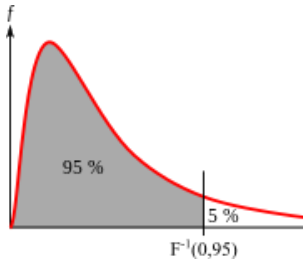
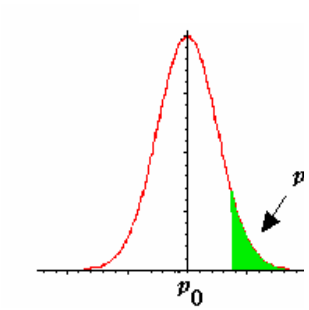
COURS DE STATISTIQUE MATHEMATIQUE (1)

Conforme au Canevas - Master Mathématiques Appliquées

Préparé par Farida LAOUDJ CHEKRAOUI
Maître de Conférences B

Université Mohamed Sedik ben Yahia - Jijel
Faculté des Sciences exactes et Informatique
Département des Mathématiques et Informatique





COURS DE STATISTIQUE MATHEMATIQUE (1)

Conforme au Canevas – Master Mathématiques appliquées

Préparé par Farida LAOUDJ CHEKRAOUI

Maître de conférences B – Faculté des sciences exactes et Informatique

Département des Mathématiques et Informatique

Université Mohamed Sedik ben Yahia - Jijel

Table des matières

Ce cours de la statistique mathématique (1), enseigné aux étudiants du Master Mathématiques Appliquées (EDP et probabilités) au septième semestre, traite les différents aspects de la statistique inférentielle d'une façon simple afin d'introduire le cours de la statistique mathématique (2) enseigné au huitième semestre.

Il est rédigé sur la base de plusieurs ouvrages et aussi d'un document internet intéressant (voir la bibliographie).

Introduction

La démarche statistique consiste à traiter et à interpréter les données et les informations recueillies. Elle inclut deux aspects : un aspect descriptif (ou exploratoire) et un aspect inférentiel (ou décisionnel).

La statistique descriptive vise à synthétiser, résumer et structurer les données et les informations collectées. Elle les représente sous forme de graphiques, de tableaux et d'indicateurs statistiques. La statistique descriptive s'est enrichie de nombreuses méthodes exploratoires ; de visualisation de données multidimensionnelles telles que l'analyse des composantes principales, l'analyse factorielle des correspondances, l'analyse des composantes multiples…

La statistique inférentielle (ou la statistique mathématique) vise à étendre les propriétés constatées sur un échantillon à la population toute entière. Il s'agit d'estimer un paramètre inconnu dans la population à l'aide d'un échantillon, de valider ou affirmer une hypothèse dite « hypothèse de travail » formulée après une phase exploratoire et descriptive. Tout au long de cette phase, le calcul des probabilités joue un rôle très important et l'analyse des propriétés mathématiques des estimateurs est systématiquement présentée dans ce cours.

La modélisation et la prévision statistique est possible de l'inclure dans l'aspect inférentiel et décisionnel**.** Elle a pour but de schématiser une réalité par un modèle et de rechercher une relation approximative entre une variable et plusieurs autres. La modélisation à souvent deux objectifs : explication d'un phénomène et/ou réalisation des prévisions à court, à moyen ou à long terme.

Ce cours est organisé autour de quatre chapitres. Le premier sera consacré aux rappels sur les lois de probabilité usuelles et le comportement asymptotique. Il est intitulé « Rappels sur les lois de probabilité usuelles et le comportement asymptotique ». Dans le deuxième chapitre, on présente les méthodes d'estimation statistique et il est intitulé « Estimation statistique ». Le troisième chapitre traite quelques tests statistiques paramétriques et il est intitulé « Les tests statistiques paramétriques ». On finit ce cours par deux tests non paramétriques, à savoir le test de Khi-deux d'indépendance et le test de khi-deux d'adéquation, qui seront exposés dans le chapitre 04.

Chapitre 01 : Rappels sur les lois de probabilité usuelles et le comportement asymptotique

Nous allons au préalable présenter quelques lois et leurs propriétés qui seront utiles pour la suite de ce cours.

I. Lois discrètes d'usage courant

1. *Loi Bernoulli de paramètre p*

Soit A $\in \mathcal{A}$ un évènement quelconque ($\mathcal{A}$ tribu de parties de l'ensemble fondamental Ω) ; on appelle variable indicatrice de A, la variable aléatoire définie par X = $1\mathrm{I}_A$:

$$X(w) = 1\mathrm{I}_A(w) = \begin{cases} 1 \ si \ w \in A \\ 0 \ si \ w \in \bar{A} \end{cases}$$

Ainsi X(Ω) = {0,1}, il s'agit alors d'une expérience à deux issues seulement (échec, succès) avec :

$P(X = 1) = P\{w \in \Omega / X(w) = 1\} = P(A) = p$

$P(X = 0) = P\{w \in \Omega / X(w) = 0\} = P(\bar{A}) = 1\text{-}p$

On dit que X suit la loi Bernoulli de paramètre p = P(A) et on la note $X \rightsquigarrow B(p)$

L'espérance mathématique de X est égale à :

$E(X) = \sum_{i=1}^{2} xi \ P(X = xi) = 1 . p + 0 . (1\text{-}p) = p$

La variance de X est égale à :

$V(X) = E(X^2) - E(X)^2 = \sum_{i=1}^{2} xi^2 \ P(X = xi) - p^2 = 1^2 . p + 0^2 . (1\text{-}p) - p^2 = p - p^2$

$V(X) = p(1\text{-}p) = pq$

2. *Loi Binomiale de paramètres n et p*

Supposons que l'on répète n fois dans les mêmes conditions une expérience aléatoire Bernoulli, dont l'issue se traduit par l'apparition ou la non apparition de A de probabilité p, le résultat d'une expérience étant indépendant des résultats précédents ; soit X le nombre d'apparitions de A parmi ces n expériences, X(Ω) = {0,1,…..n}. On dit que X une variable aléatoire binomiale ; notée $X \rightsquigarrow B(n, p)$.

La loi binomiale se caractérise aussi par la constance au cours des épreuves élémentaires de la probabilité d'apparition de l'évènement élémentaire A; ce qu'implique l'indépendance des épreuves.

Pour $K \in X(\Omega)$:

$$P(X=k) = C_n^k p^k (1-p)^{n-k} \quad = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$$

Puisque les Xi sont indépendants :

L'espérance mathématique d'une variable aléatoire qui suit une loi binomiale est :

$$E(X) = n\,p$$

Et la variance d'une variable aléatoire qui suit une loi binomiale est :

$$V(X) = n\,p\,(1-p) = n\,pq$$

Remarques

- La somme de deux variables qui suivent la loi binomiale suit aussi une loi binomiale : Soient X_1 et X_2 deux variables aléatoires indépendantes et de lois binomiales de paramètres (n_1, p) et (n_2, p) respectivement ; alors $X_1 + X_2 \rightsquigarrow B(n_1+n_2, p)$.
- Si l'on considère la variable aléatoire X/n, que l'on notera F_n (dite une fréquence), et si la valeur particulière prise par X était k, alors la valeur correspondante de F_n sera k/n notée f et

$$P(F_n = f) = C_n^k p^k (1-p)^{n-k} \quad = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$$

L'espérance mathématique de la fréquence :

$$E(F_n) = E\left(\frac{X}{n}\right) = \frac{1}{n} E(X) = p$$

La variance de la fréquence :

$$V(F_n) = V\left(\frac{X}{n}\right) = \frac{1}{n^2} V(X) = \frac{1}{n^2} n\,p\,(1-p) = \frac{p\,q}{n}$$

- Voici un résultat utile pour l'utilisation des tables : Si X suit une loi binomiale $B(n, p)$, alors la variable aléatoire (n-X) suit une loi binomiale de paramètres n et q : $B(n, q)$.

3. Loi de Poisson de paramètre λ

Lorsqu'on étudie un évènement rare dont on ne connait que la moyenne d'apparition dans une unité d'espace ou de temps, on utilise la loi de poisson. Elle peut être considérée comme une approximation de la loi binomiale, mais elle se justifie par elle-même. C'est la loi d'une variable aléatoire X entière positive ou nulle qui vérifie et satisfait : $P(X = k) = e^{-\lambda}\frac{\lambda^k}{k!}$ où $k \in IN$.

Les propriétés importantes de la loi de poisson sont :

- L'espérance mathématique et la variance de X sont identiques : E(X) = V(X) = λ
- Le rapport de la probabilité d'avoir k événements, à celle d'avoir k-1 événements est égal au quotient du paramètre de poisson λ au nombre k événements :

$$\frac{P(X=k)}{P(X=k-1)} = \frac{e^{-\lambda}\frac{\lambda^k}{k!}}{e^{-\lambda}\frac{\lambda^{k-1}}{(k-1)!}} = \frac{e^{-\lambda}\frac{\lambda\ \lambda^{k-1}}{k(k-1)!}}{e^{-\lambda}\frac{\lambda^{k-1}}{(k-1)!}} = \frac{\lambda}{k}$$

d'où : $\frac{P(X=k)}{P(X=k-1)} = \frac{\lambda}{k}$, $k \geq 1$

- La somme de deux variables aléatoires X_1 et X_2 indépendantes et suivent la loi de poisson de paramètres λ_1 et λ_2 respectivement suit également une loi de poisson de paramètre $\lambda_1 + \lambda_2$: $X_1 + X_2 \rightsquigarrow P(\lambda_1 + \lambda_2)$.

II. Lois continues d'usage courant

1. Loi exponentielle de paramètre θ

La loi exponentielle est généralement utilisée pour modéliser le temps de vie d'un phénomène notée $X \rightsquigarrow \mathcal{E}(\theta)$, $\theta > 0$. Sa densité est :

$$f(x) = \begin{cases} \theta e^{-\theta x} & si\ x > 0 \\ 0 & si\ non \end{cases}$$

L'espérance mathématique de la variable aléatoire X est égale à : $E(X) = \frac{1}{\theta}$

Sa variance est égale à : $V(X) = \frac{1}{\theta^2}$

2. Loi Gamma de paramètre r

La loi exponentielle est un cas particulier d'une famille de loi Gamma γ.

Soit une variable aléatoire positive X suit une loi gamma de paramètre r, notée $X \rightsquigarrow \gamma(r,1)$

Si sa densité est :

$f(x) = \frac{1}{\Gamma(r)} exp(-x)x^{r-1}$, r entier ≥ 1

La fonction $\Gamma(r) = \int_0^{+\infty} x^{r-1}\ e^{-x}\ dx$

L'espérance de X est : $E(X) = r$

Puisque :

$E(X) = \int_{-\infty}^{+\infty} x\, f(x)dx = \frac{1}{\Gamma(r)} \int_0^{+\infty} x^r\ e^{-x}\ dx = \frac{\Gamma(r+1)}{\Gamma(r)} = r$

La variance de X est : $V(X) = r$

Puisque :

$V(X) = E(X^2) - E(X)^2 = \frac{1}{\Gamma(r)} \int_x^{+\infty} x^{r+1}\ e^{-x}\ dx - r^2$

$V(X) = \frac{\Gamma(r+2)}{\Gamma(r)} - r^2 = (r+1)\ \frac{\Gamma(r+1)}{\Gamma(r)} - r^2 = r(r+1) - r^2 = r$

3. Loi normale ou loi de Laplace –Gauss LG

Un grand nombre de variables quantitatives suivent une loi normale (appelée aussi loi de Gauss-Laplace).

Les paramètres qui suffisent à définir mathématiquement cette loi sont la moyenne m et l'écart-type σ de cette variable aléatoire. On note : $X \rightsquigarrow N(m,\sigma)$

La fonction de densité de probabilité est définit par :

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

La loi normale joue un rôle fondamental en probabilités et en statistique mathématique car c'est un modèle fréquemment utilisé dans plusieurs domaines.

Les propriétés importantes de la loi normale sont :

- L'espérance mathématique de la variable aléatoire X est égale à : E(X) = m
- Sa variance est égale à : V(X) = σ^2
- La fonction $f(x)$ est symétrique.
- La somme (ou la différence) de deux variables aléatoires indépendantes X_1 et X_2 suivant respectivement les lois normales $N(m_1,\sigma_1)$, $N(m_2,\sigma_2)$ suit elle-même une loi normale de moyenne $m_1 \pm m_2$ et d'écart-type $\sqrt{\sigma_1^2 + \sigma_2^2}$

Avec le changement de variable aléatoire Z centrée et réduite telle que $Z = \frac{X-m}{\sigma}$, la densité de la variable aléatoire Z est : $f(z) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2} z^2)$ et $Z \rightsquigarrow N(0,1)$ loi normale centrée et réduite.

Le passage d'une loi normale de paramètre m et l'écart-type σ à la loi normale centrée et réduite est utile pour le calcul des probabilités puisque cette dernière est tabulée.

Démonstration :

$$E(Z) = E(\frac{X-m}{\sigma}) = \frac{1}{\sigma} E(X-m) = \frac{1}{\sigma} E(X) - \frac{1}{\sigma} m = \frac{1}{\sigma} m - \frac{1}{\sigma} m = 0$$

$$V(Z) = \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} z^2 \ exp(-\frac{1}{2} z^2)\, dz = \frac{2}{\sqrt{2\pi}} \int_0^{+\infty} z^2 \ exp(-\frac{1}{2} z^2)\, dz$$

Posant : $t = \frac{z^2}{2}$, on a : $z\, dz = dt$

$$V(Z) = \frac{2}{\sqrt{\pi}} \int_0^{+\infty} exp(-t)\, dt = \frac{2}{\sqrt{\pi}} \Gamma(\frac{3}{2}) = \frac{2}{\sqrt{\pi}} \frac{1}{2} \Gamma(\frac{1}{2}) = \frac{2}{\sqrt{\pi}} \frac{1}{2} \sqrt{\pi} = 1$$

4. *Lois déduites de la loi normale (lois empiriques)*

a) ***Loi de Khi-deux χ^2 ou loi de Pearson***

On peut déduire la loi de Khi-deux de la loi normale.

Soient X_1, X_2, ……X_n n variables aléatoires indépendantes de même loi (identiquement distribuées idd) : $X \rightsquigarrow N(m,\sigma)$

On définit une autre variable aléatoire Z telle que $Z = \frac{X-m}{\sigma}$

On a :

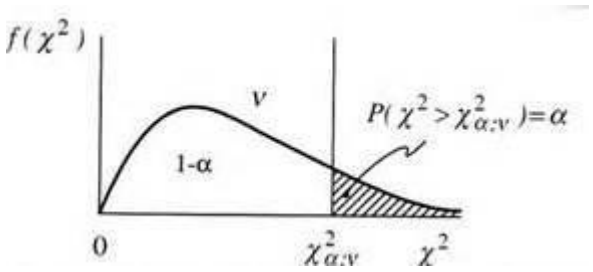
$Z_i \rightsquigarrow N(0,1)$ pour i $=\overline{1,n}$.

$Z_i^2 \rightsquigarrow \chi^2$ à 1 degré de liberté (ddl)

$Z_1^2 + Z_2^2 \rightsquigarrow \chi^2$ à 2 ddl

$\sum_{i=1}^{n} Z_i^2 \rightsquigarrow \chi^2$ à n ddl

La loi de Khi-deux n'est pas symétrique et sa distribution est sous la forme suivante :



La loi de khi-deux à n degré de liberté (ddl) est la loi de γ(n/2, ½) où n est un entier positif, donc de densité pour x>0 :

$$f(x) = \frac{1}{2^{\frac{n}{2}}\Gamma(n/2)} exp\left(-\frac{x}{2}\right) x^{\frac{n}{2}-1}$$

Les moments de la loi de Khi-deux se déduisent alors de la loi gamma. L'espérance mathématique d'une variable aléatoire qui suit la loi de χ^2 est égale à :

$E(\chi^2) = \frac{n/2}{1/2} = n$ et sa variance est égale à : $V(\chi^2) = \frac{n/2}{1/4} = 2n$

b) ***Loi de Student***

Soit U une variable aléatoire qui suit une loi normale centrée et réduite $U \rightsquigarrow N(0,1)$

Soient n variables aléatoires indépendantes telles que $\sum_{i=1}^{n} U_i^2$ suit une loi de Khi-deux à n degré de liberté : $\sum_{i=1}^{n} U_i^2 \rightsquigarrow \chi_n^2$

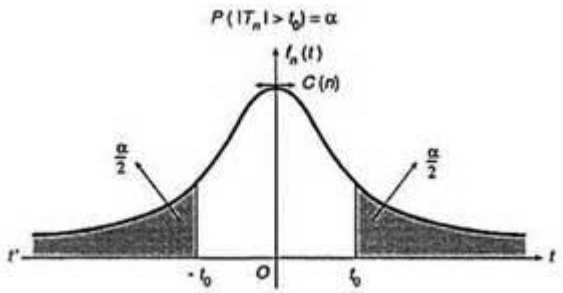
Si les deux variables aléatoires U et χ_n^2 sont indépendantes entre elles, alors T est une nouvelle variable aléatoire telle que $T = \dfrac{U}{\sqrt{\dfrac{\chi_n^2}{n}}}$,

T suit une loi de student à n degré de liberté, on note $T \rightsquigarrow t_{n\ ddl}$

L'espérance mathématique d'une variable aléatoire qui suit la loi de student est égale à :

E(T) = 0 si $n > 1$, et sa variance est égale à : $V(T) = \dfrac{n}{n-2}$ si $n > 2$

Les valeurs (absolues) de *t* ayant la probabilité α d'être dépassées sont présentées comme suit :

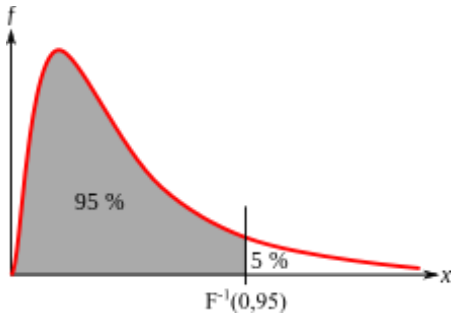


c) ***Loi de Fisher-Snédécor***

Soient χ_n^2 et χ_m^2 deux variables aléatoires indépendantes de Khi-deux à *n* et *m* ddl respectivement. Le rapport $\dfrac{\chi_n^2/n}{\chi_m^2/m}$ suit une loi de Fisher-Snédécor à *n et m* degrés de liberté notée *F(n,m)*. On a :

$\dfrac{\chi_n^2/n}{\chi_m^2/m} \rightsquigarrow$ *F(n,m) ddl*

Les valeurs de la variable de Fisher-Snédécor ayant la probabilité 0,05 d'être dépassées sont présentées comme suit :



III. Le comportement asymptotique

Nous définissons essentiellement trois convergences stochastiques, parmi les nombreuses existantes, la convergence en probabilité, la convergence en loi et la convergence presque sûre.

1. *Convergence en probabilité*

La définition de la convergence en probabilité emploie une suite numérique de probabilités dont la convergence sera souvent établit grâce à l'inégalité de Bienaymé-Tchebychev qui lie une probabilité et une variance.

2. *Inégalité de Bienaymé-Tchebychev*

Soit X une variable aléatoire positive dont l'espérance mathématique et la variance existent, l'inégalité de Bienaymé-Tchebychev est établit pour tout $\varepsilon > 0$:

$$P\left(|X - E(X)| \geq \varepsilon\right) \leq \frac{V(X)}{\varepsilon^2}$$

3. *Définition de la convergence en probabilité*

On dit que la suite de variables aléatoires X_n converge en probabilité vers une variable aléatoire X si :

$$\forall \varepsilon > 0, P\{\langle \ |X_n - X| > \varepsilon\rangle\} \to 0 \quad \text{quand n} \to +\infty$$

On écrit : $X_n \xrightarrow[P]{} X$

4. *Loi faible des grands nombres*

Soient n variables aléatoires $X_1, X_2, \ldots X_n$, 2 à 2 indépendantes de même loi (identiquement distribuées) dont :

E(X) existe. Alors : $\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow[P]{} E(X)$ quand $n \to +\infty$

Théorème

Si (X_n) est une suite de variables aléatoires mutuellement indépendantes qui admettent les mêmes moments d'ordres un et deux, c'est-à-dire avec pour tout entier n, $E(X_n) = m$ et $V(X_n) = \sigma^2$, alors $\bar{X}_n \xrightarrow[P]{} m$ quand $n \to +\infty$

5. Convergence en loi

a) ***Définition de la convergence en loi***

On dit que la suite de variables aléatoires(X_n) de fonctions de répartitions (F_n) converge en loi vers une variable aléatoire X de fonction de répartition F si la suite $\{F_n(X)\}$ converge vers F(x) en tout point x où F est continue ; on écrit alors : $X_n \xrightarrow[loi]{} X$

b) ***Lien entre la convergence en loi et la convergence en probabilité***

Théorème : La convergence en probabilité d'une suite de variables aléatoires (X_n) implique sa convergence en loi : $X_n \xrightarrow[P]{} X \Rightarrow X_n \xrightarrow[loi]{} X$

c) ***Théorème central-limite***

Soient X_1, X_2, ……X_n , n variables aléatoires indépendantes mutuellement indépendantes, de même loi, admettant deux premiers moments d'ordre 1 et 2 : $E(X_n) = m$ et $V(X_n) = \sigma^2$. Lorsque n tend vers l'infini, alors :

$$\sqrt{n}\frac{\bar{X}_n - m}{\sigma} \to N(0,1)$$

avec $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$

6. Convergence presque sûre

On dit que la suite (X_n) converge vers la variable aléatoire X presque sûrement si :

$P\{\omega \in \Omega / \lim_n X_n(\omega) = X(\omega)\} = 1$ et on a donc : $X_n \xrightarrow[P.S]{} X$, $n \to \infty$

Lorsque : $X_n \xrightarrow[P.S]{} X$ on a les équivalences suivantes :

$$X_n \xrightarrow[P.S]{} X \Leftrightarrow \forall \varepsilon > 0 , \lim_n P\{Sup_{k\geq n}|X_k - X| > \varepsilon\} = 0$$

$$\Leftrightarrow \forall \varepsilon > 0 , \lim_n P\{\bigcup_{k\geq n}(|X_k - X| > \varepsilon)\} = 0$$

$$\Leftrightarrow \forall \varepsilon > 0 , \lim_n P\{\bigcap_{k\geq n}(|X_k - X| < \varepsilon)\} = 1$$

La première condition implique $Sup_{k\geq n}|X_k - X| \xrightarrow[P]{} 0$ quand n → ∞ et implique aussi $|X_n - X| \xrightarrow[P]{} 0$, cela signifie que la convergence presque sûre est plus forte que la convergence en probabilité, donc on peut conclure que :

$X_n \xrightarrow[P.S]{} X \Rightarrow X_n \xrightarrow[P]{} X \Rightarrow X_n \xrightarrow[loi]{} X$

Théorème : Loi forte des grands nombres

Soit (X_n) une suite de variables aléatoires indépendantes et de même loi, admettant une espérance notée m , alors : $\bar{X}_n \xrightarrow[P.S]{} m$, n → ∞

7. Convergence de quelques lois usuelles

a) ***Convergence de la loi binomiale vers la loi de Poisson***

Soit un évènement aléatoire A de probabilité p (avec p < 0,1) que l'on essaie d'obtenir en répétant n fois l'expérience aléatoire (avec n est grand). Le nombre de réalisations de A suit une loi binomiale de paramètres *n* et *p* ; notée B(*n*, *p*).

Sous ces conditions ; la loi de Poisson est considérée comme une approximation de la loi binomiale : B(*n*, *p*) ~ P(*np*)

Théorème :

Soit (X_n) une suite de variables aléatoires binomiales $B(n, p)$ telle que n→ ∞ et p → 0 de manière à ce que le produit np tend vers une limite finie λ , alors la suite de variables aléatoires (X_n) converge vers une variable de Poisson $P(\lambda)$

Démonstration :

$$C_n^k p^k (1-p)^{n-k} = \frac{n(n-1)\dots(n-k+1)}{k!} p^k (1-p)^{n-k}$$

$$= \frac{(pn)^k}{k!}\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\dots\left(1-\frac{k-1}{n}\right) \quad (1-p)^{n-k}$$

Quand $n \to \infty$, les termes $\left(\mathbf{1}-\frac{\mathbf{1}}{\boldsymbol{n}}\right)\left(\mathbf{1}-\frac{\mathbf{2}}{\boldsymbol{n}}\right) \ldots \left(\mathbf{1}-\frac{\boldsymbol{k}-\mathbf{1}}{\boldsymbol{n}}\right) \to 1$, leur produit tend vers 1 car ils sont en nombre fini.

Décomposons $(\mathbf{1}-\boldsymbol{p})^{\boldsymbol{n}-\boldsymbol{k}}$ en $(\mathbf{1}-\boldsymbol{p})^{\boldsymbol{n}}(\mathbf{1}-\boldsymbol{p})^{-\boldsymbol{k}}$, $(\mathbf{1}-\boldsymbol{p})^{-\boldsymbol{k}} \to 1$ car p → 0

En mettant $\lambda = np$, $(\mathbf{1}-\boldsymbol{p})^{\boldsymbol{n}} \sim (1-\frac{\lambda}{\boldsymbol{n}})^{n}$ et $(1-\frac{\lambda}{\boldsymbol{n}})^{n} \to \boldsymbol{e}^{-\lambda}$

On a donc :

$$C_{\boldsymbol{n}}^{\boldsymbol{k}}\, \boldsymbol{p}^{\boldsymbol{k}}\, (\mathbf{1}-\boldsymbol{p})^{\boldsymbol{n}-\boldsymbol{k}} \to \frac{\boldsymbol{\lambda}^{\boldsymbol{k}}}{\boldsymbol{k}!}\, \boldsymbol{e}^{-\lambda}$$

b) ***Convergence de la loi binomiale vers la loi normale***

Si *p* reste fixe quand $n \to \infty$, on utilise l'approximation normale en écrivant la loi binomiale comme somme de variables aléatoires Xi indépendantes et de même loi de Bernoulli de paramètre p : $X = \sum_{i=1}^{n} X_i$. Comme E(Xi) = p et V(Xi) = pq on déduit du théorème centrale limite :

$$\sqrt{n}\frac{\bar{X}_n - p}{\sqrt{pq}} \underset{\boldsymbol{loi}}{\to} N(0,1)$$

On peut donc approximer la loi binomiale B(*n*, *p*) par la loi normale $N(np, \sqrt{npq})$. On considère que l'approximation est valable pour $n \geq 30, np \geq 5\ et\ nq \geq 5$.

c) ***Convergence de la loi de poisson vers la loi normale***

Si X suit une loi poisson dont le paramètre λ tend vers l'infini, alors :

$$\frac{X-\lambda}{\sqrt{\lambda}} \underset{loi}{\to} N(0,1)$$

d) ***Convergence de la loi gamma vers la loi normale***

Si X suit une loi gamma dont le paramètre r tend vers l'infini, alors :

$$\frac{X-r}{\sqrt{r}} \underset{\boldsymbol{loi}}{\to} N(0,1)$$

e) ***Convergence de la loi du khi deux vers la loi normale***

Si $\boldsymbol{\chi^2}$ suit une loi du Khi deux dont le nombre de degré de liberté v tend vers l'infini (généralement au-delà de 100), alors : $\frac{\chi 2 - v}{\sqrt{2v}} \underset{loi}{\to} N(0,1)$

Mais lorsque v est relativement élevé ($30 \leq v \leq 100$), la meilleure approximation est :

$$\sqrt{2\chi 2} - \sqrt{2v-1} \underset{loi}{\rightarrow} N(0,1)$$

f) ***Convergence de la loi de Student vers la loi normale***

Si T suit une loi de Student dont le nombre de degré de liberté n tend vers l'infini, alors :

$$T \underset{\boldsymbol{loi}}{\rightarrow} N(0,1)$$

Exercices - chapitre 01

Exercice 1 :

Soit U une variable aléatoire suit la loi normale centrée réduite, calculer les probabilités suivantes : $P(U < -2)$, $P(-1 < U < 0{,}5)$, $P(4U \geq -3)$.

1. Déterminer u_0 et v_0 telles que : $P(|U| < u_0) = 0{,}82$ $P(U < v_0) = 0{,}61$

Soit X une variable aléatoire suit la loi normale $X \sim N(-1, 2)$.

2. Calculer $P(X < -1)$, $P(X \geq 1)$, $P(-3 < X < 1)$,
3. Déterminer x_0 telle que : $P(X > x_0) = 0{,}4$

Exercice 2 :

On admet que la probabilité qu'un voyageur oublie ses bagages dans le train est 0,005.Un train transporte 850 voyageurs. On admettra que ces voyageurs se sont regroupés au hasard et que leur comportement, par rapport à leurs bagages, sont indépendants les uns des autres.

On désigne par X la variable aléatoire qui prend pour valeur le nombre de voyageurs ayant oublié leurs bagages dans le train.

1. Quelle est la loi de probabilité de X ? calculer son espérance mathématique et sa variance
2. Donner, en justifiant la réponse, une loi de probabilité permettant d'approcher la loi trouvée à la question précédente. En utilisant cette loi approchée, calculer une valeur approchée de la probabilité des événements suivants :

- Aucun voyageur n'a oublié ses bagages
- Cinq voyageurs au moins ont oublié leurs bagages

Exercice 3 :

Après observation de très nombreux relevés, on l'estime que la consommation électrique d'un abonné à Sonelgaz en heure de pointe est bien représentée par une variable aléatoire X qui suit une loi normale de paramètres m=5kW et écart-type 1,3kW. Un secteur comporte 100 abonnés dont les consommations en heure de pointe sont des variables aléatoires Xi supposées indépendantes et toutes de même loi N(5 ; 1,3).

On montre que la consommation totale Y du secteur en heure de pointe suit une loi normale.

1. Déterminer la moyenne et l'écart-type de Y.
2. Calculer la puissance minimale que Sonelgaz doit fournir au secteur en heure de pointe pour satisfaire à la demande avec une probabilité supérieure ou égale à 0,99.

Exercice 4 :

On envisage de construire dans une école primaire un abri dans lequel pourra s'abriter les élèves en cas d'intempéries. La taille des élèves est distribuée selon une loi normale de moyenne 130cm et d'écart-type 7cm.

A quelle hauteur minimale doit se trouver le toit de cet abri pour qu'au moins 95% des élèves puissent s'y tenir debout ?

Exercice 5:

Dans un groupe d'assurances on s'intéresse aux sinistres susceptibles de survenir, une année donnée, aux véhicules d'une importante entreprise de transport.

1. Soit X la variable aléatoire qui, à tout véhicule tiré au hasard dans un des parcs, associe le nombre de sinistres survenant pendant l'année considérée. On admet que X suit la loi de Poisson de paramètre $\lambda = 0,28$

a. Calculer la probabilité de l'évènement A : « un véhicule tiré au hasard dans le parc n'a aucun sinistre pendant l'année considérée »

B : « un véhicule tiré au hasard dans le parc a au plus deux sinistres pendant l'année considérée »

2. On note E l'évènement : « un conducteur tiré au hasard dans 'ensemble des conducteurs de l'entreprise n'a pas eu de sinistre pendant l'année considérée ».

On suppose que la probabilité de E est 0,6. On tire au hasard 15 conducteurs dans l'effectif des conducteurs de l'entreprise Cet effectif est assez important pour que l'on puisse assimiler ce prélèvement à un tirage avec remise de 15 conducteurs.

On considère que la variable aléatoire Y qui, à tout prélèvement de 15 conducteurs, associe le nombre de conducteurs n'ayant pas de sinistre pendant l'année considérée.

a. Justifier que Y suit la loi binomiale et déterminer ses paramètres

b. Calculer la probabilité que, dans un tel prélèvement, 10 conducteurs n'aient pas de sinistre pendant l'année considérée.

3. Dans ce qui suit, on s'intéresse au coût d'une certaine catégorie de sinistres survenus dans l'entreprise pendant l'année considérée. On considère la variable aléatoire C qui, à chaque sinistre tiré au hasard parmi les sinistres de cette catégorie, associe son coût en euros. On suppose que C suit la loi normale de moyenne 1200 et d'écart type 200.

Calculer la probabilité qu'un sinistre tiré au hasard parmi les sinistres de ce type coûte entre 1000 et 1500 euros.

Exercice 6 :

Une usine utilise une machine automatique pour remplir des flacons contenant un certain produit en poudre. Par suite de variations aléatoires dans le mécanisme, le poids de poudre par flacon est une variable aléatoire de loi normale de moyenne ***m*** et d'écart-type 1,1 mg.

Les flacons sont vendus comme contenant 100 mg de produit.

1) La machine est réglée sur m=101,2 mg. Quelle est la probabilité que le poids de produit dans un flacon soit inférieur au poids annoncé de 100 mg ? Quelle est la probabilité que le poids de produit dans un flacon soit compris entre 80 mg et 90 mg ?

2) Sur quelle valeur de ***m*** faut-il régler la machine pour qu'au plus 4% des flacons aient un poids inférieur au poids annoncé de 100 mg ?

Exercice 7 :

Une société française d'import-export spécialisée dans les tissus du luxe expédie d'une ville africaine 500 colis du tissu vers Marseille. Chacun de ces colis avait, emballage compris, un poids de 5,020 kg et écart-type de 0,300 kg.

A- Lors d'un contrôle douanier, les services portuaires décident (pour gagner du temps) de vérifier le poids de seulement 100 colis tirés au hasard sur l'ensemble.

Supposant qu'un colis ait un poids qui suit une loi normale dont les paramètres sont ceux précédemment cités, quelle est la probabilité que l'ensemble de ces 100 colis ait un poids total :

- Compris entre 496 kg et 500 Kg ?
- Supérieurs à 510kg ?
- Inférieure à 496 kg ?

B- Le personnel de l'aéroport a pour mission d'apposer 5 étiquettes auto-adhésives sur chaque colis. Mais, pour des raisons diverses et indépendantes les unes des autres, beaucoup se décollent lors de nombreuses manutentions.

A l'arrivée à Marseille, un examen minutieux de 172 colis tirés au hasard a donné les statistiques suivantes :

Nombre d'étiquettes présentes sur un colis	Nombre de colis
0	33
1	62
2	50
3	18
4	7
5	2

1- Au regard de ces informations, quelle est la probabilité qu'une étiquette reste collée ?

2- Quel type de modèle théorique probabiliste cela inspire-t-il pour ajuster le phénomène « étiquetage des colis ». Justifier la réponse ?

3- Donner la loi de probabilité de cette variable aléatoire ?

Exercice 8 :

1. Calculer le fractile d'ordre 0.95 de la loi khi deux de 40 ddl en utilisant les convergences.

2. Soit (X_n) une suite de variables aléatoires indépendantes et de même loi centrée et de variance σ^2 . Etudier la convergence en loi des variables aléatoires :

$$D_n = \frac{1}{n}\sum_{i=1}^{n}|X_i| \quad et \quad S_n^2 = \frac{1}{n}\sum_{i=1}^{n}X_i^2$$

3. Soit (X_n) une suite de variables aléatoires mutuellement indépendantes et de même loi de Gumbel de densité : $f(x) = e^{(x-e^x)}$

Déterminer la loi de probabilité de la variable aléatoire $Z_n = e^{X_n}$, puis étudier la convergence en loi de la suite (Y_n) avec : $Y_n = -\ln(\frac{1}{n}\sum_{i=1}^{n}Z_i)$

Chapitre 02 : Estimation statistique

L'estimation statistique consiste à attribuer des valeurs approchées aux paramètres d'une population (moyenne, variance, etc…) à l'aide d'un échantillon, vérifiant l'hypothèse d'échantillonnage aléatoire simple, de n observations issues de cette population.

Autrement dit, dans un échantillon de taille n, on suppose qu'une série statistique $x_1, x_2, \ldots\ldots x_n$ correspond à des réalisations de variables aléatoires $X_1, X_2, \ldots. X_n$; il s'agit de trouver une estimation d'un paramètre inconnu de la population totale en utilisant ces n réalisations.

I. Définitions et propriétés

1. *Exemples élémentaires*

Selon la loi des grands nombres $\bar{X}$ et S^2 sont les estimateurs de la moyenne théorique *m* et la variance théorique σ^2 respectivement et d'après le théorème de loi forte des grands nombres $\bar{\boldsymbol{X}}_{\boldsymbol{n}} \xrightarrow[\boldsymbol{P.S}]{} \boldsymbol{m}$

Et aussi d'après le théorème de loi forte des grands nombres :

$\frac{1}{n}(\sum_{i=1}^{n} X_i^2) \xrightarrow[P.S]{} E(X^2)$ *et* $\bar{X}_n^2 \xrightarrow[P.S]{} E^2(\mathrm{X})$

Donc :

$\frac{1}{n}(\sum_{i=1}^{n} X_i^2) - \bar{X}_n^2 \xrightarrow[P.S]{} E(X^2) - E^2(\mathrm{X})$

Alors : $S_n^2 \xrightarrow[\boldsymbol{P.S}]{} \sigma^2$

De même la fréquence empirique *f* d'un évènement est un estimateur de sa probabilité p.

Les variables aléatoires $\bar{X}_n$, S_n^2 et F_n sont appelées alors les estimateurs de m, σ^2 et p.

2. *Définition d'un estimateur*

Soit X une variable aléatoire dont la loi dépend d'un paramètre inconnu θ, élément d'un sous ensemble donné Θ de IR appelé espace des paramètres. On cherche à estimer θ à partir d'un échantillon $(X_1, X_2, \ldots. X_n)$ de variables aléatoires indépendantes de même loi de X, on notera ($x_1, x_2, \ldots\ldots x_n$)l'échantillon observé. Un estimateur T_n de θ sera une variable aléatoire T_n qui dépend

de X telle que $T_n = T_n(X_1, X_2, \ldots X_n)$ et chaque réalisation $T_n(x_1, x_2, \ldots x_n)$ est estimateur de θ.

Choisir une seule valeur pour estimer θ est un problème d'estimation ponctuelle, choisir un sous ensemble de Θ, dénommée région de confiance, est un problème d'estimation par intervalle dans IR. Ce type de problème sera résolu par l'application $T_n : E^n \to F$ qui associera une (ou plusieurs) variable(s) aléatoire(s) à valeur(s) numérique(s) $(F \subset IR \text{ ou } IR^k)$ à un n-échantillon $(X_1, X_2, \ldots X_n)$, application que nous nommerons une **statistique.** Il s'agit du modèle d'échantillonnage noté $(E, B, (P_\theta;\ \theta \in \Theta))$ *et* B est la tribu borélienne associée.

Définition : Un estimateur de θ est une application T_n de E^n dans F qui à un échantillon $(X_1, X_2, \ldots X_n)$ de la loi P_θ associe une variable aléatoire réelle (ou plusieurs dans le cas d'un paramètre multidimensionnel) dont on peut déterminer la loi de probabilité.

Exemple : L'estimateur classique de la moyenne théorique est la moyenne empirique :

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

Sous la condition d'indépendance, pour un échantillon avec remise :

$E(\bar{X}_n) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\, n\, E(X_i) = E(X_i)$

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

Pour un échantillon sans remise : $V(\bar{X}_n) = \frac{V(X)}{n}\left(1 - \frac{n-1}{N-1}\right)$

3. Propriétés d'un estimateur

Afin de choisir entre plusieurs estimateurs possibles d'un même paramètre, il faut satisfaire certaines propriétés.

a) ***Sans biais :***

L'estimateur T_n est une variable aléatoire. Lorsque l'on estime θ par T_n on peut commettre une erreur d'estimation qui est la différence entre T_n et θ. $T_n - \theta$ est donc une variable aléatoire que l'on peut décomposée de façon élémentaire en :

$T_n - E(T_n) + E(T_n) - \theta$ où $E(T_n)$ est l'espérance mathématique de T_n.

Le premier terme $T_n - E(T_n)$ représente les fluctuations aléatoires de T_n autour de sa moyenne et le second terme $E(T_n) - \theta$ représente une erreur asymétrique impliquée par le fait que T_n varie autour de sa moyenne et non autour de θ.

Par conséquent, la quantité $E(T_n) - \theta$ s'appelle **le biais.**

Il est donc souhaitable d'utiliser les estimateurs **sans biais.**

Définition 1 :

on dit qu'un estimateur est sans biais si l'espérance mathématique de cet estimateur est égale au paramètre estimé : $E(T_n) = \theta \quad \forall\, \theta \in \Theta$.

Définition 2 :

On dit qu'un estimateur est "asymptotiquement sans biais" si : $\forall\, \theta \in \Theta \quad E(T_n) \to \theta$ ***quand*** $n \to +\infty$

b) ***Convergence :***

La deuxième qualité d'un estimateur est d'être « convergent ». Il est souhaitable que si $n \to \infty$, T_n tend vers θ. C'est le cas de $\bar{X}_n$, S_n^2 *et* F_n**.** Les estimateurs ne convergent pas nécessairement à la même vitesse, ceci dépend, pour une taille d'échantillon donnée, de la notion de précision d'un estimateur.

On dit aussi qu'un estimateur T_n est convergent lorsque sa variance tend vers 0 quand n tend vers l'infini.

Théorème 1 : Tout estimateur sans biais dont la variance tend vers 0 est convergent :

$$E(T_n) = \theta \quad et \quad V(T_n) \to 0 \Rightarrow T_n \xrightarrow[P]{} \theta\,,\ n \to +\infty$$

Ce résultat se déduit directement d'Inégalité de Bienaymé-Tchebychev

$$P\left(|Tn - \theta| > \varepsilon\right) \leq \frac{V(Tn)}{\varepsilon^2} \quad \to 0\,,\ \text{pour tout } \varepsilon > 0,\ n \to +\infty$$

Théorème 2 : Tout estimateur asymptotiquement sans biais dont la variance tend vers 0 est convergent :

$$E(T_n) \to \theta \quad et \quad V(T_n) \to 0 \Rightarrow T_n \xrightarrow[P]{} \theta\,,\ n \to +\infty$$

Exemple : (X_n) est une suite de variables aléatoires normales N(m,σ), mutuellement indépendantes, de même loi :

1) $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$

$E(\bar{X}_n) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\, n\, E(X_i) = E(X_i) = m$

$\bar{X}_n$ est donc un estimateur sans biais de *m*.

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

Lorsque $n \to \infty, V(\bar{X}) \to 0, \bar{X}$ est un estimateur convergent de *m*.

2) $S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ est la variance empirique de X dans l'échantillon. Cette variance est un estimateur biaisé de σ^2 et le biais vaut $\frac{\sigma^2}{n}$.

Démonstration :

Pour calculer $E(S_n^2)$ décomposons S_n^2 :

On a : $X_i - m = X_i - \bar{X}_n + \bar{X}_n - m$

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

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

Comme $\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right) = \sum_{i=1}^{n} X_i - n\bar{X}_n = \sum_{i=1}^{n} X_i - \sum_{i=1}^{n} X_i = 0$

Alors :

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

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

D'où : $S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - m)^2 - \left(\bar{X}_n - m\right)^2$

Et : $E(S_n^2) = E\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - m)^2 - \left(\bar{X}_n - m\right)^2\right)$

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

$E(S_n^2) = \frac{1}{n}\sum_{i=1}^{n}\sigma^2 - \frac{\sigma^2}{n} = \sigma^2 - \frac{\sigma^2}{n}$

Le biais vaut $\frac{\sigma^2}{n}$

$E(S_n^2) \to \sigma^2$ quand $n \to +\infty$ alors S_n^2 est un estimateur asymptotiquement sans biais de σ^2

3) L'estimateur sans biais de σ^2 est $S_n^{*2} = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2 = \frac{n}{n-1} S_n^2$

Il est recommandé d'utiliser S_n^{*2} pour estimer σ^2.

c) ***Estimateur optimal***

✓ Précision d'un estimateur ou qualité d'un estimateur

La précision d'un estimateur T_n se mesure par l'erreur quadratique moyenne :

$EQ(T_n) = E[(T_n - \theta)^2] = V(T_n) + b_n^2(\theta)$ avec $b_n^2(\theta) = [E(T_n) - \theta]^2$

Démonstration :

$E[(T_n - \theta)^2] = E[\{T_n - E(T_n) + E(T_n) - \theta\}^2]$

$= E[(T_n - E(T_n))^2] + 2E[(T_n - E(T_n))(E(T_n) - \theta)] + E[(E(T_n) - \theta)^2]$

Comme : $E(T_n) - \theta$ est une constante $E[(E(T_n) - \theta)^2] = (E(T_n) - \theta)^2$

et que : $E[(T_n - E(T_n))] = 0$

alors : $E[(T_n - \theta)^2] = V(T_n) + (E(T_n) - \theta)^2 = EQ(T_n)$.

✓ Variance minimale

Parmi les estimateurs sans biais de θ, le plus précis est celui qui a une plus petite variance.

Soient deux estimateurs sans biais T_n et T'_n. T_n est meilleur que T'_n si $V(T_n) \leq V(T'_n)$

Lorsque on pourrait trouver un troisième estimateur T''_n ayant une variance plus petite que $V(T_n)$ il faut poursuivre la recherche mais on ne peut pas améliorer indéfiniment un estimateur !! Nous allons voir comment peut-on régler ce problème.

✓ Inégalité de Fréchet-Darmois-Cramer-Rao (FDCR)

Si X prend ses valeurs dans un ensemble qui ne dépend pas de θ, si la densité $f(x, \theta)$ est deux fois continûment dérivable par rapport à θ, et sous certaines conditions de régularité, tout estimateur T_n sans biais de θ dont la variance existe vérifié l'inégalité FDCR :

$\forall \theta \in$, $V(T_n) \geq \frac{1}{I_n(\theta)}$ où $I_n(\theta)$ est la quantité d'information de Fisher définie par :

$I_n(\theta) = E(\frac{\partial lnL}{\partial \theta})^2 = E(-\frac{\partial^2 lnL}{\partial \theta^2})$ (L : est la vraisemblance).

Définition :

On appelle vraisemblance (likelihood) de l'échantillon ($X_1, X_2, \ldots X_n$) la loi de probabilité de ce n-uple, notée L($x_1, x_2, \ldots x_n$), et définie par :

$\boldsymbol{L\ \ (x_1, x_2 \ldots\ldots x_n\ ;\ \theta) = \prod_{i=1}^{n} P(X_i = x_i \backslash \theta)}$ ***(pour X v.a discrète)***

$\boldsymbol{L\ \ (x_1, x_2 \ldots\ldots x_n\ ;\ \theta) = \prod_{i=1}^{n} f(x_i;\ \theta)}$ ***(pour X v.a continue)***

Les conditions de régularité sont :

- On suppose que l'ensemble des estimateurs Θ est un ensemble ouvert sur lequel la densité $f(x, \theta)$ ne s'annule en aucun point x et est dérivable par rapport à θ.
- On suppose aussi que l'on peut intervenir dérivation par rapport à θ et intégration, et que la quantité d'information de Fisher est strictement positive.

✓ Efficacité :

La borne inférieure pour la variance des estimateurs sans biais peut être atteinte ou non. Si cette borne est effectivement atteinte par un estimateur, il sera donc le meilleur, selon ce critère parmi l'ensemble des estimateurs sans biais. Cette optimalité est traduite par la définition suivante :

Définition :

Un estimateur sans biais T_n est efficace si sa variance est égale à la borne inférieure de FDCR : $V(T_n) = \frac{1}{I_n(\theta)}$

Exemple : Soit X suit la loi exponentielle de paramètre $1/\theta$ avec θ >0 de densité :

$$f(x, \theta) = \frac{1}{\theta} e^{-x/\theta} \quad , x>0$$

$E(X) = \theta$ *et* $V(X) = \theta^2$

$\bar{X}_n$ est un estimateur sans biais, convergent mais aussi efficace.

Démonstration de l'efficacité :

Soient X_1, X_2, …., X_n , n variables exponentielles indépendantes de idd.

$$L\ \ (x_1, x_2 \ldots\ldots x_n\ ;\ \theta) = \prod_{i=1}^{n} f(x_i;\ \theta) = \frac{1}{\theta^n} \exp(-\frac{1}{\theta} \sum_{i=1}^{n} x_i)$$

$$\ln\ \ L\ \ (x_1, x_2 \ldots\ldots x_n\ ;\ \theta) = -n \ln \theta\ \ - \frac{1}{\theta} \sum_{i=1}^{n} x_i$$

Nous dérivons par rapport au paramètre θ :

$$\frac{\partial \ln \quad L \ (x_1, x_2 \ldots\ldots x_n \ ; \ \theta)}{\partial \theta} = -\frac{n}{\theta} + \frac{1}{\theta^2} \sum_{i=1}^{n} x_i$$

On a :

$$\left\{\frac{\partial \ln \quad L \ (x_1, x_2 \ldots\ldots x_n \ ; \ \theta)}{\partial \theta}\right\}^2 = \frac{1}{\theta^2}\left\{n^2 - 2\frac{n}{\theta}\sum_{i=1}^{n} x_i + \frac{1}{\theta^2}\left(\sum_{i=1}^{n} x_i\right)^2\right\}$$

On pose : Y = $\sum_{i=1}^{n} X_i$

$$I_n(\theta) = \mathrm{E}\left\{\frac{\partial Ln\, L(x_1, x_2 \ldots\ldots x_n \ ; \theta)}{\partial \theta}\right\}^2 = \frac{1}{\theta^2}\left\{n^2 - 2\frac{n}{\theta}\, E(Y) + \frac{1}{\theta^2} E(Y^2) \right\}$$

Avec : $E(Y) = n\,E(X) = n\theta$ *et* $V(Y) = n\,V(X) = n\theta^2$

On déduit : $E(Y^2) = V(Y) + E^2(Y) = n\theta^2 + n^2\theta^2 = n(1+n)\,\theta^2$

On obtient :

$$I_n(\theta) = \frac{1}{\theta^2}\left\{n^2 - 2\frac{n}{\theta}\, n\theta + \frac{1}{\theta^2} n(1+n)\,\theta^2\right\} = \frac{n}{\theta^2}$$

$$I_n(\theta) = \frac{1}{V(\bar{X}_n)} \quad \text{puisque :} \quad V(\bar{X}_n) = \frac{V(X)}{n} = \frac{\theta^2}{n}$$

$\bar{X}_n$ est un estimateur efficace.

4. *Exhaustivité*

Soit un n-échantillon d'une variable aléatoire X. Soit T une statistique fonction de $X_1, X_2 \ldots\ldots X_n$ de loi g(t, θ) (densité dans le cas continu et P(T = t) dans le cas discret).

Définition :

T sera dite exhaustive si l'on a L(x ; θ) = g(t, θ)h(x) (principe de factorisation) en d'autres termes si la densité conditionnelle de l'échantillon est indépendante du paramètre.

Ceci signifie qu'une fois T est connu, aucune valeur de l'échantillon ni aucune autre statistique n'apportera d'informations supplémentaires sur le paramètre inconnu θ.

Théorème de Neyman et Fisher (théorème de factorisation) : Une statistique T_n est exhaustive s'il existe deux applications mesurables positives g et h telles que la densité L de l'échantillon puisse se factoriser sous la forme :

$$L\ (x_1, x_2 \ldots\ldots x_n\ ;\ \theta) = g(t\ ;\ \theta)\ h(x_1, x_2 \ldots\ldots x_n).$$

Exemple : Loi de poisson de paramètre λ inconnu

$$L_n(x_1, x_2 \ldots\ldots x_n\ ;\lambda\ ) = \prod_{i=1}^{n} P(X_i = x_i) = \prod_{i=1}^{n} \frac{\lambda^{xi}}{xi!}\ e^{-\lambda} = e^{-n\lambda}\ \frac{\lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!}$$

T = $\sum_{i=1}^{n} X_i$ est une statistique exhaustive car T suit une loi de poisson de paramètre $n\lambda$ car la densité $L_n(x_1, x_2 \ldots\ldots x_n\ ;\lambda\ )$ peut se factiriser de la façon suivante :

$$L_n(x_1, x_2 \ldots\ldots x_n\ ;\lambda\ ) = \mathrm{g(t\ ;}\ \lambda\ ).h(\mathrm{x})$$

Avec : $\mathrm{g(t\ ;}\ \lambda\ ) = e^{-n\lambda}\ \frac{(n\lambda)^t}{t!}$ et $h(x) = \frac{t!}{n^t \prod_{i=1}^{n} x_i!}$

II. Méthodes d'estimation statistique

Dans ce chapitre, on expose deux types d'estimations : l'estimation ponctuelle et l'estimation par intervalle de confiance.

1. Estimation ponctuelle

Quand il n'a y a pas d'estimateurs ponctuels évidents tels que la moyenne empirique et la variance empirique, on les construit par deux méthodes : méthode du maximum de vraisemblance et méthode des moments.

a) *Méthode d'estimation du maximum de vraisemblance (EMV)*

Supposons que $x_1, x_2 \ldots\ldots x_n$ sont les réalisations des variables aléatoires indépendantes $X_1, X_2 \ldots\ldots X_n$ de lois de probabilité inconnues mais identiques. Nous cherchons à estimer cette loi P inconnue à partir des observations $x_1, x_2 \ldots\ldots x_n$. La méthode d'estimation du maximum de

vraisemblance (EMV) est basée sur la vraisemblance, qui est la probabilité conjointe de la série $x_1, x_2 \ldots\ldots x_n$:

$$L_n(x_1, x_2 \ldots\ldots x_n\ ;\ \theta) = \prod_{i=1}^{n} P(X_i = x_i)$$

Une fois qu'on a obtenu la vraisemblance $L_n(\theta)$, on cherche à la maximiser. La maximisation de $L_n(\theta)$ est identique à la maximisation de son logarithme ($\ln L_n(\theta)$). L'estimateur qui maximise la vraisemblance c'est celui qui satisfait les conditions suivantes :

$$\begin{cases} \dfrac{\partial \ln L_n(x_1, x_2 \ldots\ldots x_n\ ;\ \theta)}{\partial \theta} = 0 \\ \dfrac{\partial^2 \ln L_n(x_1, x_2 \ldots\ldots x_n\ ;\ \theta)}{\partial^2 \theta} < 0 \end{cases}$$

On prend comme estimateur de θ la solution de l'équation $\dfrac{\partial \ln L_n(x_1, x_2 \ldots\ldots x_n\ ;\ \theta)}{\partial \theta} = 0$ et qui vérifie $\dfrac{\partial^2 \ln L_n(x_1, x_2 \ldots\ldots x_n\ ;\ \theta)}{\partial^2 \theta} < 0$.

Définition : On appelle estimateur du maximum de vraisemblance EMV toute fonction $\widehat{\boldsymbol{\theta}}_{\boldsymbol{n}}$ de (x_1, x_2, …., x_n) qui vérifie :

$$\boldsymbol{L}\left(x_1, x_2 \ldots\ldots x_n\ ;\ \widehat{\boldsymbol{\theta}}_{\boldsymbol{n}}\right) = \boldsymbol{max}_{\boldsymbol{\theta \epsilon\, \Theta}} \quad \boldsymbol{L}\ \left(x_1, x_2 \ldots\ldots x_n\ ,\boldsymbol{\theta}\right)$$

Propriété 1 : S'il existe une statistique exhaustive T, alors l'estimateur du maximum de vraisemblance en dépend.

Puisque pour une statistique exhaustive ; $L_n\left(x_1, x_2 \ldots\ldots x_n\ ;\theta\ \right) = g(t\ ;\theta\ ).h(x)$

Donc résoudre l'équation $\dfrac{\partial \ln L_n(x_1, x_2 \ldots\ldots x_n\ ;\ \theta)}{\partial \theta} = 0$ revient à résoudre $\dfrac{\partial \ln g_n(t_1, t_2 \ldots\ldots t_n\ ;\ \theta)}{\partial \theta} = 0$ donc l'estimateur de θ ; $\hat{\theta} = f(t)$.

Propriété 2 : Si $\hat{\theta}$ est un estimateur de maximum de vraisemblance de θ, $f(\hat{\theta})$ est l'estimateur de maximum de vraisemblance de $f(\theta)$.

Exemple : Estimation par la méthode EMV de la moyenne m de loi normale

Soient $X_1, X_2 \ldots\ldots X_n$, n variables aléatoires de lois normales et indépendantes. On suppose que la variance σ^2 est connue.

$X_i \rightsquigarrow N(m,\sigma)$

La fonction de densité de probabilité est :

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

La vraisemblance de la loi normale est :

$$L_n(x_1, x_2 \ldots\ldots x_n\ ;\ m) = \prod_{i=1}^{n} f(x_i) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}}\ e^{-\frac{1}{2\sigma^2}(x_i - m)^2}$$

$$L_n(x_1, x_2 \ldots\ldots x_n\ ; m) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - m)^2}$$

Une fois qu'on a obtenu la vraisemblance $L_n(m)$, on cherche à la maximiser. La maximisation de $L_n(m)$ est identique à la maximisation de son logarithme ($\ln L_n(m)$).

$$\begin{cases} \dfrac{\partial \ln L}{\partial m} = 0 \\ \dfrac{\partial^2 \ln L}{\partial^2 m} < 0 \end{cases}$$

$$L_n(x_1, x_2 \ldots\ldots x_n\ ;\ m) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n\ e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - m)^2}$$

$$\ln\ L_n(x_1, x_2 \ldots\ldots x_n\ ;\ m) = -\ n \ln(\sigma\sqrt{2\pi}) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - m)^2$$

$$\frac{\partial \ln L}{\partial m} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - m)$$

$$\frac{\partial \ln L}{\partial m} = 0 \Leftrightarrow \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - m) = 0 \Leftrightarrow \sum_{i=1}^{n}(x_i - m) = 0 \Leftrightarrow \sum_{i=1}^{n} x_i - n\,m = 0$$

La solution pour cette équation est : $\widehat{m} = \frac{1}{n}\sum_{i=1}^{n} x_i$

$\frac{\partial^2 \ln L}{\partial^2 m} = \frac{-n}{\sigma^2}$ donc $\frac{\partial^2 \ln L}{\partial^2 m} < 0$ car σ^2 et n sont toujours positifs.

Alors $\widehat{m} = \frac{1}{n}\sum_{i=1}^{n} x_i$ s'agit bien d'un maximum.

1) $\widehat{m}$ est-il sans biais ?

$E(\widehat{m}) = E(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n}\, n\, E(X_i) = E(X_i) = m$

$\widehat{m}$ est sans biais.

2) $\widehat{m}$ Est-il convergent ?

$\lim_{n\to\infty} V(\widehat{m}) = \lim_{n\to\infty} \frac{\sigma^2}{n} = 0$, alors $\widehat{m}$ est convergent.

3) $\widehat{m}$ Est-il efficace ?

$V(\widehat{m}) = V(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n^2}\, n\, V(X_i) = \frac{\sigma^2}{n}$

$I(\widehat{m}) = E(\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - m)\,)^2) = E\left\{\frac{1}{\sigma^4}[\ \sum_{i=1}^{n}(X_i - m)]^2\right\}$

On pose : $y = \sum_{i=1}^{n}(X_i - m)$

$I(\widehat{m}) = E\left\{\frac{1}{\sigma^4}[\ \sum_{i=1}^{n}(X_i - m)]^2\right\} = \frac{1}{\sigma^4}E(y^2) = \frac{1}{\sigma^4}(V(y) + E^2(y))$

$= \frac{1}{\sigma^4}[V(\sum_{i=1}^{n} X_i - nm) + E^2(\sum_{i=1}^{n} X_i - nm)$

$= \frac{1}{\sigma^4}[V(\sum_{i=1}^{n} X_i) + E^2(\sum_{i=1}^{n} X_i) - E^2(nm)]$ car $V(nm) = 0$

$= \frac{1}{\sigma^4}(n\sigma^2 + (nm)^2 - (nm)^2)$

$= \frac{n}{\sigma^2}$

$I(\widehat{m}) = \frac{n}{\sigma^2} = \frac{1}{V(\widehat{m})}$ alors $\widehat{m}$ est efficace.

b) ***Estimation par la méthode des moments***

La méthode des moments consiste à estimer les paramètres inconnus en utilisant les moments d'ordre 1 et 2 : E(X) et $E(X^2)$. Il s'agit de résoudre le système d'équations en égalant les moments théoriques aux moments empiriques en fonctions des paramètres inconnues. La solution des équations si elle existe et est unique, sera appelée estimateur obtenu par la méthode des moments.

Exemple : Soit la fonction de répartition suivante

$$f_{\theta(x)=}\begin{cases}\frac{1}{\theta}e^{\frac{-1}{\theta}(x-\gamma)} & si\ x>\gamma\ , \qquad \theta>0\\ \\ 0 \qquad si\ non\end{cases}$$

En posant Y = $X-\gamma$

On obtient :

$$f_{\theta(y)=}\begin{cases}\frac{1}{\theta}e^{\frac{-1}{\theta}y} & si\ y>0\ ,\theta>0\\ \\ 0 \qquad sinon\end{cases}$$

Y suit donc la loi exponentielle de paramètre $\frac{1}{\theta}$, donc :

$$\begin{cases}E(Y)=\ \theta\\ \\ V(Y)=\ \theta^2\end{cases}$$

Il s'agit de calculer E(X) et E(X^2).

On a :

$$\begin{cases}E(Y)=\ \theta\\ \\ V(Y)=\ \theta^2\end{cases}\Leftrightarrow\begin{cases}E(X-\gamma)=\ \theta\\ \\ V(X-\gamma)=\ \theta^2\end{cases}\Leftrightarrow\begin{cases}E(X)=\ \theta+\gamma\\ \\ V(X)=\quad \theta^2\end{cases}$$

$$\Leftrightarrow\begin{cases}E(X)=\quad \theta+\gamma\\ \\ E(X^2)-\ E^2(X\ )=\quad \theta^2\end{cases}\Leftrightarrow\begin{cases}E(X)=\quad \theta+\gamma\\ \\ E(X^2)=\ (\theta+\gamma)^2+\theta^2\end{cases}$$

En égalant les moments théoriques aux moments empiriques en fonctions des paramètres inconnues θ et γ :

$$\begin{cases}E(X)=\ \frac{1}{n}\ \sum_{i=1}^{n}x_i\\ \\ E(X^2)=\ \frac{1}{n}\ \sum_{i=1}^{n}x_i^2\end{cases}$$

On trouve :

$$\begin{cases}\hat{\theta}+\hat{\gamma}\ =\bar{x}\\ \\ (\hat{\theta}+\hat{\gamma})^2+\ \ \hat{\theta}^2\quad =\ \frac{1}{n}\ \sum_{i=1}^{n}x_i^2\end{cases}\Leftrightarrow\begin{cases}\hat{\theta}=\ \bar{x}\ -\hat{\gamma}\\ \\ \bar{x}^2+\ \ \hat{\theta}^2\quad =\ \frac{1}{n}\ \sum_{i=1}^{n}x_i^2\end{cases}$$

$$\Leftrightarrow \begin{cases} \hat{\theta} = \bar{x} - \hat{\gamma} \\ \hat{\theta}^2 = \frac{1}{n} \sum_{i=1}^{n} x_i^2 - \bar{x}^2 \end{cases} \Leftrightarrow \begin{cases} \hat{\gamma} = \bar{x} - \hat{\theta} \\ \hat{\theta}^2 = s^2 \end{cases} \Leftrightarrow \begin{cases} \hat{\gamma} = \bar{x} - s \\ \hat{\theta} = s \end{cases}$$

2. *Méthode d'estimation par intervalle de confiance*

Jusqu'à maintenant nous avons déterminé les estimateurs ponctuels c'est-à-dire l'estimateur EMV et l'estimateur par la méthode des moments.

Nous allons traiter dans cette section la méthode d'estimation par intervalle de confiance. Cette méthode est fréquemment utilisée et cela dans plusieurs domaines : la biologie, l'épidémiologie, la finance …

Nous présentons ici les intervalles de confiances les plus utilisés.

Définition : Soit X une variable aléatoire dont la loi dépend d'un paramètre réel θ inconnu et $\alpha \in [0,1]$ un nombre donné. On appelle « intervalle de confiance » pour le paramètre θ, de niveau de confiance 1- α, un intervalle qui a la probabilité 1- α de contenir la vraie valeur du paramètre θ .

a) *Intervalle de confiance pour la moyenne d'une loi normale d'écart type théorique connu*

Supposons qu'une variable aléatoire X suit la loi normale de moyenne ***m*** inconnue et de variance connue $\boldsymbol{\sigma^2}$: $X \rightsquigarrow N(m, \sigma^2)$

m est donc le paramètre inconnu que l'on cherche à estimer.

On cherche à estimer m et l'encadrer entre deux valeurs à un certain niveau de confiance $1-\alpha$ tel que : $P(a \leq m \leq b) = 1-\alpha$.

Soient $X_1, X_2, \ldots, X_n$, n variables aléatoires de lois normales et indépendantes.

La statistique utilisée pour construire cet intervalle de confiance et pour trouver les valeurs a et b est : $\frac{\bar{X}_n - E(\bar{X}_n)}{\sigma/\sqrt{n}} = \frac{\bar{X}_n - m}{\sigma/\sqrt{n}}$.

avec $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$

Son choix est justifié par le fait que $\bar{X}_n$ est l'estimateur ponctuel sans biais de la moyenne m.

Elle provient du théorème central limite et lorsque l'échantillon est grand (n $\geq$30) :

$$\frac{\bar{X}_n - m}{\sigma/\sqrt{n}} \rightsquigarrow \mathrm{N}(0,1)$$

On a donc : $P(-z\alpha_{/2} \leq \frac{\bar{X}_n - m}{\sigma/\sqrt{n}} \leq z\alpha_{/2}) = 1-\alpha$

$$\Leftrightarrow \quad P(\bar{X}_n - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq m \leq \bar{X}_n + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}) = 1-\alpha$$

Avec $z_{\frac{\alpha}{2}}$ est le fractile de la loi N(0,1) d'ordre 1- $\frac{\alpha}{2}$

On obtient finalement l'intervalle de confiance au niveau $1-\alpha$ de la moyenne *m* :

$IC_{1-\alpha}$ pour la moyenne m est : $[\bar{x}_n \mp z\alpha_{/2} \frac{\sigma}{\sqrt{n}}]$

Exemple : pour un niveau de confiance $1-\alpha$ = 0,95 , $z_{\frac{\alpha}{2}}$ = 1,96 qui est le fractile de la loi N(0,1) d'ordre 0.975.

b) ***Intervalle de confiance pour la moyenne d'une loi normale d'écart type théorique inconnu***

La statistique utilisée dans le cas précédent, et dont la loi était connue, était la variable normale centrée et réduite. Si maintenant σ est inconnue ; la loi $\bar{X}$ dépendant de σ , on utilise comme estimateur sans biais de σ^2 ; la variance empirique $\hat{\sigma}^2 = S_n^{*2} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$. La statistique $\frac{\bar{X}_n - m}{S_n^*/\sqrt{n}}$ suit une loi de Student à n-1 degré de liberté permettant de déterminer la valeur de ***t*** telle que ***:*** $\mathrm{P}(-t_{\frac{\alpha}{2}} \leq \frac{\bar{X}_n - m}{S_n^*/\sqrt{n}} \leq t_{\frac{\alpha}{2}}) = 1-\alpha$

Avec $t_{\frac{\alpha}{2}}$ est le fractile de la loi de student à n-1 ddl d'ordre1- $\frac{\alpha}{2}$

Ce qui conduit à l'intervalle de confiance bilatéral symétrique à $1-\alpha$ de la moyenne *m* :

$IC_{1-\alpha}$ pour la moyenne m est : $[\bar{x}_n \mp t_{\frac{\alpha}{2}}(\mathrm{n}-1) \frac{s_n^*}{\sqrt{n}}]$

Avec $s_n^{*2} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ est calculée à partir de l'échantillon de taille n tiré au hasard.

Exemple : pour une taille d'échantillon n= 26 et pour un niveau de confiance $1-\alpha$ = 0,95, $t_{\frac{\alpha}{2}}$ = 2,060 qui est le fractile de la loi student à 25 ddl d'ordre 0.975.

c) ***Intervalle de confiance pour la variance d'une loi normale d'espérance connue***

Lorsque la moyenne théorique *m* est connue, la statistique qui conduit à la construction de l'intervalle de confiance de la variance au niveau $1-\alpha$ est : $\frac{n}{\sigma^2} S_n^2 \rightsquigarrow \chi^2_{1-\alpha}$ (n)

avec $S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - m)^2$

On cherche :

$$P\left(\chi^2_{1-(\alpha/2)} \leq \frac{n}{\sigma^2} S_n^2 \leq \chi^2_{(\alpha/2)}\right) = 1-\alpha$$

$$\Leftrightarrow \quad P\left(\frac{\chi^2_{1-(\alpha/2)}}{nS_n^2} \leq \frac{1}{\sigma^2} \leq \frac{\chi^2_{(\alpha/2)}}{nS_n^2}\right) = 1-\alpha$$

$$\Leftrightarrow \quad P\left(\frac{nS_n^2}{\chi^2_{(\alpha/2)}} \leq \sigma^2 \leq \frac{nS_n^2}{\chi^2_{1-(\alpha/2)}}\right) = 1-\alpha$$

On obtient finalement l'intervalle de confiance à $1-\alpha$:

$IC_{1-\alpha}$ pour la variance σ^2 est : $\left[\frac{ns_n^2}{\chi^2_{\left(\frac{\alpha}{2}\right)}}, \frac{ns_n^2}{\chi^2_{1-\left(\frac{\alpha}{2}\right)}}\right]$

$\chi 2\left(\frac{\alpha}{2}\right)$ *et* $\chi 2\left(1-\frac{\alpha}{2}\right)$ étant les fractiles d'ordre respectifs $\frac{\alpha}{2}$ et $1-\frac{\alpha}{2}$ de la loi du khi-deux à ***n*** degré de liberté et $s_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - m)^2$ est calculée à partir de l'échantillon de taille n tiré au hasard.

Remarque : Soient $X_1, X_2, \ldots, X_n$, n variables aléatoires de lois normales et indépendantes :

$X_i \rightsquigarrow N(m,\sigma)$ et $Z_i = \frac{X_i - m}{\sigma} \rightsquigarrow N(0,1)$ et on définit $\sum_{i=1}^{n} Z_i^2 \rightsquigarrow \chi^2$ (n)

$$\sum_{i=1}\left(\frac{X_i - m}{\sigma}\right)^2 \rightsquigarrow \chi^2 \text{ (n)} \quad \Leftrightarrow \quad \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - m)^2 \rightsquigarrow \chi^2 \text{ (n)}$$

$$\Leftrightarrow \quad \frac{n}{\sigma^2}\frac{1}{n}\sum_{i=1}^{n}(X_i - m)^2 \rightsquigarrow \chi^2 \text{ (n)} \quad \Leftrightarrow \quad \frac{n}{\sigma^2} S_n^2 \rightsquigarrow \chi^2 \text{ (n)}$$

avec $S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - m)^2$

d) ***Intervalle de confiance pour la variance d'une loi normale d'espérance inconnue***

Lorsque la moyenne théorique *m* est inconnue, la statistique qui conduit à la construction de l'intervalle de confiance de la variance au niveau $1-\alpha$ est : $\frac{n-1}{\sigma^2} S_n^{*2} \rightsquigarrow \chi^2_{1-\alpha}$ (n-1)

avec $S_n^{*2} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$ estimateur sans biais de la variance lorsque m est inconnue.

On cherche :

$$P\left(\chi^2_{1-(\alpha/2)} \leq \frac{n-1}{\sigma^2} S_n^{*2} \leq \chi^2_{(\alpha/2)}\right) = 1-\alpha$$

$$\Leftrightarrow \quad P\left(\frac{\chi^2_{1-(\alpha/2)}}{(n-1)S_n^{*2}} \leq \frac{1}{\sigma^2} \leq \frac{\chi^2_{(\alpha/2)}}{(n-1)S_n^{*2}}\right) = 1-\alpha$$

$$\Leftrightarrow \quad P\left(\frac{(n-1)S_n^{*2}}{\chi^2_{(\alpha/2)}} \leq \sigma^2 \leq \frac{(n-1)S_n^{*2}}{\chi^2_{1-(\alpha/2)}}\right) = 1-\alpha$$

On obtient finalement l'intervalle de confiance à $1-\alpha$:

$IC_{1-\alpha}$ pour la variance σ^2 est : $\left[\frac{(n-1)S_n^{*2}}{\chi^2_{\left(\frac{\alpha}{2}\right)}}, \frac{(n-1)S_n^{*2}}{\chi^2_{1-\left(\frac{\alpha}{2}\right)}}\right]$

$\chi 2\left(\frac{\alpha}{2}\right)$ *et* $\chi 2\left(1-\frac{\alpha}{2}\right)$ étant les fractiles d'ordre respectifs $\frac{\alpha}{2}$ et $1-\frac{\alpha}{2}$ de la loi du khi-deux à ***n*** -1 degré de liberté et $s_n^{*2} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2$ est calculée à partir de l'échantillon de taille n tiré au hasard.

e) ***Intervalle de confiance pour une proportion p inconnue***

Pour construire l'intervalle de confiance d'une proportion *p* (inconnue) des individus possédant un certain caractère appartenant à une population infinie (ou finie si le tirage s'effectue avec remise), on utilise f_n ; proportion calculée dans un échantillon de taille n. Si n est faible on utilise les tables de la loi binomiale puisque : n $F_n \rightsquigarrow$ B(n, *p*) , ou on utilise l'abaque.

Si n est suffisamment grand généralement n>30 : $F_n \rightsquigarrow N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$

où $F_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ avec $X_i \rightsquigarrow B(p)$

et $Xi = \begin{cases} 1 \ si\ succés\ \ avec\ p \\ 0 \ si\ échec\ avec\ 1-p \end{cases}$

$X_i \rightsquigarrow B(p) \quad \forall\, i = \overline{1,n}$ et sont indépendantes.

$E(F_n) = E(\frac{1}{n}\sum_{i=1}^{n} X_i) = E(X_i) = p$

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

Si on ne dispose pas d'abaque permettant d'obtenir sans calculs l'intervalle de confiance exact de la proportion p, et si la taille de l'échantillon n est suffisamment grande, on obtiendra un intervalle approché (i.e de niveau voisin de $1-\alpha$) en utilisant la loi asymptotique de F_n *(qui l'estimateur de p)* déduite du théorème central limite : $\frac{F_n - P}{\sqrt{\frac{p\,(1-p)}{n}}} \rightsquigarrow N(0,1)$.

On retient alors un intervalle symétrique, à partie de la valeur $z_{\frac{\alpha}{2}}$ lue dans la table de la loi normale centrée réduite, telle que :

$$P\left(-z\alpha_{/2} \leq \frac{F_n - P}{\sqrt{\frac{p\,(1-p)}{n}}} \leq z\alpha_{/2}\right) = 1-\alpha$$

$$\Leftrightarrow \quad P\left(F_n - z_{\frac{\alpha}{2}}\sqrt{\frac{p\,(1-p)}{n}} \leq p \leq F_n - z_{\frac{\alpha}{2}}\sqrt{\frac{p\,(1-p)}{n}}\right) = 1-\alpha$$

Pour en déduire un intervalle pour p, il faudra résoudre ces deux inéquations du second degré en p. Il est plus simple de faire une autre approximation en remplaçant p par son estimateur dans les bornes de l'intervalle approché. A partir de l'échantillon, on a :

$IC_{1-\alpha}$ pour la proportion p est : $\left[f_n \mp z_{\frac{\alpha}{2}}\sqrt{\frac{f_n(1-f_n)}{n}}\right]$

avec $z_{\frac{\alpha}{2}}$ est le fractile de la loi $N(0,1)$ d'ordre $1-\frac{\alpha}{2}$.

On considère que cette approximation est acceptable pour $npq \geq 3$.

Remarque importantes : Ces formules sont applicables pour X suivant une normale et n est suffisamment grand.

Exercices – Chapitre 02

Exercice 1 : Soient X1, . . . , X_n n variables aléatoires i.i.d. B(p).

Déterminer l'estimateur du maximum de vraisemblance pour le paramètre p ? Est-il biaisé ?

Soit maintenant X1, . . . , X_n un ´échantillon aléatoire simple issu d'une population de densité :

$f(x, \theta) = \frac{\theta}{1-\theta} \; x^{\frac{2\theta-1}{1-\theta}}$ avec ½ < θ<1 et $0 < x < 1$

Déterminer l'estimateur du maximum de vraisemblance pour le paramètre θ ?

Déterminer l'estimateur par la méthode des moments pour le paramètre θ ?

Exercice 2 : Soit X une variable aléatoire de densité :

$f(x, \theta) = \frac{1}{2\theta\sqrt{x}} \; e^{\frac{-\sqrt{x}}{\theta}}$ avec x>0 et θ >0

Déterminer l'estimateur du maximum de vraisemblance pour le paramètre θ et étudier ses propriétés ?

Construire un intervalle de confiance de niveau 0.90 pour θ dans le cas où on a :

$\sum_{i=1}^{20} \sqrt{x_i}$ = 47,7

Exercice 3 : Donnez un estimateur de β suivant la méthode des moments pour :

$f(y; \beta) = (\beta + 1) y^{-\beta-2}$ *avec* $0 < y < 1$

Exercice 4 :

La force de compression d'un type de béton est modélisée par une variable gaussienne d'espérance μ et de variance σ^2. L'unité de mesure est le *psi* (pound per square inch). Dans les questions de 1. à 4, on supposera la variance□□ □□ est connue et égale à 1000. Sur un échantillon de 12 mesures, on a observé une moyenne empirique de 3250 psi.

1. Donner un intervalle de confiance de niveau 0.95 pour μ.

2. Donner un intervalle de confiance de niveau 0.99 pour μ. Comparer sa largeur avec celle de l'intervalle précédent.

3. Si avec le même échantillon on donnait un intervalle de confiance de largeur 30 psi, quel serait son niveau de confiance ?

4. On souhaite maintenant estimer μ avec une précision de ± 15 psi, avec un niveau de confiance de 0.95. Quelle taille minimum doit avoir l'échantillon ?

5. La variance théorique est désormais supposée inconnue. On dispose de la donnée suivante (sur le même échantillon de taille 12) :

$$\sum_{i=1}^{12} x_i^2 = 126761700 .$$

Donnez pour μ un intervalle de confiance de niveau 0.95 et comparez-le avec celui de la question 1, puis un intervalle de confiance de niveau 0.99 et comparez-le avec celui de la question 2.

6. Donner un intervalle de confiance de niveau 0.95 pour la variance, et pour l'écart type.

Exercice 5 :

On se propose d'estimer la précision d'un thermomètre dit 'ultra-sensible'. Pour cela, on réalise 15 mesures X_i (i = 1, …15) de la température d'un liquide maintenu pendant le temps des mesures à une température rigoureusement constante égale à 20°C.

On admet que les erreurs d'observations $\varepsilon_i = X_i - 20$ suivent indépendamment une loi $N(0, \sigma)$

Construire un intervalle de confiance pour σ^2 de niveau 0,99, sachant que :

$$\frac{1}{15}\sum_{i=1}^{15}(x_i - 20)^2 = 18$$

Exercice 6 :

On veut déterminer le poids P d'un objet à l'aide d'une balance à deux plateaux. Le poids marqué à l'équilibre est une variable aléatoire X compte tenu de l'imprécision et de déréglage possible de la balance. On supposera que $P+\varepsilon$, où ε de loi $N(\mu, \sigma)$ représente la précision de la balance.

On réalise 20 pesées $(X_1, X_2, \ldots, X_{20})$ du même objet et on obtient

$$\sum_{i=1}^{20}(x_i - \overline{x_{20}})^2 = 250$$

On admet que la balance est bien réglée et donc $\mu = 0$. Donner un intervalle de confiance de niveau 0,95 pour σ^2 .

Exercice 7 : Soit $(X_1, \ldots, X_n)$ un échantillon d'une loi de Poisson de paramètre λ. Déterminer l'estimateur du maximum de vraisemblance $\mathbf{T_n}$ du paramètre λ et étudier ses propriétés.

Exercice 8 : Soit $(X_1, \ldots, X_n)$ un échantillon d'une loi normale de paramètres moyenne **m** et d'écart-type $\boldsymbol{\sigma}$.

1. Déterminer l'estimateur du maximum de vraisemblance $\mathbf{T_n}$ du paramètre m et étudier ses propriétés.
2. Le fait que $\boldsymbol{\sigma}$ est connu ou non modifie-t-il le résultat ?
3. Supposant maintenant que la moyenne **m** est connue. Déterminer l'estimateur du maximum de vraisemblance $\mathbf{T'_n}$ du paramètre $\boldsymbol{\sigma^2}$ et déduire ses propriétés.
4. En déduire un estimateur du paramètre $\boldsymbol{\sigma}$. Cet estimateur est-il sans biais ? calculer la borne de FDCR.
5. Dans le cas où **m** est inconnue, proposer un estimateur sans biais de $\mathbf{S^2_n}$ de $\boldsymbol{\sigma^2}$ Calculer sa variance.

Exercice 9 : Le tableau suivant représente, pour un atelier de fabrication de pièces de fonderie d'une certaine qualité, les résultats de l'inspection de 100 pièces par jour pendant 30 jours consécutifs du travail.

1 .Donner une estimation ponctuelle de la proportion de pièces défectueuses de cette fabrication.

2. On se propose de tirer à nouveau un échantillon de 100 pièces parmi la fabrication de cette période. Sur la base d'un coefficient de confiance de niveau 0,9973, construire un intervalle à l'intérieur duquel devrait se situer le nombre de pièces défectueuses constatées dans cet échantillon.

3. Confronter le résultat du tableau avec ceux des deux premières questions. Quelles conclusions en tirer ?

(*La lecture du cahier journalier de l'atelier révèle que le 6ème jour du mauvais sable a provoqué une très mauvaise coulée, tandis que deux ouvriers nouveaux ont été introduits dans l'équipe, les jours 9 et 25).*

jour	*Nombre de rebuts*	*Jour*	*Nombre de rebuts*
1	*6*	*16*	*14*
2	*11*	*17*	*13*
3	*20*	*18*	*5*
4	*22*	*19*	*7*
5	*9*	*20*	*9*
6	*40*	*21*	*12*
7	*12*	*22*	*4*
8	*10*	*23*	*23*
9	*31*	*24*	*27*
10	*30*	*25*	*31*
11	*33*	*26*	*33*
12	*39*	*27*	*16*
13	*25*	*28*	*14*
14	*18*	*29*	*11*
15	*17*	*30*	*7*

Exercice 10 :

Donnez un estimateur de θ suivant la méthode des moments pour :

f(x; θ)= $\frac{1}{\theta} e^{-\frac{1}{\theta}(x-\gamma)}$ avec x > γ et θ >0

Exercice 11 : Le coût d'un certain type de sinistre peut être considéré comme une v.a X qui suit une loi normale de paramètres m et **σ**

On observe dans une compagnie d'assurance, n dossiers de sinistres indépendants.

1. On suppose que l'écart-type est connu **σ = 15** euros.

a) Calculer l'intervalle bilatéral de confiance à 98% pour **m.**

Application numérique : pour 20 dossiers, la moyenne des coûts observée est de 120 euros. Dans quelle fourchette placez-vous **m ?**

b) Combien de dossiers doit-on examiner pour que la longueur de l'intervalle de confiance 98% soit inférieure ou égal à 10 euros ?

2. L'écart-type **σ** n'est en fait pas connu. Comment l'intervalle est-il modifié ?

Application numérique : pour 20 dossiers, la moyenne des coûts observée est de 120 euros. Et l'estimateur sans biais de $\sigma^2 = 15^2$

Exercice 12 :

Ayant besoin d'un grand nombre d'axes de 400 mm de diamètre (tolérance ± 3mm), un industriel s'adresse à un fabricant. Ce dernier décide, avec son ingénieur de fabrication d'utiliser pour cette commande une machine pour laquelle on sait que la distribution des axes usinés est normale, avec un écart-type égal à 1mm. L'ingénieur décide de contrôler la fabrication par échantillon de 9 pièces.

1. Ayant réglé la machine sur **m**=400mm, l'ingénieur se donne comme règle de n'intervenir sur le réglage que si un échantillon donne une moyenne extérieure à l'intervalle centré autour **m** et ayant 99% de chances de contenir cette moyenne. Déterminer les limites a et b de cet intervalle (a<b).

2. La machine se dérègle à l'insu de l'ingénieur et se met à usiner des pièces d'un diamètre moyen $\mathbf{m}'$ = 401,5mm, la dispersion de la fabrication n'étant pas modifiée. Quelle est la probabilité qu'en prélevant alors un échantillon, l'ingénieur intervienne sur le réglage de la machine

3. La machine ayant travaillé longtemps, l'ingénieur décide de vérifier la stabilité de la variance de la fabrication. Un échantillon de 9 pièces, prélevé aléatoirement à cet effet, a fourni : $\sum_i (xi - \bar{x})^2 = 12$, où xi est le diamètre d'une pièce fabriquée. En déduire une estimation $\mathbf{s^2}$ de la variance de fabrication au moment de prélèvement de cet échantillon.

Chapitre 03 : Les tests statistiques paramétriques

La théorie des tests est la seconde branche de la statistique mathématique après l'estimation statistique. Elle permet de confronter deux hypothèses, formulées a priori, relatives à une ou plusieurs populations.

Ces deux hypothèses font l'objet d'un test d'hypothèses qui permet de contrôler leur validité (accepter ou rejeter) à partir de l'étude d'un ou plusieurs échantillons aléatoires.

Les méthodes de l'inférence statistique permet de déterminer, avec une probabilité à priori fixée, si les différences constatées au niveau des échantillons peuvent être imputables au hasard ou si elles sont suffisamment importantes pour signifier que les échantillons proviennent de populations vraisemblablement différentes.

Les tests d'hypothèses se basent sur un certain nombre d'hypothèses concernant la nature de la population dont provient l'échantillon étudié (normalité de la variable, égalité des variances, … etc).

Il existe plusieurs types de tests et diffèrent en fonction de l'hypothèse de travail :

1. Les tests destinés à vérifier si un échantillon peut être considéré comme extrait d'une population donnée, vis-à-vis d'un paramètre comme la moyenne ou la fréquence observée, sont nommés « **tests de conformité** ».
2. Les tests destinés à comparer plusieurs populations à l'aide d'un nombre équivalent d'échantillons ; ce sont les tests d'égalité nommés **« test d'homogénéité ».**
3. On peut ajouter à cette catégorie **le test d'indépendance** qui cherche à tester l'indépendance entre deux caractères qualitatifs et **le test d'adéquation** qui cherche à vérifier si une variable aléatoire a pour une fonction de répartition F ou autre fonction de répartition.

On peut classer les tests selon leurs objectifs (ajustement, indépendance, de moyenne, de variance, etc.), ainsi qu'il est fait dans la suite ou selon leurs propriétés mathématiques : on parle ainsi de tests paramétriques ou non, de tests robustes, de tests libres…

On parle de **tests paramétriques** lorsque l'on stipule que les données sont issues d'une distribution paramétrée. Dans ce cas, les caractéristiques des données peuvent être résumées à l'aide de paramètres estimés sur l'échantillon, la procédure de test subséquente ne porte alors que sur ces paramètres. L'hypothèse de normalité sous-jacente des données est le plus souvent utilisée, la moyenne et la variance suffisent pour caractériser complètement la distribution.

Concernant les tests d'homogénéité par exemple, pour éprouver l'égalité des distributions, il suffira de comparer les moyennes et/ou les variances.

Les **tests non paramétriques** ne font aucune hypothèse sur la distribution sous-jacente des données. On les qualifie souvent de tests *distribution free*. L'étape préalable consistant à estimer les paramètres des distributions avant de procéder au test d'hypothèse proprement dit n'est plus nécessaire.

Lorsque les données sont quantitatives, les tests non paramétriques transforment les valeurs en rangs. L'appellation **tests de rangs** est souvent rencontrée. Lorsque les données sont qualitatives, seuls les tests non paramétriques sont utilisables.

La distinction paramétrique – non paramétrique est essentielle. Elle est systématiquement mise en avant dans la littérature. Les tests non paramétriques, en ne faisant aucune hypothèse sur les distributions des données, élargissent le champ d'application des procédures statistiques. En contrepartie, ils sont moins puissants lorsque ces hypothèses sont compatibles avec les données.

Revenant aux tests paramétriques, on distingue généralement entre hypothèses simples et hypothèses composites :

- Une hypothèse simple est du type H : $\theta = \theta_0$ où θ_0 est une valeur isolée du paramètre.
- Une hypothèse composite ou multiple est du type H : $\theta \in A$ où A est une partie de $\mathbb{R}$ non réduite à un élément et la plupart des hypothèses composites se ramènent aux cas : $\theta > \theta_0$ ou $\theta < \theta_0$ ou $\theta \neq \theta_0$.

En fait, on construira les régions critiques en utilisant la valeur θ_0 seule. Lorsque l'hypothèse alternative est composite, la puissance du test est variable et on parle de fonction puissance $1-\beta(\theta)$.

Les étapes à suivre pour tester une hypothèse sont les suivantes :

(1) définir l'**hypothèse nulle** (notée H_0) à contrôler,

(2) choisir un test statistique ou **une statistique** pour contrôler H_0,

(3) définir la distribution de la statistique sous l'hypothèse « H_0 est réalisée »,

(4) définir **le niveau de signification du test** ou région critique notée α,

(5) calculer, à partir des données fournies par l'échantillon, la valeur de la statistique

(6) prendre une décision concernant l'hypothèse posée et conclure.

I. Notions générales sur les tests statistiques

Un test est un mécanisme qui permet de choisir une hypothèse H_0 ou H_1 au vu des résultats d'un échantillon.

1. *Risque d'erreur de première espèce* α

L'erreur de première espèce α consiste à rejeter l'hypothèse nulle H_0 alors qu'elle est vraie, soit rejeter à tort H_0 . Le risque d'erreur α est donc la probabilité que la valeur calculée de la statistique *D* appartienne à la région critique si H_0 est vrai. Dans ce cas H_0 est rejetée et H_1 est considérée comme vraie.

On écrit : $\alpha = P($ rejeter H_0 / H_0 est vraie) = $P($ accepter H_1 / H_1 est fausse).

La valeur α doit être fixée *a priori* par l'expérimentateur et jamais en fonction des données. C'est un compromis entre le risque de conclure à tort et la faculté de conclure.

Toutes choses étant égales par ailleurs, la région critique diminue lorsque α décroît et donc on rejette moins fréquemment H_0.

2. *Risque d'erreur de deuxième espèce* β

L'erreur de seconde espèce β consiste à accepter l'hypothèse nulle H_0 alors qu'elle est fausse, soit accepter à tort H_0 . L'erreur de seconde espèce β est la probabilité que la valeur calculée de la statistique *D* n'appartienne pas à la région critique si H_1est vraie. Dans ce cas H_0 est acceptée et H_1 est considérée comme fausse.

On écrit : $\beta = P($ accepter H_0/ H_0 est fausse) = $P($ accepter H_0 / H_1 est vraie)=$P($ rejeter H_1 / H_1 est vraie).

Pour quantifier le risque β, il faut connaître la loi de probabilité de la statistique D sous l'hypothèse H_1 , ce qui est un peu difficile. C'est pour cette raison l'erreur utilisée plus fréquemment est celle de première espèce.

3. *La puissance et la robustesse d'un test (1 - β)*

Les tests ne sont pas faits pour « démontrer » H_0 mais pour « rejeter » H_0. L'aptitude d'un test à rejeter H_0 alors qu'elle est fausse constitue la puissance du test.

La puissance d'un test est : 1 - β = $P($ rejeter H_0 / H_0 est fausse) = P(accepter H_1/ H_1 est vraie).

La puissance d'un test est fonction de la nature de H_1, un test unilatéral est plus puissant qu'un test bilatéral.

La puissance d'un test augmente avec la taille de l'échantillon *n* étudié à valeur de α constant. La puissance d'un test diminue lorsque α diminue.

La décision aboutira à choisir H_0 ou H_1 . Il ya donc 4 cas possibles schématisés dans le tableau ci-dessous avec les probabilités correspondantes :

Réalité / Décision	H_0 est vraie	H_1 est vraie
H_0 est retenue	1- α	β
H_1 est retenue	α	1- β

4. Test bilatéral, test unilatéral, statistique du test et région critique

Le choix entre le test bilatéral et unilatéral dépend de la nature des données, du type d'hypothèse que l'on désire à contrôler, des affirmations que l'on peut admettre concernant la nature des populations étudiées (normalité, égalité des variances).

La statistique du test est une fonction des variables aléatoires représentant l'échantillon dont la valeur numérique obtenue pour l'échantillon considéré permet de distinguer entre H_0 vraie et H_0 fausse. Dans la mesure où la loi de probabilité suivie par le paramètre θ_0 au niveau de la population, en général, est connue, on peut ainsi établir la loi de probabilité de la statistique *D* telle que : $D = \theta - \theta_0$ avec $H_0 : \theta = \theta_0$

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

$P(W|H_0) = \alpha$

La région d'acceptation est son complémentaire $\overline{W}$ et l'on a donc :

$P(\overline{W}|H_0) = 1-\alpha$ et $P(W|H_1) = 1-\beta$

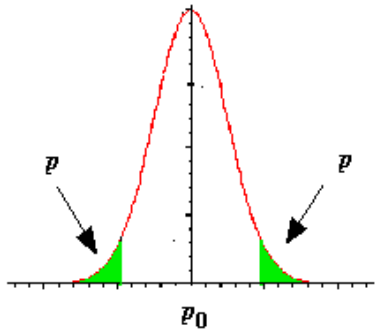
La région critique dépend aussi de la nature de test à réaliser. La nature de H_0 détermine la façon de formuler H_1 et par conséquence, et selon la nature unilatérale ou bilatérale du test, la définition de la région critique varie.

a) ***Test bilatéral***

Si H_0 consiste à supposer que la population féminine avec une proportion de chômage « p » est représentative de la population totale avec une proportion de chômeurs « p_0 », on pose le test suivant :

$H_0: p = p_0$ et $H_1 : p \neq p_0$

Le test sera bilatéral car on considère que la fréquence p peut être supérieure ou inférieure à la fréquence p_0 . La région critique α correspond à une probabilité 2 fois α/2 (avec α/2 = p *sur le graphique)*

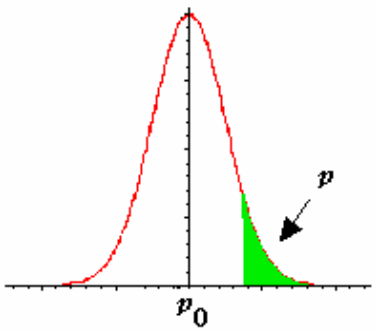


b) ***Test unilatéral***

Si l'on fait l'hypothèse que la proportion de chômage dans la population féminine p est supérieure à la proportion de chômeurs dans la population p_0, on pose alors : $H_0 : p = p0$

et $H_1 : p > p0$

Le test sera unilatéral car on considère que la fréquence p ne peut être que supérieure à la fréquence p_0. La région critique α correspond à une probabilité α (avec α = p sur le graphique).



Le raisonnement inverse peut être formulé avec l'hypothèse suivante :

$H_0 : p = p_0$ et $H_1 : p < p_0$

Puisque on peut trouver la loi de probabilité suivie par la statistique D sous l'hypothèse nulle H_0, on peut déterminer une valeur seuil (théorique) ; D_{seuil} de la statistique pour une probabilité donnée appelée le niveau de signification du test : 1 - α

La région critique correspond à l'ensemble des valeurs telles que $D >$ Dseuil.

Le niveau de signification est telle que : $P(D > \text{Dseuil}) = \alpha$ avec $P(D \leq \text{Dseuil}) = 1 - \alpha$

c) ***Règles de décision***

Une fois les hypothèses du test sont posées (test unilatéral ou test bilatéral), nous devons choisir la statistique pour le réaliser. C'est en comparant la valeur de cette statistique observée dans l'échantillon à sa valeur sous l'hypothèse H_0 que nous pourrons prendre une décision c'est-à-dire donner la conclusion du test.

Règle de décision 1 **:** Sous l'hypothèse H_0 et pour un seuil de signification 1-α fixé :

✓ si la valeur de la statistique D calculée ou observée est supérieure à la valeur seuil de *D* théorique : $D_{\text{observée}} > D_{\text{seuil}}$, alors l'hypothèse H_0 est rejetée au risque d'erreur α et l'hypothèse H_1 est acceptée au risque d'erreur α.

✓ si la valeur de la statistique D calculée est inférieure à la valeur seuil D théorique : $D_{\text{observée}} \leq D_{\text{seuil}}$, alors l'hypothèse H_0 ne peut être rejetée au risque d'erreur α.

Règles de décision 2 **:** La probabilité critique α telle que $P(D \geq D_{\text{observée}}) = \alpha_{\text{observé}}$ est évaluée. Si le risque d'erreur α est fixé à 0,05 :

✓ si $\alpha_{\text{observé}} \geq 0{,}05$ l'hypothèse H_0 ne peut être rejetée car le risque d'erreur de rejeter H_0 alors qu'elle est vraie est trop important.

✓ si $\alpha_{\text{observé}} < 0{,}05$ l'hypothèse H_0 est rejetée car le risque d'erreur de rejeter H_0 alors qu'elle est vraie est très faible.

5. Règle de décision de NEYMAN et PEARSON

Dans l'optique de Neyman et Pearson, on accroît la dissymétrie du problème de test en considérant que l'erreur la plus grave consiste à rejeter à tort l'hypothèse nulle. On fixe donc un seuil maximum α_0 au risque de première espèce et on cherche un test qui minimise le risque de seconde espèce.

Le test d'hypothèses est $H_0 : \theta = \theta_0$ et $H_1 : \theta = \theta_1$

La règle de décision proposée par Neyman et Pearson consiste à déterminer la région critique pour laquelle la puissance (la probabilité de rejeter l'hypothèse nulle avec raison) est maximum sous la contrainte $\alpha \leq \alpha_0$.

Théorème de Neyman et Pearson :

Pour un risque de première espèce α_0 fixé dans [0,1], le test de puissance maximum entre les hypothèses simples ci-dessus est défini par la région critique :

$$W = \left\{ (x_1, x_2, \dots x_n) : \frac{L_0(x_1, x_2, \dots x_n)}{L_1(x_1, x_2, \dots x_n)} \leq k \right\}$$

où la valeur de la constante k est déterminée par le risque α_0 tel que $\alpha_0 = P(W / \theta = \theta_0)$, ayant posé :

$L_0(x_1, x_2, \dots x_n) = L\ (x_1, x_2, \dots x_n, \theta_0)$ et $L_1(x_1, x_2, \dots x_n) = L\ (x_1, x_2, \dots x_n, \theta_1)$.

Exemple :

Une variable aléatoire X suit une loi normale d'écart-type connu $\sigma = 1$. Au vu d'un échantillon $(X_1, X_2, \dots X_n)$ de la loi de X, on veut choisir entre deux hypothèses :

$$\begin{cases} H_0 : m = 1 \\ H_1 : m = 1.5 \end{cases}$$

La vraisemblance sous H_0 est :

$$L(x_1, x_2, \dots x_n ; m) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \quad e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i - m)^2}$$

Le théorème de Neyman et Pearson fournit la forme de la région critique : $\frac{L_0}{L_1} \leq k$

En passant aux logarithmes, on a l'inégalité :

$$-\frac{1}{2}\left[\sum_{i=1}^{n}(x_i - 1)^2 - \sum_{i=1}^{n}(x_i - 1.5)^2\right] \leq Ln\ k \Leftrightarrow$$

$$-\sum_{i=1}^{n} x_i + 5n/4 \leq 2Ln\ k \Leftrightarrow \sum_{i=1}^{n} x_i \geq 5n/4 - 2Ln\ k$$

$$\Leftrightarrow \frac{1}{n}\sum_{i=1}^{n} x_i \geq 5/4 - \frac{2}{n} Ln\ k \Leftrightarrow \bar{x}_n \geq C$$

La région critique est : $W = \left\{ (x_1, x_2, \dots x_n) : \bar{x}_n \geq C \right\}$

Avec : $\alpha = P(\bar{X}_n \geq C \setminus m = 1)$, soit $\frac{C-1}{1/\sqrt{n}} = z_\alpha$ où z_α est le fractile d'ordre 1-α de la loi normale centrée et réduite N(0,1). On déduit $C = 1 + \frac{z\alpha}{\sqrt{n}}$

Pour $\alpha = 0{,}05$; $z_\alpha = 1{,}6449$ lue sur la table N(0,1) ; C = 1.33 pour n=25

La région critique est : $W = \left\{ (x_1, x_2, \dots x_{25}) : \bar{x}_{25} \geq 1.33 \right\}$

On calcul $\bar{x}_{25}$ et on la compare à 1,33.

✓ si la valeur de $\bar{x}_{25}$ est supérieure à la valeur seuil 1.33, alors l'hypothèse H_0 (m=1) est rejetée au risque d'erreur 5% et l'hypothèse H_1 (m=1.5) est acceptée au risque d'erreur 5%

✓ • si la valeur de $\bar{x}_{25}$ est inférieure à la valeur seuil 1.33, alors l'hypothèse H_0 (m=1) ne peut être rejetée au risque d'erreur 5%.

Lorsque maintenant le test est sous la forme multiple :

$$\begin{cases} H_0: \theta = \theta_0 \\ H_1: \theta < \theta_0 \end{cases} \quad \text{ou} \quad \begin{cases} H_0: \theta = \theta_0 \\ H_1: \theta > \theta_0 \end{cases}$$

On détermine la région critique par la méthode Neyman et Pearson du test :

$$\begin{cases} H_0: \theta = \theta_0 \\ H_1: \theta = \theta_1 \end{cases}$$

où $\theta_1$vest une valeur fixée quelconque mais vérifiant H_1. Si la région critique W obtenue pour ce test simple ne dépend pas de la valeur de θ_1 alors on dit qu'on a obtenu un test uniformément le plus puissant UPP pour le test initial. Cela veut dire que pour toute autre région critique W' on aura $P_\theta\,(W \setminus \theta \in \Theta_1) \geq P_\theta\,(W' \setminus \theta \in \Theta_1) \quad \forall\ \theta \in \Theta_1$.

Théorème de Lehmann

Il existe un test uniformément le plus puissant dont la région critique W est l'ensemble des points $(x_1, x_2, \ldots x_n)$ tels que : $T_n\,(x_1, x_2, \ldots x_n) > k$

Où la valeur de la constante k est déterminée par le risque fixé $\alpha_0 = P_\theta\,(W \setminus \theta = \theta_0)$

6. Méthode de Bayes

On effectue des probabilités a priori p_0 et $p_1=1-p_0$ attribuées à chacune des hypothèses respectives H_0 et H_1 et que on associe un coût à chaque décision, en fonction de l'hypothèse qui est effectivement réalisée. Le tableau ci-après contient ces coûts, la décision prise figurant en colonne et l'hypothèse vraie en ligne :

	Décision D0	Décision D1
H_0 (p_0)	C_{00}	C_{01}
H_1 (p_1)	C_{10}	C_{11}

Une bonne décision peut avoir également un coût et donc on aura généralement :

$C_{00} > 0$ et $C_{11} > 0$.

Après la réalisation $(x_1, x_2, \dots x_n)$ on peut calculer à l'aide du théorème de Bayes les probabilités a posteriori $\pi 0$ et $\pi 1$ telles que :

$$\pi 0 = \frac{p_0 L_0}{p_0 L_0 + p_1 L_1} \quad \text{et } \pi 1 = \frac{p_1 L_1}{p_0 L_0 + p_1 L_1}$$

Avec les vraisemblances : $L_0(x_1, x_2, \dots x_n ; \theta)$ est la valeur de la vraisemblance quand $\theta \in \Theta_0$ *et* $L_1(x_1, x_2, \dots x_n , \theta)$ est la valeur de la vraisemblance quand $\theta \in \Theta_1$

On calcule ensuite l'espérance du coût de chaque décision pour la distribution posteriori :

$E(C(D_0)) = C_{00} . \pi 0 + C_{10} . \pi 1$

$E(C(D_1)) = C_{01} . \pi 0 + C_{11} . \pi 1$

La règle de décision de Bayes est celle qui associe à la réalisation $(x_1, x_2, \dots x_n)$ la décision dont l'espérance du coût est la plus faible.

II. Tests paramétriques les plus courants

1. *Tests de conformité*

Le test de conformité permet de vérifier si un échantillon est représentatif ou non d'une population vis-à-vis d'un paramètre donné (la moyenne, la variance ou la fréquence…). Pour effectuer ce test, la loi théorique doit être connue au niveau de la population.

Quand il s'agit des petits échantillons (***n* < 30)**, la variable aléatoire *X* étudiée doit suivre une loi normale N(μ,σ). En revanche, quand la taille de l'échantillon est suffisamment grande ($n \geq 30$), la loi de la variable aléatoire X converge vers une loi normale et le test est peut-être appliqué en raison du théorème central-limite.

a) ***Comparaison d'une moyenne observée et une moyenne théorique***

Soit *X* une variable aléatoire observée dans une population, suivant une loi normale de paramètres μ et σ. Soit un échantillon tiré aléatoirement de cette population. Soit (X_n) une suite de variables aléatoires indépendantes et de même loi (loi normale).

L'objectif est de savoir si un échantillon de moyenne $\bar{X}$, estimateur sans biais de μ, appartient à une population de référence connue d'espérance μ_0 (H_0 vraie) et ne diffère de μ_0 que par des fluctuations d'échantillonnage, ou bien appartient à une autre population inconnue d'espérance μ (H_1 vraie). Les hypothèses sont exprimées de la façon suivantes :

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{cases}$$

La statistique du test varie selon si la variance σ^2 de la population de référence est connue ou non.

✓ **<u>Si la variance de la population est connue</u>**

Soit une variable aléatoire X suit la loi normale de moyenne μ inconnue et de variance connue σ^2 ; On la note $X \rightsquigarrow N(\mu, \sigma)$

Soit (X_n) une suite de variables aléatoires indépendantes et de même loi gaussienne.

On a : $\bar{X} \rightsquigarrow N(\mu, \sigma/\sqrt{n})$

La statistique du test est : $Z = \dfrac{\bar{X} - E(\bar{X})}{\sigma/\sqrt{n}}$

Sous H_0 : $Z = \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}$

Grâce au théorème central limite la variable Z centrée réduite suit la loi normale centrée réduite : $Z \rightsquigarrow N(0,1)$

Une valeur $z_{observée}$ de la variable aléatoire Z est calculée à partir de l'échantillon :

$z_{observée} = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$

Le test est bilatéral, $z_{observée}$ est comparée, en valeur absolue, à la valeur z_{seuil} lue sur la table de la loi normale centrée réduite pour un risque d'erreur α fixé. z_{seuil} est le fractile de la loi normale centrée réduite d'ordre 1- $\frac{\alpha}{2}$.

Par exemple, pour un risque de première espèce α= 5%, la valeur de z_{seuil} est le fractile de la loi normale centrée réduite d'ordre 0,975 est égale à 1,96.

• si $|z_{observée}| > z_{seuil}$ l 'hypothèse H_0 est rejetée au risque d'erreur α : l'échantillon appartient à une population d'espérance μ et n'est pas représentatif de la population de référence d'espérance μ_0 au risque d'erreur α.

• si $|z_{observée}| \leq z_{seuil}$ l'hypothèse H_0 est acceptée au risque d'erreur α : l'échantillon est représentatif de la population de référence d'espérance μ_0 au risque d'erreur α.

✓ **<u>Si la variance de la population inconnue</u>**

Soit une variable aléatoire X suit la loi normale de moyenne μ inconnue et de variance inconnue σ^2 : $X \rightsquigarrow N(\mu, \sigma)$.

Soit (X_n) une suite de variables aléatoires indépendantes et de même loi (loi normale).

$\bar{X}$ la moyenne dans la population suit une loi normale telle que : $\bar{X} \rightsquigarrow N(\mu, \sigma/\sqrt{n})$

La variance de la population n'étant pas connue, elle est estimée par : $\hat{\sigma}^2 = S_n^{*2} = \frac{n}{n-1} S_n^2$

avec : $S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ et $S_n^{*2} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$

La statistique du test est la variable aléatoire T = $\frac{\bar{X}-E(\bar{X})}{\frac{\hat{\sigma}}{\sqrt{n}}}$

Sous H_0 : $T = \dfrac{\bar{X}-\mu_0}{\frac{\hat{\sigma}}{\sqrt{n}}}$

Nous pouvons établir grâce au théorème central limite la variable T qui suit la loi de student à n-1 degré de liberté.

La valeur de T calculée sur l'échantillon est $t_{observée} = \dfrac{\bar{x}-\mu_0}{\hat{\sigma}\ /\sqrt{n}} = \dfrac{\bar{x}-\mu_0}{s_n\ /\sqrt{n-1}}$

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

$t_{observée}$ est comparée, en valeur absolue, à la valeur t_{seuil} lue sur la table de la loi de student à n-1 ddl pour un risque d'erreur α fixé.

t_{seuil} est le fractile de la loi student à n-1 ddl d'ordre 1-$\frac{\alpha}{2}$ car le test est bilatéral. Par exemple, pour un risque de première espèce α = 5%, et pour une taille d'échantillon égale à 33, la valeur de t_{seuil} est le fractile de la loi student à 32 ddl d'ordre de 0,975 est égale à 2,037.

Remarque : Lorsque n → $+\infty$, les fractiles de la loi de student sont égaux aux fractiles de la loi N(0,1) pour un même risque d'erreur α fixé.

• si $|t_{observée}|$ > t_{seuil} l'hypothèse H_0 est rejetée au risque d'erreur α : l'échantillon appartient à une population d'espérance μ et n'est pas représentatif de la population de référence d'espérance μ_0 au risque d'erreur α

• si $|t_{observée}|$ ≤ t_{seuil} l'hypothèse H_0 ne peut pas être rejetée au risque d'erreur α: l'échantillon est représentatif de la population de référence d'espérance μ_0 au risque d'erreur α.

b) ***Comparaison d'une fréquence observée et une fréquence théorique***

Soit X une variable qualitative prenant deux modalités (succès X=1, échec X=0) ; observée sur une population et un échantillon tiré au hasard de cette population. La probabilité p est la probabilité du succès P(X = 1) = p et la probabilité d'échec est P(X = 0) = 1-p .

Le but est de savoir si un échantillon (de grande taille) de fréquence observée $f=K/n$, estimateur de p, appartient à une population de référence connue de fréquence p_0 (H_0 vraie) ou à une autre population inconnue de fréquence p (H_1 vraie).

$$\begin{cases} H_0 : p = p_0 \\ H_1 : p \neq p_0 \end{cases}$$

f est la réalisation de la variable aléatoire F qui est la fréquence empirique (l'équivalence de $\bar{X}$ lorsque la v.a X est quantitative) et qui suit **approximativement** une loi normale :

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

On peut définir une variable aléatoire Z telle que :

$$Z = \frac{F - E(F)}{\sqrt{V(F)}} = \frac{F - p}{\sqrt{\frac{p(1-p)}{n}}}$$

Sous H_0 : $p = p_0$

$$Z = \frac{F - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

Grace au théorème central-limite, cette variable aléatoire Z suit approximativement la loi normale centrée réduite $Z \rightsquigarrow N(0,1)$ si seulement si n >30, np_0 et $nq_0 \geq 5$.

A partir de l'échantillon, on calcule $z_{observée} = \frac{f - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$

• si $|z_{observée}| > z_{seuil}$ l'hypothèse H_0 est rejetée au risque d'erreur α : l'échantillon appartient à une population d'espérance p et n'est pas représentatif de la population de référence d'espérance p_0 au risque d'erreur α

• si $|z_{observée}| \leq z_{seuil}$ l'hypothèse H_0 ne peut pas être rejetée au risque d'erreur α: l'échantillon est représentatif de la population de référence d'espérance p_0 au risque d'erreur α.

2. Tests d'égalité ou d'homogénéité des populations

Il s'agit de comparer deux populations à l'aide d'un paramètre donné tel que la moyenne, la variance et la fréquence…

a) ***Comparaison de deux variances de deux échantillons gaussiens indépendants***

Soit X une variable aléatoire observée sur deux populations indépendantes suivant une loi normale. On tire deux échantillons de ces deux différentes populations, indépendants l'un de l'autre, de tailles respectives $n_1$et n_2.

On veut savoir si les variances, dans ces deux populations, sont égales (ou les différences potentielles sont statistiquement non significatives au seuil α fixé) ou bien ces différences sont statistiquement significatives au seuil α fixé.

Le test à réaliser est :

$H_0 : \sigma_1^2 = \sigma_2^2$ contre $H_1 : \sigma_1^2 \neq \sigma_2^2$

La statistique utilisée pour réaliser ce test est la statistique de Fisher-Snédécor F.

Sous $H_0 : \sigma_1^2 = \sigma_2^2$

$F = \dfrac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} = \dfrac{\frac{n_1}{n_1-1}S_1^2}{\frac{n_2}{n_2-1}S_2^2}$ suit une loi de Fisher-Snedecor à $(n_1 - 1, n_2 - 1)$ degré de liberté

avec $S_1^2 = \frac{1}{n_1}\sum_{i=1}^{n_1}(X_{1i} - \bar{X}_1)^2$ et $S_2^2 = \frac{1}{n_2}\sum_{j=1}^{n_2}(X_{2j} - \bar{X}_2)^2$

On pratique, on met toujours au numérateur la plus grande des deux variances ($\hat{\sigma}_1^2 > \hat{\sigma}_2^2$) pour que le rapport des variances soit supérieur à 1.

Pour prendre une décision, la valeur de F observée est comparée à la valeur théorique F_{seuil} lue dans la table de Fisher-Snédeéor de degré de liberté $(n_1 - 1, n_2 - 1)$ et pour un risque d'erreur α fixé.

• si $F_{observée} > F_{seuil}$ l 'hypothèse H_0 est rejetée au risque d'erreur α : Les deux échantillons sont issues des deux populations ayant des variances significativement différentes au risque d'erreur α ; $\hat{\sigma}_1^2$ et $\hat{\sigma}_2^2$.

• si $F_{observée} \leq F_{seuil}$ l'hypothèse H_0 ne peut pas être rejetée au risque d'erreur α : Les deux échantillons sont issues des deux populations ayant des variances statistiquement égales à $\hat{\sigma}^2$ au risque d'erreur α.

b) ***Comparaison de deux moyennes de deux échantillons gaussiens indépendants***

Soient X_1 et X_2 deux variables aléatoires quantitatives continues observées sur deux populations indépendantes suivant une loi normale de paramètres (μ_1, σ_1) et (μ_2, σ_2) respectivement. On

tire deux grands échantillons indépendants l'un de l'autre de ces deux populations de tailles n_1 et n_2 respectivement et n_1 et n_2 sont supposées supérieures à 30.

On suppose que dans les deux populations les espérances sont égales. L'hypothèse nulle à tester est donc : $H_0 : \mu_1 = \mu_2$ contre $H_1 : \mu_1 \neq \mu_2$.

✓ **Si les variances théoriques sont connues**

$X_1 \rightsquigarrow N(\mu_1 , \sigma_1)$ et $X_2 \rightsquigarrow N(\mu_2 , \sigma_2)$ et X_1 et X_2 sont indépendantes.

Soit (X_{n_1}) une suite de variables aléatoires indépendantes et de même loi (loi normale).

Soit (X_{n_2}) une suite de variables aléatoires indépendantes et de même loi (loi normale).

$\bar{X}_1$ la moyenne dans la première population suit une loi normale telle que :

$\bar{X}_1 \rightsquigarrow N(\mu_1 , \sigma_1/\sqrt{n_1})$

$\bar{X}_2$ la moyenne dans la seconde population suit une loi normale telle que :

$\bar{X}_2 \rightsquigarrow N(\mu_2 , \sigma_2/\sqrt{n_2})$

$\bar{X}_1$ et $\bar{X}_2$ étant indépendantes, on peut établir la loi de probabilité de la différence :

$$\bar{X}_1 - \bar{X}_2 \rightsquigarrow N\left(\mu_1 - \mu_2 , \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$$

On peut définir une autre variable aléatoire Z telle que :

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - E(\bar{X}_1 - \bar{X}_2)}{\sqrt{V(\bar{X}_1 - \bar{X}_2)}} = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Puisque : $E(\bar{X}_1 - \bar{X}_2) = \mu_1 - \mu_2$ et $V(\bar{X}_1 - \bar{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$

Sous $H_0 : \mu_1 = \mu_2$

$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$ et grâce au théorème central-limite la variable Z suit une loi normale centrée réduite N(0,1).

Il s'agit d'un test bilatéral. La statistique Z calculée à partir des deux échantillons est :

$$Z_{\text{observée}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

La valeur de Z observée est comparée, en valeur absolue, à la valeur théorique z_{seuil} lue dans la table de la loi normale centrée réduite pour un risque d'erreur α fixé. z_{seuil} est le fractile de la loi normale centrée réduite d'ordre 1- $\frac{\alpha}{2}$ car le test est bilatéral.

• si $|z_{observée}| > z_{seuil}$ l 'hypothèse H_0 est rejetée au risque d'erreur α : Les deux échantillons sont issues des deux populations ayant des espérances statistiquement et significativement différentes $\mu_1 et\ \mu_2$ au risque d'erreur α.

• si $|z_{observée}| \leq z_{seuil}$ l'hypothèse H_0 ne peut pas être rejetée au risque d'erreur α : Les deux échantillons sont issues des deux populations ayant des espérances statistiquement égales à μ au risque d'erreur α.

Pour un risque de première espèce α = 5%, la valeur de z_{seuil} est le fractile de la loi normale centrée réduite d'ordre 1-$\frac{\alpha}{2}$ soit 0.975 est égale à 1,96.

Remarque : Lorsque le test est unilatéral ($H_1: \mu_1 > \mu_2$ par exemple), la valeur de Z observée est comparée sans valeur absolue à la valeur théorique z_{seuil} lue dans la table de la loi normale centrée réduite pour un risque d'erreur α fixé. z_{seuil} est le fractile de la loi normale centrée réduite d'ordre 1-α. Pour un risque de première espèce α = 5%, la valeur de z_{seuil} est le fractile de la loi normale centrée réduite d'ordre 1-α soit 0,95 est égale à 1,645.

✓ **Si les variances théoriques sont inconnues et égales**

$X_1 \rightsquigarrow N(\mu_1, \sigma)$ et $X_2 \rightsquigarrow N(\mu_2, \sigma)$ et X_1 et X_2 sont indépendantes.

Soit (X_{n_1}) une suite de variables aléatoires indépendantes et de même loi (loi normale).

Soit (X_{n_2}) une suite de variables aléatoires indépendantes et de même loi (loi normale).

$\bar{X}_1$ la moyenne dans la population suit une loi normale telle que : $\bar{X}_1 \rightsquigarrow N(\mu_1, \sigma / \sqrt{n_1})$

$\bar{X}_2$ la moyenne dans la population suit une loi normale telle que : $\bar{X}_2 \rightsquigarrow N(\mu_2, \sigma / \sqrt{n_2})$

$\bar{X}_1$ et $\bar{X}_2$ étant indépendantes, on peut établir la loi de probabilité de la différence :

$$\bar{X}_1 - \bar{X}_2 \rightsquigarrow N(\mu_1 - \mu_2, \sqrt{\sigma^2 (\frac{1}{n_1} + \frac{1}{n_2})})$$

On peut définir une autre variable aléatoire T telle que :

$$T=\frac{(\bar{X}_1-\bar{X}_2)-E(\bar{X}_1-\bar{X}_2)}{\sqrt{V(\bar{X}_1-\bar{X}_2)}}=\frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{\sqrt{\sigma^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

Puisque : $E(\bar{X}_1-\bar{X}_2)=\mu_1-\mu_2$ et $V(\bar{X}_1-\bar{X}_2)=\frac{\sigma^2}{n_1}+\frac{\sigma^2}{n_2}=\sigma^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)$

Sous H_0 : $\mu_1=\mu_2$

$$T=\frac{\bar{X}_1-\bar{X}_2}{\sqrt{\hat{\sigma}^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

T suit une loi student à (n_1+n_2-2) degré de liberté car la variance théorique est inconnue et remplacée par son estimateur $\hat{\sigma}^2$.

Il s'agit d'un test bilatéral. La statistique T calculée à partir des deux échantillons est :

$$t_{observée}=\frac{\bar{x}_1-\bar{x}_2}{\sqrt{\hat{\sigma}^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \quad \text{avec} \quad \hat{\sigma}^2=\frac{n_1s_1^2+n_2s_2^2}{n_1+n_2-2}$$

avec : $s_1^2=\frac{1}{n_1}\sum_{i=1}^{n_1}(x_{1i}-\bar{x}_1)^2$ et $s_2^2=\frac{1}{n_2}\sum_{j=1}^{n_2}\left(x_{2j}-\bar{x}_2\right)^2$

La valeur de t observée est comparée, en valeur absolue, à la valeur théorique $\mathbf{t_{seuil}}$ lue dans la table de la loi de student à (n_1+n_2-2) degré de liberté et pour un risque d'erreur α fixé. $\mathbf{t_{seuil}}$ est le fractile de la loi student à (n_1+n_2-2) ddl d'ordre 1- $\frac{\alpha}{2}$ car le test est bilatéral.

• si $|t_{observée}| > \mathbf{t_{seuil}}$ l'hypothèse H_0 est rejetée au risque d'erreur α: Les deux échantillons sont issues des deux populations ayant des espérances statistiquement et significativement différentes (μ_1 et μ_2) au risque d'erreur α.

• si $|t_{observée}| \leq \mathbf{t_{seuil}}$ l'hypothèse H_0 ne peut pas être rejetée au risque d'erreur α: Les deux échantillons sont issues des deux populations ayant des espérances statistiquement égales à μ au risque d'erreur α.

✓ **<u>Si les variances théoriques sont inconnues et inégales</u>**

Soient $X_1 \rightsquigarrow N(\mu_1, \sigma_1)$ et $X_1 \rightsquigarrow N(\mu_2, \sigma_2)$ et X_1 et X_2 sont indépendantes.

Soit (X_{n_1}) une suite de variables aléatoires indépendantes et de même loi (loi normale).

Soit (X_{n_2}) une suite de variables aléatoires indépendantes et de même loi (loi normale).

$\bar{X}_1$ la moyenne dans la population suit une loi normale telle que : $\bar{X}_1 \rightsquigarrow N(\mu_1, \sigma_1/\sqrt{n_1})$

$\bar{X}_2$ la moyenne dans la population suit une loi normale telle que : $\bar{X}_2 \rightsquigarrow N(\mu_2, \sigma_2/\sqrt{n_2})$

$\bar{X}_1$ et $\bar{X}_2$ étant indépendantes, on peut établir la loi de probabilité de la différence :

$$\bar{X}_1 - \bar{X}_2 \rightsquigarrow N(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}})$$

On peut définir une autre variable aléatoire T telle que :

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - E(\bar{X}_1 - \bar{X}_2)}{\sqrt{V(\bar{X}_1 - \bar{X}_2)}} = \frac{(\bar{X}1 - \bar{X}2) - (\mu 1 - \mu 2)}{\sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}}$$

Car : $E(\bar{X}_1 - \bar{X}_2) = \mu_1 - \mu_2$ et $V(\bar{X}_1 - \bar{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$

Sous H_0 : $$T = \frac{\bar{X}1 - \bar{X}2}{\sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}}$$

T suit une loi student à (n_1+n_2-2) degré de liberté car les variances théoriques sont inconnues et inégales. Elles seront remplacées par leurs estimateurs $\hat{\sigma}_1^2$ et $\hat{\sigma}_2^2$.

On a : $\hat{\sigma}_1^2 = \frac{n_1}{n_1-1} s_1^2$, $\hat{\sigma}_2^2 = \frac{n_2}{n_2-1} s_2^2$

Avec : $S_1^2 = \frac{1}{n_1}\sum_{i=1}^{n_1}(X_{1i} - \bar{X}_1)^2$, $S_2^2 = \frac{1}{n_2}\sum_{j=1}^{n_2}(X_{2j} - \bar{X}_2)^2$

La statistique t calculée à partir des deux échantillons indépendants est :

$$t_{observée} = \frac{\bar{x}1 - \bar{x}2}{\sqrt{\frac{s_1^2}{n_1-1} + \frac{s_2^2}{n_2-1}}}$$

La valeur de t observée est comparée, en valeur absolue, à la valeur théorique $\mathbf{t_{seuil}}$ lue dans la table de la loi de student $(n_1 + n_2 - 2)$ degré de liberté pour un risque d'erreur α fixé. $\mathbf{t_{seuil}}$ est le fractile de la loi student à $(n_1 + n_2 - 2)$ ddl d'ordre 1- $\frac{\alpha}{2}$ car le test est bilatéral.

. si $|t_{observée}| > \mathbf{t_{seuil}}$ l'hypothèse H_0 est rejetée au risque d'erreur α : Les deux échantillons sont issues des deux populations ayant des espérances statistiquement et significativement différentes μ_1 et μ_2 , au risque d'erreur α

• si $|t_{observée}| \leq \mathbf{t_{seuil}}$ l'hypothèse H_0 ne peut pas être rejetée au risque d'erreur α : Les deux échantillons sont issues des deux populations ayant des espérances statistiquement égales μ au risque d'erreur α.

c) ***Comparaison de deux proportions (grands échantillons)***

Soit p_1 et p_2 deux proportions d'individus d'une certaine modalité A dans deux populations M_1 et M_2 respectivement. On tire un échantillon de taille n_1 de M_1 et un échantillon de taille n_2 de M_2. n_1 et n_2 sont supposés suffisamment grands (supérieurs à 30).

On estime p_1 par $f1$ et p_2 par $f2$ dans les deux échantillons respectifs.

On cherche à tester :

$$\begin{cases} H_0: P_1 = P_2 = P \\ H_1: P_1 \neq P_2 \end{cases}$$

$f1$ et $f2$ sont des réalisations de deux variables aléatoires indépendantes F_1 et F_2 respectivement, suivant les lois approximatives :

$$F_1 \rightsquigarrow \mathrm{N}(p_1 , \sqrt{\frac{p_1(1-p_1)}{n_1}})$$

$$F_2 \rightsquigarrow \mathrm{N}(p_2 , \sqrt{\frac{p_2(1-p_2)}{n_2}})$$

On peut donc établir la loi de probabilité de la différence :

$$F_1 - F_2 \rightsquigarrow \mathrm{N}(p_1 - p_2 \ , \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \ )$$

Puisque :

$\mathrm{E}(F_1 - F_2) = p_1 - p_2$

$\mathrm{V}(F_1 - F_2) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$

Et on peut définir une variable aléatoire Z telle que :

$$Z = \frac{(F_1 - F_2) - \mathrm{E}(F_1 - F_2)}{\sqrt{\mathrm{V}(F_1 - F_2)}} = \frac{(F_1 - F_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$$

Sous H_0 : $p_1 = p_2 = p$ avec $p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$ et $q = 1-p$

$$Z = \frac{F_1 - F_2}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} = \frac{F_1 - F_2}{\sqrt{pq(\frac{1}{n_1} + \frac{1}{n_2})}}$$

Grâce au théorème central limite, la variable aléatoire Z suit une loi normale centrée réduite $N(0,1)$.

Si la valeur de *p* probabilité de succès commune aux deux populations n'est en réalité pas connue, on l'estime npar les résultats observés dans les deux échantillons par son estimation :

$\hat{p} = \frac{n_1 f_1 + n_2 f_2}{n_1 + n_2} = \frac{k_1 + k_2}{n_1 + n_2}$ et $\hat{q} = 1 - \hat{p}$

avec $f_1 = \frac{k_1}{n_1}$ et $f_2 = \frac{k_2}{n_2}$ et k_1 et k_2 représentent le nombre de succès observés dans le l'échantillon 1 et 2 respectivement.

Il s'agit d'un test bilatéral puisque ($H_1 : P_1 \neq P_2$) et la statistique Z calculée à l'aide des échantillons est égale à :

$$Z_{\text{observée}} = \frac{\frac{k_1}{n_1} - \frac{k_2}{n_2}}{\sqrt{\hat{p}\,\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

La valeur de Z observée est comparée, en valeur absolue, à la valeur théorique z_{seuil} lue dans la table de la loi normale centrée réduite pour un risque d'erreur α fixé. z_{seuil} est le fractile de la loi normale centrée réduite d'ordre $1-\frac{\alpha}{2}$.

• si $|Z_{observée}| > z_{seuil}$, l'hypothèse H_0 est rejetée au risque d'erreur α : Les deux échantillons sont issues des deux populations ayant des probabilités de succès statistiquement et significativement différentes $p_1 et\ p_2$, au risque d'erreur α

• si $|Z_{observée}| \leq z_{seuil}$ l'hypothèse H_0 ne peut pas être rejetée au risque d'erreur α : Les deux échantillons sont issues des deux populations ayant des probabilités de succès statistiquement égales à p au risque d'erreur α.

Pour un risque de première espèce α = 5%, la valeur de z_{seuil} est le fractile de la loi normale centrée réduite d'ordre $1-\frac{\alpha}{2}$ soit 0.975 est égale à 1,96.

Remarque : Lorsque le test est unilatéral ($H_1 : P_1 > P_2$), la valeur de Z observée est comparée sans valeur absolue à la valeur théorique z_{seuil} lue dans la table de la loi normale centrée réduite

pour un risque d'erreur α fixé. z_{seuil} est le fractile de la loi normale centrée réduite d'ordre $1-\alpha$. Pour un risque de première espèce $\alpha = 5\%$, la valeur de z_{seuil} est le fractile de la loi normale centrée réduite d'ordre $1-\alpha$ soit 0.95 est égale à 1,645.

Exercices – Chapitre 03

Exercice 1 :

Une v.a X suit une loi normale N(m, σ). Au vu d'un échantillon de **n** v.a (X1, … X_n**)** de loi de X, on veut choisir entre les hypothèses : $\begin{cases} H_0 : \mathrm{m} = 2 \\ H_1 : \mathrm{m} = 3 \end{cases}$

a . *test de l'espérance d'une loi normale d'écart type connu* σ=2

1. Résoudre ce problème par la méthode de Neyman et Pearson.
2. Dans le cas où n=100 et α = 0,05 calculer la puissance de ce test. Qu'en concluez-vous ?
3. Quelle doit-être la taille d'échantillon minimum $\boldsymbol{n_0}$pour que la puissance soit supérieure à 0,95% ?

b . test de l'espérance d'une loi normale d'écart type inconnu

On suppose maintenant que d'écart type est inconnu.

Déterminer la région critique de ce test quand α = 0,05 et n=25. Calculer sa puissance.

Exercice 2 :

On dispose d'un échantillon de taille n=15 d'une variable aléatoire de loi normale centrée et de variance $1/\theta$ pour choisir entre les deux hypothèses :

$$\begin{cases} H_0 : \theta = 1 \\ H_1 : \theta > 1 \end{cases}$$

Déterminer la région critique d'un test UPP de risque de première espèce α et préciser sa fonction puissance. Calculer cette puissance dans le cas où n=15, $\theta = 3$ et $\alpha = 0.05$.

Exercice 3 :

On dispose d'un échantillon de taille n=12 d'une variable aléatoire X qui suit loi normale d'espérance inconnue m et d'écart type inconnu σ pour choisir entre les deux hypothèses :

$$\begin{cases} H_0 : m \leq 6 \\ H_1 : m > 6 \end{cases}$$

Déterminer la région critique d'un test de niveau $\alpha = 0.05$. Peut-on déterminer sa fonction puissance.

Exercice 4 :

Une entreprise, spécialisée dans la fabrication des pâtes « aux œufs frais », fonde sa publicité sur l'affirmation suivante : « 3 œufs frais au kilo, soit 14 % de la composition ». Le producteur s'engage ainsi à respecter certaines normes de qualité de fabrication portant sur les caractéristiques de la v.a. X représentant le pourcentage d'œufs dans la composition d'un paquet, supposée de loi $N(\mu, \sigma)$.

1) On suppose $\sigma=1$. La machine à fabriquer les pâtes est bien réglée si m= 14. le fabriquant considère qu'un paquet et non conforme à sa norme de qualité si $x < 14$. Calculer la probabilité $\boldsymbol{p}$ pour qu'un paquet ne soit pas conforme lorsque la machine est bien réglée.
Quelle doit être la valeur minimale de $\boldsymbol{m}$ pour que la probabilité p(m) qu'un paquet ne soit pas conforme ne dépasse pas 5% ?

2) Une grande surface reçoit un lot de ces pâtes, et désire vérifier la qualité du fournisseur. Un échantillon indépendant de 76 paquets est tiré ; soit $(X_1, \ldots, X_{76})$ les observations après analyse ; $\square$ est inconnu.
On suppose $\boldsymbol{m}$ =16 et on donne les informations suivantes : $\sum_1^{76} xi = 140$, $\sum_1^{76} xi^2 = 17\,195$. Tester, au risque de 5% :

$H_0 : \sigma = 1$ contre $1 : \sigma = 2$ Puis $H_0 : \sigma = 2$ contre $H_1 : \sigma = 1$. Que conclure ?

Exercice 5 :

Avant le second tour d'une élection présidentielle, un candidat commande un sondage à une société spécialisée pour savoir s'il a une chance d'être élu. Si **p** est la proportion d'électeurs qui lui est favorable dans la population, il souhaite résoudre le problème de test suivant :

H_0 : p = 0,48 contre H_1 : p = 0,52

1) Quelle est la signification du choix de p = 0,48 comme hypothèse nulle ?

2) Déterminer la région critique w pour $\alpha = 0,10$ si le sondage est effectué auprès de n=100 électeurs ? Que penser du résultat ?

3) Indique comment varie la région critique et la puissance du test en fonction de n ? calculer la valeur de cette puissance pour n=100, 500 puis 1000.

On considère maintenant le test : H_0 : p = 0,49 contre H_1 : p = 0,51

Calculer la taille d'échantillon minimum n_0 pour que la puissance soit supérieure à 0,90. Quelle est alors la région critique

Exercice 6 : Le salaire moyen des professeurs des universités américaines est de 61 650 dollars par an en 1995. Le salaire moyen d'un échantillon de 36 professeurs d'Ecoles de Commerce est de 72 800 dollars et l'écart-type de l'échantillon est égal à 5000 dollars.

1. Construisez un intervalle de confiance à 95% pour le salaire de ces derniers.
2. Tester les hypothèses : $H_0 : m = 61650$ et $H_1 : m \neq 61650$.
3. Utilisez l'intervalle de confiance trouvé en Q1 pour tester les hypothèses précédentes

Exercice 7 : Dans la population générale, on sait que 25% des individus réussissent à un certain test (ex : capacité à reproduire parfaitement un exercice). On soumet un échantillon de taille 200 de cette population à un même stress (veille prolongée) et on renouvelle l'expérience : seulement 42 individus réussissent le test.

Peut-on affirmer à une erreur de 1ere espèce de 5% que le stress fait diminuer les performances au test des individus ?

Exercice 8 : Parmi 3000 entreprises nouvelles déclarées au cours du premier trimestre 1998, il y a 2000 entreprises provenant d'une création pure (les autres provenant de la reprise ou de la réactivation d'une entreprise existante).

Cinq ans plus tard, nous observons qu'il reste 1490 entreprises actives parmi lesquelles 960 proviennent d'une création pure. Votre collègue vous affirme que le taux de survie à cinq ans est de 48% pour les créations pures contre 49,6% pour l'ensemble des entreprises et que " le fait d'avoir été une création pure ne représente pas un handicap significatif pour les entreprises ". Vous n'êtes pas d'accord avec cette analyse.

1. En quoi l'analyse de votre collègue est-elle faussée ?
2. Effectuez le test de seuil 5% pour décider si le fait d'avoir été une création pure représente ou non, un handicap pour les entreprises.

Exercice 9: Dans un atelier de traitements thermiques, on met en service 80 paniers de type I servant à la trempe d'arbre de boite à vitesse et 60 de type II servant à la trempe de pignons de boite à vitesse. Six mois plus tard, il reste en service respectivement 50 paniers sur 80 du type I et 40 du type II. La résistance à l'usure des deux séries de paniers peut-elle être considérée comme identique ?

Exercice 10 :

Un médecin propose un régime amaigrissant sans en connaître la durée optimale de mise en œuvre. Pour estimer cette durée, il réalise un suivi de 5 groupes de personnes adultes souhaitant perdre du poids : le premier groupe utilise le régime pendant 1 semaine, le second 4 semaines, le troisième 5 semaines, le quatrième 6 semaines et le dernier 8 semaines. Le tableau suivant donne les résultats de ce suivi :

Durée de mise en œuvre du régime	1 semaine	4 semaines	5 semaines	6 semaines	8 semaines
Nombre de personnes concernés	144	36	36	36	252
Moyenne de poids perdu (en Kg)	2,55	2,6	2,7	2,7	2,75
Ecart type estimé du poids (kg)	0,8	0,9	0,8	1,1	1,3

1. Estimer par intervalle de confiance au niveau 99% le poids moyen perdu après 1 semaine de régime
2. Tester à la confiance 95% l'hypothèse selon laquelle le régime durant 8 semaines n'est pas plus efficace que la moyenne des autres régimes ?
3. Quel devrait être l'effectif de l'ensemble de l'échantillon pour pouvoir mettre en évidence une différence de 250 g de perte de poids entre le régime durant 8 semaines et l'ensemble des autres régimes (95% de puissance et de confiance) ?

Exercice 11 :

Dans une expérience, on s'est intéressé au rôle du contexte sur les performances cognitives dans des tâches scolaires.

Deux groupes de 33 élèves (chacun) de cinquième ont été constitués.

Les élèves avaient à résoudre un certain nombre de problèmes se rattachant à la géométrie. Dans le premier groupe (G1), on leur demandait de résoudre ces problèmes dans le cadre de cours de mathématiques. Les élèves du second groupe (G2) devaient résoudre les mêmes problèmes dans le cadre de cours de dessin.

Chaque sujet recevait ensuite une note sur 20 en fonction du nombre de problèmes résolus. Les moyennes et les variances observées dans les deux groupes avaient été les suivantes :

	Moyenne	variance
G1	09,23	02,64
G2	11,46	04,40

1- Pour le premier groupe G1 :

a- Donner les estimations ponctuelles non biaisées de la moyenne m_1 et de la variance σ^2

b- Donner un intervalle de confiance pour la moyenne théorique au seuil 0,05 ?

c- En déduire la conclusion au test $H_0 : m_1 = 10$ contre $H_1 : m_1 \neq 10$

2- Les résultats des deux groupes et dans les deux conditions diffèrent-ils significativement au seuil 0,05 ? Préciser la condition nécessaire qui permet la réalisation de ce test.

Exercice 12 :

Une législation impose aux aéroports une intensité maximum de bruit égale à 80 décibels au décollage et à l'atterrissage des avions. Au-delà de cette limite, l'aéroport doit indemniser les riverains. On admet que la v.a. X dont les valeurs représentent l'intensité du bruit causé par un avion d'un certain type obéit à une loi normale N(m, σ). Les habitants d'un village proche de l'aéroport assurent que la limite de 80db est dépassée en moyenne et demandent une expertise. On décide de faire au seuil 1% le test :

$H_0 : m = 80$ contre $H_1 : m < 80$

On suppose que σ est connu et égal à 7

1°) Montrer que le test préserve les intérêts des riverains, le risque d'autoriser à voler des avions trop bruyants étant maîtrisé. Préciser à quoi correspond le risque de $2^{ème}$ espèce .

2°) On enregistre atterrissages-décollages sur un échantillon de 40 avions du type considéré. Sous l'hypothèse H_0 quelle est la loi de la v.a. X ? Déterminer la région critique au seuil de 1%. Enoncer la règle de décision du test.

3°) La compagnie qui commercialise ce type d'avions ; affirme que l'intensité moyenne du bruit occasionné par ces avions est de 78db (avec un écart-type de 7). Si cette affirmation était vraie, quelle serait la probabilité, pour l'aéroport, de verser à tort des indemnités aux riverains à la suite du test ?

4°) L'échantillon de 40 enregistrements a donné une intensité moyenne x de 79db ; quelle est la conclusion du test au seuil 1% ?

5°) Quel doit être le nombre minimal d'enregistrement à effectuer pour que, dans ce test, les risques soient les suivants :

- les riverains ne perçoivent pas l'indemnité qui leur est due, avec une probabilité au plus égale à 0,01

- la vraie valeur de m est bien de 78db comme l'affirme la compagnie, et l'aéroport verse (à tort) des indemnités aux riverains, avec une probabilité au plus égale à 0,05.

Chapitre 04 : Tests non paramétriques -

Test de Khi-deux d'indépendance et test de Khi-deux d'adéquation

Le test de Khi-deux est un test non-paramétrique, car il n'est pas basé sur les prémisses des paramètres de la distribution de la variable dans la population (moyenne, écart-type et normalité). Il existe d'autres tests non-paramétriques, mais nous ne les verrons pas dans ce cours.

I. Test de Khi-deux d'indépendance

1. Définition du test d'indépendance

Le test d'indépendance est utilisé pour tester l'hypothèse nulle d'absence de relation entre deux variables qualitatives. On peut également dire que ce test vérifie l'hypothèse d'indépendance de ces variables. Si deux variables dépendent l'une de l'autre ; la variation de l'une influence la variation de l'autre.

2. L'hypothèse nulle du test

L'hypothèse nulle est l'hypothèse de l'absence de relation entre deux variables qualitatives.
C'est l'hypothèse d'indépendance. On suppose que la valeur d'une des deux variables ne nous donne aucune information sur la valeur possible de l'autre variable. Lorsqu'il n'existe aucune relation entre deux variables on dit que les variables sont **indépendantes** l'une de l'autre. Il ne faut pas confondre cette expression avec l'appellation « variable indépendante ».
L'hypothèse alternative est donc qu'il existe une relation entre les variables ou que les deux variables sont dépendantes.

$$\begin{cases} H_0 : Les\ deux\ variables\ sont\ indépendantes \\ H_1 : \quad Les\ deux\ variables\ sont\ dépendantes \end{cases}$$

3. *Conditions du test*

Les observations doivent être indépendantes, ce qui signifie que les individus statistiques apparaissent une fois dans le tableau et que les catégories des variables sont mutuellement exclusives.

La majorité des effectifs attendus (théoriques) d'un tableau croisé doivent être supérieures ou égales à 5 et aucun effectif attendu ne doit être inférieure à 1.

4. *Statistique du test*

Lorsque l'on a voulu tester l'hypothèse nulle de l'égalité des moyennes de deux échantillons indépendants, nous avons calculé la statistique Z (ou T si les variances sont inconnues). Puis, à l'aide de la distribution de la statistique du test, nous avons déterminé dans quelle mesure la valeur z obtenue était « inhabituelle » si l'hypothèse nulle était vraie.

Dans le cas de tableau croisé où l'on travaille avec des effectifs (ou occurrences), nous allons calculer la statistique Khi-deux χ^2 et comparer sa valeur à l'aide de la distribution Khi-deux dans le but de déterminer dans quelle mesure cette valeur est « inhabituelle » si l'hypothèse nulle est vraie.

La procédure statistique que nous allons employer pour tester l'hypothèse nulle compare les effectifs observés (celles déjà dans le tableau de données) avec les effectifs attendus ou théoriques. L'effectif attendu est simplement l'effectif que l'on devrait trouver dans une cellule si l'hypothèse nulle était vraie.

Les étapes du test seront présentées en traitant l'exemple illustratif suivant :

Au niveau national, un examen est ouvert à des étudiants de spécialités différentes (Economie et Mathématiques). Les responsables de l'examen désirent savoir si la formation initiale d'un étudiant influence sa réussite à cet examen. Les résultats sont enregistrés dans le tableau suivant :

		Spécialité		Total
		Economie	Mathématiques	
Réussite	*effectif*	268	195	463
	%	57,9%	42,1%	100,0%
Echec	*effectif*	232	240	472
	%	49,2%	50,8%	100,0%
Total	*effectif*	500	435	935
	%	53,5%	46,5%	100,0%

a) ***Etape 1 : calculer l'occurrence théorique***

Si l'hypothèse nulle est vraie, on s'attend à ce que les pourcentages du tableau soient les mêmes pour les étudiants ayant réussis l'examen et les étudiants n'ayant pas réussis l'examen. Dans le tableau croisé ci-dessus, nous remarquons que 53,5 % des étudiants sont en économie et que 46,5 % en Mathématiques.

La façon la plus simple pour calculer les effectifs théoriques d'une cellule est de multiplier l'effectif total de la ligne de cette cellule $n_{i.}$ par l'effectif total de la colonne de cette même cellule $n_{.j}$ et de diviser par le nombre total d'effectif observés du tableau n : $\frac{n_{i.}n_{.j}}{n}$ . .

Par exemple, pour les étudiants en économie n'ayant pas réussi à l'examen national, l'effectif attendu est : (500 * 472)/ 935 = 252,4.

pour les étudiants en économie ayant réussi l'examen national, l'effectif attendu est : (500 * 463)/ 935 = 247,6.

pour les étudiants en mathématique s n'ayant pas réussis à l'examen national l'effectif attendu est : (435 * 472)/ 935 = 219,6.

pour les étudiants en mathématiques ayant réussi l'examen national, l'effectif attendu est est : (435 * 463)/ 935 = 215,4.

Remarque : vous pouvez aussi utiliser les pourcentages de la façon suivante :

Pour les étudiants en économie n'ayant pas réussis, la fréquence attendue est : 53,5 % x 472 = **252,4**.

b) ***Etape 2 : Trouvez la différence entre l'effectif observé et l'effectif attendu, nommée le résiduel.***

Un résiduel positif indique qu'il y a plus d'occurrences comparativement à ce qu'on s'attendrait à observer si l'hypothèse nulle était vraie. Ceci est aussi vrai à l'inverse pour les résiduels négatifs.

Tableau croisé Résultat de l'examen*spécialité

		Spécialité		Total
		Economie	Mathématiques	
Réussite	*effectif observé*	268	195	463
	effectif attendu	247,6	215,4	463
	différence	20,4	-20,4	
Echec	*effectif observé*	232	240	472
	effectif attendu	252,4	219,6	472
	différence	-20,4	20,4	
Total	*effectif observé*	500	435	935
	effectif attendu	500	435	935
	différence			

c) ***Élevez les différences entre l'effectif observé de la cellule n_{ij} et l'effectif attendu $\frac{n_{i.}n_{.j}}{n}$ au carré***

d) ***Divisez cette différence au carré par l'effectif attendu***

e) ***Trouvez la quantité khi-deux observée:***

$$\chi^2_{observé} = \sum_{i=1}^{l} \sum_{j=1}^{c} \frac{(n_{ij} - \frac{n_{i.}n_{.j}}{n})^2}{\frac{n_{i.}n_{.j}}{n}}$$

$$\chi^2_{observé} = \frac{(268-247,6)^2}{247,6} + \frac{(195-215,4)^2}{215,4} + \frac{(232-252,5)^2}{252,5} + \frac{(240-219,6)^2}{219,6} = 7,16$$

f) ***Décision***

Le degré de liberté de la distribution de khi-deux ne dépend pas du nombre d'individus statistiques, mais plutôt du nombre de lignes et de colonnes du tableau croisé.

Degré de liberté = (nombre de rangées – 1) X (nombre de colonnes – 1)

Dans ce cas, le degré de liberté de la distribution khi-deux est de 1. Il suffit maintenant de comparer cette statistique khi-deux observée à la valeur de khi-deux théorique ou taulée lue dans

la table de distribution khi-deux paramétrée par le degré de liberté en fonction du niveau de signification choisi. Il sera alors possible ou non de rejeter l'hypothèse nulle d'absence de relation. Dans notre exemple, on rejette H_0 au risque d'erreur 5% puisque 7,16 > 3,84. Cela signifie que la réussite des étudiants à cet examen dépend de leur spécialité au risque 5%.

II. Test de khi-deux d'adéquation

Soit (X_1, X_2, ….. X_n) un n-échantillon d'une variable aléatoire X et F une fonction de répartition donnée.

On désire tester :

$$\begin{cases} H_0 : X \text{ a pour fonction de répartition } F \\ H_1 : X \text{ n'a pour fonction de répartition } F \end{cases}$$

Pour cela on répartit les n observations en k classes $[a_{i-1}, a_i[$ d'effectif observé n_i ($1 \leq i \leq k$) et on calcule $p_i = F(a_i) - F(a_{i-1})$ pour trouver ensuite les effectifs attendus (ou théoriques) np_i .

On peut obtenir ainsi la valeur observée de la variable aléatoire D qui suit la loi de khi-deux à k-1 ddl : $D_{observée} = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}$

Avec $n = \sum_{i=1}^{k} n_i$

La région critique du test de seuil α est : $D \geq D_{seuil}$

La valeur de D_{seuil} étant lue dans la table de Khi-deux avec $\alpha = P(D \geq D_{seuil} / H_0)$ et D de loi approchée de khi-deux à k-1 ddl sous H_0.

L'effectif np_i de la classe $[a_{i-1}, a_i[$ doit être supérieur ou égal à 5 ; sinon on regroupe deux (ou plus) classes consécutives.

Exemple : On souhaite étudier la circulation en un point fixe d'une autoroute en comptant, pendant deux heures, le nombre de voitures passant par minute devant un observateur. Le tableau suivant résume les données obtenues :

débit en voitures par minute	Fréquences observées ni
00	04
01	09
02	24
03	25
04	22
05	18
06	06
07	05
08	03
09	02
10	01
11	01
Total	120

Tester l'adéquation de la loi empirique observée à une loi théorique simple au seuil $\alpha = 0{,}10$?

Soit X une variable aléatoire représentant le trafic des voitures par minute. Le nombre moyen $\bar{x}$ de voitures par minute est égal à 3,70 et la variance empirique $s^2 = 4{,}37$. Les deux valeurs sont proches à 4.

On peut penser ajuster la loi de X à une loi de Poisson de paramètre $\lambda = 4$.

On obtient les probabilités p_i = P(X=x) , x ≥ 0 en lisant la table de Poisson P(4). Par exemple pour X=3 ; p_4 = P(X=3) = 0,1954 ; np_4 = 120 * 0,1954 = 23,448 arrondi à 23 voitures.

débit en voitures par minute	Fréquences observées ni	Fréquences théoriques np_i
00	04	02
01	09	09
02	24	18
03	25	23
04	22	23
05	18	19
06	06	13
07	05	07
08	03	03
09	02	02
10	01	01
11	01	00
Total	120	120

On constate que certains effectifs np_i sont inférieurs à 5. On regroupe les deux premières classes et aussi les quatre dernières classes, on obtient :

débit en voitures par minute	Fréquences observées ni	Fréquences théoriques np_i
[00-02[	13	11
02	24	18
03	25	23
04	22	23
05	18	19
06	06	13
07	05	07
[08-12[	07	06
Total	120	120

La valeur de $D_{observée} = \frac{(13-11)^2}{11} + \frac{(24-18)^2}{18} + \frac{(25-23)^2}{23} + \frac{(22-23)^2}{23} + \frac{(18-19)^2}{19} + \frac{(6-13)^2}{13} + \frac{(5-7)^2}{7} + \frac{(7-6)^2}{6} = 7{,}141$

On compare cette valeur de 7,141 à la valeur du fractile de la loi de khi-deux à 7 ddl (car on a 8 classes) d'ordre 0.90, lue dans la table de khi-deux, qui est égal à 12,017. On est donc dans la région d'acceptation de H_0 et par conséquent on ne rejette pas l'ajustement à la loi de P(4).

Exercices – chapitre 04

Exercice 1 :

Nous souhaitons étudier la relation entre la répartition des chèques bancaires (selon leur montant) et la ville française (Paris, Lyon, Marseille). La distribution observée est la suivante :

	Répartition des chèques selon leur montant		
	< 3500	[3500-7500[	≥ 7500
PARIS	2	18	5
LYON	3	12	2
MARSEILLE	10	30	3

Existe-t-il une dépendance entre la ville et la répartition des chèques bancaires en France au seuil 5%?

Exercice 2 :

A la sortie d'une chaine de fabrication d'une pièce mécanique, on prélève, toutes les trente minutes, 20 pièces. On compte, pour chaque groupe de 20 pièces ainsi contrôlé, le nombre de pièces défectueuses. Après observation de 200 échantillons indépendants, on obtient le tableau suivant :

Nombre de déchets X par échantillon de 20 pièces	Nombre de fois où l'on a rencontré X déchets
0	26
1	52
2	56
3	40
4	20
5	2
6	0
7	4

X désignera la variable « nombre de pièces défectueuses » par échantillon de 20 pièces. A quelle(s) loi(s) peut-on ajuster la loi empirique observée ? $\alpha = 0{,}05$.

Références bibliographiques

1. BICKEL P.J., K.ADoksum. Mathematical statistics. Prentice Hall 1977/2001.
2. DACUNCHA D. Castelle et M. Duflo, Probabilités et statistiques, Tome 1, Masson 1983
3. LECOUTRE J.P, Probabilités et statistique : Exercices corrigés avec rappels de cours. 5eme édition. Edition MASSON, juin 2014
4. MEOT A., Introduction aux statistiques inférentielles : de la logique à la pratique avec exercices et corrigés. édition (de boeck), 2003
5. MONFORT A., Cours de statistique mathématique. Economie, 1997
6. SAPORTA G., Probabilités analyse des données et statistique. Editions TECHNIP, 1990
7. Le cours de D. Mouchiroud sur les tests diffusé sur Internet, notamment pour le paragraphe « Tests paramétriques les plus courants ».