Cours de Biostatistiques
$2^{ère}$partie
Statistiques Inférentielles

Abdallah Benaissa [1]

14 Avril 2020

[1]Professeur en mathématiques, chargé du module de biostatistiques à la faculté de médecine de l'université de Batna - Algérie

Table des matières

Preface

Ce polycopié est adressé aux étudiants de la première année en médecine. Il est conçu selon le nouveau programme officiel élaboré en Juin 2017. Il contient la totalité du programme concernant les statistiques inférentielles. Pour rendre ce document adapté aux programmes des étudiants en médecine en particulier et en sciences appliquées en général, on a évité au maximum les démonstrations mathématiques laborieuses et on s'est basé surtout sur la variété d'exemples.

0.0.1

Chapitre 1

Moyenne Arithmétique

1.1 Théorie d'échantillonnage

La théorie d'échantillonnage englobe les méthodes de sélection d'un échantillon à partir d'une population statistique. Les

méthodes d'échantillonnage qui nous intéressent dans ce cours sont les méthodes dites probabilistes. Dans ces méthodes

probabilistes, les valeurs des paramètres, calculés sur l'échantillon, peuvent être considérées comme des estimations

des mêmes paramètres de la population. La précision de ces estimations est quantifiée en utilisant la théorie des probabilités.

les techniques de passage de l'échantillon à la population mére s'appellent, comme on l'a déja mentionné au début de ce cours,

les statistiques inférentielles.

1.1.1 Les différentes méthodes d'échantillonnage aléatoire.

On estime que l'échantillonnage aléatoire le plus adéquat pour les statistiques inférentielle est le suivant :

Echantillonnage aléatoire simple. Le principe de ce genre d'échantillonnage est le suivant :

chaque individu de la population a les mêmes chances d'appartenir à l'échantillon. Dans ce genre d'échantillonnage, on utilise des ordinateurs ou des machines générant des nombres aléatoires pour former l'échantillon. C'est généralement ce genre d'échantillonnage qui est utilisé dans les statistiques inférentielles (estimations et tests statistiques).

Les autres échantillonnages sont :

- l'échantillonnage par échantillons stratifiés. Utilisé dans le cas où la population est formée de groupes homogènes.

- l'échantillonnage par échantillons en grappes. Il est utilisé dans le cas où la population est formée de groupes non homogènes. Dans ce cas des groupes entiers peuvent appartenir à l'échantillon.

1.1.2 Définitions

Définition

On appelle échantillon aléatoire de taille n (ou n-échantillon) une suite de n variables aléatoires $X_1, X_2, ...X_n$ indépendantes et de même loi. Cette loi est appelée la loi mère de l'´echantillon.

Définition

On appelle statistique h, liée à un n-échantillon $X_1, X_2, ...X_n$, une variable aléatoire $h_n(X_1, X_2, ...X_n)$ fonction des n variables aléatoires $X_1, X_2, ...X_n$.

Définition

On appelle moyenne de l'échantillon ou moyenne empirique, la statistique notée $\overline{X}_n$ et définie par

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{1.1}$$

Définition

On appelle variance empirique la statistique, notée $\left(\widetilde{S}_n\right)^2$, définie par

$$\left(\widetilde{S}_n\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2 \tag{1.2}$$

Définition

On appelle variance de l'échantillon la statistique, notée $(S_n)^2$, et définie par

$$(S_n)^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 \tag{1.3}$$

1.1.3 Exemple.

On considère la variable aléatoire X définie par l'expérience du jet d'une pièce de monnaie. les deux valeurs possibles de X sont pile ou face, l'une codée par 1 et l'autre par zéro. La loi de X est donnée par $P_r(X=0) = P_r(X=1) = \frac{1}{2}$. Considérons l'échantillon aléatoire de taille 2 : $X_1 = X_2 = X$. La loi de probabilité de l'échantillon $(X_1 \times X_2)$ est donnée par le tableau

$(X_1 \times X_2)$	$(0,0)$	$(0,1)$	$(1,0)$	$(1,1)$	$\sum$
p_{ij}	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	1

On note bien que la variable aléatoire moyenne arithmétique $\overline{X}_2$ est définie sur l'ensemble fondamental $X_1 \times X_2$ dont la loi de probabilité est définie ci-dessus.

Pour calculer l'espérance $E\left(\overline{X}_2\right)$ et la variance $V\left(\overline{X}_2\right)$ de la variable aléatoire

$$\overline{X}_2 = \frac{X_1 + X_2}{2},$$

On utilise un tableau définissant la loi de $\overline{X}_2$:

$\overline{X}_2 = \overline{x}_{2i}$	0	$\frac{1}{2}$	1	$\sum$
$p_i = P_r\left(\overline{X}_2 = \overline{x}_{2i}\right)$	$\frac{1}{4}$	$\frac{1}{4}+\frac{1}{4}=\frac{1}{2}$	$\frac{1}{4}$	1
$p_i\overline{x}_{2i}$	0	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{2}$
$p_i\left(\overline{x}_{2i}\right)^2$	0	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{3}{8}$

Donc

$$\begin{aligned} E\left(\overline{X}_2\right) &= p_1\overline{x}_{21} + p_2\overline{x}_{22} + p_3\overline{x}_{23} \\ &= \frac{1}{4}.0 + \frac{1}{2}.\frac{1}{2} + \frac{1}{4}.1 \\ &= \frac{1}{2}. \end{aligned}$$

et

$$\begin{aligned} V\left(\overline{X}_2\right) &= \sum_{i=1}^{3} p_i\left(\overline{x}_{2i}\right)^2 - \left(E\left(\overline{X}_2\right)\right)^2 \\ &= \frac{3}{8} - \frac{1}{4} \\ &= \frac{1}{8}. \end{aligned}$$

Calculons par le même procédé l'espérance $E\left(\left(S_2\right)^2\right)$ de la variable aléatoire

$$\left(S_2\right)^2 = \left(X_1 - \overline{X}_2\right)^2 + \left(X_2 - \overline{X}_2\right)^2.$$

$(X_1 \times X_2)$	$(0,0)$	$(0,1)$	$(1,0)$	$(1,1)$	$\sum$
p_{ij}	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	1
$\overline{X}_2$	0	$\frac{1}{2}$	$\frac{1}{2}$	1	
$\left(X_1 - \overline{X}_2\right)^2$	0	$\frac{1}{4}$	$\frac{1}{4}$	0	
$\left(X_2 - \overline{X}_2\right)^2$	0	$\frac{1}{4}$	$\frac{1}{4}$	0	
$\left(S_2\right)^2$	0	$\frac{1}{2}$	$\frac{1}{2}$	0	
$\left(\widetilde{S}_2\right)^2$	0	$\frac{1}{4}$	$\frac{1}{4}$	0	

On en déduit la loi de probabilité de la variable aléatoire $(S_2)^2$ et de son espérance $E\left((S_2)^2\right)$

$(S_2)^2 = (s_2)_i^2$	0	$\frac{1}{2}$	$\sum$
p_i	$\frac{1}{4}+\frac{1}{4}=\frac{1}{2}$	$\frac{1}{4}+\frac{1}{4}=\frac{1}{2}$	1
$\mathrm{p}_i\,(s_2)_i^2$	0	$\frac{1}{4}$	$\frac{1}{4}$

Donc

$$\begin{aligned} E\left((S_2)^2\right) &= p_1\,(s_2)_1^2 + p_2\,(s_2)_2^2 \\ &= \frac{1}{2}.0 + \frac{1}{2}.\frac{1}{2} \\ &= \frac{1}{4}. \end{aligned}$$

De la même façon on déduit la loi de probabilité de $\left(\widetilde{S}_2\right)^2$ et de son espérance $E\left(\left(\widetilde{S}_2\right)^2\right)$:

$\left(\widetilde{S}_2\right)^2 = (\widetilde{s}_2)_i^2$	0	$\frac{1}{4}$	$\sum$
p_i	$\frac{1}{4}+\frac{1}{4}=\frac{1}{2}$	$\frac{1}{4}+\frac{1}{4}=\frac{1}{2}$	1
$\mathrm{p}_i\,(\widetilde{s}_2)_i^2$	0	$\frac{1}{8}$	$\frac{1}{8}$

et

$$E\left(\left(\widetilde{S}_2\right)^2\right) = \frac{1}{8}.$$

On refait maintenant les mêmes calculs pour l'échantillon de taille 3.
La loi du produit $(X_1 \times X_2 \times X_3)$ est donnée par le tableau

$(X_1 \times X_2 \times X_3)$	$(0,0,0)$	$(0,0,1)$	$(0,1,0)$	$(0,1,1)$	$(1,0,0)$	$(1,0,1)$	$(1,1,0)$	$(1,1,1)$	$\sum$
p_{ijk}	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	1

La loi de

$$\overline{X}_3 = \frac{X_1 + X_2 + X_3}{3},$$

son espérance $E\left(\overline{X}_3\right)$ et sa variance $V\left(\overline{X}_3\right)$ sont calculés dans le tableau

$\overline{X}_3 = \overline{x}_{3i}$	0	$\frac{1}{3}$	$\frac{2}{3}$	1	$\sum$
$p_i = P_r\left(\overline{X}_3 = \overline{x}_{3i}\right)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$	1
$p_i\overline{x}_{3i}$	0	$\frac{3}{24}$	$\frac{6}{24}$	$\frac{1}{8}$	$\frac{1}{2}$
$p_i\left(\overline{x}_{3i}\right)^2$	0	$\frac{1}{24}$	$\frac{1}{6}$	$\frac{1}{8}$	$\frac{1}{3}$

En effet

$$E\left(\overline{X}_3\right) = \sum p_i\overline{x}_{3i} = \frac{1}{2}$$

et

$$\begin{aligned} V\left(\overline{X}_3\right) &= \sum p_i\left(\overline{x}_{3i}\right)^2 - \left(E\left(\overline{X}_3\right)\right)^2 \\ &= \frac{1}{3} - \frac{1}{4} = \frac{1}{12}. \end{aligned}$$

Le tableau suivant nous facilite la construction de la loi de probabilité de $(S_3)^2$ et de $\left(\widetilde{S}_3\right)^2$.

$(X_1 \times X_2 \times X_3)$	$(0,0,0)$	$(0,0,1)$	$(0,1,0)$	$(0,1,1)$	$(1,0,0)$	$(1,0,1)$	$(1,1,0)$	$(1,1,1)$
p_{ijk}	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$
$\overline{X}_3$	0	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{2}{3}$	$\frac{1}{3}$	$\frac{2}{3}$	$\frac{1}{3}$	1
$\left(X_1 - \overline{X}_3\right)^2$	0	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{4}{9}$	$\frac{4}{9}$	$\frac{1}{9}$	$\frac{1}{9}$	0
$\left(X_2 - \overline{X}_3\right)^2$	0	$\frac{1}{9}$	$\frac{4}{9}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{4}{9}$	$\frac{1}{9}$	0
$\left(X_3 - \overline{X}_3\right)^2$	0	$\frac{4}{9}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{4}{9}$	0
$(S_3)^2$	0	$\frac{3}{9}$	$\frac{3}{9}$	$\frac{3}{9}$	$\frac{3}{9}$	$\frac{3}{9}$	$\frac{3}{9}$	0
$\left(\widetilde{S}_3\right)^2$	0	$\frac{2}{9}$	$\frac{2}{9}$	$\frac{2}{9}$	$\frac{2}{9}$	$\frac{2}{9}$	$\frac{2}{9}$	0

A partir du tableau ci-dessus on déduit

$$\begin{aligned} E\left(\left(S_3\right)^2\right) &= p_1\left(s_3\right)_1^2 + p_2\left(s_3\right)_2^2 \\ &= \frac{2}{8}.0 + \frac{6}{8}.\frac{3}{9} = \frac{1}{4}. \end{aligned}$$

et

$$\begin{aligned} E\left(\left(\widetilde{S}_3\right)^2\right) &= p_1\left(\widetilde{s}_3\right)_1^2 + p_2\left(\widetilde{s}_3\right)_2^2 \\ &= \frac{2}{8}.0 + \frac{6}{8}.\frac{2}{9} \\ &= \frac{1}{6}. \end{aligned}$$

On remarque à travers l'exemple précédent que, pour $n = 1, n = 2$ et $n = 3$.

1- l'espérance $E\left(\overline{X}_n\right)$ de la moyenne $\overline{X}_n$ de l'échantillon de taille n est égale à l'espérance $\frac{1}{2}$ de la loi mère :

$$E\left(\overline{X}_1\right) = E\left(\overline{X}_2\right) = \frac{1}{2}.$$

2- la variance $V\left(\overline{X}_n\right)$ de l'échantillon de taille n est égale a la variance de la population mère, ici $\frac{1}{4}$, divisée par la

taille n de l'échantillon :

$$\begin{aligned} V\left(\overline{X}_2\right) &= \frac{\frac{1}{4}}{2} = \frac{1}{8}. \\ V\left(\overline{X}_3\right) &= \frac{\frac{1}{4}}{3} = \frac{1}{12}. \end{aligned}$$

3- l'espérance $E\left(\left(S_n\right)^2\right)$ de la variance de l'échantillon $\left(S_n\right)^2$ est égale à la variance de la population mère, ici égale $\frac{1}{4}$:

$$E\left(\left(S_2\right)^2\right) = E\left(\left(S_3\right)^2\right) = \frac{1}{4}.$$

4- l'espérance $E\left(\left(\widetilde{S}_n\right)^2\right)$ de la variance empirique $\left(\widetilde{S}_n\right)^2$ est égale à la variance de la population mère, ici égale $\frac{1}{4}$, multipliée par $\frac{n-1}{n}$ de l'échantillon :

$$\begin{aligned} E\left(\left(\widetilde{S}_2\right)^2\right) &= \frac{1}{4} \times \frac{(2-1)}{2} = \frac{1}{8}. \\ E\left(\left(\widetilde{S}_3\right)^2\right) &= \frac{1}{4} \times \frac{(3-1)}{3} = \frac{1}{6}. \end{aligned}$$

Les relations entre les lois des statistiques de l'échantillon constatées sur l'exemple précédent sont valables dans le cas générale. On résume ces relations dans la proposition suivante.

Proposition

Sous les notations du présent paragraphe, les relations suivantes sont vérifiées pour tout entier n.

1-

$$E\left(\overline{X}_n\right) = \mu, \text{ et } V\left(\overline{X}_n\right) = \frac{\sigma^2}{n}. \tag{1.4}$$

2-

$$E\left(\left(\widetilde{S}_n\right)^2\right) = \frac{n-1}{n}\sigma^2. \tag{1.5}$$

3-

$$E\left(\left(S_n\right)^2\right) = \sigma^2. \tag{1.6}$$

1.2 La loi de probabilité de la moyenne arithmétique

1.2.1 Proposition

Si la loi mère est normale, c'est-à-dire $X \rightsquigarrow \mathcal{N}(\mu, \sigma^2)$, alors, pour tout entier naturel n, la loi de la moyenne arithmétique $\overline{X}_n$ est normale indépendamment de la taille n de l'échantillon, précisément

$$\overline{X}_n \rightsquigarrow \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right), \tag{1.7}$$

par conséquent,

$$\frac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \rightsquigarrow \mathcal{N}(0,1). \tag{1.8}$$

1.2.2 Théorème central limite.

le théorème central limite nous renseigne sur le comportement asymptotique de la loi de la moyenne arithmétique $\overline{X}_n$, plus précisément sur la nature de cette loi pour des valeurs assez grandes de l'entier n.

Ce théorème est l'outil principal dans le concept (très important en application) d'inférence statistique, en particulier, les estimations et les tests statistiques.

Formulation du théorème.

Si la taille n de l'échantillon $X_1, X_2, ..., X_n$ est assez grande, alors, la loi de probabilité de la moyenne arithmétique réduite centrée est presque identique à la loi normale réduite centrée $\mathcal{N}(0,1)$. On exprime ce résultat par

$$\frac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0,1) \text{ (à peu prés).}$$

Dans la pratique, si $n \geq 30$, on remplace la loi de

$$\frac{\sqrt{n}\left(\overline{X}_n - \mu\right)}{\sigma}$$

par la loi normale réduite centrée

$$Z \sim \mathcal{N}(0,1).$$

1.2.3 Remarque

Il est utile de bien noter la force et le résultat extraordinaire du théorème central limite. Ce théorème atteste que la moyenne arithmétique $\overline{X}_n$ d'un échantillon aléatoire assez grand d'une variable aléatoire X, dont on connait la moyenne μ et la variance σ^2, est (à peu prés) une loi normale de moyenne μ et d'écart type $\frac{\sigma}{\sqrt{n}}$, cela équivaut à ce que la loi de la moyenne réduite centrée

$$\frac{\sqrt{n}\left(\overline{X}_n - \mu\right)}{\sigma}$$

est pour les grands échantillons, (à peu prés) identique à la loi normale réduite centrée $\mathcal{N}(0,1)$.

1.2.4 Exemple 1

Dans une ville la glycémie X d'une population est distribuée selon une loi normale, d'écart-type $\sigma = 0,4$ et d'espérance $\mu = 1.5$.

Trouvez le nombre positif h tel que

$$P_r\left(\left|\overline{X}_{10} - 1.5\right| \geq h\right) = \alpha,$$

avec $\alpha = 0.05$.

Solution.

On a

$$P_r\left(\left|\overline{X}_{10} - 1.5\right| \geq h\right) = P_r\left(\left|\frac{\overline{X}_{10} - 1.5}{\frac{0.4}{\sqrt{10}}}\right| \geq \frac{h}{\frac{0.4}{\sqrt{10}}}\right) = 0.05,$$

Puisque

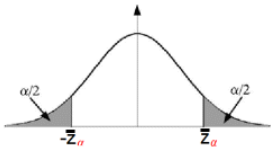
$$\frac{\overline{X}_{10} - 1.5}{\frac{0.4}{\sqrt{10}}}$$

suit la loi normale réduite centrée, on peut utiliser la table 2, de l'écart réduit pour trouver la valeur de

$$\frac{h}{\frac{0.4}{\sqrt{10}}} = \overline{z}_{0.05},$$

on aura

$$\frac{h}{\frac{0.4}{\sqrt{10}}} = 1.96,$$



α	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,00	∞	2,576	2,326	2,170	2,054	1,960	1,881	1,812	1,751	1,695
0,10	1,645	1,598	1,555	1,514	1,476	1,440	1,405	1,372	1,341	1,311

Détermination de $\overline{Z}_{0.05}$ dans la table 2 de l'écart réduit de la loi normale réduite centrée. La valeur de $\overline{Z}_{0.05}$ correspondant à $\alpha = 0.05$ se trouve dans la cellule, intersection de la ligne de la partie décimale 0.00 et la colonne de la partie centième 0.05.

par conséquent

$$h = 1.96 \times \frac{0.4}{\sqrt{10}} = \mathbf{0.248}.$$

1.2.5 Exemple 2

Soit X une variable aléatoire quelconque, de moyenne $\mu = 90$ et de variance $\sigma^2 = 4$. Calculer h tel que

$$P_r\left(\left|\overline{X}_{100} - 90\right| \leq h\right) = 0.95. \tag{1.9}$$

Solution.

La relation (1.9) équivaut à que

$$P_r\left(\left|\overline{X}_{100} - 90\right| \geq h\right) = 1 - 0.95 = 0.05.$$

D'où

$$\begin{aligned} P_r\left(\left|\overline{X}_{100} - 90\right| \geq h\right) &= P_r\left(\left|\frac{\overline{X}_{100} - 90}{\frac{2}{\sqrt{100}}}\right| \geq \frac{h}{\frac{2}{\sqrt{100}}}\right) \\ &= 0.05. \end{aligned}$$

Puisque $n = 100 \geq 30$, alors la variable aléatoire

$$T_n = \frac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\overline{X}_{100} - 90}{\frac{2}{\sqrt{100}}}$$

suit (approximativement) la loi normale réduite centrée. Cela nous permet d'utiliser la table 2 de l'écart réduit pour avoir

$$\frac{h}{\frac{2}{\sqrt{100}}} = \overline{z}_{0.05} = 1.96$$

d'où

$$h = 1.96 \times \frac{2}{\sqrt{100}} = \mathbf{0.392}.$$

Chapitre 2

Théorie de l'estimation

Soit X une variable aléatoire quantitative ou qualitative. La théorie de l'estimation consiste en l'évaluation d'un paramètre donné de la variable aléatoire X, comme par exemple sa moyenne μ ou sa variance σ^2, à partir de la valeur du même paramètre d'un échantillon de X. Il y a deux genres d'estimation, estimation ponctuelle et estimation par intervalle de confiance. On parle d'estimation ponctuelle quand le paramètre est estimé par une valeur unique, et on parle d'estimation par intervalle de confiance quand on construit un intervalle et on affirme que le paramètre en question appartient à cet intervalle avec une probabilité donnée. Dans cette section, X désigne toujours la variable aléatoire de la loi mère, μ sa moyenne et σ^2 sa variance (si X est quantitative). En outre, on note son paramètre à estimer par θ.

2.1 Estimation ponctuelle

Soit $X_1, X_2, ..., X_n$ un échantillon aléatoire de taille n et soit $\varphi_n(X_1, X_2, ..., X_n)$ une variable aléatoire fonction de $X_1, X_2, ..., X_n$. On dit que $\varphi_n(X_1, X_2, ..., X_n)$ est un estimateur ponctuel du paramètre θ si les valeurs de $\varphi_n(X_1, X_2, ..., X_n)$ se rapprochent (dans un sens à préciser) pour les grandes valeurs de l'entier n vers le paramètre θ.

La qualité d'un estimateur est caractérisée par la variance de l'estimateur et par ce qu'on appelle le biais.

2.1.1 Biais

le biais d'un estimateur ponctuel Φ d'un paramètre θ, noté $B(\Phi)$ est défini par

$$B(\Phi) = E(\Phi - \theta) = E(\Phi) - \theta. \tag{2.1}$$

Il est évident que la qualité de l'estimateur est déterminée par la petitesse des deux grandeurs, la valeur absolue du biais $|B(\Phi)|$ et la variance $\sigma^2(\Phi)$ de l'estimateur Φ. Si $B(\Phi) = 0$, on dit que Φ est un estimateur sans biais du paramètre θ, sinon on dit que c'est un estimateur biaisé.

2.1.2 Exemple

a- La moyenne arithmétique $\overline{X}_n$ est un estimateur sans biais de la moyenne μ de la loi mère, puisque

$$\begin{aligned} E\left(\overline{X}_n - \mu\right) &= E\left(\overline{X}_n\right) - \mu \\ &= \mu - \mu = 0. \end{aligned}$$

b- S_n^2 est un estimateur sans biais de la variance σ^2 de X, puisque

$$\begin{aligned} B\left(S_n^2 - \sigma^2\right) &= E\left(S_n^2 - \sigma^2\right) \\ &= E\left(S_n^2\right) - \sigma^2 \\ &= \sigma^2 - \sigma^2 \\ &= 0. \end{aligned}$$

c- $\widetilde{S}_n^2$ est un estimateur biaisé de la variance σ^2 de X, puisque

$$\begin{aligned} B\left(\widetilde{S}_n^2 - \sigma^2\right) &= E\left(\widetilde{S}_n^2\right) - \sigma^2 \\ &= \frac{n-1}{n}\sigma^2 - \sigma^2 \\ &\neq 0. \end{aligned}$$

2.2 Estimation par intervalle de confiance

Il est plus raisonnable d'estimer un paramètre θ d'une variable aléatoire X par un intervalle I et cela à partir de mesures $x_1, ..., x_n$ effectuées sur un échantillon de taille n. Cela se fait de la manière suivante :

A partir des valeurs $x_1, ..., x_n$, on construit un intervalle I tel qu'on peut affirmer avec la probabilité $1-\alpha$ que la valeur du paramètre θ se situe dans l'intervalle I. La probabilité $0 < \alpha < 1$ fixée d'avance s'appelle le risque ou le seuil de signifiance, et la probabilité $1-\alpha$ s'appelle le niveau de confiance. On dit que I est l'intervalle de confiance au risque α (ou au seuil de signfiance α ou de niveau $1-\alpha$, ou seulement à $(1-\alpha)\,100\%$).

Donc, la confirmation "I est un intervalle de confiance du paramètre θ, au risque α" s'exprime par la formule

$$P_r\left(\theta \in I\right) = 1 - \alpha.$$

Cette formule équivaut à

$$P_r\left(\theta \notin I\right) = \alpha.$$

2.2.1 Signification de l'intervalle de confiance

Pour plus de clarté, on explique cette caractérisation pour $\alpha = 0.05$. Si $\alpha = 0.05$, l'intervalle de confiance de la moyenne μ est caractérisé par la propriété suivante :

si on répète 100 fois l'opération consistant à prendre n mesures de X et calculer la moyenne arithmétique m, alors dans 95 cas la moyenne vraie μ se situe dans l'intervalle de confiance I et dans les cinq qui restent elle se situe en dehors de cet intervalle.

2.2.2 Taille de la population

Dans l'étude développée dans ce cours pour les intervalles de confiance et pour les tests statistiques, on suppose que la taille de la population est trés grande devant la taille de l'échantillon considéré (taille infinie de la population). Ce choix est motivé par deux raisons.

Premièrement, en pratique, l'étude est importante surtout pour les populations de très grande taille, car pour les populations de petite taille, le développement des machines de calculs a rendu accessible le calcul direct des paramètres de toute la population sans passer par des échantillons.

Deuxièmement, cela nous évite la complication des formules obtenues par un autre terme appelé coefficient d'exhaustivité.

2.2.3 Taille de l'échantillon

Dans le calcul d'intervalles de confiance ou pour les tests statistiques, il vaut mieux considérer de grands échantillons ($n > 30$), car cela permet l'utilisation de la loi normale et simplifie ainsi les formules. Mais dans le domaine médical, ce n'est pas toujours facile de considérer de grands échantillons, car pour certains problèmes, ça nécessite des manipulations pour chaque observation.

2.2.4 Intervalle de pari (ou de fluctuation).

1- L'intervalle de confiance donne une estimation d'un paramètre de la population mère à partir de paramètres (observés) d'un échantillon, tandis que l'intervalle de pari consiste en l'estimation d'un paramètre de l'échantillon à partir de paramètres de la loi mère. L'intervalle de pari a un lien direct avec le concept de tests d'hypothèse qu'on va étudier dans les chapitres qui suivent.

2- L'intervalle de pari de la moyenne arithmétique $\overline{X}_n$ dépend uniquement de la taille n et du risque α. Donc pour chaque valeur de la taille n et du risque α, il y a un seul intervalle de pari. Tandis que l'intervalle de confiance de la moyenne vraie μ dépend de la taille n et du risque α, mais aussi des mesures effectuées $x_1, ...x_n$. Ainsi, pour un risque donné α et une taille n de l'échantillon,

on peut construire plusieurs intervalles de confiance.

3- L' intervalle de pari et l'intervalle de confiance sont de même longueur. Cette longueur augmente quand le risque α diminue et diminue quand la taille n augmente.

2.2.5 Intervalle de pari pour $\overline{X}_n$

Soit X une variable aléatoire de variance σ^2 et d'espérance μ finie, et soit $0 < \alpha < 1$. Soit l'échantillon aléatoire $X_1, ..., X_n$ de taille n.

Théorème.

Si l'une au moins des deux conditions," la loi de X est normale" ou "$n > 30$", est vérifiée, alors

$$P_r\left(\left|\overline{X}_n - \mu\right| \geq \overline{z}_\alpha \frac{\sigma}{\sqrt{n}}\right) = \alpha, \tag{2.2}$$

où $\overline{z}_\alpha$ est calculé dans la table 2 de l'écart réduit de la loi normale réduite centrée.

Cette formule est une conséquence du théorème centrale limite.

2.3 Intervalles de confiance pour μ

Soient $x_1, ..., x_n$ des mesures effectuées sur l'échantillon aléatoire $X_1, ..., X_n$ d'une variable aléatoire X.

En d'autres termes, $x_1, ..., x_n$ sont n valeurs de la variable X.

A partir de ces mesures on construit un intervalle de confiance au risque $0 < \alpha < 1$ de la moyenne vraie μ, centré sur la moyenne observée m, de la façon suivante.

La moyenne observée m est la moyenne arithmétiques des n valeurs $x_1, ..., x_n$ de X :

$$m = \frac{x_1 + ... + x_n}{n}. \tag{2.3}$$

Les formules, nous permettant de calculer effectivement l'intervalle de confiance pour la moyenne μ de la population mère en fonction de la moyenne observée m et le risque α, dépendent de la loi de la population mère X, de son écart-type σ et de la taille n de l'échantillon. On explicitera dans ce qui suit les formules utilisées.

2.3.1 Cas de grands échantillons

D'après le théorème central limite, pour les grands échantillons (de taille $n > 30$), la variable aléatoire moyenne arithmétique de l'échantillon

$$\overline{X}_n = \frac{X_1 + ... + X_n}{n},$$

(où $X_1 = ... = X_n = X$), suit (approximativement) la loi normale de moyenne μ et d'écart type $\frac{\sigma}{\sqrt{n}}$, même dans le cas où la loi de la population mère n'est pas normale.

A)- Cela nous permet de construire un intervalle de confiance I_α pour μ, au risque α, sous la forme

$$I_\alpha = \left[m - \overline{z}_\alpha \frac{\sigma}{\sqrt{n}},\ m + \overline{z}_\alpha \frac{\sigma}{\sqrt{n}} \right]$$

B)- Si l'écart type σ de la population n'est pas connu, on le remplace par l'écart type S de l'échantillon, l'intervalle de confiance I_α devient

$$I_\alpha = \left[m - \overline{z}_\alpha \frac{S}{\sqrt{n}},\ m + \overline{z}_\alpha \frac{S}{\sqrt{n}} \right].$$

Exemple

Un échantillon de 75 stimulateurs cardiaques étudié a donné les résultats suivants : La moyenne est $m = 0.31$ et l'écart-type est $S = 0.015$.

1) Donner un intervalle de confiance à 0.95 % pour la moyenne des stimulateurs cardiaques.

2) Quelle sera la taille n de l'échantillon si l'erreur absolue est inférieure à 0.001, et l'écart type S est le même ?

Solution

1) Puisque la taille de l'échantillon $n = 75 > 30$, et l'écart type σ de la population est inconnu, l'intervalle de confiance

$I_{0.05}$ est donné par

$$\begin{aligned} I_{0.05} &= \left[m - \overline{z}_{0.05} \frac{S}{\sqrt{n}},\ m + \overline{z}_{0.05} \frac{S}{\sqrt{n}} \right] \\ &= \left[0.31 - 1.96 \frac{0.015}{\sqrt{75}},\ 0.31 + 1.96 \frac{0.015}{\sqrt{75}} \right] \\ & \quad [0.31 - 0.0039,\ 0.31 + 0.0039] \\ &= [0.3066, 0.3134]. \end{aligned}$$

On représente également cet intervalle de confiance sous la forme

$$\mu = 0.31 \pm 0.0039$$

2) On a

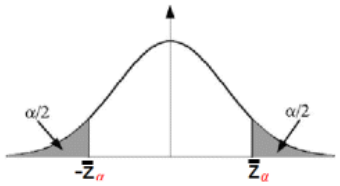
$$\overline{z}_{0.05} \frac{S}{\sqrt{n}} = 1.96 \frac{0.015}{\sqrt{n}} < 0.001.$$

Donc

$$864.36 = \left(1.96 \frac{0.015}{0.001} \right)^2 < n,$$

donc, puisque n est un entier,

$$n \geq 865.$$



α	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,00	∞	2,576	2,326	2,170	2,054	1,960	1,881	1,812	1,751	1,695
0,10	1,645	1,598	1,555	1,514	1,476	1,440	1,405	1,372	1,341	1,311

Utilisation de la table 2 de l'écart réduit de la loi normale réduite centrée pour déterminer la valeur de $\overline{z}_\alpha$ (ici $\alpha = 0.05$) : la valeur de $\overline{z}_{0.05}$ se trouve dans la cellule, intersection de la ligne de la partie décimale de 0.05 (donc la première ligne) et la colonne de la partie centième (donc la sixième colonne).

2.3.2 Cas de petits échantillons, quand X suit la loi normale et σ est inconnu

Dans le cas où l'écart type σ de la population est inconnu, en remplaçant σ par l'écart type S de l'échantillon, la variable aléatoire

$$\frac{\overline{X}_n - \mu}{\frac{S}{\sqrt{n}}} \tag{2.4}$$

ne suit pas approximativement une loi normale pour les petits échantillons $(n \leq 30)$, elle suit plutôt une loi appelée loi de Student, dépendant d'un seul paramètre $ddl = n - 1$, appelé degré de liberté. On écrit alors

$$\frac{\overline{X}_n - \mu}{\frac{S}{\sqrt{n}}} \sim T_{n-1}. \tag{2.5}$$

L'intervalle de confiance au risque α, de la moyenne μ de la population, centré sur la moyenne m de l'échantillon est, dans ce cas, donné par la formule

$$\mu = m \pm \overline{t}_\alpha \frac{S}{\sqrt{n}}, \tag{2.6}$$

où, $\overline{t}_\alpha$ est calculé dans la table 3 de l'écart réduit de la loi de Student T_{n-1} de degré de liberté $ddl = n - 1$, à partir de la formule

$$P_r\left(|T_{n-1}| \geq \overline{t}_\alpha\right) = \alpha.$$

Exemple.

Un échantillon aléatoire de taille $n = 16$, de moyenne $m = 27.9$ et d'écart type $S = 3.23$, est pris d'une population distribuée normalement, dont l'écart type σ est inconnu.

Quelle est, au risque $\alpha = 0.05$, l'intervalle de confiance de la moyenne μ de cette population, centré sur la moyenne m de l'échantillon.

Solution.

Puisque l'écart type de la population est inconnu et l'échantillon est de petite taille $(n = 16 \leq 30)$, l'intervalle de confiance est

$$\begin{aligned} \mu &= m \pm \bar{t}_\alpha \frac{S}{\sqrt{n}} \\ &= 27.9 \pm \bar{t}_{0.05} \frac{3.23}{\sqrt{16}} \\ &= 27.9 \pm 2.131 \frac{3.23}{\sqrt{16}} \\ &= 27.9 \pm 1.72. \end{aligned}$$

α / ddl	0,90	0,50	0,30	0,20	0,10	0,05	0,02	0,01	0,001
1	0,158	1,000	1,963	3,078	6,314	12,706	31,821	63,657	636,619
2	0,142	0,816	1,386	1,886	2,920	4,303	6,965	9,925	31,598
3	0,137	0,765	1,250	1,638	2,353	3,182	4,541	5,841	12,924
4	0,134	0,741	1,190	1,533	2,132	2,776	3,747	4,604	8,610
5	0,132	0,727	1,156	1,476	2,015	2,571	3,365	4,032	6,869
6	0,131	0,718	1,134	1,440	1,943	2,447	3,143	3,707	5,959
7	0,130	0,711	1,119	1,415	1,895	2,365	2,998	3,499	5,408
8	0,130	0,706	1,108	1,397	1,860	2,306	2,896	3,355	5,041
9	0,129	0,703	1,100	1,383	1,833	2,262	2,821	3,250	4,781
10	0,129	0,700	1,093	1,372	1,812	2,228	2,764	3,169	4,587
11	0,129	0,697	1,088	1,363	1,796	2,201	2,718	3,106	4,437
12	0,128	0,695	1,083	1,356	1,782	2,179	2,681	3,055	4,318
13	0,128	0,694	1,079	1,350	1,771	2,160	2,650	3,012	4,221
14	0,128	0,692	1,076	1,345	1,761	2,145	2,624	2,977	4,140
15	0,128	0,691	1,074	1,341	1,753	2,131	2,602	2,947	4,073
16	0,128	0,690	1,071	1,337	1,746	2,120	2,583	2,921	4,015
17	0,128	0,689	1,069	1,333	1,740	2,110	2,567	2,898	3,965
18	0,127	0,688	1,067	1,330	1,734	2,101	2,552	2,878	3,922
19	0,127	0,688	1,066	1,328	1,729	2,093	2,539	2,861	3,883
20	0,127	0,687	1,064	1,325	1,725	2,086	2,528	2,845	3,850

Utilisation de la table 3 de l'écart réduit de la loi de Student pour déterminer $\bar{t}_{0.05}$: la valeur de $\bar{t}_{0.05}$ se trouve dans la cellule, intersection de la ligne du ddl (ici ddl=n-1=15) et la colonne de $\alpha = 0.05$.

2.4 Intervalle de confiance pour la différence $\mu_1 - \mu_2$

Soient

1)- $X_1, X_2, ..., X_m$ un échantillon aléatoire d'une population de moyenne μ_1 et d'écart-type σ_1.

2)- $Y_1, Y_2, ..., Y_n$ un échanillons aléatoire d'une population Y de moyenne μ_2 et d'écart-type σ_2.

3)- Les deux échantillons sont indépendants.

4)- $\overline{x}$ une valeur observée de la moyenne arithmétique $\overline{X}$ de l'echanillon de taille m de X et $\overline{y}$ une valeur observée de la moyenne arithmétique$\overline{Y}$ de l'echanillon de taille n de$Y$.

Une estimation naturelle de $\mu_1 - \mu_2$ est $\overline{x} - \overline{y}$ la différence entre les moyennes des deux échantillons.

Puisque les échantillons sont indépendants, on a

$$\begin{aligned} \mu_{\overline{X}-\overline{Y}} &= \mu_1 - \mu_2, \\ \sigma^2_{\overline{X}-\overline{Y}} &= \frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}. \end{aligned}$$

Le calcul d'intervalle de confiance pour $\mu_1 - \mu_2$, et également le calcul d'intervalle de pari, repose sur la détermination de la loi de probabilité de la variable aléatoire

$$\frac{\left(\overline{X} - \overline{Y}\right) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}}. \tag{2.7}$$

Cette variable aléatoire est la différence standarisée des deux variables aléatoires $\overline{X}$ et $\overline{Y}$.

2.4.1 Cas de grands échantillons ($m > 30$ et $n > 30$)

Dans le cas où les deux échantillons sont de grande taille (> 30) , le TCL assure que la variable aléatoire (2.7) suit (à peu prés) la loi normale réduite centrée, même dans le cas où les variances σ_X^2 et σ_Y^2 des populations mères sont remplacées par les variances S_X^2 et S_Y^2 des échantillons. Ainsi,

a)- dans le cas où σ_X^2 et σ_Y^2 sont connues,

$$\frac{\left(\overline{X} - \overline{Y}\right) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_X^2}{m} + \frac{\sigma_X^2}{n}}} \sim N(0,1) \tag{2.8}$$

b)- dans le cas où σ_X^2 et σ_Y^2 sont inconnues,

$$\frac{\left(\overline{X}-\overline{Y}\right)-(\mu_1-\mu_2)}{\sqrt{\frac{S_X^2}{m}+\frac{S_Y^2}{n}}} \sim N(0,1). \tag{2.9}$$

Intervalle de confiance pour $\mu_1-\mu_2$.

Ainsi, dans le cas de grands échantillons, l'intervalle de confiance au risque $0<\alpha<1$ pour la différence de moyennes $\mu_1-\mu_2$, calculé à partir des valeurs observées $\overline{x}$ et $\overline{y}$ de $\overline{X}$ et $\overline{Y}$, est

$$(\mu_1-\mu_2)=(\overline{x}-\overline{y})\pm\overline{z}_\alpha\sqrt{\frac{\sigma_X^2}{m}+\frac{\sigma_Y^2}{n}} \tag{2.10}$$

si les variances σ_X^2 et σ_Y^2 sont connues, et

$$\overline{x}-\overline{y}\pm\overline{z}_\alpha\sqrt{\frac{S_X^2}{m}+\frac{S_Y^2}{n}} \tag{2.11}$$

si les variances σ_X^2 et σ_Y^2 sont inconnues.

Intervalle de pari pour $\overline{X}-\overline{Y}$.

L'intervalle de pari au risque de $0<\alpha<1$, centré sur $(\mu_1-\mu_2)$, pour $\overline{X}-\overline{Y}$, est

$$(\overline{x}-\overline{y})=(\mu_1-\mu_2)\pm\overline{z}_\alpha\sqrt{\frac{\sigma_X^2}{m}+\frac{\sigma_Y^2}{n}}. \tag{2.12}$$

2.4.2 Cas où les lois mères sont normales, de variances σ_1^2 et σ_2^2 connues

Dans ce cas, comme dans le cas de grands échantillons, la variable aléatoire

$$\frac{\left(\overline{X}-\overline{Y}\right)-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{m}+\frac{\sigma_2^2}{n}}}, \tag{2.13}$$

différence standarisée des moyennes des deux échantillons, suit également la loi normale réduite centrée, par conséquent, le calcul d'intervalles de confiance et d'intervalles de pari se fait de la même façon.

Rappel

La variable standarisée (ou réduite centrée) d'une variable aléatoire X de moyenne μ et d'écart type σ (finis et $\sigma \neq 0$) est par définition la variable aléatoire

$$\frac{X-\mu}{\sigma}.$$

La moyenne de cette variable est zéro et sa variance est un.

2.4.3 Cas où les lois mères sont normales, mais de variances σ_1^2 et σ_2^2 inconnues

Contrairement au cas d'échantillons de grande taille, dans ce cas (cas où au moins m ou n n'est pas supérieur à 30), quand on remplace, dans l'expression (2.7) de la statistique de test, les variances σ_X^2 et σ_Y^2 des populations mères par les variances S_X^2 et S_Y^2 des échantillons, la variable aléatoire de l'expression (2.9) ne suit pas sensiblement la loi normale réduite centrée.

On se limite dans le présent cours au cas où les deux variances σ_1^2 et σ_2^2 sont identiques. Dans cette situation particulière, les mathématiciens ont montré que

1)- la variable aléatoire différence standarisée des moyennes des deux échantillons suit la loi de Student de degré de liberté $ddl = m + n - 2$.

2) Cette variable aléatoire est donnée par l'expression

$$\frac{\left(\overline{X} - \overline{Y}\right) - (\mu_1 - \mu_2)}{S_c\sqrt{\frac{1}{m} + \frac{1}{n}}}, \tag{2.14}$$

où S_c^2, appelée variance commune est donnée par la formule

$$S_c^2 = \frac{(m-1)\,S_X^2 + (n-1)\,S_Y^2}{m+n-2}. \tag{2.15}$$

Par conséquent, on a les formules suivantes pour les intervalles de confiance et les intervalles de pari, dans le cas où les lois mères sont normales, mais de variances σ_1^2 et σ_2^2 inconnues :

Intervalle de confiance L'intervalle de confiance pour la différence des moyennes $(\mu_1 - \mu_2)$, centré sur la différence des moyennes observées $\overline{x} - \overline{y}$ des échantillons, est

$$(\mu_1 - \mu_2) = (\overline{x} - \overline{y}) \pm \overline{t}_\alpha S_c \sqrt{\frac{1}{m} + \frac{1}{n}}, \tag{2.16}$$

où $\overline{t}_\alpha$ est calculé dans la table 3 de l'écart réduit de la loi de Student de dégré de liberté $ddl = m + n - 2$, selon la règle

$$P_r\left(|T_{m+n-2}| \geq \overline{t}_\alpha\right) = \alpha. \tag{2.17}$$

Intervalle de pari L'Intervalle de pari pour la différence $\overline{X} - \overline{Y}$ des moyennes des échantillons, centré sur la différence $(\mu_1 - \mu_2)$ des moyennes vraies, est

$$(\overline{x} - \overline{y}) = (\mu_1 - \mu_2) \pm \overline{t}_\alpha S_c \sqrt{\frac{1}{m} + \frac{1}{n}}. \tag{2.18}$$

2.5 Intervalle de confiance pour la fréquence vraie P

2.5.1 Introduction

On considère un caractère à deux modalités (succès-échec) sur une grande population (éventuellement infinie), on considère ensuite un échantillon de cette population de taille n et on note par X $(X \leq n)$ le nombre des succès de ce caractère dans cet échantillon. On note par P la fréquence des succès de ce caractère dans la population et par $\widetilde{P} = \frac{X}{n}$ la fréquence des succès du même caractère dans l'échantillon. On sait que $\widetilde{P}$ est un estimateur sans biais de P. On sait également que, pour n assez grand, X suit la loi binomiale de paramètres n et P, et que cette loi binomiale peut être approximée par la loi normale de moyenne $\mu = nP$ et d'écart type $\sigma = \sqrt{np\,(1-p)}$.

Ce résultat nous permet de construire un intervalle de confiance de la fréquence P de la population mère, centré sur la fréquence $\widetilde{P} = \frac{X}{n}$ de l'échantillon, par la formulation suivante :

Si la population mère est infinie (pratiquement trés grande) et la taille de l'échantillon est assez grande $(n > 30)$, un intervalle de confiance I_α de la fréquence P de la population mère, centré sur la fréquence $\widetilde{P}$ de l'échantillon, est donné par la formule

$$P = \widetilde{p} \pm \overline{z}_\alpha \sqrt{\frac{\widetilde{p}\,(1-\widetilde{p})}{n}}, \tag{2.19}$$

où $\overline{z}_\alpha$ est calculée dans la table 2.

Remarque.

1) La confection d'intervalles de confiances et d'intervalles de pari pour les fréquences dans le cas de petits échantillons ($n \leq 30$) est plus compliquée, car on ne peut pas y appliquer la loi normale. Le calcul d'intervalles de confiance et les intervalles de pari dans le cas de petits échantillons pour les fréquences n'est pas développé dans ce cours.

2) L'expression de l'intervalle de confiance développée ici est valable pour les grands échantillons (en pratique $n > 30$) et quand $n\widetilde{p}$ et $n\left(1-\widetilde{p}\right)$ ne sont pas voisins de zéro (en pratique supérieures à 5).

2.5.2 Exemple.

Dans un échantillon aléatoire de 85 patients d'une population de diabétiques traités par un certain médicament pour stabiliser le taux de glycémie, 10 patients avaient une réaction positive.

Utilisez le résultat obtenu sur cet échantillon pour trouver un intervalle de confiance au risque $\alpha = 0.05$, de la fréquence (ou proportion) P des patients ayant réagi positivement au traitement par ce médicament.

Solution.

On a

$$\widetilde{P} = \frac{10}{85} = 0.12 \text{ et } \overline{z}_{0.05} = 1.96.$$

Donc

$$I_{0.05} = 0.12 \pm 1.96\sqrt{\frac{0.12\left(1-0.12\right)}{85}} = 0.12 \pm 0.07$$

2.5.3 Intervalle de pari pour la fréquence $\widetilde{P}$ d'un échantillon

Dans le cas de grand échantillon ($n > 30$) et si nP et $n\left(1-P\right)$ ne sont pas voisins de zéro (en pratique supérieures à 5), un intervalle de pari $\widetilde{I}_\alpha$ pour $\widetilde{P}$ est donné par la formule

$$\widetilde{I}_\alpha = P \pm \overline{z}_\alpha\sqrt{\frac{P\left(1-P\right)}{n}}, \tag{2.20}$$

où $\overline{z}_\alpha$ est calculé de la table 2 de l'écart réduit de la loi normale réduite centrée.

Tableau récapitulatif

Intervalles de confiances : moyennes et fréquences

Données	Intervalle de confiance
- n>30 - σ connu	de la moyenne $\mu = m \pm \bar{z}_{\propto} \frac{\sigma}{\sqrt{n}}$
- n>30 - σ inconnu	$\mu = m \pm \bar{z}_{\propto} \frac{S}{\sqrt{n}}$
- $n \leq 30$ - la loi de la population est normale - σ connu	de la moyenne $\mu = m \pm \bar{z}_{\propto} \frac{\sigma}{\sqrt{n}}$
- $n \leq 30$ - la loi de la population est normale - σ inconnu	de la moyenne $\mu = m \pm \bar{t}_{\propto} \frac{S}{\sqrt{n}}$
- n>30 - $n\widetilde{P} > 5$ - $n(1-\widetilde{P}) > 5$	De la fréquence $P = \widetilde{P} \pm \bar{z}_{\propto} \sqrt{\frac{\widetilde{P}(1-\widetilde{P})}{n}}$

FIG. 2.1 – **Tableau récapitulatif sur les formules des intervalles de confiance.** On a utilisé les notation suivantes : n = taille de l'échantillon, S = écart-type de l'échantillon, μ = moyenne de la population, σ = écart-type de la population, P = fréquence de la population, $\widetilde{p}$ = fréquence de l'échantillon, $\overline{z}_{\alpha}$ est à calculer (pour α donné) dans la table 2 de l'écart réduit de la loi NRC, $\bar{t}_{\alpha}$ est à calculer (pour α donné) dans la table 3 de l'écart réduit de la loi de Student de degré de liberté ddl = n-1.

Chapitre 3

Tests d'hypothèses de conformité

3.1 Conformité de m à μ

Le concept de test d'hypothèses en statistique est une illustration frappante de l'utilité des sciences rationnelles dans la vie quotidienne.

3.1.1 Principe

Connaissant la moyenne vraie μ d'une certaine population et la moyenne m d'un échantillon aléatoire, peut-on accepter, avec une probabilité donnée $1-\alpha$, l'hypothèse (qu'on note H_0 et qu'on appelle hypothèse nulle) que l'échantillon est issu de cette population. Le principe général est le suivant : si m appartient à l'intervalle de pari $\widetilde{I}_\alpha$, on accepte l'hypothèse H_0, sinon on l'a rejette.

Soit donc une variable aléatoire X d'espérance μ et d'écart type σ, et soit $x_1, x_1, ...x_n$ une suite de n nombres de moyenne m et d'écart type S.

3.1.2 Cas de grand échantillons $(n > 30)$

Dans ce cas l'intervalle de pari est donné par

$$\widetilde{I}_\alpha = \left[\mu - \overline{z}_\alpha \frac{\sigma}{\sqrt{n}}, \mu + \overline{z}_\alpha \frac{\sigma}{\sqrt{n}}\right]. \tag{3.1}$$

L'intervalle de pari $\widetilde{I}_\alpha$ s'appelle dans ce contexte l'intervalle d'acceptation de l'hypothèse nulle H_0 et son complémentaire, l'intervalle de rejet de H_0.

Les étapes du test

1- On définit l'hypothèse $H_0 : m = \mu$, appelée hypothèse nulle, et l'hypothèse alternative $H_a : m \neq \mu$.

2- On cherche la valeur de $\overline{z}_\alpha = z_{1-\frac{\alpha}{2}}$ (appelé seuil critique) sur la table 2 de l'écart réduit ou sur la table 1 de la fonction de répartition, de la loi normale réduite centré.

3- On calcule ce qu'on appelle la statistique de test observée T_O par la formule

$$T_O = \frac{m - \mu}{\frac{\sigma}{\sqrt{n}}}. \tag{3.2}$$

4- Décision :

- si T_O appartient à l'intervalle d'acceptation $[-\overline{z}_\alpha, \overline{z}_\alpha]$, "au risque α, on accepte l'hypothèse H_0"

- si T_O n'appartient pas à l'intervalle d'acceptation $[-\overline{z}_\alpha, \overline{z}_\alpha]$, "on rejette l'hypothèse H_0 et on accepte l'hypothèse alternative H_a". On peut remplacer l'expression "au risque α" par l'expression "au niveau de signifiance α" ou par l'expression "au niveau de confiance $1 - \alpha$".

Dans le cas d'un grand échantillon, si σ est inconnu, on procède de la même façon, seulement, on remplace dans l'expression (3.2), permettant le calcul de T_O, l'écart type σ de la population mère par l'écart type S de l'échantillon.

Ces quatre étapes pour réaliser un test statistique sont communes aux différents genres de tests statistiques.

Mais, il faut faire attention ! La formule permettant le calcul de T_O observée n'est pas la même : chaque genre de test a sa formule. De même pour le seuil critique.

Dans la suite, on précisera pour chaque situation le seuil critique et la formule permettant le calcul de T_O observée.

3.1.3 Cas où la loi de X est normale et σ est connu

Dans ce cas la procédure est la même pour les grands ou les petits échantillons.

Le seuil critique est, comme dans le cas précédent, $\overline{z}_\alpha$, et donc l'intervalle d'acceptation est $[-\overline{z}_\alpha, \overline{z}_\alpha]$.

Egalement, l'expression permettant le calcul du T_O observée est la même, c'est-à-dire

$$T_O = \frac{m - \mu}{\frac{\sigma}{\sqrt{n}}}. \tag{3.3}$$

3.1.4 Test de conformité unilatéral

Si dans le test de comparaison de la moyenne m de l'échantillon à la moyenne μ de la population, l'hypothèse alternative est $m > \mu$ (ou $m < \mu$), au lieu de $m \neq \mu$, on dit qu'il s'agit de test unilatéral à droite (ou à gauche), au lieu de test bilatéral.

un test unilatéral est, comme le test bilatéral, composé de quatre étapes :

1. Hypothèse nulle $H_0 : m = \mu$,

Hypothèse alternative $H_a : m > \mu$ pour un test unilatéral à droite, (ou $m < \mu$ pour un test unilatéral à gauche).

2. la statistique de test observée T_O est la même que dans le cas bilatéral.

3. le seuil critique est $\overline{z}_{2\alpha}$ pour un test à droite ou $-\overline{z}_{2\alpha}$ pour un test à gauche (au lieu de $\overline{z}_\alpha$ pour un test bilatéral). Ce principe s'applique à tous les tests unilatéraux : on remplace α par 2α.

4. Décision.

a)- test à droite : si $T_O \leq \overline{z}_{2\alpha}$, au risque α, on accepte l'hypothèse nulle $H_0 : m = \mu$, sinon on rejette l'hypothèse nulle H_0 et on accepte l'hypothèse alternative $H_a : m > \mu$.

b)- test à gauche : si $T_O \geq -\overline{z}_{2\alpha}$, au risque α, on accepte l'hypothèse nulle $H_0 : m = \mu$, sinon on rejette l'hypothèse nulle H_0 et on accepte l'hypothèse alternative $H_a : m < \mu$.

Remarque importante.

Le test unilatéral à droite s'applique uniquement si, on a en hypothèse $m > \mu$, sinon le test est inutile.

Le test unilatéral à gauche s'applique uniquement si, on a en hypothèse $m < \mu$, sinon le test est inutile.

3.1.5 Exemple.

On reprend les données de l'exemple précédent.

En 1980, le temps de réaction moyen à des stimuli visuels simples pour des étudiants en psychologie (âge moyenne = 20 ans) était de 200 millisecondes. On choisit cette année (année 2019) au hasard un échantillon de 36 étudiants

de 20 ans qu'on soumet à la même expérimentation, afin de tester l'effet de Flynn. Le temps de réaction moyen observé pour l'échantillon est de 180 millisecondes avec un écart-type de 10 millisecondes.

Tester au seuil de signifiance de 5% l'hypothèse selon laquelle le temps de réaction moyen en 2019 est inférieur à celui de 1980 !

Solution

1)- On teste l'hypothèse nulle $H_O : m = \mu$ (où $m = 180$ et $\mu = 200$), contre l'hypothèse alternative $H_a : m < \mu$.

2)- Puisque l'échantillon est de grande taille $(n = 36 > 30)$, le seuil critique est $\overline{z}_{2\alpha} = \overline{z}_{0.1} = 1.645$.

3)- La statistique de test observée T_O est la même que pour le test bilatéral : $T_O = -12$.

4)- Décision : puisque $T_O = -12 < -1.645$, on rejette l'hypothèse nulle H_0 et on accepte l'hypothèse alternative H_a : on peut affirmer au seuil de signifiance de 5%, que le temps de réaction moyen en 2019 est inférieur à celui de 1980.

3.1.6 Cas où la loi de X est normale, $n < 30$ et σ est inconnu

Dans ce cas

1- la statistique de test suit la loi de Student de degré de liberté $ddl = n-1$, et le seuil critique noté $\overline{t}_\alpha$ est obtenu à partir de la table 3, de l'écart réduit de la loi de Student-Fisher, liant $\overline{t}_\alpha$ à α par la relation $P(|T| \geq \alpha) = \overline{t}_\alpha$.

2- l'expression permettant le calcul de T observée est

$$T_O = \frac{m - \mu}{\frac{S}{\sqrt{n}}} \tag{3.4}$$

où S l'écart type de l'échantillon a remplacé l'écart type σ (inconnu) de la loi mère.

3.1.7 Exemple.

On prélève au hasard dix comprimés d'un grand lot d'un certain médicament, et on les pèse. les poids de ces 10 comprimés en grammes sont :

0.81 ; 0.84 ; 0.83 ; 0.80 ; 0.85 ; 0.81 ; 0.85 ; 0.83 ; 0.84 ; 0.80.

A- Le poids moyen observé est-il compatible au niveau de confiance de 98 % avec le poids moyen de 0.84 g annoncé par le producteur ?

B- Au risque 5% de se tromper, le poids moyen observé est-il inférieur à celui annoncé par le producteur ?

Solution.

A-

On réalise un test bilatéral de conformité.

Etape 1. H_0 : le poids moyen observé est compatible avec le poids annoncé par le producteur ($m = \mu$),

H_a : le poids moyen observé n'est pas compatible avec le poids annoncé par le producteur ($m \neq \mu$).

Etape 2. La statistique de test observée est

$$T_O = \frac{m - \mu}{\frac{S}{\sqrt{n}}}.$$

Pour obtenir sa valeur numérique on calcule la moyenne m de l'échantillon et son écart type S :

$$\begin{aligned} m &= \frac{(2 \times 0.80) + (2 \times 0.81) + (2 \times 0.83) + (2 \times 0.84) + (2 \times 85)}{10} \\ &= \mathbf{0.826} \end{aligned}$$

$$\begin{aligned} S &= \sqrt{\frac{\sum_{i=1}^{10} n_i \left(x_i - m\right)^2}{N - 1}} \\ &= \sqrt{\frac{2\left(0.80 - 0.826\right)^2 + ... + 2\left(85 - 0.826\right)^2}{9}} \\ &= \mathbf{0.0196}. \end{aligned}$$

Remplaçant m et S par leurs valeurs on aura

$$T_O = \frac{0.826 - 0.84}{\frac{0.0196}{\sqrt{10}}} = \mathbf{-2.265}$$

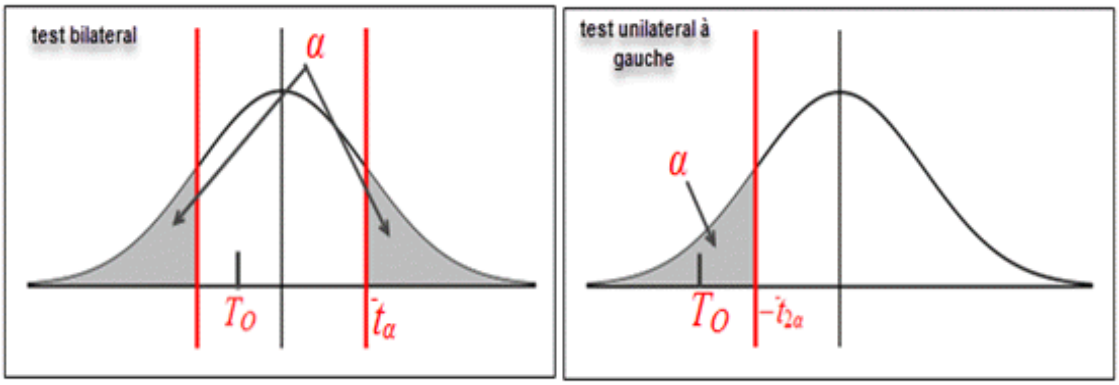


FIG. 3.1 – A)- Test bilatéral : T_0 est dans la zone d'acceptation de l'hypothèse nulle H_0, par conséquent, on accepte H_0 et on rejette l'hypotèse alternative H_a. B)-Test unilatéral à gauche : T_0 n'est pas dans la zone d'acceptation de l'hypothèse nulle H_0, par conséquent, on rejette H_0 et on accepte l'hypotèse alternative H_a

Etape 3. On calcule le seuil critique (loi de Student avec $ddl = n - 1 = 9$) : le seuil critique est

$$\bar{t}_\alpha = \bar{t}_{0.02} = \mathbf{2.821}$$

Etape 4. Décision

Puisque la statistique de test observée $T_O = \mathbf{-2.265}$ appartient à l'intervalle $[\mathbf{-2.821}; \mathbf{2.821}]$ d'acceptation de H_0, on rejette l'hypothèse alternative H_a et on accepte l'hypothèse nulle H_0 : Au niveau de confiance de 98 %, le poids moyen observé est compatible avec le poids annoncé par le producteur.

B-

On réalise un test unilatéral à gauche.

1. Hypothèse nulle $H_0 : m = \mu$, contre l'hypothèse alternative $H_1 : m < \mu$.

2. La statistique de test observée T_O est la même que pour le test bilatéral : $T_O = -2.265$.

3. le seuil critique est $\bar{t}_{2\alpha} = \bar{t}_{0.1} = 1.833$ (table 3 de student $ddl = 9$).

4. Décision : puisque $T_O = -2.265 < -\bar{t}_{2\alpha} = -1.833$, on rejette l'hypothèse nulle H_0 et on accepte l'hypothèse alternative H_1 : au risque 5% de se tromper, le poids moyen observé est inférieure à celui annoncé par le producteur.

3.2 Comparaison de $\widetilde{p}$ calculée et P vraie

3.2.1 Test bilatéral de l'hypothèse $\left(\widetilde{P} = P\right)$

Soit P la proportion des succès d'un caractère à deux modalités (succès-échec) dans une grande population, et soit $\widetilde{P}$ la proportion des succès du même caractère dans un échantillon fini (pas forcément issu de la même population). Peut-on affirmer avec une probabilité α de se tromper que l'échantillon est issu de cette population (ou d'une population avec la même proportion P). Pour répondre à cette question on réalise un test de l'hypothèse $\widetilde{P} = P$ contre l'hypothèse alternative $\widetilde{P} \neq P$. Ce test se fait à travers les étapes suivantes.

1- On définit l'hypothèse nulle $H_0 : \widetilde{P} = P$ et l'hypothèse alternative $H_1 : \widetilde{P} \neq P$

2- La statistique de test suit la loi normale réduite centrée, On calcule le T_O observé par la formule

$$T_O = \frac{\widetilde{P} - P}{\sqrt{\frac{\widetilde{P}(1-\widetilde{P})}{n}}}. \tag{3.5}$$

3- On calcule le seuil critique $\overline{z}_\alpha$ dans la table 2 de l'écart réduit de la loi normale réduite centrée.

4- Décision. Si

$$-\overline{z}_\alpha \leq T_O \leq \overline{z}_\alpha,$$

au risque α, on accepte l'hypothèse $\widetilde{P} = P$, sinon, on rejette l'hypothèse nulle $\widetilde{P} = P$ et on accepte l'hypothèse alternative $\widetilde{P} \neq P$.

3.2.2 Test unilatéral de l'hypothèse $(\widetilde{p} = P)$

Les quatre étapes restent les mêmes, avec les changements suivants :

l'hypothèse alternative devient $H_{a:} : \widetilde{p} > P$ pour un test à droite ou $H_{a:} : \widetilde{P} < P$ pour un test à gauche.

Pour calculer le seuil critique on considère 2α au lieu de α.

Décision : a)- si $T_O \leq \overline{z}_{2\alpha}$, on accepte l'hypothèse nulle, sinon on la rejette (test à droite)

b)- si $T_O \geq -\overline{z}_{2\alpha}$, on accepte l'hypothèse nulle, sinon on la rejette (test à gauche).

Remarque.

Le test présenté ici est valable dans le cas de grand échantillon ($n > 30$) et si nP et $n(1-P)$ ne sont pas voisins de zéro (en pratique supérieures à 5).

3.2.3 Exemple.

Une anomalie génétique touche dans un certain pays 1/1000 des individus. Dans une région donnée de ce pays, on a enregistré 57 personnes ayant cette anomalie sur 50 000 naissances.

A)- Cette région est-elle représentative du pays entier au risque 5 % ?

B)- La proportion de cette anomalie génétique enregistrée dans cette région est-elle plus importante que celle du pays entier ?.

Solution.

A)-

On répond à cette question en réalisant un test de comparaison d'une proportion calculée $\widetilde{P} = \frac{57}{5}10^{-4}$ et une proportion vraie $P = 10^{-3}$.

1- L'hypothèse nulle $H_0 : \widetilde{P} = P$ (cette région est représentative du pays entier au risque 5 %),

l'hypothèse alternative $H_a : \widetilde{P} \neq P$ (cette région n'est pas représentative du pays entier)

2- La statistique de test suit la loi normale réduite centrée, On calcule la statistiqueT_O observée par la formule

$$T_O = \frac{\widetilde{P} - P}{\sqrt{\frac{P(1-P)}{n}}}.$$

Puisque

$$n = 5.10^4, \widetilde{p} = \frac{57}{5}10^{-4} \text{ et } P = 10^{-3},$$

on a

$$T_O = \frac{\frac{57}{5}.10^{-4} - 10^{-3}}{\sqrt{\frac{10^{-3}(1-10^{-3})}{5.10^4}}} = \mathbf{0.99}.$$

3- Le seuil critique est

$$\overline{z}_\alpha = z_{0.05} = \mathbf{1.96}.$$

4- Décision. Puisque

$$-1.96 \leq 0.99 \leq 1.96,$$

au risque 0.05, on accepte l'hypothèse $\widetilde{P} = P$.

B)-

On répond à cette question en réalisant un test unilatéral à droite, de comparaison d'une proportion calculée $\widetilde{P} = \frac{57}{5}.10^{-4}$ et une proportion vraie $P = 10^{-3}$.

L'hypothèse nulle $H_0 : \widetilde{p} = P$, l'hypothèse alternative $H_a : \widetilde{p} > P$.

La statistique de test observée T_O reste la même : 0.99

Le seuil critique dans le cas unilatéral est $\overline{z}_{2\alpha}$ au lieu de $\overline{z}_\alpha$:

Donc le seuil critique est

$$\overline{z}_{2\alpha} = \overline{z}_{0.1} = \mathbf{1.645}.$$

Décision. Puisque la T observée $T_O = \mathbf{0.99}$ n'est pas supérieur au seuil critique $\overline{z}_{0.1} = \mathbf{1.645}$, on accepte H_0 au risque 0.05.

Tableau récapitulatif 2 : Tests Statistiques de conformité: moyennes et fréquences

Genre de test	Données	Statistique de test observée T_0	Seuil critique SC		
			Test bilatéral	Test unilatéral à droite	Test unilatéral à gauche
Comparaison de la moyenne m d'un échantillon et la moyenne vraie μ.	n>30 σ connu	$T_0 = \frac{\sqrt{n}(m-\mu)}{\sigma}$	$\bar{z}_\alpha$	$\bar{z}_{2\alpha}$	$-\bar{z}_{2\alpha}$
	n>30 σ inconnu	$T_0 = \frac{\sqrt{n}(m-\mu)}{s}$	$\bar{z}_\alpha$	$\bar{z}_{2\alpha}$	$-\bar{z}_{2\alpha}$
	n ≤ 30, la loi de la population est normale et σ connu	$T_0 = \frac{\sqrt{n}(m-\mu)}{\sigma}$	$\bar{z}_\alpha$	$\bar{z}_{2\alpha}$	$-\bar{z}_{2\alpha}$
	n ≤ 30, la loi de la population est normale et σ inconnu	$T_0 = \frac{\sqrt{n}(m-\mu)}{s}$	$\bar{t}_\alpha$	$\bar{t}_{2\alpha}$	$-\bar{t}_{2\alpha}$
Comparaison d'une proportion $\tilde{p}$ calculée sur un échantillon et une proportion vraie P	n > 30, $nP > 5$ $n(1-P) > 5$	$T_0 = = \frac{\sqrt{n}(\tilde{p}-\mathrm{P})}{\sqrt{P(1-P)}}$	$\bar{z}_\alpha$	$\bar{z}_{2\alpha}$	$-\bar{z}_{2\alpha}$

Notation. n = taille de l'échantillon, m = moyenne de l'échantillon, $\mu = moyenne\ de\ la\ population$, P = Proportion vraie, $\tilde{p} = proportion\ de\ l'échantillon$ $\alpha = le\ risque$, S = écart type de l'échantillon, σ = écart type de la population, $\bar{z}_\alpha$ et $\bar{z}_{2\alpha}$ à calculer dans la table 2 de l'écart réduit de la loi normale, $\bar{t}_{2\alpha}$ à calculer dans la table 3 de l'écart réduit de la loi de Student de ddl = n-1, SC = seuil critique.

Décision Test bilatéral : si $|T_0| \leq SC$ on accepte l'hypothèse nulle H_0, sinon on la rejette.
Décision Test unilatéral à droite : si $T_0 \leq SC$ on accepte l'hypothèse nulle H_0, sinon on la rejette.
Décision Test unilatéral à gauche: si $T_0 \geq SC$ on accepte l'hypothèse nulle H_0, sinon on la rejette.

FIG. 3.2 – Tableau récapitulatif : tests de conformité, moyennes et fréquences

Chapitre 4

Tests d'hypothèses d'homogénéité

4.1 Comparaison de deux moyennes μ_1 et μ_2

Soient

1)- $X_1, X_2, ..., X_m$ un échantillon aléatoire d'une population de moyenne μ_X et d'écart-type σ_X.

2)- $Y_1, Y_2, ..., Y_n$ un échantillon aléatoire d'une population Y de moyenne μ_Y et d'écart-type σ_Y.

3)- Les deux échantillons sont indépendants.

4)- $\overline{x}$ une valeur observée de la moyenne arithmétique $\overline{X}$ de l'échantillon de la première population et $\overline{y}$ une valeur observée de la moyenne arithmétique $\overline{Y}$ de l'échantillon de la deuxième population.

A partir de ces données on se propose de tester l'hypothèse nulle "$\mu_X = \mu_Y$", contre l'hypothèse alternative " $\mu_X \neq \mu_Y$".

4.1.1 Cas de grands échantillons ($m > 30$ et $n > 30$)

a)- dans le cas où σ_X^2 et σ_Y^2 sont connues, la statistique de test, pour tester l'hypothèse nulle $\mu_X - \mu_Y = 0$ est donnée par la formule

$$\frac{\left(\overline{X} - \overline{Y}\right)}{\sqrt{\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}}}, \tag{4.1}$$

et elle suit la loi normale réduite centrée $Z \sim N(0, 1)$.

b)- dans le cas où σ_X^2 et σ_Y^2 sont inconnues, on les remplace par les variances S_X^2 et S_Y^2 des échantillons. La statistique de test devient

$$\frac{\left(\overline{X}-\overline{Y}\right)}{\sqrt{\frac{S_X^2}{m}+\frac{S_Y^2}{n}}}. \tag{4.2}$$

La statistique de test suit encore dans ce cas la loi normale réduite centrée.

4.1.2 Test bilatéral de l'hypothèse $\mu_X - \mu_Y = 0$.

Etape 1. On définit l'hypothèse nulle $H_0 : \mu_X - \mu_Y = 0$ (ou $\mu_X = \mu_Y$) et l'hypothèse alternative $H_a : \mu_X - \mu_Y \neq 0$.

Etape 2. La statistique de test observée T_O est donnée par l'expression

$$T_O = \frac{(\overline{x}-\overline{y})}{\sqrt{\frac{\sigma_X^2}{m}+\frac{\sigma_Y^2}{n}}} \tag{4.3}$$

dans le cas où σ_X^2 et σ_Y^2 sont connues, et

$$T_O = \frac{(\overline{x}-\overline{y})}{\sqrt{\frac{S_X^2}{m}+\frac{S_Y^2}{n}}} \tag{4.4}$$

dans le cas où σ_X^2 et σ_Y^2 sont inconnues.

Etape 3. On calcule le seuil critique $\overline{z}_\alpha$, en utilisant la table 2. de l'écart réduit de la loi normale réduite centrée.

Etape 4. Décision : si $T_O \in [-\overline{z}_\alpha, \overline{z}_\alpha]$, au risque α, on accepte l'hypothèse H_0. Sinon, on rejette l'hypothèse nulle H_0 et on accepte l'hypothèse alternative H_1.

4.1.3 Test unilatéral de l'hypothèse $\mu_X - \mu_Y = 0$.

1. L'hypothèse nulle $H_0 : \mu_X - \mu_Y = 0$ et l'hypothèse alternative $H_1 : \mu_X - \mu_Y > 0$ pour un test à droite, ou $H_1 : \mu_X - \mu_Y < 0$ pour un test à gauche.

2. On calcule la statistique de test observée T_O, qui est la même que pour le test bilatéral.

3. Seuil critique, qui est dans ce cas $\overline{z}_{2\alpha}$ (au lieu de $\overline{z}_\alpha$ dans le test bilatéral) pour un test à droite, ou $-\overline{z}_{2\alpha}$ pour un test à gauche.

4. Décision :

a)- test à droite : si $T_O \leq \overline{z}_{2\alpha}$, au risque α, on accepte l'hypothèse nulle $H_0 : \mu_X - \mu_Y = 0$, sinon (i.e si $T_O > \overline{z}_{2\alpha}$), on rejette l'hypothèse nulle H_0 et on accepte l'hypothèse alternative $H_a : \mu_X - \mu_Y > 0$

b)– test à gauche : si $T_O \geq -\overline{z}_{2\alpha}$, au risque α, on accepte l'hypothèse nulle $H_0 : \mu_X - \mu_Y = 0$, sinon (i.e si $T_O < -\overline{z}_{2\alpha}$), on rejette l'hypothèse nulle H_0 et on accepte l'hypothèse alternative $H_a : \mu_X - \mu_Y < 0$.

4.1.4 Exemple.

Dans une maternité pour deux échantillons de nouveau-nés de sexes différents on a obtenu les résultats suivants :

51 garçons : taille moyenne $51cm$ et écart-type des tailles $S_1 = 3cm$

59 filles : taille moyenne $49cm$ et écart-type des tailles $S_2 = 3.2cm$

A. Au risque 5 % peut-on déduire de ces données une différence significative entre les moyennes des tailles des nouveau-nés suivant le sexe.

B. Au risque 5 % peut-on déduire de ces données que la moyenne des tailles des nouveau-nés (garçons) est plus grande que celle des nouveau-nées (filles).

Solution.

A.

1. Hypothèse nulle $H_0 : \mu_1 = \mu_2$ (pas de différence significative) , contre Hypothèse alternative $H_a : \mu_1 \neq \mu_2$, (il existe une différence significative), c'est un test bilatéral.

2. Puisque les échantillons sont grands $n_1 = 51 > 30$ et $n_2 = 59 > 30$, et les écarts types des populations sont inconnus, la statistique de test observée est

$$z_O = \frac{(m_1 - m_2)}{\sqrt{\frac{S_1^2}{m} + \frac{S_2^2}{n}}} = \frac{(51 - 49)}{\sqrt{\frac{9}{51} + \frac{10.24}{59}}} = \mathbf{3.38}$$

3. On calcule le seuil critique $\overline{z}_\alpha$:

La table 2 donne

$$\overline{z}_\alpha = \overline{z}_{0.05} = 1.96$$

4. Décision.

$$T_O = 3.38 \notin [-1.96; 1.96]\,.$$

Puisque la statistique de test observée T_O n'appartient pas à l'intervalle d'acceptation de H_0, on rejette H_0 et on accepte H_a : on rejette l'hypothèse qu'il n'y a pas une différence significative entre les moyennes des tailles selon le sexe et on accepte l'hypothèse alternative qu' il y a une différence significative entre les moyennes des tailles selon le sexe.

B.

1. Hypothèse nulle $H_0 : \mu_1 = \mu_2$ (pas de différence significative), contre Hypothèse alternative $H_a : \mu_1 > \mu_2$, c'est un test unilatéral à droite.

2. la statistique de test observée est la même, $T_O = \mathbf{3.38}$.

3. Le seuil critique est

$$\overline{z}_{2\alpha} = \overline{z}_{0.1} = \mathbf{1.645}$$

4. Décision : Puisque

$$T_O = \mathbf{3.38} > \mathbf{1.645},$$

on rejette l'hypothèse nulle et on accepte l'hypothèse alternative : on peut déduire de ces indications que la taille moyenne des nouveau-nés (garçons) est plus grande que celle des nouveau-nées (filles).

Remarques.

1)- Dans les tests unilatéraux, le qualificatif à gauche (ou à droite) est lié à la zone de rejet de l'hypothèse nulle H_0 : si la zone de rejet est à gauche de zéro, on dit test à gauche, sinon on dit test à droite.

2)- Le seuil critique pour un test unilatéral se calcule de la même façon que dans le cas de test bilatéral, mais en remplaçant α par 2α. En plus pour le test à gauche, le seuil critique est pris avec le signe négatif.

4.1.5 Cas de lois normales et variances σ_X^2 et σ_Y^2 connues

Dans ce cas, la statistique de test, comme dans le cas de grand échantillon, a la même expression et suit la même loi. Par conséquent, le calcul de tests se fait de la même façon.

4.1.6 Cas de lois normales, et variances σ_X^2 et σ_Y^2 inconnues

Contrairement au cas d'échantillons de grandes tailles, dans ce cas (cas où au moins m ou n n'est pas plus grande que 30), quand on remplace, dans l'expression de la statistique de test, les variances σ_X^2 et σ_Y^2 des populations mères par les variances S_X^2 et S_Y^2 des échantillons, la statistique de test ne suit pas sensiblement la loi normale réduite centrée. Dans ce cas il n'y a pas une méthode unifiée pour traiter le problème. On se limite dans le présent exposé au cas où les deux variances σ_1^2 et σ_2^2 sont identiques. Dans cette situation particulière, les mathématiciens ont montré que

1) La statistique de test est

$$\frac{\left(\overline{X}-\overline{Y}\right)}{S_c\sqrt{\frac{1}{m}+\frac{1}{n}}}, \tag{4.5}$$

où S_c^2, appelée variance commune est donnée par la formule suivante

$$S_c^2=\frac{(m-1)\,S_X^2+(n-1)\,S_Y^2}{m+n-2}. \tag{4.6}$$

2)- La statistique de test suit la loi de Student de degré de liberté $ddl = m+n-2$.

Remarque

Si l'égalité des variances des deux populations n'est pas donnée en hypothèse, on peut éventuellement appliquer ce test aprés avoir vérifier l'égalité des variance par l'application du test de Fisher Snedecor (qui sera présenté ultérieurement).

4.1.7 Exemple

On veut tester l'efficacité d'un nouveau traitement en mesurant la durée de survie de souris atteintes par une maladie. On suppose que cette durée de survie est une variable aléatoire normale. Un premier échantillon de 9 souris est soigné à l'aide de l'ancien traitement et un autre effectif de 12 à l'aide d'un nouveau traitement.

Pour le premier échantillon, on a obtenu $m_1 = 21$ jours et $S_1^2 = 11.39$ jours.

Pour le second échantillon, on a obtenu $m_2 = 24$ jours et $S_2^2 = 10.12$ jours.

A.

Sachant que la variance est la même pour les deux traitements, peut-on affirmer en prenant un risque de 5 % qu'il y a une différence significative entre les deux traitements ?

B. Sachant que la variance est la même pour les deux traitements, peut-on affirmer en prenant un risque de 5 % que l'ancien traitement est moins efficace que le nouveau traitement ?

Solution

A.

On applique un test d'hypothèse.

Etape 1 : Hypothèse nulle $H_0 : \mu_1 = \mu_2$ contre $H_a : \mu_1 \neq \mu_2$; c'est un test bilatéral

Etape 2 : la variance commune S_C^2 est

$$\begin{aligned} S_C^2 &= \frac{8 \times 11.39 + 11 \times 10.12}{9 + 12 - 2} = 10.65. \\ S_C &= \sqrt{10.65} = 3.26 \end{aligned}$$

Donc, la statistique de test observée est

$$T_O = \frac{(m_1 - m_2)}{S_c\sqrt{\frac{1}{m} + \frac{1}{n}}} = \frac{(21 - 24)}{3.26\sqrt{\frac{1}{9} + \frac{1}{12}}} = \mathbf{-2.09}.$$

Etape 3 : Le seuil critique est $\bar{t}_{0.05}$, et il se calcule de la table 3. de l'écart réduit de la loi de Student de $ddl = 19$.

$$\bar{t}_{0.05} = \mathbf{2.093}.$$

Etape 4 : Décision : $T_O = \mathbf{-2.09} \geq -2.093$, donc $T_O \in [-2.093; 2.093]$. On rejette l'hypothèse alternative $H_1 (\mu_1 \neq \mu_2)$ et on accepte l'hypothèse nulle $H_0 (\mu_1 = \mu_2)$.

On peut donc affirmer, au risque 5% qu'il n'y a pas une différence significative entre les deux traitements.

B.

Hypothèse nulle $H_0 : \mu_1 = \mu_2$ contre $H_1 : \mu_1 < \mu_2$; c'est un test unilatéral à gauche.

La statistique de test observée est la même : $T_O = \mathbf{-2.09}$.

Le seuil critique est $-\bar{t}_{2\alpha} = -t_{0.1} = -1.729$.

Décision : puisque $T_O = \mathbf{-2.09} < -1.729$, on rejette l'hypothèse nulle et on accepte alternative H_a : au risque de 5 %, l'ancien traitement est moins efficace que le nouveau traitement ?

4.2 Comparaison de P_1 et P_2

4.2.1 Introduction

On considère un même caractère avec deux modalités (succès - échec) dans deux populations. Soient P_1 la proportion des succès de ce caractère dans la première population et P_2 la proportion des succès dans la deuxième population. On considère un échantillon de la première population de taille m et un autre échantillon de la deuxième population de taille n. Si X et Y sont respectivement le nombre de succès dans le premier et dans le deuxième échantillon, alors $\widetilde{P}_1 = \frac{X}{m}$ et $\widetilde{P}_2 = \frac{Y}{n}$ sont respectivement les proportions du succès dans le premier et dans le deuxième échantillon.

Si les tailles m et n des échantillons sont petites devant les tailles des populations, alors les variables aléatoires X et Y suivent approximativement la loi binomiale :

$$X \sim B\left(m; P_1\right) \text{ et } Y \sim B\left(n; P_2\right).$$

En outre, si m et n sont assez grands (> 30), les lois binomiales peuvent êre approximées par une loi normale. Cela ramène à déduire que la variable aléatoire

$$\frac{\left(\widetilde{P}_1 - \widetilde{P}_2\right) - \left(P_1 - P_2\right)}{\sqrt{\frac{P_1(1-P_1)}{m} + \frac{P_2(1-P_2)}{n}}} \tag{4.7}$$

suit la loi normale réduite centrée. On peut également estimer dans le dénominateur les proportions vraies P_1 et P_2 par les proportions $\widetilde{p}_1$ et $\widetilde{p}_2$ des échantillons et dire que

$$\frac{\left(\widetilde{P}_1 - \widetilde{P}_2\right) - \left(P_1 - P_2\right)}{\sqrt{\frac{\widetilde{p}_1(1-\widetilde{p}_1)}{m} + \frac{\widetilde{p}_2(1-\widetilde{p}_2)}{n}}} \tag{4.8}$$

suit aussi la loi normale réduite centrée.

Le raisonnement étalé ci-dessus est valables dans le cas de grands échantillons. Dans le cas de petits échantillons, la situation se complique un peu. En ce qui nous concerne, par souci de ne pas trop compliquer les choses, on se limite au cas de grands échantillons.

Intervalles de confiance pour $P_1 - P_2$

Le calcul d' intervalles de confiance pour la différence $P_1 - P_2$ des proportions vraies, à partir de la différence $\widetilde{P}_1 - \widetilde{P}_2$ des proportions observées des échantillons, est basé sur la loi de la statistique (4.8) .

L'intervalle de confiance au risque $0 < \alpha < 1$ de la différence $P_1 - P_2$ des proportions vraies, basé sur la différence $\widetilde{p}_1 - \widetilde{p}_2$ des proportions calculées des échantillons est

$$(P_1 - P_2) = (\widetilde{p}_1 - \widetilde{p}_2) \pm \overline{z}_\alpha \sqrt{\frac{\widetilde{p}_1 (1 - \widetilde{p}_1)}{m} + \frac{\widetilde{p}_2 (1 - \widetilde{p}_2)}{n}}. \tag{4.9}$$

Test d'hypothèse

A partir des proportions calculées $\widetilde{p}_1$ et $\widetilde{p}_2$ des échantillons, on se propose de tester l'hypothèse nulle $P_1 = P_2$ contre une hypothèse alternative bilatérale ou unilatérale.

On rappelle que la statistique de test est construite sous l'hypothèse nulle (ici $H_0 : P_1 = P_2$).

Sous l'hypothèse H_0 la statistique de test donnée par (4.8) se simplifie de la façon suivante. Sous H_0 les deux échantillons sont de la même population de proportion P $((P_1 = P_2 = P))$ dont L'union forme un échantillon. La proportion observée des succès dans ce grand échantillon est

$$\widetilde{P} = \frac{X + Y}{m + n},$$

qu'on appelle proportion commune. La statistique de test devient donc

$$\frac{\left(\widetilde{P}_1 - \widetilde{P}_2\right)}{\sqrt{\widetilde{P}\left(1 - \widetilde{P}\right)\left(\frac{1}{m} + \frac{1}{n}\right)}}. \tag{4.10}$$

4.2.2 Test bilatéral de l'hypothèse $P_1 = P_2$

1. L'hypothèse nulle est $H_0 : P_1 = P_2$ et l'hypothèse alternative est $H_1 : P_1 \neq P_2$.

2. La valeur observée de la statistique de test est

$$T_O = \frac{(\widetilde{p}_1 - \widetilde{p}_2)}{\sqrt{\widetilde{p}\,(1-\widetilde{p})\left(\frac{1}{m}+\frac{1}{n}\right)}}. \tag{4.11}$$

3. Le seuil critique est $\overline{z}_\alpha$ à calculer dans la table 2 de l'écart réduit de la loi normale réduite centrée.

4. Décision :

- si T_O appartient à la zone d'acceptation $(-\overline{z}_\alpha \leq T_O \leq \overline{z}_\alpha)$, on accepte l'hypothèse H_0,

- sinon $(T_O < -\overline{z}_\alpha$ ou $T_O > \overline{z}_\alpha)$, on rejette l'hypothèse H_0 et on accepte l'hypothèse alternative H_a.

4.2.3 Test unilatéral de l'hypothèse $P_1 = P_2$

1. L'hypothèse nulle est $H_0 : P_1 = P_2$ et l'hypothèse alternative est $H_1 : P_1 > P_2$ pour un test à droite et $H_1 : P_1 < P_2$ pour un test à gauche.

2. La valeur observée T_O de la statistique de test est la même que dans le cas de test bilatéral.

3. Le seuil critique est $\overline{z}_{2\alpha}$ pour un test à droite, et $-\overline{z}_{2\alpha}$ pour un test à gauche, à calculer sur la table 2 de l'écart réduit de la loi normale réduite centrée.

4. Décision :

a) test à droite : si T_O appartient à la zone d'acceptation de H_0 $(T_O \leq \overline{z}_{2\alpha})$, on accepte l'hypothèse H_0, sinon $(T_O > \overline{z}_{2\alpha})$, on rejette l'hypothèse H_0 et on accepte l'hypothèse alternative H_a.

b)- test à gauche : si T_O appartient à la zone d'acceptation $(T_O \geq -\overline{z}_{2\alpha})$, on accepte l'hypothèse H_0, sinon $(T_O < -\overline{z}_{2\alpha})$, on rejette l'hypothèse H_0 et on accepte l'hypothèse alternative H_a.

4.2.4 Exemple.

On a obtenu les résultats suivants après avoir suivi pendant 20 ans un groupe de 200 sujets fumeurs et un groupe de 200 sujets non fumeurs.

	non fumeurs	fumeurs
apparition d'un cancer	20	40
pas de cancer	180	160

A. Peut-on accepter au risque $\alpha = 0.1$ que les différences observées sont significatives ?.

B. Peut-on accepter au risque $\alpha = 0.1$ que l'abstention de fumer démunie le risque d'attraper un cancer ?

Solution

On applique un test de comparaison de deux fréquences.

On note par P_1 et par P_2 la proportion des cancéreux non fumeurs et celle des cancéreux fumeurs respectivement.

A.

1. Hypothèse nulle $H_0 : P_1 = P_2$, hypothèse alternative $H_a : P_1 \neq P_2$, il s'agit de test bilatéral

2. La proportion commune $\widetilde{P}$ est

$$\widetilde{P} = \frac{X+Y}{m+n} = \frac{20+40}{200+200} = 0.15,$$

la valeur observée de la statistique de test est

$$\begin{aligned} T_O &= \frac{(\widetilde{p}_1 - \widetilde{p}_2)}{\sqrt{\widetilde{p}\,(1-\widetilde{p})\left(\frac{1}{m}+\frac{1}{n}\right)}} \\ &= \frac{\left(\frac{20}{200} - \frac{40}{200}\right)}{\sqrt{(0.15)\,(0.85)\left(\frac{1}{200}+\frac{1}{200}\right)}} \\ &= -2.801. \end{aligned}$$

3. Le seuil critique est

$$\overline{z}_\alpha = \overline{z}_{0.1} = 1.645 \text{ (table 2)}$$

4. Décision : Puisque $|T_O| = |-2.801| = 2.801 > z_{0.1} = 1.645$, on rejette l'hypothèse nulle H_0, et on accepte l'hypothèse alternative H_1 : Au risque $\alpha = 0.1$, on accepte que les différences observées sont significatives.

B.

1. Hypothèse nulle $H_0 : P_1 = P_2$, hypothèse alternative $H_a : P_1 < P_2$, il s'agit de test unilatéral à gauche.

2. La statistique de test observée T_O est la même : $T_O = -2.801$.
3. Le seuil critique est $-\overline{z}_{2\alpha} = -\overline{z}_{0.2} = -1.282$.
4. Décision, puisque $T_O = -2.801 < \overline{z}_{2\alpha} = -\overline{z}_{0.2} = -1.282$, on rejette l'hypothèse nulle H_0 et on accepte l'hypothèse alternative H_a : l'abstention de fumer démunie le risque d'attraper un cancer.

4.3 La p-valeur

Dans un test statistique concernant les moyennes ou les fréquences, la p-valeur (appelée aussi probabilité critique) est une valeur de probabilité liée à La statistique de test observée T_O.

4.3.1 Cas de test bilatéral

Dans le cas d'un test bilatéral la valeur de la p-valeur est définie par

$$\text{P-valeur} = P\left(|T| \geq |T_O|\right).$$

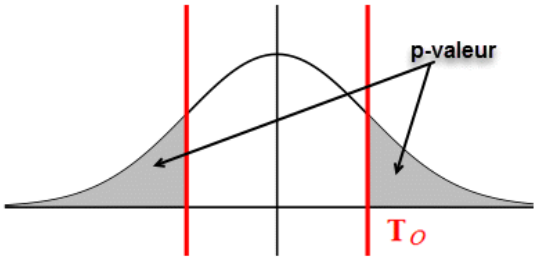


FIG. 4.1 – La p-valeur pour un test bilatéral est représentée par la surface en gris

4.3.2 Cas de test unilatéral à droite

Dans le cas d'un test unilatéral à droite, la p-valeur est définie par

$$\text{p-valeur} = P\left(T \geq T_O\right)$$

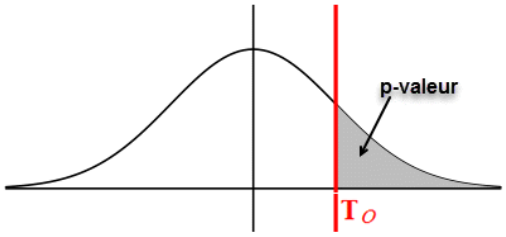


FIG. 4.2 – La p-valeur pour un test unilatéral à droite est représentée par la surface en gris

4.3.3 Cas de test unilatéral à gauche

Dans le cas d'un test unilatéral à gauche, la p-valeur est définie par

$$\text{p-valeur} = P\left(T \leq T_O\right)$$

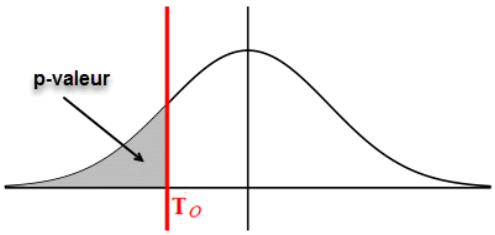


FIG. 4.3 – Dans ce shéma, La p-valeur pour un test unilatéral à gauche est représentée par la surface en gris

4.3.4 La p-valeur et la décision

Dans un test statistique, on peut dans l'étape de la décision, utiliser la p-valeur au lieu de la statistique de test observée T_O :

a) - Dans un test bilatéral, on peut remplacer l'expression "$|T_0| \leq \overline{z}_\alpha$" ou l'expression "$|T_0| \leq \overline{t}_\alpha$" par

l'expression "la p-valeur $\geq \alpha$".

b) - Dans un test unilatéral à droite, on peut remplacer l'expression "$T_0 \leq \overline{z}_{2\alpha}$" par l'expression "la p-valeur $\geq \alpha$".

c) - Dans un test unilatéral à gauche, on peut remplacer l'expression "$T_0 \geq -\overline{z}_{2\alpha}$" par l'expression "la p-valeur $\leq \alpha$".

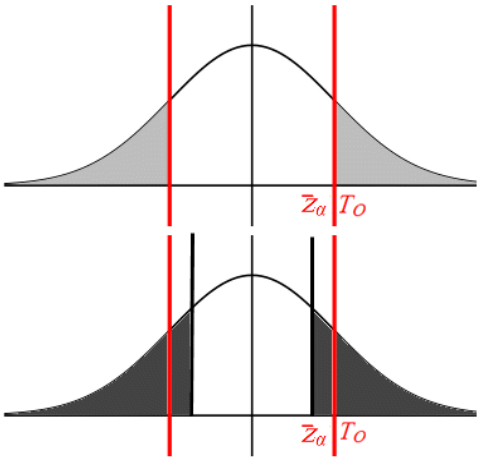


FIG. 4.4 – Dans ce shéma, est illusté le cas d'un test bilatéral où le seuil critique est $\overline{z}_\alpha$. Ici, la p-valeur est représentée par la surface en gris clair et α est représentée par la surface en gris foncé, la statistique observée T_O appartient à la zone de rejet de l'hypothèse nulle, ce qui équivaut à ce que la p-valeur $<$ au risque α. Donc, dans ce test, l'hypothèse nulle H_0 est rejetée

Tableau récapitulatif 3 : Tests Statistiques d'homogénéité : moyennes et fréquences

Genre de test	Données	Statistique de test observée T_O	Seuil critique SC		
			Test bilat.	Test unil. à droite	Test unil. à gauche
Comparaison des moyennes μ_1 et μ_2 de deux populations, à partir des moyennes $\bar{x}$ et $\bar{y}$ calculées sur des échantillons issues de ces deux populations.	m>30 et n>30, σ_X et σ_Y connus	$\frac{(\bar{x}-\bar{y})}{\sqrt{\frac{\sigma^2{}_X}{m}+\frac{\sigma^2{}_Y}{n}}}$	$\bar{z}_\alpha$	$\bar{z}_{2\alpha}$	$-\bar{z}_{2\alpha}$
	m>30 et n>30, σ_X et (ou) σ_Y Inconnus	$\frac{(\bar{x}-\bar{y})}{\sqrt{\frac{S^2{}_X}{m}+\frac{S^2{}_Y}{n}}}$	$\bar{z}_\alpha$	$\bar{z}_{2\alpha}$	$-\bar{z}_{2\alpha}$
	m≤30 ou (et) n≤30, les lois des deux populations normales, σ_X et σ_Y connus	$\frac{(\bar{x}-\bar{y})}{\sqrt{\frac{\sigma^2{}_X}{m}+\frac{\sigma^2{}_Y}{n}}}$	$\bar{z}_\alpha$	$\bar{z}_{2\alpha}$	$-\bar{z}_{2\alpha}$
	m≤30 ou (et) n≤30, les lois des deux populations normales, σ_X et σ_Y inconnus mais égaux	$\frac{(\bar{x}-\bar{y})}{S_C\sqrt{\frac{1}{m}+\frac{1}{n}}}$	$\bar{t}_\alpha$	$\bar{t}_{2\alpha}$	$-\bar{t}_{2\alpha}$
Comparaison de deux proportions vraies P_1 et P_2 à partir de deux proportions calculées $\tilde{p}_1$ et $\tilde{p}_2$	$m > 30, mP_1 > 5$ $m(1-P_1) > 5$ $n > 30, nP_2 > 5$ $n(1-P_2) > 5$	$\frac{(\tilde{p}_1-\tilde{p}_2)}{\sqrt{\tilde{p}(1-\tilde{p})(\frac{1}{m}+\frac{1}{n})}}$	$\bar{z}_\alpha$	$\bar{z}_{2\alpha}$	$-\bar{z}_{2\alpha}$

Notation. m et n = tailles des échantillons, $\bar{x}$ et $\bar{y}$ = moyennes des échantillons, S_X et S_Y = écarts types des échantillons, μ_1 et μ_2 = $moyennes$ des $populations$, σ_X et σ_Y = écarts types des populations, $\tilde{p}_1$ et $\tilde{p}_2$ = proportions calculées sur les échantillons, X et Y = nombres de succès dans le premier et dans le deuxième échantillon, P_1 et P_2 = proportions vraies, $\bar{z}_\alpha$ et $\bar{z}_{2\alpha}$ à calculer de la table 2 de l'écart réduit de la loi normale réduite centrée, $\bar{t}_\alpha$ et $\bar{t}_{2\alpha}$ à calculer de la table 3 de l'écart réduit de la loi de Student de ddl = m+n-2

$S_C{}^2$ (Variance commune) = $\frac{(m-1)S^2{}_X+(n-1)S^2{}_Y}{m+n-2}$, $\tilde{p}$ (Proportion commune) = $\frac{X+Y}{m+n}$

Décision Test bilatéral : si $|T_O| \leq SC$ on accepte l'hypothèse nulle H_0, sinon on la rejette.
Décision Test unilatéral à droite : si $T_O \leq SC$ on accepte l'hypothèse nulle H_0, sinon on la rejette.
Décision Test unilatéral à gauche: si $T_O \geq SC$ on accepte l'hypothèse nulle H_0, sinon on la rejette.

FIG. 4.5 – Tableau récapitulatif : tests d'homogénéité, moyennes et fréquences.

Chapitre 5

Tests de Khi deux

On considère ici trois genres de tests : test de conformité, test d'homogénéité et test d'indépendance. L'outil utilisé est le test de khi deux introduit par le staticien Pearson au début du vingtième siècle.

5.1 Test de Khi deux de conformité

5.1.1 Introduction

On considère sur une population Ω une variable aléatoire qualitative ou quantitative qui a l modalités ou classes de modalités $\Omega_1, \Omega_2, ..., \Omega_l$.

Les probabilités de ces modalités sont notées

$$\alpha_1 = P_r(\Omega_1), \alpha_2 = P_r(\Omega_2), ...\alpha_l = P_r(\Omega_l).$$

On considère maintenant un échantillon aléatoire de taille n d'une population, pas forcément la première, sur laquelle est définie la même variable aléatoire. Soient

$$O_1, O_2, ...O_l$$

les effectifs des éléments de cette échantillon vérifiant les modalités

$$\Omega_1, \Omega_2, ..., \Omega_l.$$

La question à laquelle on se propose de répondre est la suivante.

Peut-on confirmer, avec une probabilité α de se tromper, que l'échantillon en question provient de la population Ω.

La réponse à cette question revient à réaliser un test de conformité de la distribution observée de l'échantillon à la distribution de probabilité de la population Ω.

L'étape principale de ce test consiste en la détermination de la statistique de test. On le fait de la manière suivante.

5.1.2 Les effectifs observés et effectifs théoriques.

Les effectifs observés sont les effectifs $O_1, O_2, ...O_l$ des modalités dans l'échantillon

Les effectifs théoriques $A_1, A_2, ..., A_l$ de l'échantillon sont calculés en supposant que la distribution de probabilité de l'échantillon est conforme à celle de la population. Ils sont donnés par les expressions

$$A_1 = \alpha_1 n, \quad A_2 = \alpha_2 n_2, \quad A_l = \alpha_l n_l. \tag{5.1}$$

La statistique de test observée est le nombre

$$\begin{aligned} T_O &= \sum_{i=1}^{l} \frac{(O_i - A_i)^2}{A_i} \\ &= \frac{(O_1 - A_1)^2}{A_1} + ... + \frac{(O_l - A_l)^2}{A_l}. \end{aligned} \tag{5.2}$$

5.1.3 Les étapes du test.

Etape 1. H_0 : La distribution de l'échantillon est conforme à celle de la population Ω au risque α.

H_a : L'échantillon n'est pas tiré d'une population qui suit la loi en question.

Etape 2. Le seuil critique est calculé selon la loi de khi deux de degré de liberté $\nu = l - 1$. Ce seuil critique est le quantile d'ordre $1 - \alpha$ de cette loi :

$$\chi^2_{\nu,1-\alpha}$$

qu'on calcule dans la table 4 de la loi de Khi deux.

Etape 3. La statistique de test observée est

$$T_O = \sum_{i=1}^{l} \frac{(O_i - A_i)^2}{A_i}. \tag{5.3}$$

Etape 4. Décision.

Si $T \leq \chi^2_{\nu,1-\alpha}$, on accepte l'hypothèse H_0 : la distribution de l'échantillon est conforme à celle de la population au risque α.

Si $T > \chi^2_{\nu,1-\alpha}$, on rejette l'hypothèse H_0 et on accepte l'hypothèse alternative H_a : la distribution de l'échantillon n'est pas conforme à celle de la population.

5.1.4 Validité de l'application du test.

Attention ! pour pouvoir appliquer ce test il faut que chacun des effectifs théoriques A_i soit supérieur à cinq. Dans le cas contraire on fusionne les cellules qui ne vérifient pas cette condition.

5.1.5 Exemple

La distribution de la couleur des cheveux dans une certaine population est donnée dans le tableau suivant.

couleur des cheveux	blond	brune	rousse	**total**
%	40	30	30	100

Dans un ensemble de 37 personnes la répartition de la couleur des cheveux est représentée dans le tableau suivant.

couleur des cheveux	blond	brune	rousse	**total**
effectif	25	9	3	37

Cet échantillon de personnes est-il représentatif de la population au risque $\alpha = 5\%$.

Solution.

On réalise un test de khi deux de conformité. On dresse un tableau sur lequel on fait les calculs.

couleur des cheveux	blond	brune	rousse	**total**
effectif observé O_i	25	9	3	37
Proportion théorique α_i	0.4	0.3	0.3	1
effectif théorique $A_i = \alpha_i n$	14.8	11.1	11.1	37
$\frac{(O_i - A_i)^2}{A_i}$	$\frac{(25-14.8)^2}{14.8}$	$\frac{(9-11.1)^2}{11.1}$	$\frac{(3-11.1)^2}{11.1}$	**13.3**

Puisque chacun des effectifs théoriques A_i est au moins égal à cinq, le test est valable.

Etapes du test.

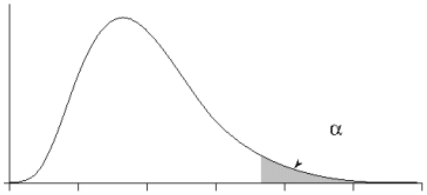
1)- Hypothèse nulle H_0 : cet échantillon est représentatif de la population au risque de 5%.

Hypothèse alternative H_a : cet échantillon n'est pas représentatif de la population.

2)- Le seuil critique est $\chi^2_{\nu=2,1-\alpha=0.95}$, qu'on calcule dans la table 4 de la loi de χ^2 avec degré de liberté $\nu = 3 - 1 = 2$, la table donne

$$\chi^2_{\nu=2,1-\alpha=0.95} = \mathbf{5.99}.$$

Table 4. Loi de Kh 2



ν	0,995	0,975	0,95	0,9	0,1	0,05	0,025	0,005
1	0,000	0,001	0,004	0,016	2,706	3,841	5,024	7,879
2	0,010	0,051	0,103	0,211	4,605	5,991	7,378	10,597
3	0,072	0,216	0,352	0,584	6,251	7,815	9,348	12,838
4	0,207	0,484	0,711	1,064	7,779	9,488	11,143	14,860

Le quantile $\chi_{dll,1-\alpha}$ se trouve dans la cellule, intersection de la ligne du dll (ici $dll = 2$) et la colonne de α (ici $\alpha = 0.05$).

3)- La statistique de test observée est

$$T_O = \sum_{i=1}^{3} \frac{(O_i - A_i)^2}{A_i} = \mathbf{13.3}.$$

4)- Décision. Puisque

$$T_O = 13.3 > 5.99$$

on rejette l'hypothèse nulle H_0 et on accepte l'hypothèse alternative H_1 : cet échantillon n'est pas représentatif de la population.

5.1.6 Exemple

Les salaires d'un échantillon de 100 personnes choisies au hasard d'une grande population sont répartis en classes comme suit

classes	$[40;60[$	$[60;80[$	$[80;100[$	$[100;120[$	$[120;140[$
effectifs	10	25	25	20	20

Au vu de ces résultats peut-on décider que les salaires de cette population sont répartis uniformément sur ces classes.

Solution.

On doit réaliser un test de khi deux (degré de liberté $\nu = 5 - 1 = 4$). Puisque le risque α n'est pas donné, on le fixe à $\alpha = 0.05$.

La répartition uniforme des salaires signifie que les cinq classes contiennent la même proportion, c'est-à-dire la proportion

$$\frac{1}{5} = 0.20.$$

Etapes du test.

1)- H_0 : la répartition des salaires est uniforme, H_a : la répartition des salaires n'est pas uniforme.

2)- Le seuil critique est calculé dans la table 4 de la loi de χ^2 de degré de liberté $\nu = 4$: ce seuil est égal à

$$\chi^2_{\nu=4;1-\alpha=0.95} = 9.49$$

3)- On calcule la statistique observée de test sur un tableau de la façon suivante

classes	$[40;60[$	$[60;80[$	$[80;100[$	$[100;120[$	$[120;140[$	Total
effectifs observés O_j	10	25	25	20	20	100
Proportions théoriques α_j	0.2	0.2	0.2	0.2	0.2	1
effectifs théoriques A_j	20	20	20	20	20	100
$\frac{(O_j-A_j)^2}{A_j}$	5	1.25	1.25	0	0	$T_O = \mathbf{7.5}$

4)- Décision. Puisque la statistique observée (égale à 7.5) est inférieure au seuil (égale à 9.49) on accepte l'hypothèse H_0 : La répartition de la population est distribuée uniformément sur les cinq classes de salaire au risque de 5% de se tromper.

5.1.7 Exemple (Cas particulier)

L'exemple suivant concerne la comparaison d'une proportion observée à une proportion théorique. On utilise ici un cas particulier du test de khi deux correspond au degré de liberté $\nu = 1$.

Une anomalie génétique touche dans un certain pays 1/1000 des individus. Dans une région donnée de ce pays, on a enregistré 57 personnes ayant cette anomalie sur 50 000 naissances. Est ce que cette région est représentative du pays tout entier au risque de 5 %. On a vu précédemment qu'on peut répondre à cette question en réalisant un test de l'écart réduit. Ici on va utiliser un test de Khi deux.

Etapes du test

1)-

Hypothèse nulle H_0 : cette région est représentative du pays tout entier au risque de $\alpha = 0.05$.

Hypothèse alternative H_1 : cette région n'est pas représentative du pays tout entier.

2) le seuil critique est calculé sur la table 4 de la loi de khi deux de degré de liberté $\nu = 1$. Ce seuil est égal à **3.841.**

3)- la statistique observée de test est calculée dans le tableau suivant, sa valeur est **1.96** .

modalité	anomalie	pas d'anomalie	total
Proportions théoriques α_j	0.001	0.999	1
effectifs observés O_j	57	49943	50000
effectifs théoriques A_j	50	49950	50000
$\frac{(O_j - A_j)^2}{A_j}$	0.98	0.98	**1.96**

4)- Décision. puisque la statistique observée est inférieure au seuil ($\mathbf{1.96 < 3.841}$) , on accepte donc l'hypothèse nulle H_0 et on rejette l'hypothèse alternative H_a.

Remarque.

1)-On remarque que l'application du test de khi à cet exemple a donné le même résultat que l'application du test de l'écart réduit. Ce résultat est général : le test de l'écart réduit et le test de khi deux donnent le même résultat pour le test de conformité d'une proportion observée et une proportion théorique.

2)- Le test de khi deux ne peut pas s'appliquer pour réaliser des test unilatéraux.

5.2 Test d'homogénéité

5.2.1 Introduction

Comme dans le test de conformité, on considère un caractère qualitatif ou quantitatif ayant l modalité (ou classes de modalités), $\Omega_1, \Omega_2, ..., \Omega_l$, et on considère k échantillons $E_1, E_2, ..., E_k$ issus de k populations. Pour chaque $1 \leq j \leq l$ et chaque $1 \leq i \leq k$, Le nombre d'éléments de l'échantillon E_i vérifiant la modalité Ω_j est noté O_{ij}. On connait les effectifs O_{ij} et on veut tester l'hypothèse H_0 : les échantillons $E_1, E_2, ..., E_k$ sont tirés de la même population.

La différence entre le test de conformité déjà présenté et le test d'homogénéité est la suivante.

Dans le test de conformité la loi de probabilité à laquelle on doit comparer la distribution observée est connue, par contre, pour le test d'homogénéité on construit cette loi de probabilité à partir des distributions observées.

Pour chaque $1 \leq i \leq k$, On note l'effectif de l'échantillon E_i par N_i :

$$N_i = \sum_{j=1}^{l} O_{i1} = O_{i2} + O_{i3} + ... + O_{il}. \tag{5.4}$$

et par N la somme des effectif des k échantillons :

$$N = \sum_{i=1}^{k} N_i = N_1 + N_2 + ... + N_k. \tag{5.5}$$

Pour chaque $1 \leq j \leq l$; on définit

$$\alpha_j = \frac{\sum_{i=1}^{k} O_{ij}}{N}. \tag{5.6}$$

On note que α_j est la proportion de la modalité Ω_j dans le grand échantillon (union des k échantillons.

On a évidemment

$$\sum_{j=1}^{l} \alpha_j = 1.$$

Effectifs théoriques

Pour chaque $1 \leq i \leq k$ et $1 \leq j \leq l$ **,** le nombre

$$A_{ij} = \alpha_j N_i. \tag{5.7}$$

est l'effectif de la modalité Ω_j attendu dans l'échantillon E_i .

La statistique de test observée est donnée par la formule

$$T = \sum_{j=1}^{l} \sum_{i=1}^{k} \frac{(O_{ij} - A_{ij})^2}{A_{ij}}. \tag{5.8}$$

Seuil critique

Pour le seuil critique, on utilise la loi de khi deux à $(k-1)(l-1)$ degré de liberté.

5.2.2 Exemple

les résultats de l'évolution d'une maladie M, à la suite de l'emploi de l'un ou l'autre des traitements A ou B, figurent dans le tableau suivant, qui donne le nombre de malades appartenant à chacune des catégories.

catégorie	guérison	amélioration	état stationnaire	totaux
A	280	210	110	600
B	220	90	90	400
totaux	500	300	200	1000

Peut-on- dire que les traitements A et B sont identiques ?

Solution.

On fait un test de khi deux d'homogénéité.

1)- hypothèse nulle H_0 : les deux traitements sont identiques (au risque de 5 %).

Hypothèse alternative H_a : les deux traitements ne sont pas identiques (au risque de 5 %).

2)- Le seuil critique est calculé sur la table 4 de la loi de khi deux de degré de liberté $\nu = 3 - 1 = 2$. Ce seuil est égal à **5.99.**

3)- On calcule la statistique observée de test.

On commence par calculer les distributions théoriques α_i de probabilité sur un tableau à deux lignes :

catégorie	guérison	amélioration	état stationnaire	total
probabilité α_i	$\frac{500}{1000} = 0.50$	$\frac{300}{1000} = 0.3$	$\frac{200}{1000} = 0.20$	1

On calcule la distribution et les effectifs théoriques, on le fait sur un tableau.

catégorie	guérison		amélioration		état stationnaire		totaux
	eff. obs.	effe. théor.	eff. obs.	eff. théor.	eff. obs.	eff. théor.	
A	280	300	210	180	110	120	600
B	220	200	90	120	90	80	400
totaux	500	500	300	300	200	200	1000

La statistique observée de test T_O est donc

$$\begin{aligned} T_O &= \frac{(280-300)^2}{300} + \frac{(210-180)^2}{180} + \frac{(110-120)^2}{120} \\ &\quad + \frac{(220-200)^2}{200} + \frac{(90-120)^2}{120} + \frac{(90-80)^2}{8170} \\ &= \mathbf{17.917}. \end{aligned}$$

4)- Décision.

Puisque la statistique observée est supérieure au seuil critique ($\mathbf{17.917} \geq \mathbf{5.99}$), on rejette H_0, et on accepte H_a : les deux traitements ne sont pas identiques.

5.2.3 Exemple

Même problème que celui dans l'exemple précédent, mais avec trois traitements A, B et C, non uniquement deux.

Les trois traitements avaient donné les résultats décrits sur le tableau suivant.

catégorie	guérison	amélioration	état stationnaire	totaux
A	500	150	50	700
B	400	140	60	600
C	500	130	70	700
totaux	1400	420	180	2000

Peut-on dire que ces trois traitements donnent le même résultat.

Solution.

On réalise un test d'homogéniété de khi deux.

1)- Hypothèse nulle H_0 : les trois traitements donnent le même résultat.

Hypothèse alternative H_a : les trois traitements ne donnent pas le même résultat.

2)- Le seuil critique est à calculé sur la table 4 de la la loi de khi deux de degré de liberté $\nu = (3-1)(3-1) = 4$ et $\alpha = 0.05$. Ce seuil critique est **9.49**

3) - on calcule la statistique observée de test T_O.

On calcule la distribution de probabilité α_j, liée au trois échantillons, sur le tableau suivant

catégorie	guérison	amélioration	stationnaire	
α_j	$\frac{1400}{2000} = 0.70$	$\frac{420}{2000} = 0.21$	$\frac{180}{2000} = 0.09$	1

On calcule ensuite les effectifs théoriques

catégorie	guérison		amélioration		état stationnaire		totaux
	obs.	theo.	obs.	theo.	obs.	theo.	
A	500	490	150	147	50	63	700
B	400	420	140	126	60	54	600
C	500	490	130	147	70	63	700
totaux	1400	1400	420	420	180	180	2000

A partir de ce tableau, on calcule la valeur de la statistique de test observée T_O, on trouve

$$T_O = \mathbf{9.07}$$

4)- Décision. Puisque $\mathbf{9.07} < \mathbf{9.49}$, on accepte l'hypothèse H_0 : ces trois traitement donnent le même résultat.

5.3 Test d'indépendance

5.3.1 Introduction

On considère deux variables aléatoires X et Y sur la même population. la variable aléatoire X à K modalités $\Omega_1, \Omega_2, ..., \Omega_k$ modalités et la variable aléatoire Y a l modalités $\Psi_1, \Psi_2, ..., \Psi_l$. On dispose d'un échantillon E contenant N éléments de la population. Pour chaque $1 \leq i \leq k$ et $1 \leq j \leq l$, on connait le nombre des éléments de l'échantillon E vérifiant la modalité Ω_i de X et la modalités Ψ_j de Y, on note ce nombre O_{ij}.

Ces données nous permettent de réaliser un test de khi deux d'indépendance des deux variables aléatoires X et Y.

Les étapes du test.

1)- Hypothèse nulle H_0 : les deux variables aléatoires X et Y aléatoires sont indépendantes.

Hypothèse alternative H_1 : les deux variables aléatoires X et Y ne sont pas indépendantes.

2)- Le seuil critique est à calculer sur la table de loi de khi deux de degré de liberté $(k-1)(l-1)$, pour $\alpha = 0.05$.

3)- La statistique observée de test. Pour la calculer, on commence par dresser un tableau de contingence :

$X \diagdown Y$	Ψ_1	Ψ_2	...	Ψ_l	totaux
Ω_1	O_{11}	O_{12}	...	O_{1l}	$n_{1\bullet}$
Ω_2	O_{21}	O_{22}	...	O_{2l}	$n_{2\bullet}$
...	...	...	...	...	...
Ω_k	O_{k1}	O_{k2}	...	O_{kl}	$n_{k\bullet}$
totaux	$n_{\bullet 1}$	$n_{\bullet 2}$	...	$n_{\bullet l}$	$n_{\bullet\bullet} = N$

La statistique observée de test T_O se calcule par la même formule utilisée pour le test d'homogénéité :

$$T = \sum_{j=1}^{l} \sum_{i=1}^{k} \frac{(O_{ij} - A_{ij})^2}{A_{ij}}, \tag{5.9}$$

où, A_{ij} est l'effectif théorique du couple (A_i, B_j), donnée par la formule

$$A_{ij} = \frac{n_{i\bullet}}{N}\frac{n_{\bullet j}}{N}N = \frac{n_{i\bullet}n_{\bullet j}}{N}. \tag{5.10}$$

On note que $n_{i\bullet}$ est la somme des cellules de la ième ligne et $n_{\bullet j}$ est la somme des cellules de la jème colonne.

4)- Décision. Si la statistique de test observée est inférieure au seuil critique, on accepte H_0, sinon on la rejette.

5.3.2 Exemple

Dans une certaine population, on s'intéresse à la question suivante.

Est-ce-que la couleur des yeux et celle des cheveux sont indépendantes ?.

Pour répondre à cette question, on réalise un test d'indépendance, en considérant un échantillon aléatoire de cette population de taille $N = 124$. Les données concernant la couleur des yeux et celle des cheveux sont rassemblées dans le tableau de contingence suivant.

yeux\cheveux	blonds	bruns	roux	noirs
bleu	25	9	7	3
gris	13	17	7	10
marrons	7	13	5	8

D'après ce tableau, on a

$$k = 3 \text{ et } l = 4.$$

Pour faciliter les calculs on complète le tableau de contingence par les effectifs théoriques pour chaque couple de modalités :

yeux\cheveux	blonds		bruns		roux		noirs		totaux
	obs.	theo.	obs.	theo.	obs	theo.	obs.	theo.	
bleu	25	15.97	9	13.84	7	6.74	3	7.45	$n_{1\bullet}$=44
gris	13	17.06	17	14.78	7	7.20	10	7.96	$n_{2\bullet}$=47
marrons	7	11.98	13	10.38	5	5.06	8	5.59	$n_{3\bullet}$=33
totaux	$n_{\bullet 1}$=45	45	$n_{\bullet 2}$=39	39	$n_{\bullet 3}$=19	19	$n_{\bullet 4}$=21	21	$n_{\bullet\bullet}$=124

Puisque tous les effectifs théoriques A_{ij} sont supérieurs ou égaux à cinq, le test de khi deux est valable.

On rappelle que pour remplir le tableau ci-dessus, on a calculé les effectifs théoriques A_{ij} par la formule $A_{ij} = \frac{n_{i\bullet} n_{\bullet j}}{N}$.

Par exemple l'effectif théorique du couple (marrons, bruns) est

$$A_{32} = \frac{n_{3\bullet} n_{\bullet 2}}{n_{\bullet\bullet}} = \frac{33 \times 39}{124} = 10.38.$$

La statistique de test observée T_O est dans ce cas

$$T = \sum_{j=1}^{4} \sum_{i=1}^{3} \frac{(O_{ij} - A_{ij})^2}{A_{ij}} = \mathbf{15.1}$$

Le seuil critique est à calculer sur la table 4 de la loi de khi deux de degré de liberté $\upsilon = (k-1)(l-1) = 6$.

Pour $\alpha = 0.05$, ce seuil est égal à **12.6**

Décision. puisque

$$\mathbf{T = 15.1 > 12.6},$$

on rejette l'hypothèse H_0, et on accepte l'hypothèse H_a : la couleur des yeux et celle des cheveux ne sont pas indépendantes.

Chapitre 6

Analyse de la variance (ANOVA)

On a jusqu'ici présenter des tests pour comparer deux moyennes théoriques $\mu_1$et μ_2 à partir de deux échantillons, ou pour comparer une moyennes observée et une moyenne théorique μ.

Dans cette section on utilise une classe de test (analyse de variance) pour réaliser deux tests, le premier concerne la comparaison de moyennes de plusieurs populations (deux populations ou plus) à partir d'échantillons de ces populations, le deuxième test concerne la comparaison de variances de deux populations à partir de deux échantillons de ces populations.

Ce genre de tests est indiqué sous le qualificatif " analyse de variance" parce qu'il se base sur la dispersion des populations considérées.

6.1 Comparaison de plusieurs moyennes.

6.1.1 Conditions de validité du test.

Deux conditions sont nécessaires pour pouvoir appliquer la méthode qu'on va présenter : les populations doivent être normales et ayant la même variance.

On explique la façon de réaliser ce genre de test à travers l'exemple suivant.

6.1.2 Exemple

Une importante entreprise de conservation alimentaire réalise une étude économique relative à la transformation des haricots verts. Une enquête de terrain est réalisée pour étudier l'influence éventuelle du facteur variétal sur le diamètre des haricots ; ce dernier paramètre est en effet un critère important puisqu'il permet de classer les haricots selon diverses catégories (fins, extra-fins, etc).

On se limite à quatre variétés V_1, V_2, V_3 et V_4 qui offrent une bonne résistance aux maladies et sont donc fréquemment cultivées dans la région étudiée. On considère des haricots issus de sols comparables et de techniques culturales proches.

On prélève des échantillons aléatoires de chacune des quatre variétés et l'on observe les résultats indiqués sur le tableau suivant où sont mentionnés les diamètres en cm.

		totaux
V_1	8,8 7,1 3,7 4,5 8,3 9,2 7,5 4,9 5,5 5,5 7,8 10 5,7 8,1 5,8 7,3 6,0 8,6 6,4 6,8 7,0	
V_2	6,8 3,5 5,5 6,2 6,0 8,0 6,3 6,3 8,0 7,7 5,9 8,2 7,5 5,7 6,2 3,0 7,0 3,5 7,8 4,0 7,5 4,2 7,3 4,3 5,9 4,4 5,7 4,6 5,8 4,8 5,9 5,0 5,0 6,1 5,1	
V_3	5,2 6,3 5,3 6,4 5,4 6,5 5,5 6,6 4,8 6,7 5,0 5,8 5,3 5,7 5,5 6,5 5,6 6,7 3,2 3,0 3,1 6,1 6,8 6,6 8,6 6,9 6,9 8,6 7,6 4,8 5,7 6,7 7,7 7,4 4,1 9,9 8,8 5,6 5,9 4,3 7,7 5,4	
V_4	6,1 6,8 6,6 8,6 6,9 6,9 8,6 7,6 4,8 5,7 6,7 7,7 7,4 4,1 9,9 8,8 5,6 5,9 4,3 7,7 5,4 6,0 9,0 8,0 6,0 5,0 6,0 10 6,2 8,0 8,6 6,4 8,2	

Question. Peut-on considérer qu'en moyenne les quatre variétés ont le même diamètre ?

On sait qu'une étude préalable a permis d'accepter l'hypothèse de la normalité ainsi que l'hypothèse de l'égalité des variances des variables aléatoires "diamètre des haricots verts" pour les quatre variétés.

Tester cette hypothèse au niveau 95%.

Solution

Notation générale.

E_i $(i = 1, 2, 3, 4)$ est un échantillon aléatoire de la population V_i de moyenne μ_i.

Le nombre d'échantillons E_1, E_2, E_3 et E_4 est $k = 4$.

La taille de l'échantillon E_i $(i = 1, 2, 3, 4)$ est noté N_i :

$$N_1 = 21,\ N_2 = 35,\ N_3 = 42,\ N_4 = 33.$$

x_{ij} est l'observation numéro j de l'échantillon numéro i, par exemple

$$\begin{aligned} x_{11} &= 8.8,\ x_{14} = 4.5,\ x_{1\,21} = 7.0,\ x_{2\,6} = 8.0, \\ x_{2\,10} &= 7.7,\ x_{3\,2} = 6.3,\ x_{35} = 5.4,\ x_{4\,5} = 5.0. \end{aligned}$$

$\overline{x}_i$ $(i = 1, 2, 3, 4)$ est la moyenne de l'échantillon E_i :

$$\overline{x}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} x_{ij}.$$

$\overline{x}$ est la moyenne observée sur l'ensemble des observations de tous les échantillons :

$$\overline{x} = \frac{1}{N} \sum_{i=1}^{4} \sum_{j=1}^{N_i} x_{ij} = \frac{1}{N} \sum_{i=1}^{4} N_i \overline{x}_i.$$

La somme des carrés totale ou dispersion totale, notée SCE_t, est définie par

$$SCE_t = \sum_{i=1}^{4} \sum_{j=1}^{N_i} (x_{ij} - \overline{x})^2 .$$

La somme des carrés résiduelle ou dispersion intra-échantillons, noté SCE_r, est définie par

$$SCE_r = \sum_{i=1}^{4} \sum_{j=1}^{N_i} (x_{ij} - \overline{x}_i)^2 .$$

La somme des carrés factorielle ou dispersion inter- échantillons, noté SCE_f, est définie par

$$SCE_{fa} = \sum_{i=1}^{4} N_i (\overline{x}_i - \overline{x})^2 .$$

La variance interclasse ou carré moyen factoriel, notée CM_{fa}, est définie par

$$CM_{fa} = \frac{SCE_{fa}}{k-1} = \frac{SCE_{fa}}{3}.$$

La variance intra-classe ou carré moyen résiduel, notée CM_r, est définie par

$$CM_r = \frac{SCE_r}{N-k} = \frac{SCE_r}{127}.$$

Equation de l'analyse de la variance

On peut montrer l'équation suivante, appelée équation de la variance.

SCE_t	=	SCE_r	+	SCE_{fa}
Variabilité totale		Variabilité résiduelle		variabilité factorielle
		(intra-échantillons)		(inter-échantillons)

6.1.3 Les étapes du test.

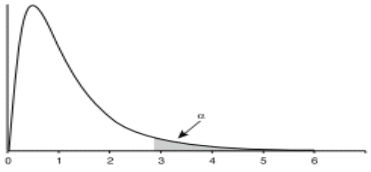
1)- Hypothèse nulle $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$,

l'hypothèse alternative H_1 : au moins deux moyennes ne sont pas égales.

ici μ_i $(i = 1, 2, 3, 4)$ est le diamètre moyen de la variété i.

2) Le seuil critique est à calculer sur la loi de Fisher $F_{3,127}$ de ddl (ν_1, ν_2) où $\nu_1 = k - 1 = 3$ et $\nu_2 = N - k = 127$. Ce seuil est le quantile $F_{3,127,0.95}$ d'ordre $1 - \alpha = 0.95$. Dans la table 5 de la loi de Fisher fournie dans notre cours pour $\alpha = 0.05$, $\nu_2 = 127$ ne figure pas, on prend la valeur la plus proche de 127 qui est la valeur 100:

Table 5. Loi de fisher.



$\Pr(F_{\nu_1,\nu_2} > c) = \mathbf{0,05}$

ν_1 \ ν_2	18	20	22	24	26	28	30	40	50	60	100	200
1	4,41	4,35	4,30	4,26	4,23	4,20	4,17	4,08	4,03	4,00	3,94	3,89
2	3,55	3,49	3,44	3,40	3,37	3,34	3,32	3,23	3,18	3,15	3,09	3,04
3	3,16	3,10	3,05	3,01	2,98	2,95	2,92	2,84	2,79	2,76	2,70	2,65
4	2,93	2,87	2,82	2,78	2,74	2,71	2,69	2,61	2,56	2,53	2,46	2,42
5	2,77	2,71	2,66	2,62	2,59	2,56	2,53	2,45	2,40	2,37	2,31	2,26

$$F_{3,127,0.95} \simeq F_{3,100,0.95} = 2.\mathbf{70}$$

3)- La statistique de test observée T_O est donnée par la formule

$$T_O = \frac{CM_{fa}}{CM_r} = \frac{17.53}{2.17} = \mathbf{8.08}$$

4)- Décision. Puisque

$$T_O = \mathbf{8.08} > F_{3,127,0.99} = 2.\mathbf{70},$$

on rejette l'hypothèse nulle H_0 et on accepte l'hypothèse alternative H_1 : au moins les diamètres moyennes de deux variétés ne sont pas égaux.

6.2 Comparaison de deux variances

Ce test appelé test de Fisher-Snedecor, concerne la comparaison des variances σ_1 et σ_2 de deux populations. La condition nécessaire pour réaliser ce test est la normalité des deux populations.

On explique ce test à travers l'exemple suivant.

6.2.1 Exemple

On considère les trois échantillons suivants.

Ech 1.	12	15	14	16	22	17	25	9	18		
Ech 2.	7	18	9	9	18	27	12	10	32	6	37
Ech 3.	10	13	13	15	17	10	10	15	4	24	

A partir de ces données, est-ce-que la population d'où est tirée l'échantillon 2 est plus hétérogène que celle d'où est tirée l'échantillon 1 ?

Remarque.

On a vu dans ce cours que pour appliquer un test de student à la comparaison de deux moyennes de populations normales, partant de petits échantillons, la condition d'égalité des variances des deux populations est nécessaire. Ainsi, le test présenté ici de comparaison de deux variances est une étape préliminaire dans le test de Student de comparaison de deux moyennes dans le cas de petits échantillons.

6.2.2 Exemple

Lors d'une expérience pédagogique, on s'intéresse à l'effet comparé de deux pédagogies des mathématiques chez deux groupes de 10 sujets : pédagogie traditionnelle (p 1) et pédagogie moderne (p 2).

On note la performance à une épreuve de combinatoire.

Pédagogie traditionnelle

sujet	1	2	3	4	5	6	7	8	9	10
note	5.0	4.0	1.5	6.0	3.0	3.5	3.0	2.5	1.5	2.5

Pédagogie moderne

sujet	11	12	13	14	15	16	17	18	19	20
note	4.0	5.5	4.5	6.5	4.5	5.5	1.0	2.0	4.5	4.5

Avant d'appliquer un test de comparaison de moyennes, on veut s'assurer que l'on peut supposer que les variances sont égales dans les populations parentes. Procéder à un test de comparaison de variances permettant de s'en assurer ?

Solution.

Test de comparaison de deux variances

1- $H_O : \sigma_1^2 = \sigma_2^2$, $H_a : \sigma_2^2 > \sigma_1^2$.

2- Le seuil critique est calculé dans la table 5 de la loi de Fisher de $ddl_1 = ddl_2 = 9$, pour $\alpha = 0.05$.

Ce seuil est **3.18**.

3- La statistique de test observée est

$$T_O = \frac{S_2^2}{S_1^2} = \frac{2.681}{2.069} = \mathbf{1.30}$$

4- Décision : puisque $T_O = 1.30 \leq 3.18$, on accepte H_0.

Chapitre 7

Tests non paramétriques

Nous avons déjà présenté des tests sur la comparaison de deux moyennes, utilisant la loi de Student ou la loi normale. La validité de l'application de ces tests exige certaines conditions, en particulier, la normalité des lois mères pour les petits échantillons. Bien que ces tests sont robustes relativement à la normalité des lois mères, c'est-à- dire ils sont peu sensible à la perturbation de la normalité des lois mères, ils sont très sensibles aux valeurs aberrantes, c'est-à dire les valeurs anormalement petites ou grandes.

Les tests qu'on va présenter dans ce chapitre sont peu sensibles aux valeurs aberrantes puisqu'ils n'utilisent pas les valeurs effectives des variables aléatoires mais plutôt leurs rangs. En outre, ils n'exigent pas en particulier la normalité. Ils sont qualifiés de "non paramétriques" parce qu'ils n'utilisent pas les paramètres des populations comme la moyenne et la variance. La spécificité de ces tests réside dans le fait qu'ils n'agissent pas sur les valeurs effectives des populations mais plutôt sur leurs rangs.

7.1 Test de la somme des rangs de Wilcoxon

Ce test est utilisé pour comparer les moyennes de deux populations à partir des moyennes observées.

7.1.1 Exemple illustratif

On considère deux variables quantitatives continues X et Y dont les fonctions densités ayant la même allure. On note les moyennes de X et celle de

Y par μ_1 et μ_2 respectivement.

Soient n observations $x_1, x_2, ..., x_n$ de X et m observations $y_1, y_2, ..., y_m$ de Y. Pour simplifier l'exposé, on suppose par exemple que $n = m = 4$.

On forme un grand échantillon Z union des deux échantillons : $Z = x_1, x_2, ..., x_n, y_1, y_2, ..., y_m$. L'échantillon Z est composé de $n + m = 8$ observations. On ordonne ces huit observations de la plus petite à la plus grande.

Cas extrême. On suppose que, dans cet ordre les quatre observations x_1, x_2, x_3, x_4 de X ont les plus petits rangs c'est-à-dire les rangs 1, 2, 3 et 4 et les quatre observations y_1, y_2, y_3, y_4 de Y ont les plus grands rangs 5, 6, 7 et 8.

La somme des rangs de l'échantillon de X est $R_X = 1 + 2 + 3 + 4 = 10$.

La somme des rangs de l'échantillon de Y est $R_Y = 5 + 6 + 7 + 8 = 26$

La somme des rangs du grand échantillon Z est égale $R = 1 + 2 + 3 + 4 + 5 + 6 + 7 + 8 = 36$.

Puisque la somme des rangs des x_i est très petite comparativement à la somme des rangs des y_i, on peut sans connaitre les valeurs des x_i et celles des y_i, affirmer que la moyenne

$$\overline{x} = \frac{x_1 + x_2 + x_3 + x_4}{4}$$

est plus petite que la moyenne

$$\overline{y} = \frac{y_1 + y_2 + y_3 + y_4}{4}$$

et déduire par suite que probablement, la moyenne $\mu_1$de X est plus petite que la moyenne μ_2 de Y.

Cas médiane. Si par contre, on trouve que la somme des rangs des x_i n'est pas très différente de celle des rangs des y_i, il n'y a pas de raison d'affirmer que la moyenne des x_i est différente de celle des y_i.

Cet exemple illustre le principe sur lequel repose le test de la somme des rangs de Wilcoxon.

Le test de la somme des rangs peut être réalisé pour des très petits échantillons, dans ce cas la région d'acceptation de l'hypothèse nulle est exprimée à l'aide d'un tableau de probabilité de lois discrète. On se limite dans ce cours à présenter le test dans le cas d'échantillons de taille modérée, pratiquement supérieure à huit, cas nous permettant d'utiliser l'approximation par la loi normale. On note que généralement ce test est utile surtout pour les échantillons de taille ne dépassant pas trente, car dans le cas contraire, on peut, comme on l'a déjà vu, utiliser des tests basés sur la loi normale.

7.1.2 Test de la somme des rangs de Wilcoxon

On considère deux variables aléatoires quantitatives continues X de moyenne μ_X et Y de moyenne μ_Y. Et, un échantillon aléatoire, $x_1, x_{2,}..., x_{n_1}$ de X de taille n_1 et un échantillon $y_1, y_{2,}..., y_{n_2}$ de Y de taille n_2.

A partir de ces deux échantillons on espère tester l'hypothèse nulle $\mu_X = \mu_Y$.

On rassemble ces deux échantillons pour former un grand échantillon $x_1, x_{2,}..., x_{n_1}$, $y_1, y_{2,}..., y_{n_2}$, on ordonne les éléments de ce grand échantillon de taille $n_1 + n_2$, du plus petit au plus grand.

On note par R la somme des rangs des x_i :

$$R = R(x_1) + R(x_2) + ... + R(x_{n_1}).$$

On pose

$$T = \frac{R - \mu_R}{\sigma_R},$$

où

$$\mu_R = \frac{n_1(n_1 + n_2 + 1)}{2}$$

et

$$\sigma_R = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}.$$

Sous l'hypothèse nulle $\mu_X = \mu_Y$, pour n_1 assez grand ($n_1 \geq 8$ et $n_2 \geq 8$), la variable aléatoire R définie ici suit (approximativement) la loi normale de moyenne μ_R et d'écart type σ_R, et par conséquent T suit la loi normale réduite centrée.

Etapes du test

1. Hypothèse nulle H_0 : $\mu_X = \mu_Y$ contre l'hypothèse alternative H_1 : $\mu_X \neq \mu_Y$ pour un test bilatéral, ou $\mu_X > \mu_Y$ pour un test unilatéral à droite, ou $\mu_X < \mu_Y$ pour un test unilatéral à gauche.

2. La statistique de test observée T_O est donnée par la formule

$$T_O = \frac{r - \mu_R}{\sigma_R},$$

r étant la valeur de la variable aléatoire R pour les deux échantillons $x_1, ...x_{n_1}$ et $y_1, ..., y_{n_2}$.

3. Seuil critique. Le seuil critique est

a) $\overline{z}_\alpha = z_{1-\frac{\alpha}{2}}$ pour un test bilatéral, à calculer dans la table 2 de l'écart réduit de la loi NRC ou dans la table 1 de la loi normale,

b) $\overline{z}_{2\alpha} = z_{1-\alpha}$ pour un test unilatéral à droite,

c) $-\overline{z}_{2\alpha} = -z_{1-\alpha}$ pour un test unilatéral à gauche.

4. Décision

a) test bilatéral. Si $-\overline{z}_\alpha \leq T_O \leq \overline{z}_\alpha$, on accepte $H_0 : \mu_X = \mu_Y$, sinon on la rejette et on accepte l'hypothèse alternative $H_1 : \mu_X \neq \mu_Y$.

b) test unilatéral à droite. si $T_O \leq \overline{z}_{2\alpha}$, on accepte $H_0 : \mu_X = \mu_Y$, sinon on la rejette et on accepte l'hypothèse alternative $H_1 : \mu_X > \mu_Y$.

c) b) test unilatéral à gauche. si $T_O \geq -\overline{z}_{2\alpha}$, on accepte $H_0 : \mu_X = \mu_Y$, sinon on la rejette et on accepte l'hypothèse alternative $H_1 : \mu_X < \mu_Y$.

7.1.3 (Calculs des rangs dans le cas des ex aequo)

S'il y a peu d'ex aequo dans l'échantillon Z formé des deux échantillons X et Y (c'est-à dire peu de valeurs identiques on utilise le rang moyen.

Par exemple si les échantillons X et Y sont.

$x_1 = -3$, $x_2 = -2$, $x_3 = -2$, $x_4 = 1$, $x_5 = 5$, $x_6 = 5$, $x_7 = 7$, $x_8 = 9$, $x_9 = 11$

$y_1 = -4$, $y_2 = -3$, $y_3 = 2$, $y_4 = 4$, $y_5 = 5$, $y_6 = 8$, $y_7 = 10$, $y_8 = 11$, $y_9 = 12.5$, $y_{10} = 14$

alors on calcule les rangs du grand échantillon Z dans le tableau suivant

a) On ordonne les éléments de Z du plus petit au plus grand, on obtient

y_1 y_2 x_1 x_2 x_3 x_4 y_3 y_4 y_5 x_5 x_6 x_7 y_6 x_8 y_7 y_8 x_9 y_9 y_{10}.

b) On assigne ensuite les rangs, on aura

$R(y_1) = 1$, $R(y_2) = R(x_1) = (2+3)/2 = 2.5$,

$R(x_2) = R(x_3) = (4+5)/2 = 4.5$, $R(x_4) = 6$, $R(y_3) = 7$, $R(y_4) = 8$, $R(y_5) = R(x_5) = R(x_6) = (9+10+11)/3 = 10$, $R(x_7) = 12$, $R(y_6) = 13$, $R(x_8) = 14, R(y_7) = 15$,

$R(x_9) = R(y_8) = (16+17)/2 = 16.5$, $R(y_9) = 18$, $R(y_9) = 19$.

On a enfin
$R = R(x_1) + R(x_2) + R(x_3) + R(x_4) + R(x_5) + R(x_6) +$
$R(x_7) + R(x_8) + R(x_9) = 2.5 + 4.5 + 4.5 + 6 + 10 +$
$10 + 12 + 14 + 16.5 = \mathbf{80}$

7.1.4 Exemples

L'article "Histamine Content in Sputum from Allergic and Non-Allergic Individuals (J. of Appl. Physiology, 1969 : 535–539)" rapporte les données suivantes sur le niveau d'histamine dans les expectorations (mg / g de poids sec d'expectorations) pour un échantillon de 9 individus classés comme allergiques et un autre échantillon de 13 individus classés comme non allergiques :

Allergiqus	67.6 39.6 1651.0 100.0 65.9 1112.0 31.0 102.4 64.7
Non allergiques	34.3 27.3 35.4 48.1 5.2 29.1 4.7 41.7 48.0 6.6 18.9 32.4 45.5

Les données indiquent-elles qu'il existe une différence significative entre le niveau moyen d'histamine des expectorations entre les allergiques et les non allergiques ? Utiliser un test de la somme des rang de Wilcoxon.

Solution

Puisque les tailles des deux échantillons sont supérieures à 8, l'approximation par la loi normale est appropriée.

On réalise un test bilatéral :

1. L'hypothèse nulle $H_0 : \mu_1 = \mu_2$ contre l'hypothèse alternative $H_1 : \mu_1 \neq \mu_2$, où μ_1 et μ_2 sont respectivement les niveaux moyens d'histamine des expectorations pour les allergiques et les non allergiques.

2. Le seuil critique pour $\alpha = 0.05$ est $\overline{z}_{0.05} = \mathbf{1,960}$ (selon la table 2 de l'écart réduit de la loi NRC)

3. Statistique de test observée T_O.

Les tailles des deux échantillons sont $n_1 = 9$ et $n_2 = 13$.

Les rangs des neuf valeurs observées d'histamine des expectorations des allergiques sont $r_1 = 18$, $r_2 = 11$, $r_3 = 22$, $r_4 = 19$, $r_5 = 17$, $r_6 = 21$, $r_7 = 7$, $r_8 = 20$ et $r_9 = 16$. Et, la somme des rangs est $r = \sum r_i = \mathbf{151}$

La statistique de test observée T_O est donc

$$\begin{aligned} T_O &= \frac{r - n_1 (n_1 + n_2 + 1) \diagup 2}{\sqrt{n_1 n_2 (n_1 + n_2 + 1) \diagup 12}} \\ &= \frac{151 - 9 (9 + 13 + 1) \diagup 2}{\sqrt{9 \times 13 (9 + 13 + 1) \diagup 12}} \\ &= \frac{151 - 103.5}{\sqrt{224.25}} = \mathbf{3.17} \end{aligned}$$

4. Décision. Puisque $T_O = \mathbf{3.17} > \mathbf{1.96}$ n'appartient pas à la zone d'acceptation de H_0 on rejette H_0 et on accepte H_1 : il y a une différence significative au seuil de 5%, entre les niveaux moyens d'histamine dans les expectorations entre les allergiques et les non allergiques.

7.2 Test des rangs signés de Wilcoxon

Ce test est utilisé pour comparer la moyenne d'une population à une valeur donnée à partir d'un échantillon de la population ou également pour comparer les moyennes de deux populations à partir d'échantillons appariés. C'est l'équivalent du test de la somme des rangs de Wilcoxon présenté dans la section précédente.

7.2.1 Exemple illustratif (comparaison d'une moyenne à une valeur donnée)

Un fabricant de fers à repasser électriques, souhaitant tester la précision de la commande du thermostat au réglage de 500 ° F, charge un ingénieur de test d'obtenir les températures réelles à ce réglage pour $n = 15$ fers à l'aide d'un thermocouple. Les mesures résultantes sont les suivantes :

494.6 510.8 487.5 493.2 502.6 485.0 495.9 498.2 501.6 497.3 492.0 504.3 499.2 493.5 505.8

On explique à travers cet exemple le principe de réaliser un test des rangs signés de Wilcoxon pour tester l'hypothèse nulle $H_0 : \mu = 500°F$ contre l'hypothèse alternative $\mu \neq 500°F$, où μ est la température moyenne au réglage du thermostat à $500°F$.

1. La soustraction de 500 de chacun des valeurs donne les valeurs suivantes :

-5.6 10.8 -12.5 -6.8 2.6 -15.0 -4.1 -1.8 1.6 -2.7 -8.0 4.3 -0.8 -6.5 5.8

2. On ordonne ces valeurs du plus petit en valeur absolue au plus grand (sans prendre en compte le signe + ou -), le rang du plus petit est 1, le rang du deuxième est 2, ... , le rang du dixième est 10. On assigne le signe - au rangs des nombres négatifs et le signe + au rangs des nombres positifs, on obtient alors le tableau suivant

valeur absolue	0.8	1.6	1.8	2.6	2.7	4.1	4.3	5.6
rang	1	2	3	4	5	6	7	8
signe	-	+	-	+	-	-	+	-

valeur absolue	5.8	6.5	6.8	8.0	10.8	12.5	15.0
rang	9	10	11	12	13	14	15
signe	+	-	-	-	+	-	-

3. On somme les rangs de signes positifs

$$s_+ = 2 + 4 + 7 + 9 + 13 = 35.$$

La somme des rangs négatifs est

$$s_- = 1 + 3 + 5 + 6 + 8 + 10 + 11 + 12 + 14 + 15 = 120.$$

Le Test des rangs signés de Wilcoxon utilise le principe suivant : si les deux sommes s_+ et s_- ont des valeurs proches alors probablement, la médiane est zéro (donc la moyenne est également zéro pour une distribution symétrique). On note que l'égalité des valeurs de s_+ et s_- équivaut à ce que

$$s_+ = s_- = n\,(n+1)\,/4 = \frac{15 \times 16}{4} = 60,$$

car la somme totale des rangs de signe positif et de signe négatif est égale à

$$n\,(n+1)\,/2 = \frac{15 \times 16}{2} = 120.$$

Dans cet exemple les deux sommes ne sont pas proches l'une de l'autre, donc probablement la valeur moyenne des températures n'est pas proche de

$500°F$. Le test des rangs signés de Wilcoxon peut se réaliser dans le cas de petits échantillons. Dans ce cas La détermination de la région d'acceptation de l'hypothèse nulle est construite, comme pour le test de la somme des rangs, à l'aide d'un tableau de probabilité finie. Ici, on présente ce test dans le cas de l'approximation par la loi normale, c'est-à-dire pour des échantillons de taille supérieure à vingt.

On utilise maintenant les notations de cet exemple pour présenter le test des rangs signés de Wilcoxon

7.2.2 Test des rangs signés de Wilcoxon

Soit X une variable aléatoire continue symétrique (c'est-à-dire sa fonction densité à une axe de symétrie), et soit x_1, $x_2, ..., x_n$

un échantillon aléatoire de taille n de X, et soit μ la moyenne supposée finie de X.

Le test de l'hypothèse H_0 se fait à travers les étapes suivantes.

1. Hypothèse nulle $H_0 : \mu = 0$ contre $H_1 : \mu \neq 0$ pour un test bilatéral (ou $\mu > 0$ pour un test unilatéral à droite et $\mu < 0$ pour un test unilatéral à gauche)

2. Statistique de test observée T_O, pour $n > 20$, la statistique de test observée est donnée par la formule

$$T_O = \frac{s_+ - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}}.$$

3. Le seuil critique est $\overline{z}_\alpha$ à calculer de la table 2 de l'écart réduit de la loi NRC pour le test bilatéral (ou $\overline{z}_{2\alpha}$ pour un test unilatéral à droite et $-\overline{z}_{2\alpha}$ pour un test unilatéral à gauche).

4. Décision.

a) Pour un test bilatéral, on accepte H_0 Si $T_O \in [-\overline{z}_\alpha, \overline{z}_\alpha]$.

b) Pour un test unilatéral à droite, on accepte H_0 Si $T_O \leq \overline{z}_{2\alpha.}$

c) Pour un test unilatéral à gauche, on accepte H_0 Si $T_O \geq -\overline{z}_{2\alpha.}$

7.2.3 Exemple (Echantillons appariés)

Des échographies ont été réalisées au moment de la transplantation hépatique et de nouveau 5 à 10 ans plus tard pour déterminer la pression systolique de l'artère hépatique.

Les résultats de 21 greffes pour 21 enfants sont présentés dans le tableau suivant.

Enfant	1	2	3	4	5	6	7	8	9	10	11
au moment de la greffe	35	40	58	71	33	79	20	19	56	26	44
5 ans plus tard	46	40	50	50	41	70	35	40	56	30	30

Enfant	12	13	14	15	16	17	18	19	20	21
au moment de la transplantation	90	43	42	55	60	62	26	60	27	31
5 ans plus tard	60	43	45	40	50	66	45	40	35	25

A) Dans un tableau, calculer

a)- les valeurs absolues des différences des systoliques, 5 ans après et au moment de la transplantation.

b)- les rangs de ces différences.

c)- les rangs signés.

B) Peut-on affirmer au seuil de 5%, qu'il y a une différence significative en moyennes entre les systoliques au moment de la transplantation et 5 ans après.

Solution

A)

Enfant	1	2	3	4	5	6	7	8	9	10	11
au moment de la greffe	35	40	58	71	33	79	20	19	56	26	44
5 ans plus tard	46	40	50	50	41	70	35	40	56	30	30
Différences	-11	0	8	19	-8	9	-15	–21	0	-4	14
Rangs	13	2	9	17.5	9	11	15.5	20	2	5.5	14
Rangs signés	-13	0	9	17.5	-9	11	-15.5	-20	0	-5.5	14

Enfant	12	13	14	15	16	17	18	19	20	21
au moment de la greffe	90	43	42	55	60	62	26	60	27	31
5 ans plus tard	60	43	45	40	50	66	45	40	35	25
différences	30	0	-3	15	10	-4	-19	20	8	6
Rangs	21	2	4	15.5	12	5.5	17.5	19	9	7
Rangs signés	21	0	-4	15.5	12	-5.5	-17.5	19	9	7

B) on réalise un test des rangs signés de Wilcoxon.

On note μ_1 et μ_2 respectivement les moyennes des systoliques au moment de la greffe et cinq ans plus tart.

On calcule s^+ :

$$s^+ = 13 + 9 + 15.5 + 20 + 5.5 + 4 + 5.5 + 17.5 = 90.$$

Hypothèse nulle $H_0 : \mu_1 = \mu_2$ contre l'hypothèse alternative $H_1 : \mu_1 \neq \mu_2$.

La statistique de test observée est

$$\begin{aligned} T_O &= \frac{s_+ - n\,(n+1)\,/4}{\sqrt{n\,(n+1)\,(2n+1)\,/24}} \\ &= \frac{90 - (21 \times 22)\,/4}{\sqrt{(21 \times 22 \times 43)\,/4}} \\ &= \frac{90 - 115.5}{28.77} = \mathbf{-0.89} \end{aligned}$$

Le seuil critique est $\overline{z}_\alpha = \overline{z}_{0.05} = \mathbf{1.96}$

Décision. Puisque $|T_O| = 0.89 \leq 1.96$, il n'y a pas une différence significative au seuil de 5% entre μ_1 et μ_2.

Remarque : traitement des ex aequo

Comme pour le cas du test de la somme des rangs, dans le cas des ex aequo des différences, on adopte le rang moyen. En plus, si la différence est zéro le signe est zéro. Dans l'exemple précédent, la différence pour les enfants numéros 2, 9 et 13 est zéro, ainsi ils ont le même rang égal à la moyenne $(1 + 2 + 3)\,/3 = 2$, et ont également le même rang signé zéro.

Chapitre 8

Corrélations linéaires

8.1 Introduction

Soient (X, Y) un couple de variables aléatoires et soient $(x_1, y_1), (x_2, y_2), ... (x_N, y_N)$ une série numérique double formée de N observations de ce couple de variables aléatoires. Le coefficient de corrélation linéaire, noté ρ, de ce couple de variables aléatoires est définie par la formule

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

où

$$Cov(X, Y) = \sum_j \sum_i (x_i - E(X))(y_j - E(X)) p(x_i, y_j) \text{ dans le cas discret}$$

et

$$Cov(X, Y) = \int_y \int_x (x - E(X))(y - E(X)) f(x, y) dxdy \text{ dans le cas continu.}$$

Ce coefficient de corrélation ρ est un paramètre du couple aléatoire (X, Y) comme les autres paramètres, dont on a déjà pris connaissance, comme les moyennes μ_X et μ_Y ou les variances σ_X^2 et σ_Y^2. Ainsi, on peut également réaliser des tests statistiques sur ce paramètre ρ.

La variable aléatoire

$$R = \frac{\sum_{i=1}^{N} \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sqrt{\sum \left(X_i^2 - \overline{X}\right)^2 \sum \left(Y_i^2 - \overline{Y}\right)^2}}$$

est un estimateur sans biais de ρ, et ainsi, le coefficient de corrélation linéaire r de la série statistique associée, défini par

$$r = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{N} (x_i - \overline{x})^2 \sum_{i=1}^{N} (y_i - \overline{y})^2}}$$

est une estimation ponctuelle du coefficient de corrélation linéaire ρ du couple aléatoire (X, Y).

8.2 Notations

Les notations suivantes seront utilisées dans cette section.

1)-la somme des carrés des résidus e_i, notée SSE est définie par

$$SSE = \sum_{i=1}^{N} e_i^2,$$

où les résidus e_i sont les différences entre les valeurs observées y_i et les valeurs

$$\widehat{y}_i = bx_i + a$$

obtenues par l'ajustement linéaire de la série statistique double $(x_i, y_i), i = 1, ..., N$:

$$\begin{aligned} e_i &= (y_i - \widehat{y}_i) \\ &= (y_i - (by_i + a)). \end{aligned}$$

2)-

$$S_{xx} = \sum_{i=1}^{N} (x_i - \overline{x})^2,$$

où

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

est la valeur moyenne des nombres x_i.

3)-

$$\widehat{\sigma} = \frac{\sum_{i=1}^{N} e_i^2}{N-2} = \frac{SSE}{N-2}.$$

8.3 Régression linéaire

8.3.1 Rappel

La série statistique double $(x_i, y_i), i = 1, ..., N$, est formée à partir de N observations du couple aléatoire (X, Y).

On a vu au début de ce cours qu'il existe deux nombres uniques a et b qui minimisent la sommes $\sum_{i=1}^{N} e_i^2$ des carrés des résidus e_i, et on a appelé la droite D_y, d'équation $y = ax + b$, droite de régression de la variable statistique Y en la variable statistique X. On a également défini le coefficient de corrélation linéaire r qui mesure la qualité de la régression linéaire entre les variables statistiques X et Y.

Cette procédure se généralise, d'une façon naturelle à un couple (X, Y) de variables aléatoires.

8.3.2 Ajustement linéaire.

Pour une valeur donnée $X = x$ de la variable aléatoire X, la valeur de Y est une variable aléatoire qu'on note ici $Y|_x$. L'ajustement linéaire du couple (X, Y) signifie la modélisation de la valeur moyenne $E(Y|_x)$ (c'est-à-dire l'espérance) de $Y|_x$ par une expression affine de la valeur x de la variable X.

Les staticiens ont montré qu'il existent deux nombres α et β tels que, pour chaque valeur x de X, l'expression $\alpha x + \beta$ donne la meilleure approximation de $E(Y|_x)$.

La droite d'équation

$$E(Y|_x) = \beta x + \alpha$$

S'appelle droite de régression linéaire de (la variable aléatoire) Y en (la variable aléatoire) X.

Les coefficients α et β sont donnés par les formules suivantes.

$$\begin{aligned} \beta &= \frac{Cov(X, Y)}{\sigma_X^2}, \\ \alpha &= \overline{Y} - \beta \overline{X}. \end{aligned}$$

On montre également que les coefficients obtenus expérimentalement

$$\begin{aligned} b &= \frac{Cov(x, y)}{\sigma_x^2} \text{ et} \\ a &= \overline{y} - b\overline{x} \end{aligned}$$

sont des estimations ponctuelles sans biais des coefficients β et α.

8.3.3 Intervalle de confiance pour $E\left(Y|_{x_0}\right)$

On montre que bx_0+a est une estimation ponctuelle sans biais de $E\left(Y|_{x_0}\right)$, où b et a sont les paramètres de la régression linéaire de l'échantillon $(x_1, y_1), ..., (x_N, y_N)$. A partir de ce résultat, pour x_0 donné, on construit l'intervalle de confiance de $E\left(Y|_{x_0}\right)$, au niveau de risque α, suivant.

$$E\left(Y|_{x_0}\right) = \widehat{y}_0 \pm \overline{t}_{\theta,N-2}\sqrt{\widehat{\sigma}^2\left[\frac{1}{N} + \frac{(x_0-\overline{x})^2}{S_{xx}}\right]},$$

où $\overline{t}_{\theta,N-2}$ est calculée dans la table 3 de l'écart réduit de la loi de Student.

8.4 Test sur le coefficient de corrélation ρ

8.4.1 Test d'indépendance : $\rho = 0$

L'indépendance entre les deux variables X et Y, qui signifie l'absence de corrélation entre X et Y, est équivalente à la condition

$$\rho = 0$$

Etapes du test.

1)- Hypothèse nulle $H_0 : \rho = 0$

Hypothèse alternative H_1. L'une des trois suivantes :

a)- $H_1 : \rho = 0$ (test bilatéral),

b)- $H_1 : \rho > 0$ (test unilatéral à droite),

c)- $H_1 : \rho < 0$ (test unilatéral à gauche).

2)- La statistique du test est

$$T = \frac{R\sqrt{N-2}}{\sqrt{1-R^2}},$$

et elle suit la loi de Student de $dll = N-2$. Sa valeur observée est

$$T_O = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}}.$$

3)- Le seuil critique est à calculer sur la loi de Student de $dll = N - 2$, table 3.

4)- Décision.

a) - (test bilatéral). Le seuil critique est $\bar{t}_{\theta,N-2}$, on accepte H_0 si

$$|T_O| \leq \bar{t}_{\theta,N-2}.$$

b)- (test unilatéral à droite). le seuil critique est $t_{N-2,1-\alpha}$, on accepte H_0 si

$$T_O \leq \bar{t}_{2\theta,N-2}.$$

c)- (test unilatéral à gauche). le seuil critique est $-\bar{t}_{2\theta,N-2}$, on accepte H_0 si

$$T_O \geq -\bar{t}_{2\theta,N-2}.$$

On rappelle que $\bar{t}_{2\theta,N-2}$ est calculé dans la table 3 de l'écart réduit de la loi Student de $ddl = N - 2$.

8.4.2 Exemple

On étudie la corrélation entre les activités de deux enzymes sériques. On a obtenu :

- dans l'espèce humaine, on a obtenu, $r = -0.296$ pour un échantillon de 30 individus,

- dans l'espèce bovine, on a obtenu, $r = -0.452$ pour un échantillon de taille 21.

Pour chacune des deux espèces, au vu de ces résultats, leurs corrélations ρ_1 et ρ_2 sont-elles significativement différent de $\rho = 0$?

Solution.

Il s'agit de test bilatéral.

1)- L'hypothèse nulle est $H_0 : \rho_1 = 0$ contre l'hypothèse alternative $H_1 : \rho_1 \neq 0$ pour l'espèce humaine.

L'hypothèse nulle est $H_0 : \rho_2 = 0$ contre l'hypothèse alternative $H_1 : \rho_2 \neq 0$ pour l'espèce bovine.

Le risque θ n'est pas précisé, on pose $\theta = 0.05$.

2)- La statistique de test observée est

$$T_O = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}} = \frac{-0.296\sqrt{30-2}}{\sqrt{1-(-0.296)^2}} \simeq -1.64$$

pour l'espèce humain, et

$$T_O = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}} = \frac{-0.452\sqrt{21-2}}{\sqrt{1-(-0.452)^2}} \simeq 2.21$$

Pour l'espèce bovine.

3)- Le seuil critique pour l'espèce humaine ($\theta = 0.05, N = 30$) est

$$\overline{t}_{28,0.05} = \mathbf{2.048}$$

Le seuil critique pour l'espèce bovine ($\theta = 0.05, N = 21$) est

$$\overline{t}_{19,0.05} = \mathbf{2.093}$$

Décision.

- Pour l'espèce humaine : puisque

$$|T_O| = \mathbf{1.64} \leq t_{28,0.975} = \mathbf{2.048},$$

On accepte $H_0 : \rho_1 = 0$, les activités des deux enzymes sériques étudiés sont indépendantes.

- Pour l'espèce bovine : puisque

$$|T_O| = \mathbf{2.21} > t_{19,0.975} = \mathbf{2.093},$$

On rejette H_0 et on accepte $H_1 : \rho_1 \neq 0$, les activités des deux enzymes sériques étudiés ne sont pas indépendantes.

8.4.3 Exemple

On considère la série double suivante, échantillon d'une variable aléatoire double (X, Y) .

x	46	48	55	57	60	72	81	85	94
y	2.18	2.10	2.13	2.28	2.34	2.53	2.28	2.62	2.63

x	109	121	132	137	148	149	184	185	187
y	2.50	2.66	2.79	2.80	3.01	2.98	3.34	3.49	3.26

a)- Calculer le coefficient de corrélation r de cette série double.

b)- Tester, au risque $\theta = 0.05$, l'hypothèse " il y a une corrélation négative entre X et Y.

Solution.

1)-L'hypothèse nulle $H_0 : \rho = 0$, l'hypothèse alternative $H_1 : \rho < 0$, test unilatéral à gauche.

On a

$$\begin{aligned} N &= 18, \ \sum x_i = 1950, \ \sum y_i = 49.92, \\ \sum x_i^2 &= 251.970, \ \sum y_i^2 = 130.6074, \ \sum x_i y_i = 5530.92 \end{aligned}$$

$$\begin{aligned} r &= \frac{\sum_{i=1}^{18} x_i y_i - \overline{x} \sum_{i=1}^{18} y_i}{\sqrt{\sum_{i=1}^{18} \left(x_i^2 - \overline{x} \sum_{i=1}^{18} x_i\right) \sum_{i=1}^{18} \left(y_i^2 - \overline{y} \sum_{i=1}^{18} y_i\right)}} \\ r &= \frac{5530.92 - \frac{1950}{18} 49.92}{\sqrt{\left(251.970 - \frac{1950}{18} 1950\right) \left(130.6074 - \frac{49.92}{18} 49.92\right)}}. \\ &= \mathbf{0.0956} \end{aligned}$$

b)-

La statistique de test observée T_O est

$$T_O = \frac{0.0956 \times 4}{\sqrt{1 - 0.0956^2}} = \mathbf{0.4021}$$

Le seuil critique est

$$\overline{t}_{N-2,2\theta} = \overline{t}_{16,0.1} = \mathbf{2.120}$$

Décision.

Puisque

$$T_O = \mathbf{0.4021} > -\overline{t}_{N-2,2\alpha} = \mathbf{-2.120},$$

on ne peut pas rejeter l'hypothèse nulle $H_0 : \rho = 0$, il n'y a pas de corrélation négative entre les variables aléatoires X et Y.

8.4.4 Comparaison de ρ à une valeur donnée ρ_0

On conserve les données de la section précédente. Le test de l'hypothèse

$$\rho = \rho_0$$

où ρ_0 est différent de zéro est basé sur la transformation suivante appelée transformation de Fisher.

$$V = \frac{1}{2}\ln\left(\frac{1+R}{1-R}\right).$$

La variable aléatoire V suit approximativement une loi normale de moyenne et de variance

$$\mu_V = \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right),\ \sigma_V^2 = \frac{1}{N-3}.$$

La statistique de test pour l'hypothèse $H_0 : \rho = \rho_0$ (qui suit la loi normale réduite centrée) est

$$Z = \frac{V - \frac{1}{2}\ln\left(\frac{1+\rho_0}{1-\rho_0}\right)}{\frac{1}{\sqrt{N-3}}},$$

et par suite la valeur observée de la statistique de test est

$$Z_O = \frac{\frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) - \frac{1}{2}\ln\left(\frac{1+\rho_0}{1-\rho_0}\right)}{\frac{1}{\sqrt{N-3}}},$$

L'intervalle de rejet de l'hypothèse nulle H_0 dépend de l'hypothèse alternative H_1. Les trois cas possibles sont résumés dans le tableau suivant.

Hypothèse alternative H_1	intervalle de rejet de H_0
$\rho \neq \rho_0$	$\lvert Z_O\rvert \geq \overline{z}_\alpha$
$\rho < \rho_0$	$Z_O < -\overline{z}_{2\alpha}$
$\rho > \rho_0$	$Z_O > \overline{z}_{2\alpha}$

8.5 Tests sur la droite d'ajustement

1)- Soit un couple (X, Y) de variables aléatoires et soit

$$y = \beta x + \alpha$$

l'équation de la droite de régression de la variable aléatoire Y en la variable aléatoire X.

2)- Soit

$$(x_1, y_1), (x_2, y_2), ..., (x_N, y_N),$$

N observations de (X, Y), et soit

$$y = bx + a$$

l'équation de la droite de régression liée à cette série statistique double.

3)- Soit β_0 et α_0 deux nombres réels.

On explique dans ce qui suit comment, à partir de la série statistique double (x_i, y_i), utiliser des tests statistiques pour comparer la pente β de la droite de régression de la variable aléatoire Y en la variable aléatoire X et le nombre β_0, et aussi, la même chose pour la constante α de cette droite et le nombre α_0.

8.5.1 Test de l'hypothèse $\beta = \beta_0$.

Etapes du test.

1)- L'hypothèse nulle $H_0 : \beta = \beta_0$ (la pente β de la droite de régression de la variable aléatoire Y en la variable aléatoire X est égale à la valeur donnée β_0),

l'hypothèse alternative H_1 est l'une des trois hypothèses

a)- $\beta \neq \beta_0$ (test bilatéral),

b)- $\beta > \beta_0$ (test unilatéral à droite),

c)- $\beta < \beta_0$ (test unilatéral à gauche).

2)- La statistique de test observée est

$$T_O = \frac{b - \beta_0}{\sqrt{\frac{\widehat{\sigma}}{S_{xx}}}}.$$

3)- Le seuil critique est

a)-

$$\overline{t}_{N-2}, \theta,$$

à calculer dans la table 3 de l'écart réduit de la loi de Student.

b)-

$$\overline{t}_{N-2,2\theta}$$

c)-

$$-\overline{t}_{N-2,2\theta}.$$

4)- Décision.

a)- pour le test bilatéral $(H_1 : \beta \neq \beta_0)$, On accepte H_0 si

$$|T_O| \leq \overline{t}_{N-2,\theta}.$$

b)- pour le test unilatéral à droite $(H_1 : \beta > \beta_0)$, On accepte H_0 si

$$T_O \leq \overline{t}_{N-2,2\theta}.$$

c)- pour le test unilatéral à gauche $(H_1 : \beta < \beta_0)$, On accepte H_0 si

$$T_O \geq -\overline{t}_{N-2,2\theta}.$$

Cas particulier.

On peut tester l'hypothèse $H_0 : \beta = 0$. Si H_0 est rejetée, on dit parfois que la régression est significative.

8.5.2 Test de l'hypothèse $\alpha = \alpha_0$.

La statistique observée pour ce test est

$$T_O = \frac{a - \alpha_0}{\sqrt{\widehat{\sigma}^2 \left[\frac{1}{N} + \frac{\overline{x}^2}{S_{xx}}\right]}}$$

Décision.

a) pour le test alternative $H_1 : \alpha \neq \alpha_0$, on rejette l'hypothèse nulle $H_0 : \alpha = \alpha_0$ si

$$|T_O| \geq \overline{t}_{N-2,\theta},$$

b)- pour le test unilatéral à droite $(H_1 : \alpha > \alpha_0)$, on rejette l'hypothèse nulle $H_0 : \alpha = \alpha_0$ si

$$T_O > \overline{t}_{N-2,2\theta},$$

c)- pour le test unilatéral à gauche $(H_1 : \alpha < \alpha_0)$, on rejette l'hypothèse nulle $H_0 : \alpha = \alpha_0$ si

$$T_O < -\overline{t}_{N-2,2\theta}.$$

8.5.3 Exemples

Exemple 1.

Vingt observations d'un couple (X, Y) de variables aléatoires sont représentées dans le tableau suivant

numéro de l'observation	x	y
1	0.99	90.01
2	1.02	89.05
3	1.15	91.43
4	1.29	93.74
5	1.46	96.73
6	1.36	94.45
7	0.87	87.59
8	1.23	91.77
9	1.55	99.42
10	1.40	93.65
11	1.19	93.54
12	1.15	92.52
13	0.98	90.56
14	1.01	89.54
15	1.11	89.85
16	1.20	90.39
17	1.26	93.25
18	1.32	93.41
19	1.43	94.98
20	.095	87.33

Les calculs donnent pour la série statistique double (x_i, y_i), les valeurs suivantes.

$$\begin{aligned}\overline{x} &= \frac{1}{20}\left(\sum_{i=1}^{20} x_i\right) = \frac{23.92}{20} = 1.196,\\ \overline{y} &= \frac{1}{20}\left(\sum_{i=1}^{20} y_i\right) = \frac{1843.21}{20} = 92.1605\end{aligned}$$

$$\begin{aligned}S_{xx} &= \sum_{i=1}^{20} x_i^2 - \frac{1}{20}\left(\sum_{i=1}^{20} x_i\right)^2\\ &= 29.2892 - \frac{(23.92)^2}{20}\\ &= 0.68088.\end{aligned}$$

$$\begin{aligned}S_{xy} &= \sum_{i=1}^{20} x_i y_i - \frac{1}{20}\left(\sum_{i=1}^{20} x_i\right)\left(\sum_{i=1}^{20} y_i\right)\\ &= 2,214.6566 - \frac{(23.92)\,(1843.21)}{20}\\ &= 10.17744.\end{aligned}$$

Ainsi les coefficients a et b de la droite de régression de la série statistique double sont

$$\begin{aligned}b &= \frac{S_{xy}}{S_{xx}} = \frac{10.17744}{0.68088} = 14.94748,\\ a &= \overline{y} - b\overline{x}\\ &= 92.1605 - (14.94748)\,(1.196)\\ &= 74.28331.\end{aligned}$$

L'équation de la droite de régression de la série statistique (y_i) en la série statistique (x_i) est

$$y = 14.94748x + 74.28331$$

Les calculs donnent également

$$\begin{aligned} SSE &= \sum_{i=1}^{N} e_i^2 \\ &= \sum_{i=1}^{N} (y_i - \widehat{y}_i)^2 \\ &= \sum_{i=1}^{N} (y_i - (bx_i + a))^2 \\ &= \sum_{i=1}^{N} (y_i - (14.94748x_i + 74.28331))^2 \\ &= 21.25. \end{aligned}$$

Il en résulte

$$\widehat{\sigma}^2 = \frac{SSE}{N-2} = \frac{21.25}{18} = 1.180.$$

On applique maintenant des tests statistiques sur les paramètres de la régression linéaire du couple aléatoire (X, Y) au seuil de signification $\theta = 0.05$.

Test sur la pente β.

D'après l'échantillon (x_i, y_i) précédent, peut-on décider que la pente β est supérieure 12 ?

Solution.

On doit faire un test statistique unilatéral à droite.

L'hypothèse nulle est $H_0 : \beta = 12$. L'hypothèse alternative est $H_1 : \beta > 13$.

La statistique de test observée est

$$\begin{aligned} T_O &= \frac{b - \beta_0}{\sqrt{\frac{\widehat{\sigma}^2}{S_{xx}}}} \\ &= \frac{14.94748 - 12}{\sqrt{\frac{1.180}{0.68088}}} \\ &= 2.239. \end{aligned}$$

Le seuil critique est calculé sur la table de l'écart réduit de la loi de Student :

$$\overline{t}_{18,0.05} = 1.33.$$

Décision.

Puisque

$$T_O = 2.239 > \bar{t}_{18,0.05} = 1.33,$$

on rejette l'hypothèse nulle H_0 et on accepte l'hypothèse alternative H_1 : Au risque de $\theta = 0.05$, on peut confirmer que la pente β de la régression linéaire est supérieure à 12.

Test sur la constante α.

D'après l'échantillon (x_i, y_i) , peut-on décider que la constante α est inférieure à 75.

Solution.

On réalise le test unilatéral à gauche.

Hypothèse nulle $H_0 : \alpha = 76$

Hypothèse alternative $H_1 : \alpha < 75$.

La statistique de test observée est

$$\begin{aligned} T_O &= \frac{a - \alpha_0}{\sqrt{\widehat{\sigma}^2 \left[\frac{1}{N} + \frac{\bar{x}^2}{S_{xx}}\right]}} \\ &= \frac{74.28331 - 75}{\sqrt{1.18 \left[\frac{1}{20} + \frac{1.43}{0.680}\right]}} \\ &= -0.4496. \end{aligned}$$

Le seuil critique est

$$-\bar{t}_{18,0.05} = -1.33.$$

Décision.

Puisque

$$T_O = -0.4496 < -\bar{t}_{18,0.05} = -1.33,$$

on rejette H_0 est on accepte H_1 : la constante α est inférieure à 75.

Exemple 2.

On a mesuré l'absorption de la lumière par des solutions de 4-nitrophénol, de concentrations croissantes. On a obtenu les résultats suivants (pour une longueur d'onde 400 nm) :

concentration C (en mol$/l$)	1×10^{-5}	2×10^{-5}	3×10^{-5}	4×10^{-5}	5×10^{-5}
absorbance	0.1865	0.3616	0.5370	0.7359	0.9238

1)- peut-on admettre que la relation entre l'absorbance et la concentration est linéaire, c'est-à-dire que, la droite de régression passe par l'origine (au risque de 5%).

2)- comparer la valeur de la pente b obtenue à la valeur $\beta = 18100\ l/mol$ fournie par les ouvrages de référence sur le sujet ($\theta = 0.05$).

Solution.

1)- On répond à cette question par le test statistique $\alpha = 0$.

La statistique de test observée est

$$\begin{aligned} T_O &= \frac{a - \alpha_0}{\sqrt{\widehat{\sigma}^2 \left[\frac{1}{N} + \frac{\overline{x}^2}{S_{xx}}\right]}} \\ &= \frac{-0.00571 - 0}{\sqrt{\widehat{\sigma}^2 \left[\frac{1}{N} + \frac{\overline{x}^2}{S_{xx}}\right]}} \\ &= -0.62 \end{aligned}$$

et le seuil critique est

$$\overline{t}_{N-2,\theta} = \overline{t}_{3,0.05} = 3.182.$$

Puisque

$$|T_O| = 0.62 \leq \overline{t}_{3,0.05} = 3.182,$$

On accepte l'hypothèse nulle $\alpha = 0$: au risque de 5%, on admet que la relation entre l'absorbance A et la concentration C est linéaire.

2)- On réalise le test statistique $\beta = 18100\ l/mol$.

La statistique de test observée est

$$\begin{aligned} T_O &= \frac{b - \beta_0}{\sqrt{\frac{\widehat{\sigma}}{S_{xx}}}} \\ &= \frac{18489 - 18100}{\sqrt{\frac{\widehat{\sigma}}{S_{xx}}}} \\ &= 1.40. \end{aligned}$$

Puisque

$$T_O = 1.40 < \overline{t}_{3,0.05} = 3.182,$$

on accepte l'hypothèse nulle $\beta = 18100\ l/mol$.

Bibliographie

[1] D.C. Montgomery & G.C. Runger, Applied Statistics and Probability for Engineers (Third Edition), John Wiley & Sons, Inc. (2003).

[2] J. L. Devore, Probability and Statistics for Engineering and the Sciences (Seventh edition), Brooks-Cole Cengage Learning, (2009).

[3] G. H. Heiman, Basic Statistics for the Behavioral Sciences (Sixth Edition), Wandsworth Cengage Learning, 2011).

[4] E. S. Keeping, Introduction to statistical inference, Dover Edition, (1995).

[5] M. R. Chernick & R. H. Friis, Introductory Biostatistics for the Health Sciences, John Wiley & Sons, (2003).

[6] Michel Lejeune, Statistique, la théorie et ses applications (2ème Edition), Springer-Verlag France, collection statistiques et probabilités appliquées, (2010).

[7] M. L. Samuels, J. A. Witmer & A. A. Schaffner, Statistics for the life sciences (Fourth Edition), Pearson Education, Inc., (2012).

[8] R. L. Ott, An Introduction to Statistical Methods and Data Analysis (Sixth Edition), Brooks-Kole Cengage Leaning (2010).

[9] Renée Veysseyre, Aide-mémoire Statistique et probabilités pour l'ingénieur (2ème édition), Dunod, Paris (2006).

[10] R. R. Sokal and F. J. Rohlf, Introduction to Biostatistics (second edition), D over Publication, Inc, Mineola, New York (2009).

[11] S. Bernstein and R.Bernstein, Elements of Statistics II, Interential statistics, Schaum's Outline Series, McGRaw-Hill, New York... (1999).

[12] W. J. DeCoursey, Statistics and Probability for Engineering Applications With Microsoft ® Excel, Elsevier Science, USA (2003).

[13] R. E. Walpole, R. H. Myers, S. L. Myers & K. Ye, Probability & Statistics for Engineers & Scientists (Ninth edition), Prentice Hall, Boston...(2012).

[14] O. J. Dunn & V. A. Clark, Basic Statistics, A Primer for the Biomedical Sciences (Fourth Edition), J. Wiley & Sons Inc. Publication, New Jersey (2009).

[15] M. Laviéville, Statistique et probabilités, Rappels de cours et exercices corrigés, Dunod, Paris, (1996).

[16] F. Couty, J. Debord & D. Fredon, Probabilités et statistiques, résumé de cours et 157 exercices et problèmes corrigés, Dunod, Paris (1999.)

[17] C. T. Le, Introductory biostatistics, J. Wiley & Sons Publications, Hoboken New jerey (2003).

[18] A. Vidal, Statistique descriptive et inférentielle avec Exel, approche par l'exemple, Presses universitaires de Rennes Collection « Didact Statistique », Rennes (2004).

[19] G. Jergaud, Module Statistique 1, Département Bioscience végétales, INP. ENSAT (2006).

[20] F. Carrat, A. Mallet & V. Morice, Biostatistique PACES - UE4, Faculté de Médecine - Université Pierre et Marie Curie, Paris (2013).

[21] T. LE. Chap, Introductory Biostatistics, John Wiley & Sons publications, Inc., Hoboken, New Jersey (2003).

[22] P. Sprent & N.C. Smeeton, Applied nonparametric statistical methods, Chapman & Hall/CRC, London...(2001).