{X_n}{n}$ la fréquence relative (pourcentage) des succès.

- Cas des prélèvements sans remise :

Dans ce cas la variable aléatoire Xn suit une loi hypergéométrique de paramètre N, n et p.

On sait que :

$$E(Xn) = n\,p \quad \text{et} \quad V(Xn) = \frac{N-n}{N-1}\,n\,p\,q$$

On démontre :

$$E(F_n) = E(\frac{X_n}{n}) = \frac{1}{n}E(X_n) = \frac{1}{n}\,n\,p = p$$

$$V(F_n) = V(\frac{X_n}{n}) = \frac{1}{n^2}V(X_n) = \frac{1}{n^2}\,\frac{N-n}{N-1}\,n\,p\,q = \frac{N-n}{N-1}\,\frac{pq}{n}$$

$$\lim_{n\to\infty} V(F_n) = 0$$

La fréquence relative F_n converge en probabilité vers p.

- Cas des prélèvements avec remise :

Dans ce cas la variable aléatoire Xn suit une loi binomiale de paramètre n et p.

On sait que :

$$E(Xn) = n\,p \quad \text{et} \quad V(Xn) = n\,p\,q$$

On démontre :

$$E(F_n) = E(\frac{X_n}{n}) = \frac{1}{n}E(X_n) = \frac{1}{n}\,n\,p = p$$

$$V(F_n) = V(\frac{X_n}{n}) = \frac{1}{n^2}V(X_n) = \frac{1}{n^2}\,n\,p\,q = \frac{pq}{n}$$

$$\lim_{n\to\infty} V(F_n) = 0$$

La fréquence relative F_n converge en probabilité vers p.

Exemple 2 :

Soient Xi (i=1 à n) n variables aléatoires indépendantes et ayant la même loi de probabilité.

E(Xi) = m et V(Xi) = σ^2

Désignons par : $\bar{X}_n = \frac{\sum_{i=1}^{n} Xi}{n}$ la moyenne calculée à partir d'un échantillon de taille n.

- Cas des prélèvements sans remise :

On démontre :

$$E(\bar{X}_n) = E(\frac{\sum_{i=1}^{n} Xi}{n}) = \frac{1}{n} \times \sum_{i=1}^{n} E(Xi) = \frac{1}{n} \times n \times m = m$$

$$V(\bar{X}_n) = V(\frac{\sum_{i=1}^{n} Xi}{n}) = \frac{1}{n^2} \times \sum_{i=1}^{n} V(Xi) = \frac{1}{n^2} \times n \times \frac{N-n}{N-1} \times \sigma^2 = \frac{N-n}{N-1} \frac{\sigma^2}{n}$$

$$\lim_{n \to \infty} V(\bar{X}_n) = 0$$

La moyenne $\bar{X}_n = \frac{\sum_{i=1}^{n} Xi}{n}$ calculée à partir d'un échantillon de taille n converge en probabilité vers m.

- Cas des prélèvements avec remise :

On démontre :

$$E(\bar{X}_n) = E(\frac{\sum_{i=1}^{n} Xi}{n}) = \frac{1}{n} \times \sum_{i=1}^{n} E(Xi) = \frac{1}{n} \times n \times m = m$$

$$V(\bar{X}_n) = V(\frac{\sum_{i=1}^{n} Xi}{n}) = \frac{1}{n^2} \times \sum_{i=1}^{n} V(Xi) = \frac{1}{n^2} \times n \times \sigma^2 = \frac{\sigma^2}{n}$$

$$\lim_{n \to \infty} V(\bar{X}_n) = 0$$

La moyenne $\bar{X}_n = \frac{\sum_{i=1}^{n} Xi}{n}$ calculée à partir d'un échantillon de taille n converge en probabilité vers m.

IV. INEGALITE DE BIENAYME TCHEBYCHEFF

Cette inégalité concerne des probabilités relatives à des écarts par rapport à l'espérance mathématique supérieurs à k fois écart type, c'est à dire à des écarts centrés réduits $\frac{X - E(X)}{\sigma}$.

Quelle que soit la variable aléatoire X, la probabilité d'un intervalle [E(X)-kσ , E(X)+kσ] a pour borne inférieure $1 - \frac{1}{k^2}$.

$$P(E(X)-k\sigma < X < E(X)+k\sigma) \geq 1 - \frac{1}{k^2}$$

Si on pose k = $\frac{\varepsilon}{\sigma}$ l'inégalité peut être écrite :

$$P(E(X)-\varepsilon < X < E(X)+\varepsilon) \geq 1 - \frac{V(X)}{\varepsilon^2} \quad \text{ou} \quad P(|X - E(X)| < \varepsilon) \geq 1 - \frac{V(X)}{\varepsilon^2}$$

Demonstration :

$$V(X) = \sum (x - E(X))^2 p(x)$$

On peut décomposer la variance en trois sommes :

$$V(X) = S1 + S2 + S3$$

avec :

- S1 = $\sum (x - E(X))^2 p(x)$ pour $x < E(X)-k\sigma$
- S2 = $\sum (x - E(X))^2 p(x)$ pour $E(X)-k\sigma \leq x \leq E(X)+\sigma$
- S3 = $\sum (x - E(X))^2 p(x)$ pour $x > E(X)+\sigma$

$$V(X) = S1 + S2 + S3$$

$$V(X) \geq S1 + S3$$

- Pour S1 $x < E(X) - k\sigma$

  $x - E(X) < - k\sigma$

  $(x - E(X))^2 > k^2\sigma^2$

  $$\sum (x - E(X))^2 p_1(x) \geq \sum k^2\sigma^2 p_1(x)$$

  $$S1 \geq k^2\sigma^2 \sum p_1(x)$$

- Pour S3 $x > E(X) + k\sigma$

$x - E(X) > k\sigma$

$(x - E(X))^2 > k^2\sigma^2$

$$\sum (x - E(X))^2 p_3(x) \geq \sum k^2\sigma^2 p_3(x)$$

$$S3 \geq k^2\sigma^2 \sum p_3(x)$$

$$V(X) \geq S1 + S3$$

$$V(X) \geq k^2\sigma^2 \sum p_1(x) + k^2\sigma^2 \sum p_3(x)$$

$$V(X) \geq k^2\sigma^2 \times (\sum p_1(x) + \sum p_3(x))$$

$$\sum p_1(x) + \sum p_3(x) = 1 - \sum p_2(x)$$

On note : $\sum p_2(x) = p$

$$\sum p_2(x) = p(E(X) - k\sigma \leq X \leq E(X) + k\sigma)$$

Or $V(X) = \sigma^2$

On a donc :

$$\sigma^2 \geq k^2\sigma^2 \times (1 - p)$$

$$1 \geq k^2 \times (1 - p)$$

$$\frac{1}{k^2} \geq 1 - p$$

$$p \geq 1 - \frac{1}{k^2}$$

L'inégalité de Biénaymé Tchebycheff est donc :

$$p(E(X) - k\sigma \leq X \leq E(X) + k\sigma) \geq 1 - \frac{1}{k^2}$$

ou encore :

$P(E(X)-\varepsilon < X < E(X)+\varepsilon) \geq 1-\frac{V(X)}{\varepsilon^2}$ ou $P(|X-E(X)| < \varepsilon) \geq 1-\frac{V(X)}{\varepsilon^2}$

En appliquant L'inégalité de Biénaymé Tchebycheff à la fréquence relative $f_n=\frac{X_n}{n}$ et à la moyenne $\bar{X}_n=\frac{\sum_{i=1}^{n} Xi}{n}$ on obtient :

$$P(|f_n-p| < \varepsilon) \geq 1-\frac{pq}{n\varepsilon^2} \quad \text{et} \quad P(\left|\bar{X}-m\right| < \varepsilon) \geq 1-\frac{\sigma^2}{n\varepsilon^2}$$

LOIS THEORIQUES DISCRETES

I. INTRODUCTION

Le but des lois théoriques est la description des phénomènes statistiques dont le but de calculer la probabilité de certains événements et donc d'avoir une certaine représentation de l'avenir.

Nous étudierons au cours de ce chapitre les lois de probabilités les plus courantes qui vont nous permettre la description d'un phénomène aléatoire déterminé. Nous présenterons ainsi la loi de Bernoulli, la loi binomiale, la loi hypergéométrique, et la loi de poisson.

II. LOI DE BERNOULLI

La loi de Bernoulli intervient dans le cas d'une seule expérience aléatoire à laquelle on associe un événement aléatoire quelconque.

La réalisation de l'événement au cours de cette expérience est appelée succès et la probabilité de réalisation est dite probabilité de succès, désignée par p. Par contre la non-réalisation de l'événement est appelée échec et la probabilité de non-réalisation est dite probabilité d'échec, désignée par q.

$$q = 1 - p$$

La variable aléatoire X qui caractérise le nombre de succès au cours d'une seule expérience aléatoire est appelée variable de Bernoulli, elle prend les valeurs entières 0 et 1 avec les probabilités respectives q et p.

Loi de probabilité d'une variable Bernoulli

x	p(x)
0	q
1	P
Total	1

Les caractéristiques d'une variable Bernoulli sont :

- **Espérance mathématique**

$$E(X) = \sum xp(x) = 0 \times q + 1 \times p = p$$

- **Variance**

$$E(X^2) = \sum x^2 p(x) = 0^2 \times q + 1^2 \times p = p$$

$$V(X) = E(X^2) - E(X)^2 = p - p^2 = p\,(1 - p) = pq$$

Exemple :

On lance une pièce de monnaie une seule fois. Soit X la variable aléatoire qui caractérise le nombre de piles obtenues. X est une variable de Bernoulli, elle prend les valeurs entières 0 et 1 avec la probabilité constante 0,5.

Loi de probabilité de X

x	p(x)
0	0,5
1	0,5
Total	1

III. LOI BINOMIALE

3.1. Définition

La loi binomiale intervient dans le cas de plusieurs expériences aléatoires identiques et indépendantes aux quelles on associe un événement aléatoire quelconque.

La réalisation de l'événement au cours de chacune des expériences est appelée succès et la probabilité de réalisation est dite probabilité de succès, désignée par p. Par contre la non-réalisation de l'événement est appelée échec et la probabilité de non-réalisation est dite probabilité d'échec, désignée par q.

$$q = 1 - p$$

Les probabilités p et q restent constantes au cours d'une suite d'expériences aléatoires. C'est le cas des prélèvements d'individus au hasard dans une population infinie ou le prélèvement d'individus dans une population finie, lorsque les individus sont remis en place au fur et à mesure des prélèvements.

La variable aléatoire X qui caractérise le nombre de succès au cours de n expériences aléatoires indépendantes est appelée variable binomiale, elle prend les valeurs entières de 0 à n.
La probabilité d'obtenir x succès et donc (n-x) échecs au cours de n expériences aléatoires indépendantes est, pour x = 0, 1, ..., n :

$$p(x) = C_n^x p^x q^{n-x}$$

La loi binomiale dépend de deux paramètres :

- n = nombre d'expériences aléatoires indépendantes ;
- p = probabilité de succès au cours de chacune des n expériences aléatoires, p doit rester constante.

Une variable aléatoire X qui sui une loi binomiale de paramètres n et p, est désignée par :

$$X = B(n , p)$$

3.2. Caractéristiques d'une variable binomiale

La variable Bernoulli est un cas particulier de la loi binomiale, elle correspond à la loi binomiale de paramètres 1 et p.

Une variable binomiale de paramètres n et p, peut être considérée comme étant la somme de n variables de Bernoulli identiques et indépendantes de même paramètre p.

$$X = B(n , p)$$

$$X = X1 + X2 + \ldots + Xn$$

Avec Xi (i=1 à n) est une variable Bernoulli tel que :

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

- **Espérance mathématique**

En appliquant la propriété de l'espérance d'une somme on peut écrire :

$$E(X) = E(X1 + X2 + \ldots + Xn)$$

$$E(X) = E(X1) + E(X2) + \ldots + E(Xn)$$

$$E(X) = p + p + \ldots + p$$

$$E(X) = np$$

- **Variance et écart-type**

En appliquant la propriété de la variance d'une somme de variables aléatoires indépendantes on peut écrire :

$$V(X) = V(X1 + X2 + \ldots + Xn)$$

$$V(X) = V(X1) + V(X2) + \ldots + V(Xn)$$

$$V(X) = pq + pq + \ldots + pq$$

$$V(X) = npq$$

Ecart type : $\sigma = \sqrt{npq}$

Exemple :

Dans un lot important de pièces, dont 10 % sont défectueuses, on prélève un échantillon de 20 pièces. Quelle est la probabilité d'obtenir plus de deux pièces défectueuses ?

On définit la variable aléatoire X comme étant le nombre de pièces défectueuses qu'on peut obtenir dans l'échantillon. La variable X peut prendre les valeurs entières de 0 à 20.

La population des pièces peut être considérée comme une population pratiquement infinie. La probabilité de succès, c'est à dire la probabilité qu'une pièce choisie soit défectueuse, est constante et égale à 0,1. La variable aléatoire X suit donc une loi binomiale de paramètre 20 et 0,1.

$$X = B(20 ; 0,1)$$

La probabilité d'avoir plus de deux pièces défectueuses dans l'échantillon est :

$$P(X > 2) = 1 - p(X \leq 2) = 1 - p(0) - p(1) - p(2)$$

$$p(X > 2) = 1 - C_{20}^{0} 0,1^{0} \times 0,9^{20} - C_{20}^{1} 0,1^{1} \times 0,9^{19} - C_{20}^{2} 0,1^{2} \times 0,9^{18}$$

$$p(X > 2) = 1 - 0,1501 - 0,2702 - 0,2852 = 0,2945$$

L'espérance mathématique :

$$E(X) = np = 20 \times 0,1 = 2 \text{ pièces défectueuses.}$$

Dans un échantillon de 20 pièces, on peut s'attendre à avoir deux pièces défectueuses.

La variance :

$$V(X) = npq = 20 \times 0,1 \times 0,9 = 1,8$$

3.3. Propriétés

- **Additivité**

La somme de deux ou plusieurs variables binomiales indépendantes de même paramètres p est elle-même une variable binomiale.

$$X_1 = B(n_1 , p) \qquad X_2 = B(n_2 , p) \qquad \ldots \qquad X_k = B(n_k , p)$$

$$X_1 + X_2 + \ldots + X_k = B(n_1 + n_2 + \ldots + n_k , p)$$

- **Formule de récurrence**

En effectuant le rapport de deux probabilités successives, on obtient :

$$p(x+1) = \frac{p(n-x)}{q(x+1)} p(x)$$

- Les distributions binomiales sont symétriques lorsque p = q = 1/2, la dissymétrie est d'autant plus grande que p et q sont plus différents de 1/2.

Exemple :

Distribution de la variable B(4 , 1/2)

x	p(x)
0	0,0625
1	0,2500
2	0,3750
3	0,2500
4	0,0625
Total	1

IV. LOI HYPERGEOMETRIQUE

4.1. Définition

La loi hypergéométrique intervient dans le cas de plusieurs expériences aléatoires dépendantes aux quelles on associe un caractère étudié quelconque.

La probabilité de succès varie d'une expérience aléatoire à l'autre. C'est le cas des prélèvements d'individus au hasard dans une population finie, lorsque les individus ne sont pas remis en place au fur et à mesure des prélèvements.

Désignons par N l'effectif total de la population dans laquelle on prélève au hasard et sans remise n individus. La population est composée d'individus qui possèdent le caractère étudié, le nombre de ces individus sera désigné par n_1. n_2 désigne le nombre d'individus de la population qui ne possèdent pas le caractère étudié.

$$N = n_1 + n_2$$

La variable aléatoire X, qui caractérise le nombre d'individus prélevés qui possèdent le caractère étudié, est appelée variable hypergéométrique, elle prend les valeurs entières de 0 à n.

La probabilité d'obtenir x individus possédant le caractère étudié parmi les n individus prélevés et donc (n-x) individus ne possédant pas le caractère étudié est, pour x = 0, 1, ..., n :

$$p(x) = \frac{C_{n_1}^{x} C_{n_2}^{n-x}}{C_N^n}$$

La loi hypergéométrique dépend de trois paramètres :

- N = effectif total de la population ;
- n_1 = nombre d'individus de la population qui possèdent le caractère étudié ;
- n = nombre d'individus prélevés sans remise.

Une variable aléatoire X qui sui une loi hypergéométrique de paramètres N, n_1, et n est désignée par :

$$X = H(N, n_1 , n)$$

4.2. Caractéristiques d'une variable hypergéométrique

Les distributions hypergéométriques possèdent des propriétés semblables à celles des distributions binomiales.

La proportion des individus de la population qui possèdent le caractère étudié est :

$$p = \frac{n_1}{N}$$

La proportion des individus de la population qui ne possèdent pas le caractère étudié est :

$$q = \frac{n_2}{N}$$

- **Espérance mathématique** : $E(X) = np$
- **Variance et écart-type** : $V(X) = \frac{N-n}{N-1}\ npq$ et $\sigma = \sqrt{\frac{N-n}{N-1}}\sqrt{npq}$

Exemple :

Dans une population de 40 personnes, dont 6 personnes sont originaires du Sud, 14 du Nord, 12 de l'Est et 8 de l'Ouest, on choisit au hasard un échantillon de 4 personnes.

La variable aléatoire X désigne le nombre d'individus de l'échantillon qui sont originaire du Nord.

La population étant finie et les prélèvements s'effectuent sans remise, la variable X suit donc une loi hypergéométrique de paramètres :

- N = effectif total de la population = 40
- n_1 = nombre d'individus de la population qui sont originaires du Nord = 14
- n = nombre d'individus prélevés sans remise = 4

$$X = H(40, 14, 4)$$

La distribution de cette variable est telle que, pour x = 0, 1, 2, 3, 4 :

$$p(0)=\frac{C_{14}^{0}C_{26}^{4}}{C_{40}^{4}}=0{,}1636$$

$$p(1)=\frac{C_{14}^{1}C_{26}^{3}}{C_{40}^{4}}=0{,}3983$$

$$p(2)=\frac{C_{14}^{2}C_{26}^{2}}{C_{40}^{4}}=0{,}3236$$

$$p(3)=\frac{C_{14}^{3}C_{26}^{1}}{C_{40}^{4}}=0{,}1036$$

$$p(4)=\frac{C_{14}^{4}C_{26}^{0}}{C_{40}^{4}}=0{,}0110$$

Distribution de probabilité de X

x	p(x)
0	0,1636
1	0,3983
2	0,3236
3	0,1036
4	0,0110
Total	1

La proportion des individus de la population qui sont originaires du Nord est :

$$p=\frac{14}{40}=0{,}35$$

La proportion des individus de la population qui ne sont pas originaires du Nord est :

$$q=\frac{26}{40}=0{,}65$$

- Espérance mathématique : $E(X) = np = 4 \times 0{,}35 = 1{,}4$
- Variance et écart-type : $V(X) = \frac{N-n}{N-1}\ npq = \frac{40-4}{40-1}\ \text{x}\ 4\times0{,}35\times0{,}65 = 0{,}84$
- Ecart type : $\sigma=\sqrt{0{,}84}=0{,}92$

4.3. Approximation de la loi hypergéométrique par la loi binomiale

Dès que l'effectif N de la population devient important, le calcul de $p(x)=\frac{C_{n_1}^{x} C_{n_2}^{n-x}}{C_N^n}$ devient fastidieux. On peut démonter dans ce cas que lorsque l'effectif de la population (N) tend vers l'infini et la proportion des individus possédant le caractère étudié (p) est constante ou tend vers une constante, la loi hypergéométrique tend vers une loi binomiale de paramètre n et p. On peut dans ce cas effectuer les calculs de probabilités de façon approximatives à l'aide de la formule de la loi binomiale. En pratique, l'approximation est satisfaisante dés que la proportion des individus prélevés est inférieure à 5 %.

$$\frac{n}{N} < 0{,}05 \quad \text{ou} \quad N > 20\,n$$

Exemple :

Soit la variable hypergéométrique H(100, 30, 4)

La distribution de cette variable est telle que, pour x = 0, 1, 2, 3, 4 :

$$p(x)=\frac{C_{30}^{x} C_{70}^{4-x}}{C_{100}^{4}}$$

Distribution de probabilité de X = H(100, 30, 4)

x	p(x)
0	0,2338
1	0,4188
2	0,2679
3	0,0725
4	0,0070
Total	1

La distribution de cette variable peut être calculée à l'aide de l'approximation par la loi binomiale de paramètres 4 et 0,3. Les probabilités approximatives sont telle que, pour x = 0, 1, 2, 3, 4 :

$$p(x)=C_4^x 0{,}3^x \times 0{,}7^{4-x}$$

Distribution de probabilité de X = B(4 ; 0,3)

x	p(x)
0	0,2401
1	0,4116
2	0,2646
3	0,0756
4	0,0081
Total	1

On constate que l'approximation est satisfaisante.

V. LOI DE POISSON

5.1. Définition

La loi de poisson intervient pour des phénomènes statistiques dont le nombre de réalisation varie de 0 à l'infini et dont la fréquence moyenne de réalisation est connue.

Exemple :

Nombre d'appels reçus par un standard téléphonique.
Nombre d'accidents de la circulation.
Nombre de visiteur d'un centre commercial.

La variable aléatoire X qui caractérise le nombre de réalisations de ce phénomène est appelée variable de poisson, elle prend les valeurs entières 0,1, 2, …etc.

La probabilité d'obtenir x réalisations est, pour x = 0, 1, 2, ... :

$$p(x) = \frac{e^{-m} \times m^{x}}{x!}$$

La loi binomiale dépend d'un seul paramètre :

- m = fréquence moyenne du phénomène étudié.

Une variable aléatoire X qui suit une loi de poisson de paramètre m est désignée par :

$$X = P(m)$$

Exemple :

Un port a les moyens techniques de recevoir au maximum 4 bateaux pétroliers par jour. Le reste est envoyé vers un autre port. Quelle est la probabilité qu'un jour donné, le port ne puisse recevoir tous les bateaux qui se présentent, si on sait qu'en moyenne 3 bateaux se présentent par jour.

Désignons par la variable aléatoire X, le nombre de bateaux qui se présentent un jour donné. X suit une loi de poisson de paramètre 3.

$$X = P(3)$$

La probabilité qu'un jour donné, le port ne puisse recevoir tous les bateaux qui se présentent est :

$$P(X > 4) = 1 - p(X \leq 4) = 1 - p(0) - p(1) - p(2) - p(3) - p(4)$$

$$p(X > 4) = 1 - \frac{e^{-3} \times 3^0}{0!} - \frac{e^{-3} \times 3^1}{1!} - \frac{e^{-3} \times 3^2}{2!} - \frac{e^{-3} \times 3^3}{3!} - \frac{e^{-3} \times 3^4}{4!}$$

$$p(X > 4) = 1 - 0{,}0498 - 0{,}1494 - 0{,}2240 - 0{,}2240 - 0{,}1680 = 0{,}1840$$

5.2. Caractéristiques d'une variable de poisson

On peut démontrer que l'espérance mathématique d'une variable de poisson est égale à sa variance est égale au paramètre m :

$$E(X) = V(X) = m$$

5.3. Propriété d'additivité

La somme de deux ou plusieurs variables de poisson indépendantes de paramètres respectives m_1, m_2, …, m_k est elle-même une variable de poisson de paramètre la somme des paramètres m_i.

$$X_1 = P(m_1) \qquad X_2 = P(m_2) \qquad \ldots \qquad X_k = P(m_k)$$

$$X_1 + X_2 + \ldots + X_k = P(m_1 + m_2 + \ldots + m_k)$$

5.4. Formule de récurrence

En effectuant le rapport de deux probabilités successives, on obtient :

$$p(x+1) = p(x) \times \frac{m}{x+1}$$

Exemple :

Soit la distribution de poisson de paramètre 3.

$$X = P(3)$$

La distribution de cette variable est telle que, pour x = 0, 1, 2, 3, 4, …

$$p(x) = \frac{e^{-3} \times 3^x}{x!}$$

Les probabilités p(x) peuvent être calculées par récurrence de la manière suivante :

$$p(0) = e^{-3} = 0,0498$$

$$p(1) = 0,0498 \times \frac{3}{1} = 0,1494$$

$$p(2) = 0,1494 \times \frac{3}{2} = 0,2240$$

$$p(3) = 0,2240 \times \frac{3}{3} = 0,2240$$

$$p(4) = 0,2240 \times \frac{3}{4} = 0,1680$$

5.5. Approximation de la loi binomiale par la loi de poisson

Dès que le paramètre n de la loi binomiale devient grand, le calcul de $p(x) = C_n^x p^x q^{n-x}$ devient fastidieux. On peut démonter dans ce cas que lorsque le nombre d'expériences indépendantes (n) tend vers l'infini et la probabilité de succès tend vers zéro de telle sorte que le produit np tend vers une constante, la loi binomiale de paramètre n et p tend vers une loi de poisson de paramètre np. On peut dans ce cas effectuer les calculs de probabilités de façon approximatives à l'aide de la formule de la loi de poisson. En pratique, l'approximation est satisfaisante lorsque la probabilité p est inférieure à 0,1 et le produit np est inférieur à 5.

Exemple :

Une machine fabrique des ampoules avec une proportion d'ampoules défectueuses de 5 %. Pour contrôler la qualité des ampoules, on a prélevé au hasard, dans un lot important d'ampoules, un échantillon de 20 ampoules.

Quelle est la probabilité que sur les 20 ampoules prélevées, on ait plus d'une ampoule défectueuse ?

Désignons par la variable aléatoire X, le nombre d'ampoules défectueuses dans l'échantillon. La variable X peut prendre les valeurs entières de 0 à 20.

La population des ampoules peut être considérée comme une population pratiquement infinie. La probabilité de succès, c'est à dire la probabilité qu'une ampoule choisie soit défectueuse, est constante et égale à 0,05. La variable aléatoire X suit donc une loi binomiale de paramètre 20 et 0,05.

$$X = B(20 ; 0,05)$$

La probabilité d'avoir plus d'une ampoule défectueuse dans l'échantillon est :

$$p(X > 1) = 1 - p(X \leq 1) = 1 - p(0) - p(1)$$

$$p(X > 1) = 1 - C_{20}^{0}\, 0{,}05^{0} \times 0{,}95^{20} - C_{20}^{1}\, 0{,}05^{1} \times 0{,}95^{19}$$

$$p(X > 1) = 1 - 0{,}3585 - 0{,}3774 = 0{,}2641$$

La probabilité d'avoir plus d'une ampoule défectueuse dans l'échantillon peut être calculée de façon approximative à l'aide de la loi de poisson de paramètre $20 \times 0{,}05 = 1$, puisque la probabilité p est inférieure à 0,1 (0,05) et le produit np est inférieur à 5 ($20 \times 0{,}05 = 1$) :

$$p(X > 1) = 1 - p(X \leq 1) = 1 - p(0) - p(1)$$

$$p(X > 1) = 1 - \frac{e^{-1} \times 1^{0}}{0!} - \frac{e^{-1} \times 1^{1}}{1!}$$

$$p(X > 1) = 1 - 0{,}3679 - 0{,}3679 = 0{,}2642$$

On constate que l'approximation est très satisfaisante.

LOIS THEORIQUES CONTINUES

I. INTRODUCTION

Le but des lois théoriques est la description des phénomènes statistiques. Nous étudierons au cours de ce chapitre les lois de probabilités continues les plus courantes. Nous présenterons ainsi la loi Normale dont le principal but est de calculer la probabilité de certains événements et donc d'avoir une certaine représentation des phénomènes. La loi Khi deux de Pearson, la loi de Student et la loi de Fisher qui ont un rôle très important dans les problèmes d'estimation et les tests d'hypothèses.

II. LOI NORMALE

2.1. Définition

La loi normale est la loi continue la plus importante et la plus utilisée dans le calcul de probabilité. Elle est aussi appelée loi de LAPLACE GAUSS[1].

On appelle variable normale toute variable aléatoire continue X définie dans l'intervalle $]-\infty,+\infty[$ par la fonction de densité de probabilité suivante :

$$f(x) = \frac{1}{\sigma\sqrt{2\Pi}} e^{-\frac{1}{2}(\frac{x-m}{\sigma})^2}$$

m et σ sont des paramètres quelconques qui représentent respectivement la moyenne et l'écart type de la variable.

On peut vérifier que :

$$\int_{-\infty}^{+\infty} f(x)dx = 1$$

La loi normale dépend de deux paramètres m et σ. Une variable aléatoire X qui suit une loi normale de paramètres m et σ est désignée par :

$$X = N(m , \sigma)$$

2.2. Loi normale réduite

On appelle variable normale réduite toute variable aléatoire normale Z de paramètres m = 0 et σ = 1.

$$Z = N(0 , 1)$$

[1] **Laplace, Pierre Simon** (1749-1827)

Une variable normale réduite est définie par la fonction de densité de probabilité suivante :

$$f(z)=\frac{1}{\sqrt{2\Pi}}e^{-\frac{z^2}{2}}$$

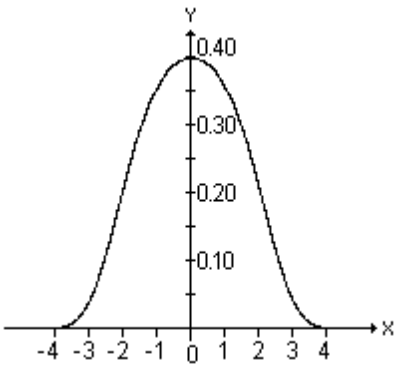
Toute variable normale X de paramètres m et σ peut être transformée en une variable normale réduite par le changement de variable suivant :

$$Z=\frac{X-m}{\sigma}$$

2.3. Forme de la loi normale

La représentation graphique de la fonction de densité de probabilité d'une variable normale est une courbe en forme de cloche symétrique par rapport à la moyenne m et caractérisée par l'existence d'un maximum en x = 0 et f(x) = $\frac{1}{\sigma\sqrt{2\Pi}}$.

En particulier la loi normale réduite est symétrique par rapport à l'axe des abscisses et caractérisée par l'existence d'un maximum en z = 0 et f(z) = $\frac{1}{\sqrt{2\Pi}} \approx 0{,}40$.



Courbe normale de Gauss

La fonction de répartition correspond à l'aire comprise entre cette courbe et l'axe des abscisses.

2.4. Détermination pratique des probabilités

Pour le calcul de probabilités sans utiliser la fonction de densité, des tables de la loi normale réduite ont été élaborées. On distingue deux tables de la loi normale réduite, relatives l'une à la fonction de densité de probabilité et l'autre à la fonction de répartition. En raison de la symétrie de la distribution, ces tables sont limitées aux valeurs positives de z.

Par le changement de variable $Z=\frac{X-m}{\sigma}$ toutes les variables normales se ramènent à la loi normale réduite.

Table de la fonction de répartition

Cette table donne les valeurs de la fonction de répartition Π(z) pour des valeurs positives z d'une variable normale réduite. En raison de la symétrie de f(z), on peut déduire les valeurs Π(z) pour les valeurs négatives de z :

$$\Pi(-z) = p(Z \leq -z) = p(Z > z) = 1 - p(Z \leq z) = 1 - \Pi(z)$$

$$\Pi(-z) = 1 - \Pi(z)$$

Pour une variable normale quelconque X de paramètre m et σ :

$$F(x) = p(X \leq x) = p\left(\frac{X-m}{\sigma} \leq \frac{x-m}{\sigma}\right) = p(Z \leq z) = \Pi(z)$$

$$F(x) = \Pi(z)$$

Pour lire une valeur Π(z) dans la table, il suffit de lire l'intersection entre la ligne correspondante à la valeur de z et la colonne correspondante au deuxième chiffre après la virgule de z.

TABLE DE LA FONCTION DE REPARTITION DE LA LOI NORMALE REDUITE

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	05359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	05753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	06141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	06517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	06879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	07224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	07549
0,7	0,7580	0,7611	0,7642	0,7673	0,7703	0,7734	0,7764	0,7794	0,7823	07852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	08133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	08389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	08621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	08830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	090147
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	0,92785	0,92922	0,93056	0,93189
1,5	0,93319	0,93448	0,93574	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	0,95543	0,95637	0,95728	0,95818	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	0,97932	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
3,5	0,99977	0,99978	0,99978	0,99979	0,99980	0,99981	0,99981	0,99982	0,99983	0,99983
3,6	0,99984	0,99985	0,99985	0,99986	0,99986	0,99987	0,99987	0,99988	0,99988	0,99989
3,7	0,99989	0,99990	0,99990	0,99990	0,99991	0,99991	0,99992	0,99992	0,99992	0,99992
3,8	0,99993	0,99993	0,99993	0,99994	0,99994	0,99994	0,99994	0,99995	0,99995	0,99995
3,9	0,99995	0,99995	0,99996	0,99996	0,99996	0,99996	0,99996	0,99996	0,99997	0,99997

Exemple :

La valeur de Π(1,36) correspond à l'intersection entre la ligne correspondante à 1,3 et la colonne correspondante à 0,06, on peut lire la valeur 0,91309.

$$\Pi(-2,24) = 1 - \Pi(2,24) = 1 - 0,98745 = 0,01255$$

Exemple :

Pour qu'une pièce fabriquée par une machine soit utilisable, sa longueur doit être comprise entre 14,7 et 15,3 cm, sinon elle est rejetée. Sachant que la longueur de cette pièce est une variable normale de paramètres 15 cm et 0,2 cm, quelle proportion de pièces peuvent être rejetées.

Si on désigne par la variable X la longueur des pièces, X suit une loi normale :

$$X = N(15 ; 0,2)$$

La probabilité de rejet d'une pièce est :

$$p(\text{rejet}) = 1 - p(\text{accepter})$$

$$p(\text{accepter}) = p(14,7 \leq X \leq 15,3) = p(X \leq 15,3) - p(X \leq 14,7)$$

$$p(\text{accepter}) = p\left(\frac{X-15}{0,2} \leq \frac{15,3-15}{0,2}\right) - p\left(\frac{X-15}{0,2} \leq \frac{14,7-15}{0,2}\right)$$

$$p(\text{accepter}) = p(Z \leq 1,50) - p(Z \leq -1,50)$$

$$p(\text{accepter}) = \Pi(1,50) - \Pi(-1,50)$$

$$p(\text{accepter}) = \Pi(1,50) - (1 - \Pi(1,50)) = 2 \times \Pi(1,50) - 1$$

$$p(\text{accepter}) = 2x\ 0,93319 - 1 = 0,86638$$

Chaque pièce a une probabilité de 0,13362 d'être rejetée ou il y a un risque de rejet de 13% des pièces fabriquées.

2.5. Propriété d'additivité

La somme de deux ou plusieurs variables normales indépendantes est une variable normale de moyenne la somme des moyennes et d'écart type la racine carrée de la somme des variances des variables initiales.

Soient X_1, X_2, …,X_n n variables normales de paramètres respectivement m_1, m_2, …, m_n et σ_1, σ_2, …,σ_n.

$$X_1 + X_2 + \ldots + X_n = N(m_1 + m_2 + \ldots + m_n, \sqrt{\sigma_1{}^2 + \sigma_2{}^2 + \ldots + \sigma_n{}^2})$$

Exemple :

Pour se rendre à son travail un ouvrier prend deux bus. La durée du trajet du premier bus est une variable normale de paramètres 27 minutes et 5 minutes. La durée du trajet du deuxième bus est une variable normale de paramètres 30 minutes et 2 minutes. Quelle est la probabilité que cet ouvrier n'arrive pas en retard s'il dispose d'une heure ?

- Désignons par X_1 La durée du trajet du premier bus : $X_1 = N(27 ; 5)$.
- Désignons par X_2 La durée du trajet du deuxième bus : $X_2 = N(30 ; 2)$.
- Désignons par X la durée totale des deux trajets : $X = X_1 + X_2$.

La variable X est la somme de deux variables normales indépendantes, elle suit donc une loi normale :

$$X = N(30+27 ; \sqrt{5^2 + 2^2}) = N(57 ; 5,4)$$

Pour ne pas arriver en retard la durée totale des deux trajets ne doit pas dépasser 60 minutes.

$$p(X \leq 60) = p(\frac{X-57}{5,4} \leq \frac{60-57}{5,4}) = p(Z \leq 0,56)$$

$$p(X \leq 60) = \prod(0,56) = 0,7123$$

L'ouvrier a donc 71% de chance de ne pas arriver en retard ou il a un risque de 29 % d'arriver en retard.

2.6. Le théorème central limite

Le théorème central limite est une généralisation de la propriété d'additivité. Toute somme de variables aléatoires indépendantes tend à suivre une loi normale quelles que soient les lois de probabilités suivies par ces variables.

Quelles que soient les variables aléatoires indépendantes X_1, X_2, …, X_n de moyennes respectivement m_1, m_2, …, m_n et d'écarts type respectivement σ_1, σ_2, …, σ_n. La somme de ces variables tend à suivre une loi normale de moyenne la somme des moyennes et d'écart type la racine carrée de la somme des variances des variables initiales.

$$X1+X2+\ldots+Xn \approx N(m1+m2+\ldots+mn, \sqrt{\sigma 1^2+\sigma 2^2+\ldots+\sigma n^2})$$

Exemple :

Une caisse d'assurance maladie reçoit 120 personnes pour l'obtention de remboursements. On suppose que la somme à rembourser à chaque personne est une variable aléatoire de moyenne 1000 dirhams et d'écart type 600 dirhams. La caisse dispose de 130000 dirhams. Quelle est le risque que cette somme ne soit pas suffisante pour rembourser toutes les personnes ?

Désignons par X_i (i = 1 à 120) la somme à rembourser à chaque personne.
Désignons par X la somme totale que la caisse doit payer aux 120 personnes.

$$X = X_1 + X_2 + \ldots + X_{120}$$

D'après le théorème central limite, on peut affirmer que X suit une loi normale de moyenne la somme des moyennes et d'écart type la racine carrée de la somme des variances.

$$X = N(120 \times 1000; \sqrt{120 \times 600^2}) = N(120000; 6572,67)$$

La somme de 130000 dh ne sera pas suffisante si la somme totale à rembourser aux 120 personnes dépasse 130000 dh :

$$p(X > 130000) = 1 - p(X \leq 130000) = 1 - p(\frac{X - 120000}{6572,67} \leq \frac{130000 - 120000}{6572,67})$$

$$p(X > 130000) = 1 - p(Z \leq 1,52) = 1 - \prod(1,52) = 1 - 0,93574 = 0,0643$$

Il y a donc un risque de 6,5 % que la somme de 130000 dirhams ne soit pas suffisante pour rembourser toutes les personnes.

2.7. Approximation de la loi binomiale par la loi normale

Parfois les problèmes relatifs à la loi binomiale se rapportent aux calculs de probabilités dans un ou plusieurs intervalles donnés :

$p(X < x)$ $p(X > x)$ ou $p(x1 < X < x2)$

La recherche de ces probabilités est souvent longue, car il faut déterminer individuellement et d'additionner les différentes probabilités $p(X = x)$.

$$p(X < 10) = p(0)+p(1)+p(2)+p(3)+p(4)+p(5)+p(6)+p(7)+p(8)+p(9)$$

Lorsque le paramètre n de la loi binomiale est grand et les probabilités de succès p et d'échec q ne sont pas trop petites, on peut effectuer ce calcul d'une manière approchée à l'aide de la loi normale de paramètres np et $\sqrt{npq}$.

En pratique l'approximation est satisfaisante lorsque les produits np et nq sont supérieurs à 5 :

$$B(n ; p) \approx N(np ; \sqrt{npq})$$

Pour améliorer la qualité de l'approximation de la loi binomiale, qui est discrète, par la loi normale, qui est continue, on introduit généralement une correction de continuité de 0,5. Les différentes probabilités deviennent :

- $p(X < x - 0,5)$ au lieu de $p(X < x)$
- $p(X > x + 0,5)$ au lieu de $p(X > x)$
- $p(x1 - 0,5 < X < x2 + 0,5)$ au lieu de $p(x1 < X < x2)$

Exemple :

On suppose que la probabilité qu'un étudiant réussisse un examen est de 0,8. Quelle est la probabilité qu'au moins 75 étudiants parmi 100 étudiants réussissent l'examen ?

Désignons par X le nombre d'étudiants qui réussissent l'examen.

X est une variable discrète qui prend les valeurs entières de 0 à 100. Elle suit une loi binomiale de paramètres 100 et 0,8.

$$X = B(100 ; 0,8)$$

La probabilité qu'au moins 75 étudiants parmi 100 étudiants réussissent l'examen est :

$$p(X \geq 75)$$

Les produits np et nq sont respectivement $100 \times 0,8 = 80$ et $100 \times 0,2 = 20$, ils sont supérieurs à 5. On peut donc effectuer le calcul de cette probabilité d'une manière approchée à l'aide de la loi normale de paramètres $np = 80$ et $\sqrt{npq} = 4$.

$$X = B(100 ; 0,8) \approx N(80 ; 4)$$

Pour améliorer la qualité de l'approximation on introduit la correction de continuité, la probabilité $p(X \geq 75)$ devient :

$$p(X \geq 75 + 0,5) = 1 - p(X < 75,5)$$

$$p(X \geq 75,5) = 1 - p\left(\frac{X - 80}{4} < \frac{75,5 - 80}{4}\right) = 1 - p(Z < -1,13)$$

$$p(X \geq 75,5) = 1 - \Pi(-1,13) = \Pi(1,13) = 0,8708$$

$$p(X \geq 75) \approx 0,8708$$

La probabilité qu'au moins 75 étudiants parmi 100 étudiants réussissent l'examen est à peu près 0,8708.

Le calcul exact à partir de la loi binomiale donne un résultat de 0,8686. On constate que l'approximation est très satisfaisante.

III. LOIS DERIVEES DE LA LOI NORMALE

Cet ensemble de lois de répartition est particulièrement utile dans les problèmes d'estimations et les tests statistiques.

3.1. La loi Khi deux de Pearson

3.1.1. Définition

On appelle variable Khi deux de Pearson, la variable χ^2 qui varie entre 0 et $+\infty$ et définie par la fonction de densité de probabilité :

$$f(x) = c \times x^{\frac{k}{2}-1} e^{-\frac{x}{2}}$$

Le paramètre k est une constante entière positive appelée nombre de degrés de liberté, on dit variable Khi carré à k degré de liberté, désignée par $\chi^2_{\text{à k dl}}$.

c est une constante telle que : $\int_0^{+\infty} f(x)dx = 1$

La variable Khi deux de Pearson correspond aussi à la somme des carrés de k variables normales réduites indépendantes.

Soient Z_1, Z_2, …, Z_k k variables normales réduites indépendantes, on peut démontrer :

$$\chi^2_{\text{à k dl}} = Z_1^2 + Z_2^2 + \ldots + Z_k^2$$

3.1.2. Caractéristiques de la loi $\chi^2_{\text{à k dl}}$

On peut démontrer que :

- Espérance mathématique : $E(\chi^2_{\text{à k dl}}) = k$
- Variance : $V(\chi^2_{\text{à k dl}}) = 2\ k$

3.1.3. Propriété d'additivité

La somme de deux ou plusieurs variables Khi carré indépendantes est une variable Khi carrée.

Soient n variables Khi deux de degrés de liberté respectivement k_1, k_2, …, k_n :

$$\chi^2_{\text{à k1 dl}} + \chi^2_{\text{à k2 dl}} + \ldots + \chi^2_{\text{à kn dl}} = \chi^2_{\text{à (k1+k2+…+kn) dl}}$$

Une variable Khi carré à k degré de liberté peut donc être considérée comme étant la somme

de k variables Khi carré à 1 degré de liberté indépendantes.

3.1.4. Table de la loi Khi deux de Pearson

La table de la loi Khi carré dépend du paramètre k, elle donne les valeurs de $\chi^2_{\text{à k dl}}$ pour les valeurs de la fonction de répartition $F(\chi^2_{\text{à k dl}})$.

TABLE DE LA LOI KHI DEUX DE PEARSON

k / p	0,0005	0,001	0,005	0,01	0,025	0,05	0,1	0,2	0,3	0,4
1	$0,0^{6}393$	$0,0^{5}157$	$0,0^{4}393$	$0,0^{3}157$	$0,0^{3}982$	$0,0^{2}393$	0,0158	0,0642	0,148	0,275
2	$0,0^{2}100$	$0,0^{2}200$	0,0100	0,0201	0,0506	0,103	0,211	0,446	0,713	1,02
3	0,0153	0,0243	0,0717	0,115	0,216	0,352	0,584	1,00	1,42	1,87
4	0,0639	0,0908	0,207	0,297	0,484	0,711	1,06	1,65	2,19	2,75
5	0,158	0,210	0,412	0,554	0,831	1,15	1,61	2,34	3,00	3,66
6	0,299	0,381	0,676	0,872	1,24	1,64	2,20	3,07	3,83	4,57
7	0,485	0,598	0,989	1,24	1,69	2,17	2,83	3,82	4,67	5,49
8	0,710	0,857	1,34	1,65	2,18	2,73	3,49	4,59	5,53	6,42
9	0,972	1,15	1,73	2,09	2,70	3,33	4,17	5,38	6,39	7,36
10	1,26	1,48	2,16	2,56	3,25	3,94	4,87	6,18	7,27	8,30
11	1,59	1,83	2,60	3,05	3,82	4,57	5,58	6,99	8,15	9,24
12	1,93	2,21	3,07	3,57	4,40	5,23	6,30	7,81	9,03	10,2
13	2,31	2,62	3,57	4,11	5,01	5,89	7,04	8,63	9,93	11,1
14	2,70	3,04	4,07	4,66	5,63	6,57	7,79	9,47	10,8	12,1
15	3,11	3,48	4,60	5,23	6,26	7,26	8,55	10,3	11,7	13,0
16	3,54	3,94	5,14	5,81	6,91	7,96	9,31	11,2	12,6	14,0
17	3,98	4,42	5,70	6,41	7,56	8,67	10,1	12,0	13,5	14,9
18	4,44	4,90	6,26	7,01	8,23	9,39	10,9	12,9	14,4	15,9
19	4,91	5,41	6,84	7,63	8,91	10,1	11,7	13,7	15,4	16,9
20	5,40	5,92	7,43	8,26	9,59	10,9	12,4	14,6	16,3	17,8
21	5,90	6,45	8,03	8,90	10,3	11,6	13,2	15,4	17,2	18,8
22	6,40	6,98	8,64	9,54	11,0	12,3	14,0	16,3	18,1	19,7
23	6,92	7,53	9,26	10,2	11,7	13,1	14,8	17,2	19,0	20,7
24	7,45	8,08	9,89	10,9	12,4	13,8	15,7	18,1	19,9	21,7
25	7,99	8,65	10,5	11,5	13,1	14,6	16,5	18,9	20,9	22,6
26	8,54	9,22	11,2	12,2	13,8	15,4	17,3	19,8	21,8	23,6
27	9,09	9,80	11,8	12,9	14,6	16,2	18,1	20,7	22,7	24,5
28	9,66	10,4	12,5	13,6	15,3	16,9	18,9	21,6	23,6	25,5
29	10,2	11,0	13,1	14,3	16,0	17,7	19,8	22,5	24,6	26,5
30	10,8	11,6	13,8	15,0	16,8	18,5	20,6	23,4	25,5	27,4

TABLE DE LA LOI KHI DEUX DE PEARSON (SUITE)

k / p	0,5	0,6	0,7	0,8	0,9	0,95	0,975	0,99	0,995	0,999	0,9995
1	0,455	0,708	1,07	1,64	2,71	3,84	5,02	6,63	7,88	10,8	12,1
2	1,39	1,83	2,41	3,22	4,61	5,99	7,38	9,21	10,6	13,8	15,2
3	2,37	2,95	3,67	4,64	6,25	7,81	9,35	11,3	12,8	16,3	17,7
4	3,36	4,04	4,88	5,99	7,78	9,49	11,1	13,3	14,9	18,5	20,0
5	4,35	5,13	6,06	7,29	9,24	11,1	12,8	15,1	16,7	20,5	22,1
6	5,35	6,21	7,23	8,56	10,6	12,6	14,4	16,8	18,5	22,5	24,1
7	6,35	7,28	8,38	9,80	12,0	14,1	16,0	18,5	20,3	24,3	26,0
8	7,34	8,35	9,52	11,0	13,4	15,5	17,5	20,1	22,0	26,1	27,9
9	8,34	9,41	10,7	12,2	14,7	16,9	19,0	21,7	23,6	27,9	29,7
10	9,34	10,5	11,8	13,4	16,0	18,3	20,5	23,2	25,2	29,6	31,4
11	10,3	11,5	12,9	14,6	17,3	19,7	21,9	24,7	26,8	31,3	33,1
12	11,3	12,6	14,0	15,8	18,5	21,0	23,3	26,2	28,3	32,9	34,8
13	12,3	13,6	15,1	17,0	19,8	22,4	24,7	27,7	29,8	34,5	36,5
14	13,3	14,7	16,2	18,2	21,1	23,7	26,1	29,1	31,3	36,1	38,1
15	14,3	15,7	17,3	19,3	22,3	25,0	27,5	30,6	32,8	37,7	39,7
16	15,3	16,8	18,4	20,5	23,5	26,3	28,8	32,0	34,3	39,3	41,3
17	16,3	17,8	19,5	21,6	24,8	27,6	30,2	33,4	35,7	40,8	42,9
18	17,3	18,9	20,6	22,8	26,0	28,9	31,5	34,8	37,2	42,3	44,4
19	18,3	19,9	21,7	23,9	27,2	30,1	32,9	36,2	38,6	43,8	46,0
20	19,3	21,0	22,8	25,0	28,4	31,4	34,2	37,6	40,0	45,3	47,5
21	20,3	22,0	23,9	26,2	29,6	32,7	35,5	38,9	41,4	46,8	49,0
22	21,3	23,0	24,9	27,3	30,8	33,9	36,8	40,3	42,8	48,3	50,5
23	22,3	24,1	26,0	28,4	32,0	35,2	38,1	41,6	44,2	49,7	52,0
24	23,3	25,1	27,1	29,6	33,2	36,4	39,4	43,0	45,6	51,2	53,5
25	24,3	26,1	28,2	30,7	34,4	37,7	40,6	44,3	46,9	52,6	54,9
26	25,3	27,2	29,2	31,8	35,6	38,9	41,9	45,6	48,3	54,1	56,4
27	26,3	28,2	30,3	32,9	36,7	40,1	43,2	47,0	49,6	55,5	57,9
28	27,3	29,2	31,4	34,0	37,9	41,3	44,5	48,3	51,0	56,9	59,3
29	28,3	30,3	32,5	35,1	39,1	42,6	45,7	49,6	52,3	58,3	60,7
30	29,3	31,3	33,5	36,3	40,3	43,8	47,0	50,9	53,7	59,7	62,2

Pour lire une valeur $\chi^2_{\text{à k dl}}$ dans la table, il suffit de lire l'intersection entre la colonne correspondante à la valeur de la probabilité cumulée $F(\chi^2_{\text{à k dl}})$ et la ligne correspondante aux degrés de liberté k.

Exemple :

La valeur de $\chi^2_{\text{à 10 dl}}$ pour une probabilité de 0,95 correspond à l'intersection entre la colonne correspondante à 0,95 et la ligne correspondante à 10, on peut lire la valeur 18,3.

$$\chi^2_{0,95 \text{ à } 10 \text{ dl}} = 18,3$$

$$\chi^2_{0,05 \text{ à } 20 \text{ dl}} = 10,9$$

3.1.5. Approximation de la loi Khi deux par la loi normale

Une variable Khi carré à k degrés de liberté peut donc être considérée comme étant la somme de k variables Khi carré à 1 degré de liberté indépendantes.

De ce fait, et par application du théorème central limite, on peut affirmer que la loi Khi deux tend vers une loi normale de paramètres k et $\sqrt{2k}$. Ce qui permet de résoudre les problèmes relatifs aux distributions χ^2 de nombre de degrés de liberté k élevé. Toutefois, la convergence vers la loi normale est relativement lente, l'approximation est généralement satisfaisante lorsque k est supérieur à 100. pour un nombre de degré de liberté compris entre 30 et 100, on préfère faire usage de la racine carrée. On peut en effet démonter que la transformation :

$$Z = \sqrt{2\chi^2} - \sqrt{2k-1}$$

est très proche de la loi normale centrée réduite. On peut aussi utiliser la transformation inverse :

$$\chi^2 = \frac{(Z+\sqrt{2k-1})^2}{2}$$

<u>Exemple 1</u> :

La lecture de la table Khi deux donne :

$$\chi^2_{0,95 \text{ à } 30 \text{ dl}} = 43,8$$

En utilisant l'approximation de la loi Khi deux par la transformation ci dessus on obtient :

$$\chi^2 = \frac{(Z_{0.95}+\sqrt{2\times 30-1})^2}{2}$$

La lecture de la table de la fonction de répartition de la loi normale réduite montre que la valeur de z pour F(z) = 0,95 est égale à 1,65.

$$\chi^2 = \frac{(1.65+\sqrt{59})^2}{2} = 43.8$$

On constate que l'approximation est très satisfaisante.

Exemple 2 :

La valeur de $\chi^2_{0,95 \text{ à } 150 \text{ dl}}$ ne se trouve pas dans la table statistique. Le nombre de degrés de liberté étant très grand, on peut utiliser l'approximation par la loi normale de moyenne 150 et d'écart type $\sqrt{2\times150}=17,32$.

En passant à la loi normale centrée réduite on obtient :

$$\frac{\chi^2 0,95 à 150 dl - 150}{17,32} = Z_{0,95}$$

d'où :

$$\chi^2_{0,95 \text{ à } 30 \text{ dl}} = Z_{0,95} \text{ x } 17,32 + 150$$

$$\chi^2_{0,95 \text{ à } 30 \text{ dl}} = 1,65 \text{ x } 17,32 + 150 = 178,58$$

3.2. La loi t de Student

3.2.1. Définition

On appelle variable t de Student, la variable t qui varie entre -∞ et +∞ et définie par la fonction de densité de probabilité :

$$f(t) = c \times (1 + \frac{t^2}{k})^{-\frac{k+1}{2}}$$

Le paramètre k est une constante entière positive appelée nombre de degrés de liberté, on dit variable t à k degré de liberté, désignée par $t_{\text{à k dl}}$.

c est une constante telle que : $\int_{-\infty}^{+\infty} f(t)dt = 1$

La variable t de Student correspond aussi au quotient d'une variable normale réduite par la racine carrée d'une variable $\chi^2_{\text{à k dl}}$ indépendante de la première variable.

Soient Z une variable normale réduite et $\chi^2_{\text{à k dl}}$ une variable Khi carré à k degrés de liberté, indépendantes. On peut démontrer :

$$t à k dl = \frac{Z}{\sqrt{\frac{\chi^2 à k dl}{k}}}$$

3.2.2. Caractéristiques de la loi $t_{\text{à k dl}}$

On peut démontrer que :

- Espérance mathématique : $E(t_{\text{à k dl}}) = 0$
- Variance : $V(t_{\text{à k dl}}) = k / (k-2)$ pour $k_2 > 2$.

3.2.3. Table de la loi t de Student

La table de la loi t de Student dépend du paramètre k, elle donne les valeurs de $t_{\text{à k dl}}$ pour les valeurs de la fonction de répartition $F(t_{\text{à k dl}})$.

TABLE DE LA LOI T DE STUDENT

k / p	0,6	0,7	0,8	0,9	0,95	0,975	0,99	0,995	0,999	0,9995
1	0,325	0,727	1,376	3,078	6,314	12,71	31,82	63,66	318,3	636,6
2	0,289	0,617	1,061	1,886	2,920	4,303	6,965	9,925	22,33	31,60
3	0,277	0,584	0,978	1,638	2,353	3,182	4,541	5,841	10,22	12,94
4	0,271	0,569	0,941	1,533	2,132	2,776	3,747	4,604	7,173	8,610
5	0,267	0,559	0,920	1,476	2,015	2,571	3,365	4,032	5,893	6,859
6	0,265	0,553	0,906	1,440	1,943	2,447	3,143	3,707	5,208	5,959
7	0,263	0,549	0,896	1,415	1,895	2,365	2,998	3,499	4,785	5,405
8	0,262	0,546	0,889	1,397	1,860	2,306	2,896	3,355	4,501	5,041
9	0,261	0,543	0,883	1,383	1,833	2,262	2,821	3,250	4,297	4,781
10	0,260	0,542	0,879	1,372	1,812	2,228	2,764	3,169	4,144	4,587
11	0,260	0,540	0,876	1,363	1,796	2,201	2,718	3,106	4,025	4,437
12	0,259	0,539	0,873	1,356	1,782	2,179	2,681	3,055	3,930	4,318
13	0,259	0,538	0,870	1,350	1,771	2,160	2,650	3,012	3,852	4,221
14	0,258	0,537	0,868	1,345	1,761	2,145	2,624	2,977	3,787	4,140
15	0,258	0,536	0,866	1,341	1,753	2,131	2,602	2,947	3,733	4,073
16	0,258	0,535	0,865	1,337	1,746	2,120	2,583	2,921	3,686	4,015
17	0,257	0,534	0,863	1,333	1,740	2,110	2,567	2,898	3,646	3,965
18	0,257	0,534	0,862	1,330	1,734	2,101	2,552	2,878	3,611	3,922
19	0,257	0,533	0,861	1,328	1,729	2,093	2,539	2,861	3,579	3,883
20	0,257	0,533	0,860	1,325	1,725	2,086	2,528	2,845	3,552	3,850
21	0,257	0,532	0,859	1,323	1,721	2,080	2,518	2,831	3,527	3,819
22	0,256	0,532	0,858	1,321	1,717	2,074	2,508	2,819	3,505	3,792
23	0,256	0,532	0,858	1,319	1,714	2,069	2,500	2,807	3,485	3,767
24	0,256	0,531	0,857	1,318	1,711	2,064	2,492	2,797	3,467	3,745
25	0,256	0,531	0,856	1,316	1,708	2,060	2,485	2,787	3,450	3,725
26	0,256	0,531	0,856	1,315	1,706	2,056	2,479	2,779	3,435	3,707
27	0,256	0,531	0,855	1,314	1,703	2,052	2,473	2,771	3,421	3,690
28	0,256	0,530	0,855	1,313	1,701	2,048	2,467	2,763	3,408	3,674
29	0,256	0,530	0,854	1,311	1,699	2,045	2,462	2,756	3,396	3,659
30	0,256	0,530	0,854	1,310	1,697	2,042	2,457	2,750	3,385	3,646
40	0,255	0,529	0,851	1,303	1,684	2,021	2,423	2,704	3,307	3,551
60	0,254	0,527	0,848	1,296	1,671	2,000	2,390	2,660	3,232	3,460
80	0,254	0,527	0,846	1,292	1,664	1,990	2,374	2,639	3,195	3,415
100	0,254	0,526	0,845	1,290	1,660	1,984	2,365	2,626	3,174	3,389
200	0,254	0,525	0,843	1,286	1,653	1,972	2,345	2,601	3,131	3,339
500	0,253	0,525	0,842	1,283	1,648	1,965	2,334	2,586	3,106	3,310
∞	0,253	0,524	0,842	1,282	1,645	1,960	2,326	2,576	3,090	3,291

Pour lire une valeur $t_{àkdl}$ dans la table, il suffit de lire l'intersection entre la colonne correspondante à la valeur de la probabilité cumulée $F(t_{à\ k\ dl})$ et la ligne correspondante aux degrés de liberté k.

Exemple :

La valeur de $t_{à\ 10\ dl}$ pour une probabilité de 0,95 correspond à l'intersection entre la colonne correspondante à 0,95 et la ligne correspondante à 10, on peut lire la valeur 1,812.

$$t_{0,95\ à\ 10\ dl} = 1,812$$

$$t_{0,7\ à\ 20\ dl} = 0,533$$

3.2.4. Approximation de la loi t de Student par la loi normale

Lorsque le nombre de degrés de liberté k est très élevé, la loi t de Student peut être directement assimilée à la loi normale réduite sans effectuer aucun changement de variable. Ce qui permet de résoudre les problèmes relatifs aux distributions t de nombre de degrés de liberté élevé. L'approximation est généralement satisfaisante lorsque k est supérieur à 30.

Exemple :

La lecture de la table t donne :

$$t_{0,95 \text{ à } 80 \text{ dl}} = 1,664 \qquad \text{et} \qquad t_{0,8 \text{ à } 80 \text{ dl}} = 0,846$$

En utilisant l'approximation de la loi t par la loi normale réduite, on peut lire dans la table de la fonction de répartition de la loi normale réduite la valeur de z pour F(z) = 0,95 qui est égale à 1,65.

La lecture de la table de la fonction de répartition de la loi normale réduite montre que la valeur de z pour F(z) = 0,80 est égale à 0,84.

On constate que l'approximation est satisfaisante.

3.3. La loi F de Fisher Snédécor

3.3.1. Définition

On appelle variable F de Fisher, la variable F qui varie entre 0 et $+\infty$ et définie par la fonction de densité de probabilité :

$$f(x) = c \times x^{\frac{k1}{2}-1} \times (k_1 x + k_2)^{-\frac{k1+k2}{2}}$$

Les paramètres k_1 et k_2 sont deux constantes entières positives appelées nombre de degrés de liberté, on dit variable F à k_1 et k_2 degrés de liberté, désignée par $F_{\text{à k1 et k2 dl}}$.

c est une constante telle que : $\int_0^{+\infty} f(x)dx = 1$

La variable F de Fisher correspond aussi au quotient de 2 variables Khi deux respectivement à k_1 et k_2 degrés de liberté $\chi^2_{\text{à k1 dl}}$ et $\chi^2_{\text{à k2 dl}}$ indépendantes.

Soient deux variables Khi deux $\chi^2_{\text{à k1 dl}}$ et $\chi^2_{\text{à k2 dl}}$ indépendantes. On peut démontrer :

$$F_{\text{àk1etk2dl}} = \frac{\chi^2_{\text{àk1dl}} / k1}{\chi^2_{\text{àk2dl}} / k2}$$

Il en résulte que si F est une variable $F_{\text{à k1 et k2 dl}}$, son inverse $\frac{1}{F}$ est une variable $F_{\text{à k2 et k1 dl}}$.

3.3.2. Caractéristiques de la loi $F_{\text{à k1 et k2 dl}}$

On peut démontrer que :

- Espérance mathématique : $E(F_{\text{à k1 et k2 dl}}) = \frac{k_2}{k_2 - 2}$ pour $k_2 > 2$.
- Variance : $V(F_{\text{à k1 et k2 dl}}) = \frac{2k_2{}^2 \times (k_1 + k_2)}{k_1(k_2 - 2)^2(k_2 - 4)}$ pour $k_2 > 4$.

3.3.3. Tables de la loi F de Fisher

Il y a plusieurs tables de la loi F de Fisher pour différentes valeurs de la fonction de répartition $F(F_{\text{à k1 et k2 dl}})$.

Chaque table de la loi F de Fisher dépend des paramètres k_1 et k_2, elle donne les valeurs de $F_{\text{à k1 et k2 dl}}$ pour la valeur de la fonction de répartition $F(F_{\text{à k1 et k2 dl}})$.

TABLE DE LA LOI F DE FISHER (p = 0,95)

K1 k2	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	4,96	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	4,94	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	4,89	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	4,86	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

TABLE DE LA LOI F DE FISHER (p = 0,975)

K1 k2	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	39,2	39,3	39,3	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

Pour lire une valeur $F_{\text{à k1 et k2 dl}}$ dans la table, il suffit de lire l'intersection entre la colonne correspondante à la valeur de k_1 et la ligne correspondante à la valeur de k_2.

Exemple :

La valeur de $F_{\text{à 10 et 15 dl}}$ pour une probabilité de 0,95 correspond dans la table de la loi F pour p=0,95, à l'intersection entre la colonne correspondante à 10 et la ligne correspondante à 15, on peut lire la valeur 2,54.

$$F_{0,95 \text{ à } 10 \text{ et } 15 \text{ dl}} = 2,54$$

$$F_{0,975 \text{ à } 15 \text{ et } 20 \text{ dl}} = 2,57$$

EXERCICES SUR LES LOIS DE PROBABILITE

1. Une confiture peut être qualifiée de "pure sucre" si elle contient entre 440 et 520 grammes de sucre par kilogramme de confiture. Un fabricant vérifie 200 pots de confiture de 1 kilogramme chacun. Il trouve que le poids moyen de sucre est de 480 grammes avec un écart type de 20 grammes. Sachant que le poids en sucre est distribué normalement, calculer le pourcentage de la production du fabriquant qui ne doit pas porter la mention "pur sucre" en considérant que l'échantillon des 200 pots est représentatif de la production globale.

2. Une machine met du sucre en poudre en sachet. Elle peut être réglée au moyen d'un dispositif gradué en gramme, tel que lorsque la machine est réglée sur le poids moyen par sachet m, la probabilité que les sachets pèsent au moins 1 Kg est égale à 98,5 %. Sachant que le poids par sachet suit une loi normale d'écart type 10 grammes, sur quelle valeur m faut-il régler le dispositif ?

3. Une machine est réglée pour faire remplir des bouteilles d'un volume moyen de 255 cm3. Si la distribution des volumes est normale et l'écart type est égal à 4 cm3 : (a) dans quelle proportion des cas le volume sera inférieur à 250 cm3 ? (b) quelle valeur faut-il donner au volume moyen pour que cette proportion soit de 5 % ?

4. Dans le cadre de la gestion d'un stock de marchandise, on doit lancer une commande destinée à couvrir quatre semaines de fourniture d'un produit donné. On admet que la demande hebdomadaire de ce produit suit une loi normale de moyenne, 50 et d'écart type 10. Combien d'unités doit-on commander pour que la probabilité d'être en rupture de stock soit inférieure à 1 % si on considère que les demandes des semaines successives sont indépendantes ?

5. Trouver la probabilité qu'au moins 70 de 1 00 moustiques seront tués par un nouvel insecticide si l'on sait que la probabilité que n'importe quel moustique soit tué est voisine de 0,75.

6. Si Ul et U2 Sont deux variables aléatoires normales centrées, réduites et indépendantes, calculer : (a) p(u1>u2), (b) p(u1+2u2>5), (c) calculer k tel que p(Ul+kU2>2)= 0,05.

7. Quelle est la valeur de la variable aléatoire X si $p(X<x) = 0,975$ et si la variable aléatoire X est : (a) une variable normale centrée réduite; (b) une variable normale de moyenne 10 et d'écart type 2 ; (c) une variable de Student à 50 degrés de liberté ; (d) une variable Khi deux à 60 degrés de liberté ; (e) une variable de Fisher à 25 et 20 degrés de Libertés.

8. Si Z1, Z2, …, Zk sont k variables aléatoires normales réduites indépendantes, que valent la moyenne et la variance de la variable :

$$X = \frac{Z1}{\sqrt{\sum_{i=2}^{k} Zi^2}}$$

et quelle est, pour k = 10, la valeur de x telle que : $P(X > x) = 0,1$?

9. Déterminez la valeur de la médiane de la distribution Khi carré à deux degrés de liberté.

10. Pour une variable Khi carré à 40 degrés de liberté, déterminez les valeurs $\chi^2 1$ et $\chi^2 2$ telles que : $F(\chi^2 1) = 0,05$ et $F(\chi^2 2) = 0,95$.

PREMIERE PARTIE

THEORIE D'ECHANTILLONNAGE

THEORIE D'ECHANTILLONNAGE

I. Rôle de l'échantillonnage

Lorsqu'on souhaite collecter les informations sur une population, deux possibilités s'offrent : La première solution consiste à observer ou interroger tous les éléments de la population, c'est ce qu'on appelle une enquête complète ou enquête exhaustive ou recensement. La seconde solution consiste à observer ou interroger une partie de la population, c'est ce qu'on appelle enquête partielle ou sondage. Les éléments de la population qui sont réellement observés constituent l'échantillon et l'opération qui consiste à choisir ces éléments est appelée échantillonnage.

L'alternative décrite ci-dessus se présente dans beaucoup de situations et le recours à la deuxième solution c'est à dire l'enquête partielle et la pratique la plus courante.

Par rapport à l'enquête complète, l'enquête partielle offre une série d'avantages. Le coût global de l'enquête partielle est en général plus réduit que le coût global d'une enquête complète. L'enquête par sondage est plus rapide que l'enquête complète, surtout lorsque la caractéristique étudiée présente des modifications assez importantes au cours du temps. Les erreurs d'observations sont plus réduites que dans l'enquête exhaustive. En fin dans certaines situations particulières, l'enquête partielle est la seule solution possible, c'est le cas lorsque l'observation présente un caractère destructif.

II. VOCABULAIRE

Enquête : ensemble des opérations de collecte et de traitement de données relatives à quelques domaines que ce soit.

Population : rassemblement de tous les cas qui répondent à un ensemble de caractères spécifiques. Appelée aussi univers ou ensemble statistique, c'est l'ensemble des éléments auxquels on s'intéresse.

Unité de base : unité d'échantillonnage ou unité de sondage, c'est l'élément pris en considération dans l'enquête.

Recensement : Enquête complète ou enquête exhaustive, c'est une enquête au cours de laquelle toutes les unités de base de la population sont observées.

Sondage : Enquête incomplète, enquête partielle ou enquête par échantillonnage, c'est une enquête au cours de laquelle seulement une partie des unités de base de la population sont observée.

Echantillon : ensemble des unités de base sélectionnées et réellement observées au cours d'un sondage.

Echantillonnage : ensemble des opérations qui permettent de sélectionner de façon organisée les éléments de l'échantillon.

Base de sondage : énumération ou présentation ordonnée de toutes les unités de base constituant la population.

Erreur d'échantillonnage : écart entre les résultats obtenus auprès d'un échantillon et ce que nous apprendrait un recensement comparable de la population. Plus la taille de l'échantillon est grande plus l'erreur d'échantillonnage diminue.

Fraction ou taux de sondage : proportion des unités de la population qui font partie de l'échantillon. C'est le rapport entre la taille de l'échantillon n, et la taille de la population N.

$$f = \frac{n}{N} \text{x}100$$

III. METHODES D'ECHANTILLONNAGE

Pour que les résultats d'une enquête par sondage puissent être extrapolés à l'ensemble de la population faisant l'objet de l'étude, il est indispensable que cette enquête soit conduite selon des règles bien définies et que les calculs conduisant à ces extrapolations soient conformes à la procédure d'échantillonnage utilisée.

L'échantillon choisi doit être le plus représentatif possible de la population étudiée, c'est à dire le degré de correspondance entre l'information recueillie et ce que nous apprendrait un recensement comparable de la population dépend en grande partie de la façon dont l'échantillon a été choisi.

La théorie moderne de l'échantillonnage nous propose une distinction fondamentale entre échantillons basés sur la probabilité : échantillons probabilistes; et échantillons non basés sur la probabilité : échantillons non probabilistes ou empiriques.

3.1. METHODES D'ECHANTILLONNAGE PROBABILISTES

3.1.1. Echantillonnage aléatoire et simple

Un échantillonnage est aléatoire si tous les individus de la population ont la même chance de faire partie de l'échantillon; il est simple si les prélèvements des individus sont réalisés indépendamment les uns des autres.

En particulier, si la population est finie, cette définition correspond au tirage aléatoire avec remise, qui permet de traiter les populations finies comme des populations infinies.

Pour prélever un échantillon aléatoire et simple il faut :

- Constituer la base de sondage qui correspond à la liste complète et sans répétition des éléments de la population ;
- Numéroter ces éléments de 1à N ;
- Procéder, à l'aide d'une table de nombres aléatoires ou d'un générateur de nombres pseudo aléatoires à la sélection des unités différentes qui constitueront l'échantillon.

Exemple :

On souhaite avoir un échantillon aléatoire et simple de 5 entreprises parmi une population de 22 entreprises. On dispose de la base de sondage c'est à dire la liste complète et sans répétitions des 22 entreprises numérotées de 1 à 22. On prend un extrait d'une table de nombre aléatoire par exemple :

10480	15011	01536	02011	81647	91646
22368	46573	25595	85393	30995	89198
24130	48390	22527	97265	76393	64809
42167	93093	06243	61680	07856	16376
37570	39975	81837	16656	**06121**	91782
77921	06907	11008	42751	27756	53498

On choisit au hasard un nombre de la table, supposons ce nombre 06121. Comme N= 22, on va retenir le premier groupe de 2 chiffres, ce qui donne les N° :
06, ensuite 12 ; 19 ; 17 ; les nombres (82,77 et 92) sont inutilisables. La cinquième entreprise sera le N° 10.

3.1.2. Echantillonnage stratifié

L'échantillonnage stratifié est une technique qui consiste à subdiviser une population hétérogène, d'effectif N, en P sous populations ou « strates » plus homogènes d'effectif Ni de telle sorte que $N= N_1+N_2+\ldots\ldots +Np$. Un échantillon, d'effectif ni, est par la suite, prélevé indépendamment au sein de chacune des strates en appliquant un plan d'échantillonnage au choix de l'utilisateur. Le plus souvent, on procédera par un échantillonnage aléatoire et simple à l'intérieur de chaque strate.

La stratification peut entraîner des gains de précision appréciables, elle facilite en outre les opérations de collecte des données et fournit des informations pour différentes parties de la population.

Pour la répartition de l'effectif total, n, de l'échantillon dans les différentes strates, La première solution, dite proportionnelle, consiste à conserver la même fraction d'échantillonnage dans chaque strate. Une seconde solution, dite optimale, tient compte du budget de l'enquête.

a) Répartition proportionnelle

La répartition optimale consiste à répartir la taille de l'échantillon n en utilisant la même fraction de sondage f dans chacune des strates. Cette solution tient compte d'un seul facteur qui est le poids de chaque strate.

Désignons par w_i le poids de la strate et par f la fraction de sondage constante.

$$f=\frac{n}{N} \qquad\qquad wi = \frac{Ni}{N}$$

le nombre d'unités à choisir dans chacune des strates est donc :

$$ni = wi \times n = f \times Ni$$

Exemple :

Dans une population de 10000 entreprises, réparties en 500 petites entreprises, 3000 moyennes entreprises et 2000 grandes entreprises, on souhaite avoir un échantillon de 500 entreprises.

Fraction de sondage constante : f = 500 / 1000 = 0.05 %

Strate	Effectif de la strate	Taille de l'échantillon
Petite	5000	5000 * 0,05 = 250
Moyenne	3000	3000 * 0,05 = 150
Grande	2000	2000 * 0,05 = 100
Total	10000	500

b) Répartition optimale

Cette deuxième solution consiste à répartir l'effort d'échantillonnage de façon inégale dans les différentes strates. Elle tient compte de quatre facteurs :

- Budget total de l'enquête, G
- Poids de la strate, w_i
- Coût de la collecte de l'information dans la strate, c_i
- Dispersion à l'intérieur de la strate, mesurée par l'écart type σ_i.

le nombre d'unités à choisir dans chacune des strates est donné par :

$$ni = k \frac{wi\ \sigma i}{\sqrt{ci}} \qquad \text{avec} \quad k = \frac{G}{\sum wi\ \sigma i\ \sqrt{ci}}$$

Exemple :

Dans la population des 10000 entreprises, on a pu avoir les informations suivantes :

Strate	Poids de la strate w_i	Coût de la collecte de l'information dans la strate, c_i	Dispersion à l'intérieur de la strate, mesurée par l'écart type σ_i.
Petite	0,5	50	0,8
Moyenne	0,3	75	1,5
Grande	0,2	100	2,2

le nombre d'entreprises à choisir dans chacune des strates est donné par :

$$k = \frac{G}{\sum wi\ \sigma i\ \sqrt{ci}} = \frac{5000}{0,5\times0,8\times\sqrt{50} + 0,3\times1,5\times\sqrt{75} + 0,2\times2,2\times\sqrt{100}} = 449,42$$

$$n1 = 449,42 \times \frac{0,5\times0,8}{\sqrt{50}} = 26 \text{ petites entreprises}$$

$$n1 = 449,42 \times \frac{0,3\times1,5}{\sqrt{75}} = 24 \text{ moyennes entreprises}$$

$$n1 = 449,42 \times \frac{0,2\times2,2}{\sqrt{100}} = 20 \text{ grandes entreprises}$$

3.1.3. ECHANTILLONNAGE PAR DEGRES

L'échantillonnage par degrés regroupe toute une série de plans d'échantillonnage caractérisés par un système ramifié et hiérarchisé d'unités.

Dans le cas de deux degrés, par exemple, on considère que la population est constituée d'un certain nombre d'unités de sondage du premier degré (unités primaires), chacune de ces unités étant constituée d'un certain nombre d'unités du second degré. (unités secondaires)
On réalise d'abord un échantillonnage d'unités du premier degré. Ensuite, dans chaque unité sélectionnée au premier degré, on prélève un échantillon d'unités du second degré. Le mode de sélection pouvant varier d'un degré à l'autre.

L'échantillonnage par degrés s'impose lorsqu'il est impossible d'inventorier les éléments de toute la population et qu'il est possible d'énumérer les unités prélevées au premier degré. Il permet une concentration du travail sur le terrain et donc une réduction des coûts.

Pour un même nombre total d'observations, il faut citer sa plus faible efficacité que l'échantillonnage aléatoire et simple.

Exemple :

Pour étudier le niveau de consommation des ménages d'une ville, on a tiré aléatoirement 5 quartiers. Dans chaque quartier sélectionné, on retient une rue sur 5, dans chaque rue retenue, on retient un immeuble sur 3, et dans chaque immeuble, un ménage par étage sera questionné.

3.1.4. Echantillonnage systématique

L'échantillonnage systématique est une technique qui consiste à prélever des unités d'échantillonnage situées à intervalles égaux. Le choix du premier individu détermine la composition de tout l'échantillon.

Si on connaît l'effectif total de la population N et qu'on souhaite prélever un échantillon d'effectif n, l'intervalle entre deux unités successives à sélectionner est donné par :

$$k=\frac{N}{n}$$ (arrondi à l'entier le plus proche)

Connaissant k, on choisit le plus souvent, pour débuter, un nombre aléatoire, i, compris entre 1 et k. le rang des unités sélectionnées est alors i, i+2k, i+3k, …

L'échantillonnage systématique est facile à préparer et, en général facile à exécuter, il réduit le temps consacré à la localisation des unités sélectionnées.

Si les éléments de la population se présentent dans un ordre aléatoire (pas de tendance) l'échantillonnage systématique est équivalent à l'échantillonnage aléatoire et simple. Par contre si les éléments de la population présentent une tendance, l'échantillonnage systématique est plus précis que l'échantillonnage aléatoire.

Exemple :

On veut sélectionner un échantillon de 30 entreprises au sein d'une population de 1800 entreprises.

$$k=\frac{1800}{30}=60$$

Ainsi on va tirer une entreprise toutes les 60 en partant d'un nombre tiré aléatoirement entre 1 et 60.

Supposons ce nombre est le 15. On va donc sélectionner la 15^{éme} entreprise puis la 75^{éme,} la 135$^{éme.}$ jusqu'à la 1755^{éme} ce qui nous donnera l'échantillon de 30 entreprises.

3.2. METHODES D'ECHANTILLONNAGE EMPIRIQUES

3.2.1 Echantillonnage accidentel (De convenance)

Il s'agit d'un échantillon constitué d'individus qui se trouvaient accidentellement à l'endroit et au moment où l'information a été collectée.

Exemple :

- Enquêtés réalisées dans la rue, les lieux publics, en sortie de super marché …
- Questionnaires figurant dans les magasines et renvoyés spontanément.

Les échantillons accidentels ne peuvent être considérés représentatifs d'aucune population. Il est risqué de généraliser à une population donnée des résultats obtenus par un échantillon accidentel.

3.2.2. Echantillonnage à priori

C'est un échantillonnage par jugement à priori. Il consiste à sélectionner des individus dont on pense, avant de les interroger, qu'ils peuvent détenir l'information.

Le risque de ce type d'échantillonnage est de considérer des individus, apparemment représentatifs de la population étudiée.

3.2.3. Echantillonnage « Boule de neige »

Cette méthode est réservée aux populations composées d'individus dont l'identification est difficile ou qui possèdent des caractéristiques rares.

La méthode consiste à faire construire l'échantillon par les individus eux-mêmes. Il suffit d'en identifier un petit nombre initial et de leur demander de faire appel à d'autres individus possédant les mêmes caractéristiques.

3.2.4. Echantillonnage par Quotas.

L'échantillonnage par quotas est l'échantillonnage non probabiliste le plus connu, et finalement le mieux accepté comme substitut aux méthodes probabilistes dans le cas où ces dernières rencontreraient des contraintes de base de sondage. Mais la représentativité de la population étudiée reste douteuse.

L'échantillonnage par quotas consiste à étudier la structure de la population selon des critères choisis (quotas) empiriquement. L'échantillon est ensuite construit de manière à constituer une reproduction en miniature de la population sur ces critères.

L'échantillonnage par quotas est une forme simplifiée de l'échantillonnage stratifié à fraction de sondage constante. Les quotas représentent les variables de stratification.

Une fois les quotas sont fixés, les individus sont sélectionnés à la convenance de l'enquêteur.

Les critères servant de base à la définition des quotas ne doivent pas être nombreux. Au-delà de 3 critères, la démarche devient complexe. Les quotas doivent être construits sur une base de données fiables (statistiques disponibles) indiquant la répartition de la population sur les critères choisis. Les critères les plus utilisés dans les études de marché sont économiques et socio-démographiques en particulier l'âge, le sexe, la catégorie socioprofessionnelle, …

Exemple :

On souhaite avoir un échantillon de 1000 individus. La structure de la population selon trois critères est la suivante :

1) Age

Age	Structure de la population	Répartition de l'échantillon
20 à 29 ans	40 %	400
30 à 49 ans	35 %	350
50 à 60 ans	25 %	250
Total	100 %	1000

2) Sexe x Age

Structure de la population

Age Sexe	Masculin	Féminin	Total
20 à 29 ans	48 %	52 %	100 %
30 à 49 ans	49 %	51 %	100 %
50 à 60 ans	45 %	55 %	100 %

Répartition de l'échantillon

Age Sexe	Masculin	Féminin	Total
20 à 29 ans	192	208	400
30 à 49 ans	172	178	350
50 à 60 ans	113	137	250

3) Age x Sexe x Catégorie socioprofessionnelle

Structure de la population

AGE	CSP / Sexe	Sans	Etudiant	Agric	Artisans	Prof libérales	Employés	Ouvriers	Total
20-29	M	10%	30%	5%	6%	9%	25%	15%	100%
	F	15%	25%	2%	10%	8%	30%	10%	100%
30-49	M	8%	5%	15%	22%	15%	15%	20%	100%
	F	20%	4%	10%	16%	14%	24%	12%	100%
50-60	M	6%	2%	25%	22%	18%	17%	10%	100%
	F	35%	1%	20%	20%	6%	13%	5%	100%

Répartition de l'échantillon

AGE	CSP / Sexe	Sans	Etudiant	Agric	Artisans	Prof libérales	Employés	Ouvriers	Total
20-29	M	19	58	10	12	17	48	28	192
	F	31	52	4	21	17	62	21	208
30-49	M	14	9	26	38	26	26	33	172
	F	36	7	18	28	25	43	21	178
50-60	M	7	2	28	25	20	19	12	113
	F	48	1	27	27	8	18	8	137

IV. DETERMINATION DE LA TAILLE DE L'ECHANTILLON

Le nombre n'est pas une garantie absolue de représentativité. La détermination de la taille d'échantillon dépend essentiellement de deux facteurs :

- La précision souhaitée : plus on souhaite des résultats précis, plus l'échantillon nécessaire est important.
- Le budget disponible : plus on augmente la taille, plus le coût de l'enquête s'accroît.

La taille de l'échantillon doit être celle qui permet d'atteindre le meilleur équilibre entre le risque de commettre des erreurs d'échantillonnage, le coût induit par ces erreurs, et le coût de l'échantillonnage lui-même.

Afin de déterminer la taille de l'échantillon, nous utiliserons l'inégalité de Bienaymé Tchebycheff ou la loi normale.

4.1. UTILISATION DE L'INEGALITE DE BIENAYME TCHEBYCHEFF

Cette inégalité n'est utilisée que si la loi de la variable aléatoire est complètement inconnue. Elle aboutit à des échantillons de taille élevée.

4.1.1. Taille d'échantillon pour estimer une moyenne.

- La taille de l'échantillon dépend de la précision souhaitée pour la généralisation des résultats.
- La précision (ou erreur d'échantillonnage) s'exprime en valeur absolue ou relative. Elle représente la largeur de l'intervalle de confiance de la moyenne. Soit ε la moitié de cette largeur.

L'inégalité de Biénaymé Tchebycheff dans le cas de la moyenne s'écrit :

$$P(\left|\bar{X}-m\right| < \varepsilon) \geq 1-\frac{\sigma^2}{n\varepsilon^2}$$

avec :

n : taille de l'échantillon ;
ε : précision souhaitée ;
$\bar{X}$: moyenne de l'échantillon ;
m : moyenne de la population.
σ : Ecart- type d'échantillon, il est souvent inconnu, il faut avoir des informations antérieures ou mener une étude pilote.

Pour obtenir un maximum de fiabilité dans les résultats, on commence par se fixer une marge d'erreur "ε" que l'on accepte. On se fixe ensuite un seuil de confiance (1-α), qui représente la probabilité minimale pour que la moyenne calculée à partir de l'échantillon ne s'écarte pas de la moyenne de la population de plus de ε. Ceci s'écrit :

$$P(\left|\bar{X}-m\right| < \varepsilon) \geq 1-\alpha$$

En rapprochant les deux formules on obtient :

$$1-\frac{\sigma^2}{n\varepsilon^2} = 1-\alpha$$

et donc :

$$n = \frac{\sigma^2}{\varepsilon^2 \times \alpha}$$

Exemple :

Un parc de loisirs souhaite estimer à 10dh prés le montant moyen d'achats effectués par chaque visiteur, c'est à dire on se fixe une marge d'erreur de 10 dans l'analyse des résultats :

ε = 10

Une étude pilote menée sur 50 visiteurs choisis au hasard a montré que l'écart- type des achats est : σ = 100 dh.

Si on se fixe un seuil de confiance (1-α) = 95%, La taille de l'échantillon est donc :

$$n = \frac{100^2}{10^2 \times 0{,}05} = 2000$$

4.1.2. Taille d'échantillon pour estimer une proportion

- La taille de l'échantillon dépend de la précision souhaitée pour la généralisation des résultats.
- La précision (ou erreur d'échantillonnage) s'exprime en valeur absolue ou relative. Elle représente la largeur de l'intervalle de confiance de la proportion. Soit ε la moitié de cette largeur.

l'inégalité de Biénaymé Tchebycheff dans le cas de la proportion s'écrit :

$$P(|f_n - p| < \varepsilon) \geq 1 - \frac{pq}{n\varepsilon^2}$$

avec :

n : taille de l'échantillon ;
ε : précision souhaitée ;
f_n : proportion ou fréquence relative dans l'échantillon ;
p : proportion dans la population (q = 1 – p). Elle est souvent inconnue, il faut avoir des informations antérieures ou mener une étude pilote, sinon on utilise une proportion de 50 %.

Pour obtenir un maximum de fiabilité dans les résultats, on commence par se fixer une marge d'erreur "ε" que l'on accepte. On se fixe ensuite un seuil de confiance (1-α), qui représente la probabilité minimale pour que la fréquence calculée à partir de l'échantillon ne s'écarte pas de la proportion dans la population de plus de ε. Ceci s'écrit :

$$P(|f_n - p| < \varepsilon) \geq 1-\alpha$$

En rapprochant les deux formules on obtient : $1 - \frac{pq}{n\varepsilon^2} = 1-\alpha$

et donc :

$$n = \frac{pq}{\varepsilon^2 \times \alpha}$$

Exemple :

Le parc souhaite estimer la proportion des visiteurs qui font des achats à cinq points prés, c'est à dire on se fixe une marge d'erreur de 5% dans l'analyse des résultats :

ε = 0,05

L'enquête pilote a estimé cette proportion à 65%, c'est à dire p = 0,65

Si on se fixe un seuil de confiance (1-α) = 95%, la taille de l'échantillon est donc :

$$n = \frac{0,65 \times 0,35}{0,05^2 \times 0,05} = 1820$$

4.2. UTILISATION DE LA LOI NORMALE

On applique cette méthode si la variable suit une loi normale ou si elle peut être approchée par la loi normale.

4.2.1. Taille d'échantillon pour estimer une moyenne

a) Cas des prélèvements dans une population finie avec remise ou dans une population infinie sans remise :

Pour obtenir un maximum de fiabilité dans les résultats, on commence par se fixer une marge d'erreur "ε" que l'on accepte. On se fixe ensuite un seuil de confiance (1-α), qui représente la probabilité minimale pour que la moyenne calculée à partir de l'échantillon ne s'écarte pas de la moyenne de la population de plus de ε. Ceci s'écrit :

$$P(\left|\bar{X} - m\right| < \varepsilon) \geq 1-\alpha$$

avec :

ε : précision souhaitée ;
$\bar{X}$: moyenne de l'échantillon ;
m : moyenne de la population.

D'après le théorème central limite, la variable aléatoire $\bar{X}$ suit une loi normale dont les paramètres sont :

$$E(\bar{X}_n) = m$$

$$V(\bar{X}_n) = \frac{\sigma^2}{n}$$

L'écart type de la moyenne est donc : $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$

Déterminer la taille de l'échantillon consiste à résoudre l'équation :

$$P(\left|\bar{X} - m\right| < \varepsilon) \geq 1-\alpha$$

$$P(-\varepsilon < \bar{X} - m < \varepsilon) \geq 1-\alpha$$

$$P(-\frac{\varepsilon}{\frac{\sigma}{\sqrt{n}}} < \frac{\bar{X}-m}{\frac{\sigma}{\sqrt{n}}} < \frac{\varepsilon}{\frac{\sigma}{\sqrt{n}}}) \geq 1-\alpha$$

$$P(-\frac{\varepsilon\sqrt{n}}{\sigma} < Z < \frac{\varepsilon\sqrt{n}}{\sigma}) \geq 1-\alpha$$

$$\Pi(\frac{\varepsilon\sqrt{n}}{\sigma}) - \Pi(-\frac{\varepsilon\sqrt{n}}{\sigma}) \geq 1-\alpha$$

$$\Pi(\frac{\varepsilon\sqrt{n}}{\sigma}) - [1 - \Pi(\frac{\varepsilon\sqrt{n}}{\sigma})] \geq 1-\alpha$$

$$2\Pi(\frac{\varepsilon\sqrt{n}}{\sigma}) - 1 \geq 1-\alpha$$

$$\Pi(\frac{\varepsilon\sqrt{n}}{\sigma}) \geq 1 - \frac{\alpha}{2}$$

On se reporte à la table de distribution de la loi Normale centrée réduite, et on cherche la valeur correspondante à une probabilité égale à 1- $\frac{\alpha}{2}$, cette valeur de z sera désignée par $Z_{1-\frac{\alpha}{2}}$
On a alors :

$$\frac{\varepsilon\sqrt{n}}{\sigma} = Z_{1-\frac{\alpha}{2}}$$

$$n = Z^2_{1-\frac{\alpha}{2}} \frac{\sigma^2}{\varepsilon^2}$$

Exemple :

Reprenons l'exemple du parc de loisirs qui souhaite estimer à 10dh prés le montant moyen d'achats effectués par chaque visiteur, c'est à dire on se fixe une marge d'erreur de 10 dans l'analyse des résultats : $\varepsilon = 10$

Une étude pilote menée sur 50 visiteurs choisis au hasard a montré que l'écart- type des achats est : $\sigma = 100$ dh.

Si on se fixe un seuil de confiance $(1-\alpha) = 95\%$, La taille de l'échantillon est donc :

$$n = 1{,}96^2 \frac{100^2}{10^2} = 384{,}16 = 385$$

b) Cas des prélèvements dans une population finie sans remise :

$$E(\bar{X}_n) = m$$

$$V(\bar{X}_n) = \frac{N-n}{N-1}\frac{\sigma^2}{n}$$

L'écart type de la moyenne est donc : $\sigma_{\bar{X}} = \sqrt{\frac{N-n}{N-1}}\frac{\sigma}{\sqrt{n}} \approx \frac{\sigma}{\sqrt{n}}\sqrt{1-\frac{n}{N}}$

De la même manière, on arrive à :

$$\frac{\varepsilon\sqrt{n}}{\sigma}\sqrt{\frac{N}{N-n}} = Z_{1-\frac{\alpha}{2}}$$

$$\sqrt{\frac{n}{N-n}} = Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\varepsilon\sqrt{N}}$$

$$\frac{n}{N-n} = Z^2_{1-\frac{\alpha}{2}} \times \frac{\sigma^2}{\varepsilon^2 N}$$

$$n = Z^2_{1-\frac{\alpha}{2}}\frac{\sigma^2}{\varepsilon^2} - n\, Z^2_{1-\frac{\alpha}{2}}\frac{\sigma^2}{\varepsilon^2 N}$$

$$n\left(1 + Z^2_{1-\frac{\alpha}{2}}\frac{\sigma^2}{\varepsilon^2 N}\right) = Z^2_{1-\frac{\alpha}{2}}\frac{\sigma^2}{\varepsilon^2}$$

$$n = \frac{Z^2_{1-\frac{\alpha}{2}}\ \sigma^2\ N}{\varepsilon^2\ N + Z^2_{1-\frac{\alpha}{2}}\ \sigma^2}$$

4.2.2. Taille d'échantillon pour estimer une proportion.

Pour obtenir un maximum de fiabilité dans les résultats, on commence par se fixer une marge d'erreur "ε" que l'on accepte. On se fixe ensuite un seuil de confiance (1-α), qui représente la probabilité minimale pour que la fréquence calculée à partir de l'échantillon ne s'écarte pas de la proportion dans la population de plus de ε. Ceci s'écrit :

$$P(|f_n - p| < \varepsilon) \geq 1-\alpha$$

avec :

n : taille de l'échantillon ;
ε : précision souhaitée ;
f_n : proportion ou fréquence relative dans l'échantillon ;
p : proportion dans la population (q = 1 – p). Elle est souvent inconnue, il faut avoir des informations antérieures ou mener une étude pilote, sinon on utilise une proportion de 50 %.

D'après le théorème central limite, la variable aléatoire f_n suit une loi normale dont les paramètres sont :

a) Cas des prélèvements dans une population finie avec remise ou dans une population infinie sans remise :

$$E(f_n) = p$$

$$V(f_n) = \frac{pq}{n}$$

L'écart type de la fréquence est donc : $\sigma_{f_n} = \frac{\sqrt{pq}}{\sqrt{n}}$

Déterminer la taille de l'échantillon consiste à résoudre l'équation :

$$P(|f_n - p| < \varepsilon) \geq 1-\alpha$$

$$P(-\varepsilon < f_n - p < \varepsilon) \geq 1-\alpha$$

$$P\left(-\frac{\varepsilon}{\frac{\sqrt{pq}}{\sqrt{n}}} < \frac{f_n - p}{\frac{\sqrt{pq}}{\sqrt{n}}} - < \frac{\varepsilon}{\frac{\sqrt{pq}}{\sqrt{n}}}\right) \geq 1-\alpha$$

$$P\left(-\frac{\varepsilon\sqrt{n}}{\sqrt{pq}} < Z < \frac{\varepsilon\sqrt{n}}{\sqrt{pq}}\right) \geq 1-\alpha$$

$$\Pi\left(\frac{\varepsilon\sqrt{n}}{\sqrt{pq}}\right) - \Pi\left(-\frac{\varepsilon\sqrt{n}}{\sqrt{pq}}\right) \geq 1-\alpha$$

$$\Pi\left(\frac{\varepsilon\sqrt{n}}{\sqrt{pq}}\right) - [1 - \Pi\left(\frac{\varepsilon\sqrt{n}}{\sqrt{pq}}\right)] \geq 1-\alpha$$

$$2\Pi\left(\frac{\varepsilon\sqrt{n}}{\sqrt{pq}}\right) - 1 \geq 1-\alpha$$

$$\Pi\left(\frac{\varepsilon\sqrt{n}}{\sqrt{pq}}\right) \geq 1 - \frac{\alpha}{2}$$

On se reporte à la table de distribution de la loi Normale centrée réduite, et on cherche la valeur correspondante à une probabilité égale à $1 - \frac{\alpha}{2}$, cette valeur de z sera désignée par $Z_{1-\frac{\alpha}{2}}$

On a alors :

$$\frac{\varepsilon\sqrt{n}}{\sqrt{pq}} = Z_{1-\frac{\alpha}{2}}$$

$$n = Z^2_{1-\frac{\alpha}{2}} \frac{pq}{\varepsilon^2}$$

Exemple :

Reprenons l'exemple du parc de loisirs qui souhaite estimer la proportion des visiteurs qui font des achats à cinq points prés, c'est à dire on se fixe une marge d'erreur de 5% dans l'analyse des résultats :

ε = 0,05

L'enquête pilote a estimé cette proportion à 65%, c'est à dire p = 0,65

Si on se fixe un seuil de confiance (1-α) = 95%, on se reporte à la table de distribution de la loi Normale, et on cherche la valeur correspondante à une probabilité (1-α/2) = 0,975, ce qui donne Z = 1,96.

La taille de l'échantillon est donc :

$$n = 1{,}96^2 \frac{0{,}65\times0{,}35}{0{,}05^2} = 349{,}58 = 350$$

b) Cas des prélèvements dans une population finie sans remise :

$$E(f_n) = p$$

$$V(f_n) = \frac{N-n}{N-1}\frac{pq}{n}$$

L'écart type de la fréquence est donc : $\sigma_{f_n} = \frac{\sqrt{pq}}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} \approx \frac{\sqrt{pq}}{\sqrt{n}}\sqrt{1-\frac{n}{N}}$

De la même manière, on arrive à :

$$\frac{\varepsilon\sqrt{n}}{\sqrt{pq}}\frac{\sqrt{N}}{\sqrt{N-n}} = Z_{1-\frac{\alpha}{2}}$$

$$\sqrt{\frac{n}{N-n}} = Z_{1-\frac{\alpha}{2}}\frac{\sqrt{pq}}{\varepsilon\sqrt{N}}$$

$$\frac{n}{N-n} = Z^2_{1-\frac{\alpha}{2}} \times \frac{pq}{\varepsilon^2 N}$$

$$n = Z^2_{1-\frac{\alpha}{2}} \frac{pq}{\varepsilon^2} - n\, Z^2_{1-\frac{\alpha}{2}} \frac{pq}{\varepsilon^2 N}$$

$$n\left(1 + Z^2_{1-\frac{\alpha}{2}} \frac{pq}{\varepsilon^2 N}\right) = Z^2_{1-\frac{\alpha}{2}} \frac{pq}{\varepsilon^2}$$

$$n = \frac{Z^2_{1-\frac{\alpha}{2}}\ p\,q\,N}{\varepsilon^2\ N + Z^2_{1-\frac{\alpha}{2}}\ p\,q}$$

V. DISTRIBUTIONS D'ECHANTILLONNAGE

La notion de distribution d'échantillonnage est à la base des méthodes d'inférence statistique dont les deux principales applications sont les problèmes d'estimation et les tests d'hypothèses. Les premiers ont pour but d'estimer, à partir d'un échantillon, la valeur numérique d'un ou de plusieurs paramètres de la population, et de déterminer la précision de cette ou de ces estimations. Les seconds ont pour but de vérifier la véracité d'une hypothèse émise au départ au sujet d'une ou de plusieurs populations.

A tout paramètre de population θ, on peut associer une série infinie de valeurs observées t, t', t '', …, calculées à partir d'échantillons successifs de même effectif, prélevés dans des conditions identiques. Ces valeurs peuvent être considérées comme des valeurs observées d'une même variable aléatoire T, et cette variable est fonction des différentes variables aléatoires correspondant à chacun des individus de l'échantillon :

$$T = f(X_1, X_2, \ldots, Xn)$$

En supposant que l'échantillon est aléatoire et simple, la variable aléatoire T possède une distribution de probabilité, dite distribution d 'échantillonnage. On peut donc calculer l'espérance E(T) et la variance V(T) de cette distribution.

La distribution d'échantillonnage est donc la distribution des différentes valeurs que peut prendre la variable aléatoire T, pour les différents échantillons possibles. Son écart type σ_T est appelé erreur standard.

Les principales distributions d'échantillonnage sont la distribution d'échantillonnage de la moyenne, la distribution d'échantillonnage de la variance et la distribution d'échantillonnage de la proportion.

5.1. DISTRIBUTION D'ECHANTILLONNAGE DE LA MOYENNE

Supposons que dans une population infinie quelconque, on ait prélevé au hasard un premier échantillon de n observations :

$$x_1, x_2, x_3, \ldots\ldots\ldots, x_n$$

et qu'on ait calculé la moyenne : $\bar{x} = \frac{\sum_{i=1}^{n} xi}{n}$

Si on prélève, dans les mêmes conditions, un deuxième échantillon de même effectif :

$$x_1', x_2', x_3', \ldots\ldots\ldots, x_n'$$

La moyenne correspondante $\bar{x}' = \frac{\sum_{i=1}^{n} x'i}{n}$ sera généralement différente de la première moyenne observée.

Il en sera de même pour les moyennes d'autres échantillons prélevés dans les mêmes conditions :

$$x_1'', x_2'', x_3'', \ldots\ldots\ldots, x_n''$$

$$\bar{x}'' = \frac{\sum_{i=1}^{n} xi''}{n}$$

On peut considérer la suite des premières observations x_1, x_1', x_1'', … des différents échantillons comme des valeurs observées d'une même variable aléatoire X_1, la suite des deuxièmes observations des différents échantillons comme des valeurs observées d'une même variable aléatoire X_2, etc.

Les moyennes observées $\bar{x}$, $\bar{x}'$, $\bar{x}''$, … sont alors des valeurs observées d'une même variable aléatoire $\bar{X}$ qui est fonction de X_1, X_2, …, Xn.

$$\bar{X} = \frac{\sum_{i=1}^{n} Xi}{n}$$

Comme X_1, X_2, …, Xn, la variable aléatoire $\bar{X}$ possède une distribution de probabilité, dite distribution d 'échantillonnage de la moyenne. On peut donc calculer l'espérance et la variance de cette distribution, en supposant que l'échantillon est aléatoire et simple, les variables aléatoires X_1, X_2, …, Xn ont toutes la même distribution de probabilité, dont la moyenne est désignée par m et la variance par σ^2.

$E(Xi) = m$ et $V(Xi) = \sigma^2$

On démontre alors :

$$E(\bar{X}) = E\left(\frac{\sum_{i=1}^{n} Xi}{n}\right) = \frac{1}{n}\times\sum_{i=1}^{n} E(Xi) = \frac{1}{n}\times n\times m = m$$

$$V(\bar{X}) = V\left(\frac{\sum_{i=1}^{n} Xi}{n}\right) = \frac{1}{n^2}\times\sum_{i=1}^{n} V(Xi) = \frac{1}{n^2}\times n\times \sigma^2 = \frac{\sigma^2}{n}$$

$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$ est appelé erreur standard de la moyenne d'un échantillon aléatoire est simple

Dans le cas d'une population finie d'effectif N, au sein de laquelle est prélevé, sans remise, un échantillon aléatoire est simple d'effectif n, la variance de la moyenne est :

$$V(\bar{X}) = \frac{N-n}{N-1}\,\frac{\sigma^2}{n}$$

L'erreur standard est alors : $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$

En plus des caractéristiques de la distribution d'échantillonnage de la moyenne, on peut aussi rechercher la forme de cette distribution.

Si par exemple, la population parent possède une distribution normale, on peut affirmer que la distribution de la moyenne est elle-même normale de moyenne m et d'écart type $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$.

Si la distribution de la population parent est inconnue, le théorème central limite permet d'affirmer que la distribution de la moyenne est asymptotiquement normale. Pour un effectif suffisamment élevé, la moyenne d'un échantillon peut toujours être considérée comme une variable approximativement normale. C'est généralement le cas lorsque l'effectif est supérieur à 30. Dans le cas contraire (n < 30), la moyenne d'un échantillon peut toujours être considérée comme une variable de Student à (n-1) degré de liberté.

5.2. DISTRIBUTION D'ECHANTILLONNAGE DE LA VARIANCE

De la même manière que la moyenne, chacun des échantillons possède une variance :

$$v(x) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$$

$$v(x') = \frac{\sum_{i=1}^{n}(x_i' - \bar{x}')^2}{n}$$

$$v(x'') = \frac{\sum_{i=1}^{n}(x_i'' - \bar{x}')^2}{n}$$

Ces variances peuvent être considérées comme des valeurs observées d'une même variable aléatoire :

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

Comme X_1, X_2, …, Xn, la variable aléatoire V(X) possède une distribution de probabilité, dite distribution d 'échantillonnage de la variance. On peut donc calculer l'espérance mathématique et la variance de cette distribution, en supposant que l'échantillon est aléatoire et simple, les variables aléatoires X_1, X_2, …, Xn ont toutes la même distribution de probabilité, dont la moyenne est désignée par m et la variance par σ^2.

$E(Xi) = m$ et $V(Xi) = \sigma^2$

on peut démontrer alors :

$$E(V(X)) = E\left(\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n}\right) = E\left(\frac{\sum_{i=1}^{n}(X_i - m - \bar{X} + m)^2}{n}\right) = E\left(\frac{\sum_{i=1}^{n}[(X_i - m) - (\bar{X} - m)]^2}{n}\right)$$

$$E(V(X)) = E\left(\frac{\sum_{i=1}^{n}[(X_i - m)^2 - 2(X_i - m)(\bar{X} - m) + (\bar{X} - m)^2]}{n}\right)$$

$$E(V(X)) = E\left(\frac{\sum_{i=1}^{n}(X_i - m)^2 - 2(\bar{X} - m)\sum_{i=1}^{n}(X_i - m) + \sum_{i=1}^{n}(\bar{X} - m)^2}{n}\right)$$

$$E(V(X)) = E\left(\frac{\sum_{i=1}^{n}(Xi - m)^2}{n} - 2(\bar{X} - m)\frac{\sum_{i=1}^{n}(Xi - m)}{n} + \frac{\sum_{i=1}^{n}(\bar{X} - m)^2}{n}\right)$$

$$E(V(X)) = E\left(\frac{\sum_{i=1}^{n}(Xi - m)^2}{n} - 2(\bar{X} - m)(\bar{X} - m) + (\bar{X} - m)^2\right)$$

$$E(V(X)) = E(\sigma^2 - (\bar{X} - m)^2)$$

$E(V(X)) = E(\sigma^2) - E((\bar{X}-m)^2)$

$E(V(X)) = \sigma^2 - \frac{\sigma^2}{n}$

$$E(V(X)) = \frac{n-1}{n} \times \sigma^2$$

Pour la variance de la distribution d'échantillonnage de la variance, on démontre, dans le cas d'une population normale :

$$V(V(X)) = E[(V(X) - E(V(X)))^2] = E[(V(X) - \frac{n-1}{n}\sigma^2)^2] = \frac{2(n-1)}{n^2}\ \sigma^4.$$

Dans le cas d'une population finie d'effectif N, au sein de laquelle est prélevé, sans remise, un échantillon aléatoire est simple d'effectif n, l'espérance mathématique de la variance est :

$$E(V(X)) = \frac{N}{N-1} \times \frac{n-1}{n} \times \sigma^2$$

En ce qui concerne la forme de la distribution d'échantillonnage de la variance, on peut démontrer que dans le cas particulier d'une population normale, la variable aléatoire $\frac{\sum_{i=1}^{n}(xi-\bar{x})^2}{\sigma^2}$ possède une distribution khi deux à (n-1) degré de liberté.

5.3. DISTRIBUTION D'ECHANTILLONNAGE DE LA PROPORTION

Si on considère une population infinie et si on y prélève un échantillon aléatoire et simple d'effectif n, on désigne par x le nombre d'individus possédant, dans l'échantillon, le caractère étudié.

$f_n = \frac{X_n}{n}$ est la fréquence ou proportion des individus possédant, dans l'échantillon, le caractère étudié.

On désigne par p la proportion des individus possédant, dans la population, le caractère étudié.

De la même manière que la moyenne et la variance, chacun des échantillons possède une fréquence :

$$fn = \frac{Xn}{n}$$

$$f_n' = \frac{X_n'}{n}$$

$$f_n'' = \frac{X_n''}{n}$$

Ces fréquences peuvent être considérées comme des valeurs observées d'une même variable aléatoire :

$$F_n = \frac{X_n}{n}$$

La variable aléatoire F_n possède une distribution de probabilité, dite distribution d 'échantillonnage de la proportion. On peut donc calculer l'espérance et la variance de cette distribution, en supposant que l'échantillon est aléatoire et simple.

On peut démontrer alors :

$$E(F_n) = E(\frac{X_n}{n}) = \frac{1}{n} E(X_n) = \frac{1}{n} n\ p = p$$

$$V(F_n) = V(\frac{X_n}{n}) = \frac{1}{n^2} V(X_n) = \frac{1}{n^2} n\ p\ q = \frac{pq}{n}$$

$\sigma_{Fn} = \frac{\sqrt{pq}}{\sqrt{n}}$ est appelé erreur standard de la fréquence d'un échantillon aléatoire est simple

Dans le cas d'une population finie d'effectif N, au sein de laquelle est prélevé, sans remise, un échantillon aléatoire est simple d'effectif n, la variance de la fréquence est :

$$V(F_n) = V(\frac{X_n}{n}) = \frac{1}{n^2} V(X_n) = \frac{1}{n^2} \frac{N-n}{N-1} n\ p\ q = \frac{N-n}{N-1} \frac{pq}{n}$$

L'erreur standard est alors : $\sigma_{Fn} = \sqrt{\frac{N-n}{N-1}} \times \frac{\sqrt{pq}}{\sqrt{n}}$

En ce qui concerne la forme de cette distribution, on peut affirmer que la distribution de la proportion suit une loi normale de moyenne p et d'écart type $\sigma_{Fn} = \frac{\sqrt{pq}}{\sqrt{n}}$ à condition que la taille de l'échantillon soit supérieure ou égale à 30 ($n \geq 30$) et le produit $n\ p \geq 5$.

EXERCICES SUR LA THEORIE D'ECHANTILLONNAGE

1. Quelle est l'espérance mathématique et quelle est la variance des résultats qu'on peut obtenir quand on choisit au hasard et indépendamment dix nombres entiers de 1 à 9 et qu'on en calcule la moyenne, en supposant que chacun des nombres de 1 à 9 a une même probabilité d'être choisi et qu'un même nombre peut être choisi plusieurs fois sans aucune restriction ?

2. Quelle est la probabilité que la moyenne d'un échantillon de 12 observations provenant d'une population de distribution uniforme définie dans l'intervalle (0, 1) soit comprise entre 0,4 et 0,6 ?

3. Calculez la moyenne et l'écart type de la variance S^2, ainsi que la probabilité $P(10 < S^2 < 20)$, en supposant que S^2 désigne la variance observée d'échantillons aléatoires et simples d'effectif 10 extraits d'une population normale de moyenne égale à 15 et écart type égal à 4.

4. On suppose que les poids de 3000 étudiants d'une université suivent une loi normale de moyenne 68,0 kilogrammes et écart type 3,0 kilogrammes. Si l'on extrait 80 échantillons de 25 étudiants chacun, quelle est la moyenne et écart type théoriques de la distribution d'échantillonnage des moyennes pour (a) un échantillonnage non exhaustif, (b) un échantillonnage exhaustif ?

5. Pour combien d'échantillons du Problème 4 peut-on s'attendre à trouver une moyenne (a) comprise entre 66,8 et 68,3 kilogrammes, (b) inférieure à 66,4 kilogrammes ?

6. 500 pignons ont un poids moyen de 5,02 grammes et un écart type de 0,30 grammes. Trouver la probabilité pour qu'un échantillon de 100 pignons choisi au hasard ait un poids total (a) compris entre 496 et 500 grammes. (b) plus grand que 510 grammes.

7. Chacune des personnes d'un groupe de 500 individus lance 120 fois une pièce de monnaie parfaite. Combien de personnes signaleront-elles que (a) le nombre de faces qu'elles obtiennent se trouve compris entre 40 et 60. (b) 5 sur 8 ou plus de leurs jets correspondent à des faces ?

8. Lors d'élections, les résultats ont montré qu'un des candidats a obtenu 46 % des voix. Déterminer la probabilité pour que le vote de (a) 200 (b) 1000 personnes choisies au hasard parmi le corps électoral donne une majorité de voix en faveur de ce candidat.

9. Les ampoules électriques d'un fabricant A ont une durée de vie moyenne de 1400 heures avec un écart-type de 200 heures, et celles d'un fabricant B ont une durée de vie moyenne de 1200 heures avec un écart-type de 100 heures. Si l'on teste des échantillons de 125 ampoules pour chaque marque, quelle est la probabilité pour que la marque d'ampoules A ait une durée de vie moyenne qui soit au moins supérieure de (a) 160 heures, (b) 250 heures à celle de la marque d'ampoules B ?

10. Les pignons d'une marque donnée pèsent 0,50 gramme avec un écart-type de 0,02 gramme. Quelle est la probabilité pour que deux lots de 1000 pignons chacun diffèrent entre eux de plus de 2 grammes ?

11. Un certain type d'ampoule électrique a une durée de vie moyenne de 1500 heures et un écart type de 150 heures. Trois ampoules sont branchées de telle manière que, si l'une d'elles est grillée, les autres continuent à fonctionner. En supposant que les durées de vie suivent une loi de Laplace Gauss, quelle est la probabilité pour que l'éclairage fonctionne (a) au moins pendant 5000 heures. (b) au plus pendant 4200 heures ?

12. L'écart type des poids d'une très grande population de personnes est 10 kg On extrait de cette population des échantillons de 200 personnes chacun. On calcule alors les écarts types pour chaque échantillon. (a) Trouver la moyenne et l'écart type de la distribution d'échantillonnage des écarts types. (b) Quel est le pourcentage d'échantillons qui a un écart type plus grand que 11 Kg ?

13. Les poids de 1500 pignons suivent une loi de Laplace-Gauss de moyenne 22,40 kg et écart type 0,048 kg Déterminer pour 300 échantillons aléatoires de taille 36 de cette population la moyenne et l'écart-type théoriques de la distribution d'échantillonnage des moyennes, l'échantillonnage étant (a) non exhaustif, (b) exhaustif.

14. Combien d'échantillons aléatoires du Problème 13 ont-ils leur moyenne (a) comprise entre 22,39 et 22,41 Kg, (b) plus grande que 22,42 Kg, *(c)* plus petite que 22,37 Kg, (d) plus petite que 22,38 ou plus grande que 22,41 Kg ?

15. Les poids des colis reçus dans un grand magasin ont une moyenne de 300 kg et un écart-type de 50 kg, Quelle est la probabilité pour que 25 colis reçus au hasard et chargés sur un monte-charge dépassent la limite de sécurité du monte-charge, qui est 8200 kilogrammes.

16. Trouver la probabilité pour que parmi les 200 prochains enfants à naître (a) il y ait moins de 40 % de garçons, (b) il y ait entre 43 % et 57 % de filles, (c) il y ait plus de 54 % de garçons. On supposera que la naissance d'un garçon et la naissance d'une fille sont équiprobables.

17. Etant donné 1000 échantillons de 200 enfants chacun, pour combien d'échantillons a-t-on une chance de trouver (a) moins de 40 % de garçons, (b) entre 40 % et 60 % de filles, (c) 53 % ou plus de filles ?

18. Un fabricant expédie 1000 lots de 100 ampoules électriques chacun. Si *5* % des ampoules sont normalement défectueuses, dans combien de lots peut-on avoir (a) moins de 90 bonnes ampoules, (b) 98 bonnes ampoules ou davantage ?

19. A et B fabriquent deux types de câbles ayant comme charges de rupture respectives 4000 et 4500 kilogrammes avec des écarts-types de 300 et 200 kilogrammes. Si l'on teste 100 câbles de la marque A et 50 câbles de la marque B, quelle est la probabilité pour que la résistance de rupture moyenne de B *ait* (a) au moins 600 kilogrammes de plus que A, (b) au moins 450 kilogrammes de plus que A ?

20. Les résultats d'une élection montrent qu'un des candidats a obtenu 65 % des voix. Trouver la probabilité pour que deux échantillons aléatoires, chacun correspondant à 200 votants, indiquent plus de 10 % de différence dans les proportions de gens qui ont voté pour ce candidat.

21. Le voltage moyen d'une batterie est 15,0 volts avec un écart-type de 0,2 volt. Quelle est la probabilité pour que quatre batteries de ce type, branchées en série, aient un voltage combiné de 60,8 volts ou plus ?

22. Une firme fabrique un bien dont la durée de vie est en moyenne 1800 heures avec un écart type de 200 heures. (a) Trouver la probabilité qu'un échantillon aléatoire de 100 unités de ce bien a une moyenne de vie supérieure à 1825. (b) Trouver la probabilité qu'un échantillon aléatoire de 100 Unités de ce bien à une moyenne de vie de pas plus de 1775 et pas moins de 1760.

23. Une population est constituée des cinq nombres 2, 3, 6, 8, 11. On considère tous les échantillons non exhaustifs possibles de taille deux de cette population. Trouver (a) la moyenne de la population, (b) écart type de la population, (c) la moyenne de la distribution d'échantillonnage des moyennes, (d) écart type de la distribution d'échantillonnage des moyennes, c'est-à-dire l'erreur quadratique moyenne des moyennes.

24. résoudre le problème 23 dans le cas d'un échantillon exhaustif.

25. Dans le but d'étudier l'intention d'achat d'un produit, on décide de réaliser un sondage. Combien de personnes doit-on interroger pour que la fréquence empirique ne s'éloigne pas de la vraie proportion de 1% et ce avec une probabilité au moins égale à 95%?

26. Des sachets de sucre granulé, dont le poids moyen est de 1,01 kg avec un écart type de 50 grammes, sont mis dans des cartons contenant chacun 100 sachets. Le poids d'un carton vide est de 500 grammes. On procède par sondage au contrôle du poids des sachets de sucre granulé. (a) en utilisant l'IBT, déterminer le nombre de sachets de sucre granulé qu'on doit contrôler pour que le poids moyen de l'échantillon ne soit pas loin de la vraie moyenne de plus ou moins 20 grammes, avec une probabilité au moins égale à 0,99. (b) Reprendre la même question en supposant que le poids moyen est distribué normalement, et que l'échantillon sera tiré d'un stock de 4000 sachets. (c) On choisit au hasard un carton rempli, quelle est la probabilité que le poids de ce carton soit inférieur à 100 kg ?

27. Un avion (Boeing 747) peut transporter 100 passagers et leurs bagages, Il pèse 120 tonnes sans bagages, ni passagers mais équipage compris et plein de carburant. les consignes de sécurité imposent au commandant de bord de ne pas décoller si le poids de l'appareil chargé dépasse 129,42 tonnes. les 100 places ont été réservées. Le poids d'un voyageur est une variable aléatoire d'espérance mathématique 70 kg et de variance 100 kg Le poids de ses bagages est une V.A. de moyenne 20 kg et de variance 100 kg Toutes les variables sont supposées indépendantes. (a) L'espérance mathématique du poids de l'appareil au moment du décollage est-elle conforme aux normes de sécurité ? (b) Calculer l'écart type du poids total de l'appareil. (c) En admettant 1'IBT, quelle est la probabilité maximale pour que le poids réel de l'appareil au moment du décollage dépasse 129,42 tonnes ?

28. Une enquête sur l'emploi a pour but d'estimer le taux d'activité dans un pays. Dans les statistiques disponibles, la population active du pays est estimée à 10000000 personnes sur une population totale de 40 millions de personnes. Détermines la taille de l'échantillon si l'on accepte une erreur de 1% . avec une probabilité de 0,95.

29. Le rendement de la main d'œuvre d'une usine est chiffré par une production moyenne par jour et par ouvrier de 72 unités avec un écart type de 6 unités. (a) on a observé la production journalière d'un échantillon aléatoire de 25 ouvriers. Déterminer la loi et les paramètres de la moyenne de l'échantillon. (*b)* Quelle est la probabilité pour que la moyenne de cet échantillon soit inférieure à 63 ? (c) Quelle est la probabilité pour que l'écart entre la moyenne de cet échantillon et celle de la population soit supérieur à 3 ?

30. Un standard téléphonique reçoit en moyenne 400 appels par jour avec un écart type de 9,5. (a) Quelle est la probabilité pour qu'en une journée donnée, le nombre d'appels soit compris entre 360 et 440. (b) Quelle est la probabilité pour que le nombre moyen d'appels par jour en une période d'un mois soit compris entre 380 et 420 ?

31. Afin d'estimer le revenu mensuel moyen dans un secteur de production. Quelle doit être la taille de l'échantillon de salariés à interroger pour que la moyenne empirique ne s'éloigne pas de la moyenne de la population de 100 dh avec une probabilité au moins égale à 0,95 sachant que l'écart type est de 500 dh par salarié ?

32. On souhaite réaliser une enquête sur la consommation des ménages afin d'estimer la dépense moyenne par ménage. Quelle doit être la taille de l'échantillon de ménages si la population est composée de 5 millions de ménages et que l'erreur admise ne doit pas dépasser 100 dh avec une probabilité de 0,99 ? l'écart type de la dépense des ménages est de 2000 dh.

33. On souhaite réaliser une enquête sur l'emploi afin d'estimer le taux de chômage. La population active est de 5 millions de personnes. Quelle doit être la taille de l'échantillon pour que la fréquence empirique ne s'éloigne pas du vrai taux de chômage et ce avec une probabilité de 0,95 de 2%. Une enquête récente avait donné un taux de chômage de 12 %

34. Dans le cadre d'une étude socio-économique, on s'intéresse aux habitants de 18 unités urbaines, réparties en deux régions. L'enquête devrait comporter 500 interviews. Comme on dispose de 10 enquêteurs et qu'on souhaite que chaque enquêteur n'opère que dans une seule unité urbaine, on souhaite se limiter à l'étude de 10 unités urbaines. On considère qu'un enquêteur peut réaliser 10 interviews dans la même journée. En fonction de la répartition des unités urbaines par région et de leurs nombres d'habitants, expliquez, de façon aussi détaillée que possible la manière dont on pourrait organiser cette enquête, en précisant notamment dans quelles unités urbaines il y aurait lieu d'envoyer les enquêteurs.

Région 1		Région 2	
Unités urbaines	Nombres d'habitants	Unités urbaines	Nombres d'habitants
1	93600	9	117100
2	45400	10	107100
3	38900	11	61200
4	36500	12	51000
5	35100	13	43800
6	32900	14	38900
7	28100	15	37800
8	26400	16	33500
		17	25800
		18	25300

35. Dans une région regroupant environ 3 millions d'habitants réunis en un peu plus de 1500 communes, on désire réaliser une enquête au cours de laquelle 0,5 pour mille des habitants devraient être interrogés. En effectuant une stratification basée sur la distribution de fréquences donnée ci-dessous, combien d'interviews devrait-on réaliser dans chacune des catégories de communes. Si de plus pour des raisons de facilité, on décidait de ne pas effectuer moins de 10 interviews par commune, dans combien de communes différentes de chacune des catégories les enquêteurs devraient-ils se rendre ?

Nombre d'habitants	Nombre de communes
Moins de 1000	900
1000 – 2000	300
2000 – 5000	200
5000 – 10000	80
10000 – 20000	40
plus de 20000	10
Total	1530

36. Un sondage vise à étudier la notoriété d'une marque. Pour cela on dispose de 12 enquêteurs durant un mois. (a) Sachant que le rendement par jour et par enquêteur est distribué selon une loi normale de moyenne 5, et écart type 1, déterminer la taille de l'échantillon retenue no telle que : $P(n > no) = 0{,}025$. (b) On propose de stratifier la population selon l'âge. Sachant que la population se répartit comme suit, déterminer la répartition de l'échantillon:

Age	moins de 20 ans	entre 20 et 30 ans	entre 30 et 60 ans	plus de 60 ans
Effectifs	5500 000	2500 000	1250 000	250 000

37. On s'intéresse au pourcentage de fusibles défectueux dans un lot de 50 sacs contenant chacun 10000 fusibles. Les sacs proviennent de différents fournisseurs qui affirment en général que le proportion de fusibles défectueux ne dépasse pas 1%. L'erreur acceptée sur ce pourcentage est de 0,1% au niveau de confiance 0,95. (a) Déterminer la taille de cet échantillon en utilisant l'IBT, et en supposant la normalité de la variable. Laquelle de ces deux tailles doit-on retenir ? et pourquoi ? (b) Préciser de quel type de sondage s'agit-il : Si on tire n fusibles en prélevant n/50 par sac. Si on choisit d'abord K sacs et on tire ensuite ni fusibles par sac. Si on mélange le contenu des 50 sacs, et on tire n fusibles. (c) Quel est le procédé de tirage, le mieux adapté ?

38. Un sondage vise une population d'entreprises réparties en quatre régions contenant respectivement 360, 840, 600 et 1200 entreprises. Le budget réservé pour cette enquête est de 44 320 DH, Les écart-types sont estimés à 0,2 ; 0,1 ; 0,2 ; 0,4 respectivement pour les quatre régions. Les coûts de réalisation par questionnaire sont respectivement de 225 DH, 196 DH, 400H. et 324 DH. (a) Etablir une stratification optimale de l'échantillon à déterminer. (b) Préciser le niveau d'erreur que l'on doit accepter avec la taille de l'échantillon calculée, en admettant un niveau de confiance de 0,99 et une proportion théorique de 0,3.

39. Le budget allouée à une enquête est de 132500 dh. Cette enquête est destinée à estimer le taux de chômage qu'on a estimé à priori égal à 10 %. Les frais de déplacement quotidien sont évalués à 1000 dh par enquêteur. La rémunération d'un enquêteur est de 170 dh par jour. Les charges fixes sont de 20000 dh. (a) Déterminer la taille de l'échantillon si en tolère une erreur de moins de 1 % avec un niveau de confiance de 95% (b) Déterminer la taille maximale permise par le budget allouée si le rendement par enquêteur est de 6 questionnaires par jour. (c) Quel niveau d'erreur faut-il accepter si on réalise l'enquête avec le budget alloué ?

40. Une machine automatique fabrique des entretoises destinées à un montage de roulements. La longueur de ces entretoises doit être comprise, au sens large, entre 37,45 et 37,55 mm. La variable aléatoire X, qui associe à chaque entretoise sa longueur, est une variable gaussienne de moyenne 37,50 mm.

    1) Quel doit être l'écart type de la variable aléatoire X pour que 998 sur 1000 des pièces fabriquées soient bonnes ?

    2) On prélève un échantillon non exhaustif dans la production. Quel doit être l'effectif de cet échantillon pour que la moyenne des longueurs des pièces prélevées appartienne à l'intervalle [37,495 ; 37,505] avec une probabilité de 0,95 ?

41. Une machine fabrique des disques pleins en grande série. On suppose que la variable aléatoire X qui, à chaque disque tiré au hasard, associe son diamètre suit la loi normale de moyenne 12,8 mm et d'écart type 2,1 mm.

    a) Quelle loi suit la variable aléatoire, qui à tout échantillon aléatoire non exhaustif de taille 49, associe la moyenne des diamètres des disques de cet échantillon ?

    b) Déterminer un intervalle centré en 12,8 tel que la moyenne des diamètres prendra ses valeurs dans cet intervalle avec la probabilité 0,95.

    c) On se propose de prélever un échantillon aléatoire non exhaustif de taille n. Déterminer n pour que la moyenne des diamètres des disques prélevés ne s'écarte pas de la vrai moyenne de la population de plus de 0,2 mm avec une probabilité de 0,95.

DEUXIEME PARTIE

LES PROBLEMES D'ESTIMATION

LES PROBLEMES D'ESTIMATION

Les premiers problèmes d'inférence statistique auxquels s'applique la théorie des distributions d'échantillonnage sont les problèmes d'estimations. Le but poursuivi est d'estimer, à partir d'un échantillon, la ou les valeurs numériques d'un ou de plusieurs paramètres de la population considérée et de déterminer la précision de cette ou de ces estimations.

On distingue deux formes d'estimations : l'estimation ponctuelle et l'estimation par intervalle de confiance.

I. ESTIMATION PONCTUELLE

L'estimation ponctuelle ou l'estimation de point d'un paramètre est la connaissance de la seule valeur estimée de ce paramètre. Les paramètres les plus recherchés sont la moyenne, la variance et la proportion.

1.1. PRINCIPES GENERAUX DE L'ESTIMATION

Soit une population quelconque, dont la distribution de probabilité L(X) est fonction d'un paramètre θ : $L(X) = f(X, \theta)$ et un échantillon aléatoire et simple d'effectif n extrait de cette population.

On appelle estimateur du paramètre θ, toute fonction aléatoire des valeurs observées, X_1, X_2, …, Xn, susceptibles de servir à estimer θ.

$$\mathbf{Tn = f (X_1, X_2, \ldots, Xn)}$$

On appelle estimations les valeurs numériques t_1, t_2, …de cette variable aléatoire Tn.

1.1.1. Les principales qualités d'un estimateur

a) l'absence de biais

La première qualité d'un bon estimateur est l'absence d'erreur systématique ou de biais. Cette qualité implique que la vraie valeur θ doit être retrouvée en moyenne :

$$\mathbf{E(Tn) = \theta}$$

Tout estimateur qui satisfait cette condition est dit sans biais ou non biaisé.

b) la variance minimale

Une deuxième qualité d'un bon estimateur est de posséder une précision suffisante. Cette précision peut être mesurée par le moment d'ordre deux par rapport à θ.

$$E[(Tn - \theta)^2]$$

Pour les estimateurs non biaisés, ce moment se confond avec la variance :

$$E[(Tn - \theta)^2] = V(Tn)$$

On peut démonter qu'à tout paramètre θ correspond une valeur minimum de $E[(Tn - \theta)^2]$.

La fonction qui correspond à ce minimum définit l'estimateur de variance minimum.

Dans le cas des estimateurs non biaisés, cette variance vaut :

$$\frac{1}{nE[(\frac{d\log f(x,\theta)}{d\theta})^2]}$$

Un estimateur non biaisé dont la variance est égale à ce minimum est appelé estimateur non biaisé de variance minimum ou estimateur efficace.

c) convergence en probabilité

un estimateur Tn converge en probabilité vers θ si :

$$\forall\ \varepsilon > 0, \qquad \lim_{n\to\infty} P(|T_n - \theta| > \varepsilon) = 0$$

Ce ci signifie que l'écart entre le paramètre calculé à partir de l'échantillon et la vraie valeur du paramètre de la population est très faible quand la taille de l'échantillon est grande. Cet écart peut être mesuré par la variance. Ainsi on parle de convergence en probabilité si :

$$\lim_{n\to\infty} V(T_n) = 0$$

Un estimateur qui converge en probabilité est dit consistant.

1.1.2. la méthode du maximum de vraisemblance

Ayant définit les principales qualités des estimateurs, la méthode du maximum de vraisemblance permet le plus souvent d'obtenir des estimateurs possédant ces qualités. Le principe de cette méthode est de choisir comme estimation de tout paramètre θ la valeur la plus vraisemblable, c'est à dire celle qui a la plus grande probabilité de provoquer l'apparition des valeurs observées dans l'échantillon. Cette probabilité est appelée fonction de vraisemblance. C'est la probabilité ou la densité de probabilité relative aux valeurs observées $x_1, x_2, \ldots, x_n$, exprimée en fonction du paramètre de la population.

Pour un échantillon aléatoire et simple et pour une population définie par un seul paramètre θ, la fonction de vraisemblance est :

$$L(\theta) = p(x_1, x_2, \ldots, x_n ; \theta) = p(x_1 ; \theta)\, p(x_2 ; \theta) \ldots p(x_n ; \theta)$$

Ou

$$L(\theta) = f(x_1, x_2, \ldots, x_n ; \theta) = f(x_1 ; \theta)\, f(x_2 ; \theta) \ldots f(x_n ; \theta)$$

Les estimateurs du maximum de vraisemblance correspondent par définition au maximum de cette fonction. On cherche ce maximum en annulant la dérivée de la fonction par rapport à θ :

$$\frac{dL(\theta)}{d\theta}=0$$

ou en annulant la dérivée de son logarithme :

$$\frac{d\log L(\theta)}{d\theta}=0$$

1.2. Estimation de la moyenne

La meilleure estimation de la moyenne m d'une population, qui puisse être déduite d'un échantillon aléatoire et simple, est la moyenne de l'échantillon.

$$\hat{m} = \bar{x}$$

La dispersion des différentes estimations possibles autour de cette moyenne générale, est mesurée par l'erreur standard de la moyenne :

$$\sigma_{\bar{X}}=\frac{\sigma}{\sqrt{n}}$$

Estimateur du maximum de vraisemblance :

Pour une population normale, la densité de probabilité est :

$$f(x,m)=\frac{1}{\sigma\sqrt{2\Pi}}e^{-\frac{1}{2}(\frac{x-m}{\sigma})^2}$$

La fonction de vraisemblance est :

$$L(m) = \frac{1}{\sigma\sqrt{2\Pi}} e^{-\frac{1}{2}(\frac{x1-m}{\sigma})^2} \frac{1}{\sigma\sqrt{2\Pi}} e^{-\frac{1}{2}(\frac{x2-m}{\sigma})^2} \dots \frac{1}{\sigma\sqrt{2\Pi}} e^{-\frac{1}{2}(\frac{xn-m}{\sigma})^2}$$

$$L(m) = (\frac{1}{\sigma\sqrt{2\Pi}})^n \; e^{-\frac{1}{2}\sum_{i=1}^{n}(\frac{xi-m}{\sigma})^2}$$

$$\text{Log } L(m) = \frac{-n}{2}\log(2\Pi\sigma^2) - \frac{1}{2}\sum_{i=1}^{n}(\frac{xi-m}{\sigma})^2$$

La dérivée de cette fonction par rapport à m est :

$$\frac{d\log L(m)}{dm} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(xi-m)$$

l'estimation du maximum de vraisemblance de la moyenne de la population, est telle que :

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(xi-m) = 0$$

$$\sum_{i=1}^{n}(xi-m) = \sum_{i=1}^{n}xi - n\times m = 0$$

$$\hat{m} = \frac{\sum_{i=1}^{n}xi}{n} = \bar{x}$$

On retrouve la moyenne de l'échantillon définie précédemment.

LES QUALITES DE CET ESTIMATEUR

a) l'absence de biais

$\bar{X}=\frac{\sum_{i=1}^{n} X_i}{n}$ est un estimateur non biaisé de la moyenne m de la population puisqu'on a démontré que :

$$E(\bar{X}_n) = E(\frac{\sum_{i=1}^{n} Xi}{n}) = \frac{1}{n}\times\sum_{i=1}^{n} E(Xi) = \frac{1}{n}\times n\times m = m$$

En effet, pour l'ensemble des échantillons qui peuvent être rencontrés, on doit retrouver, en moyenne, la vraie valeur de la population.

b) la variance minimale

Pour une population normale, la densité de probabilité est :

$$f(x,m)=\frac{1}{\sigma\sqrt{2\Pi}}e^{-\frac{1}{2}(\frac{x-m}{\sigma})^2}$$

$$\log f(x , m) = -\log (\sigma \sqrt{2\Pi}) - \frac{1}{2}(\frac{x-m}{\sigma})^2$$

$$\frac{d\log f(x,m)}{dm} = \frac{x-m}{\sigma^2}$$

$$nE[(\frac{d\log f(x,m)}{dm})^2] = n\, E[(\frac{x-m}{\sigma^2})^2] = \frac{n}{\sigma_4} E[(X-m)^2] = \frac{n}{\sigma^2}$$

le minimum de la variance des estimateurs de la moyenne est donc :

$$\frac{1}{nE[(\frac{d\log f(x,m)}{dm})^2]} = \frac{\sigma^2}{n}$$

Comme cette valeur est aussi la variance de la distribution d'échantillonnage de la moyenne, il en résulte que la moyenne $\bar{X}=\frac{\sum_{i=1}^{n} X_i}{n}$ d'un échantillon aléatoire et simple est un estimateur de variance minimale. Il est donc un estimateur efficace de la moyenne m de la population.

c) convergence en probabilité

$\bar{X}=\frac{\sum_{i=1}^{n} X_i}{n}$ est un estimateur consistant de la moyenne m de la population puisqu'on a démontré que :

$$V(\bar{X}_n) = \frac{\sigma^2}{n}$$

$$\lim_{n\to\infty} V(\bar{X}_n) = 0$$

la moyenne $\bar{X}_n=\frac{\sum_{i=1}^{n} Xi}{n}$ calculée à partir d'un échantillon de taille n converge en probabilité vers m.

1.3. Estimation de la variance

Estimateur du maximum de vraisemblance :

Pour une population normale, la densité de probabilité est :

$$f(x,\sigma)=\frac{1}{\sqrt{2\Pi\sigma^2}}e^{-\frac{1}{2}(\frac{x-m}{\sigma})^2}$$

la fonction de vraisemblance est :

$$L(\sigma^2) = \frac{1}{\sqrt{2\Pi\sigma^2}}e^{-\frac{1}{2}(\frac{x1-m}{\sigma})^2} \frac{1}{\sqrt{2\Pi\sigma^2}}e^{-\frac{1}{2}(\frac{x2-m}{\sigma})^2} \dots \frac{1}{\sqrt{2\Pi\sigma^2}}e^{-\frac{1}{2}(\frac{xn-m}{\sigma})^2}$$

$$L(\sigma^2) = (\frac{1}{\sqrt{2\Pi\sigma^2}})^n \; e^{-\frac{\sum_{i=1}^{n}(xi-m)^2}{2\sigma^2}}$$

$$\text{Log } L(\sigma^2) = \frac{-n}{2}\log(2\Pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(xi-m)^2$$

La dérivée de cette fonction par rapport à σ^2 est :

$$\frac{d\log L(\sigma^2)}{d\sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(xi-m)^2$$

l'estimation du maximum de vraisemblance de la variance de la population, est telle que :

$$-\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(xi-m)^2 = 0$$

$$-n\sigma^2+\sum_{i=1}^{n}(x_i-m)^2 = 0$$

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(x_i-m)^2}{n}$$

On retrouve la variance de l'échantillon V(X).

LES QUALITES DE CET ESTIMATEUR

$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(x_i-m)^2}{n}$ est un estimateur biaisé de la variance σ^2 de la population puisqu'on a démontré que :

$$E(V(X))=\frac{n-1}{n}\times\sigma^2$$

Contrairement à la moyenne, la meilleure estimation de la variance σ^2 d'une population, qui puisse être déduite d'un échantillon aléatoire et simple, n'est pas la variance de l'échantillon v(x). En effet, pour l'ensemble des échantillons qui peuvent être rencontrés, on ne retrouve pas, en moyenne, la vraie valeur de la population, on obtient ainsi, en moyenne, une valeur inférieure à la variance de la population.

le biais est :

$$E(V(X)) - \sigma^2 = \frac{-\sigma^2}{n}$$

Ce biais peut être corrigé en multipliant la variance de l'échantillon par le facteur $\frac{n}{n-1}$. On obtient alors l'estimation :

$$\hat{\sigma}^2 = \frac{n}{n-1} \times v(x) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

dont l'espérance mathématique est bien σ^2.

$$E(\hat{\sigma}^2) = E(\frac{n}{n-1} \times v(x)) = \frac{n}{n-1} E(v(x)) = \frac{n}{n-1} \frac{n-1}{n} \times \sigma^2 = \sigma^2$$

$\hat{\sigma}^2 = \frac{n}{n-1} \times v(x)$ est appelée quasi-variance, c'est un estimateur sans biais de la variance σ^2 de la population. La quasi-variancee est désignée par $\sigma^2 n-1$

L'erreur standard de cette estimation est, dans le cas d'une population normale :

$$\sqrt{v(\hat{\sigma}^2)} = \sqrt{v(\frac{n}{n-1}\sigma^2)} = \frac{n}{n-1}\sqrt{\frac{2(n-1)\sigma^4}{n^2}} = \sigma^2\sqrt{\frac{2}{n-1}}$$

1.4. Estimation de la proportion

La meilleure estimation de la proportion p d'une population, qui puisse être déduite d'un échantillon aléatoire et simple, est la fréquence de l'échantillon f_n.

$$\hat{p} = f_n$$

La dispersion des différentes estimations possibles autour de cette proportion générale, est mesurée par l'erreur standard de la proportion :

$$\sigma_{\bar{f}_n} = \sqrt{\frac{f_n(1-f_n)}{n}}$$

Estimateur du maximum de vraisemblance :

Pour un échantillon aléatoire et simple d'effectif n, dont x individus possèdent le caractère étudié, la fonction de vraisemblance est :

$$L(p) = C_n^x \, p^x \, (1-p)^{n-x}$$

$$\text{Log } L(p) = \log C_n^x + x \log p + (n-x) \log (1-p)$$

La dérivée de cette fonction par rapport à p est :

$$\frac{d \log L(p)}{dp} = \frac{x}{p} - \frac{n-x}{1-p}$$

l'estimation du maximum de vraisemblance de la variance de la population, est telle que :

$$\frac{x}{p} - \frac{n-x}{1-p} = 0$$

$$(1-p)\, x - p\,(n-x) = 0$$

$$x - np = 0$$

$$\hat{p} = \frac{x}{n}$$

La fréquence f_n de l'échantillon est donc un estimateur du maximum de vraisemblance de la proportion de la population.

Les qualités de cet estimateur

a) l'absence de biais

$F_n = \frac{X_n}{n}$ est un estimateur non biaisé de la proportion p de la population puisqu'on a démontré que :

$$E(F_n) = p$$

En effet, pour l'ensemble des échantillons qui peuvent être rencontrés, on doit retrouver, en moyenne, la vraie valeur de la population.

b) convergence en probabilité

$F_n = \frac{X_n}{n}$ est un estimateur consistant de la proportion p de la population puisqu'on a démontré que :

$$V(f_n) = \frac{pq}{n}$$

$$\lim_{n \to \infty} V(f_n) = 0$$

la fréquence relative $F_n = \frac{X_n}{n}$ calculée à partir d'un échantillon de taille n converge en probabilité vers p.

II. ESTIMATION PAR INTERVALLE DE CONFIANCE

L'estimation par intervalle de confiance consiste à déterminer autour de la valeur estimée un intervalle dont on a de fortes chances de croire qu'il contient la vraie valeur du paramètre recherché.

Si on s'intéresse à un paramètre θ, dont on possède un estimateur T, l'estimation par intervalle de confiance consiste à déterminer de part et d'autre de T les bornes T1 et T2 d'un intervalle qui a une forte probabilité de contenir θ. Cette probabilité est appelée niveau de confiance et désignée par (1-α). α est alors un risque d'erreur.

Les limites T1 et T2 sont telles que :

$$\mathbf{p(T1 \leq \theta \leq T2) = 1 - \alpha}$$

L'intervalle [T1 , T2] est appelé intervalle de confiance.

La probabilité que le paramètre θ se trouve à l'extérieur de cet intervalle est donc :

$$\mathbf{p(\theta < T1) + p(\theta > T2) = \alpha}$$

Le risque total α peut être réparti d'une infinité de manière. Généralement, on divise le risque α en deux parties égales, Les limites T1 et T2 sont telles que :

$$\mathbf{p(\theta < T1) = p(\theta > T2) = \alpha/2}$$

2.1. Intervalle de confiance de la moyenne

2.1.1. cas d'une population normale

Si on s'intéresse à la moyenne inconnue m d'une population normale d'écart type connu σ, l'estimation par intervalle de confiance consiste à déterminer de part et d'autre de l'estimateur $\bar{X}$ les bornes $\bar{X}_1$ et $\bar{X}_2$ d'un intervalle qui a un niveau de confiance (1-α) de contenir m.

Les limites $\bar{X}_1$ et $\bar{X}_2$ sont telles que :

$$p(\bar{X}_1 \leq m \leq \bar{X}_2) = 1 - \alpha$$

ou d'une autre façon :

$$p(m < \bar{X}_1) = p(m > \bar{X}_2) = \alpha/2$$

les limites de confiance peuvent être écrites :

$$\bar{X}_1 = \bar{X} - d1 \quad \text{et} \quad \bar{X}_2 = \bar{X} + d2$$

on peut alors écrire :

$$p(m < \bar{X} - d1) = p(m > \bar{X} + d2) = \alpha/2$$

$$p(\bar{X} - m > d1) = p(m - \bar{X} > d2) = \alpha/2$$

Comme, pour une population normale, la variable $\bar{X}$ est elle-même normale de moyenne m et d'écart type $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$, on peut écrire :

$$p\left(\frac{\bar{x}-m}{\frac{\sigma}{\sqrt{n}}} > \frac{d1}{\frac{\sigma}{\sqrt{n}}}\right) = p\left(\frac{m-\bar{x}}{\frac{\sigma}{\sqrt{n}}} > \frac{d2}{\frac{\sigma}{\sqrt{n}}}\right) = \frac{\alpha}{2}$$

$$p\left(Z_1 > \frac{d1}{\frac{\sigma}{\sqrt{n}}}\right) = p\left(Z_2 > \frac{d2}{\frac{\sigma}{\sqrt{n}}}\right) = \frac{\alpha}{2}$$

$$p\left(Z_1 < \frac{d1}{\frac{\sigma}{\sqrt{n}}}\right) = p\left(Z_2 < \frac{d2}{\frac{\sigma}{\sqrt{n}}}\right) = 1 - \frac{\alpha}{2}$$

$$\prod\left(\frac{d1}{\frac{\sigma}{\sqrt{n}}}\right) = \prod\left(\frac{d2}{\frac{\sigma}{\sqrt{n}}}\right) = 1 - \frac{\alpha}{2}$$

Si on désigne par $Z_{1-\frac{\alpha}{2}}$ la valeur de la variable normale réduite lue dans la table :

$$\frac{d1}{\frac{\sigma}{\sqrt{n}}} = \frac{d2}{\frac{\sigma}{\sqrt{n}}} = Z_{1-\frac{\alpha}{2}}$$

il en résulte :

$$d1 = d2 = Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

Les limites de confiances sont donc :

$$\bar{X}_1 = \bar{X} - Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \qquad \textbf{et} \qquad \bar{X}_2 = \bar{X} + Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

On notera l'intervalle de confiance :

$$\bar{X} \pm Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

C'est un intervalle symétrique par rapport à la moyenne.

2.1.2. cas d'une population de distribution inconnue

Pour une population de distribution de probabilité inconnue (écart type σ inconnu), on utilise la quasi-variance comme estimation de la variance de la population. L'intervalle de confiance de la moyenne sera défini selon les cas.

Cas d'un échantillon d'effectif inférieur à 30 (n < 30) :

Dans ce cas, la moyenne d'un échantillon peut toujours être considérée comme une variable T de Student à (n-1) degré de liberté. La valeur $Z_{1-\frac{\alpha}{2}}$ sera remplacée par la valeur $T_{1-\frac{\alpha}{2}}$ à (n-1) degré de liberté. L'intervalle de confiance est alors :

$$\bar{X} \pm T_{1-\frac{\alpha}{2}} \frac{\hat{\sigma}}{\sqrt{n}}$$

Cas d'un échantillon d'effectif supérieur ou égal à 30 (n ≥ 30) :

Dans ce cas, la moyenne d'un échantillon peut toujours être considérée comme une variable approximativement normale. L'intervalle de confiance est alors :

$$\bar{X} \pm Z_{1-\frac{\alpha}{2}} \frac{\hat{\sigma}}{\sqrt{n}}$$

2.2. Intervalle de confiance de la variance

Si on s'intéresse à la variance σ^2 d'une population normale, l'estimation par intervalle de confiance consiste à déterminer les bornes σ^2_1 et σ^2_2 d'un intervalle qui a un niveau de confiance $(1-\alpha)$ de contenir σ^2.

Les limites σ^2_1 et σ^2_2 sont telles que :

$$p(\sigma^2_1 \leq \sigma^2 \leq \sigma^2_2) = 1 - \alpha$$

Comme, pour une population normale, la variable aléatoire $\frac{\sum_{i=1}^{n}(xi-\bar{x})^2}{\sigma^2}$ possède une distribution khi deux à (n-1) degré de liberté, on peut alors écrire :

$$p\left(\frac{\sum_{i=1}^{n}(xi-\bar{x})^2}{\sigma^2 2} \leq \frac{\sum_{i=1}^{n}(xi-\bar{x})^2}{\sigma^2} \leq \frac{\sum_{i=1}^{n}(xi-\bar{x})^2}{\sigma^2 1}\right) = 1 - \alpha$$

ou encore :

$$p\left(\frac{\sum_{i=1}^{n}(xi-\bar{x})^2}{\sigma^2} < \frac{\sum_{i=1}^{n}(xi-\bar{x})^2}{\sigma^2 2}\right) = p\left(\frac{\sum_{i=1}^{n}(xi-\bar{x})^2}{\sigma^2} > \frac{\sum_{i=1}^{n}(xi-\bar{x})^2}{\sigma^2 1}\right) = \alpha/2$$

$$p\left(\frac{\sum_{i=1}^{n}(xi-\bar{x})^2}{\sigma^2} \leq \frac{\sum_{i=1}^{n}(xi-\bar{x})^2}{\sigma^2 1}\right) = 1 - \alpha/2 \quad \Rightarrow \quad \frac{\sum_{i=1}^{n}(xi-\bar{x})^2}{\sigma^2 1} = \chi^2_{1-\frac{\alpha}{2}}$$

$$p\left(\frac{\sum_{i=1}^{n}(xi-\bar{x})^2}{\sigma^2} < \frac{\sum_{i=1}^{n}(xi-\bar{x})^2}{\sigma^2 2}\right) = \alpha/2 \quad \Rightarrow \quad \frac{\sum_{i=1}^{n}(xi-\bar{x})^2}{\sigma^2 2} = \chi^2_{\frac{\alpha}{2}}$$

Les limites de confiances sont alors :

$$\sigma^2_1 = \frac{\sum_{i=1}^{n}(xi-\bar{x})^2}{\chi^2_{1-\frac{\alpha}{2}}} \quad \text{et} \quad \sigma^2_2 = \frac{\sum_{i=1}^{n}(xi-\bar{x})^2}{\chi^2_{\frac{\alpha}{2}}}$$

Les valeurs de $\chi^2_{\frac{\alpha}{2}}$ et $\chi^2_{1-\frac{\alpha}{2}}$ sont à (n-1) degré de liberté.

2.3. Intervalle de confiance de la proportion

Si on s'intéresse à la proportion p, l'estimation par intervalle de confiance consiste à déterminer de part et d'autre de l'estimateur Fn les bornes p_1 et p_2 d'un intervalle qui a un niveau de confiance (1-α) de contenir p.

Les limites p_1 et p_2 sont telles que :

$$p(p_1 \leq p \leq p_2) = 1 - \alpha$$

ou d'une autre façon :

$$p(p < p_1) = p(p > p_2) = \alpha/2$$

les limites de confiance peuvent être écrites :

$$p_1 = f_n - d1 \quad \text{et} \quad p_2 = f_n + d2$$

on peut alors écrire :

$$p(p < f_n - d1) = p(p > f_n + d2) = \alpha/2$$

$$p(f_n - p > d1) = p(p - f_n > d2) = \alpha/2$$

Comme, la distribution de la proportion suit une loi normale de moyenne p et d'écart type $\sigma_{Fn} = \frac{\sqrt{pq}}{\sqrt{n}}$ à condition que la taille de l'échantillon soit supérieure ou égale à 30 ($n \geq 30$) et le produit $n\,p \geq 5$, on peut écrire :

$$p\left(\frac{f_n - p}{\frac{\sqrt{p(1-p)}}{\sqrt{n}}} > \frac{d1}{\frac{\sqrt{p(1-p)}}{\sqrt{n}}}\right) = p\left(\frac{p - f_n}{\frac{\sqrt{p(1-p)}}{\sqrt{n}}} > \frac{d2}{\frac{\sqrt{p(1-p)}}{\sqrt{n}}}\right) = \frac{\alpha}{2}$$

$$p\left(Z_1 > \frac{d1}{\frac{\sqrt{p(1-p)}}{\sqrt{n}}}\right) = p\left(Z_2 > \frac{d2}{\frac{\sqrt{p(1-p)}}{\sqrt{n}}}\right) = \frac{\alpha}{2}$$

$$p\left(Z_1 < \frac{d1}{\frac{\sqrt{p(1-p)}}{\sqrt{n}}}\right) = p\left(Z_2 < \frac{d2}{\frac{\sqrt{p(1-p)}}{\sqrt{n}}}\right) = 1 - \frac{\alpha}{2}$$

$$\Pi\left(\frac{d1}{\frac{\sqrt{p(1-p)}}{\sqrt{n}}}\right) = \Pi\left(\frac{d2}{\frac{\sqrt{p(1-p)}}{\sqrt{n}}}\right) = 1 - \frac{\alpha}{2}$$

Si on désigne par $Z_{1-\frac{\alpha}{2}}$ la valeur de la variable normale réduite lue dans la table :

$$\frac{d1}{\frac{\sqrt{p(1-p)}}{\sqrt{n}}} = \frac{d2}{\frac{\sqrt{p(1-p)}}{\sqrt{n}}} = Z_{1-\frac{\alpha}{2}}$$

il en résulte :

$$d1 = d2 = Z_{1-\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}}$$

Les limites de confiances sont donc :

$$p_1 = f_n - Z_{1-\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}} \qquad \text{et} \qquad p_2 = f_n + Z_{1-\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}}$$

On notera l'intervalle de confiance :

$$f_n \pm Z_{1-\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}}$$

La proportion p de la population sera estimée par la fréquence f_n de l'échantillon. On obtient ainsi un intervalle symétrique par rapport à la proportion.

<u>Exemple 1</u> : intervalle de confiance de la moyenne et de l'écart type

Dans une entreprise produisant un article déterminé on veut estimer sa durée de vie en heures. À cette fin on a observé un échantillon aléatoire et simple de 16 unités dont les résultats sont (en 1000 heures) :

1,10	1,05	1,25	1,08	1,35	1,15	1,30	1,25
1,30	1,35	1,15	1,32	1,05	1,25	1,10	1,15

L'estimation ponctuelle de la moyenne de la population est :

$$\hat{m} = \bar{x} = \frac{\sum_{i=1}^{16} x_i}{16} = 1{,}2$$

L'estimation ponctuelle de l'écart type de la population de la population est :

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{16} (x_i - \bar{x})^2}{16-1}} = 0{,}11$$

L'intervalle de confiance de la moyenne à un niveau de confiance de 95 % (∝=5%):

La distribution de la population parent étant inconnue et la taille de l'échantillon inférieure à 30, l'intervalle de confiance de la moyenne est défini par :

$$\bar{X} \pm T_{1-\frac{\alpha}{2}} \frac{\hat{\sigma}}{\sqrt{n}}$$

La valeur de $T_{1-\frac{\alpha}{2}}$ à 15 degrés de liberté est : $t_{0,975} = 2,131$

l'intervalle de confiance est :

$$\bar{X} \pm T_{1-\frac{\alpha}{2}} \frac{\hat{\sigma}}{\sqrt{n}} = 1,2 \pm 2,131 \frac{0,11}{\sqrt{16}}$$

$$\bar{X}_1 = 1,2 - 2,131 \frac{0,11}{\sqrt{16}} = 1,14 \qquad \text{et} \qquad \bar{X}_2 = 1,2 + 2,131 \frac{0,11}{\sqrt{16}} = 1,26$$

L'intervalle [1,14 ; 1,26] a une probabilité de 95 % de contenir la vraie valeur de la moyenne de la population.

L'intervalle de confiance de l'écart type à un niveau de confiance de 95 % (∝=5%):

Les limites de confiances de la variance sont :

$$\sigma^2_1 = \frac{\sum_{i=1}^{n}(xi-\bar{x})^2}{\chi^2_{1-\frac{\alpha}{2}}} \qquad \text{et} \qquad \sigma^2_2 = \frac{\sum_{i=1}^{n}(xi-\bar{x})^2}{\chi^2_{\frac{\alpha}{2}}}$$

les valeurs de $\chi^2_{\frac{\alpha}{2}}$ et $\chi^2_{1-\frac{\alpha}{2}}$ sont à 15 degrés de liberté :

$$\chi^2_{0,025} = 6,26 \qquad \text{et} \qquad \chi^2_{0,975} = 27,49$$

L'écart type est la racine carrée de la variance, ses limites de confiance sont donc :

$$\hat{\sigma}_1 = \sqrt{\frac{\sum_{i=1}^{16}(x_i - \bar{x})^2}{\chi^2_{1-\frac{\alpha}{2}}}} = \sqrt{\frac{0,11^2 \times 15}{27,49}} = 0,08 \qquad \hat{\sigma}_2 = \sqrt{\frac{\sum_{i=1}^{16}(x_i - \bar{x})^2}{\chi^2_{\frac{\alpha}{2}}}} = \sqrt{\frac{0,11^2 \times 15}{6,26}} = 0,17$$

<u>Exemple 2</u> : intervalle de confiance de la proportion

On étudie le pourcentage d'utilisation d'une machine. 400 observations ont été effectuées qui ont donné le résultat suivant :

- Machine marche : 320 observations.
- Machine arrêtée : 80 observations.

L'estimation ponctuelle de la proportion d'utilisation de la machine est :

$$\hat{p} = f_n = \frac{320}{400} = 0,8$$

Le taux d'utilisation de la machine est estimé à 80 %.

L'intervalle de confiance de la proportion à un niveau de confiance de 95 % est défini par :

$$f_n \pm Z_{1-\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}}$$

La valeur de $Z_{1-\frac{\alpha}{2}}$ est : $Z_{0,975} = 1,96$

Les limites de confiances de la proportion sont :

$$p_1 = f_n - Z_{1-\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}} = 0,80 - 1,96 \sqrt{\frac{0,8(1-0,8)}{400}} = 0,76$$

$$p_2 = f_n + Z_{1-\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}} = 0,80 + 1,96 \sqrt{\frac{0,8(1-0,8)}{400}} = 0,84$$

L'intervalle [76 % ; 84 %] a une probabilité de 95% de contenir le vrai taux d'utilisation de la machine.

EXERCICES SUR LES PROBLEMES DE L'ESTIMATION

Ex 1 : Soit X une variable de Poisson de paramètre (inconnu) m et (XI, … Xn) les observations d'un échantillon de taille n. Écrire la fonction du maximum de vraisemblance associée à la moyenne. Quel est l'estimateur du maximum de vraisemblance de la moyenne de la population ? Cet estimateur précédent est-il un estimateur efficace ?

Ex 2 : Soit X une variable aléatoire dont la densité de probabilité f est ainsi définie:

$$f(x,\lambda) = \frac{1}{\lambda}\exp(-\frac{x}{\lambda}) \qquad \text{si } x > 0$$

$$f(x,\lambda) = 0 \qquad \text{si } x < 0$$

Où λ est le paramètre (positif) de la loi.

a) Calculer l'espérance mathématique et la variance de X.
b) Pour estimer le paramètre λ, on considère un échantillon aléatoire de taille n. Quel est l'estimateur du maximum de vraisemblance de λ ?
c) L'estimateur de λ est-il un estimateur efficace ?

Ex 3 : Le tableau suivant donne la distribution du nombre de pannes observées dans le fonctionnement d'une machine au cours de 100 journées de travail. Déduisez-en une estimation du nombre moyen de pannes par jour, en supposant que la distribution théorique du nombre de pannes est une loi de poisson. Donner l'erreur standard du résultat obtenu.

Nombres de pannes	Nombres de jours
0	53
1	32
2	11
3	3
4	1
Total	100

Ex 4 : lors d'un concours radiophonique, on note X le nombre de réponses reçues chaque jour. On suppose que X suit une loi normale de paramètres m et σ. Durant les 10 premiers jours, on a obtenu : x1 = 200 ; x2 = 240 ; x3 = 190 ; x4 = 150 ; x5 = 220 ; x6 = 180 ; x7 = 170 ; x8 = 230 ; x9 = 210 et x10 = 210. Déterminer une estimation ponctuelle de m et σ.

Ex 5 : Un échantillon de 15 étudiants d'une faculté a donné les notes suivantes :

13 ; 06 ; 12 ; 10 ; 10 ; 16 ; 02 ; 04 ; 11 ; 12 ; 12 ; 05 ; 07 ; 08 ; 13

a) Estimer la note moyenne et l'écart type des notes pour l'ensemble des étudiants de la faculté.
b) Donner des estimations par intervalle de confiance pour la moyenne et l'écart type. ($\propto$=5%).

Ex 6 : Dans une entreprise produisant un article déterminé on veut estimer sa durée de vie en heures. À cette fin on a observé un échantillon de 16 unités dont les résultats sont (en 1000 heures) :

1,10	1,05	1,25	1,08	1,35	1,15	1,30	1,25
1,30	1,35	1,15	1,32	1,05	1,25	1,10	1,15

a) Estimer la durée de vie moyenne et l'écart type d'un article.
b) Donner des estimations par intervalle de confiance pour la moyenne et l'écart type. ($\propto$=5%).

Ex 7 : dans une population d'étudiants en sociologie, on a prélevé, indépendamment, deux échantillons de taille n1 = 120 et n2 = 150. On constate que 48 étudiants de l'échantillon 1 et 66 étudiants de l'échantillon 2 ont une formation secondaire scientifique; Soit p la proportion d'étudiants de la population ayant une formation scientifique ; calculer trois estimations ponctuelles de p.

Ex 8 : dans une station service, on suppose que le montant des chèques essence suit une loi normale de paramètres m et σ. On considère un échantillon de taille n = 50 et on obtient une moyenne de 130 Dh et un écart-type de 28 Dh. Donner une estimation de m et σ par un intervalle de confiance au niveau de confiance 95%.

Ex 9 : on donne la répartition des masses de 219 ressorts provenant d'une même fabrication :

masses (g)	[8,2 ; 8,4[	[8,4 ; 8,6[	[8,6 ; 8,8[	[8,8 ; 9[	[9 ; 9,2[	[9,2 ; 9,4[	[9,4 ; 9,6[
Nbre de ressorts	9	21	39	63	45	27	15

X donnant le poids d'un ressort provenant de cette fabrication, donner une estimation de E(X) et V(X). Donner pour E(X) et V(X) un intervalle de confiance au niveau de confiance 95%.

Ex 10 : on veut estimer l'espérance mathématique m d'une variable aléatoire gaussienne X dont on connaît l'écart type σ = 2,3. Quelle est la taille minimum de l'échantillon de X qui est à prendre si l'on veut obtenir pour m un intervalle de confiance de seuil 0,95 et dont la longueur ne dépasse pas 0,1 ?

Ex 11 : un confiseur vend des boites de bonbons d'un certain modèle. On note X la masse d'une boite pleine. Les pesées de 8 boites ont conduit aux masses (en kg) :

1,22 ; 1,23 ; 1,21 ; 1,19 ; 1,23 ; 1,24 ; 1,18 ; 1,21.

a) Donner pour E(X) un intervalle de confiance au risque de 5%.
b) En supposant que la variance de X soit connue et égale à la variance observée, donner pour E(X) un intervalle de confiance au seuil de confiance 95% et comparer avec le a).
c) On suppose maintenant que l'on a trouvé la même moyenne et la même variance qu'observées mais avec 16 observations au lieu de 8. Reprendre les questions a) et b).

Ex 12 : après avoir pesé 12 pamplemousses d'une même provenance, on donne pour l'espérance mathématique m du poids X d'un pamplemousse, l'intervalle de confiance au niveau de confiance 95% : 390 g ≤ m ≤ 520 g. En déduire la moyenne observée et l'écart type observé.

Ex 13 : Un promoteur désire étudier le nombre de garage qu'il est souhaitable de construire avec un ensemble de logements, afin que les occupants puissent y ranger leur voiture. Pour cela il fait effectuer une enquête par sondage auprès d'un échantillon de ménages susceptibles d'habiter ces appartements.

a) On interroge un échantillon de 3238 ménages. On trouve parmi eux 1943 possesseurs d'une voiture. Estimez, à partir de cet échantillon, la proportion des ménages ayant une voiture. Degré de confiance 99 %.
b) À partir de la proportion estimée, combien de ménages faudrait-il interroger pour construire, avec un risque d'erreur de 5 %, un intervalle de confiance d'amplitude 0,04 ?

Ex 14 : On étudie le pourcentage d'utilisation d'une machine. 400 observations ont été effectuées qui ont donné le résultat suivant :

- Machine marche : 320 observations.
- Machine arrêtée : 80 observations.

a) Entre quelles limites peut-on fixer le taux d'utilisation de la machine avec un degré de confiance de 95 % ?
b) On fait un plus grand nombre d'observations. On obtient le même pourcentage d'utilisation ce qui permet, avec un risque d'erreur de 5 %, de fixer les limites de confiance à [78,4 % ; 81,6 %]. Combien a-t-on fait d'observations ?

Ex 15 : Un échantillon aléatoire de 50 notes (sur 100) dans une population de 200 a donné une moyenne de 75 et un écart type de 10.

a) Quelles sont les limites de confiance à 95 % pour estimer la moyenne des 200 notes ?
b) Avec quel degré de confiance peut-on dire que la moyenne des 200 notes est de 75 plus ou moins 1 ?

Ex 16 : Un échantillon de 150 lampes de marque A a donné une durée de vie moyenne de 1400 heures et un écart type de 120 heures. Un échantillon de 200 lampes de marque B a donné une durée de vie moyenne de 1200 heures et un écart type de 80 heures. Déterminer les limites de confiances à 95 % de la différence des durées de vie moyennes des marques A et B.

Ex 17 : Sur un échantillon de 400 adultes et de 600 adolescents ayant regardé un certain programme de télévision, 100 adultes et de 300 adolescents l'ont apprécié. Calculer les limites de confiances à 99 % de la différence des fréquences des adultes et des adolescents qui ont regardé et apprécié le programme.

Ex 18 : Une compagnie fabrique des roulements à billes ayant un poids moyen de 0,638 Kg et un écart type de 0,012 Kg Calculer les limites de confiance à 95 % des poids de lots comprenant 100 roulements chacun.

Ex 19 : Dans une population de 579 individus, divisée en quatre strates comprenant respectivement 53 ; 190 ; 231 ; et 105 individus, on a prélevé un échantillon de 58 individus, dont 10 dans la première strate, 14 dans la deuxième, 21 dans la troisième et 13 dans la quatrième. En fonction des résultats suivants, estimez la moyenne de la population globale et l'erreur standard de cette moyenne, en considérant l'échantillon comme :

a) Aléatoire et simple ;
b) Stratifié.

Strates	Sommes	Sommes des carrés
1	54	1004
2	127	3081
3	388	13270
4	553	39667

Ex 20 : En vue d'estimer la note moyenne des élèves d'une école, on a choisi de façon aléatoire et simple six classes, et dans chacune de ces classes on a choisit aléatoirement 4 élèves. En fonction des résultats obtenus et repris ci-dessous :

a) Estimer la note moyenne des élèves de l'école ;
b) Déterminer l'intervalle de confiance à 95 % de cette estimation.

Classes Elèves	1	2	3	4	5	6
1	11,69	11,79	11,84	12,30	11,83	11,95
2	12,32	11,97	11,59	11,91	11,77	11,87
3	12,32	12,07	11,25	12,05	12,15	11,65
4	11,90	12,06	11,80	12,23	11,66	11,87

Ex 21 : Soit une variable aléatoire X de densité de probabilité f(x,λ) définie par :

$$f(x,\lambda) = \frac{1}{\sqrt{2\pi\lambda}} \; e^{-\frac{x^2}{2\lambda}}$$ pour tout nombre réel x.

a) Reconnaître la loi de la variable X et en déduire, sans calcul, l'espérance mathématique et la variance de X.
b) Déterminer un estimateur de maximum de vraisemblance de λ associé à un échantillon aléatoire de taille n.
c) L'estimateur précédent est-il un estimateur sans biais ?

TROISIEME PARTIE

LES TESTS STATISTIQUES

LES TESTS STATISTIQUES

I. INTRODUCTION

Un test statistique est une méthode permettant de prendre une **décision** à partir d'informations fournies par un **échantillo**n.

Les tests statistiques ou les tests d'hypothèses ont pour but de vérifier, à partir de données observées dans un ou plusieurs échantillons, la validité de certaines hypothèses relatives à une ou plusieurs populations.

On peut distinguer différents types de tests, en fonction des hypothèses auxquelles on a affaire.

Les tests de comparaison à une norme ou tests de conformité sont destinés à comparer entre eux une population théorique et un échantillon observé. Ils servent à vérifier si un échantillon donné peut être considéré comme extrait d'une population possédant telle caractéristique particulière (telle moyenne, telle variance, …). Le test se fait en vérifiant si la différence entre la valeur observée et la valeur théorique du paramètre considéré peut être attribuée au hasard ou non.

Les tests d'homogénéité ou d'égalité ont pour but de comparer entre elles un certain nombre de populations, à l'aide d'un même nombre d'échantillons.

Les tests d'ajustement sont destinés à vérifier si un échantillon observé peut être extrait d'une population donnée.

Les tests d'indépendance ont pour but de contrôler, à partir d'un échantillon, l'indépendance de deux ou plusieurs critères de classification, généralement qualitatifs.

II. LE PRINCIPE D'UN TEST STATISTIQUE

Pour commencer, on émet une certaine hypothèse à tester, appelée **hypothèse nulle**, généralement désignée par H_0. Celle-ci suppose toujours l'égalité des caractéristiques comparées.

L'hypothèse qui diffère de H_0 est dite **hypothèse alternative**, généralement désignée par H_1.

On mesure ensuite l'écart observé entre les caractéristiques comparées, et on calcule la probabilité d'observer, si l'hypothèse nulle est vraie, un écart aussi important.

Si cette probabilité est relativement élevée, on considère L'hypothèse nulle comme plausible et on l'accepte. Par contre si la probabilité calculée est faible, l'écart observé apparaît comme peu compatible avec l'hypothèse nulle et on rejette celle-ci.

L'ensemble des valeurs observées pour lesquelles l'hypothèse nulle est admissible forme la région d'acceptation. Les autres valeurs constituent la région de rejet. Les valeurs limites sont appelées valeurs critiques.

La décision dépend donc de l'échantillon. Ainsi qu'elle que soit la décision prise, le hasard de l'échantillonnage peut fausser les conclusions. Quatre situations doivent en effet être envisagées:

L'acceptation de l'hypothèse nulle alors qu'elle est vraie, le rejet de l'hypothèse nulle alors qu'elle est vraie, l'acceptation de l'hypothèse nulle alors qu'elle est fausse, le rejet de l'hypothèse nulle alors qu'elle est fausse.

Dans le premier et le dernier cas, la conclusion obtenue est correcte, mais il n'en est malheureusement pas de même dans les deux cas intermédiaires. L'erreur qui consiste à rejeter une hypothèse vraie est appelée erreur de première espèce et désignée par RH_0/H_0. Accepter une hypothèse fausse est une erreur de seconde espèce, elle est désignée par AH_0/H_1.

Les probabilités d'aboutir à de telles conclusions erronées sont les risques de première et de deuxième espèce, désignés respectivement par α et β.

$$\alpha = p(RH_0/H_0) \qquad \beta = p(AH_0/H_1)$$

Le risque de première espèce α est appelé aussi seuil de signification du test, fixé très souvent à 5 %. La probabilité contraire de α désigne le niveau de confiance du test.

$$1-\alpha = p(AH_0/H_0)$$

La probabilité contraire de β désigne la **puissance du test**.

$$1-\beta = p(RH_0/H_1)$$

On peut présenter une table de décision comme suit :

		Décision prise	
		Accepter H_0	Accepter H_1
Hypothèse vraie	H_0	$1-\alpha$ Niveau de confiance	α : erreur de première espèce
	H_1	β : erreur de deuxième espèce	$1-\beta$ Puissance du test

La détermination des valeurs limites de la région d'acceptation de l'hypothèse nulle dépend de l'hypothèse alternative H_1, ainsi on distingue le test bilatéral et le test unilatéral.

2.1. Test bilatéral

Un test est dit bilatéral si la condition de rejet est indépendante du signe de l'écart observé entre les caractéristiques comparées. Les hypothèses formulées du test bilatéral sont :

$$\mathbf{H_0 : \theta = t_0 \quad et \quad H_1 : \theta \neq t_0}$$

θ et t_0 sont les caractéristiques comparées.
La règle de décision peut être représentée ainsi :

$\theta \neq t_0$	$\theta = t_0$	$\theta \neq t_0$
Région de rejet de H_0	Région d'acceptation de H_0	Région de rejet de H_0
	A1 A2	

A1 et A2 sont les valeurs critiques qui délimitent la région d'acceptation.

La région d'acceptation est donc l'intervalle [A1 ; A2].

$$\mathbf{p(A1 \leq t_0 \leq A2) = 1 - \alpha}$$

$$\mathbf{p(t_0 < A1) = p(t_0 > A2) = \alpha/2}$$

2.2. Test unilatéral

Un test est dit unilatéral si l'hypothèse alternative désigne qu'une caractéristique est strictement supérieure ou inférieure à l'autre. On parle respectivement de test unilatéral à droite ou à gauche.

2.2.1 Test unilatéral à droite

Les hypothèses formulées du test unilatéral à droite sont :

$$\mathbf{H_0 : \theta = t_0 \quad et \quad H_1 : \theta > t_0}$$

La règle de décision peut être représentée ainsi :

$\theta \leq t_0$	$\theta > t_0$
Région d'acceptation de H_0	Région de rejet de H_0
A	

A désigne la valeur critique qui délimite la région d'acceptation.

La région d'acceptation est donc l'intervalle]-∞ ; A].

$$\mathbf{p(t_0 \leq A) = 1 - \alpha}$$

$$\mathbf{p(t_0 > A) = \alpha}$$

2.2.2. Test unilatéral à gauche

Les hypothèses formulées du test unilatéral à gauche sont :

$$\mathbf{H_0 : \theta = t_0 \quad et \quad H_1 : \theta < t_0}$$

La règle de décision peut être représentée ainsi :

$\theta < t_0$	$\theta \geq t_0$
Région de rejet de H_0	Région d'acceptation de H_0

A

A désigne la valeur critique qui délimite la région d'acceptation.
La région d'acceptation est donc l'intervalle [A ; +∞ [.

$$\mathbf{p(t_0 < A) = \alpha}$$

$$\mathbf{p(t_0 \geq A) = 1 - \alpha}$$

pour récapituler, la démarche d'un test statistique est formée des étapes suivantes :

1. Formuler les hypothèses H_0 et H_1 ;
2. Fixer le seuil de signification α ;
3. Préciser la loi de probabilité de l'écart observé, appelé aussi variable de décision ;
4. Calculer la valeur numérique de la variable de décision ;
5. Déterminer les valeurs critiques qui délimitent la région d'acceptation ;
6. Prendre la décision et conclure.

III. TESTS STATISTIQUES SUR LES MOYENNES

3.1. Test de conformité d'une moyenne

Formulation de l'hypothèse nulle :

On attribue la valeur m_0 pour moyenne dans une population dont la vraie moyenne m est inconnue, et on veut juger la validité de cette hypothèse.

Ce test a pour but de vérifier si la moyenne m d'une population est ou n'est pas égale à une valeur donnée m_0, appelée norme.

L'hypothèse nulle est donc : $\mathbf{H_0 \quad m = m_0}$

Variable de décision :

On extrait de la population un échantillon aléatoire et simple dans lequel la moyenne observée $\bar{x}$ est en général différente de m_0, il s'agit d'expliquer cette différence.

La variable de décision du test correspond à l'estimation de m qui est la moyenne de l'échantillon :

$$VD = \bar{x}$$

Pour une population normale d'écart type σ connu, la variable de décision est elle-même normale de moyenne m_0 et d'écart type. La variable de décision centrée réduite est donc :

$$VDR = \frac{\bar{x} - m0}{\frac{\sigma}{\sqrt{n}}}$$

VDR est alors une variable normale réduite N(0 ; 1).

Si la distribution de la population parent est inconnue, la quasi-variance sera utilisée comme estimation de la variance de la population. Pour un effectif suffisamment élevé, la variable de décision peut toujours être considérée comme une variable approximativement normale. C'est généralement le cas lorsque l'effectif est supérieur à 30. Dans le cas contraire (n < 30), la variable de décision réduite VDR peut toujours être considérée comme une variable de Student à (n-1) degré de liberté.

Région d'acceptation :

La région d'acceptation dépend de l'hypothèse alternative H_1.

a) Test bilatéral :

$$H_0 : m = m_0 \quad \text{et} \quad H_1 : m \neq m_0$$

Les valeurs critiques qui délimitent la région d'acceptation sont, pour une distribution normale réduite ou asymptotiquement normale réduite, Z1 et Z2 telles que :

$$p(Z1 \leq VDR \leq Z2) = 1 - \alpha$$

$$p(VDR < Z1) = \alpha/2 \quad \Rightarrow \quad Z1 = Z_{\frac{\alpha}{2}}$$

$$p(VDR > Z2) = \alpha/2 \quad \Rightarrow \quad p(VDR \leq Z2) = 1-\alpha/2 \Rightarrow \quad Z2 = Z_{1-\frac{\alpha}{2}}$$

La région d'acceptation est donc l'intervalle [$Z_{\frac{\alpha}{2}}$; $Z_{1-\frac{\alpha}{2}}$].

On accepte l'hypothèse nulle si la variable de décision réduite appartient à la région d'acceptation. Sinon, c'est l'hypothèse alternative qui est acceptée.

Remarque :

Puisque la région d'acceptation est symétrique, on rejette l'hypothèse nulle si :

$$|VDR| > Z_{1-\frac{\alpha}{2}}$$

b) Test unilatéral à droite :

$$H_0 : m = m_0 \text{ et } \quad H_1 : m > m_0$$

La valeur critique qui délimitent la région d'acceptation est, pour une distribution normale réduite ou asymptotiquement normale réduite, Z telle que :

$$p(VDR \leq Z) = 1 - \alpha \quad \Rightarrow \quad Z = Z_{1-\alpha}$$

La région d'acceptation est donc l'intervalle]-∞ ; $Z_{1-\alpha}$].

c) Test unilatéral à gauche :

$$H_0 : m = m_0 \text{ et } \quad H_1 : m < m_0$$

La valeur critique qui délimitent la région d'acceptation est, pour une distribution normale réduite ou asymptotiquement normale réduite, Z telle que :

$$p(VDR < Z) = \alpha \quad \Rightarrow \quad Z = Z_{\alpha}$$

La région d'acceptation est donc l'intervalle [Z_{α} ; +∞[.

Remarque :

Pour une distribution de probabilité inconnue, et lorsque l'effectif de l'échantillon est inférieur à 30, la variable de décision réduite VDR peut toujours être considérée comme une variable de Student à (n-1) degré de liberté. Les valeurs de Z sont remplacées par les valeurs de T de la loi de Student avec (n-1) degré de liberté.

Exemple :

Le diamètre des billes fabriquées par une machine est en moyenne de 6 mm. Pour contrôler si la machine est bien réglée, on a prélevé un échantillon de 50 billes et on a mesuré leur diamètre. On a trouvé :

$$\sum x_i = 350 \qquad \sum x_i^2 = 2462$$

La machine est-elle bien réglée au seuil de signification de 95 %?

Pour répondre à cette question, on doit vérifier si le diamètre moyen des 50 billes observées, est conforme à la norme de 6 mm. Il s'agit donc de faire un test de conformité de la moyenne.

Hypothèse nulle :

Il s'agit d'un test bilatéral H_0 $m = 6$ $H_1 : m \neq 6$

Variable de décision :

La variable de décision du test correspond à l'estimation de m qui est la moyenne de l'échantillon :

$$VD = \frac{\sum xi}{50} = \frac{350}{50} = 7$$

La variable de décision peut être considérée comme une variable approximativement normale.

La variance de la population peut être estimée par la quasi-variance.

$$\hat{\sigma}^2 = \frac{\sum (xi - \bar{x})^2}{n-1} = \frac{n\sigma^2}{n-1} = \frac{50}{49}\left(\frac{2462}{50} - 7^2\right) = 0{,}24$$

$$\hat{\sigma} = \sqrt{0{,}24} = 0{,}49$$

$$VDR = \frac{\bar{x} - m0}{\frac{\hat{\sigma}}{\sqrt{n}}} = \frac{7-6}{\frac{0.49}{\sqrt{50}}} = 14{,}43$$

Région d'acceptation :

La région d'acceptation est l'intervalle $[\, Z_{\frac{\alpha}{2}} \; ; \; Z_{1-\frac{\alpha}{2}} \,]$.

Au seuil de signification de 95 % (α = 0,05), les valeurs critiques qui délimitent la région d'acceptation sont :

$$Z_{\frac{\alpha}{2}} = Z_{0,025} = -1,96$$

$$Z_{1-\frac{\alpha}{2}} = Z_{0,975} = 1,96$$

La région d'acceptation est donc l'intervalle [-1,96 ; 1,96].

On rejette l'hypothèse nulle car la variable de décision réduite n'appartient pas à la région d'acceptation. La machine n'est donc pas bien réglée au seuil de signification de 95 %

3.2. Test de comparaison des moyennes de deux échantillons indépendants

Ce test a pour but de comparer les moyennes de deux populations à l'aide de deux échantillons.

Soient deux échantillons aléatoires et non exhaustifs prélevés respectivement dans une population 1 de moyenne inconnue m_1 et dans une population 2 de moyenne inconnue m_2. les moyennes observées des deux échantillons $\bar{x}_1$ et $\bar{x}_2$ sont en général différentes, il s'agit d'expliquer cette différence.

Formulation de l'hypothèse nulle :

Ce test a pour but de vérifier si la moyenne m_1 d'une population est ou n'est pas égale à la moyenne m_2 d'une autre population.

L'hypothèse nulle est donc : $\mathbf{H_0}$ $\mathbf{m_1 = m_2}$

Variable de décision :

La variable de décision du test correspond à la différence entre les moyennes observées des deux échantillons :

$$VD = \bar{x}_1 - \bar{x}_2$$

Une distinction est faite entre le cas de deux populations de variances inégales et le cas de deux populations de variances égales.

a) cas de deux populations de variances inégales

Pour des populations normales (variances connues), les variables $\bar{x}_1 - \bar{x}_2$ sont des variables normales de moyennes respectivement m_1 et m_2 et d'écarts type respectivement $\frac{\sigma 1}{\sqrt{n1}}$ et $\frac{\sigma 2}{\sqrt{n2}}$.

La variable de décision est elle-même normale de moyenne (m_1-m_2) et d'écart type $\sqrt{\frac{\sigma^2 1}{n1}+\frac{\sigma^2 2}{n2}}$.

Sous l'hypothèse nulle, (m_1-m_2)=0. La variable de décision centrée réduite :

$$VDR = \frac{(\bar{x}_1-\bar{x}_2)}{\sqrt{\frac{\sigma^2 1}{n1}+\frac{\sigma^2 2}{n2}}}$$

Est donc une variable normale réduite N(0 ; 1).

Si les distributions des populations parents sont inconnues, pour des effectifs suffisamment élevés, la variable de décision peut toujours être considérée comme une variable approximativement normale. C'est généralement le cas lorsque les effectifs sont supérieurs à 30. Dans le cas contraire, la variable de décision réduite VDR peut toujours être considérée comme une variable de Student à (n_1 + n_2 - 2) degré de liberté.

b) cas de deux populations de variances inégales

Dans le cas où les populations sont de variances égales, une estimation de la variance commune aux deux populations est donnée par :

$$\hat{\sigma}^2 = \frac{\sum(xi-\bar{x}_1)^2+\sum(xi-\bar{x}_2)^2}{n1+n2-2}$$

la variable de décision réduite devient :

$$VDR = \frac{(\bar{x}_1-\bar{x}_2)}{\sqrt{\frac{\hat{\sigma}^2}{n1}+\frac{\hat{\sigma}^2}{n2}}} = \frac{(\bar{x}_1-\bar{x}_2)}{\sqrt{\hat{\sigma}^2\ (\frac{1}{n1}+\frac{1}{n2})}}$$

$$VDR = \frac{(\bar{x}_1-\bar{x}_2)}{\sqrt{\frac{\sum(xi-\bar{x}_1)^2+\sum(xi-\bar{x}_2)^2}{n1+n2-2}(\frac{1}{n1}+\frac{1}{n2})}}$$

Si les distributions des populations parents sont inconnues, pour des effectifs suffisamment élevés, la variable de décision peut toujours être considérée comme une variable approximativement normale. C'est généralement le cas lorsque les effectifs sont supérieurs à 30. Dans le cas contraire, la variable de décision réduite VDR peut toujours être considérée comme une variable de Student à (n_1 + n_2 - 2) degré de liberté.

Région d'acceptation :

La région d'acceptation dépend de l'hypothèse alternative H_1.

a) Test bilatéral :

$H_0 : m_1 = m_2$ et $H_1 : m_1 \neq m_2$

Les valeurs critiques qui délimitent la région d'acceptation sont, pour des distributions normales réduites ou asymptotiquement normales réduites, Z1 et Z2 telles que :

$$p(Z1 \leq VDR \leq Z2) = 1 - \alpha$$

$$p(VDR < Z1) = \alpha/2 \quad \Rightarrow \quad Z1 = Z_{\frac{\alpha}{2}}$$

$$p(VDR > Z2) = \alpha/2 \quad \Rightarrow \quad p(VDR \leq Z2) = 1-\alpha/2 \Rightarrow \quad Z2 = Z_{1-\frac{\alpha}{2}}$$

La région d'acceptation est donc l'intervalle $[\, Z_{\frac{\alpha}{2}} \; ; \; Z_{1-\frac{\alpha}{2}} \,]$.

On accepte l'hypothèse nulle si la variable de décision réduite appartient à la région d'acceptation. Sinon, c'est l'hypothèse alternative qui est acceptée.

Remarque :

Puisque la région d'acceptation est symétrique, on rejette l'hypothèse nulle si :

$$|VDR| > Z_{1-\frac{\alpha}{2}}$$

b) Test unilatéral à droite :

$H_0 : m_1 = m_2$ et $H_1 : m_1 > m_2$

La valeur critique qui délimitent la région d'acceptation est, pour des distributions normales réduites ou asymptotiquement normales réduites, Z telle que :

$$p(VDR \leq Z) = 1 - \alpha \quad \Rightarrow \quad Z = Z_{1-\alpha}$$

La région d'acceptation est donc l'intervalle $]-\infty \; ; \; Z_{1-\alpha}\,]$.

c) Test unilatéral à gauche :

$H_0 : m_1 = m_2$ et $H_1 : m_1 < m_2$

La valeur critique qui délimitent la région d'acceptation est, pour des distributions normales réduites ou asymptotiquement normales réduites, Z telle que :

$$p(VDR < Z) = \alpha \quad \Rightarrow \quad Z = Z_{\alpha}$$

La région d'acceptation est donc l'intervalle [$Z\alpha$; +∞[.

Remarque :

Pour des distributions de probabilités inconnues, et lorsque les effectifs des échantillons sont inférieurs à 30, la variable de décision réduite VDR peut toujours être considérée comme une variable de Student à (n-1) degré de liberté. Les valeurs de Z sont remplacées par les valeurs de T de la loi de Student avec (n-1) degré de liberté.

Exemple :

Pour savoir s'il existe une différence d'assiduité entre les filles et les garçons, on a choisi de manière aléatoire et simple un premier échantillon de 10 filles et de façon indépendante, un deuxième échantillon de 10 garçons. En fonction des résultats ci-dessous relatifs aux notes d'assiduités (note sur 100), et en supposant que les variances des deux populations sont égales, peut-on conclure, au seuil de 5 %, à l'existence d'une différence significative entre les deux sexes ?

Assiduité des filles	72	67	52	54	46	58	59	54	58	63
Assiduité des garçons	66	59	54	57	63	55	61	55	66	75

Pour répondre à cette question, on doit réaliser un test de comparaison de deux moyennes.

Hypothèse nulle :

Ce test a pour but de vérifier si l'assiduité moyenne m_1 des filles est ou n'est pas égale à l'assiduité moyenne m_2 des garçons.

Il s'agit d'un test bilatéral :

$$H_0 : m_1 = m_2 \quad \text{et} \quad H_1 : m_1 \neq m_2$$

Variable de décision :

Les deux échantillons sont indépendants, les populations sont de variances égales, la variable de décision centrée réduite est donc:

$$VDR = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{\sum(xi-\bar{x}_1)^2+\sum(xi-\bar{x}_2)^2}{n1+n2-2}\left(\frac{1}{n1}+\frac{1}{n2}\right)}} = \frac{58,3-61,1}{\sqrt{\frac{514,1+390,9}{10+10-2}\left(\frac{1}{10}+\frac{1}{10}\right)}} = -0,88$$

Région d'acceptation :

$$|VDR| = 0,88$$

Pour $\alpha = 0,05$, la valeur de $t_{1-\frac{\alpha}{2}}$ avec 18 degrés de liberté est : $t_{0,975} = 2,101$

$|VDR| < t_{1-\frac{\alpha}{2}}$, on accepte donc l'hypothèse nulle. C'est à dire, il n'y a pas de différence significative entre l'assiduité des deux sexes.

3.3. Test de comparaison des moyennes de deux échantillons appariées

Ce test a pour but de comparer les moyennes de deux populations à l'aide de deux échantillons associés par paires. C'est le cas où on soumet les mêmes individus, choisis dans une population donnée, à deux types d'observations.

Formulation de l'hypothèse nulle :

Ce test a pour but de vérifier si la moyenne m_1 d'une population sous une forme donnée est ou n'est pas égale à la moyenne m_2 de la même population sous une autre forme.

L'hypothèse nulle est donc : $\mathbf{H_0 \quad m_1 = m_2}$

Variable de décision :

Soient deux séries de n observations chacune, $x_1, x_2, \ldots, x_n$, et $y_1, y_2, \ldots, y_n$. On travaille avec la série des différences :

$$\mathbf{d_i = x_i - y_i}$$

La variable de décision du test correspond à la moyenne des différences :

$$VD = \bar{d}$$

Pour une population normale, la variable de décision est elle-même normale de moyenne. La variable de décision centrée réduite est donc :

$$VDR = \frac{\bar{d}}{\frac{\sigma_d}{\sqrt{n}}}$$

VDR est alors une variable normale réduite N(0 ; 1).

Si la distribution de la population parent est inconnue, pour un effectif suffisamment élevé, la variable de décision peut toujours être considérée comme une variable approximativement normale. C'est généralement le cas lorsque l'effectif est supérieur à 30. Dans le cas contraire (n < 30), la variable de décision réduite VDR peut toujours être considérée comme une variable de Student à (n-1) degré de liberté.

Région d'acceptation :

La région d'acceptation est identique à celle du test précédent. Elle dépend toujours de l'hypothèse alternative H_1.

a) Test bilatéral :

$$H_0 : m_1 = m_2 \quad \text{et} \qquad H_1 : m_1 \neq m_2$$

La région d'acceptation est l'intervalle $[\ Z_{\frac{\alpha}{2}}\ ;\ Z_{1-\frac{\alpha}{2}}\ ]$.

On accepte l'hypothèse nulle si la variable de décision réduite appartient à la région d'acceptation. Sinon, c'est l'hypothèse alternative qui est acceptée.

Remarque :

Puisque la région d'acceptation est symétrique, on rejette l'hypothèse nulle si :

$$|VDR| > Z_{1-\frac{\alpha}{2}}$$

b) Test unilatéral à droite :

$H_0 : m_1 = m_2$ et $H_1 : m_1 > m_2$

La région d'acceptation est l'intervalle $]-\infty\ ;\ Z_{1-\alpha}\ ]$.

c) Test unilatéral à gauche :

$$H_0 : m_1 = m_2 \quad \text{et} \qquad H_1 : m_1 < m_2$$

La région d'acceptation est l'intervalle $[\ Z_{\alpha}\ ;\ +\infty[$.

Remarque :

Pour des distributions de probabilités inconnues, et lorsque les effectifs des échantillons sont inférieurs à 30, la variable de décision réduite VDR peut toujours être considérée comme une variable de Student à (n-1) degré de liberté. Les valeurs de Z sont remplacées par les valeurs de T de la loi de Student avec (n-1) degré de liberté.

Exemple :

Un chef de produit souhaite tester l'effet d'un nouvel emballage sur les ventes d'un produit. Un échantillon aléatoire de 20 magasins est constitué, puis scindé en deux échantillons de 10 unités, couplés sur la base de leurs ventes hebdomadaires. L'un des magasins de chaque couple propose le produit dans son nouvel emballage, tandis que l'autre magasin présente le produit dans l'ancien emballage. Les ventes enregistrées sont indiquées dans le tableau ci-dessous. Peut-on parler d'un effet positif du nouvel emballage ?

Couple	Nouvel emballage	Ancien emballage	Différence (di)
1	4580	3970	610
2	5190	4880	310
3	3940	4090	-150
4	6320	5870	450
5	7680	6930	750
6	3480	4000	-520
7	5720	5080	640
8	7040	6950	90
9	5270	4960	310
10	5840	5130	710

Pour répondre à cette question, on doit réaliser un test de comparaison de deux moyennes.

Hypothèse nulle :

Ce test a pour but de vérifier si, en moyenne, les ventes enregistrées avec le nouvel emballage m_1 sont ou ne sont pas égales aux ventes enregistrées avec l'ancien emballage m_2.

Il s'agit d'un test unilatéral à droite :

$$H_0 : m_1 = m_2 \quad \text{et} \qquad H_1 : m_1 > m_2$$

Variable de décision :

Les deux échantillons sont associés par paires, la variable de décision centrée réduite est donc:

$$VDR = \frac{\bar{d}}{\frac{\hat{\sigma}_d}{\sqrt{n}}} = \frac{320}{\frac{410,96}{\sqrt{10}}} = 2,462$$

Région d'acceptation :

$$|VDR| = 2,462$$

Pour $\alpha = 0,05$, la valeur de $t_{1-\alpha}$ avec 9 degrés de liberté est : $t_{0,95} = 1,833$

$|VDR| > t_{1-\alpha}$, on rejette donc l'hypothèse nulle. C'est à dire, on peut conclure que le nouvel emballage est plus performant que l'ancien.

3.4. Analyse de la variance

C'est une m**éthode statistique pour tester l'égalité de plusieurs moyennes**. La méthode repose sur les postulats suivants: les échantillons aléatoires proviennent de populations distribuées normalement et ayant la même variance. Comme ces suppositions de base ne sont pas toujours satisfaites en pratique, l'analyste dispose aussi de méthodes dites non paramétriques pour comparer les échantillons entre eux.

Formulation de l'hypothèse nulle

L'analyse de variance, sert à effectuer le test de l'égalité de plusieurs moyennes. On écrit comme suit les hypothèses:

H_0: $m_1 = m_2 = ... = m_J$

H_1: au moins une des moyennes est différente des autres.

En effet, l'analyse de variance est une technique d'analyse statistique qui permet de tester globalement l'égalité des moyennes de J populations normales dans lesquelles on suppose que les variances sont égales ($\sigma_1^2 = \sigma_2^2 = ... = \sigma_J^2 = \sigma^2$), même si elles demeurent inconnues. L'analyse de variance constitue une extension à J populations normalement distribuées, J ≥ 2, du test de comparaison des moyennes de deux échantillons indépendants.

Modèles d'analyse de variance

Les modèles varient selon le nombre de facteurs contrôlés. On aura ainsi le **modèle à un facteur**, le **modèle à 2 facteurs sans interaction** et le **modèle à 2 facteurs avec interaction**.

3.4.1. ANALYSE DE VARIANCE À UN FACTEUR

On essaie de découvrir si un seul facteur peut expliquer ou non les **variations** constatées dans les observations Y_{ij}. Au départ, on dispose d'échantillons prélevés aléatoirement dans des populations normales dans lesquelles les variances sont supposées égales ($\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = ... = \sigma_J^2$). Le tableau suivant illustre la notation indicée: par exemple, $\mathbf{Y_{21}}$ représente la deuxième observation prélevée de la première population. Dans chaque échantillon, on a aussi calculé le total des observations, la moyenne et la variance.

Matrice des données

Population	P1 : N(m1,σ1)	P1 : N(m2,σ2)	…	P1 : N(mj,σj)
	Y_{11} $\mathbf{Y_{21}}$ … y_{n11}	Y_{12} Y_{22} … y_{n22}	…	Y_{1j} Y_{2j} … $y_{nj}j$
Total	T1	T2	…	Tj
Moyenne	$\bar{Y}_1$	$\bar{Y}_2$	…	$\bar{Y}_j$
Variance	$S^2 1$	$S^2 2$	…	$S^2 j$

$N = n_1 + n_2 + \ldots + n_J$ Grand total = T Moyenne générale = $\overline{Y}$

Équation fondamentale de l'analyse de la variance

L'analyse de la variance développée par Fisher repose sur la comparaison de deux estimateurs de la variance σ^2 commune aux J populations normales.

a) Estimation de σ^2 par $\hat{\sigma}_T^2$

Un premier estimateur de σ^2, noté $\hat{\sigma}_T^2$, est obtenu à partir de l'ensemble des $N = n_1 + n_2 + \ldots + n_J$ observations en divisant la somme totale des carrés, STC, par ses degrés de liberté, soit (N-1). La statistique qui en découle est donnée par l'expression suivante:

$$\hat{\sigma}_T^2 = \frac{STC}{N-1} = \frac{\sum_{j=1}^{J}\sum_{i=1}^{n_j}\left(Y_{ij} - \overline{Y}\right)^2}{N-1}$$

b) Estimation de σ^2 par $\hat{\sigma}_M^2$

Un deuxième estimateur de σ^2, noté $\hat{\sigma}_M^2$, est obtenu cette fois en mesurant la variabilité existante entre les moyennes des échantillons. On l'appelle parfois la moyenne des carrés inter-groupes, ou la moyenne des carrés due aux traitements. Dans ce qui suit, on la nomme la moyenne des carrés due au facteur (MCF); elle est calculée en divisant la somme des carrés due au facteur (SCF) par ses degrés de liberté, (J-1):

$$\hat{\sigma}^2_M = MCF = \frac{SCF}{J-1} = \frac{\sum_{j=1}^{J} n_j \left(\overline{Y}_j - \overline{Y}\right)^2}{J-1}$$

c) Estimation de σ^2 par $\hat{\sigma}^2_C$

Un troisième estimateur de σ^2 est obtenu cette fois en combinant les variances intra-échantillons (S_1^2, S_2^2, ..., S_J^2) déjà présentées dans le tableau des données. La pondération attribuée à S_j^2 sera égale aux degrés de liberté de cette statistique, soit $(n_j - 1)$, j=1, 2, ..., J. L'estimateur est appelé la moyenne des carrés due à l'erreur (MCE) et il est donné par les expressions équivalentes suivantes:

$$\hat{\sigma}^2_C = MCE = \frac{SCE}{N-J} = \frac{\sum_{j=1}^{J}(n_j - 1)S_j^2}{N-J} = \frac{\sum_{j=1}^{J}\sum_{i=1}^{n_j}\left(Y_{ij} - \overline{Y}_j\right)^2}{N-J}$$

Les trois sommes de carrés présentées plus haut ne sont pas totalement indépendantes les unes des autres. Il existe en effet un résultat important qui montre que la somme totale des carrés est égale à la somme des deux autres sommes de carrés:

STC = SCF + SCE

C'est cette relation qui s'appelle l'équation fondamentale de l'analyse de la variance. La variabilité totale entre les observations est décomposée en une part due aux différences entre les modalités du facteur et une part de variabilité résiduelle.

Formules équivalentes

Pour effectuer les calculs à l'aide d'une calculatrice électronique, il est préférable d'utiliser les formules suivantes qui sont algébriquement équivalentes aux précédentes:

$$STC = \sum_{j=i}^{J}\sum_{i=1}^{n_j} Y_{ij}^2 - \frac{T^2}{N}$$

$$SCF = \sum_{j=i}^{J} \frac{T_j^2}{n_j} - \frac{T^2}{N}$$

$$SCE = STC - SCF$$

Tableau d'analyse de variance à un seul facteur

Il est d'usage de présenter les résultats d'une analyse de variance à un seul facteur dans un tableau comme celui-ci:

Analyse de variance à un facteur

Source de variation	Somme des carré	Degrés de libert	Moyenne des carré	
Facteur	SCF	J-1	MCF	$\frac{MCF}{MCE}$
Erreur	SCE	N-J	MCE	
Totale	SCT	N-1		

Quand H_o est vraie, MCF et MCE constituent deux estimateurs indépendants de σ^2 de sorte que le rapport $\tilde{F} = \frac{MCF}{MCE}$ obéit à une loi de Fisher avec (J-1) et (N-J) degrés de liberté. En vertu même de la construction du rapport $\tilde{F}$, on devra rejeter l'hypothèse nulle de l'égalité des moyennes $H_o : \mu_1 = \mu_2 = ... = \mu_J$ au seuil α si et seulement si la valeur de $\tilde{F} = \frac{MCF}{MCE}$ est **plus grande** que la valeur critique de la table $F_{1-\alpha}$ ◊ (J-1) et (N-J) dl·

Exemple :

Un manufacturier japonais de puces électroniques songe à implanter une nouvelle usine au Maroc afin de desservir tout le marché nord-africain. Il hésite entre trois villes: Tanger, Casablanca et Eljadida. Selon son point de vue, le critère le plus important à prendre en considération pour déterminer l'emplacement de cette nouvelle usine est l'assiduité au travail des ouvriers.
Le manufacturier a visité au hasard dans chacune des villes considérées cinq grandes usines de fabrication et il a obtenu des administrateurs le taux d'absentéisme par 3500 journées de travail. Les résultats sont reproduits dans le tableau ci-dessous.

Données numériques

Ville	Echantillon	Total	Moyenne	Variance
Tanger	141; 127 ; 111; 124 ; 144	T1 = 647	$\bar{Y}_1 = 129,4$	$S^2_1 = 180,3$
Casablanca	157; 131; 105; 132 ; 163	T2 = 688	$\bar{Y}_2 = 137,6$	$S^2_2 = 539,8$
Eljadida	183; 161; 145 ; 157 ; 189	T3 = 835	$\bar{Y}_3 = 167$	$S^2_3 = 340$
J = 3	N = 15	T = 2170	$\bar{Y} = 144,67$	

A un seuil de 5%, peut-on conclure que le taux d'absentéisme au travail est le même **en moyenne** dans ces 3 villes?

On calcule en premier lieu les trois sommes des carrés:

$$STC = \sum_{j=i}^{J}\sum_{i=1}^{n_j} Y_{ij}^2 - \frac{T^2}{N} = 141^2 + 127^2 + ... + 189^2 - \frac{2170^2}{15} = 8149{,}33$$

$$SCF = \sum_{j=i}^{J} \frac{T_j^2}{nj} - \frac{T^2}{N} = \frac{647^2}{5} + \frac{688^2}{5} + \frac{835^2}{5} - \frac{2170^2}{15} = 3908{,}93$$

$$SCE = 8149{,}33 - 3908{,}93 = 4240{,}40$$

Ce qui permet la construction du tableau de l'analyse de variance.

Tableau d'analyse de variance à un facteur

Source de variation	Somme des carrés	Degrés de liberté	Moyenne des carrés	
Facteur	3908,93	2	1954,467	5,53
Erreur	4240,40	12	353,367	
Totale	8149,33	14		

A un seuil $\alpha = 5\%$, on ne peut pas conclure que l'assiduité des travailleurs à leur travail soit la même en moyenne dans ces 3 villes puisque la valeur observée 5,53 de F est supérieure à la valeur critique $F_{0,95 \text{ à } 2 \text{ et } 12 \text{ dl}} = 3{,}89$ obtenue de la distribution de Fisher à 2 et 12 degrés de liberté.

3.4.2. ANALYSE DE VARIANCE À DEUX FACTEURS SANS INTERACTION

On essaiera dans ce chapitre-ci de découvrir si deux facteurs A et B peuvent expliquer ou non les variations constatées dans les observations aléatoires Y_{ij}.

La matrice des données

Au départ, l'analyste dispose d'échantillons prélevés aléatoirement de populations **normales** dans lesquelles les variances sont présumées égales. Le tableau ci-dessous illustre la notation indicée. Ainsi, $\mathbf{Y_{32}}$ représente la valeur de l'observation prélevée quand le premier facteur est à son troisième niveau (ou modalité) et que le second facteur est à son deuxième niveau; par ailleurs,

$T_{2\bullet}$ et $\overline{Y}_{2\bullet}$ désignent le total et la moyenne des observations quand le premier facteur est maintenu à son deuxième niveau (l'indice sur lequel la sommation a été effectuée est remplacé par un •).
Toutes les combinaisons possibles des modalités des facteurs donnent lieu à IJ «**traitements**». A remarquer qu'il n'y a qu'une seule observation pour chaque traitement, c'est-à-dire une seule valeur numérique dans chacune des cellules du tableau.

Matrice des données

Facteur A \ Facteur B	1	2	.....	J	Total	Moyenne
1	Y_{11}	Y_{12}	.....	Y_{1J}	$T_{1\bullet}$	$\overline{Y}_{1\bullet}$
2	Y_{21}	Y_{22}	.....	Y_{2J}	$T_{2\bullet}$	$\overline{Y}_{2\bullet}$
3	....	$\mathbf{Y_{32}}$	.....			
..... etc.....						
I	Y_{I1}	Y_{I2}	....	Y_{IJ}	$T_{I\bullet}$	$\overline{Y}_{I\bullet}$
Total	$T_{\bullet 1}$	$T_{\bullet 2}$	....	$T_{\bullet J}$	T	
Moyenne	$\overline{Y}_{\bullet 1}$	$\overline{Y}_{\bullet 2}$	....	$\overline{Y}_{\bullet J}$		$\overline{Y}$

Tableau d'analyse de variance à deux facteurs sans répétition

Les résultats d'une analyse de variance à deux facteurs sans répétition se présentent dans un tableau comme celui-ci:

Analyse de variance à deux facteurs sans répétition

Source de variation	Somme des carrés	D.L.	Moyenne des carrés	$\tilde{F}$
Facteur A	SCF_A	I-1	MCF_A	MCF_A / MCE
Facteur B	SCF_B	J-1	MCF_B	MCF_B / MCE
Erreur	SCE	(I-1)(J-1)	MCE	
Totale	STC	IJ-1		

Les diverses sommes des carrés et moyennes des carrés sont calculées à l'aide des formules suivantes:

$$STC = \sum_{i=1}^{I}\sum_{j=1}^{J} Y_{ij}^2 - \frac{T^2}{IJ}$$

$$SCF_A = \sum_{i=1}^{I} \frac{T_{i\bullet}^2}{J} - \frac{T^2}{IJ}$$

$$SCF_B = \sum_{j=1}^{J} \frac{T_{\bullet j}^2}{I} - \frac{T^2}{IJ}$$

$$SCE = STC - SCF_A - SCF_B$$

En se basant sur les résultats présentés au tableau, on déduit que les tests sur le facteur A et sur le facteur B s'effectuent exactement comme dans le cas de l'analyse de variance à un facteur, à savoir au moyen des statistiques :

$$\tilde{F}_A = \frac{MCA}{MCE}$$
$$\tilde{F}_B = \frac{MCB}{MCE}$$

Exemple :

Sur le marché, il existe quatre machines différentes, M_1, M_2, M_3, et M_4 pouvant servir à l'assemblage d'un produit à haute teneur technologique. On a alors décidé de toutes les essayer et d'utiliser les opérateurs qualifiés pour comparer les dites machines. Comme ce travail exige beaucoup de dextérité manuelle de la part de l'usager, on s'attend à ce qu'il y ait des différences importantes entre opérateurs et peut-être aussi entre machines. C'est dans un ordre aléatoire et en laissant écouler beaucoup de temps entre les tests que les opérateurs ont été assignés aux machines afin de contrôler l'effet d'apprentissage. Voici les temps (en minutes) mesurés lors de ces tests.

Données numériques

Opérateurs \ Machines	M_1	M_2	M_3	M_4	Total	Moyenne
O_1	42	45	55	50	192	48
O_2	39	41	52	46	178	44,5
O_3	38	39	48	42	167	41,75
O_4	43	45	54	48	190	47,5
O_5	44	45	56	49	194	48,5
Total	206	215	265	235	921	
Moyenne	41,2	43	53	47		46,05

Y a-t-il des différences significatives au niveau 5% entre les cinq opérateurs d'une part et entre les quatre machines d'autre part quant au temps **moyen** nécessaire à l'assemblage de ce produit?

On calcule en premier lieu les quatre sommes des carrés:

$$STC = \sum_{i=1}^{I}\sum_{j=1}^{J} Y_{ij}^2 - \frac{T^2}{IJ} = 42^2 + 45^2 + ... + 49^2 - \frac{921^2}{20} = 548{,}95$$

$$SCF_A = \sum_{i=1}^{I} \frac{T_{i\bullet}^2}{J} - \frac{T^2}{IJ} = \frac{192^2 + ... + 194^2}{4} - \frac{921^2}{20} = 131{,}20$$

$$SCF_B = \sum_{j=1}^{J} \frac{T_{\bullet j}^2}{I} - \frac{T^2}{IJ} = \frac{206^2 + ... + 235^2}{5} - \frac{921^2}{20} = 410{,}15$$

$$SCE = 548{,}95 - 131{,}20 - 410{,}15 = 7{,}60$$

Ce qui permet la construction du tableau de l'analyse de variance ci-dessous.

Analyse de variance à deux facteurs sans répétition:

Source de variation	Somme des carrés	D.L.	Moyenne des carrés	F
Facteur A= Hommes	131,2	4	32,8	51,79
Facteur B = Machines	410,15	3	136,72	215,87
Erreur	7,6	12	0,63	
Totale	548,95	19		

En examinant les valeurs F observées 51,79 et 215,87 qui sont toutes deux supérieurs aux valeurs théoriques $F_{0,95 \text{ à } 4 \text{ et } 12 \text{ dl}} = 3{,}26$ et $F_{0,95 \text{ à } 3 \text{ et } 12 \text{ dl}} = 3{,}49$ on peut rejeter les deux hypothèses nulles et conclure qu'il y a d'une part, des différences significatives entre les cinq opérateurs quant au temps moyen nécessaire à l'assemblage de ce produit et d'autre part, des différences significatives entre les quatre machines.

3.4.3. ANALYSE DE VARIANCE À DEUX FACTEURS AVEC INTERACTION

Bien des recherches ont pour but d'étudier l'impact de plusieurs facteurs sur le résultat d'une expérience. Dans ce qui suit on tentera de découvrir si deux facteurs **A** et **B** peuvent expliquer ou non les **variations** constatées dans les observations $Y_{ijk.}$

On dispose de IJ échantillons de taille K (K>1) prélevés aléatoirement de populations **normales** dans lesquelles les variances sont présumées égales. Le tableau suivant illustre la notation indicée: par exemple, $\mathbf{Y_{324}}$ renvoie à la quatrième observation prélevée quand le facteur A est à son troisième niveau (ou modalité) et que le facteur B est à son deuxième niveau. Ainsi, $T_{2\bullet\bullet}$ représente le total des observations quand le premier facteur est maintenu à son deuxième niveau, alors que $\overline{Y}_{\bullet 3\bullet}$ désigne la moyenne des observations quand le second facteur est maintenu à sa troisième modalité.
Toutes les combinaisons possibles des modalités des facteurs donnent lieu à IJ «traitements». A remarquer enfin qu'il y a ici le même nombre d'observations dans chacune des IJ cellules, soit K, et cette valeur est supérieure à l'unité.

Matrice des données

Facteur A \ Facteur B	1	2		J	Total	Moyenne
1	Y_{111} Y_{112} ... Y_{11K}	Y_{121} Y_{122} ... Y_{12K}	...	Y_{1J1} Y_{1J2} ... Y_{1JK}	$T_{1\bullet\bullet}$	$\overline{Y}_{1\bullet\bullet}$
...	...	...	...	...	...	...
2	Y_{211} Y_{212} ... Y_{21K}	Y_{221} Y_{222} ... Y_{22K}		Y_{IJ1} Y_{IJ2} ... Y_{IJK}	$T_{I\bullet\bullet}$	$\overline{Y}_{I\bullet\bullet}$
Total	$T_{\bullet 1\bullet}$	$T_{\bullet 2\bullet}$	...	$T_{\bullet J\bullet}$	T = grand total	
Moyenne	$\overline{Y}_{\bullet 1\bullet}$	$\overline{Y}_{\bullet 2\bullet}$		$\overline{Y}_{\bullet J\bullet}$		$\overline{Y}$ = moyenne générale

Tableau d'analyse de variance à deux facteurs avec répétitions

Les résultats d'une analyse de la variance à deux facteurs avec répétitions sont habituellement présentés dans un tableau comme celui-ci

.

Analyse de variance à deux facteurs avec répétitions

Source de variation	Somme des carrés	Degrés de liberté	Moyenne des carrés	$\tilde{F}$
Facteur A	SCF_A	I-1	MCF_A	MCF_A / MCE
Facteur B	SCF_B	J-1	MCF_B	MCF_B / MCE
Interaction	SCI	(I-1)(J-1)	MCI	MCI / MCE
Erreur	SCE	IJ(K-1)	MCE	
Totale	STC	IJK-1		

Les sommes des carrés et les moyennes des carrés sont calculées à l'aide des formules suivantes:

$$STC = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} Y_{ijk}^2 - \frac{T^2}{IJK}$$

$$SCF_A = \sum_{i=1}^{I} \frac{T_{i\bullet\bullet}^2}{JK} - \frac{T^2}{IJK}$$

$$SCF_B = \sum_{j=1}^{J} \frac{T_{\bullet j\bullet}^2}{IK} - \frac{T^2}{IJK}$$

$$SCI = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{T_{ij\bullet}^2}{K} - \sum_{i=1}^{I} \frac{T_{i\bullet\bullet}^2}{JK} - \sum_{j=1}^{J} \frac{T_{\bullet j\bullet}^2}{IK} + \frac{T^2}{IJK}$$

$$SCE = STC - SCF_A - SCF_B - SCI$$

En se basant sur les résultats présentés au tableau, on déduit que les tests sur la présence d'interaction, sur le facteur A et sur le facteur B s'effectuent exactement comme dans le cas de l'analyse de variance à un facteur, à savoir au moyen des statistiques :

$$\tilde{F}_I = \frac{MCI}{MCE}$$

$$\tilde{F}_A = \frac{MCA}{MCE}$$

$$\tilde{F}_B = \frac{MCB}{MCE}$$

L'analyse de variance doit vérifier en premier lieu si l'interaction entre les deux facteurs est importante; si la réponse est négative, on pourra considérer ensuite les deux autres tests disponibles dans le tableau de l'analyse de la variance.

La présence d'interaction entre les deux facteurs signifie que les résultats sous les niveaux d'un facteur se comportent différemment selon les différents niveaux de l'autre facteur.

Exemple :

Il est difficile de prédire le temps nécessaire pour apprendre à programmer en langage C++. On a demandé à 24 programmeurs qui ne connaissaient pas ce langage de prédire le nombre d'heures nécessaires pour apprendre les principales commandes en langage C++ et effectuer ensuite un certain projet. Les programmeurs ont été classifiés selon leur type d'expérience et leur nombre d'années d'expérience. Quand le projet fut terminé, tous sans exception avaient sous-estimé le temps effectivement requis pour accomplir cette tâche. Dans le tableau qui suit, on a ces erreurs de prévision (en heures).

Données numériques

TYPE D'EXPÉRIENCE	NOMBRE D'ANNÉES D'EXPÉRIENCE			
	Moins de 2 ans	Entre 2 et 5 ans	Plus de 5 ans	Total
Sur petits systèmes seulement	25 22 18 20	12 10 14 8	10 9 11 8	167
Sur gros systèmes seulement	30 38 45 44	20 28 29 28	14 15 26 24	341
Total	242	149	117	508

Que ce soit sous l'angle «Type d'expérience» ou «Nombre d'années d'expérience», existe-t-il globalement des différences significatives entre les groupes?

L'analyse de ces données doit vérifier en premier lieu si l'interaction entre les deux facteurs est importante; si la réponse est négative, on pourra considérer ensuite les deux autres tests disponibles dans le tableau de l'analyse de la variance et répondre aux deux questions ci-dessus.

Calculons d'abord les quatre sommes des carrés:

$$STC = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} Y_{ijk}^2 - \frac{T^2}{IJK} = 25^2 + 12^2 + ... + 24^2 - \frac{508^2}{24} = 2737{,}33$$

$$SCFA = \sum_{i=1}^{I} \frac{T_{i\bullet\bullet}^2}{JK} - \frac{T^2}{IJK} = \frac{167^2 + 341^2}{12} - \frac{508^2}{24} = 1261{,}50$$

$$SCFB = \sum_{j=1}^{J} \frac{T_{\bullet j \bullet}^2}{IK} - \frac{T^2}{IJK} = \frac{242^2}{8} + \frac{149^2}{8} + \frac{117^2}{8} - \frac{508^2}{24} = 1054{,}08$$

$$SCI = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{T_{ij\bullet}^2}{K} - \sum_{i=1}^{I} \frac{T_{i\bullet\bullet}^2}{JK} - \sum_{j=1}^{J} \frac{T_{\bullet j \bullet}^2}{IK} + \frac{T^2}{IJK}$$

$$SCI = \frac{85^2 + ... + 79^2}{4} - \frac{167^2 + 341^2}{12} - \frac{242^2 + 149^2 + 117^2}{8} + \frac{508^2}{24} = 61{,}75$$

$$SCE = STC - SCFA - SCFB - SCI = 2737{,}33 - 1261{,}50 - 1054{,}08 - 61{,}75 = 360$$

ce qui permet la construction du tableau de l'analyse de variance suivant :

Analyse de variance à deux facteurs avec répétitions

Source de variation	Somme des carrés	D.L.	Moyenne des carrés	F
Facteur A: Type d'expérience	1261,5	1	1261,5	63,075
Facteur B: Nombre d'années d'expérience	1054,08	2	527,04	26,35
Interaction	61,75	2	30,875	1,54
Erreur	360	18	20	
Totale	2737,33	23		

En examinant en tout premier lieu le test sur l'interaction, on peut vérifier que la valeur F_I = 1,54 est inférieure à la valeur critique de la table, soit $F_{0,95 \text{ à } 2 \text{ et } 18 \text{ dl}} = 3{,}55$. on doit conclure qu'il n'y a pas d'interaction significative entre les deux facteurs Type d'expérience et Nombre d'années d'expérience.

Cette constatation justifie la poursuite de l'analyse de la variance. Comme les valeurs F_A= 63,075 et F_B =26,35 sont supérieures respectivement aux valeurs critiques de la table $F_{0,95 \text{ à } 1 \text{ et } 18 \text{ dl}} = 4{,}41$ et $F_{0,95 \text{ à } 2 \text{ et } 18 \text{ dl}} = 3{,}55$, on doit conclure qu'aussi bien sous l'angle «Type d'expérience» que «Nombre d'années d'expérience», il existe globalement des différences significatives entre les groupes.

IV. TESTS STATISTIQUES SUR LES VARIANCES

4.1. Test de conformité d'une variance

Formulation de l'hypothèse nulle :

Ce test a pour but de vérifier si la variance σ^2 d'une population est ou n'est pas égale à une valeur donnée σ^2_0, appelée norme.

L'hypothèse nulle est donc : $\mathbf{H_0} \quad \sigma^2 = \sigma^2_0$

Variable de décision :

On extrait un échantillon aléatoire non exhaustif de taille n. La variable de décision du test correspond à :

$$VD = \frac{\sum_{i=1}^{n}(xi-\bar{x})^2}{\sigma^2 0}$$

La variable de décision possède une distribution khi deux à (n-1) degrés de liberté.

Région d'acceptation :

La région d'acceptation dépend de l'hypothèse alternative H_1.

Test bilatéral :

$$H_0 : \sigma^2 = \sigma^2_0 \quad \text{et} \qquad H_1 : \sigma^2 \neq \sigma^2_0$$

Les valeurs critiques qui délimitent la région d'acceptation sont χ^2_1 et χ^2_2 telles que :

$$p(\chi^2_1 \leq VD \leq \chi^2_2) = 1 - \alpha$$

$$p(VD < \chi^2_1) = \alpha/2 \quad \Rightarrow \quad \chi^2_1 = \chi^2_{\frac{\alpha}{2}}$$

$$p(VD > \chi^2_2) = \alpha/2 \quad \Rightarrow \quad p(VD \leq \chi^2_2) = 1-\alpha/2 \quad \Rightarrow \quad \chi^2_2 = \chi^2_{1-\frac{\alpha}{2}}$$

La région d'acceptation est donc l'intervalle $[\ \chi^2_{\frac{\alpha}{2}} \ ; \ \chi^2_{1-\frac{\alpha}{2}}\ ]$.

On accepte l'hypothèse nulle si la variable de décision appartient à la région d'acceptation. Sinon, c'est l'hypothèse alternative qui est acceptée.

Test unilatéral à droite :

$$H_0 : \sigma^2 = \sigma^2_0 \quad \text{et} \qquad H_1 : \sigma^2 > \sigma^2_0$$

La valeur critique qui délimitent la région d'acceptation est χ^2 telle que :

$$p(VD \leq \chi^2) = 1 - \alpha \quad \Rightarrow \quad \chi^2 = \chi^2_{1-\alpha}$$

La région d'acceptation est donc l'intervalle $]0 ; \chi^2_{1-\alpha}]$.

Test unilatéral à gauche :

$$H_0 : \sigma^2 = \sigma^2_0 \quad \text{et} \qquad H_1 : \sigma^2 < \sigma^2_0$$

La valeur critique qui délimite la région d'acceptation est χ^2 telle que :

$$p(VD < \chi^2) = \alpha \quad \Rightarrow \quad \chi^2 = \chi^2_{\alpha}$$

La région d'acceptation est donc l'intervalle $[\chi^2_{\alpha} ; +\infty[$.

Exemple :

On souhaite vérifier, au seuil de signification de 95 %, si le peuplement, dans lequel on a mesuré la hauteur d'un échantillon de 12 arbres, appartient à un type de forêt dont l'écart type est de 1,4 m. Les résultats en mètre sont :

5,1 5,2 5,2 5,4 5,9 6,3 6,3 6,8 6,9 6,9 7,0 7,0

Pour répondre à cette question, on doit réaliser un test de conformité de la variance.

Hypothèse nulle :

Il s'agit d'un test bilatéral.

$$H_0 \quad \sigma^2 = 1{,}4^2 = 1{,}96 \qquad H_1 : \sigma^2 \neq 1{,}96$$

Variable de décision :

La variable de décision du test correspond à :

$$VD = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sigma^2_0} = \frac{6{,}6}{1{,}96} = 3{,}37$$

La variable de décision possède une distribution khi deux à 11 degrés de liberté.

Région d'acceptation :

Les valeurs critiques qui délimitent la région d'acceptation sont : $\chi^2_{\frac{\alpha}{2}}$ et $\chi^2_{1-\frac{\alpha}{2}}$

Au seuil de signification de 95 % ($\alpha = 0,05$)

$$\chi^2_{0,025} = 3,82 \qquad \text{et} \qquad \chi^2_{0,975} = 21,9$$

La région d'acceptation est donc l'intervalle [3,82 ; 21,9].

On rejette l'hypothèse nulle car la variable de décision n'appartient pas à la région d'acceptation.

4.2. Test de comparaison des deux variances de deux échantillons indépendants

Ce test a pour but de comparer les variances de deux populations à l'aide de deux échantillons indépendants.

Formulation de l'hypothèse nulle :

Ce test a pour but de vérifier si la variance σ^2_1 d'une population est ou n'est pas égale à la variance σ^2_2 d'une autre population.

L'hypothèse nulle est donc : $\mathbf{H_0} \qquad \boldsymbol{\sigma^2_1 = \sigma^2_2}$

Variable de décision :

Soient deux échantillons aléatoires et non exhaustifs prélevés dans les deux populations. La variable de décision du test correspond au rapport des deux variances observées des deux échantillons :

$$VD = \frac{\hat{\sigma}^2 1}{\hat{\sigma}^2 2}$$

La variable de décision sui une loi de Fisher avec (n_1-1) et (n_2-1) degré de liberté.

Les tables de la loi de Fisher ne donnent que des valeurs supérieures à l'unité. C'est la raison pour laquelle la variable de décision correspond au rapport de variances qui est supérieur à l'unité, d'où l'échantillon 1 est celui qui a la plus grande variance.

Région d'acceptation :

Le test d'égalité de deux variances est en général un test bilatéral. Il précède généralement le test de comparaison des moyennes de deux échantillons indépendants.

$H_0 \quad \sigma^2_1 = \sigma^2_2 \quad$ et $\quad H_1 \quad \sigma^2_1 \neq \sigma^2_2$

Les valeurs critiques qui délimitent la région d'acceptation sont F1 et F2 telles que :

$p(F1 \leq VD \leq F2) = 1 - \alpha$

$p(VD < F1) = \alpha/2 \quad \Rightarrow \quad F1 = F_{\frac{\alpha}{2}}$

$p(VD > F2) = \alpha/2 \quad \Rightarrow \quad p(VD \leq F2) = 1-\alpha/2 \quad \Rightarrow \quad F2 = F_{1-\frac{\alpha}{2}}$

La région d'acceptation est donc l'intervalle $[\, F_{\frac{\alpha}{2}} \; ; \; F_{1-\frac{\alpha}{2}} \,]$.

Les tables de la loi de Fisher ne donnent que des valeurs supérieures à l'unité, de telle sorte que seule est possible la comparaison avec $F_{1-\frac{\alpha}{2}}$, et on rejette l'hypothèse nulle si la variable de décision est supérieure ou égale à $F_{1-\frac{\alpha}{2}}$.

Exemple :

Pour savoir si les filles sont plus assidues que les garçons ou non, on a choisi de manière aléatoire et simple un premier échantillon de 10 filles et de façon indépendante, un deuxième échantillon de 10 garçons. En fonction des résultats ci-dessous relatifs aux notes d'assiduités (note sur 100), peut-on supposer, au seuil de 5 %, que les variances des deux populations sont égales ?

Assiduité des filles	72	67	52	54	46	58	59	54	58	63
Assiduité des garçons	66	59	54	57	63	55	61	55	66	75

Pour répondre à cette question, on doit réaliser un test de comparaison de deux variances.

Hypothèse nulle :

Ce test a pour but de vérifier si la variance σ^2_1 de la population des filles est ou n'est pas égale à la variance σ^2_2 de la population des garçons.

Il s'agit d'un test bilatéral : $H_0 \quad \sigma^2_1 = \sigma^2_2 \quad$ et $\quad H_1 \quad \sigma^2_1 \neq \sigma^2_2$

Variable de décision :

$$VD = \frac{\hat{\sigma}^2 1}{\hat{\sigma}^2 2} = \frac{57,12}{43,43} = 1,31$$

Région d'acceptation :

Pour $\alpha = 0,05$ la valeur de $F_{1-\frac{\alpha}{2}}$ avec 9 et 9 degrés de liberté est :

$$F_{0,975} = 4,03$$

La variable de décision est inférieure à $F_{1-\frac{\alpha}{2}}$, on accepte donc l'hypothèse d'égalité des variances des deux populations.

V. TESTS STATISTIQUES SUR LES PROPORTIONS

5.1. Test de conformité d'une proportion

Formulation de l'hypothèse nulle :

On attribue la valeur p_0 pour proportion dans une population dont la vraie proportion p est inconnue, et on veut juger la validité de cette hypothèse.

Ce test a pour but de vérifier si la proportion p d'une population est ou n'est pas égale à une valeur donnée p_0, appelée norme.

L'hypothèse nulle est donc : $\mathbf{H_0}$ $\mathbf{p = p_0}$

Variable de décision :

On extrait de la population un échantillon aléatoire et simple dans lequel la proportion observée f_n est en général différente de p_0, il s'agit d'expliquer cette différence.

La variable de décision du test correspond à l'estimation de p qui est la fréquence de l'échantillon :

$$\mathbf{VD = f_n}$$

Comme, la distribution de la proportion suit une loi normale de moyenne p et d'écart type $\frac{\sqrt{p(1-p)}}{\sqrt{n}}$ à condition que la taille de l'échantillon soit supérieure ou égale à 30 ($n \geq 30$) et le produit $n\,p \geq 5$, la variable de décision réduite :

$$VDR = \frac{fn - p0}{\sqrt{\frac{p0(1-p0)}{n}}}$$

est donc une variable normale réduite N(0 ; 1).

Région d'acceptation :

La région d'acceptation dépend de l'hypothèse alternative H_1.

Test bilatéral :

$H_0 : p = p_0$ et $H_1 : p \neq p_0$

Les valeurs critiques qui délimitent la région d'acceptation sont les valeurs d'une variable normale réduite Z1 et Z2 telles que :

$p(Z1 \leq VDR \leq Z2) = 1 - \alpha$

$p(VDR < Z1) = \alpha/2 \Rightarrow Z1 = Z_{\frac{\alpha}{2}}$

$p(VDR > Z2) = \alpha/2 \Rightarrow p(VDR \leq Z2) = 1-\alpha/2 \Rightarrow Z2 = Z_{1-\frac{\alpha}{2}}$

La région d'acceptation est donc l'intervalle $[\, Z_{\frac{\alpha}{2}} \; ; \; Z_{1-\frac{\alpha}{2}} \,]$.

On accepte l'hypothèse nulle si la variable de décision réduite appartient à la région d'acceptation. Sinon, c'est l'hypothèse alternative qui est acceptée.

Remarque :

Puisque la région d'acceptation est symétrique, on rejette l'hypothèse nulle si :

$$|VDR| > Z_{1-\frac{\alpha}{2}}$$

Test unilatéral à droite :

$H_0 : p = p_0$ et $H_1 : p > p_0$

La valeur critique qui délimitent la région d'acceptation est la valeur d'une variable normale réduite Z telle que :

$$p(VDR \leq Z) = 1 - \alpha \quad \Rightarrow \quad Z = Z_{1-\alpha}$$

La région d'acceptation est donc l'intervalle $]-\infty ; Z_{1-\alpha}]$.

Test unilatéral à gauche :

$H_0 : p = p_0$ et $H_1 : p < p_0$

La valeur critique qui délimitent la région d'acceptation est la valeur d'une variable normale réduite Z telle que :

$$p(VDR < Z) = \alpha \quad \Rightarrow \quad Z = Z_{\alpha}$$

La région d'acceptation est donc l'intervalle $[Z_{\alpha} ; +\infty[$.

Exemple :

Au cours des élections, un candidat est élu avec 52 % des voix. Plusieurs mois après l'élection, un institut de sondage interroge 1600 électeurs, dont 800 déclarent qu'ils voteraient en cas d'élection, pour le même candidat. Ce résultat est-il ou non significatif d'une désaffection des électeurs pour l'élu ?

Pour répondre à cette question, on doit vérifier si le nouveau pourcentage obtenu par le sondage, n'est pas inférieur à la norme de 52 %. Il s'agit donc de faire un test de conformité de la proportion.

Hypothèse nulle :

Il s'agit d'un test unilatéral à gauche H_0 $p = 0,52$ $H_1 : p < 0,52$

Variable de décision :

La variable de décision du test correspond à la fréquence f_n de l'échantillon :

$$VD = f_n = \frac{800}{1600} = 0,50$$

La distribution de la proportion suit une loi normale de moyenne p et d'écart type $\frac{\sqrt{pq}}{\sqrt{n}}$ (la taille de l'échantillon est supérieure à 30 et le produit n p > 5).

La variable de décision réduite est :

$$VDR = \frac{f_n - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0{,}50-0{,}52}{\sqrt{\frac{0{,}52(1-0{,}52)}{1600}}} = -1{,}60$$

Région d'acceptation :

La région d'acceptation est l'intervalle $[Z\alpha ; +\infty[$.

Au seuil de signification de 95 % ($\alpha = 0{,}05$) : $Z\alpha = Z_{0,05} = -1{,}65$

La région d'acceptation est donc l'intervalle [-1,65 ; +∞[.

On accepte l'hypothèse nulle car la variable de décision réduite appartient à la région d'acceptation. Ce résultat n'est donc pas significatif d'une désaffection des électeurs pour ce candidat.

5.2. Test de comparaison des proportions de deux échantillons indépendants

Ce test a pour but de comparer les proportions de deux populations à l'aide de deux échantillons indépendants.

Formulation de l'hypothèse nulle :

Ce test a pour but de vérifier si la proportion p_1 d'une population est ou n'est pas égale à la proportion p_2 d'une autre population.

L'hypothèse nulle est donc **: H_0 $p_1 = p_2$**

Variable de décision :

Il s'agit de comparer deux proportions observées. Soient deux échantillons aléatoires de taille respectivement n_1 et n_2 extraits de deux populations. Les fréquences observées f_{n1} et f_{n2}
Sont généralement différentes, il s'agit d'expliquer cette différence.

$$f_{n1} = \frac{X1}{n1} \quad \text{et} \quad f_{n2} = \frac{X2}{n2}$$

La variable de décision du test correspond à la différence entre les fréquences observées des deux échantillons :

$$VD = f_{n1} - f_{n2}$$

Comme, les distributions des deux proportions suivent des lois normales de moyennes respectivement p_1 et p_2 et d'écarts types respectifs $\frac{\sqrt{p_1 (1 - p_1)}}{\sqrt{n_1}}$ et $\frac{\sqrt{p_2 (1 - p_2)}}{\sqrt{n_2}}$ à condition que la taille de l'échantillon soit supérieure ou égale à 30 ($n \geq 30$) et le produit $n\,p \geq 5$, la variable de décision est elle-même normale de moyenne (p_1-p_2) et d'écart type $\sqrt{\frac{p_1 (1-p_1)}{n_1} + \frac{p_2 (1-p_2)}{n_2}}$.

Sous l'hypothèse nulle $p_1 = p_2$, il y a la même proportion inconnue p dans les deux populations. Cette proportion peut être estimée par la fréquence observée f_{n1+n2} dans l'échantillon unique qui est la réunion des deux échantillons.

$$f_{n1+n2} = \frac{X_1+X_2}{n_1+n_2} = \frac{n_1 f_{n1}+n_2 f_{n2}}{n_1+n_2}$$

Sous l'hypothèse nulle, la variable de décision suit une loi normale de moyenne $(p_1-p_2) = 0$ et d'écart type :

$$\sqrt{\frac{p (1-p)}{n_1} + \frac{p (1-p)}{n_2}} = \sqrt{f_{n1+n2} (1- f_{n1+n2})(\frac{1}{n_1} + \frac{1}{n_2})}$$

La variable de décision centrée réduite :

$$VDR = \frac{f_{n1} - f_{n2}}{\sqrt{f_{n1+n2}(1- f_{n1+n2})(\frac{1}{n_1}+\frac{1}{n_2})}}$$

est donc une variable normale réduite N(0 ; 1).

Région d'acceptation :

La région d'acceptation est identique à celle du test de conformité d'une proportion, elle dépend de l'hypothèse alternative H_1.

Test bilatéral :

$$H_0 : p_1 = p_2 \quad \text{et} \quad H_1 : p_1 \neq p_2$$

La région d'acceptation est l'intervalle $[\, Z_{\frac{\alpha}{2}} \; ; \; Z_{1-\frac{\alpha}{2}} \,]$.

On accepte l'hypothèse nulle si la variable de décision réduite appartient à la région d'acceptation. Sinon, c'est l'hypothèse alternative qui est acceptée.

Remarque :

Puisque la région d'acceptation est symétrique, on rejette l'hypothèse nulle si :

$$|VDR| > Z_{1-\frac{\alpha}{2}}$$

Test unilatéral à droite :

$$H_0 : p_1 = p_2 \quad \text{et} \quad H_1 : p_1 > p_2$$

La région d'acceptation est donc l'intervalle $]-\infty \; ; \; Z_{1-\alpha}\,]$.

Test unilatéral à gauche :

$$H_0 : p_1 = p_2 \quad \text{et} \quad H_1 : p_1 < p_2$$

La région d'acceptation est donc l'intervalle $[\, Z_{\alpha} \; ; +\infty[$.

Exemple :

Une enquête sur l'emploi a concerné 220 personnes dont 115 dans le milieu rural et 105 dans le milieu urbain. Sur les 115 ruraux enquêtés, 74 se sont révélés actifs, alors que pour les enquêtés urbains, 81 sont actifs. Peut-on admettre, au seuil de 5 %, qu'il n'y a pas de différence significative entre les taux d'activités dans les deux milieux ?

Pour répondre à cette question, on doit réaliser un test de comparaison de deux proportions.

Hypothèse nulle :

Ce test a pour but de vérifier si la proportion p_1 des personnes actives dans le milieu rural est ou n'est pas égale à la proportion p_2 des personnes actives dans le milieu urbain.

Il s'agit d'un test bilatéral : $H_0 : p_1 = p_2$ et $H_1 : p_1 \neq p_2$

Variable de décision :

D'après les données :

$$f_{n1} = \frac{74}{115} = 0,64 \qquad f_{n2} = \frac{81}{105} = 0,77 \qquad f_{n2+n2} = \frac{155}{220} = 0,70$$

La variable de décision centrée réduite est :

$$\text{VDR} = \frac{f_{n1} - f_{n2}}{\sqrt{f_{n1+n2}(1 - f_{n1+n2})(\frac{1}{n_1} + \frac{1}{n_2})}} = \frac{0,64 - 0,77}{\sqrt{0,70(1-0,70)(\frac{1}{115} + \frac{1}{105})}} = -2,10$$

Région d'acceptation :

Pour $\alpha = 0,05$ la valeur de $Z_{1-\frac{\alpha}{2}}$ est : $Z_{0,975} = 1,96$

$|VDR| > Z_{1-\frac{\alpha}{2}}$, on rejette donc l'hypothèse nulle. C'est à dire, il y a une différence significative entre les taux d'activités dans les deux milieux.

5.3. Test de comparaison des proportions de plusieurs échantillons indépendants

Ce test a pour but de comparer les proportions d'un certain nombre de populations à l'aide du même nombre d'échantillons indépendants.

Formulation de l'hypothèse nulle

Ce test a pour but de vérifier si les proportions p_1, p_2, ... p_k de k populations sont égales. On écrit comme suit les hypothèses:

H_o: $p_1 = p_2 = ... = pk$

H_1: au moins une des proportions est différente des autres.

Variable de décision :

Soient k échantillons aléatoires de taille respectivement n_1, n_2 , …, n_k extraits de k populations. Il s'agit de comparer les effectifs observés n_{ij} dans les k échantillons et les effectifs attendus ou théoriques sous l'hypothèse nulle.

Effectifs observés

	Echantillon 1	Echantillon 2	…	Echantillon k
Avoir le caractère étudié	n11	n21	…	nk1
Ne pas avoir le caractère étudié	n12	n22	…	nk12
Total	**n1.**	**n2.**	…	**nk.**

Sous l'hypothèse nulle $p_1 = p_2 = ... = pk$, il y a la même proportion inconnue p dans les k populations. Cette proportion peut être estimée par la fréquence observée f dans l'échantillon unique qui est la réunion des k échantillons.

$$f = \frac{n11+n21+...+nk1}{n1.+n2.+...+nk.}$$

sous l'hypothèse nulle, les effectifs théoriques sont :

Effectifs théoriques

	Echantillon 1	Echantillon 2	…	Echantillon k
Avoir le caractère étudié	f n1.	f n2.	…	f nk.
Ne pas avoir le caractère étudié	(1 – f) n1.	(1 – f) n2.	…	(1 – f) nk1.
Total	**n1.**	**n2.**	…	**nk.**

On est amené à confronter les effectifs observés et les effectifs théoriques. On calcule la variable de décision VD :

$$VD = \Sigma(\text{effectif observé-effectif théorique})^2 / \text{effectif théorique}$$

$$VD = \sum_{i=1}^{k}\left[\frac{(ni1-fni.)^2}{fni.}+\frac{(ni2-(1-f)ni.)^2}{(1-f)ni.}\right]$$

On peut démonter que la variable de décision est une variable aléatoire Khi deux avec (k-1) degré de liberté.

Région d'acceptation :

La variable de décision est nulle lorsque les effectifs observés sont touts égales aux effectifs attendus, c'est à dire, lorsqu'il y a concordance absolue entre la distribution observée et la distribution théorique. La valeur de la variable de décision est d'autant plus grande que les écarts entre les effectifs observés et attendus sont plus grands. La valeur critique qui délimite la région d'acceptation est χ^2 telle que :

$$p(VD < \chi^2) = 1 - \alpha \quad \Rightarrow \quad \chi^2 = \chi^2_{1-\alpha}$$

Le test étant toujours unilatéral, la région d'acceptation est donc l'intervalle $[0 ; \chi^2_{1-\alpha}[$.

On rejettera donc l'hypothèse nulle lorsque la valeur de la variable de décision est supérieure ou égale à $\chi^2_{1-\alpha}$ avec (k-1) degrés de liberté.

Exemple :

Lors d'une campagne électorale, un parti politique a effectué un sondage pour évaluer les intentions de vote en faveur de ce parti. Quatre échantillons indépendants ont été choisis dans quatre villes différentes. On a obtenu les résultats suivants :

	Rabat	Tanger	Oujda	Agadir
Voteront pour le parti	94	58	60	43
Ne Voteront pas pour le parti	240	230	252	197
Total	334	288	312	240

Au seuil de signification de 5 %, la proportion de la population des électeurs qui ont l'intention de voter pour ce parti est-elle identique dans les quatre villes ?

Formulation de l'hypothèse nulle

H_o: $p_1 = p_2 = p_3 = p_4$

H_1: au moins une des proportions est différente des autres.

Variable de décision :

Sous l'hypothèse nulle : $p_1 = p_2 = p_3 = p_4$, il y a la même proportion inconnue p dans les 4 villes. Cette proportion peut être estimée par la fréquence observée f dans l'échantillon unique qui est la réunion des 4 échantillons.

$$f = \frac{94+58+60+43}{334+288+312+240} = 0,22$$

Sous l'hypothèse nulle, les effectifs théoriques sont :

Effectifs théoriques

	Rabat	Tanger	Oujda	Agadir
Voteront pour le parti	73,48	63,36	68,64	52,8
Ne Voteront pas pour le parti	260,52	224,64	243,36	187,2
Total	334	288	312	240

On calcule la variable de décision VD :

$$VD = \frac{(94-73{,}48)^2}{73{,}48}+\frac{(240-260{,}52)^2}{260{,}52}+\frac{(58-63{,}36)^2}{63{,}36}+\frac{(230-224{,}64)^2}{224{,}64}+\frac{(60-68{,}64)^2}{68{,}64}+\frac{(252-243{,}36)^2}{243{,}36}$$

$$\frac{(43-52{,}8)^2}{52{,}8}+\frac{(197-187{,}2)^2}{187{,}2} = 11{,}65$$

La variable de décision est une variable aléatoire Khi deux avec 3 degrés de liberté.

Région d'acceptation :

La région d'acceptation est donc l'intervalle $[0 ; \chi^2_{1-\alpha}[$.

Au seuil de signification de 5 %, la valeur $\chi^2_{0,95}$ à 3 degrés de liberté est égale à 7,81.

La valeur de la variable de décision est supérieure à la valeur $\chi^2_{0,95}$ à 3 degrés de liberté, on rejettera donc l'hypothèse nulle, c'est à dire au seuil de signification de 5 %, la proportion de la population des électeurs qui ont l'intention de voter pour ce parti n'est pas identique dans les quatre villes.

VI. LES TESTS D'AJUSTEMENT

Les tests d'ajustement sont destinés à comparer une distribution observée et une distribution théorique donnée. D'une façon générale, on considère d'une part, une population infinie dont les individus sont classés en k catégories, en fonction d'un critère qualitatif ou quantitatif, et d'autre part, un échantillon aléatoire et simple d'effectif n, dont les individus sont classés de la même manière. Le but du test est de vérifier si la population possède une distribution de probabilité donnée :

$p_1, p_2, p_3, \ldots, p_k$ tel que : $\sum_{i=1}^{k} p_i = 1$

Formulation de l'hypothèse nulle :

Pour comparer la distribution théorique et la distribution observée, on est amené à confronter les effectifs observés n_i et les effectifs attendus ou théoriques correspondants np_i.

L'hypothèse nulle est alors :

$$H_0 : n_i = np_i \text{ avec } \sum_{i=1}^{k} n_i = \sum_{i=1}^{k} np_i = n$$

Variable de décision :

On distingue deux cas d'application de ces tests, selon que la distribution théorique est ou n'est pas complètement définie. Dans le premier cas, la variable de décision peut être calculée immédiatement. Dans le second cas, la distribution de probabilité de la population n'est définie qu'en fonction d'un ou de plusieurs paramètres, ceux-ci doivent préalablement être estimés à partir des données de l'échantillon.

Cas d'une distribution complètement définie :

Pour comparer la distribution théorique et la distribution observée, on est amené à confronter les effectifs observés n_i et les effectifs attendus ou théoriques correspondants np_i.

Les effectifs attendus doivent être tous supérieurs ou égales à 5. quand cette condition n'est pas remplie, on peut regrouper des classes voisines, de manière à augmenter les effectifs attendus.

On calcule la variable de décision VD :

$$VD = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}$$

On peut démonter que la variable de décision est une variable aléatoire Khi deux avec (k-1) degré de liberté. k correspond au nombre de calasses après regroupement.

La variable de décision peut être calculée plus facilement par :

$$VD = \sum_{i=1}^{k} \frac{(ni - npi)^2}{npi} = \sum_{i=1}^{k} \frac{ni^2 - 2ninpi + n^2pi^2}{npi} = \sum_{i=1}^{k} \frac{ni^2}{npi} - \sum_{i=1}^{k} \frac{2ninpi}{npi} + \sum_{i=1}^{k} \frac{n^2pi^2}{npi} = \sum_{i=1}^{k} \frac{ni^2}{npi} - 2n + n$$

$$VD = \sum_{i=1}^{k} \frac{ni^2}{npi} - n$$

Cas d'une distribution incomplètement définie :

Lorsque la distribution théorique n'est pas complètement définie, le ou les paramètres qui caractérisent cette distribution doivent tout d'abord être estimés. On peut calculer ensuite les probabilités estimées $\hat{p}_i$, les effectifs attendus correspondants $n\hat{p}_i$, et la valeur de décision :

$$VD = \sum_{i=1}^{k} \frac{ni^2}{n\hat{p}_i} - n$$

Le nombre de degré de liberté (k-1) doit être réduit du nombre de paramètres estimés.

Région d'acceptation :

La variable de décision est nulle lorsque les effectifs observés sont touts égales aux effectifs attendus, c'est à dire, lorsqu'il y a concordance absolue entre la distribution observée et la distribution théorique. La valeur de la variable de décision est d'autant plus grande que les écarts entre les effectifs observés et attendus sont plus grands. La valeur critique qui délimite la région d'acceptation est χ^2 telle que :

$$p(VD < \chi^2) = 1 - \alpha \Rightarrow \quad \chi^2 = \chi^2_{1-\alpha}$$

Le test étant toujours unilatéral, la région d'acceptation est donc l'intervalle $[0 ; \chi^2_{1-\alpha}[$.

On rejettera donc l'hypothèse nulle lorsque la valeur de la variable de décision est supérieure ou égale à $\chi^2_{1-\alpha}$.

Exemple :

Le tableau suivant donne la distribution de fréquences des nombres de garçons observés dans 1600 familles de 4 enfants, considérées comme choisies au hasard au sein d'une très large population. En fonction de ces résultats, peut-on affirmer, au seuil de 5 %, que le nombre de garçons suit une loi binomiale ?

Nombre de garçons	Nombre de familles
0	113
1	367
2	576
3	426
4	118
Total	1600

Pour répondre à cette question, on doit réaliser un test d'ajustement dans le but de comparer la distribution observée à la une distribution binomiale.

Hypothèse nulle :

$$H_0 : n_i = np_i \qquad \text{avec} \qquad \sum_{i=1}^{k} ni = \sum_{i=1}^{k} npi = n$$

Variable de décision :

Pour comparer la distribution théorique et la distribution observée, on est amené à confronter les effectifs observés n_i et les effectifs attendus ou théoriques correspondants np_i. on doit calculer alors les probabilités p_i en utilisant la loi binomiale.

La probabilité d'avoir un garçon est supposée égale à 0,5, la loi binomiale qui caractérise le nombre de garçons dans une famille de 4 enfants a pour paramètre 4 et 0,5.

En utilisant la formule de la loi binomiale, on trouve les probabilités suivantes :

$$p(x) = C_n^x p^x q^{n-x}$$

Distribution de la variable B(4 , 1/2)

x	p(x)
0	0,0625
1	0,2500
2	0,3750
3	0,2500
4	0,0625
Total	1

Le tableau suivant regroupe les effectifs observés n_i et les effectifs attendus ou théoriques correspondants np_i.

x	n_i	np_i
0	113	100
1	367	400
2	576	600
3	426	400
4	118	100
Total	1600	1600

Les effectifs théoriques sont tous supérieures à 5, on peut calculer la variable de décision :

$$VD = \sum_{i=1}^{k} \frac{ni^2}{npi} - n$$

$$VD = \frac{113^2}{100}+\frac{367^2}{400}+\frac{576^2}{600}+\frac{426^2}{400}+\frac{118^2}{100} - 1600 = 10,3$$

Région d'acceptation :

La région d'acceptation est l'intervalle $[0 ; \chi^2_{1-\alpha}[$.

Pour $\alpha = 0,05$, la valeur de $\chi^2_{1-\alpha}$ avec 4 degrés de liberté est : $\chi^2_{0,95} = 9,49$

La valeur de la variable de décision est supérieure à $\chi^2_{1-\alpha}$, on rejette donc l'hypothèse nulle.

VII. LES TESTS D'INDEPENDANCE

Les tests d'indépendance ont pour but de contrôler l'indépendance stochastique de deux ou plusieurs critères de classification. Ils permettent également d'effectuer des comparaisons de proportions.

Les tests d'indépendance concernent une population subdivisée en pq classes, en fonction de deux critères de classification. La distribution de probabilité correspondante est alors une distribution à deux dimensions, et les données relatives à tout échantillon sont présentées sous la forme d'un tableau de contingence.

Pour des échantillons aléatoires et simples, si les deux critères de classification sont indépendants, les probabilités p_{ij} de la distribution à deux dimensions peuvent être estimées par :

$\hat{p}_{ij} = fi. \times f.j$ avec $f_{i.} = \frac{ni.}{n}$ et $f_{.j} = \frac{n.j}{n}$ sont les fréquences relatives marginales.

$n_{i.}$ et $n_{.j}$ sont les effectifs marginaux, et n_{ij} les effectifs conjoints.
Les effectifs attendus correspondants sont donc :

$$n \hat{p}_{ij} = n f_{i.} \times f_{.j} = n \frac{n_{i.}}{n} \frac{n_{.j}}{n} = \frac{n_{i.} \times n_{.j}}{n}$$

les effectifs attendus doivent touts être supérieurs ou égales à 5.

Formulation de l'hypothèse nulle :

Pour comparer la distribution théorique et la distribution observée, on est amené à confronter les effectifs observés n_{ij} et les effectifs attendus ou théoriques correspondants $n \hat{p}_{ij}$.

L'hypothèse nulle est l'indépendance des deux critères de classification.

$$H_0 : n_{ij} = n \hat{p}_{ij}$$

Variable de décision :

la comparaison des effectifs observés et attendus se fait comme pour les tests d'ajustement, en calculant la variable de décision suivante :

$$VD = \sum_{i=1}^{p} \sum_{j=1}^{q} \frac{n_{ij}^2}{n \hat{p}_{ij}} - n$$

On démontre que la variable de décision est une variable aléatoire Khi deux avec (p-1)(q-1) degré de liberté.

Région d'acceptation :

La valeur critique qui délimite la région d'acceptation est χ^2 telle que :

$p(VD < \chi^2) = 1 - \alpha \quad \Rightarrow \quad \chi^2 = \chi^2_{1-\alpha}$

Le test étant toujours unilatéral, la région d'acceptation est donc l'intervalle $[0 ; \chi^2_{1-\alpha}[$.

On rejettera donc l'hypothèse nulle lorsque la valeur de la variable de décision est supérieure ou égale à $\chi^2_{1-\alpha}$.

Exemple :

Un tour opérateur souhaite segmenter son marché. Il se demande s'il existe un lien entre le choix d'une destination de vacances et le niveau d'instruction. Les données recueillies ont été structurées sous forme de d tableau de contingence.

Niveau d'instruction	Destination de vacances			
	Mer	Montagne	Désert	Total
Primaire	300	50	100	450
Secondaire	250	80	20	350
Supérieur	50	120	30	200
Total	600	250	150	1000

Hypothèse nulle :

L'hypothèse nulle est l'indépendance des deux critères de classification.

$$H_0 : n_{ij} = n \hat{p}_{ij}$$

Variable de décision :

Les effectifs attendus sont estimés par la formule : $n \hat{p}_{ij} = \frac{n_{i.} \times n_{.j}}{n}$

Niveau d'instruction	Destination de vacances			
	Mer	Montagne	Désert	Total
Primaire	270	112,5	67,5	450
Secondaire	210	87,5	52,5	350
Supérieur	120	50	30	200
Total	600	250	150	1000

$$VD = \sum_{i=1}^{p} \sum_{j=1}^{q} \frac{n_{ij}^2}{n \hat{p}_{ij}} - n = \frac{300^2}{270} + \frac{50^2}{112,5} + \frac{100^2}{67,5} + \frac{250^2}{210} + ... + \frac{30^2}{30} - 1000 = 220,91$$

Région d'acceptation :

La région d'acceptation est l'intervalle $[0 ; \chi^2_{1-\alpha}[$.

Le nombre de degrés de liberté est égal à (3-1)(3-1) = 4.

Pour $\alpha = 0,05$, la valeur de $\chi^2_{1-\alpha}$ avec 4 degrés de liberté est $\chi^2_{0,95} = 9,49$.

La valeur de la variable de décision est supérieure à $\chi^2_{1-\alpha}$, on rejette donc l'hypothèse nulle. On conclut donc que le niveau d'instruction a une influence sur le choix d'une destination touristique.

EXERCICES SUR LES TESTS STATISTIQUES

Ex 1 : Un fabricant de tubes à essais pour laboratoire fonde sa publicité sur le fait que la durée de vie de ses tubes correspond à 1500 heures de chauffage à l'aide d'un bec Bunzen. Un laboratoire de contrôle de publicité constate que sur 100 tubes à essais, la durée moyenne de vie est de 1485 heures de chauffage avec un écart-type de 110 heures. Au risque 5%, la durée de vie des tubes à essais est-elle différente de 1500 heures de chauffage ?

Ex 2 : L'expérience suivante a été réalisée par Weldon : il a lancé un dé 315 672 fois, il a tiré 106 602 fois l'une des faces 5 ou 6 Peut-on accepter l'hypothèse selon laquelle le dé est équilibré, au risque de 5% ?

Ex 3 : Le directeur de ventes d'un laboratoire pharmaceutique veut savoir s'il existe des différences significatives entre les régions en terme de niveau d'accueil d'un nouveau produit. Les résultats suivants ont été obtenus auprès d'un échantillon aléatoire de clients :

	Régions			
Niveau d'accueil	Nord	Est	Sud	Ouest
Faible	22	35	0	5
Modéré	84	55	8	24
Elevé	25	17	22	12

Le niveau d'accueil dépend-t-il de la région ?

Ex 4 : Les moteurs des appareils électroménagers d'une marque M ont une durée de vie moyenne de 3000 heures avec un écart-type de 150 heures. À la suite d'une modification dans la fabrication des moteurs, le fabriquant affirme que les nouveaux moteurs ont une durée de vie supérieure à celle des anciens. On a testé un échantillon de 50 nouveaux moteurs et on a trouvé une durée de vie moyenne de 3250 heures avec un écart-type égal à 150 heures. Les nouveaux moteurs apportent-ils une amélioration dans la durée de vie des appareils électroménagers au risque de 1% ?

Ex 5 : Dans une grande ville d'un pays donné, une enquête a été réalisée sur les dépenses mensuelles pour les loisirs. On a observé les résultats suivants:

• Sur 280 familles habitant le centre-ville, les dépenses mensuelles pour les loisirs sont en moyenne de 640 dh avec un écart-type de 120 dh.
• Sur 300 familles habitant la banlieue, les dépenses mensuelles pour les loisirs sont en moyenne de 610 dh avec un écart-type de 100 dh.
Peut-on dire au risque de 5 % que la part du budget familial consacré aux loisirs est différente suivant que la famille habite le centre-ville ou la banlieue ?

Ex 6 : Un fabricant affirme qu'au moins 95 % de l'équipement qu'il fournit à un dépositaire est conforme au cahier des charges. L'examen d'un échantillon de 200 pièces fournies montre que 18 pièces sont défectueuses. Que penser de l'affirmation du fabricant au seuil de confiance de 5 %?

Ex 7 : On prélève dans la production d'une machine, un échantillon de 100 tiges métalliques. La moyenne des longueurs des tiges de cet échantillon est 100,04 cm avec un écart-type de 0,16 cm. La machine est réglée en principe pour obtenir des tiges de 100 cm.

1°) Au risque de 5 %, peut-on dire que la machine est bien réglée ?
2°) Reprendre la question précédente avec un risque de 1 %.

Ex 8 : Pour une élection, on effectue un sondage pour évaluer les intentions de vote en faveur du candidat M. Dans la ville de casa, sur 450 personnes interrogées, 52% ont l'intention de voter pour M. Dans la ville de rabat, sur 300 personnes interrogées, 49 % ont l'intention de voter pour M. Au risque de 5%, y a-t-il une différence d'intention de vote dans ces deux villes?

Ex 9 : Un spécialiste en marketing a fait modifier la méthode traditionnellement utilisée pour effectuer la promotion d'un certain produit. A titre expérimental, il a observé dans 10 points de vente le nombre d'unités vendues en une semaine en utilisant la méthode existante. La semaine d'après, les mêmes points de vente ont utilisé la nouvelle méthode de promotion, on a observé le nombre d'unités vendues en cette semaine en utilisant. Les données recueillies sont comme suit :

Ancienne méthode: 48, 46, 47, 43, 46, 45, 49, 46, 47, 44.
Nouvelle méthode: 56, 49, 53, 51, 48, 52, 55, 53, 49, 50.

La nouvelle méthode de promotion a-t-elle un effet positif sur les ventes ($\propto = 5\%$)?

Ex 10 : Un chercheur a découvert un procédé efficace à 90 % pour prolonger la durée de vie des ballons à eau chaude. On teste son procédé sur 200 ballons. On constate qu'il est efficace pour 160 d'entre eux. L'affirmation du chercheur est-elle légitime au seuil de signification de 0,05 ?

Ex 11 : Un laboratoire annonce que l'un de ses médicaments est efficace à 95 %. Sur un échantillon de 400 personnes le traitement s'est révélé efficace sur 355 d'entre elles. Quel risque faut-il accepter si l'on considère que l'affirmation du laboratoire est légitime ?

Ex 12 : Dans le but de contrôler le poids net des sachets d'un produit alimentaire, on a prélevé deux échantillons respectivement de 10 et 12 sachets, on a obtenu les résultats suivant (en grammes) :

Éch 1	190	200	202	195	194	208	205	196	198	206		
Éch 2	210	204	203	189	194	195	206	205	200	201	198	197

Ces deux résultats sont-ils significativement différents en ce qui concerne le poids moyen %

Ex 13 : Au concours d'entrée à une école, l'épreuve de culture générale est notée de 0 à 50. on tire au hasard un échantillon de 100 candidats et l'on relève que les notes qu'ils ont obtenues se classent en cinq tranches de la manière suivante :

Tranches de notes	Nombre de candidats
Note ≤ 10	10
10 < Note ≤ 20	20
20 < Note ≤ 30	30
30 < Note ≤ 40	20
40 < Note ≤ 50	20

Le jury se demande s'il est justifié de considérer que la distribution des notes suit une loi normale dans la population de tous les candidats.

Ex 14 : 24 têtes d'ovin ont reçu 6 alimentations différentes pour constituer 4 répétitions et on a enregistré les gains moyens quotidiens en poids suivants :

Alim. 1	Alim. 2	Alim. 3	Alim. 4	Alim. 5	Alim. 6
590	460	600	640	690	690
760	430	460	660	600	650
700	540	610	720	550	680
640	470	510	580	480	740

Au seuil de 5 %, existe-t-il une différence significative quant à l'effet des différentes alimentations sur le gain moyen quotidien en poids des ovins ?

Ex 15 : L'expérience suivante avait pour but d'analyser l'impact des 2 facteurs Sexe et Âge sur la consommation d'un certain produit de luxe. Dans chacun des 6 groupes, le produit a été offert à 100 personnes choisies au hasard. La consommation, en nombre d'unités achetées, est donnée dans le tableau qui suit:

Sexe	Catégorie d'âge			
	Moins de 20 ans	Entre 20 et 45 ans	Plus de 45 ans	Total
Féminin	27	39	54	120
Masculin	32	45	62	139
Total	59	84	116	259

On suppose que les nombres d'unités achetées obéissent à des lois normales, que les variances sont égales dans ces six populations.

Quant au nombre d'unités achetées en moyenne, peut-on affirmer au niveau □ = 0.01 qu'il y a une différence significative entre hommes et femmes d'une part, et entre les trois groupes d'âge, d'autre part?

Ex 16 : Une entreprise commerciale à succursales multiples procède à un sondage dans ses magasins de rabat et casa. A rabat, sur 1000 clients interrogés, 350 déclarent souhaiter que le magasin reste ouvert jusqu'à 21 heures tandis qu'à casa, sur 900 clients, 280 ont émis ce même vœu. L'entreprise peut-elle, au seuil de signification de 5%, considérer que sa clientèle de rabat réagit comme celle de casa ?

Ex 17 : Une machine fabrique des pièces identiques. La moyenne des poids de 50 pièces prélevées dans la production est 68,2 grammes avec un écart-type de 2,5 grammes. On effectue un réglage sur la machine. On prélève un nouvel échantillon de 50 pièces. On trouve un poids moyen de 67, 5 grammes avec un écart-type de 2, 8 grammes. Peur-on affirmer, au risque 5 % que le réglage a modifié le poids des pièces ?

Ex 18 : Les ventes quotidiennes d'ordinateurs réalisées par une société informatique durant les 3 premiers mois de 2001, du lundi au jeudi sont comme suit :

	Janvier 2010	**Février** 2010	**Mars** 2010
lundi	13 9 8 7	9 5 8 12	7 15 14 10
mardi	8 6 6 7	11 4 9 5	17 14 12 13
mercredi	6 10 7 4	10 2 8 3	6 14 12 13
jeudi	1 10 7 5	6 10 12 9	10 8 4 9

En supposant les conditions de l'analyse de la variance satisfaites, peut-on dire qu'il y a une différence significative à un seuil de 5% entre les moyennes des ventes réalisées chaque mois et entre les moyennes des ventes réalisées chaque jour ?

Ex 19 : Dans une population, soit p1, la proportion d'hommes possédant le baccalauréat et p2 la proportion de femmes possédant le baccalauréat. Le tableau suivant correspond à la répartition de 200 individus choisis au hasard dans cette population.

	hommes	femmes
Possèdent le bac	32	26
ne possèdent pas le bac	64	78

Peut-on affirmer au risque 0,05, que p1 et p2 sont significativement différents ?

Ex 20 : Dans un pays M, le gouvernement a annoncé que le taux de chômage est de 15,6 %. Contestant ce chiffre, les députés de l'opposition ont fait appel à un institut de sondage. Celui ci a réalisé une étude couvrant 4900 personnes en âge d'activité et a trouvé que le taux de chômage est de 16,4 %. Avec un niveau de confiance de 0,95 ; estimez-vous que l'opposition a raison de contester le chiffre annoncé par le gouvernement ?

Ex 21 : Une enquête a été réalisée au près d'un échantillon de 500 individus prélevé au sein d'une population cible de 4 millions d'individus. Les données que l'on possède sur cette population sont les suivantes :
Hommes 48% soit 1,92 millions d'hommes
Femmes 58% soit 2,08 millions de femmes

Sexe	Hommes		Femmes		Total
Niveau d'instruction	%	Effectifs	%	Effectifs	
Aucun	35	672000	50	1040000	1712000
Primaire	30	576000	25	520000	1096000
Secondaire	15	288000	10	208000	496000
Formation professionnelle	13	249600	10	208000	457600
Supérieur	7	134400	5	104000	238400
Total	100	1920000	100	2080000	4000000

Au dépouillement, on a trouvé que les individus qui ont formé l'échantillon ont les caractéristiques suivantes :

Sexe	Hommes		Femmes		Total
Niveau d'instruction	%	Effectifs	%	Effectifs	
Aucun	32	61	54	112	173
Primaire	28	54	23	48	102
Secondaire	18	35	12	25	60
Formation professionnelle	14	27	8	17	44
Supérieur	8	15	3	6	21
Total	100	192	100	208	400

L'échantillon prélevé est-il représentatif de la population étudiée ?

Ex 22 : Dans une population, on interroge un échantillon aléatoire de 400 personnes dont 160 sont âgées de 18 à 40 ans et 240 sont âgées de plus de 40 ans. On a trouvé que le pourcentage des personnes propriétaires de leur logement dans les deux groupes sont respectivement 35% et 45%. Ces deux résultats sont-ils significativement différents au seuil de signification de 5% ?

Ex 23 : On a enregistré plusieurs fois de suite le nombre de personnes qui se sont présenté à un guichet automatique bancaire, pendant des temps de 5 minutes.

Nombres d'arrivées	0	1	2	3	4	5	6	7	8	9	10
Fréquences absolues observées	1	4	12	18	22	17	11	6	4	3	2

Peut-on affirmer au seuil de signification de 5 % que le nombre de personnes qui se présentent à un guichet automatique bancaire, pendant un intervalle de temps de 5 minutes suit une loi de poisson ?

Ex 24 : Le tableau suivant donne le nombre d'étudiants qui ont été brillants et médiocres devant trois examinateurs :

	Examinateur1	Examinateur2	Examinateur3	Total
Brillants	50	47	56	153
Médiocres	5	14	8	27
Total	55	61	64	180

Au seuil de 5 %, testez l'hypothèse selon laquelle le nombre d'étudiants médiocres est le même pour chaque examinateur.

Ex 25 : On a mesuré la longueur, en mm, de 75 grains de blé. Les résultats obtenus ont été répartis en neuf classes;

longueur en mm	Nombre de grains
[5,25 ; 5,75[	1
[5,75 ; 6,25[	6
[6,25 ; 6,75[	6
[6,75 ; 7,25[	9
[7,25 ; 7,75[	15
[7,75 ; 8,25[	17
[8,25 ; 8,75[	10
[8,75 ; 9,25[	8
[9,25 ; 9,75[	3

Peut-on ajuster à cette distribution une la loi normale de moyenne 7,75 mm, et d'écart type 0,94 mm ? (seuil de signification de 5 %)

Ex 26 : Quelques jours avant une consultation électorale mettant deux candidats A et B en présence, une société d'étude effectue un sondage auprès des électeurs afin d'estimer le pourcentage des voix que chaque candidat est susceptible de recueillir dans l'ensemble du corps électoral.

a) 2304 personnes sont interrogées ; 1267 se prononcent en faveur du candidat A. On demande d'estimer l'intervalle de confiance contenant le pourcentage de voix que le candidat A pourrait obtenir (α = 5 %).

b) Quelques mois après deux instituts de sondage interrogent à nouveau les électeurs. Pour l'institut X, qui a interrogé 1600 personnes, le candidat A ne recueillerait que 47 % des suffrages. Pour l'institut Y, qui a interrogé 2500 personnes, A recueillerait 50 % des suffrages.

Ces deux résultats sont-ils significativement différents avec un degré de confiance de 95 % ?