وزارة التعليم العالي والبحث العلمي

Université Badji Mokhtar
Annaba

Badji Mokhtar University -
Annaba

جامعة باجي مختار

عنابة

Faculté des Sciences de la Terre
Département de Géologie

Polycopié de cours

MATHEMATIQUES 2

Par:
Dr. CHOUIA SANA

Année : 2019/2020

Table des matières

Introduction

Ce polycopié de cours "Mathématiques 02", enseigné aux étudiants de la licence mathématiques au premier semestre, département de Géologie, traite les différents aspects des statistiques descriptives de l'inférence statistique et de probabilité d'une façon simple. Il est utile aussi à toute personne souhaitant connaitre et surtout utiliser les principales méthodes.

Ce cours est organisé autour de quatre chapitres. Le premier sera consacré aux Statistiques Descriptives; dans le deuxième chapitre, on présente les différentes méthodes d'estimation statistique. Le troisième chapitre traite les tests statistiques paramétriques. On finit ce cours par quelques éléments de calcul de probabilités, dans le chapitre 04.

CHAPITRE 1

Statistiques Descriptives

Les statistiques descriptives visent à étudier les caractéristiques d'un ensemble d'observations comme les mesures obtenues lors d'une expérience. L'expérience est l'étape préliminaire à toute étude statistique.

1.1 Définitions

Epreuve statistique: L'épreuve statistique est une expérience que l'on provoque.

Population: On appelle population l'ensemble sur lequel porte notre étude statistique. Cet ensemble est noté Ω.

Unité statistique (individu): On appelle individu ou unité statistique tout élément de la population Ω, il est noté ω (ω dans Ω).

Variable statistique (Caractère): Une variable aléatoire est une fonction à valeurs numériques (réelles) définie sur un espace échantillon.

Exemple 1.1.1 *Supposons qu'on lance une pièce de monnaie; la fonction X qui associe le nombre 1 au résultat "face" et le nombre 0 au résultat "pile" est une variable aléatoire.*

Modalités: Les modalités d'une variable statistique sont les différentes valeurs que peut prendre celle-ci.

Types de Variables: Nous distinguons deux catégories de caractères : les caractères qualitatifs et les caractères quantitatifs.

Variable qualitative: La variable est dite qualitative quand les modalités sont des catégories.

Variable quantitative: Une variable est dite quantitative si toute ses valeurs possibles sont numériques.

On distingue habituellement: les variables aléatoires discrètes et les variables aléatoires continues.

Variable quantitative discrète: Si une variable aléatoire X prend un nombre de valeurs fini ou dénombrable (son ensemble de définition est inclus dans $\mathbb{N}$), on parle de variable discrète.

Définition 1.1.1 *Une variable aléatoire est dite de type discret si le nombre de valeurs différentes qu'elle peut prendre est fini ou infini dénombrable.*

On s'intéresse à définir l'ensemble des valeurs possibles et leurs probabilités associées.

Exemple 1.1.2 - *Résultat d'un jet de dé. Le résultat X est une variable aléatoire*

$$X(\Omega) = \{1, 2, 3, 4, 5, 6\}$$

- Lancer de 2 pièces de monnaies identiques dont l'issue est P (pour pile) et F (pour face). L'univers

$$\Omega = \{PP, PF, FP, FF\}$$

Variable quantitative continue: Une variable aléatoire est dite continue si elle peut prendre toutes les valeurs d'un intervalle. En particulier, dans le cas où la variable aléatoire peut prendre toute valeur réelle (son ensemble de définition contient un intervalle de $\mathbb{R}$), on parle de variable aléatoire réelle.

Définition 1.1.2 *Une variable aléatoire qui peut prendre un nombre infini non dénombrable de valeurs est dite variable aléatoire de type continu.*

Dans ce cas, il ne s'agira plus de calculer une probabilité d'apparition d'une valeur donnée mais d'un intervalle.

Exemple 1.1.3 *Considérons l'expérience aléatoire qui consiste à observer le temps T qu'une personne doit attendre à un guichet automatique avant de pouvoir s'en servir; la fonction T est une variable aléatoire continue puisque l'ensemble des valeurs possibles est l'intervalle $(0, \infty)$.*

Série statistique: On appelle série statistique la suite des valeurs prises par une variable X sur les unités d'observation. Le nombre d'unités d'observation est noté n.

Les valeurs de la variable X sont notées

$$x_1, x_2, ..., x_n$$

Exemple 1.1.4 *On s'intéresse à la variable 'état-civil' notée X et à la série statistique des valeurs prises par X sur 20 personnes. La codification est "C : célibataire, M : marié(e), V : veuf(ve), D : divorcée".*

Le domaine de la variable X est $\{C, M, V, D\}$. Considérons la série statistique suivante :

$$M - M - D - C - C - M - C - C - C - M - C - M - V - M - V - D - C - C - C - M$$

Ici, $n = 20$.

1.2 Tableau statistique

1.2.1 Effectif partiel (fréquence absolue)

Définition 1.2.1 *Le nombre d'individus qui ont le même x_i, ça s'appelle effectif partiel n_i de x_i.*

1.2.2 Effectif cumulé

Définition 1.2.2 *L'effectif cumulé N_i d'une valeur est la somme de l'effectif de cette valeur et de tous les effectifs des valeurs qui précèdent.*

Exemple 1.2.1 *Avec la série de l'exemple précédent, on obtient le tableau statistique:*

x_i	n_i	Δn_{ic}	Δn_{id}
C	*9*	*9*	*20*
M	*7*	*16*	*11*
V	*2*	*18*	*4*
D	*2*	*20*	*2*
N	*20*	/	/

1.2.3 Fréquence partielle (fréquence relative)

Définition 1.2.3 *Pour chaque valeur x_i, on pose par définition*

$$f_i = \frac{n_i}{N}$$

f_i s'appelle la fréquence partielle de x_i.

1.2.4 Fréquence cumulée

Définition 1.2.4 *Pour chaque valeur x_i, on pose par définition*

$$\Delta f_i = f_1 + f_2 + ... + f_i$$

La quantité Δf_i s'appelle la fréquence cumulée de x_i.

Exemple 1.2.2 *Avec la série de l'exemple précédent, on obtient le tableau statistique:*

x_i	f_i	Δf_{ic}	Δf_{id}
C	$\frac{9}{20} = 0.45$	*0.45*	*1*
M	*0.35*	*0.80*	*0.55*
V	*0.1*	*0.90*	*0.2*
D	*0.1*	*1*	*0.1*
N	*1*	/	/

1.3 Représentation graphique des séries statistiques

1.3.1 Variable qualitative

Le tableau statistique d'une variable qualitative peut être représenté par deux types de graphique. Les effectifs sont représentés par un diagramme en barres et les fréquences par un diagramme en secteurs.

Diagramme circulaire (diagramme en secteurs): Les diagrammes circulaires, consistent à partager un disque ou un demi-disque, en tranches, ou secteurs, correspondant aux modalités observées et dont la surface est proportionnelle à l'effectif, ou à la fréquence, de la modalité.

Diagramme en barres: À chaque marque correspond un bâton. Les hauteurs des bâtons sont proportionnelles aux effectifs.

1.3.2 Variable quantitative discrète

Diagramme en bâtonnets :Quand la variable est discrète, les effectifs sont représentés par des bâtonnets.

1.3.3 Variable quantitative continue

Histogramme: L'histogramme consiste à représenter les effectifs (resp. les fréquences) des classes par des rectangles contigus dont la surface représente l'effectif (resp. la fréquence). Pour un histogramme des effectifs, la hauteur du rectangle correspondant à la classe j.

1.4 Paramètres de position

Les indicateurs statistiques de tendance centrale (dits aussi de position) considérés fréquemment sont la moyenne, la médiane et le mode.

1.4.1 La moyenne

On appelle moyenne de X, la quantité

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} n_i.x_i$$

1.4.2 La médiane

On appelle médiane la valeur Me de la V.S X:

1^{ier} cas: Si $N = P \times 2$, on a

$$Me = \frac{X_P + X_{P+1}}{2}$$

$2^{ème}$ cas: Si $N = P \times 2 + 1$, on a

$$Me = X_P$$

1.4.3 Le mode

Le mode d'une V.S est la valeur qui a le plus grand effectif partiel (ou la plus grande fréquence partielle) et il est dénoté par Mo.

Exemple 1.4.1 *Une enquête réalisée dans un village porte sur le nombre d'enfants à charge par famille. On note X le nombre d'enfants, les résultats sont données par ce tableau:*

x_i	*0*	*1*	*2*	*3*	*4*	*5*	*6*
n_i	*18*	*32*	*66*	*41*	*32*	*9*	*2*

La moyenne est

$$\overline{X} = \frac{(0 \times 18) + (1 \times 32) + ... + (5 \times 9) + (6 \times 2)}{200} = ...$$

Le mode est $Mo = 2$;

La médiane, $N = 100 \times 2$ donc

$$Me = \frac{X_{100} + X_{101}}{2} = \frac{2+2}{2} = 2$$

1.5 Paramètres de dispersion

Les indicateurs statistiques de dispersion usuels sont l'étendue, la variance et l'écarttype.

1.5.1 L'étendue

La différence entre la plus grande valeur et la plus petite valeur du caractère, donnée par la quantité

$$e = x_{\max} - x_{\min}$$

s'appelle l'étendue de la V.S X.

1.5.2 La variance

On appelle variance d'une série statistique X,

$$Var\,(X) = \frac{1}{N}\sum_{i=1}^{n} n_i\,(x_i - \overline{x})^2 = \frac{1}{N}\sum_{i=1}^{n} n_i x_i^2 - \overline{x}^2$$

1.5.3 L'écart-type

La quantité

$$\delta_X = \sqrt{Var\,(X)}$$

s'appelle l'écart type de la V.S X.

CHAPITRE 2

Estimation des paramètres

Soit X une variable aléatoire associée à un certain phénomène aléatoire observable de façon répétée. La distribution exacte d'une variable X modélisant le caractère qui intéresse le statisticien est généralement partiellement connue. Souvent la loi de X dépend d'un paramètre inconnu. Notre objectif est de se faire une idée sur ce paramètre à partir des données observées sur un'échantillon issues de cette population.

Attribuer au paramètre une valeur numérique unique est une " **estimation ponctuelle**". Pour ce faire, on choisit une statistique dont la valeur est, après tirage aléatoire de l'échantillon, l'estimation du paramètre. Cette statistique est "**un estimateur**".

Plutôt que d'estimer un paramètre à l'aide d'un seul nombre, il arrive fréquemment que l'on fasse l'estimation en donnant un **intervalle** de valeurs. **Un intervalle d'estimation** (ou de **confiance**) est défini de telle sorte que l'on puisse affirmer avec un degré de confiance fixé que le paramètre visé se trouve dans cet intervalle.

Nous nous intéresserons dans ce chapitre à l'estimation des principales caractéristiques (ou paramètres) d'une variable alétoire dans une population, à savoir la moyenne, la variance et la fréquence (ou proportion), par **une estimation ponctuelle** et par **intervalle de confiance**.

2.1 Généralités

Considérons un échantillon de taille n extrait d'une population de taille N et pour laquelle on s'intéresse à un caractère X mesuré pour chaque individu de la population.

Le caractère X est considéré comme une variable aléatoire et l'échantillon de valeurs est constitué de n réalisations de cette variable.

On représente cette situation au moyen d'un modèle statistique qui comporte une famille de lois de probabilités parmi lesquelles se trouve la loi suivie par la variable X. Ces lois de probabilité dépendent en général d'un ou de plusieurs paramètres notés θ. Dans ce cas, on dit qu'on a un modèle statistique paramétrique.

Un des problèmes les plus courants en statistique consiste à trouver la valeur du ou des paramètres pour la population. Mais comme on ne peut pas en général avoir l'information nécessaire, on doit se contenter des valeurs fournies par l'échantillon.

On considère que chaque X_i est une variable aléatoire et on suppose qu'elles sont indépendantes entre elles. D'autre part, elles sont identiquement distribuées.

2.1.1 Estimateurs

Définition 2.1.1 *On appelle* ***estimateur*** *d'un paramètre θ d'une population, toute fonction*

$$T = f(X_1, X_2, ..., X_n)$$

et sa réalisation sera notée

$$t = f(x_1, x_2, ..., x_n)$$

Pour un même paramètre, il peut y avoir plusieurs estimateurs possibles; par exemple, Le paramètre λ d'une loi de Poisson admet comme estimateurs possibles la moyenne empirique et la variance empirique. Pour pouvoir choisir, il faut définir les qualités qui font qu'un estimateur sera meilleur.

On va voir en particulier les quantités empiriques les plus couramment utilisées :

1. La moyenne empirique.

2. La variance empirique.

3. La fréquence empirique.

La valeur empirique d'un paramètre est également une variable aléatoire car, non seulement, il est calculé à partir d'une variable aléatoire mais aussi d'un échantillon qui lui-même est aléatoirement choisi.

2.1.2 Qualité d'un estimateur

Définition 2.1.2 *On appelle* ***biais*** *d'un estimateur, la quantité*

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

qui représente l'erreur systématique.

Définition 2.1.3 *Un estimateur T de θ est dit* ***sans biais*** *si*

$$E(T) = \theta$$

Définition 2.1.4 *Un estimateur* ***sans biais*** *est dit* ***convergent*** *si*

$$V(T) \underset{n \longrightarrow \infty}{\longrightarrow} 0$$

Définition 2.1.5 *Soient T_1 et T_2 deux* ***estimateurs sans biais*** *de θ. T_1 est dit* ***plus efficace*** *que T_2 si*

$$V(T_1) \leq V(T_2)$$

2.2 Estimation ponctuelle

Estimer un paramètre, par exemple: une moyenne, une variance, une proportion, etc..., c'est chercher une valeur approchée en se basant sur les résultats d'un échantillon. Lorsqu'un paramètre est estimé par un seul nombre déduit des résultats de l'échantillon, ce nombre est appelé une estimation ponctuelle du paramètre.

2.2.1 Estimation de la moyenne

Soit X une variable alétoire dont on veut estimer la moyenne (ou espérance) $\mu = E(X)$ à partir d'un échantillon $(X_1, X_2, ..., X_n)$ de X (on ne suppose rien sur la loi de X).

Définition 2.2.1 *On appelle* moyenne empirique *de l'échantillon* $(X_1, X_2, ..., X_n)$ *de* X*, la statistique*

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

sa réalisation est

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

qui est la moyenne de l'échantillon aussi appelée moyenne observée*.*

Propriétés:

$$\begin{aligned} E\left(\overline{X}\right) &= \mu \\ V\left(\overline{X}\right) &= \frac{1}{n}\sigma^2 \end{aligned}$$

On a

$$E\left(\overline{X}\right) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\sum_{i=1}^{n} E(X) = E(X) = \mu$$

et

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

donc $\overline{X}$ est un estimateursans biais de $E\left(\overline{X}\right) = \mu$ et de plus il est convergent $V\left(\overline{X}\right) = \frac{V(X)}{n} \underset{n\longrightarrow\infty}{\longrightarrow} 0$, et $\forall T$, un autre estimateur de μ , $V(T) > V\left(\overline{T}\right)$.

Théorème central limite 1

Lorsque la variance σ^2 de la population est connue et que l'échantillon prélevé est grand ($n \geq 30$), alors la moyenne échantillonnale vérifie:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \rightsquigarrow N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Alors

$$Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \rightsquigarrow N(0,1)$$

Remarque:

- Le théorème précédent est vrai aussi lorsque la variance est connue, l'échantillon est petit et que la variable aléatoire X suit une loi normale $N(\mu, \sigma^2)$.
- Lorsque la variance σ^2 de la population est inconnue et que l'échantillon prélevé est grand ($n \geq 30$), alors

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \rightsquigarrow N\left(\mu, \frac{s}{\sqrt{n}}\right) \qquad \text{c'est à dire} \qquad Z = \frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}} \rightsquigarrow N(0,1)$$

Soit un lot de 500 chocolats. Le poids d'un chocolat est une v.a. telle que $\mu = 5g$ et $\sigma = 0.5g$. Quelle est la probabilité qu'une boîte de 50 chocolats issus de ce lot ait un poids total supérieur à $260g$?

L'échantillon étant grand ($n = 50 > 30$) donc on peut appliquer la formule:

$$\overline{X} \rightsquigarrow N\left(5; \frac{0.5}{\sqrt{50}}\right)$$

on pose

$$Y = 50\overline{X}$$

alors

$$Y \rightsquigarrow N\left(50 \times 5; 50 \times \frac{0.5}{\sqrt{50}}\right) \Rightarrow Y \rightsquigarrow N\left(250; 0.5 \times \sqrt{50}\right)$$

Ainsi

$$\begin{aligned} P(Y > 260) &= P\left(Z > \frac{260 - 250}{0.5 \times \sqrt{50}}\right) = P(U > 2.83) \\ &= 1 - P(U < 2.83) = 1 - 0.997 = 0.0023 \end{aligned}$$

Théorème central limite 2

Si la variance de la population est inconnue, si la variable X suit une distribution normale $N(\mu, \sigma^2)$, et si la taille de l'échantillon est petite ($n < 30$), alors

$$T = \frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}} \rightsquigarrow t_{(n-1)} \qquad \text{Loi de Student à } n-1 \text{ degrés de liberté ddl.}$$

Pour estimer le montant hebdomadaire moyen dépensé par les familles de 4 personnes pour leur épicerie, on tire un échantillon aléatoire de 25 personnes. On suppose que les montants dépensés sont distribués normalement avec une moyenne 120 et une variance inconnue. Si la variance de l'échantillon de taille 25 est 36, calculer la probabilité que la moyenne de l'échantillon soit supérieure à 123.

On a $n = 25 < 30$ et la variance de la population est *inconnue*, (on connaît la variance de l'échantillon $s^2 = 36$), donc

$$T = \frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}} = \frac{\overline{X} - 120}{\frac{6}{5}} \rightsquigarrow t_{25-1}$$

alors

$$\begin{aligned} P\left(\overline{X} > 123\right) &= P\left(T > \frac{123 - 120}{\frac{6}{5}}\right) \\ &= P\left(T > 2.5\right) \simeq P\left(T > 2.492\right) = 0.01 \end{aligned}$$

2.2.2 Estimation de la variance

Soit X une variable alétoire qui suit une loi normale $N(\mu, \sigma)$. On veut estimer la variance σ^2 de X.

Définition 2.2.2 *On appelle variance empirique de l'échantillon $(X_1, X_2, ..., X_n)$ de X, la statistique*

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 = \frac{1}{n}\left(\sum_{i=1}^{n} X_i^2\right) - \overline{X}^2$$

qui est la variance de l'échantillon, aussi appelée variance observée.

La variance empirique correspond donc à la moyenne des écarts à la moyenne empirique.

Propriétés

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

On a

$$\begin{aligned} E\left(S^2\right) &= E\left(\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2\right) = E\left(\frac{1}{n}\left(\sum_{i=1}^{n}X_i^2\right)-\overline{X}^2\right) \\ &= \frac{1}{n}E\left(\sum_{i=1}^{n}X_i^2\right)-E\left(\overline{X}^2\right) = \frac{1}{n}\sum_{i=1}^{n}E\left(X_i^2\right)-E\left(\overline{X}^2\right) \\ &= \frac{1}{n}\sum_{i=1}^{n}\left[V\left(X_i\right)+\left(E\left(X_i\right)\right)^2\right]-\frac{1}{n}\sigma^2-\mu^2 \\ &= V\left(X_i\right)+\left(E\left(X_i\right)\right)^2-\frac{1}{n}\sigma^2-\mu^2 = \sigma^2+\mu^2-\frac{1}{n}\sigma^2-\mu^2 \\ &= \left(1-\frac{1}{n}\right)\sigma^2 = \frac{n-1}{n}\sigma^2 \end{aligned}$$

On voit donc qu'à un coefficient près, l'espérance de la variance empirique est différente de la variance de la population. Cet estimateur est donc biaisé. D'où la nécessité de trouver un estimateur non biaisé. C'est là qu'intervient la notion de variance empirique modifiée.

Variance empirique modifiée

Soit S^{*2} la variance empirique modifiée. Elle se calcule comme suit

$$S^{*2} = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n}X_i^2\right)-\frac{n}{n-1}\overline{X}^2$$

On peut aisément montrer que :

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

et que

$$E\left(S^{*2}\right) = \sigma^2$$

Variance de la variance empirique S^2 :

L'expression de la variance de S^2 se présente comme suit :

$$Var\left(S^2\right) = \frac{n-1}{n^3}\sigma^2$$

Propriétés

Si $(X_1, ..., X_n)$ est un échantillon de variables gaussiennes (loi normale), alors les variables $\sqrt{n}\left(\frac{\overline{X}-\mu}{\sigma}\right)$ et $(n-1)\frac{S^{*2}}{\sigma^2}$ sont indépendantes et suivent respectivement la loi normale $N(0,1)$ et la loi de khi-deux à $n-1$ degrés de liberté.

Comme on a $S^{*2} = \frac{n}{n-1}S^2$ la propriété est aussi vraie pour $n\frac{S^2}{\sigma^2}$ qui suit une loi de khi-deux à n degrés de liberté.

On prélève 25 pièces dans une production industrielle. Une étude préalable a montré que le diamètre de ces pièces suivait une loi gaussienne de moyenne $10mm$ et d'écart-type $2mm$. Entre quelles valeurs a-t-on 85% de chances de trouver l'écart-type de ces pièces?

On pose

$$Y = \frac{n.S^2}{\sigma^2}$$

on détermine α et β tel que

$$\begin{aligned} P(\alpha < Y < \beta) &= P(Y < \beta) - P(Y < \alpha) \\ &= [1 - P(Y > \beta)] - [1 - P(Y > \alpha)] \\ &= P(Y > \alpha) - P(Y > \beta) \end{aligned}$$

on cherche dans la table du χ^2_{24} degrés de liberté les valeurs α et β tel que:

$$P(Y > \alpha) = 0.90 \quad \text{et} \quad P(Y > \beta) = 0.05$$

on trouve

$$\begin{cases} \alpha = 15.659 \\ \beta = 36.415 \end{cases}$$

alors

$$\begin{aligned} P\left(15.659 < \frac{25.S^2}{2^2} < 36.415\right) &= 0.85 \\ P\left(2.5054 < S^2 < 5.8264\right) &= 0.85 \\ P(1.58 < S < 2.41) &= 0.85 \end{aligned}$$

Lors d'un concours radiophonique, on note X: le npmbre de réponses reçues chaque jour. On suppose $X \sim N(\mu, \sigma)$. Durant 10 jours on a obtenu:

$$200 - 240 - 190 - 150 - 220 - 180 - 170 - 230 - 210 - 210$$

Donner une estimation ponctuelle de μ, σ^2.

On a $n = 10$

$$\overline{X} = \frac{1}{10}(X_1 + ... + X_{10})$$

est un estimateur de μ, sa réalisation

$$\overline{x} = \frac{1}{10}(x_1 + ... + x_{10}) = \frac{2000}{10} = 200$$

est une estimation ponctuelle de μ. on est dans le cas où la moyenne μ n'est pas connue, donc:

$$S^2 = \frac{1}{10}\left(X_1^2 + ... + X_{10}^2\right) - \overline{X}^2$$

est un estimateur biaisé de σ^2, sa réalisation

$$s^2 = \frac{1}{10}\left(x_1^2 + ... + x_{10}^2\right) - \overline{x}^2 = 40700 - 40000 = 700$$

est une estimation ponctuelle, biaisée de σ^2.

$$\delta^2 = \frac{n}{n-1}S^2 = \frac{10}{9}S^2$$

est un estimateur sans biais de σ^2, sa réalisation

$$\delta^2 = \frac{10}{9}700 = 778$$

est une estimation ponctuelle, sans biais de σ^2.

2.2.3 Estimation d'une proportion

Considérons une variable aléatoire qui suit une loi de Bernoulli, c'est-à-dire d'une variable aléatoire X qui ne peut prendre que deux valeurs 0 (échec) ou 1 (succès). Dans cette expérience, il s'agit d'étudier la probabilité de succès. On écrit $X \rightsquigarrow B(f)$.

Dans une expérience de Bernoulli, la moyenne empirique est appelée fréquence empirique. Elle représente l'estimateur du paramètre f. On la note F.

Définition 2.2.3 *La variable aléatoire*

$$F = \frac{K_n}{n}, \qquad \text{où } K_n \text{ représente le nombre de succès}$$

s'appelle fréquence empirique.

Loi de probabilité de F

L'espérance d'une loi de Bernoulli $B(f)$ est égale à f et la variance à $f(1-f)$. On déduit donc des calculs sur la moyenne empirique que :

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

en effet

$$E(F) = \frac{E(X_1) + ... + E(X_2)}{n} = f$$

et

$$V(F) = \frac{V(X_1) + ... + V(X_n)}{n^2} = \frac{nf(1-f)}{n^2} = \frac{f(1-f)}{n}$$

donc F est un estimateur sans biais de f.

dès que $n > 30$, $f \in [0.1; 0.9]$.

On trouve aussi

$$\frac{F-f}{\sqrt{\frac{f(1-f)}{n}}} \rightsquigarrow N(0,1)$$

avec les seules conditions $nf > 5$, $n(1-f) > 5$.

$f = 0.8$ est la proportion de Canadiens satisfaits du libre échange. Soit $n = 100$ personnes interrogées. Quelle est la probabilité que la proportion des personnes interrogées satisfaites du libre échange soit inférieure ou égale à 0.9?

$$Z = \frac{\overline{f} - f}{\sqrt{\frac{f(1-f)}{n}}} \sim N(0,1)$$

$$P\left(\overline{f} \leq 0.9\right) = P\left(Z \leq \frac{0.9 - 0.8}{\sqrt{\frac{0.8(1-0.8)}{100}}}\right) = P(Z \leq 2.5) = 0.9938.$$

Dans une population d'étudiants de UBMA, on a prélevé indépendamment deux échantillons de taille $n_1 = 120, n_2 = 150$. On constate que 48 étudiants du premier échantillon et 66 du deuxième ont une formation scientifique secondaire. Soit F la proportion d'étudiants ayant suivi une formation scientifique. Calculer 3 estimations ponctuelles de F.

On a

$$F = \frac{K}{n}$$

dons

$$F_1 = \frac{48}{120} = 0.4$$

$$F_2 = \frac{66}{150} = 0.44$$

$$F_3 = \frac{48 + 66}{120 + 150} = 0.422$$

2.3 Estimation par intervalle de confiance

Les estimations ponctuelles, bien qu'utiles, ne fournissent aucune information concernant la précision des estimations, c'est-à-dire qu'elles ne tiennent pas compte de l'erreur possible dans l'estimation dûe aux fluctuations d'échantillonnage. La théorie des intervalles de confiance (IC) consiste à construire, autour de l'estimation ponctuelle, un intervalle qui aura une grande probabilité $(1 - \alpha)$ de contenir la vraie valeur du paramètre.

Soit X une variable aléatoire dont la loi dépend d'un paramètre θ inconnu. Soit $(X_1, ..., X_n)$ un échantillon issu de X et $\alpha \in]0, 1[$.

Définition 2.3.1 *On appelle* ***intervalle de confiance*** *pour θ de niveau $1 - \alpha$ (ou de seuil α), un intervalle $[t_1, t_2]$ qui a la probabilité $1 - \alpha$ de contenir la vraie valeur de θ*

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

plus le niveau de confiance est élevé, plus la certitude est grande et que la méthode d'estimation produira une estimation contenant la vraie valeur de θ.

- Si on augmente le niveau de confiance $1-\alpha$, on augmente la longueur de l'intervalle.

2.3.1 Intervalle de confiance pour la moyenne

cas où n, la taille de l'échantillon, est petite $n < 30$

On suppose que $X \rightsquigarrow N(\mu, \sigma)$

Première cas: Lorsque la taille de l'échantillon est petite ($n < 30$) et $X \rightsquigarrow N(\mu, \sigma)$ de variance **inconnue**:

On a:

$$\frac{\overline{X} - \mu}{S\sqrt{n-1}} \rightsquigarrow t_{n-1} \qquad \text{la loi de student à } n-1 \text{ ddl}$$

On cherche dans la table de la loi de Student, α étant fixé, la valeur $t_{n-1;\left(1-\frac{\alpha}{2}\right)}$ telle que:

$$P\left(-t_{n-1;\left(1-\frac{\alpha}{2}\right)} < \frac{\overline{X} - \mu}{S\sqrt{n-1}} < t_{n-1;\left(1-\frac{\alpha}{2}\right)}\right) = 1 - \alpha$$

On a

$$P\left(\overline{X} - t_{n-1;\left(1-\frac{\alpha}{2}\right)} \cdot \frac{s^*}{\sqrt{n}} < \mu < \overline{X} + t_{n-1;\left(1-\frac{\alpha}{2}\right)} \cdot \frac{s^*}{\sqrt{n}}\right) = 1 - \alpha$$

Conclusion

si $\overline{x}$ est une réalisation de $\overline{X}$ et s une réalisation de S, l'intervalle de confiance de μ de seuil α est

$$IC_\mu = \left[\overline{x} - t_{n-1;\left(1-\frac{\alpha}{2}\right)} \cdot \frac{s^*}{\sqrt{n}}; \overline{x} + t_{n-1;\left(1-\frac{\alpha}{2}\right)} \cdot \frac{s^*}{\sqrt{n}}\right]$$

Un reporter pour un journal étudiant est en train de rédiger un article sur le coût du logement près du campus. Un échantillon de 10 appartements dans un rayon de 1 km de l'université a permis d'estimer le coût moyen du loyer mensuel à 350 par mois et un écart type de 30. Quel est l'intervalle de confiance de 95% pour la moyenne des loyers mensuels? Supposons que les loyers suivent une loi normale.

pour un coefficient de confiance de $1 - \alpha = 0,95$, on a $\alpha = 0.05$ et $\frac{\alpha}{2} = 0.025$. On a $n - 1 = 10 - 1 = 9$ degrés de liberté, alors la table de la distribution Student nous donne

$$t_{n-1;\left(1-\frac{\alpha}{2}\right)} = t_{9;0.975} = 2.262$$

Finalement

$$IC_\mu = [328.54; 371.46]$$

c'est-à-dire nous sommes confiants à 95% que la moyenne des loyers mensuels, se trouve entre 328.54 et 371.46.

Deuxième cas: Lorsque la taille de l'échantillon est petitee ($n < 30$) et la variance de la population de X est **connue**:

On a:

$$\overline{X} \rightsquigarrow N\left(\mu; \frac{\sigma}{\sqrt{n}}\right) \qquad \text{ou bien} \qquad \frac{\overline{X}-\mu}{\sigma\sqrt{n}} \rightsquigarrow N(0;1)$$

On se fixe le risque α et on cherche dans la table de la loi normale la valeur $u_{1-\frac{\alpha}{2}}$ (est le fractile d'ordre $1-\frac{\alpha}{2}$ de la loi normale centrée réduite) .telle que

$$P\left(-u_{1-\frac{\alpha}{2}} < \frac{\overline{X}-\mu}{\sigma\sqrt{n}} < u_{1-\frac{\alpha}{2}}\right) = 1-\alpha$$

$$P\left(\overline{X} - u_{1-\frac{\alpha}{2}}.\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + u_{1-\frac{\alpha}{2}}.\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

Conclusion

si $\overline{x}$ est une réalisation de $\overline{X}$, l'intervalle de confiance de μ de seuil α est

$$IC_\mu = \left[\overline{X} - u_{1-\frac{\alpha}{2}}.\frac{\sigma}{\sqrt{n}}; \overline{X} + u_{1-\frac{\alpha}{2}}.\frac{\sigma}{\sqrt{n}}\right]$$

Un échantillon de 11 observations fournit une moyenne d'échantillon de 42, et un écart-type d'échantillon de 9. On souhaite construire un intervalle de confiance pour la moyenne de la population au seuil de confiance de 90%. Quelle hypothèse doit-on faire sur la population ? Construire l'intervalle de confiance.

On suppose que $X \sim N(\mu, \sigma)$

L'intervalle de confiance de μ de seuil $\alpha = 10\%$ est

$$IC_\mu = \left[\overline{X} - u_{1-\frac{\alpha}{2}}.\frac{\sigma}{\sqrt{n}}; \overline{X} + u_{1-\frac{\alpha}{2}}.\frac{\sigma}{\sqrt{n}}\right]$$

donc pour $\alpha = 0.1$ on a $u_{1-\frac{\alpha}{2}} = u_{0.95} = 1.64$, alors

$$\begin{aligned} IC_\mu &= \left[\overline{X} - u_{0.95}.\frac{\sigma}{\sqrt{n}}; \overline{X} + u_{0.95}.\frac{\sigma}{\sqrt{n}}\right] \\ &= \left[42 - 1.64.\frac{9}{\sqrt{11}}; 42 + 1.64.\frac{9}{\sqrt{11}}\right] \\ &= [37.5497; 46.4503] \end{aligned}$$

cas où n, la taille de l'échantillon, est grande $n \geq 30$

Il n'est plus nécessaire de supposer que X est Gaussienne.

Premier cas: Lorsque la taille de l'échantillon est grande ($n \geq 30$) et $X \rightsquigarrow N(\mu, \sigma)$ de variance **connue**:

On a:

$$\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \rightsquigarrow N(0,1)$$

Si $\overline{x}$ est une réalisation de $\overline{X}$ et si s une réalisation de S, l'intervalle de confiance de μ de seuil α est

$$IC_\mu = \left[\overline{x} - u_{1-\frac{\alpha}{2}} . \frac{\sigma}{\sqrt{n}}; \overline{x} + u_{1-\frac{\alpha}{2}} . \frac{\sigma}{\sqrt{n}}\right]$$

On a observé la taille de $n = 200$ hommes marocains adultes. Après calcul , on a obtenu une moyenne de $\overline{x} = 168cm$. Si on suppose que la variance connue vaut $\sigma^2 = 1$. Donnez un intervalle de confiance à 95% de la vraie taille moyenne de la population.

Puisque

$$1 - \alpha = 0.95 \Rightarrow \alpha = 0.05$$

alors

$$\Phi(u_\alpha) = 1 - \frac{0.05}{2} = 0.975 \Rightarrow u_\alpha = 1.96$$

Finalement

$$IC_\mu = [167.86; 168.14]$$

c'est-à-dire

$$P(\mu \in [167.86; 168.14]) = 0.95$$

Deuxième cas: Lorsque la taille de l'échantillon est grande ($n \geq 30$) et la variance **inconnue**:

On a:

$$\frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}} \rightsquigarrow N(0,1)$$

On se fixe l'erreur α et on cherche dans la table de la loi normale la valeur $u_{1-\frac{\alpha}{2}}$ telle que

$$P\left(-u_{1-\frac{\alpha}{2}} < \frac{\overline{X}-\mu}{\frac{s}{\sqrt{n}}} < u_{1-\frac{\alpha}{2}}\right) = 1-\alpha$$

$$P\left(\overline{X} - u_{1-\frac{\alpha}{2}}.\frac{s^*}{\sqrt{n}} < \mu < \overline{X} + u_{1-\frac{\alpha}{2}}.\frac{s^*}{\sqrt{n}}\right) = 1-\alpha$$

si $\overline{x}$ est une réalisation de $\overline{X}$ et s une réalisation de S, l'intervalle de confiance de μ de seuil α est

$$IC_\mu = \left[\overline{x} - u_{1-\frac{\alpha}{2}}.\frac{s^*}{\sqrt{n}}; \overline{x} + u_{1-\frac{\alpha}{2}}.\frac{s^*}{\sqrt{n}}\right]$$

2.3.2 Intervalle de confiance pour la variance

On suppose que X est gaussienne.

D'après la section 2.2, on a

$$\frac{(n-1)\,S^2}{\sigma^2} = \sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2$$

Remarque.

Si la moyenne est μ inconnu, donc

$$\frac{(n-1)\,S^2}{\sigma^2} = \sum_{i=1}^{n}\left(\frac{X_i-\overline{X}}{\sigma}\right)^2$$

est une somme de n variable aléatoire indépendantes qui suivent la loi normale $N(0,1)$ et donc

$$\frac{(n-1)\,S^2}{\sigma^2} \rightsquigarrow \chi^2_{n-1}$$

On cherche les deux nombres a et b tels que

$$P\left(\frac{(n-1)\,S^2}{\sigma^2} \geq a\right) = 1-\frac{\alpha}{2}; \qquad P\left(\frac{(n-1)\,S^2}{\sigma^2} \geq b\right) = \frac{\alpha}{2}$$

L'intervalle de confiance au seuil $1-\alpha$ pour la variance inconnue σ^2 de la population est de la forme:

$$IC_{\sigma^2} = \left[\frac{(n-1)\,S^2}{b}; \frac{(n-1)\,S^2}{a}\right]$$

Dans une entreprise produisant un article déterminé on veut estimer sa durée de vie en heures. À cette fin on a observé un échantillon aléatoire et simple de 16 unités dont les résultats sont (en 1000 heures)

1.10	1.05	1.25	1.08	1.35	1.15	1.30	1.25
1.30	1.35	1.15	1.32	1.05	1.25	1.10	1.15

- Déterminer un intervalle de confiance pour la variance à 95%.

1. L'estimation ponctuelle de la moyenne et la variance:

   On a

   $$\overline{x} = \frac{1}{16}\sum_{i=1}^{16} n_i.x_i = 1.2$$

   et

   $$S^{*2} = \frac{1}{15}\sum_{i=1}^{16} n_i.\left(x_i - \overline{x}\right)^2 = 0.0121$$

   alors

   $$s^* = 0.11$$

2. On cherche les deux nombres a et b tels que

   $$P\left(\frac{(n-1)\,S^2}{\sigma^2} \geq a\right) = 0.025; \qquad P\left(\frac{(n-1)\,S^2}{\sigma^2} \geq b\right) = 0.975$$

   on trouve

   $$a = 6.26; b = 27.49$$

L'intervalle de confiance au seuil 95% pour la variance inconnue σ^2 de la population est de la forme:

$$\begin{aligned} IC_{\sigma^2} &= \left[\frac{(n-1)\,S^2}{b}; \frac{(n-1)\,S^2}{a}\right] \\ &= \left[\frac{15.0.11^2}{27.49}; \frac{15.0.11^2}{6.26}\right] \\ &= [0.0066; 0.0289] \end{aligned}$$

2.3.3 Intervalle de confiance pour la proportion

On a

$$f = \frac{K}{n}$$

est le meilleur estimateur de F où F est la proportion de la population possédant le caractère considéré.

On cherche dans la table de $N(0,1)$ la valeur $u_{1-\frac{\alpha}{2}}$ telle que:

$$P\left(-u_{1-\frac{\alpha}{2}} < \frac{F-f}{\sqrt{\frac{f(1-f)}{n}}} < -u_{1-\frac{\alpha}{2}}\right) = 1-\alpha$$

$$P\left(f - u_{1-\frac{\alpha}{2}}\sqrt{\frac{f(1-f)}{n}} < F < f + u_{1-\frac{\alpha}{2}}\sqrt{\frac{f(1-f)}{n}}\right) = 1-\alpha$$

L'intervalle de confiance au seuil $1-\alpha$ pour la proportion f de la population est de la forme:

$$IC_f = \left[f - u_{1-\frac{\alpha}{2}}\sqrt{\frac{f(1-f)}{n}}; f + u_{1-\frac{\alpha}{2}}\sqrt{\frac{f(1-f)}{n}}\right]$$

SPI est une compagnie qui se spécialise dans les sondages politiques. A l'aide de sondages téléphoniques, les interviewers demandent aux citoyens pour qui ils voteraient si les élections avaient lieu aujourd'hui. Récemment, SPI a trouvé que 220 votants sur 500 voterait pour un candidat particulier. SPI veut estimer l'intervalle de confiance à 95% pour la proportion des votants qui sont en faveur de ce candidat.

On a

$$n = 500 \quad \text{et} \quad f = \frac{220}{500} = 0.44$$

et

$$u_\alpha = 1.96$$

donc

$$IC_f = [0.3965; 0.4835]$$

Alors, SPI est confiant à 95% que la proportion des votants qui favoriseront ce candidat est entre 0.3965 et 0.4835.

CHAPITRE 3

Tests d'hypothèses

3.1 Généralités

Un test statistique est une procédure de décision qui permet de choisir entre deux hypothèses (contraires) faites sur une population ou sur un ou plusieurs paramètres au vu d'un échantillon aléatoire. On formule une hypothèse de départ, appelée hypothèse nulle et souvent notée H_0 et il s'agit de décider si on rejette ou non cette hypothèse par opposition à une contre hypothèse appelée hypothèse alternative et souvent notée H_1.

Pour effectuer le test statistique, il faudra choisir un certain risque d'erreur qui est la probabilité de se tromper en prenant la décision retenue. Il existe deux types d'erreur :

- On appelle erreur de première espèce ou erreur de type I, notée α, la probabilité de rejeter H_0 alors qu'elle est vraie. α est aussi appelé niveau ou seuil de signification.
- On appelle erreur de deuxième espèce ou erreur de type II, notée β, la probabilité d'accepter H_0 alors qu'elle est fausse.
- On appelle puissance du test pour H_1 la probabilité de retenir H_1 alors qu'elle est vraie $(1 - \beta)$

Les étapes de construction d'un test statistique

1. Il s'agit d'abord de formuler les hypothèses H_0 et H_1. On choisit en général le risque de type I , α. (souvent donné dans l'énoncé).

2. Détermination de la variable de décision: on détermine la variable de décision Z (qui est une statistique) dont on connaît la loi si H_0 est vraie.

3. On calcule la région critique ou région de rejet W qui est l'ensemble des valeurs de Z qui conduiront à rejeter H_0. Ainsi, si α est fixé, W est déterminé par $\alpha = P(Z \in W \text{ avec } H_0 \text{ vraie})$. Le complémentaire de W est appelé "région d'acceptation". Les points de jonction entre les deux régions sont les points critiques.

4. On calcule la valeur de Z à partir de l'observation de l'échantillon.

5. Conclusion du test : acceptation ou rejet de H_0 selon que la valeur de Z est ou non dans la région d'acceptation.

En fonction de l'hypothèse testée, plusieurs types de tests peuvent être réalisés :

- **Les tests de conformité**: sont destinés à vérifier si un échantillon peut être considéré comme extrait d'une population donnée ou représentatif de cette population, vis-à-vis d'un paramètre comme la moyenne, la variance ou la fréquence observée. Ceci implique que la loi théorique du paramètre est connue au niveau de la population.

- **Les tests d'homogénéité:** sont destinés à comparer plusieurs populations à l'aide d'un nombre équivalent d'échantillons. Dans ce cas la loi théorique du paramètre est inconnue au niveau des populations.

3.2 Tests de conformité

3.2.1 Construction générale

Soit X une variable aléatoire dont la loi dépend d'un paramètre inconnu θ.

- Tester l'hypothèse $H_0 : \ \theta = \theta_0$, θ_0 étant une valeur numérique; contre $H_1 : \ \theta \neq \theta_0$

- Choix de la variable de décision Z qui est l'estimateur de θ ou une fonction simple de l'estimateur de θ.

- Calcul de la région critique :

$$\alpha = P\,(\text{décider } H_1 \text{ alors que } H_0 \text{ est vraie}) \Longleftrightarrow \alpha = P\,(Z \in W \text{ alors que } \theta = \theta_0)$$

3.2.2 Comparaison d'une moyenne observée à une moyenne théorique

Soit X, une variable aléatoire observée sur une population, suivant une loi normale et un échantillon extrait de cette population.

Le but est de savoir si un échantillon de moyenne $\overline{x}$, estimateur 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).

L'hypothèse testée est la suivante:

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

Pour tester cette hypothèse, il existe deux statistiques : la variance σ^2 de la population de référence est connue ou cette variance est inconnue et il faut l'estimer.

Premier cas: la variance σ^2 de la population de référence est connue

On calcule la valeur

$$Z_{obs} = \frac{|\overline{x} - \mu_0|}{\sqrt{\frac{\sigma^2}{n}}}$$

On détermine Z_{tab} lue sur la table de la loi normale centrée réduite pour un risque d'erreur α fixé, et on décide que:

1. si $Z_{obs} > Z_{tab}$ 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 .

2. si $Z_{obs} \leq Z_{tab}$ l'hypothèse H_0 est acceptée: l'échantillon est représentatif de la population de référence d'espérance μ_0.

Deuxième cas: la variance σ_0^2 de la population de référence est inconnue

On calcule la valeur

$$T_{obs} = \frac{|\overline{x} - \mu_0|}{\sqrt{\frac{s^2}{n}}}$$

On détermine T_{tab} lue dans la table de Student pour un risque d'erreur α fixé et $(n-1)$ degrés de liberté, et on décide que:

1. si $T_{obs} > T_{tab}$ 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 .

2. si $T_{obs} \leq T_{tab}$ l'hypothèse H_0 est acceptée: l'échantillon est représentatif de la population de référence d'espérance μ_0.

Remarque: Si $n < 30$, la variable aléatoire X étudiée doit impérativement suivre une loi normale $N(\mu, \sigma)$. Pour $n \geq 30$, la variable de student T converge vers une loi normale centrée réduite Z.

Le diamètre des billes fabriquées par une machine est en moyenne de 6 mm. Pour contrôler si la machine est bien réglée, on a prélevé un échantillon de 50 billes et on a mesuré leur diamètre. On a trouvé :

$$\sum x_i = 350; \qquad \sum x_i^2 = 2462$$

La machine est-elle bien réglée au seuil de signification de 95 %?

Pour répondre à cette question, on doit vérifier si le diamètre moyen des 50 billes observées, est conforme à la norme de 6 mm. Il s'agit donc de faire un test de conformité de la moyenne. Donc

$$\begin{cases} H_0 : \mu = 6 \\ H_1 : \mu \neq 6 \end{cases}$$

On calcule la valeur

$$T_{obs} = \frac{|\overline{x} - \mu_0|}{\sqrt{\frac{s^2}{n}}}$$

Alors

$$\overline{x} = \frac{350}{50} = 7$$

et

$$S^2 = \frac{50}{49}\left(\frac{2462}{50} - 7^2\right) = 0.24$$

Donc

$$T_{obs} = \frac{|7-6|}{\sqrt{\frac{0.24}{50}}} = 14.43$$

Au seuil de signification de 95 %, on a

$$\alpha = 5\% \Rightarrow T_{tab} = 1.677$$

On a $T_{obs} > T_{tab}$, alors " l'hypothèse H_0 est rejetée au risque d'erreur 5% : l'échantillon appartient à une population d'espérance 7 et n'est pas représentatif de la population de référence d'espérance 6 .

3.2.3 Comparaison d'une variance observée à une variance théorique

Soit un échantillon $(X_1, ..., X_n)$ issu d'une population de loi normale, de moyenne μ et de variance σ^2.

Le but est de savoir si un échantillon de moyenne s^2 , estimateur de σ^2, appartient à une population de référence connue de variance σ_0^2 (H_0 vraie) et ne diffère de σ_0^2 que par des fluctuations d'échantillonnage ou bien appartient à une autre population inconnue de variance σ^2 (H_1 vraie).

L'hypothèse testée est la suivante:

$$\begin{cases} H_0 : \sigma^2 = \sigma_0^2 \\ H_1 : \sigma^2 \neq \sigma_0^2 \end{cases}$$

Pour tester cette hypothèse, il existe deux statistiques : la moyenne μ de la population de référence est connue ou cette variance est inconnue et il faut l'estimer.

Premier cas: la moyenne μ de la population de référence est connue

Lorsque la moyenne μ est connue, la statistique T^2 est la meilleure estimation de la variance

$$T^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$$

Sous l'hypothèse H_0, comme l'échantillon est gaussien, on a la statistique

$$y^2 = \frac{nT^2}{\sigma_0^2} = \sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma_0}\right)^2$$

suit une loi du χ^2 à n degrés de libertés. (en tant que somme de carrés de $N(0,1)$)

On cherche les deux nombres tabulés a et b tels que:

$$P\left(Y^2 \geq a\right) = \frac{\alpha}{2}; P\left(Y^2 \geq b\right) = 1 - \frac{\alpha}{2}$$

pour un risque d'erreur α fixé, et on décide que:

1. si $y^2 \in]a; b[$ l'hypothèse H_0 est acceptée au risque d'erreur α : l'échantillon est représentatif de la population de référence de variance σ_0 .

2. sinon l'hypothèse H_0 est rejetée: l'échantillon n'est pas représentatif de la population de référence de variance σ_0.

Deuxième cas: la moyenne μ de la population de référence est inconnue

Lorsque la moyenne μ est inconnue, la statistique S^2 est la meilleure estimation de la variance

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$

Sous l'hypothèse H_0, comme l'échantillon est gaussien, on a la statistique

$$y^2 = \frac{(n-1)\,S^2}{\sigma_0^2} = \sum_{i=1}^{n}\left(\frac{X_i - \overline{X}}{\sigma_0}\right)^2$$

suit une loi du χ^2 à $n-1$ degrés de liberté, (en tant que somme de carrés de $N(0,1)$).

On cherche les deux nombres tabulés a et b tels que:

$$P\left(Y^2 \geq a\right) = \frac{\alpha}{2}; P\left(Y^2 \geq b\right) = 1 - \frac{\alpha}{2}$$

pour un risque d'erreur α fixé, et on décide que:

1. si $y^2 \in]a;b[$ l'hypothèse H_0 est acceptée au risque d'erreur α : l'échantillon est représentatif de la population de référence de variance σ_0 .

2. sinon l'hypothèse H_0 est rejetée: l'échantillon n'est pas représentatif de la population de référence de variance σ_0.

On souhaite vérifier, au seuil de signification de 95 %, si le peuplement dans lequel on a mesuré la hauteur d'un échantillon de 12 arbres, appartient à un type de forêt dont l'écart type est de 1,4 m. Les résultats en mètre sont :

$$5,1 - 5,2 - 5,2 - 5,4 - 5,9 - 6,3 - 6,3 - 6,8 - 6,9 - 6,9 - 7,0 - 7,0$$

Pour répondre à cette question, on doit réaliser un test de conformité de la variance.

$$\begin{cases} H_0 : \sigma^2 = 1.4^2 = 1.96 \\ H_1 : \sigma^2 \neq 1.96 \end{cases}$$

comme l'échantillon est gaussien, on a la statistique

$$y^2 = \sum_{i=1}^{n} \left(\frac{X_i - \overline{X}}{\sigma_0} \right)^2 = \frac{6.6}{1.96} = 3.37$$

suit une loi du χ^2 à 11 degrés de liberté.

Au seuil de signification de 95 % , on a

$$\chi^2_{0.025} = 3.82 \quad \text{et} \quad \chi^2_{0.975} = 21.9$$

On a $y^2 \notin]a;b[$, alors "l'hypothèse H_0 est rejetée: l'échantillon n'est pas représentatif de la population de référence de variance 1.96".

3.2.4 Comparaison d'une fréquence observée à une fréquence théorique

Soit X une variable qualitative prenant deux modalités (succès $X = 1$, échec $X = 0$) observée sur une population et un échantillon extrait de cette population.

Le but est de savoir si un échantillon de fréquence observée $\frac{K}{n}$, estimateur de f, appartient à une population de référence connue de fréquence f_0 (H_0 vraie) ou à une autre population inconnue de fréquence f (H_1 vraie).

L'hypothèse testée est la suivante:

$$\begin{cases} H_0 : f = f_0 \\ H_1 : f \neq f_0 \end{cases}$$

On calcule la valeur:

$$Z_{obs} = \frac{\left|\frac{K}{n} - f_0\right|}{\sqrt{\frac{f_0(1-f_0)}{n}}}$$

suit la loi $N(0,1)$.

On détermine Z_{tab} lue sur la table de la loi normale centrée réduite pour un risque d'erreur α fixé, et on décide que:

1. si $Z_{obs} > Z_{tab}$ l'hypothèse H_0 est rejetée au risque d'erreur α: l'échantillon appartient à une population de fréquence f et n'est pas représentatif de la population de référence de fréquence f_0.

2. si $Z_{obs} \leq Z_{tab}$ l'hypothèse H_0 est acceptée: l'échantillon est représentatif de la population de référence de fréquence f_0.

Sur un échantillon de taille $n = 400$ de naissances, on a observé 206 mâles, soit une proportion de mâles de $f = \frac{206}{400} = 0.515$. On se demande s'il y a autant de mâles que de femelles, i.e., si $f_0 = 0.5$. On peut effectuer alors le test

$$\begin{cases} H_0 : f = 0.5 \\ H_1 : f \neq 0.5 \end{cases}$$

On calcule alors

$$Z_{obs} = \frac{\left|\frac{K}{n} - f_0\right|}{\sqrt{\frac{f_0(1-f_0)}{n}}} = \frac{|0.515 - 0.5|}{\sqrt{\frac{0.5(1-0.5)}{400}}} = 0.6$$

On a

$$\alpha = 5\% \Rightarrow Z_{tab} = 1.96$$

Comme $Z_{obs} \leq Z_{tab}$, alors " l'hypothèse H_0 est acceptée: l'échantillon est représentatif de la population de référence de fréquence 0.5".

3.3 Tests d'homogénéité

3.3.1 Construction générale

On considère deux variables aléatoires X_1 et X_2 définies sur deux populations P_1 et P_2 respectivement. Ces variables aléatoires dépendent d'un paramètre inconnu θ_1 et θ_2 respectivement.

- Tester l'hypothèse $H_0 : \ \theta_1 = \theta_2$, contre $H_1 : \ \theta_1 \neq \theta_2$
- On choisit le risque α.

  On dispose d'un échantillon de X_1 et d'un échantillon de X_2 qui fournissent respectivement T_1 un estimateur de θ_1 et T_2 un estimateur de θ_2.
- On détermine la variable de décision Z qui est une fonction de T_1 et T_2, et dont on connaît la loi de probabilité si H_0 est vraie.
- α étant connu, on calcule la région critique ou la région d'acceptation.
- On calcule la valeur z de Z à partir des résultats des échantillons.

3.3.2 Tests de comparaison de deux moyennes

Soit X un caractère quantitatif continu observé sur deux populations suivant une loi normale et deux échantillons indépendants extraits de ces deux populations.

On fait l'hypothèse que les deux échantillons proviennent de deux populations dont les espérances μ_1 et μ_2 sont égales.

$$\begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 \neq \mu_2 \end{cases}$$

Pour tester cette hypothèse, il existe deux statistiques : la variance σ^2 de la population de référence est connue ou cette variance est inconnue et il faut l'estimer.

Premier cas: les variances des population σ_1^2 et σ_2^2 sont connues

Sous l'hypothèse H_0 avec σ_1^2 et σ_2^2 sont connues, on a

$$\frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \approx N(0,1) \qquad \text{si} \quad \sigma_1^2 \text{ et } \sigma_2^2 \text{ sont connues}$$

On calcule

$$z_{obs} = \frac{|\overline{x}_1 - \overline{x}_2|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

On détermine Z_{tab} lue sur la table de la loi normale centrée réduite pour un risque d'erreur α fixé, et on décide que:

1. si $Z_{obs} > Z_{tab}$ l'hypothèse H_0 est rejetée au risque d'erreur α : les deux échantillons sont extraits de deux populations ayant des espérances respectivement μ_1 et μ_2.

2. si $Z_{obs} \leq Z_{tab}$ l'hypothèse H_0 est acceptée: les deux échantillons sont extraits de deux populations ayant même espérance μ.

Deuxième cas: les variances des population σ_1^2 et σ_2^2 sont inconnues et égales

Sous l'hypothèse H_0 avec $\sigma_1^2 = \sigma_2^2 = \sigma^2$, on a

$$\frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \text{ suit une loi de Student à } (n_1 + n_2 - 2) \text{ degrés de liberté}$$

La variance commune σ^2 peut être estimée par:

$$s^2 = \frac{n_1.s_1^2 + n_2.s_2^2}{n_1 + n_2 - 2}$$

On calcule

$$t_{obs} = \frac{|\overline{x}_1 - \overline{x}_2|}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

On détermine t_{tab} lue sur la table de Student pour un risque d'erreur α fixé et $(n_1 + n_2 - 2)$ degrés de liberté, et on décide que:

1. si $t_{obs} > t_{tab}$ l'hypothèse H_0 est rejetée au risque d'erreur α : les deux échantillons sont extraits de deux populations ayant des espérances respectivement μ_1 et μ_2.

2. si $t_{obs} \leq t_{tab}$ l'hypothèse H_0 est acceptée: les deux échantillons sont extraits de deux populations ayant même espérance μ.

Troisième cas: les variances des population σ_1^2 et σ_2^2 sont inconnues et inégales

Si les variances des populations ne sont pas connues et si leurs estimations à partir des échantillons sont significativement différentes, il faut considérer deux cas de figure selon la taille des échantillons comparés :

Cas où n_1 et $n_2 > 30$

La statistique utilisée est la même que pour le cas où les variances sont connues.

On calcule

$$z_{obs} = \frac{|\overline{x}_1 - \overline{x}_2|}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

On détermine Z_{tab} lue sur la table de la loi normale centrée réduite pour un risque d'erreur α fixé, et on décide que:

1. si $Z_{obs} > Z_{tab}$ l'hypothèse H_0 est rejetée au risque d'erreur α : les deux échantillons sont extraits de deux populations ayant des espérances respectivement μ_1 et μ_2.

2. si $Z_{obs} \leq Z_{tab}$ l'hypothèse H_0 est acceptée: les deux échantillons sont extraits de deux populations ayant même espérance μ.

Pour savoir s'il existe une différence d'assiduité entre les filles et les garçons, on a choisi de manière aléatoire et simple un premier échantillon de 10 filles et de façon indépendante, un deuxième échantillon de 10 garçons. En fonction des résultats ci-dessous relatifs aux notes d'assiduités (note sur 100), et en supposant que les variances des deux populations sont égales, peut-on conclure, au seuil de 5 %, à l'existence d'une différence significative entre les deux sexes ?

Assiduité des filles	72	67	52	54	46	58	59	54	58	63
Assiduité des garçons	66	59	54	57	63	55	61	55	66	75

Ce test a pour but de vérifier si l'assiduité moyenne μ_1 des filles est ou n'est pas égale à l'assiduité moyenne μ_2 des garçons.

$$\begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 \neq \mu_2 \end{cases}$$

Les deux échantillons sont indépendants, les populations sont de variances égales, alors:

$$t_{obs} = \frac{|\overline{x}_1 - \overline{x}_2|}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

On a

$$\overline{x}_1 = 58.3; \overline{x}_2 = 61.1; s^2 = 50.2778$$

donc

$$t_{obs} = 0.88$$

Pour $\alpha = 5\%$

$$t_{0.975;18} = 2.101$$

Alors $t_{obs} < t_{0.975;18}$: " l'hypothèse H_0 est acceptée: il n'y a pas de différence significative entre l'assiduité des deux sexes".

3.3.3 Tests de comparaison de deux variances

Soit X, une variable aléatoire observée sur deux populations suivant une loi normale et deux échantillons indépendants extraits de ces deux populations.

On fait l'hypothèse que les deux échantillons proviennent de deux populations dont les variances sont égales.

$$\begin{cases} H_0 : \sigma_1^2 = \sigma_2^2 \\ H_1 : \sigma_1^2 \neq \sigma_2^2 \end{cases}$$

Dans l'échantillon E_1 de taille n_1 (resp. l'échantillon E_2 de taille n_2), on estime la variance σ_1^2 (resp. σ_2^2) par:

$$s_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(x_i - \overline{x}_1)^2 \quad \text{et} \quad s_2^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n2}(x_i - \overline{x}_2)^2$$

On calcule

$$F_{obs} = \frac{s_1^2}{s_2^2}$$

telle que $F_{obs} \geq 1$ sinon on permute les échantillons de sorte que $F_{obs} \geq 1$.

On détermine F_{tab} lue sur la table de Fisher Snédécor pour un risque d'erreur α fixé avec $(n_1 - 1, n_2 - 1)$ degrés de liberté, on cherche F_{tab} tel que $P(F \geq F_{tab}) = \frac{\alpha}{2}$, et on décide que:

1. si $F_{obs} > F_{tab}$ l'hypothèse H_0 est rejetée au risque d'erreur α : les deux échantillons sont extraits de deux populations ayant des variances statistiquement différentes σ_1^2 et σ_2^2.

2. si $F_{obs} \leq F_{tab}$ l'hypothèse H_0 est acceptée: les deux échantillons sont extraits de deux populations ayant même variance σ^2.

Pour savoir si les filles sont plus assidues que les garçons ou non, on a choisi de manière aléatoire et simple un premier échantillon de 10 filles et de façon indépendante, un deuxième échantillon de 10 garçons. En fonction des résultats ci-dessous relatifs aux notes d'assiduités (note sur 100), peut-on supposer, au seuil de 5 %, que les variances des deux populations sont égales ?

Assiduité des filles	72	67	52	54	46	58	59	54	58	63
Assiduité des garçons	66	59	54	57	63	55	61	55	66	75

Ce test a pour but de vérifier si la variance σ_1^2 de la population des filles est ou n'est pas égale à la variance σ_2^2 de la population des garçons.

$$\begin{cases} H_0 : \sigma_1^2 = \sigma_2^2 \\ H_1 : \sigma_1^2 \neq \sigma_2^2 \end{cases}$$

Dans l'échantillon E_1 de taille 10 (resp. l'échantillon E_2 de taille 10), on estime la variance σ_1^2 (resp. σ_2^2) par:

$$s_1^2 = \frac{1}{9}\sum_{i=1}^{n_1} (x_i - \overline{x}_1)^2 \quad \text{et} \quad s_2^2 = \frac{1}{9}\sum_{i=1}^{n2} (x_i - \overline{x}_2)^2$$

alors

$$s_1^2 = 57.12 \quad \text{et} \quad s_2^2 = 43.43$$

On calcule

$$F_{obs} = \frac{57.12}{43.43} = 1.31$$

F_{tab} lue sur la table de Fisher Snédécor pour un risque d'erreur $\alpha = 0.05$ avec $(9, 9)$ degrés de liberté, $F_{tab} = F_{0.975} = 4.03$.

On a, $F_{obs} \leq F_{tab}$ "l'hypothèse H_0 est acceptée: les deux échantillons sont extraits de deux populations ayant même variance σ^2".

3.3.4 Tests de comparaison de deux fréquences

Soit X une variable qualitative prenant deux modalités (succès $X = 1$, échec $X = 0$) observée sur deux populations et deux échantillons indépendants extraits de ces deux populations. On fait l'hypothèse que les deux échantillons proviennent de deux populations dont les probabilités de succès sont identiques.

$$\begin{cases} H_0 : F_1 = F_2 \\ H_1 : F_1 \neq F_2 \end{cases}$$

Dans l'échantillon E_1 de taille n_1 on estime la fréquence F_1 par f_1 et dans l'échantillon E_2 de taille n_2 on estime la fréquence F_2 par f_2 et en regroupant les deux échantillons, on peut estimer F par:

$$f = \frac{n_1.f_1 + n_2.f_2}{n_1 + n_2}$$

Remarque: conditions de validité du test: $n_1.f_1 \geq 5; n_1(1 - f_1) \geq 5; n_2.f_2 \geq 5; n_2(1 - f_2) \geq 5$.

On calcule

$$z_{obs} = \frac{|f_1 - f_2|}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) f(1 - f)}}$$

On détermine z_{tab} pour la loi normale $N(0, 1)$ $\left(\Phi(z_{tab}) = 1 - \frac{\alpha}{2}\right)$, et on décide que:

1. si $z_{obs} > z_{tab}$ l'hypothèse H_0 est rejetée au risque d'erreur α : les deux échantillons sont extraits de deux populations ayant des probabilités de succès respectivement F_1 et F_2.

2. si $z_{obs} \leq z_{tab}$ l'hypothèse H_0 est acceptée: les deux échantillons sont extraits de deux populations ayant même probabilité de succès F.

Une enquête sur l'emploi a concerné 220 personnes dont 115 dans le milieu rural et 105 dans le milieu urbain. Sur les 115 ruraux enquêtés, 74 se sont révélés actifs, alors que pour les enquêtés urbains, 81 sont actifs. Peut-on admettre, au seuil de 5 %, qu'il n'y a pas de différence significative entre les taux d'activité dans les deux milieux ?

Ce test a pour but de vérifier si la proportion F_1 des personnes actives dans le milieu rural est ou n'est pas égale à la proportion F_2 des personnes actives dans le milieu urbain.

$$\begin{cases} H_0 : F_1 = F_2 \\ H_1 : F_1 \neq F_2 \end{cases}$$

Dans l'échantillon E_1 de taille 115 on estime la fréquence F_1 par f_1 et dans l'échantillon E_2 de taille 105 on estime la fréquence F_2 par f_2 et en regroupant les deux échantillons, on peut estimer F par:

$$f = \frac{n_1.f_1 + n_2.f_2}{n_1 + n_2}$$

On a

$$f_1 = \frac{74}{115} = 0.64; f_2 = \frac{81}{105} = 0.77; f = \frac{155}{220} = 0.70$$

On calcule

$$z_{obs} = \frac{|0.64 - 0.77|}{\sqrt{\left(\frac{1}{115} + \frac{1}{105}\right) 0.70\,(1 - 0.70)}} = 2.10$$

Pour $\alpha = 5\%$, $z_{tab} = z_{0.975} = 1.96$, alors $z_{obs} \geq z_{tab}$ "l'hypothèse H_0 est rejeté: les deux échantillons sont extraits de deux populations ayant même probabilité de succès F".

3.4 Test du Chi-deux d'indépendance

Le test du χ^2 d'indépendance a pour objectif d'évaluer si deux variables qualitatives X_1 et X_2 à respectivement p et k modalités sont liées, les deux variables étant observées sur un échantillon de taille N.

Les hypothèses du test du χ^2 d'indépendance sont les suivantes :

- H_0 : Les variables X_1 et X_2 sont indépendantes.
- H_1 : Il existe une liaison entre X_1 et X_2.

3.4.1 Calcul de la statistique de test:

Considérons, le tableau de contingence des effectifs observés suivant :

*Tab.*01 Tableau de contingence pour le test d'indépendance

Variable X_1	Variable X_2					Total
Modalités	M_1	M_2	.	.	M_k	t
M_1	e_{11}	e_{12}	.	.	e_{1k}	t_1
M_2	e_{21}	e_{22}	.	.	e_{2k}	t_2
.	.	.	.	.	.	.
.	.	.	.	.	.	.
M_k	e_{p1}	e_{p2}	.	.	e_{pk}	t_k
Total	n_1	n_2	.	.	n_k	N

Le principe du test du χ^2 consiste à calculer, pour chaque case du tableau, l'effectif théorique qui devrait être observé sous l'hypothèse nulle. Sous cette hypothèse, les effectifs sont répartis en proportion égale.

On définit l'effectif théorique E_{ij} associé à la case $\{i, j\}$ du tableau par la quantité suivante:

$$E_{ij} = \frac{n_j \times t_i}{N}$$

Sous l'hypothèse nulle, les effectifs observés et les effectifs théoriques doivent être sensiblement proches, donc la somme de leurs différences devrait être proche de zéro. Aussi, le principe du test du χ^2 se base sur l'évaluation de la somme de ces différences par rapport à une valeur seuil. Intuitivement, si cette somme de différences excède une certaine valeur, cela signifie que les effectifs observés et les effectifs théoriques sont différents et par conséquent l'hypothèse peut être remise en cause.

Sous H_0, le test du χ^2 a pour statistique de test:

$$\chi^2_{obs} = \sum_{i=1}^{p}\sum_{j=1}^{k}\frac{(e_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2_{(p-1)(k-1)}$$

Condition d'application du test

Le test du χ^2 est sensible aux petits effectifs. Aussi, le test est considéré comme applicable lorsque les effectifs théoriques E_{ij} sont supérieurs ou égaux à 5. En pratique, si cette condition n'est pas réalisée, la technique consiste à regrouper certaines modalités (ex : regrouper les yeux noirs et les yeux marrons) afin de, par construction, augmenter la valeurs des effectifs théoriques.

La région critique conduisant au rejet de H_0 est définie par :

$$\left[\chi^2_{(1-\alpha);(p-1)(k-1)}; +\infty\right[$$

Où $\chi^2_{(1-\alpha);(p-1)(k-1)}$ correspond au quantile d'ordre $(1-\alpha)$ de la loi du χ^2 à $(p-1)(k-1)$ degrés de liberté.

Décision:

- Si la valeur de la statistique de test χ^2 est inférieure à la valeur seuil $\chi^2_{(1-\alpha);(p-1)(k-1)}$ alors on accepte l'hypothèse nulle. Les variables X_1 et X_2 sont indépendantes. (c'est-à-dire leurs distributions sont indépendantes).

- Si la valeur de la statistique de test χ^2 est supérieure à la valeur seuil $\chi^2_{(1-\alpha);(p-1)(k-1)}$ alors on rejette l'hypothèse nulle. Il existe une liaison significative entre X_1 et X_2 (c'est-à-dire leurs distributions sont dépendantes).

La table suivante représente les résultats d'une enquête portant sur 300 étudiants à qui il a été demandé s'ils avaient une activité sportive régulière (S=Oui Sport/NS=Non Sport) et s'ils fumaient (F= Fumeur/NF= Non-fumeur).

	F	NF	Total
S	60	76	136
NS	56	108	164
Total	116	184	300

Ici, l'hypothèse H_0 est qu'il y a indépendance entre le fait de fumer et le fait de pratiquer régulièrement le sport. On va alors calculer, sous l'hypothèse H_0, les valeurs théoriques du tableau de contingence.

	F	NF	Total
S	52.59	83.41	136
NS	63.41	100.59	164
Total	116	184	300

on peut maintenant calculer la statistique χ^2

$$\chi^2_{obs} = \frac{(60-52.59)^2}{52.59} + \frac{(56-63.41)^2}{63.41} + \frac{(76-83.41)^2}{83.41} + \frac{(108-100.59)^2}{100.59} = 3.17$$

et

$$\chi^2_{(0.95);(2-2)(2-1)} = \chi^2_{(0.05);(1)} = 3.841$$

comme $\chi^2_{obs} < \chi^2_{(0.95);(1)}$, alors "on accepte H_0".

CHAPITRE 4

Eléments de calcul de probabilités

4.1 Définitions

Ensemble fondamental : considérons une expérience dont l'issue n'est pas prévisible, bien que l'issue ne soit pas connue à l'avance, admettant cependant que l'ensemble des issues possibles est connu. Cet ensemble est appelé ensemble fondamental, par convention est noté Ω.

Exemple 4.1.1 *Si le résultat d'une expérience est la détermination du sexe d'un nouveau-né, alors l'ensemble fondamental* $\Omega = \{fille, garçon\}$.

Exemple 4.1.2 *Si l'expérience consiste à jeter deux pièces de monnaies, alors l'ensemble fondamental est constitué de 4 couples suivant :* $\Omega = \{(pile, pile)\,;(face, face)\,;(pile, face)\,;(face, pile)\}$

Evénement: Tout sous ensemble E de l'ensemble fondamental Ω est appelé « événement ». De ce fait, un événement est donc un ensemble correspondant aux divers résultats possibles d'une expérience aléatoire.

4.2 Opération sur les événements

4.2.1 Union

Soit A et B deux événements de Ω, alors $A \cup B$ sera réalisé si soit A soit B l'est.

Exemple 4.2.1 *Soient* $\Omega = \{g, f\}$, $A = \{g\}$ *et* $B = \{f\}$, *alors* $A \cup B = \{g, f\} = \Omega$

4.2.2 Intersection

Soient A et B deux événements de l'ensemble fondamental Ω, le nouvel événement $A \cap B$, appelé intersection de l'événement A et l'événement B, est considéré comme l'ensemble de réalisations qui sont à la fois dans l'événement A et dans l'événement B.

Autrement dit, l'événement $A \cap B$ ne sera réalisé qui si A et B le sont à la fois.

Exemple 4.2.2 *Nous avons l'ensemble fondamental* $\Omega = \{(p,p)\,;(p,f)\,;(f,p)\,;(f,f)\}$ *tel que, p : pile et f : face.*

Soient $A = \{(p,p)\,;(p,f)\,;(f,p)\}$, *et* $B = \{(f,f)\,;(f,p)\,;(p,f)\}$, *alors : l'événement* $A \cap B = \{(f,p)\,;(p,f)\}$ *est l'événement une pièce de monnaie montre pile et l'autre donne face.*

4.2.3 Complémentarité

Pour tout événement A, le nouvel événement A^c , ou $\overline{A}$ est l'événement complémentaire. C'est-à-dire l'événement A^cdoit, par définition, contenir tous les points de l'ensemble fondamental Ω qui ne sont pas dans l'ensemble A.

Exemple 4.2.3 *On jette à la fois deux dés équilibrés, alors l'ensemble fondamental est :*

2 dés	*1*	*2*	*3*	*4*	*5*	*6*
1	$(1,1)$	$(1,2)$	$(1,3)$	$(1,4)$	$(1,5)$	$(1,6)$
2	$(2,1)$	$(2,2)$	$(2,3)$	$(2,4)$	$(2,5)$	$(2,6)$
3	$(3,1)$	$(3,2)$	$(3,3)$	$(3,4)$	$(3,5)$	$(3,6)$
4	$(4,1)$	$(4,2)$	$(4,3)$	$(4,4)$	$(4,5)$	$(4,6)$
5	$(5,1)$	$(5,2)$	$(5,3)$	$(5,4)$	$(5,5)$	$(5,6)$
6	$(6,1)$	$(6,2)$	$(6,3)$	$(6,4)$	$(6,5)$	$(6,6)$

alors $\Omega = \{(1,1)\,;(1,2)\,;....;(6,5)\,;(6,6)\}$.

Soit l'événement $A = \{(1,6)\,;(2,5)\,;(3,4)\,;(4,3)\,;(5,2)\,;(6,1)\}$ *(la somme des deux faces des é dés égale à 7); alors* A^c $\left(\overline{A}\right)$*, sera réalisé lorsque la somme des deux dés n'est pas égale à 7:*

$$A^c = \left\{ \begin{array}{cccccccccc} (1,1) & (1,2) & (1,3) & (1,4) & (1,6) & (2,1) & (2,2) & (2,3) & (2,4) & (2,6) \\ (3,1) & (3,2) & (3,3) & (3,5) & (3,6) & (4,1) & (4,2) & (4,4) & (4,5) & (4,6) \\ (5,1) & (5,3) & (5,4) & (5,5) & (5,6) & (6,2) & (6,3) & (6,4) & (6,5) & (6,6) \end{array} \right\}$$

Propriétés

Soient Ω un ensemble fondamental d'une expérience aléatoire, et A un événement de Ω, alors il existe une valeur $P(A)$ appelée probabilité de l'événement A où :

- $0 \leq P(A) \leq 1$
- $P(\Omega) = 1$
- Si A et B sont deux événements qui s'excluent mutuellement, alors : $P(A \cup B) = P(A) + P(B)$
- Pour toute suite d'événement mutuellement disjoints $E_1, E_2, E_3, ...$(c'est à dire $E_i \cap E_j = \phi, \forall i \neq j$) alors :$P\left(\overset{\infty}{\underset{i=1}{\cup}} E_i\right) = \sum_{i=1}^{\infty} P(E_i)$, tel que, $P(E_i)$ est la probabilité de l'événement E_i.

4.2.4 Théorèmes élémentaires

Théorème 4.2.1 *Soient A et $\overline{A}$ deux événements mutuellement disjoints, et $A \cup \overline{A} = \Omega$, alors :*

- $P(\Omega) = 1 = P\left(A \cup \overline{A}\right) = P(A) + P\left(\overline{A}\right) \Rightarrow P\left(\overline{A}\right) = 1 - P(A)$
- *Si* $A \subset B \Rightarrow P(A) \leq P(B)$

Théorème 4.2.2 *soient deux événements quelconques A et B, alors:*

- $P\left(\overline{B}\right) = P(A) - P(A \cap B)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

4.2.5 Calcul des probabilités

Définition 4.2.1 *La probabilité d'un évènement A est*

$$P(A) = \frac{\textit{Le nombre de cas favorables à la réalisation de l'évènement } A}{\textit{Nombre de cas possibles de l'ensemble fondamental } \Omega}$$

Exemple 4.2.4 *On lance une pièce de monnaie, en admettant que pile a autant de chances d'apparaître que face. C'est-à-dire,*

$$\begin{aligned} P(\{pile\}) &= P(\{face\}) \\ &= \frac{\textit{Le nombre de cas favorables à la réalisation de l'évènement } \{pile\} \textit{ (ou } \{face\}\textit{)}}{\textit{Nombre de cas possibles de l'ensemble fondamental } \Omega} \\ &= \frac{1}{2} \end{aligned}$$

Exemple 4.2.5 *On jette un dé équilibré, alors les probabilités*

$$P(\{1\}) = \frac{\textit{Le nombre de cas favorables à la réalisation de l'évènement } \{1\}}{\textit{Nombre de cas possibles de l'ensemble fondamental } \Omega} = \frac{1}{6}$$

d'où

$$P(\{2\}) = P(\{3\}) = P(\{4\}) = P(\{5\}) = P(\{6\}) = \frac{1}{6}$$

Par ailleurs, la probabilité B, B=" Avoir un chiffre pair" est :

$$P(B) = P(\{2,4,6\}) = P(\{2\}) + P(\{4\}) + P(\{6\}) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{3}{6} = \frac{1}{2}$$

CHAPITRE 5

Bibliographie

Bibliographie

[1] Baenet, V.(1973),Comparativa Statistical Inference, London, Wiley.

[2] Bouznit M., (2018), Cours de Probabilités, Université de Bejaia, Algérie.

[3] Chekroun A., (2018), Cours: Statistiques descriptives, Université Abou Bekr Belkaid, Tlemcen, Algérie.

[4] Cramer, H. (1974), Mathematical Methods of Statistics, 13^{th} printing, Princeton, Princeton University Press.

[5] Dagnelie P., (2011), Statistique théorique et appliquée. Tome 2. Inférence statistique à une et à deux dimensions, De Boeck, 736 p.

[6] Dagnelie, P. (1986), Théorie et méthodes statistiques, 2^e éd., 2 vol., Gembloux, Duculot.

[7] Dagnelie, P. (1992), Statistique théorique et appliqué. Tome 1: les bases théoriques, Gembloux, Presses Agronomiques de Gembloux.

[8] Dagnelie, P. (2006), Statistique théorique et appliqué. Tome 2: Inférence statistique à une et à deux dimensions, 2^e éd., Bruxelles, De Boeck Université.

[9] Evans, M., N. Hastings and Peacock (2000), Statistical Distributions, 3^{rd} ed., New York, Wiley Series in Probability and Statistics.

[10] Huntsberger, D. V. and P. Billingsley (1977), Elements of statistical Inference, Boston, Allyn and Bacon.

[11] Lecoutre J.-P., (2002), Statistiques et probabilités, manuel et exercices corrigés, Dunod, 3e édition, 296p.

[12] Mélard, G. and M. Petitfrère (1989), Manuel d'inférence statistique univariée et bivariée, 3^e éd., Presses Universitaires de Bruxelles.

[13] Tassi, P., (1989), Méthodes statistiques, 2^e éd., Paris, Economica.