Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université 8 Mai 1945 Guelma

Faculté des Mathématiques et de l'Informatique
et des Sciences de la Matière
Département de Mathématiques

Polycopié de travaux dirigés

Intitulé

Exercices Corrigés en Biostatistiques

Destiné aux étudiants de 3ème année Licence

Ecologie et Environnement

Réalisé par :

Dr. MENACEUR Amor

2024

Table des matières

La biostatistique est une discipline essentielle dans le domaine des sciences de la vie, car elle permet d'analyser des données issues d'expériences biologiques, médicales et écologiques. En troisième année de Licence, l'enseignement des biostatistiques vise à doter les étudiants des compétences nécessaires pour interpréter et analyser rigoureusement des données biologiques à l'aide d'outils statistiques. Ce polycopié de travaux dirigés (TD) est conçu pour accompagner les étudiants dans la mise en pratique des concepts théoriques abordés en cours. Il regroupe une série d'exercices pratiques qui couvrent des thématiques clés de la biostatistique, telles que :

a. Les statistiques descriptives : mesures de tendance centrale, de dispersion, et de forme.

b. Les tests d'hypothèses : tests paramétriques et non paramétriques.

c. L'analyse de variance et la régression.

d. Les méthodes d'échantillonnage et d'estimation.

Les objectifs principaux de ce polycopié sont :

1. Connaître le vocabulaire particulier de la statistique.

2. Comprendre les principes du traitement des données.

3. Développer une compréhension approfondie des concepts de base en statistique descriptive et inférentielle.

4. Appliquer des méthodes statistiques pour analyser des données issues de la recherche en écologie.

5. Renforcer les compétences en manipulation et en interprétation des données à travers des exercices pratiques.

6. Déterminer les tests statistiques les plus adéquats pour vérifier vos hypothèses.

7. Savoir présenter vos résultats et justifier vos décisions en choisissant les bons indicateurs statistiques.

Les exercices proposés dans ce polycopié ont été soigneusement sélectionnés pour illustrer les différents concepts abordés, et pour aider les étudiants à mieux comprendre les enjeux des statistiques appliquées dans le contexte des sciences de la vie. Il est vivement recommandé de compléter ces TD avec des lectures et des recherches personnelles pour approfondir la compréhension des sujets traités.

Chapitre 1

Statistiques descriptives

1.1 L'essentiel du cours

1.1.1 Généralités

a. Les statistiques, les probabilités

Les statistiques sont des ensembles de données, d'observations : recensement, cadastre...(ce ne sont donc que du chiffres).

Les probabilités forment une branche des mathématiques et sont donc rigoureuse et exactes, pour cela elles travaillent sur des objets mathématiques parfaitement définis et abstraites (bien que toujours d'origine concrète).

La statistique et donc la science qui utilise les méthodes mathématiques (venant généralement des probabilités) pour étudier et analyser et des statistiques en vue :

D'en accroitre les connaissances scientifiques.

De planifier des stratégies.

D'aider à la prise de décision.

b. Domaines d'applications de la statistique

Partout, depuis les sciences les plus "approchées", en passant par la biologie et la médecine et, contrairement à ce que l'on aurait pu croire et de plus en plus, jusqu'aux sciences dites "exactes" comme la chimie et la physique : physique quantique, mécanique ondulatoire...

c. Définition préliminaires

c.1 Population : ensemble étudié par la statistique
c.2 Echantillon : tout sous-ensemble de la population.
c.3 Individu : élément de la population.
c.4 Caractère, facteur, variable, toute caractéristiques prise par les individus d'une population.

C'est caractères, facteurs, variables présentent diverses modalités, valeurs au constituent des classes. Ils sont :
c.4.1 Qualitatifs (ivs). Auquel cas ils ne se mesurent pas (on parle alors plutôt de caractères au facteur et de modalités ou classe). Ils peuvent alors être :

-classé par attribut que *les modalités* ne sont pas ordonnées ;

-classifie ou ordonnée, c'est à dire que les modalités possèdent une relation d'ordre ;
c.4.2 Quantitative. Auquel cas ils se mesurent (on parle alors plutôt de variables). Elles peuvent alors être :

Discrètes c'est à dire que les valeurs prises par la variable sont en quantité finie ou dénombrable.

Contenues elles peuvent alors prendre toutes les valeurs dans un des intervalle de nombres réels (éventuellement $]-\infty, +\infty[$).

On parle séries statistiques, c'est à dire une correspondance entre les individus d'une population et les valeurs des facteurs qui l'on étudie sur cette population.

1.1.2 Séries statistiques à une variable

a. Définitions

Lorsque l'étude statistique d'une population concerne un seul caractère X (variable) on parle d'une série statistique à une seule variable (série statistique univariée).

La série univariée est organisée (résumée) dans un tableau dit tableau des effectifs.
a.1 Variables discrètes

Le caractère statistique peut prendre un nombre fini raisonnable de valeurs (note, nombre d'enfants, nombre de pièces, ...). Dans ce cas, le caractère

statistique étudié est alors appelé un caractère discret. On représente ces données sous la forme du tableau

x_i	x_1 x_2 .....x_p	total
n_i	n_1 n_2 .....n_p	n

a.2 Variables continues

Dans le cas d'une série statistique où le caractère étudié est continu, on peut regrouper les valeurs du caractère en classes. On représente ces données sous la forme du tableau

Classes	$[x_1, x_2[$	$[x_2, x_3[$	...	$[x_p, x_{p+1}[$
n_i	n_1	n_2	...	n_p

b. Paramètres de position

Les paramètres de position (mode, médiane, moyenne) permettent de savoir autour de quelles valeurs se situent les valeurs d'une variable statistique.

b.1 Le mode

Le mode, noté M_o, est la modalité qui admet la plus grande fréquence :

$$f(Mo) = Max\left\{f_1, f_2, ..., f_p\right\},$$

Il est parfaitement défini pour une variable qualitative ou une variable quantitative discrète.

Dans le cas des données groupées en classes, le mode se calcule par la formule :

$$M_o = a_i + \frac{\Delta_1}{\Delta_1 + \Delta_2}(a_{i+1} - a_i),$$

où a_i : borne inférieure de la classe modale (classe correspondant au plus grand effectif).

L : amplitude de la classe modale $[a_i, a_{i+1}]$ et

$$\Delta_1 = \left|n_{i.} - n_{i-1}\right|, \Delta_2 = \left|n_{i.} - n_{i+1}\right|$$

avec n_i : effectif de la classe modale.

Pour une variable quantitative continue nous parlons de **classe modale** : c'est la classe dont la densité de fréquence est maximum.

Si les classes ont même amplitude la densité est remplacée par l'effectif ou la fréquence et nous retrouvons la définition précédente.

Nous définissons le mode, pour une variable quantitative continue, en tenant compte des densités de fréquence des 2 classes adjacentes par la méthode suivante.

La classe modale $[a_i, a_{i+1}[$ étant déterminée, le mode M_o vérifie :

$$M_o = a_i + \frac{\Delta_1}{\Delta_1 + \Delta_2}(a_{i+1} - a_i),$$

où

$$\Delta_1 = |n_{i.} - n_{i-1}|, \Delta_2 = |n_{i.} - n_{i+1}|$$

b.2 Moyenne arithmétique

Soient X une variable statistique discrète et $x_1, x_2, ..., x_k$ ses valeurs, pour lesquelles correspondent les effectifs $n_1, n_2, ..., n_k$. La moyenne arithmétique notée $\bar{x}$ de cette série statistique, est définie par :

$$\bar{x} = \frac{1}{N}\sum_{j=1}^{k} n_j x_j, \text{ avec } N = \sum_{i=1}^{p} n_i$$

b.3 La médiane

La médiane $Mé$ est telle que l'effectif des observations dont les modalités sont inférieures à $Mé$ est égal à l'effectif des observations dont les modalités sont supérieures à $Mé$.

b.3.2 Médiane d'une série statistique discrète

Soit une série statistique d'effectif total n, rangée par ordre croissant $(x_1, x_2, ..., x_n)$.

Si la valeur de l'effectif total n est impaire, i.e. $n = 2p+1$, alors la médiane $Mé$ est la valeur qui se trouve à l'ordre $p + 1$,

$$Mé = x_{p+1},$$

Si la valeur de l'effectif total n est paire, i.e. $n = 2p$, alors la médiane $Mé$ est la moyenne des valeurs qui se trouve à l'ordre p et $p + 1$,

$$M_{é} = \frac{x_p + x_{p+1}}{2},$$

b.3.3 Médiane d'une série statistique continue

Dans ce cas la médiane est donnée par la formule suivante :

$$M_{é} = a_i + \frac{(0.5 - f^c_{i-1})}{f_i}L.$$

f_{i-1}^c : Fréquence relative cumulée de la classe qui précède la classe médiane $[a_i, a_{i+1}]$ (classe qui divise l'effectif en deux).

f_i : Fréquence relative de la classe médiane.

L : amplitude de la classe médiane.

b.4 Quantiles

On utilise couramment les quartiles Q_1, Q_2 et Q_3.

Q_1 est le quartile d'ordre 1/4, représente 25% de l'´echantillon ;

Q_2 est le quartile d'ordre 1/2, représente 50% de l'´echantillon ;

Q_3 est le quartile d'ordre 3/4, représente 75% de l'´echantillon.

c. Paramètres de dispersion

Les paramètres de dispersion (étendue, intervalle interquartile,...) sont calculés pour les variables statistiques quantitatives.

Ils ne donnent pas une information complète sur une variable statistique X : en effet, deux variables qui ont la même moyenne peuvent se présenter avec des dispersions très différentes.

L'histogramme, ou le diagramme, des fréquences donnent déjà une idée qualitative de la dispersion.

c.1 Etendu

On appelle étendu, notée e, la différence entre la plus grande valeur et la plus petite valeur observée.

c.2 L'interquartile est la différence $Q_3 - Q_1$

c.3 L'intervalle interquartile est l'intervalle : $[Q_1, Q_3]$

c.4 Variance : La variance, notée $V(x)$ (ou S^2) est la moyenne du carré des écarts à la moyenne.

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

-La variance pour des données rangées ou groupées en classe devient :

$$S^2 = V(x) = \frac{1}{N} \sum_{i=1}^{p} n_i (x_i - \bar{x})^2 ,$$

où n_i désigne les effectifs de chaque donnée ou de chaque classe.

c.5 Ecart-type : L'écart type, noté σ_x (ou S)est la racine carré de la moyenne du carré des écarts à la moyenne, c'est à dire la racine carrée de la variance.

c.6 Le coefficient de variation

$$CV = \frac{\sigma_x}{\bar{x}}.$$

d. Coefficient d'asymétrie

Il existe plusieurs coefficients d'asymétrie. Les principaux sont les suivants.

·Le coefficient d'asymétrie de Yule fait intervenir la médiane et les quartiles, il est défini par

$$S = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{(Q_3 - Q_2) + (Q_2 - Q_1)}.$$

·Le coefficient d'asymétrie de **Pearson** fait intervenir le mode Mo : quand il existe, il est définie par

$$P = \frac{\bar{X} - M_o}{\sigma_x}.$$

·Le coefficient d'asymétrie de Fisher fait intervenir les moments centrés, il est défini par

$$\gamma = \frac{m_3(x)}{[\sigma_x]^3},$$

où

$$m_k(x) = \frac{1}{N} \sum_{i=1}^{p} n_i \left(x_i - \bar{x}\right)^k,$$

Lorsque le coefficient d'asymétrie est positif, la distribution est plus étalée à droite : on dit qu'il y a **oblicité à gauche**.

Lorsque le coefficient d'asymétrie est négatif, la distribution est plus étalée à gauche : on dit qu'il y a **oblicité à droite**.

Il est nul pour une distribution à densité de fréquence symétrique, telle la loi de Gauss.

1.1.3 Séries statistiques à 2 variables

Définition 1 : On appelle série statistique à deux variables (ou série statistique doubles) une série statistique à deux caractères sont étudiés simultanément.

On considère deux variables statistiques x et y observées sur une même population de n individus.

On note $x_1, x_2, \ldots, x_n$ les valeurs relevées pour la variable x et $y_1, y_2, \ldots, y_n$ les valeurs relevées pour la variable y.

Les couples $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ forment une série statistique à deux variables.

x	x_1	x_2	...	x_i	...	x_n
y	y_1	y_2	...	y_i	...	y_n

Définition 2 : Dans un repère orthogonal, l'ensemble des points M_i de coordonnées (x_i, y_i), avec $1 \leq i \leq n$, est appelé *le nuage de points* associé à la série statistiques $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ à deux variables.

Dans certaines distributions statistiques bidimensionnelles il est possible de calculer les moyennes, les variances et les écart-types marginaux.

Pour les moyennes

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{n} x_i \text{ et } \bar{y} = \frac{1}{N} \sum_{i=1}^{n} y_i.$$

Pour les variances

$$V(x) = \frac{1}{N} \sum_{i=1}^{n} x_i^2 - \bar{x}^2 \text{ et } V(y) = \frac{1}{N} \sum_{i=1}^{n} y_i^2 - \bar{y}^2.$$

Elle est utilisée pour le calcul de l'écart type :

$$\sigma_x = \sqrt{V(x)}, \sigma_y = \sqrt{V(y)}$$

Définition 3 : On appelle covariance de x et de y le nombre

$$\begin{aligned} Cov(x, y) &= \frac{1}{N} \sum_{i=1}^{n} (x_i - \bar{x})^2 (y_i - \bar{y})^2 \\ &= \frac{1}{N} \sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}. \end{aligned}$$

Théorème Lors d'un ajustement affine par la méthode des moindres carrés

1. La droite de régression D de y en x a pour équation $y = ax + b$ où

$$a = \frac{Cov(x, y)}{\sigma_x^2} \text{ et } \bar{y} = a\bar{x} + b.$$

2. La droite de régression $D\prime$ de x en y a pour équation $x = a'y + b'$ où

$$a' = \frac{Cov(x,y)}{\sigma_y^2} \text{ et } \bar{x} = a'\bar{y} + b',$$

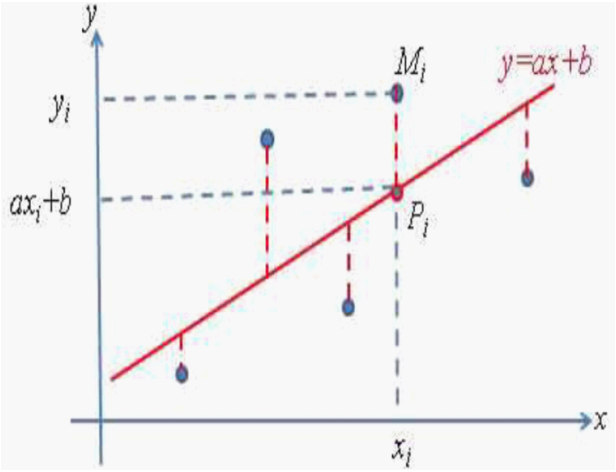


Figure 1

Définition 4 : le Coefficient de corrélation linéaire d'une série statistique à deux variables x et y est le nombre r défini par :

$$r = \frac{Cov(X,Y)}{\sigma_x \sigma_y},$$

Remarque

1. Si $r = 0$ alors il n'ya pas de corrélation entre x et y et les points (x_i, y_i) sont dispersés au hasard.

2. Si $0 < r < 1$ alors il y a une corrélation positive faible, moyenne ou forte entre x et y.

3. Si $-1 < r < 0$ alors il y a une corrélation négative faible, moyenne ou forte entre x et y.

1.2 Exercices corrigés

Exercice 1

Une crème dermatologique est testée sur un échantillon de 150 femmes également réparties en trois groupes en fonction de leur type de peau : mate, normale ou claire. On s'intéresse à l'éventuelle réaction allergique à la crème dermatologique en fonction du type de peau.

Type de peau	mate	normale	claire
Nombre d'allergies	3	7	13

1. Quelle est la proportion des femmes de cet échantillon ayant développé une allergie ?

2. Calculer cette proportion en fonction du type de peau.

3. Donner une représentation graphique en "camembert" des réactions allergisantes en fonction du type de peau.

Solution de l'exercice 1

Le type de peau (mate, normale, claire) est une variable qualitative à trois modalités.

1. Proportion de femmes ayant développé une allergie :

Nombre total de femmes : 150

Nombre total d'allergies : $3 + 7 + 13 = 23$

Proportion de femmes ayant développé une allergie : $\frac{23}{150} = 0.153333$, 15.33% des femmes de cet échantillon ont développé une allergie à cette crème.

2. Proportion de femmes ayant développé une allergie en fonction de son type de peau :

Type de peau	Nombre d'allergies	Nombre de non-allergies	total	Fréquences
mate	3	47	50	6%
normale	7	43	50	14%
claire	13	37	50	26%
total	23	127	150	15.33%

6% des femmes ayant la peau mate ont développé une allergie à cette crème.

14% des femmes ayant la peau normale ont développé une allergie à cette crème.

26% des femmes ayant la peau claire ont développé une allergie à cette crème.

3. Représentation graphique en "camembert" : les 23 individus doivent être représenté en droit parts de camembert en fonction de leur type de peau sachant que l'aire de chaque part doit être proportionnelle à l'effectif obtenu pour chaque type de peau. De plus, un camembert à un angle au centre de 360^0 devant correspondre à l'effectif total, c'est à dire 23 individus. D'où :

- L'angle au centre de la part de camembert correspondant aux 3 individus ayant le peau mate : $360\frac{3}{23} \simeq 47^0$

- L'angle au centre de la part de camembert correspondant aux 7 individus ayant le peau normale : $360\frac{7}{23} \simeq 110^0$

- L'angle au centre de la part de camembert correspondant aux 13 individus ayant le peau claire : $360\frac{13}{23} \simeq 203^0$ (Voir la figure 2)

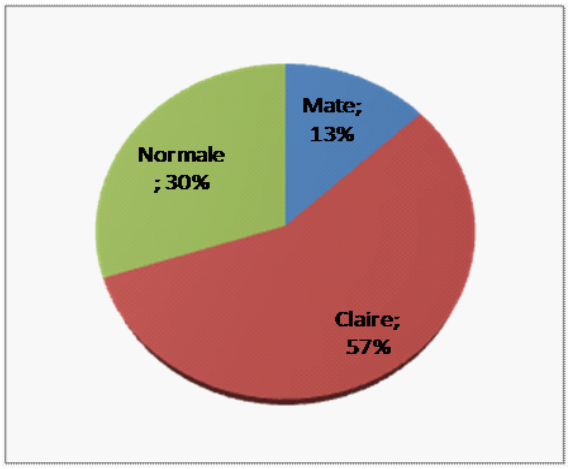


Figure 2 (Par Exel)

Exercice 2

On dispose des résultats d'une enquête concernant l'âge et les loisirs d'une population de **20** personnes :

Age	12	14	40	35	26	30	30	50	75	50
Loisir	*s*	*s*	*c*	*c*	*s*	*t*	*t*	*l*	*l*	*l*

30	45	25	55	28	25	50	40	25	35
t	*c*	*c*	*c*	*s*	*l*	*l*	*c*	*t*	*t*

Codification : s : Sport, c : Cinéma, t : Théâtre, l : Lecture

1. Faire l'étude du caractère « âge » : dresser le tableau statistique (effectifs, effectifs cumulés), calculer les valeurs de tendance centrale et ceux de la dispersion et tracez le diagramme en bâtons et la boite à moustaches de cette distribution

2. Faire l'étude du caractère « Loisir » dresser le tableau statistique, déterminer le mode et tracez le diagramme en bâtons et le diagramme à secteurs.

Solution de l'exercice 2

1. Age est une variable quantitative discrète

Age (x_i)	n_i	f_i	$f_i^c \uparrow$	$f_i x_i$
12	1	0.05	0.05	0.6
14	1	005	0.1	0.7
25	3	0.15	0.25	3.75
26	1	005	0.3	1.3
28	1	0.05	0.35	1.4
30	3	0.15	0.5	4.5
35	2	0.10	0.6	3.5
40	2	0.10	0.7	4
45	1	0.05	0.75	2.25
50	3	0.15	0.9	7.5
55	1	0.05	0.95	2.75
75	1	0.05	1	3.75
$\sum$	20	1	/	36

Les valeurs de tendance centrale (paramètre de position)

· Mode, $\mathrm{M}_0 = 25, 30, 50$

· Médiane (Q2), $M_é = Q_2 = 30$

· Moyenne, $\bar{X} = 36$

· Les quartiles $Q_1 = 25$ et $Q_3 = 45$

2. La variable loisir est une variable qualitative nominale

X	X_1	f_i
S	4	4/20
c	6	6/20
t	5	5/20
l	5	5/20
$\sum$	20	1

Le mode, c'est la modalité qui a le plus grand effectif : $M_0 = c$

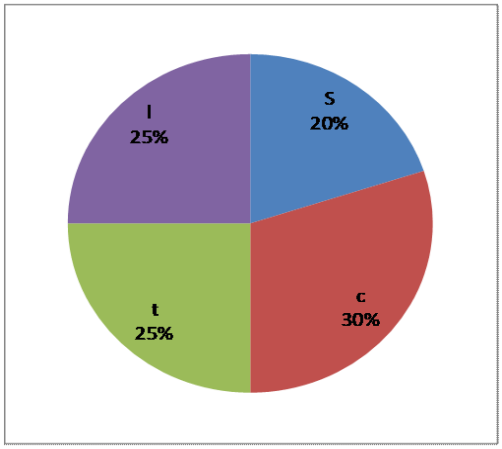


Diagramme secteurs

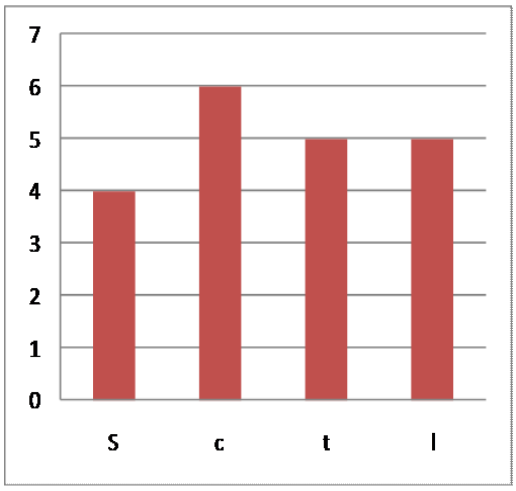


Diagramme en btons

Exercice 3

Luc a noté pendant 12 jours la température en degré Celsius, au lever du jour :

-3	-4	0	1	5	5	2	-1	-5	2	6	7

1. Calcule la moyenne de cette série.
2. **a.** Range cette série statistique dans l'ordre croissant.
b. Détermine la médiane de cette série.
c. Détermine les quartiles de cette série.
3. Calcule l'étendue de cette série de données.
Solution de l'exercice 3
1. Calcul de la moyenne
Soit $\bar{x}$ la moyenne, on a :

$$\bar{x} = \frac{1}{12}\sum_{i=1}^{12} x_i = 1.25$$

La température moyenne s'élève à **1,25** degré Celsius.
2. a. Série dans l'ordre croissant :

$$-5; -4; -3; -1; 0; 1; 2; 2; 5; 5; 6; 7$$

b. Détermination de la médiane
Comme il y a 12 valeurs, la médiane est comprise entre la 6ème et 7ème valeur qui partage la série en deux séries de 6 valeurs, soit la valeur

$$\begin{aligned} M_0 &= \frac{x_7 + x_6}{2} \\ &= \frac{1+2}{2} = 1.5. \end{aligned}$$

La médiane de cette série est $M_{\acute{e}} = 1,5°C$
c. Détermination des quartiles de cette série
On détermine le premier quartile Q_1 : On calcule $\frac{1}{4} \times 12 = 3$, alors Q_1 est la 3ème valeur de la série.
Donc: $Q_1 = -3°C$
On détermine le troisième quartile Q_3 : On calcule $\frac{3}{4} \times 12 = 9$, alors Q_3 est la 9ème valeur de la série.
Donc: $Q_3 = 5°C$, et $Q_2 = M_{\acute{e}} = 1.5°C$
3. Calcul de l'étendue
La plus petite température est -5.
La température la plus élevée est 7.
On a

$$e = 7 - (-5) = 12$$

L'étendue est donc 12.

Exercice 4

Soit X est le nombre de retards pour le premier trimestre sur une population de 200 entreprises pharmaceutiques :

x_i	0	1	2	3	4	total
n_i	24	62	52	36	26	200

1. Dresser le tableau statistique

2. Quel est le caractère étudié ? Sa nature ? justifier

3. Calculer la moyenne, la variance et en déduire l'écart-type.

4. Calculer le mode.

5. Déterminer la médiane et les quartiles.

6. Faire la représentation graphique associée à cette distribution.

Solution de l'exercice 4

1. Dresser le tableau statistique

x_i	0	1	2	3	4	total
n_i	24	62	52	36	26	200
f_i	0.12	0.31	0.26	0.18	0.13	1
$f_i^c \uparrow$	0.12	0.43	0.69	0.87	1	/
$n_i x_i$	0	62	104	108	104	378
$n_i x_i^2$	0	62	208	324	416	1010

2. Quel est le caractère étudié ? Sa nature ? justifier.

Il s'agit de nombre de retards pour le premier trimestre de 200 entreprises.

Sa nature est quantitative discrète, car les modalités sont mesurables et sont sous forme de valeurs isolées.

3. Calculer la moyenne, la variance et en déduire l'écart-typ

La moyenne

$$\begin{aligned} \bar{x} &= \frac{1}{200}\sum_{i=1}^{5} n_i x_i \\ &= \frac{378}{200} = 1.89, \end{aligned}$$

La varianceLa variance :

$$\begin{aligned} V(x) &= \frac{1}{200}\sum_{i=1}^{5} n_i x_i^2 - \bar{x}^2 \\ &= \frac{1010}{200} - (1.89)^2 \\ &= 1.478, \end{aligned}$$

L'écart-type :

$$\begin{aligned} \sigma &= \sqrt{1.478} \\ &= 1.216, \end{aligned}$$

4. Calculer le mode

Le caractère est quantitatif discret donc le mode et le caractère au plus grand effectif

$$M_o = 1,$$

5. Déterminer la médiane et les quartiles

La médiane : La première valeur de la variable x_i où la fréquence cumulée est supérieure à $0,50$, donc $M_é = 2$.

Les quartiles

Q_1 : La première valeur de la variable x_i où la fréquence cumulée est supérieure à $0,25$. Donc $Q_1 = 1$.

$$Q_2 = M_é = 2,$$

Q_3 : La première valeur de la variable x_i où la fréquence cumulée est supérieure à $0,75$. Donc $Q_3 = 3$.

6. Faire la représentation graphique associée à cette distribution

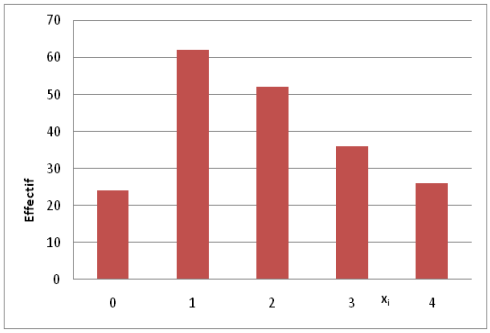


Diagramme en btons

Exercice 5

Résultats d'une enquête sur la hauteur d'un type de plante dans la forêt amazonienne

X_i	n_i
$[4, 6[$	20
$[6, 8[$	40
$[8, 10[$	80
$[10, 15[$	30
$[15, 20[$	20
$[20, 30[$	10

1. Compléter le tableau statistique (valeurs centrales, effectifs cumulés, fréquence, fréquences cumulés)

2. Déterminez les valeurs de tendance centrale de la distribution : moyenne, mode et les quartiles.

Solution de l'exercice 5

1.

Les classes	n_i	c_i	$n_i \uparrow$	f_i	$f_i^c \uparrow$	$f_i x_i$	d_i
$[4, 6[$	20	5	20	0.1	0.1	0.5	10
$[6, 8[$	40	7	60	0.2	0.3	1.4	20
$[8, 10[$	80	9	140	0.4	0.7	3.6	40
$[10, 15[$	30	12.5	170	0.15	0.85	1.875	6
$[15, 20[$	20	17.5	190	0.1	0.95	1.75	4
$[20, 30[$	10	25	200	0.05	1	1.25	1
$\sum$	200	/	/	1	/	10.375	/

2.

$$d_i = \frac{n_i}{a_{i+1} - a_i}.$$

La classe modale $[8, 10[$ (la classe qui a la plus grande densité)

$$\begin{aligned} M_0 &= x_i + \frac{\Delta_1}{\Delta_1 + \Delta_2}(a_{i+1} - a_i) \\ &= 8 + \frac{(40-20)}{(40-20)+(40-6)}(10-8) \\ &= 9.176, \end{aligned}$$

$$\begin{aligned} Q_1 &\in [6, 8[, \\ Q_1 &= x_i + \frac{0.25 - f_{i-1}^{c} \uparrow}{f_i}(a_{i+1} - a_i) \\ &= 6 + \frac{0.25 - 0.1}{0.2}(8-6) \\ &= 7.5, \end{aligned}$$

$$\begin{aligned} M_{\acute{e}} &= Q_2 \in [8, 10[, \\ Q_2 &= x_i + \frac{0.5 - f_{i-1}^{c} \uparrow}{f_i}(a_{i+1} - a_i) \\ &= 8 + \frac{0.5 - 0.3}{0.4}(10-8) \\ &= 9, \end{aligned}$$

$$\begin{aligned} Q_3 &\in [10, 15[, \\ Q_2 &= x_i + \frac{0.75 - f_{i-1}^{c} \uparrow}{f_i}(a_{i+1} - a_i) \\ &= 10 + \frac{0.75 - 0.7}{0.15}(15-10) \\ &= 11.667, \end{aligned}$$

Exercice 6

On étudie X l'âge des employés d'une entreprise, on obtient :

Age	$[20, 25[$	$[25, 30[$	$[30, 35[$	$[35, 40[$	$[40, 45[$	$[45, 50[$	$[50, 55[$	total
Effectifs n_i	150	300	600	750	450	600	150	3000

1. Dresser le tableau statistique.
2. Quel est le caractère étudié ? Sa nature ? justifier.
3. Calculer la moyenne, la variance et en déduire l'écart-type.
4. Calculer le mode.
5. Déterminer la médiane et les quartiles.
6. Faire la représentation graphique associée à cette distribution.

Solution de l'exercice 6

1. Dresser le tableau statistique :

Age	$[20, 25[$	$[25, 30[$	$[30, 35[$	$[35, 40[$	$[40, 45[$	$[45, 50[$	$[50, 55[$	total
n_i	150	300	600	750	450	600	150	3000
c_i	22.5	27.5	32.5	37.5	42.5	47.5	52.5	/
f_i	0.05	0.1	0.2	0.25	0.15	0.2	0.05	
$f_i^c \uparrow$	0.05	0.15	0.35	0.60	0.75	0.95	1	/
$n_i c_i$	3375	8250	19500	28125	19125	28500	7875	114750
$n_i c_i^2$	75937.5	226875	633750	1054687.5	812812.5	1353750	413437.5	4571250

2. Il s'agit de l'âge de 3000 employés d'une entreprise.

Sa nature est quantitative continue. Car les modalités sont mesurables et sont sous forme d'intervalles.

3. Calculer la moyenne, la variance et en déduire l'écart-type.

La moyenne :

$$\begin{aligned} \bar{x} &= \frac{1}{N}\sum_{i=1}^{7} n_i c_i \\ &= \frac{114750}{3000} = 38.25, \end{aligned}$$

La variance :

$$\begin{aligned} V(x) &= \frac{1}{N}\sum_{i=1}^{7} n_i c_i^2 - \bar{x}^2 \\ &= \frac{4571250}{3000} - (38.25)^2 \\ &= 60.688, \end{aligned}$$

L'écart-type :

$$\begin{aligned} \sigma &= \sqrt{V(x)} \\ &= 7.79, \end{aligned}$$

4. Calculer le mode

On a la classe modale celle qui a le plus grand effectif : $[35; 40[$, donc

$$\begin{aligned} M_0 &= x_i + \frac{\Delta_1}{\Delta_1 + \Delta_2}(a_{i+1} - a_i) \\ &= 35 + \frac{(750 - 600)}{(750 - 600) + (750 - 450)}(40 - 35) = 36.667, \end{aligned}$$

5. Déterminer la médiane et les quartiles

1.

$$\begin{aligned} M_{\acute{e}} &= Q_2 \in [35, 40[, \\ Q_2 &= x_i + \frac{0.5 - f_{i-1}^{c} \uparrow}{f_i}(a_{i+1} - a_i) \\ &= 35 + \frac{0.5 - 0.35}{0.25}5 = 38, \end{aligned}$$

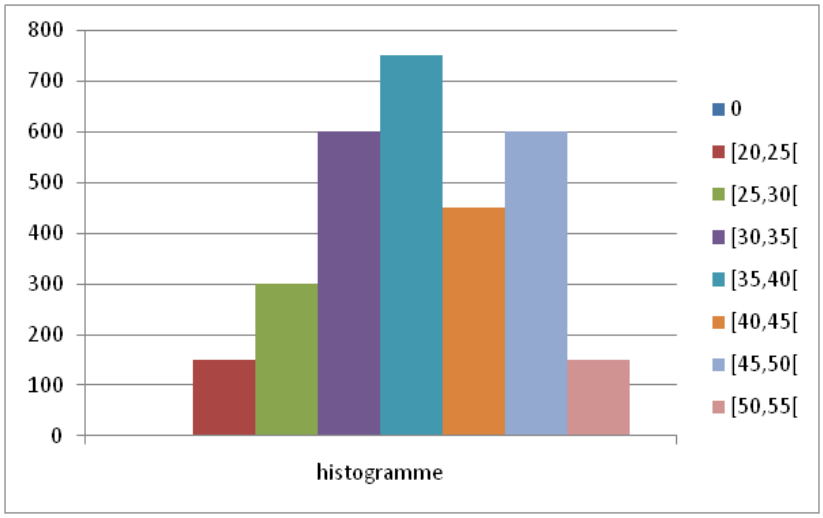
Les quartiles

$$\begin{aligned} Q_1 &\in [30, 35[, \\ Q_1 &= x_i + \frac{0.25 - f_{i-1}^{c} \uparrow}{f_i}(a_{i+1} - a_i) \\ &= 30 + \frac{0.25 - 0.15}{0.2}5 = 32.5, \end{aligned}$$

$$M_{\acute{e}} = Q_2 = 38,$$

$$\begin{aligned} Q_3 &\in [45, 50[, \\ Q_2 &= x_i + \frac{0.75 - f_{i-1}^{c} \uparrow}{f_i}(a_{i+1} - a_i) \\ &= 45 + \frac{0.75 - 0.75}{0.2}5 = 45. \end{aligned}$$

6. Faire la représentation graphique associée à cette distribution



Exercice 7

Dans la fabrication de comprimés effervescents, il est prévu que chaque comprimé doit contenir 1625 mg de bicarbonate de sodium. Afin de contrôler la fabrication de ces médicaments, on a prélevé un échantillon de 150 comprimés, et on a mesuré la quantité de bicarbonate de sodium *(BS)* pour chacun d'eux. On a obtenu les résultats suivants

BS (mg)	$[1610, 1615[$	$[1615, 1620[$	$[1620, 1625[$	$[1625, 1630[$	$[1630, 1635[$
Effectifs	17	28	42	45	18

1-Calculer les fréquences correspondantes ainsi que les fréquences cumulées croissantes.

2-Calculer les caractéristiques : classe modale, le mode, la médiane et les quartiles Q_1, Q_2et Q_3?

3-Calculer le coefficient de symétrie de Yule, conclusion ?

Solution de l'exercice 7

1.

BS (mg)	$[1610, 1615[$	$[1615, 1620[$	$[1620, 1625[$	$[1625, 1630[$	$[1630, 1635[$	total
n_i	17	28	42	45	18	150
f_i	0.113	0.186	0.280	0.300	0120	1
$f_i^c \uparrow$	0.113	0.300	0.58	0.88	1	/

2. Classe modale est une classe dons l'éffictif est maximum, donc la classe modale est $[1620, 1625[$.

$$\begin{aligned} M_0 &= x_i + \frac{\Delta_1}{\Delta_1 + \Delta_2}(a_{i+1} - a_i) \\ &= 1625 + \frac{(45-42)}{(45-42)+(45-18)}(1625-1625) \\ &= 1625.5, \end{aligned}$$

$$\begin{aligned} M_{\acute{e}} &= Q_2 \in [1620, 1625[, \\ Q_2 &= x_i + \frac{0.5 - f_{i-1}^{c} \uparrow}{f_i}(a_{i+1} - a_i) \\ &= 1620 + \frac{0.5 - 0.3}{0.28}5 \\ &= 1623.57, \end{aligned}$$

Les quartiles

$$\begin{aligned} Q_1 &\in [1615, 1620[, \\ Q_1 &= x_i + \frac{0.25 - f_{i-1}^{c} \uparrow}{f_i}(a_{i+1} - a_i) \\ &= 1615 + \frac{0.25 - 0.113}{0.186}5 \\ &= 1618.\,7, \end{aligned}$$

$$M_{\acute{e}} = Q_2 = 1623.57$$

$$\begin{aligned} Q_3 &\in [1625, 1630[, \\ Q_2 &= x_i + \frac{0.75 - f_{i-1}^{c} \uparrow}{f_i}(a_{i+1} - a_i) \\ &= 1625 + \frac{0.75 - 0.58}{0.3}5 \\ &= 1627.83, \end{aligned}$$

2. Le coefficients de symétrie de Yule

$$\begin{aligned} S &= \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{Q_3 - Q_1} \\ &\quad \frac{(1627.83 - 1623.57) - (1623.57 - 1618.\,7)}{1627.83 - 1618.\,7} \\ &= -0.06.\,681\,3 \approx 0, \end{aligned}$$

$S = 0 \Rightarrow$ symétrie parfaite.

Exercice 8

Au cours d'une fabrication de fromages de chèvres, on a relevé les masses suivantes des fromages :

Masse en grammes	$[85, 90[$	$[90, 95[$	$[95, 100[$	$[100, 105[$	$[105, 110[$	$[110, 115[$
Effectifs n_i	9	14	18	25	16	17

1-Calculer les fréquences correspondantes ainsi que les fréquences cumulées croissantes.

2-Calculer les caractéristiques :

a-classe modale et le mode M_0, **b**-la médiane $Mé$, **c** -les quartiles Q_1, Q_2, Q_3.

Solution de l'exercice 8

1.

Masse en grammes	$[85, 90[$	$[90, 95[$	$[95, 100[$	$[100, 105[$	$[105, 110[$	$[110, 115[$	total
n_i	9	14	18	25	16	17	99
f_i	0.091	0.141	0.182	0.252	0.162	0.172	1
$f_i^c \uparrow$	0.091	0.232	0.414	0.666	0.828	1	/

2. **a**. Classe modale est une classe dons l'éffictif est maximum, donc la classe modale est $[100, 105[$.

$$\begin{aligned} M_0 &= x_i + \frac{\Delta_1}{\Delta_1 + \Delta_2}(a_{i+1} - a_i) \\ &= 100 + \frac{(25-18)}{(25-18)+(25-16)}5 \\ &= 102.187 \end{aligned}$$

b.

$$\begin{aligned} M_{é} &= Q_2 \in [100, 105[, \\ Q_2 &= x_i + \frac{0.5 - f_{i-1}^c \uparrow}{f_i}(a_{i+1} - a_i) \\ &= 100 + \frac{0.5 - 0.414}{0.252}5 \\ &= 101.71 \end{aligned}$$

c. Les quartiles

$$\begin{aligned} Q_1 &\in [95, 100[, \\ Q_1 &= x_i + \frac{0.25 - f_{i-1}^{c} \uparrow}{f_i}(a_{i+1} - a_i) \\ &= 95 + \frac{0.25 - 0.232}{0.182}5 \\ &= 95.495 \end{aligned}$$

$$M_{\acute{e}} = Q_2 = 101.71$$

$$\begin{aligned} Q_3 &\in [105, 110[, \\ Q_2 &= x_i + \frac{0.75 - f_{i-1}^{c} \uparrow}{f_i}(a_{i+1} - a_i) \\ &= 105 + \frac{0.75 - 0.666}{0.162}5 \\ &= 107.59. \end{aligned}$$

Exercice 9

Lors d'un contrôle d'une chaine de médicaments, on s'intéresse au nombre de comprimés défectueux dans un lot. L'étude de 200 lots a donné les résultats suivants :

Nombre de comprimés par lot : x_i	0	1	2	3	4	5
Nombre de lots : n_i	75	53	39	23	9	1

1. Calculer la moyenne, le mode et les quartiles du nombre de comprimés défectueux pour ces 200 lots.

2. Calculer la variance, l'ecart-type et le coefficient de variation du nombre de comprimés défectueux pour ces 200 lots.

Solution de l'exercice 9

1.

x_i	0	1	2	3	4	5	$Total$
n_i	75	53	39	23	9	1	200
$n_i \uparrow$	75	128	167	190	199	200	/
f_i	37.5%	26.5%	19.5%	11.5%	4.5%	0.5%	100%
$f_i^c \uparrow$	37.5%	64%	83.5%	95%	99.5%	100%	/

$$\begin{aligned}\bar{X} &= m = \frac{1}{n}\sum_{i=1}^{6} n_i x_i \\ &= \frac{1}{200}(241) \\ &= 1.205,\end{aligned}$$

Mode : L'effectif le plus grand, c'est-à-dire 75, correspond à un nombre de comprimés égal à 0.

Quartiles : Le premier quartile Q_1 est la valeur qui permet de partager l'échantillon en deux parties de telle sorte que 25% de l'échantillon soit inférieur à Q_1 et donc 75% supérieur à Q_1. L'étude des fréquences cumulées montre que 25% correspond à " 0 comprimé défectueux" :

$$Q_1 = 0.$$

Le deuxième quartile Q_2 ou la médiane $(Q_2 = M_{\acute{e}})$, est la valeur qui permet de partager l'échantillon en deux parties de telle sorte que 50% de l'échantillon soit inférieur à Q_2 et donc 50% supérieur à Q_2. L'étude des fréquences cumulées montre que 50% correspond à " 1 comprimé défectueux" :

$$Q_2 = 1.$$

Le troisième quartile Q_3 est la valeur qui permet de partager l'échantillon en deux parties de telle sorte que 75% de l'échantillon soit inférieur à Q_3 et donc 25% supérieur à Q_3. L'étude des fréquences cumulées montre que 75% correspond à " 2 comprimé défectueux" :

$$Q_3 = 2.$$

2. Variance :

$$\begin{aligned}V(x) &= \frac{1}{n}\sum_{i=1}^{6} n_i x_i^2 - \bar{x}^2 \\ &= \frac{1}{200}(585) - 1.205^2 \\ &= 1.473.\end{aligned}$$

Ecart-type :

$$\sigma(x) = \sqrt{V(x)} = 1.213.$$

Coefficient de variation

$$\begin{aligned} CV &= \frac{\sigma}{\bar{X}} = \frac{1.213}{1.205} \\ &= 1.007, \end{aligned}$$

donc $CV \approx 100.7\%$

Exercice 10

Soit la distribution suivante des salaires d'une entreprise. On note par n_i les effectifs correspondants en nombre de salariés et F_i les fréquences cumulées croissante de cette distribution.

C_i	$[500, 700[$	$[700, 1100[$	$[1100, 1300[$	$[1300, 1500[$	$[1500, 1900[$	total
F_i	0.04	0.14	0.44	0.96	1	/

Sachant que $V(x) = 49300, \sum_{i=1}^{5} f_i x_i^2 = 1662200$, $\sum_{i=1}^{5} n_i x_i = 190500$ avec C_i : Classes des salaires.

Calculer l'effectif de chaque classe ainsi que l'effectif total ?

Solution de l'exercice 10

les classes	$[500, 700[$	$[700, 1100[$	$[1100, 1300[$	$[1300, 1500[$	$[1500, 1900[$	total
$f_i^c \uparrow$	0.04	0.14	0.44	0.96	1	/
f_i	0.04	0.10	0.30	0.52	0.04	1
n_i	n_1	n_2	n_3	n_4	n_5	N

$$V(x) = \sum f_i x_i^2 - \bar{x}^2,$$

donc

$$\begin{aligned} \bar{x}^2 &= -V(x) + \sum f_i x_i^2 = 1612900 \\ &\Rightarrow \bar{x} = m = 1270, \end{aligned}$$

et

$$\bar{x} = \frac{1}{N} \sum n_i x_i \Rightarrow N = \frac{\sum n_i x_i}{\bar{x}},$$

alors

$$N = \frac{190500}{1270} = 150.$$

L'efectif total $N = 150$ et

$$f_i = \frac{n_i}{N} \Rightarrow n_i = f_i \times N,$$

donc

$$\begin{aligned} n_1 &= f_1 \times N = 0.04 \times 150 = 6, \\ n_2 &= 15, n_3 = 45, n_4 = 78, n_5 = 6. \end{aligned}$$

Exercice 11

Archibald, le fermier, décide ce dimanche matin d'étudier le poids X de ses oeufs. Pour les 100 oeufs de sa récolte dominicale, il obtient :

X : poids (g)	$[0;40[$	$[40;60[$	$[60;70[$	$[70;80[$	$[80;100[$
n_i	10	15	20	35	20

1.Quel est le caractère étudié ? Sa nature ? justifier
2.Calculer la moyenne, la variance et en déduire l'écart-type
3.Calculer le mode
4.Déterminer la médiane et les quartiles
5. Représenter graphiquement le mode et la médiane
6. Déterminer les coefficients empiriques d'asymétrie de Pearson et yulle ? conclure ?

Solution de l'exercice 11

1.Quel est le caractère étudié ? Sa nature ? justifier

Il s'agit de poids de ses oeufs.

Sa nature est quantitative continue, car les modalités sont mesurables et sont sous forme d'intervalles.

2. Calculer la moyenne, la variance et en déduire l'écart-type

les classes	$[0;40[$	$[40;60[$	$[60;70[$	$[70;80[$	$[80;100[$	total
n_i	10	15	20	35	20	100
Centre (ci)	20	50	65	75	90	/
f_i	0.1	0.15	0.2	0.35	0.2	1
$f_i^c \uparrow$	0.1	0.25	0.45	0.8	1	/
$n_i c_i$	200	750	1300	2625	1800	6675
$n_i c_i^2$	4000	37500	84500	196875	162000	484475

$$\begin{aligned} \bar{x} &= \frac{1}{N}\sum_{i=1}^{5} n_i c_i \\ &= \frac{6675}{100} = 66.75, \end{aligned}$$

$$\begin{aligned} V(x) &= \frac{1}{N}\sum_{i=1}^{5} n_i c_i^2 - \bar{x}^2 \\ &= \frac{484475}{100} - (66.75)^2 \\ &= 389.19, \end{aligned}$$

L'écart-type :

$$\begin{aligned} \sigma &= \sqrt{389.19} \\ &= 19.73, \end{aligned}$$

3. Calculer le mode
On a la classe modale celle qui a le plus grand effectif : $[70; 80[$

$$\begin{aligned} M_0 &= x_i + \frac{\Delta_1}{\Delta_1 + \Delta_2}(a_{i+1} - a_i) \\ &= 70 + \frac{(35-20)}{(35-20)+(35-20)}10 \\ &= 75 \end{aligned}$$

4. Déterminer la médiane et les quartiles

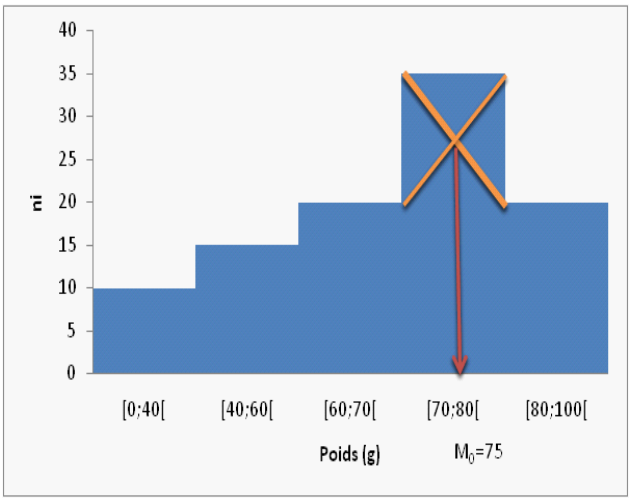
$$\begin{aligned} M_{\acute{e}} &= Q_2 \in [70, 80[, \\ Q_2 &= x_i + \frac{0.5 - f_{i-1}^{c} \uparrow}{f_i}(a_{i+1} - a_i) \\ &= 70 + \frac{0.5 - 0.45}{0.35}10 \\ &= 71.42 \end{aligned}$$

c. Les quartiles

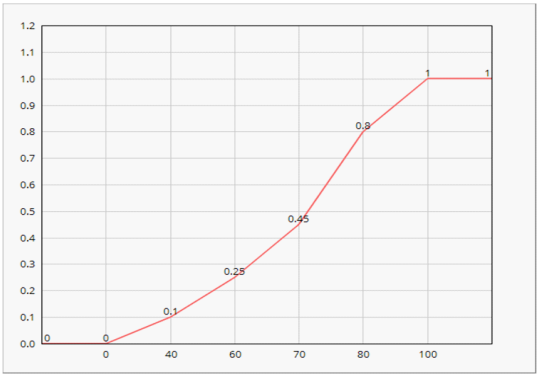
$$\begin{aligned} Q_1 &\in [60, 70[, \\ Q_1 &= x_i + \frac{0.25 - f_{i-1}^{c} \uparrow}{f_i}(a_{i+1} - a_i) \\ &= 60 + \frac{0.25 - 0.25}{0.2}10 \\ &= 60, \end{aligned}$$

$$\begin{aligned} Q_3 &\in [70, 80[, \\ Q_3 &= x_i + \frac{0.25 - f^c_{i-1}\uparrow}{f_i}(a_{i+1} - a_i) \\ &= 70 + \frac{0.75 - 0.45}{0.35}10 = 78.5, \end{aligned}$$

5



Histogramme



Courbe cumulative de $f^c_i \uparrow$

.

6. Le coefficients de symétrie de Yule

$$\begin{aligned} S &= \frac{(78.5 - 71.42) - (71.42 - 60)}{78.5 - 60} \\ &= -0.2346 < 0 \end{aligned}$$

Le coefficients de symétrie de Pearson

$$\begin{aligned} P &= \frac{\bar{x} - M_o}{\sigma_x} \\ &= \frac{66.75 - 75}{19.73} \\ &= -0.418\,14 < 0 \end{aligned}$$

Donc, la série statistiques est étalée vers la gauche.

Exercice 12

On considère deux groupes d'étudiants. Nous relevons leurs notes d'examens dans les deux tableaux suivants :

Note (groupe A)	8	9	10	11
Effectifs n_i	2	2	1	1

Note (groupe B)	6	8	9	13	14
Effectifs n_j	2	2	2	1	1

Calculer la moyenne et l'écart type de chaque groupe. Comparer les deux groupes.

Solution de l'exercice 12

Dans un premier temps, nous remarquons que l'effectif total du groupe A est égal à 6 et celui du groupe B est égal à 8.

En utilisant la formule de la moyenne, nous obtenons

$$\bar{x}_A = \frac{1}{6} \sum_{i=1}^{4} n_i x_i = 9.2,$$

$$\bar{y}_B = \frac{1}{8} \sum_{j=1}^{5} n_j y_j = 9.1,$$

. On remarque que les moyennes sont très proches. Peut-on pour autant conclure que ces deux groupes ont des niveaux identiques ?

Nous répondons à cette question après le calcule des écarts type.

$$\begin{aligned} V_A(x) &= \frac{1}{6}\sum_{i=1}^{4} n_i x_i^2 - \bar{x}_A^2 \\ &= 1.232, \end{aligned}$$

$$\begin{aligned} V_B(y) &= \frac{1}{8}\sum_{j=1}^{5} n_j y_j^2 - \bar{y}_B^2 \\ &= 7.84, \end{aligned}$$

donc

$$\sigma_A = 1.11 \text{ et } \sigma_B = 2.8.$$

Nous remarquons que même si les deux groupes ont des moyennes quasiment identiques, le groupe B est beaucoup plus dispersé que le groupe A car $\sigma_B > \sigma_A$. Les étudiants de ce groupe ont des notes plus irréguliers. On peut dire donc que le groupe B est moins homogènes que le groupe A. En observant les valeurs du tableau, on voit que c'est cohérent.

Exercice 13

On considère la série double suivante

x_i	2	5	6	10	12
y_i	83	70	70	54	49

1. Calculer la covariance.

2. Déterminer l'équation de la droite de régression $y = ax + b$.

3. Calculer le coefficient de corrélation linéaire.

Solution de l'exercice 13

1.

	x_i	y_j	x_i^2	y_j^2	$x_i y_i$
	2	83	4	83^2	2×83
	5	70	25	70^2	5×70
	6	70	36	70^2	6×70
	10	54	100	54^2	10×54
	12	49	144	49^2	12×49
total	35	326	309	22006	2064

$$\bar{x} = \frac{1}{5}\sum_{i=1}^{5} x_i = \frac{35}{5} = 7,$$

$$\bar{y} = \frac{1}{5}\sum_{j=1}^{5} y_j = \frac{326}{5} = 65.2,$$

$$\begin{aligned} Cov(x,y) &= \frac{1}{5}\sum_{i=1}^{5} x_i y_i - \bar{x}\bar{y} \\ &= -43.\,6, \end{aligned}$$

2. La droite de régression de y en x a pour de l'équation $y = ax + b$.

$$\begin{aligned} V(x) &= \frac{1}{5}\sum_{i=1}^{5} x_i^2 - \bar{x}^2 \\ &= 12.8, \end{aligned}$$

$$\begin{aligned} V(y) &= \frac{1}{5}\sum_{j=1}^{5} y_j^2 - \bar{y}^2 \\ &= \frac{22006}{5} - (65.2)^2 \\ &= 150.\,16, \end{aligned}$$

donc

$$\begin{cases} a = \frac{Cov\ (x,y)}{V(x)} \\ \bar{y} = a\bar{x} + b, \end{cases}$$

$$\Rightarrow a = -3.4 \text{ et } b = 89,$$

Alors

$$y = -3.4x + 89$$

.

3. Le coefficient de corrélation

$$\begin{aligned} r &= \frac{Cov(x,y)}{\sqrt{V(x)}\sqrt{V(y)}} \\ &= \frac{-43.6}{\sqrt{12.8}\sqrt{150.16}} \\ &= -0.994\,50, \text{ (proche de -1)} \end{aligned}$$

alors, la corrélation lineaire entre les variables x et y est forte.

Exercice 14

Une étude de la liaison entre l'âge (années) et le diamètre (cm) dans une population d'arbres a été conduite sur un échantillon de 6 plantes

X : Age	2	3	5	6	8	9
Y : Diamètre	3	9	12	13	15	16

1. Calculer $\bar{x}, \bar{y}, V(x), V(y)$ et $Cov(x, y)$.

2. Déterminer par la méthode des moindres carrés la droite de régression du diamètre d'une plante (y) en fonction de l'âge d'une plante (x).

3. Calculer le coefficient de corrélation. Conclusion ?

Solution de l'exercice 14

1.

	x_i	y_j	x_i^2	y_j^2	$x_i y_i$
	2	3	4	9	6
	3	9	9	81	27
	5	12	25	144	60
	6	13	36	169	78
	8	15	64	225	120
	9	16	81	256	144
total	33	68	219	884	435

$$\bar{x} = \frac{1}{6}\sum_{i=1}^{6} x_i = \frac{33}{6} = 5.5,$$

$$\bar{y} = \frac{1}{6}\sum_{j=1}^{6} y_j = \frac{68}{6} = 11.33,$$

$$\begin{aligned} V(x) &= \frac{1}{6}\sum_{i=1}^{6} x_i^2 - \bar{x}^2 \\ &= \frac{219}{6} - (5.5)^2 = 6.25, \end{aligned}$$

$$\begin{aligned} V(y) &= \frac{1}{6}\sum_{j=1}^{6} y_j^2 - \bar{y}^2 \\ &= \frac{884}{6} - (11.33)^2 = 18.96, \end{aligned}$$

$$\begin{aligned} Cov(x,y) &= \frac{1}{6}\sum_{i=1}^{6} x_i y_i - \bar{x}\bar{y} \\ &= \frac{435}{6} - (5.5 \times 11.33) = 10.18. \end{aligned}$$

2. La droite de régression de y en x a pour de l'équation $y = ax + b$

$$\begin{cases} a = \frac{Cov\ (x,y)}{V(x)} \\ \bar{y} = a\bar{x} + b, \end{cases}$$

$$\Rightarrow a = 1.63 \text{ et } b = 2.365,$$

donc $y = 1.63x + 2.365$.

3. Le coefficient de corrélation

$$\begin{aligned} r &= \frac{Cov(x,y)}{\sqrt{V(x)}\sqrt{V(y)}} \\ &= \frac{10.18}{\sqrt{6.25}\sqrt{18.96}} \\ &= 0.935\,17. \end{aligned}$$

La corrélation lineaire entre les variables x et y est forte.

Chapitre 2

Lois de probabilité

2.1 L'essentiel du cours

On peu schématiquement, classer les expériences en deux grandes groupes : celles d'ont l'issue est prévue avec certitude, dépendants de loi physique bien établie, ce sont les expériences déterminités, et celles dont l'issue dépend du hasard, pour lesquels on ne peut faire de prévisions rigoureuse, ce sont les expériences aléatoires.

On dépose aujourd'hui d'outils mathématiques pour étudier ces événements aléatoires, en construire des modèles simples qui vont permette d'analyser des données nombreuses ou de prévoir des résultats d'expériences prochaines. Les applications recouvrent des domaines très variés, comme l'économie, les sciences sociales, la biologie, les sciences physiques.

Définition 1 Une variable aléatoire X est dite **discrète** si l'ensemble des réalisation possibles $x_1, x_2, ..., x_n$ pour cette variable est fini ou dénombrable.

Définition 2 une variable aléatoire **continue** à pour domaine de variation l'ensemble $\mathbb{R}$ ou un intervalle de l'ensemble $\mathbb{R}$.

2.1.1 Loi de probabilité d'une variable aléatoire discrète

Soit X une variable aléatoire discrète définie sur l'univers Ω. La probabilité de l'événement $(X = x_i)$ est la probabilité de l'union des événements de

$$\Omega = \{w_1, w_2, w_3, ..., w_n\}$$

qui ont pour image x_i.

Si w_1, w_2, w_3 vérifient :

$$X(w_1) = X(w_2) = X(w_3) = x_i,$$

alors

$$P(X = x_i) = p_i = P(w_1 \cup w_2 \cup w_3) = P(w_1) + P(w_2) + P(w_3),$$

La loi de probabilité de X est l'ensemble des couples (x_i, p_i) ; elle peut être présentée dans un tableau.

x_i	x_1	x_2	...	x_i	...	x_n	total
$P(X = x_i)$	p_1	p_2	...	p_i	...	p_n	1

Une loi de probabilité discrète est définie par les couples (x_i, p_i) avec

$$P(X = x_i) = p_i \geq 0 \text{ et } \sum_{i=1}^{n} p_i = 1,$$

2.1.2 Loi de probabilité d'une variable aléatoire continue

Lorsque l'Univers Ω est constitue d'une infinité non dénombrable d'éléments, chaque élément de Ω a une probabilités nulle de se réaliser et de ce fait les réalisations isolée x d'une variable aléatoire X ont une probabilité nulle de se réaliser. Dans ce cas, il est possible de calculer les probabilités qu'une réalisation x de la variable aléatoire X appartienne à un intervalle.

Une variable aléatoire continue prend ses valeurs dans un intervalle réel. La notion d'espace probabilité (Ω, P) peut-être négligée car très sauvent l'étude concrète fait obstraction de ces notions en utilisant directement les lois de probabilités des variables aléatoires.

Définition 3 Une fonction de densité de probabilité sur un intervalle de réels I est une fonction f définie, continue et positive sur I telle que

$$\int_I f(t)dt = 1.$$

c'est-à-dire que l'aire limitée sous la courbe sur l'intervalle I est égale à 1.

Définition 4 On appelle fonction de répartition d'une variable aléatoire à densité X la fonction F définie sur $\mathbb{R}$ par

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

2.1.3 Espérance et variance d'une variable aléatoire

Définition 5

1-Si X est une variable discrète

$$E[X] = \sum_{i=1}^{n} x_i P(X = x_i).$$

2-Si X est une variable à densité f

$$E[X] = \int_{-\infty}^{+\infty} x f(x) dx.$$

Variance et écart type

$$Var(X) = E\left[X^2\right] - E[X]^2.$$

$$\sigma(X) = \sqrt{Var(X)}.$$

2.1.4 Lois discrètes usuelles

a. On dit qu'une variable aléatoire X suit une *loi de Bernoulli* de paramètre $p \in [0,1]$ lorsque X est à valeur dans $\{0,1\}$ et que

$$P(X = 1) = p; P(X = 0) = 1 - p.$$

On a alors

$$E[X] = p \text{ et } Var(X) = p(1-p).$$

b. On dit qu'une variable aléatoire X suit une *loi binomiale* de paramètres $n \in N^*$ et $p \in [0,1]$, que l'on note $B(n,p)$, lorsque X est à valeurs dans $\{0,\ldots,n\}$ avec, pour tout $k \in \{0,\ldots,n\}$,

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

On a alors

$$E[X] = np \text{ et } Var(X) = np(1-p).$$

c. Soit $\lambda > 0$. On dit que X suit une *loi de Poisson* de paramètre λ si elle est à valeurs dans $\mathbb{N}$ et si, pour tout $n \in \mathbb{N}$,

$$P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}, k \in \mathbb{N}.$$

On a alors

$$E[X] = \lambda \text{ et } Var(X) = \lambda.$$

2.1.5 Lois continues usuelles

a. *La loi normale*, loi de Gauss, ou de Laplace-Gauss, est la plus célèbre des lois de probabilité. Son succès, et son omniprésence dans les sciences de la vie, viennent du théorème central limite que nous étudierons plus loin. La loi normale de paramètres $m \in \mathbb{R}$ et $\sigma^2 \in \mathbb{R}^+$ est notée $N(m, \sigma)$. Elle a pour densité :

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

La loi normale de moyenne nulle et d'écart type unitaire, $N(0, 1)$, est appelée *loi normale centrée.*

Soit F la fonction de répartition de loi normal $N(m, \sigma)$, pour calcul $F(a) = P(X \leq a)$, on pose $X = m + \sigma z$ alors

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

où Z suit une loi normal $N(0, 1)$.on a

$$F(a) = P(X \leq a) = \Phi\left(\frac{a-m}{\sigma}\right),$$

est les valeurs de Φ sont donnés par la table de la loi normale centrée.
b. Soient k variables aléatoires $X_1, ..., X_k$ indépendantes suivant la loi normale centrée et réduite. Alors par définition la variable X définie par

$$X = \sum_{i=1}^{n} X_i^2,$$

suit une *loi du* χ^2 à k degrés de liberté. On a alors

$$E[\chi^2] = n \text{ et } V(\chi^2) = 2n.$$

c. Soit X une variable aléatoire de loi normale centrée et réduite et soit Y une variable indépendante de X et distribuée suivant la loi du χ^2 à k degrés de liberté. Par définition, la variable

$$T = \frac{X}{\sqrt{\frac{Y}{X}}},$$

suit une loi de Student à k degrés de liberté, on a alors

$$E[T] = 0, k > 1 \text{ et } V(T) = \frac{k}{k-2}, k > 2.$$

d. *La loi exponentielle* peut être utilisée pour modéliser la radioactivité ou le temps d'attente dans une queue.

Si la variable aléatoire X suit la loi exponentielle de paramètre λ , alors sa densité est :

$$P(X = x) = \begin{cases} \lambda e^{\lambda}, & \text{si } x \geq 0 \\ 0 & \text{si } x < 0, \end{cases}$$

On a alors

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

2.1.6 Échantillonnage

a. Définition générale

La théorie mathématique des probabilités suppose que, pour connaître les événements qui peuvent survenir dans une population donnée, il n'est possible d'étudier ou d'interroger qu'une petite partie de celle-ci, à condition de respecter des règles rigoureuses de sélection de cette fraction de population.

L'échantillonnage est un procédé qui permet de définir un échantillon dans un travail d'enquête. Il s'agit d'étudier une partie sélectionnée pour établir des conclusions applicables à un tout. En d'autres termes, l'échantillonnage est une sélection précise de personnes ciblées pour réaliser un entretien, un focus group, un sondage ou un questionnaire.

b. Les types d'échantillonnage

b.1 L'échantillonnage représentatif : Un échantillon représentatif, souvent utilisé dans une étude quantitative (questionnaire ou sondage), est défini comme représentatif lorsqu'il a les mêmes caractéristiques que la population étudiée (population mère). Un échantillon représentatif peut se faire à travers l'utilisation de la technique des quotas.

b.2 L'échantillonnage aléatoire : L'échantillonnage aléatoire (ou "méthode d'échantillonnage probabiliste") est déterminé à partir d'une procédure de tirage aléatoire statistique. Malgré le hasard, la représentativité de l'échantillon aléatoire est assurée par les lois statistiques de la probabilité.

2.2 Exercices corrigés

Exercice 1

La loi de probabilité d'une variable aléatoire X est donnée par le tableau suivant :

x_i	1	2	3	4	5	6
$P(X = x_i)$	a	$2a$	$3a$	$3a$	$2a$	a

Avec $a \in \mathbb{R}$

1. A quelle(s) condition(s) sur a ce tableau définit bien une loi de probabilité ?

2. Calculer $F(3)$ et $P(x > 4)$, $P(3 \leq x \leq 5)$. (F : la fonction de répartition)

3. Calculer l'espérance et la variance de X.

Solution de l'exercice 1

1. On sait bien que $P(X = x_i)$, et pour $P(X = x_i)$ soit une loi de probabilité il faut qu'elle vérifie

$$P(X = x_i) \geq 0 \text{ et } p_1 + p_2 + ... + p_6 = 1,$$

donc

$$a + 2a + 3a + 3a + 2a + a = 1,$$

$$\Longrightarrow a = \frac{1}{12},$$

2.

$$\begin{aligned} F(3) &= P(X \leq 3) = P(X = 1) + P(X = 2) + P(X = 3) \\ &= \frac{1}{2}, \end{aligned}$$

$$\begin{aligned} P(x &> 4) = P(x = 5) + P(x = 6) \\ &= \frac{1}{4} \end{aligned}$$

$$\begin{aligned} P(3 &\leq x \leq 5) = P(x = 3) + P(x = 4) + P(x = 5) \\ &= \frac{2}{3}, \end{aligned}$$

3.

$$E(x) = \sum_{i=1}^{6} P_i x_i = 3.5,$$

$$\begin{aligned} V(x) &= E(x^2) - E(x)^2 \\ &= \sum_{i=1}^{6} P_i x_i^2 - (3.5)^2 = 1.92. \end{aligned}$$

Exercice 2

Le nombre de noyades accidentelles est en moyenne de 3 par an pour une population de cent mille (100000) habitants. Calculer :

1. La probabilité que 3 noyades seraient enregistrées pour cette population durant l'année suivante.

2. La probabilité que 3 noyades seraient enregistrées pour cette population durant les deux années suivantes.

3. La probabilité qu'aucune noyade ne serait enregistrée durant l'année suivante.

Solution de l'exercice 2

Le nombre de noyade accidentelles X est une variable aléatoire de Poisson de paramètre $\lambda = 3$ pour une période de temps égale à une année.

1.

$$\begin{aligned} P(X &= k) = e^{\lambda} \frac{\lambda^k}{k!} \\ \Longrightarrow\ & P(X = 3) = e^{-3} \frac{3^3}{3!} = 0.224\,04\ (22.404\%) \end{aligned}$$

2. Posons X le nombre de noyades durant la première année et Y le nombre de noyades durant la $2^{ème}$ année, alors on a à calculer $P(X = 0)$ et $p(Y = 3)$ ou bien $P(X = 1)$ et $P(Y = 2)$ ou bien $P(X = 2)$ et $P(Y = 1)$ ou bien $P(X = 3)$ et $P(Y = 0)$. Comme il y a indépendance des noyades alors le calcul final sera :

$$\begin{aligned} &P(X = 0) \times P(Y = 3) + P(X = 1) \times P(Y = 2) \\ +&P(X = 2) \times P(Y = 1) + P(X = 3) \times P(Y = 0) = 0.08924\ (8.924\%). \end{aligned}$$

3. Soit A : noyades enregistrées pour cette population durant les deux années suivantes.

$$\begin{aligned} P(A) &= P[(X=0 \text{ et } Y=3) \cup (X=1 \text{ et } Y=2) \cup (X=2 \text{ et } Y=1) \cup (X=3 \text{ et } Y=0)] \\ &= P[(X=0)P(Y=3) + P(X=1)P(Y=2) \\ +P(X &= 2)P(Y=1) + P(X=3)P(Y=0)] \\ &= 0.892. \end{aligned}$$

Parce que les événements sont indépendants et disjoints (incompatibles)

Exercice 3

Dans une famille la probabilité de naissance d'un enfant gaucher est de $\frac{1}{5}$. On sait que cette famille a 9 enfants.

1. Quelle est la loi de probabilité suivie par la variable aléatoire X : nombre de gauchers

2. Quelle est la probabilité d'avoir exactement 2 gauchers dans cette famille.

3. Quelle est la probabilité d'avoir au moins 2 enfants gauchers.

Solution de l'exercice 3

1. Si la variable aléatoire X est le « nombre de gauchers » alors la loi que suit X est binomiale : $X \mapsto B(9, 1/5)$

2.

$$\begin{aligned} P(X &= 2) = C_9^2 \left(0.2\right)^2 \left(0.8\right)^7 \\ &= 0.30199, \ (30.199\%). \end{aligned}$$

3.

$$\begin{aligned} P(X &\geq 2) = 1 - [P(X=0) + P(X=1)] \\ &= 1 - \left((0.8)^9 + 0.30199\right) \\ &= 0.56.379. \end{aligned}$$

Exercice 4

80 personnes s'apprêtent à passer le portique de sécurité. On suppose que pour chaque personne la probabilité que le portique sonne est égale à 0.02192.

Soit X la variable aléatoire donnant le nombre de personnes faisant sonner le portique, parmi les 80 personnes de ce groupe.

1. Justifier que X suit une loi binomiale dont on précisera les paramètres.

2. Calculer l'espérance de X et interpréter le résultat.

3. Sans le justifier, donner la valeur arrondie à 10^{-3} de :

La probabilité qu'au moins une personne du groupe fasse sonner le portique.

La probabilité qu'au maximum 5 personnes fassent sonner le portique.

Sans le justifier, donner la valeur du plus petit entier n tel que

$$P(X \leq n) \geq 0.90.$$

Solution de l'exercice 4

1. On répète 80 fois la même expérience aléatoire. Toutes les "tirages" sont identiques, indépendants. Chaque expérience possède exactement deux issues : A et $\bar{A}$. De plus

$$P(A) = 0.02192,$$

X suit donc la loi binomiale de paramètres $n = 80$ et $p = 0.02192$

2.

$$\begin{aligned} E(x) &= np \\ &= 1.7536. \end{aligned}$$

Un moyenne environ 1.7 personnes feront sonner le portique.

3. La probabilité qu'au moins une personne du groupe fasse sonner le portique est :

$$\begin{aligned} P(X &\geq 1) = 1 - P(x = 0) \\ &= 1 - (1 - 0.02192)^{80} \\ &\approx 0.83. \end{aligned}$$

La probabilité qu'au maximum 5 personnes fassent sonner le portique est :

$$P(X \leq 5) \approx 0.992,$$

d'après la calculatrice.

En utilisant le mode table de la calculatrice on obtient :

$$P(X \leq 2) \approx 0.744 \text{ et } P(X \leq n) \approx 0.901,$$

Donc 3 est le plus petit entier tel que

$$P(X \leq n) \geq 0.90.$$

Exercice 5

La durée de vie, en années, d'un composant radioactif est une variable aléatoire X qui suit la loi exponentielle de paramètre $\lambda = 0,0005$. Calculer :

a. $P(X < 1500)$

b. $P(1500 < X < 2500)$

c. $P(X > 3000)$

d. Calculer la durée de vie moyenne de l'un de ces composants.

Solution de l'exercice 5

a.

$$\begin{aligned} P(X &< 1500) = P(0 < X < 1500) \\ &= e^{-0.0005\times 0} - e^{-0.0005\times 1500} \\ &= 1 - e^{-0.75} \\ &\approx 0.53 \text{ à } 0.01\text{près} \end{aligned}$$

b.

$$\begin{aligned} P(1500 &< X < 2500) = e^{-0.0005\times 1500} - e^{-0.0005\times 2500} \\ &= e^{-0.75} - e^{-1.25} \\ &\approx 0.19 \text{ à } 0.01 \text{ près} \end{aligned}$$

c.

$$\begin{aligned} P(X &> 3000) = e^{-0.0005\times 3000} \\ &= e^{-1.5} \approx 0.22 \end{aligned}$$

d. La durée de vie moyenne de l'un de ces composants est égale à

$$\begin{aligned} E(x) &= \frac{1}{\lambda} \\ &= \frac{1}{0.0005} \\ &= 2000\ ans. \end{aligned}$$

Remarque :

$$P(X \geq a) = e^{0.0005a} \text{ et } \ P(X < a) = 1 - e^{0.0005a}$$

Exercice 6

On considère la fonction f définie sur $[0;4]$ par $f(x) = \frac{1}{8}x$.

1. Vérifier que f est bien une densité de probabilité sur $[0,4]$.

2. Soit X est une variable aléatoire de densité f.

Déterminer la probabilité : $P(1 \leq X \leq 3)$ et $P(X \geq 2)$.

Solution de l'exercice 6

1. f est densité de probabilité ?

a. $\forall x \in [0,4]$ $f(x)$ est continue

b. $\forall x \in [0,4]$ $f(x) = \frac{1}{8}x \geq 0$

c.

$$\int_{-\infty}^{+\infty} f(x)dx = \int_{0}^{4} \frac{1}{8}xdx = \frac{1}{8}\frac{x^2}{2}\Big]_0^4 = 1.$$

Donc f est bien une densité de probabilité sur $[0,4]$.

2.

$$p(1 \leq X \leq 3) = \int_{1}^{3} \frac{1}{8}xdx = \frac{1}{8}\frac{x^2}{2}\Big]_1^3 = \frac{1}{2},$$

$$p(X \geq 2) = \int_{2}^{4} \frac{1}{8}xdx = \frac{1}{8}\frac{x^2}{2}\Big]_2^4 = \frac{3}{4},$$

Exercice 7

1. Soit Z une variable aléatoire de loi normale centrée réduite $N(0,1)$.

a. Calculer

$$P(Z \leq 1.34), P(Z \leq -1.72), P(1.12 \leq Z \leq 1.57),$$

b. Déterminer z tel que : $P(Z \leq z) = 0.683$;

2. Soit une variable aléatoire X suivant une loi normale $N(5,2)$

Déterminer $P(X \leq 7); P(4 \leq X \leq 7)$.

Solution de l'exercice 7

1. **a.** Z suit une variable aléatoire de loi $N(0,1)$.

$P(Z \leq 1.34)$ est lue dans la table normale centrée réduite $N(0,1)$, on note Φ la fonction de répartition de la loi normale $N(0,1)$, on a

$$P(Z \leq 1.34) = \Phi(1.34) = 0.9099$$

$$\begin{aligned} P(Z &\leq -1.72) = 1 - P(Z \leq 1.72) \\ &= 1 - \Phi(1.72) = 1 - 0.9573 \\ &= 0.0427, \end{aligned}$$

$$\begin{aligned} P(1.12 &\leq Z \leq 1.57) = \Phi(1.57) - \Phi(1.12) \\ &= 0.9418 - 0.8686 \\ &= 0.0732, \end{aligned}$$

b.

$$\begin{aligned} P(Z &\leq z) = 0.683, \\ \Phi(z) &= 0.683, \end{aligned}$$

donc $z = 0.48$.

2- X suit une loi $N(5, 2)$, siot Z suit une loi $N(0, 1)$, on pose

$$X = m + \sigma Z,$$

$$Z = \frac{X - m}{\sigma} \Rightarrow Z = \frac{X - 5}{2},$$

donc

$$\begin{aligned} P(X &\leq 7) = P(\frac{X-5}{2} \leq \frac{7-5}{2}) \\ &= P(Z \leq 1) \\ &= \Phi(1) = 0.8413, \end{aligned}$$

et

$$\begin{aligned} P(4 &\leq X \leq 7) = P(\frac{4-5}{2} \leq \frac{X-5}{2} \leq \frac{7-5}{2}) \\ &= P(-0.5 \leq Z \leq 1) \\ &= \Phi(1) - \Phi(-0.5) \\ &= \Phi(1) - [1 - \Phi(0.5)] \\ &= 0.8413 - (1 - 0.6915) = 0.5328. \end{aligned}$$

Exercice 8

On s'intéresse à la durée de vie d'un échantillon de 100 souris de laboratoire, après injection d'une molécule à la naissance. On observe que la durée de vie de ces souris est distribuée selon une loi normale autour d'une moyenne de 400 jours et avec un écart type de 8 jours.

1. Quel est le nombre attendu de souris ayant une durée de vie entre 390 et 420 jours

2. A quelle durée de vie correspond le 3ème quartile.

3. A quelle durée de vie correspond le 3ème décile.

Solution de l'exercice 8

1.

$$\begin{aligned} P(390 &\leq X \leq 420) = P(\frac{390-400}{8} \leq Z \leq \frac{420-400}{8}) \\ &= P(-1.25 \leq Z \leq 2.5) \\ &= \Phi(2.5) - \Phi(-1.25) \\ &= \Phi(2.5) - (1 - \Phi(1.25)) \\ &= 0.9938 - 1 + 0.8944 = 0.8882. \end{aligned}$$

Le nombre attendu de souris sera de $100 \times 0.8882 = 88.82$ souris ou bien 89 souris dont la vie est comprise entre 390 et 420 jours.

2. On cherche un x qui vérifie $P(X \leq x) = 0.75$

$$\begin{aligned} P(X &\leq x) = P(\frac{X-400}{8} \leq \frac{X-400}{8}) \\ &= P(Z \leq \frac{x-400}{8}) = 0.75, \end{aligned}$$

$$\Rightarrow \Phi(\frac{x-400}{8}) = 0.75.$$

La table $N(0.1)$, on a

$$\frac{x-400}{8} = 0.675,$$

donc $x = 405.4$, (≈ 405 jours)

3. On cherche un x qui vérifie $P(X \leq x) = 0.3$

$$\begin{aligned} P(X &\leq x) = P(\frac{X-400}{8} \leq \frac{x-400}{8}) \\ &= P(Z \leq \frac{x-400}{8}) = 0.30. \end{aligned}$$

Comme $0.3 < 0.5$, cette valeur de 0.3 ne figure pas dans la table N(0.1), on calcule la valeur de $-\frac{x-400}{8}$:

$$P(Z \leq -\frac{x-400}{8}) = 1 - P(Z \leq \frac{x-400}{8}) = 1 - 0.30 = 0.70.$$

La table $N(0.1)$ donne:

$$-\frac{x-400}{8} = 0.525,$$

on a

$$x = 395.8 \quad (\approx 396\ jours).$$

Exercice 9

Un Chercheur a étudié l'âge moyen auquel les premiers mots du vocabulaire apparaissent chez les enfants. Une étude effectuée auprès d'un millier d'enfants montre que les premiers mots apparaissent, en moyenne, à 2 mois et avec un écart type de 1 mois et demi. Sachant que la distribution des âges est normale, on souhaite :

1. Evaluer la proportion d'enfants ayant acquis leurs premiers mots avant 5 mois.

2. Evaluer la proportion d'enfants ayant acquis leurs premiers mots après 6 mois.

3. Evaluer la proportion d'enfants ayant acquis leurs premiers mots entre 3 et 5 mois.

Solution de l'exercice 9

1. Soit Z suit une loi $N(0,1)$

$$\begin{aligned} P(X &< 5) = P(Z < \frac{X-2}{1.5}) \\ &= P(Z < 2) \\ &= \Phi(2) \\ &= 0.9772. \end{aligned}$$

2.

$$\begin{aligned} P(X &> 6) = P(Z > \frac{6-2}{1.5}) \\ &= P(Z > 2.67) \\ &= 1 - P(Z < 2.67) \\ &= 1 - 0.9962 \\ &= 0.0038. \end{aligned}$$

3.

$$\begin{aligned} P(3 &< X < 5) = P(\frac{3-2}{1.5} < Z < \frac{5-2}{1.5}) \\ &= P(0.67 < Z < 2) \\ &= \Phi(2) - \Phi(0.67) \\ &= 0.9772 - 0.7486 \\ &= 0.2286. \end{aligned}$$

Exercice 10

Une enquête est effectuée auprès de familles de 4 personnes afin de connaitre leur achat de lait en 1 mois. Sur l'ensemble des personnes interrogées, la consommation de ce produit forme une population gaussienne avec une moyenne de 25 litres et un écart type de 6 litres.

En vue de concevoir une campagne de publicité, on souhaite connaître le pourcentage des faibles consommateurs (c'est-à-dire moins de 10 L/mois) et le pourcentage des grands consommateurs (c'est-à-dire plus de 30L/mois).

1. Calculer ces deux pourcentages.

2. Au-dessous de quel nombre de litres achetés se trouvent 75% des consommateurs ?

3. Combien de litres au maximum consomme la moitié des consommateurs ?

4. Au-dessus de quelle consommation se trouve 1/3 de la population ? et 2/3 de la population ?

Solution de l'exercice 10

1. Les faibles consommateurs $X < 10$ L et les grands consommateurs $X > 30$ L. On doit calculer les $P(X < 10)$ et $P(X > 30)$, siot Z suit une loi

$N(0,1)$

$$\begin{aligned} P(X &< 10) = P(Z < \frac{10-25}{6}) \\ &= P(Z < -2.5) \\ &= \Phi(-2.5) \\ &= 1 - \Phi(2.5) \\ &= 1 - 0.9938 = 0.0062, \end{aligned}$$

$$\begin{aligned} P(X &> 30) = P(Z > \frac{30-25}{6}) \\ &= P(Z < -0.83) \\ &= \Phi(-0.83) \\ &= 1 - \Phi(0.83) \\ &= 1 - 0.7967 = 0.2033. \end{aligned}$$

2. On cherche x qui vérifie $P(X < x) = 0.75$

$$\begin{aligned} P(X &< x) = P(\frac{X-25}{6} < \frac{x-25}{6}) \\ &= P(Z < \frac{x-25}{6}) \\ &= \Phi(\frac{x-25}{6}) = 0.75. \end{aligned}$$

La table $N(0.1)$ donne

$$\frac{x-25}{6} = 0.675x$$

on a, $x = 29.05$ litres.

3. On cherche x qui vérifie $P(X \leq x) = 0.5$

$$\begin{aligned} P(X &< x) = P(Z < \frac{x-25}{6}) \\ &= \Phi(\frac{x-25}{6}) = 0.5. \end{aligned}$$

La table $N(0.1)$ donne

$$\frac{x-25}{6} = 0,$$

on a, $x = 25$ litres

4. On cherche x qui vérifie $P(X > x) = 0.3333$

$$\begin{aligned} P(X > x) &= P(Z > \frac{x-25}{6}) \\ &= 1 - \Phi(\frac{x-25}{6}) = 0.3333, \end{aligned}$$

donc

$$\Phi(\frac{x-25}{6}) = 0.6667,$$

la table $N(0.1)$ donne $\frac{x-25}{6} = 0.435$, on a $x = 27.6$ litres

5. On cherche x qui vérifie $P(X > x) = 0.6667$

$$\begin{aligned} P(X > x) &= P(Z > \frac{x-25}{6}) \\ &= 1 - P(Z < \frac{x-25}{6}) \\ &= 1 - \Phi(\frac{x-25}{6}) = 0.6667, \end{aligned}$$

donc $\Phi(\frac{x-25}{6}) = 0.3333$, puisque $0.3333 < 0.5$ on cherche la valeur de $-\frac{x-25}{6}$

$$\begin{aligned} \Phi(-\frac{x-25}{6}) &= 1 - \Phi(\frac{x-25}{6}) \\ &= 1 - 0.3333 \\ &= 0.6667, \end{aligned}$$

La table $N(0.1)$ donne

$$-\frac{x-25}{6} = 0.435,$$

on a $x = 22.4$ litres.

Exercice 11

Dans une population masculine le taux de cholestérol X suit une loi normale $N(2, 0.5)$. Les paramètres sont exprimés en $g.l^{-1}$.

1-Calculer la probabilité $P(1.208 \leq X \leq 3.023)$.

2-Trouver les bornes de l'intervalle symétrique autour de la moyenne telles que la probabilité pour que X appartienne à cet intervalle soit égale à 0.80

Solution de l'exercice 11

1. Ici

$$Z = \frac{X-2}{0.5},$$

$$\begin{aligned} P(1.208 \leq X \leq 3.023) &= P(\frac{1.208-2}{0.5} \leq Z \leq \frac{3.023-2}{0.5}) \\ &= P(-1.584 \leq Z \leq 2.046) \\ &= \Phi(2.046) - \Phi(-1.584) \\ &= \Phi(2.046) - (1 - \Phi(1.584)) \\ &= 0.9798 - (1 - 0.9429)) = 0.9227, \end{aligned}$$

2. L'intervalle cherché est de la forme $[2-a, 2+a]$ où le réel positif a doit etre tel que

$$\begin{aligned} P(2-a \leq X \leq 2+a) &= 0.8, \\ \Rightarrow \quad P(\frac{-a}{0.5} \leq X \leq \frac{a}{0.5}) &= 0.8, \\ \Rightarrow \quad \Phi(\frac{a}{0.5}) - \Phi(-\frac{a}{0.5}) &= 0.8, \end{aligned}$$

donc

$$2\Phi(\frac{a}{0.5}) - 1 = 0.8,$$

on a $\Phi(\frac{a}{0.5}) = 0.9$, alors $\frac{a}{0.5} = 1.28$, on obtient $a = 0.64$. L'intervalle recherché est donc $[1.36, 2.64]$.

Exercice 12

La durée de vie d'un certain type d'appareil est modélisée par une variable aléatoire suivant une loi normale de moyenne et d'écart-type inconnus. Les spécifications impliquent que 80% de la production des appareils ait une durée de vie entre 120 et 200 jours et que 5% de la production ait une durée de vie inférieure à 120 jours.

1. Quelles sont les valeurs de m et σ ?

2. Quelle est alors la probabilité d'avoir un appareil dont la durée de vie soit comprise entre 200 jours et 230 jours ?

Solution de l'exercice 12

1. On note X la variable durée de vie. Les spécifications se traduisent par :

$$P(120 \leq X \leq 200) = 0,8 \text{ et } P(X < 120) = 0,05.$$

En notant toujours $Z = \frac{X-m}{\sigma}$, la variable centrée réduite, on obtient

$$\begin{cases} P(\frac{120-m}{\sigma} \leq Z \leq \frac{200-m}{\sigma}) = 0.8 \\ P(Z < \frac{120-m}{\sigma}) = 0,05, \end{cases}$$

$$\begin{aligned} P(\frac{120-m}{\sigma} &\leq Z \leq \frac{200-m}{\sigma}) = 0.8 \\ &\Rightarrow \Phi(\frac{200-m}{\sigma}) - \Phi(\frac{120-m}{\sigma}) = 0.8, \end{aligned}$$

et

$$\begin{aligned} P(Z &< \frac{120-m}{\sigma}) = 0.05 \\ &\Longrightarrow \Phi(\frac{120-m}{\sigma}) = 0.05 \\ &\Longrightarrow \Phi(-\frac{120-m}{\sigma}) = 0.95, \end{aligned}$$

donc

$$\begin{cases} \Phi(-\frac{120-m}{\sigma}) = 0.95 \\ \Phi(\frac{200-m}{\sigma}) = 0.85, \end{cases}$$

la table $N(0.1)$ donne

$$\begin{cases} -\frac{120-m}{\sigma} = 1.65 \\ \frac{200-m}{\sigma} = 1.04, \end{cases}$$

on a $m = 169.07, \sigma = 29.74$, La résolution du système donne

$$m \approx 169, \sigma \approx 30.$$

2.

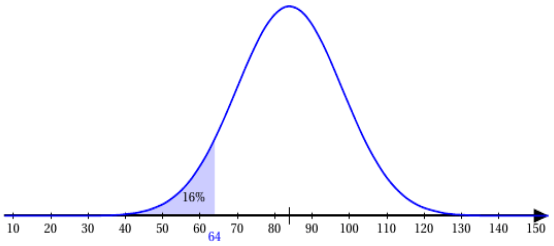
$$\begin{aligned} P(200 &\leq X \leq 230) = P(\frac{200-169}{30} \leq Z \leq \frac{230-169}{30}) \\ &= P(1.03 \leq Z \leq 2.03) \\ &= \Phi(2.03) - \Phi(1.03) \\ &= 0.9788 - 0.8485 \\ &= 0.1303, \end{aligned}$$

La probabilité que la durée de vie de l'appareil soit comprise entre 200 et 230 jours est d'environ 0.130 3.

Exercice 13

Des études statistiques ont permis de modéliser la durée de vie, en mois, d'un type de lave-vaisselle par une variable aléatoire X suivant une loi normale $N(m, \sigma^2)$ de moyenne m et d'écart-type σ De plus, on a $P(X \leq 64) = 0.16$

La représentation graphique de la fonction densité de probabilité de est donnée ci-dessous.



1. **a.** En exploitant le graphique, déterminer $P(64 \leq X \leq 104)$

b. Quelle valeur approchée entière de σ peut-on proposer ?

2. On note Z la variable aléatoire définie par $Z = \frac{X-84}{\sigma}$

a. Quelle est la loi de probabilité suivie par Z ?

b. Justifier que $P(X \leq 64) = P(Z \leq \frac{-20}{\sigma})$

c. En déduire la valeur de , arrondie à 10^{-3}

3. Dans cette question, on considère que $\sigma = 20.1$

Les probabilités demandées seront arrondies à 10^{-3}

a. Calculer la probabilité que la durée de vie du lave-vaisselle soit comprise entre 2 et 5 ans.

b. Calculer la probabilité que le lave-vaisselle ait une durée de vie supérieure à 10 ans.

Solution de l'exercice 13

1.a. par symétrie $P(X \geq 104) = 0.16$ et donc

$$\begin{aligned} P(64 &\leq X \leq 104) \\ &= 1 - 0.16 - 0.16 \\ &= 0.68. \end{aligned}$$

b. On vient donc de trouve que

$$P(m - 20 \leq X \leq m + 20) = 0.68,$$

donc $\sigma = 20$.

2.a. La variable Z est centrée et réduite : elle suit donc une loi normale centrée réduite $N(0.1)$.

b. On part de

$$\begin{aligned} P(X &\leq 64) = P(\frac{X-84}{\sigma} \leq \frac{-20}{\sigma}) \\ &= P(Z \leq \frac{-20}{\sigma}). \end{aligned}$$

c. Le résultat présédent entraine que

$$P(Z \leq \frac{-20}{\sigma}) = \Phi(\frac{-20}{\sigma}) = 0.16,$$

donc

$$1 - \Phi(\frac{20}{\sigma}) = 0.16,$$

on a

$$\Phi(\frac{20}{\sigma}) = 0.84.$$

La table $N(0.1)$ donne $\frac{20}{\sigma} = 0.9945$, alors $\sigma = 20.11$.

3. On Considère que $\sigma = 20.1$

a.

$$\begin{aligned} P(24 &\leq X \leq 60) = P(\frac{24-84}{20.1} \leq Z \leq \frac{60-84}{20.1}) \\ &= P(-2.98 \leq Z \leq -1.19) \\ &= \Phi(-1.19) - \Phi(-2.98) \\ &= 1 - \Phi(1.19) - (1 - \Phi(2.98)) \\ &= -\Phi(1.19) + \Phi(2.98), \end{aligned}$$

La table $N(0.1)$ donne

$$\begin{aligned} P(24 &\leq X \leq 60) = -0.8830 + 0.9986 \\ &= 0.116, \end{aligned}$$

b.

$$\begin{aligned} P(X &> 120) = 1 - P(X \leq 120) \\ &= 1 - P(Z \leq \frac{120-84}{20.1}) \\ &= 1 - \Phi(1.79), \end{aligned}$$

La table $N(0.1)$ donne

$$\begin{aligned} P(X &> 120) = 1 - 0.9633 \\ &\approx 0.037. \end{aligned}$$

Exercice 14

Dans une entreprise qui produit des bobines de fil pour de l'industrie textile, la longueur d'une bobine est une variable aléatoire X où X suit la loi normale $N(50, 0.2)$ de moyenne $50m$ et d'écart-type $0.2\ m$.

Calculer les probabilités suivantes

1. La longueur de la bobine est inférieure ou égala 50.19 m

2. La longueur de la bobine est supérieure 50.16 m

3. Déterminer la nombre réel positife k tel que

$$P(50 - k \leq X \leq 50 + k) = 0.9,$$

interpréter le résultat trouvé.

Solution de l'exercice 14

1. On cherche $P(X \leq 50.19)$

Soit Z suit une loi $N(0.1)$, on pose $Z = \frac{X-m}{\sigma}$

$$\begin{aligned} P(X &\leq 50.19) = P(Z \leq \frac{50.19 - 50}{0.2}) \\ &= P(Z \leq \frac{50.19 - 50}{0.2}) \\ &= \Phi(0.95), \end{aligned}$$

lue dans la table de la loi $N(0.1)$

$$P(X \leq 50.19) = 0.8289.$$

2. On cherche $P(X > 50.16)$

$$\begin{aligned} P(X &> 50.16) = 1 - P(X \leq 50.16) \\ &= 1 - P(Z \leq 0.8) \\ &\quad 1 - \Phi(0.8) \\ &\quad 1 - 0.7881 \\ &\quad 0.2119 \end{aligned}$$

3. On trouve k avec $P(50 - k \leq X \leq 50 + k) = 0.9$

$$\begin{aligned} P(50-k \leq X \leq 50+k) &= P(\frac{-k}{20} \leq Z \leq \frac{k}{20}) \\ &= \Phi(\frac{k}{20}) - \Phi(\frac{-k}{20}) \\ &= \Phi(\frac{k}{20}) - (1 - \Phi(\frac{k}{20})) \\ &= 2\Phi(\frac{k}{20}) - 1 = 0.9, \end{aligned}$$

donc

$$\Phi(\frac{k}{20}) = 0.95,$$

lue dans la table de la loi $N(0.1)$,

$$\frac{k}{20} = 1.65 \Longrightarrow k = 0.33,$$

La probabilité qu'une bobine ait une longueur comprise entre 49.67 et 50.33 est d'énvirons 90%.

Chapitre 3

Estimation

3.1 L'essentiel du cours

3.1.1 Estimation ponctuelle et biais

a. Définition d'un estimateur

En statistique, on appelle estimateur d'un paramètre de la population un paramètre, calculé à partir de l'échantillon, approchant au mieux celui de la population.

b. Définition du biais

Un estimateur est dit *sans biais* si sa valeur est en moyenne égale à la vraie valeur du paramètre de la population, pour une taille d'échantillon fixée. Dans le cas contraire, l'éstimation est dit *biaisé*.

c. Principales estimations

Estimation de la moyenne m de la population :

$$\bar{x} = \hat{m} = \frac{1}{n}\sum_{i=1}^{n} x_i, \text{ (moyenne de l'échantillon)}$$

Estimation de la variance σ^2 de la population :

-Dans le cas m (la moyenne) est connue

$$\hat{S}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - m)^2,$$

-Dans le cas m (la moyenne) est inconnue

$$\begin{aligned}\hat{S}^2 &= \frac{n}{n-1}S^2_{éch} \\ &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2,\end{aligned}$$

si n grand, $\frac{n}{n-1} \approx 1$ et $\hat{S}^2 = S^2_{éch}$ ($S^2_{éch}$ la variance de l'échantillon).

3.1.2 Méthode de maximum de vraisemblance

Soit X une variable aléatoire réelle, de loi D (loi normale $N(m, \sigma)$,loi de Bernoulli,...), de paramètre θ inconnu. On définit une fonction f selon que la loi est discrète ou continue.

Si X est une variable discrète, alors on pose $f(x, \theta) = P_\theta(X = x)$, c'est-à-dire la probabilité que X vaut x.

Si X est une variable continue, alors on pose $f(x, \theta) = f_\theta(x)$, la densité de X au point x.

On appelle **vraisemblance** de θ au vu des observations $(x_1, x_2, ..., x_n)$ d'un n-échantillon indépendamment et identiquement distribué selon la loi D, le nombre :

$$L(x_1, x_2, ..., x_n, \theta) = f(x_1, \theta) \times f(x_2, \theta) \times ... \times f(x_n, \theta)$$

La méthode de maximum de vraisemblance (M.V) consiste à choisir comme estimateur de θ, la valeur particulière de θ qui maximise la fonction de vraisemblance $L(x_1, x_2, ..., x_n, \theta)$.

Cet estimateur T est solution de l'équation

$$\frac{\partial L(x_1, x_2, ..., x_n, \theta)}{\partial \theta} = 0 \text{ ou } \frac{\partial(\ln(L(x_1, x_2, ..., x_n, \theta)))}{\partial \theta} = 0,$$

3.1.3 Estimation par intervalle

Notations :

X, variable quantitative étudiée sur la population.

n, taille de l'échantillon extrait d'une population suivant une loi normale $N(m, \sigma)$.

m et σ, moyenne et variance de X sur la population.

$\bar{x}$ moyenne de X sur l'échantillons.

$\hat{S}_1^2$ variance estimé de X pour la population, à partir d'un échantillon.

a. Intervalle de confiance de la moyennne

a.1 Intervalle de confiance d'une moyenne m avec écart-type σ connu (n>30)

m inconnu (mais estimé par $\bar{x}$) et σ connu, l'intervalle de confiance au niveau $1-\alpha$ de m est donné par :

$$\left[\bar{x} - u\frac{\sigma}{\sqrt{n}}; \bar{x} + u\frac{\sigma}{\sqrt{n}}\right],$$

où la valeur u est lue dans la table normale centrée réduite $N(0,1)$ telle que $\Phi(u) = 1 - \frac{\alpha}{2}$.

a.2 Intervalle de confiance d'une moyenne m avec écart-type σ inconnu (n>30)

m et σ^2 inconnus mais estimés par $\bar{x}$ et $\hat{S}^2$, l'intervalle de confiance au niveau $1-\alpha$ de m est donné par :

$$\left[\bar{x} - u\frac{\hat{S}}{\sqrt{n}}; \bar{x} + u\frac{\hat{S}}{\sqrt{n}}\right],$$

où la valeur u est lue dans la table normale centrée réduite $N(0,1)$ telle que $\Phi(u) = 1 - \frac{\alpha}{2}$.

a.3 Intervalle de confiance d'une moyenne m avec écart-type σ inconnu avec $n < 30$

$$\left[\bar{x} - t_{n-1,\frac{\alpha}{2}}\frac{\hat{S}}{\sqrt{n}}; \bar{x} + t_{n-1,\frac{\alpha}{2}}\frac{\hat{S}}{\sqrt{n}}\right],$$

où $\hat{S}^2$ est un estimateur sans biais de σ^2 et la valeur $t_{n-1,\frac{\alpha}{2}}$ est lue dans la table de Student à $k = n-1$ dégrés de liberté (ddl) et $\gamma = \frac{\alpha}{2}$.

b. Intervalle de confiance de la variance

b.1 Intervalle de confiance pour σ^2 lorsque m est connue

l'intervalle de confiance au niveau $1-\alpha$ de σ^2 est donné par :

$$\left[\frac{1}{\chi_{n-1,\frac{\alpha}{2}}}\sum_{i=1}^{n}(x_i-m)^2;\frac{1}{\chi_{n-1,1-\frac{\alpha}{2}}}\sum_{i=1}^{n}(x_i-m)^2\right],$$

où $\chi_{n-1,\frac{\alpha}{2}}$ et $\chi_{n-1,1-\frac{\alpha}{2}}$ désignent les valeurs du Khi-deux avec $(n-1)$ dégrées de libertés (ddl).

b.2 Intervalle de confiance pour σ^2 lorsque m est inconnue

m et σ^2 inconnus mais estimés par $\bar{x}$ et $\hat{S}^2$, l'intervalle de confiance au niveau $1-\alpha$ de σ^2 est donné par :

$$\left[\frac{(n-1)\,\hat{S}^2}{\chi_{n-1,\frac{\alpha}{2}}};\frac{(n-1)\,\hat{S}^2}{\chi_{n-1,1-\frac{\alpha}{2}}}\right],$$

où $\chi_{n-1,\frac{\alpha}{2}}$ et $\chi_{n-1,1-\frac{\alpha}{2}}$ désignent les valeurs du Khi-deux avec $(n-1)$ dégrées de libertés (ddl).

c. Intervalle de confiance d'une proportion

Soit un échantillon de taille n, T la fréquence observée et p le pourcentage sur la population.

n est supposé grand ($n \geq 30$ peut par exemple, être pris comme limite) et alors la loi binomiale $B(n,p)$ peut être assez bien approximée par la loi normale $N(np, \sqrt{np(1-p)})$.

La fréquence observée T a pour intervalle de confiance :

$$\left[p-u\frac{\sqrt{p(1-p)}}{\sqrt{n}};p+u\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right],$$

Inversement, si l'on ne connaît pas p, son intervalle de confiance sera :

$$\left[T-u\frac{\sqrt{T(1-T)}}{\sqrt{n}};T+u\frac{\sqrt{T(1-T)}}{\sqrt{n}}\right],$$

où la valeur u est lue dans la table normale centrée réduite $N(0,1)$ telle que $\Phi(u)=1-\frac{\alpha}{2}$.

3.2 Exercices corrigés

Exercice 1

Un échantillon de taille 20 a permis d'établir une estimation de la variance d'une variable X pour une population comme étant égale à 25.3.

Quel est l'intervalle de confiance de la variance de X pour cette population au risque de 5% ? On supposera que la variable aléatoire X suit une loi normale.

Solution de l'exercice 1

X : variable aléatoire étudiée.

$X \rightarrow N(.,.)$

$\hat{S}^2 = 25.3$, (estimation ponctuelle de σ^2)

Taille de l'échantillon : $n = 20$.

D'où l'intervalle de confiance de la variance σ^2 :

$$IC(\sigma^2) = \left[\frac{(n-1)\,\hat{S}^2}{\chi_{n-1,\frac{\alpha}{2}}}; \frac{(n-1)\,\hat{S}^2}{\chi_{n-1,1-\frac{\alpha}{2}}} \right],$$

dans la table khi 2 à 19 degrés de liberté

$$\chi_{n-1,\frac{\alpha}{2}} = \chi_{20-1,\frac{0.05}{2}} = \chi_{19,0.025} = 32.85,$$

$$\chi_{n-1,1-\frac{\alpha}{2}} = \chi_{19,1-\frac{0.05}{2}} = \chi_{19,0.975} = 8.91,$$

donc

$$IC(\sigma^2) = \left[\frac{19 \times 25.3}{32.85}, \frac{19 \times 25.3}{8.91} \right],$$

on a

$$IC(\sigma^2) = [14.63; 53.96],$$

Exercice 2

On suppose que la contenance d'une bouteille de lait est une variable aléatoire X qui suit une loi normale d'espérance m et d'écart type σ inconnus. Pour les estimer, on prélève un échantillon de taille $n = 25$. Soient $x_1, ..., x_{25}$ les contenances respectives de 25 bouteilles prélevées au hasard dans la production. On obtient :

$$\sum_{i=1}^{25} x_i = 24.75 \text{ litres et } \sum_{i=1}^{25} (x_i - \bar{x})^2 = 0.003456 \text{ litres} \times \text{litres}$$

1. Déterminer les estimations ponctuelles de la moyenne et l'écart-type de cette distribution ?

2. Déterminer l'intervalle de confiance de la moyenne au niveau 95%

3. Déterminer l'intervalle de confiance de la variance au niveau 95%

Solution de l'exercice 2

1. Estimation ponctuelle de m

$$\begin{aligned} \hat{m} &= \bar{x} = \frac{1}{25}\sum_{i=1}^{25} x_i \\ &= \frac{24.75}{25} \\ &= 0.99 \text{ litre.} \end{aligned}$$

Estimation ponctuelle de σ^2

$$\begin{aligned} \hat{S}^2 &= \frac{1}{25-1}\sum_{i=1}^{25}(x_i - \bar{x})^2 \\ &= \frac{0.003456}{24} \\ &= 1.44 \times 10^{-4}, \end{aligned}$$

$\hat{S} = 0.012$ soit 1.2 centilitres ou cm^3.

2. Intervalle de confiance pour la moyenne ($\alpha = 0.05$)

$$IC(m) = \left[\bar{x} - t_{25-1,\frac{0.05}{2}}\frac{\hat{S}}{\sqrt{n}}; \bar{x} + t_{25-1,\frac{0.05}{2}}\frac{\hat{S}}{\sqrt{n}}\right],$$

Dans la table de student on a

$$t_{25-1,\frac{0.05}{2}} = t_{24,0.025} = 2.064,$$

alors

$$IC(m) = \left[0.99 - 2.064\frac{0.012}{\sqrt{25}}, 0.99 + 2.064\frac{0.012}{\sqrt{25}}\right],$$

on a

$$IC(m) = [0.985\,, 0.995]\,.$$

Le volume moyen est inconnu, mais il a 95% de chances d'être compris entre 0.985 et 0.995 litres.

Exercice 3

L'indice de masse corporelle est calculé en divisant le poids d'une personne par le carré de sa taille, et est utilisé comme mesure du seuil dans laquelle la personne n'est pas en surpoids. Supposons que la distribution de l'indice de masse corporelle pour les hommes ait un écart-type de $\sigma = 3kg/m^2$, et nous souhaitons estimer la moyenne m en utilisant un échantillon de taille $n = 49$.

Trouvez la probabilité que l'erreur absolue dans cette estimation ne dépasserait pas $1kg/m^2$.

Solution de l'exercice 3

C'est un grand échantillon de taille $n = 49 > 30$, donc c'est la loi normale qui sera utilisée.

L'intervalle de confiance de la moyenne m au risque $\alpha = 5\%$ et $\Phi(u) = 1 - \frac{\alpha}{2}$ où Φ la fonction de répartition de la loi $N(0.1)$, est donné par la formule :

$$[x - u\frac{\sigma}{\sqrt{n}}, x + u\frac{\sigma}{\sqrt{n}}],$$

avec $\sigma = 3kg/m^2$ et $n = 49$. l'erreur absolue dans cette estimation ne dépasserait pas $1kg/m^2$,c-à-d

$$u\frac{\sigma}{\sqrt{n}} < 1,$$

donc

$$\begin{aligned} u &= \frac{\sqrt{n}}{\sigma} \\ &= \frac{\sqrt{49}}{3} = 2.33, \end{aligned}$$

on trouve le risque α avec $\Phi(2.33) = 1 - \frac{\alpha}{2}$

$$\begin{aligned} \Phi(2.33) &= 0.9901(\text{dans la table N(0.1)}) \\ &\Rightarrow 1 - \frac{\alpha}{2} = 0.9901, \end{aligned}$$

on a $\alpha \approx 0.02$

La probabilité que l'erreur absolue ne dépasserait pas $1kg/m^2$ est

$$1 - \alpha = 0.98.$$

Exercice 4

Les blessures auto déclarées parmi les gauchers et les droitiers ont été comparées dans une enquête auprès de 1900 étudiants. 90 étudiants parmi les 180 gauchers, et 645 étudiants parmi les 1720 droitiers ont signalé au moins une blessure au cours de la même période.

Calculez l'intervalle de confiance à 95% pour la proportion d'élèves ayant subi au moins une blessure, pour chacune des deux sous-populations d'élèves gauchers et droitiers.

Solution de l'exercice 4

· Les deux échantillons sont grands et donc c'est la loi normale qu'on utilisera. La sous-population des gauchers : la fréquence auto déclarée chez les gauchers est estimée ponctuellement par $T = \frac{90}{180} = 0.5$ et estimée par intervalle de confiance 95% .

$$IC(p) = [T - u\frac{\sqrt{T(1-T)}}{\sqrt{n}}, T + u\frac{\sqrt{T(1-T)}}{\sqrt{n}}],$$

donc

$$IC(p) = [0.5 - 1.96\frac{\sqrt{0.5(1-0.5)}}{\sqrt{180}}, 0.5 + 1.96\frac{\sqrt{0.5(1-0.5)}}{\sqrt{180}}],$$

on a

$$IC(p) = [0.427, 0.573\,].$$

· La sous-population des droitiers : la fréquence auto déclarée chez les droitiers est estimée ponctuellement par $T = \frac{645}{1720} = 0.375$ et estimée par intervalle de confiance.

$$IC(p) = [0.375 - 1.96\frac{\sqrt{0.375(1-0.375)}}{\sqrt{1720}}, 0.375 + 1.96\frac{\sqrt{0.375(1-0.375)}}{\sqrt{1720}}],$$

on a

$$IC(p) = [0.352, 0.398\,],$$

Exercice 5

Sur un échantillon de 50 individus, on a mesuré le taux d'urée sanguine et obtenu les résultats suivants : le taux d'urée moyen est de 0.30 g/L et l'écart-type est 0.05g/L.

1. déterminer l'estimation de la moyenne, de la variance et de l'écart-type du taux d'urée dans la population d'où est extrait l'échantillon.

2. Déterminer l'estimation de ces mêmes paramètres par intervalle de confiance au risque de 5%.

Solution de l'exercice 5

X : taux d'urée sanguine (variable quantitative.)

Un échantillon de taille 50 (≥ 30) avec : $\hat{m} = 0.30g/L, S_{éch} = 0.05g/L$.

1. La moyenne du taux d'urée, m, dans la population est estimée par $\hat{m}$ (notée aussi $\bar{x}$).

Donc, m est estimée par

$$\bar{x} = 0.30g/L,$$

La variance σ^2, du taux d'urée dans la population est estimée par $\hat{S}^2$

$$\begin{aligned} \hat{S}^2 &= \frac{n}{n-1} S_{éch}^2 \\ &= \frac{50}{49} (0.05)^2 \\ &= 0.00255102. \end{aligned}$$

Donc, σ^2 est estimée par

$$\hat{S}^2 \approx 0.0026 \ (g/L)^2$$

L'écart-type σ, du taux d'urée dans la population est estimée par $\hat{S}$.

$$\hat{S} = \sqrt{0.00255102} = 5.\,050\,8 \times 10^{-2},$$

Donc, σ est estimée par

$$\hat{S} \approx 0.0505 \ g/L.$$

2. Estimation par intervalle de confiance, au risque de 5%, du taux d'urée moyen dans la population avec m et σ inconnus ($n \geq 30$). La loi normale sera utilisée.

La formule donnant

$$IC(m) = [\bar{x} - u\frac{\sigma}{\sqrt{n}}, \bar{x} + u\frac{\sigma}{\sqrt{n}}],$$

et

$$\Phi(u) = 1 - \frac{\alpha}{2} = 1 - \frac{0.05}{2} = 0.975,$$

donc $u = 1.96$ (La table $N(0.1)$), alors l'intervalle de confiance demandé sera

$$IC(m) = [0.30 - 1.96\frac{0.0505}{\sqrt{50}}, 0.30 + 1.96\frac{0.0505}{\sqrt{50}}],$$

on a

$$IC(m) = [0.286\,, 0.314] \text{ (en } g/L).$$

Estimation par intervalle de cofiance au risque de 5% de σ^2 :

$$IC(\sigma^2) = \left[\frac{(n-1)\,\hat{S}^2}{\chi_{n-1,\frac{\alpha}{2}}}; \frac{(n-1)\,\hat{S}^2}{\chi_{n-1,1-\frac{\alpha}{2}}}\right],$$

dans la table de khi 2 à $n - 1 = 49$ degrés de liberté

$$\chi_{n-1,\frac{\alpha}{2}} = \chi_{50-1,\frac{0.05}{2}} = \chi_{49,0.025} \approx 71.42,$$

$$\chi_{n-1,1-\frac{\alpha}{2}} = \chi_{49,1-\frac{0.05}{2}} = \chi_{49,0.975} \approx 32.36,$$

donc

$$IC(\sigma^2) = \left[\frac{49 \times 0.0026}{71.42}, \frac{49 \times 0.0026}{32.36}\right],$$

on a

$$IC(\sigma^2) = [1.\,783\,8 \times 10^{-3}; 3.\,937\,0 \times 10^{-3}],$$

Exercice 6

Une enquête concernant la santé, et portant sur 3000 adolescents d'un certain pays européen de 12 à 20 ans, a dénombré 570 adolescents ayant pris un psychotrope au cours des 12 mois précédant l'enquête. Parmi les 1400 filles, 378 ont pris un psychotrope.

1. Donner une estimation ponctuelle de la fréquence de consommation de psychotropes chez les adolescents de ce pays.

2. Estimer par intervalle de confiance à 95% la fréquence de consommation de psychotropes chez les adolescents de ce pays.

3. Même questions chez les filles, puis les garçons.

4. Quelle devrait être la taille de l'échantillon pour que la marge d'erreur à 95% dans l'estimation de la fréquence de consommation de psychotropes chez les adolescents de ce pays soit inférieure à 1%, en supposant que la fréquence observée de consommation de psychotropes n'est pas modifiée.

Solution de l'exercice 6

1. $F = \{$Adolescents de 12 à 20 ans$\}$.

La variable X = consommation de psychotropes.

Cette variable X est qualitative dichotomique (oui et non).

p = proportion de consommateurs de psychotropes dans $\mathcal{F}$

= fréquence de consommation de psychotropes dans $\mathcal{F}$, p inconnue dans la population $\mathcal{F}$.

Grand échantillon de taille n ($= 3000 > 30$) de X issu de $\mathcal{F}$. La loi normale sera utilisée.

L'estimation ponctuelle de p est donnée par la fréquence observée

$$T = \frac{570}{3000} = 19\%.$$

Donc, La proportion d'adolescents de 12 à 20 ans consommateurs de psychotropes est estimée à 19%.

2. L'estimation par intervalle de confiance à 95% (au risque 5%) de p dans $\mathcal{F}$ est donnée par la formule du cours comme

$$IC(p) = [0.19 - 1.96\frac{\sqrt{0.19(1-0.19)}}{\sqrt{3000}}, 0.19 + 1.96\frac{\sqrt{0.19(1-0.19)}}{\sqrt{3000}}],$$

on a

$$IC(p) = [0.176, 0.204\,],$$

où $u = 1.96$ ($\Phi(u) = 1 - \frac{\alpha}{2}$) est le quantile d'ordre 0.975 de la loi $N(0,1)$

Cette approximation est justifiée puisque toutes les conditions d'application de l'estimation sont satisfaites.

3. Filles : $\mathcal{F}_1 = \{$Adolescentes de 12 à 20 ans, de sexe féminin$\}$.

La variable X = consommation de psychotropes.

Cette variable X est qualitative dichotomique (oui et non).

p_1 = proportion de consommatrices de psychotropes dans $\mathcal{F}_1$

=fréquence de consommation de psychotropes dans $\mathcal{F}_1$

p_1 inconnue dans la population $\mathcal{F}_1$.

Echantillon de taille $n_1 = 1400$ de X issu de $\mathcal{F}_1$. C'est un grand échantillon et la loi normale utilisée.

L'estimation ponctuelle de p_1 est donnée par la fréquence observée

$$T_1 = \frac{378}{1400} = 0.27 = 27\%,$$

La proportion d'adolescentes de 12 à 20 ans consommatrices de psychotropes est estimée à 27%.

L'estimation par intervalle de confiance à 95% (au risque 5%) de p_1 dans F_1 s'écrit :

$$IC(p_1) = [0.27 - 1.96\frac{\sqrt{0.27(1-0.27)}}{\sqrt{1400}}, 0.27 + 1.96\frac{\sqrt{0.27(1-0.27)}}{\sqrt{1400}}],$$

on a

$$IC(p_1) = [0.247, 0.293\,],$$

Garçons : $F_2 = \{$Adolescents de 12 à 20 ans, de sexe masculin$\}$.

La variable X = consommation de psychotropes.

Cette variable X est qualitative dichotomique (oui et non).

p_2 = proportion de consommateurs de psychotropes dans F_2 = fréquence de consommation de psychotropes dans F_2, p_2 inconnue dans la population F_2.

Grand échantillon de taille $n_2 = 3000 - 1400 = 1600$ de X issu de F_2

L'estimation ponctuelle de p_2 est donnée par la fréquence observée

$$T_2 = \frac{570 - 378}{1600} = 0.125 = 12.5\%.$$

Donc, La proportion d'adolescents de 12 à 20 ans consommateurs de psycho est estimée à 12.5%.

L'estimation par intervalle de confiance à 95% (au risque 5%) de p_2 dans F_2 s'écrit :

$$IC(p_1) = [0.125 - 1.96\frac{\sqrt{0.125(1-0.125)}}{\sqrt{1600}}, 0.125 + 1.96\frac{\sqrt{0.125(1-0.125)}}{\sqrt{1600}}],$$

on a

$$IC(p_1) = [0.109, 0.141\,]$$

4. La précision de l'intervalle de confiance à 95% est donnée par sa demi-longueur. Pour une taille d'échantillon $n = 3000$, la demi-longueur de l'intervalle $IC(p)$ est de 1.4% (voir question précédente) ; pour obtenir une demi-longueur plus faible, de 1%, il faudra donc plus de 3000 sujets. Pour n inconnu, $T = 0.19$ et $\alpha = 5\%$ connus, la demi-longueur de l'intervalle $IC(p)$ s'écrit comme suit

$$u\frac{\sqrt{T(1-T)}}{\sqrt{n}} = 1.96\frac{\sqrt{0.19(1-0.19)}}{\sqrt{n}}.$$

On cherche n tel que

$$1.96\frac{\sqrt{0.19(1-0.19)}}{\sqrt{n}} \leq 0.01,$$

c'est-à-dire

$$1.96\frac{\sqrt{0.19(1-0.19)}}{0.0001} = 76.89097736 \leq \sqrt{n},$$

d'où $n \geq 5912.2224$, on choisira donc une taille d'échantillon au moins égale à 5913 pour que la demi-longueur de l'intervalle de confiance à 95% soit inférieure à 1%.

Exercice 7

Afin de contrôler le poids (en grammes) dans trois lots de comprimés, différents dispositifs sont étudiés en fonction de la marque (A,B et C) de la machine utilisée. Pour chaque machine, un échantillon de taille 45 a été prélevé et les moyennes des trois échantillons étaient respectivement :

$$\bar{x}_A = 0.805\ g, \bar{x}_B = 0.81g \text{ et } \bar{x}_C = 0.80g,$$

L'écart-type du dernier échantillon vaut : $\hat{S}_C = 0.015g$.

On supposera que le poids des comprimés suit une loi normale.

1. La machine A est garantie par le constructeur comme faisant des comprimés de poids moyen 0.8 g avec un écart-type de 0.02 g. Est-ce exact ?

2. Pour la machine B, seul un écart-type de 0.01 g est garanti par le constructeur. En supposant ce résultat vrai, donne l'intervalle de confiance au risque de 5% du poids moyen des comprimés produits par la machine B.

3. Pour la machine C encor à l'étude, aucune n'est donnée. Donner l'intervalle de confiance au risque de 5% du poids moyen des comprimés produits par la machine B.

Solution de l'exercice 7

X : poids du comprimé

Un échantillon de taille 45 pour les trois machines : $n_A = n_B = n_C = 45$.

1. La machine A : m_A et σ_A sont connus, l'intervalle de confiance de la moyenne sur échantillon vaut

$$\left[m_A - u\frac{\sigma_A}{\sqrt{n}}, m_A + u\frac{\sigma_A}{\sqrt{n}}\right],$$

où $\Phi(u) = 1 - \frac{0.05}{2} = 0.975$ Dans la table de $N(0.1)$ on a

$$u = 1.96,$$

alors I de C est

$$\left[0.80 - 1.96\frac{0.02}{\sqrt{32}}, 0.80 + 1.96\frac{0.02}{\sqrt{32}}\right],$$

donc I de C :

$$[0.794\,, 0.807]\,,$$

La moyenne observée, $\bar{x}_A = 0.805g$, appartient à l'intervalle de pari : avec un risque d'erreur de 5%, on croit à la garantie du constructeur.

2. La machine B : $\sigma_B = 0.01$ est connu et supposé être exact. L'intervalle de confiance, au risque de 5%, du poids moyen des comprimés produits par la machine B est

$$\begin{aligned} & \left[\bar{x}_B - u\frac{\sigma_B}{\sqrt{n}}, \bar{x}_B + u\frac{\sigma_B}{\sqrt{n}}\right] \\ = & \left[0.81 - 1.96\frac{0.01}{\sqrt{32}}, 0.81 + 1.96\frac{0.01}{\sqrt{32}}\right] \\ = & [0.807\,, 0.813] \text{ (en g)} \end{aligned}$$

3. La machine C : m_C et $\sigma_C$sont inconnus. L'intervalle de confiance, au risque de 5%, du poids moyen des comprimés produits par la machine C est :

$$\begin{aligned} & \left[m_c - u\frac{\sigma_c}{\sqrt{n}}, m_c + u\frac{\sigma_c}{\sqrt{n}}\right] \\ = & \left[0.80 - 1.96\frac{0.015}{\sqrt{32}}, 0.80 + 1.96\frac{0.015}{\sqrt{32}}\right] \\ = & [0.795\,, 0.805] \text{ (en g)} \end{aligned}$$

Remarque générale : l'intervalle de confiance ne devant jamais être raccourci, l'arrondi de la borne inférieur de l'intervalle se fait par défaut et celui de la borne supérieure par excès.

Exercice 8

Dans une population, la proportion d'individus immunisés contre un virus est de 0.45. En prélevant un échantillon de taille n (> 30) dans cette population, déterminer dans quel intervalle de confiance au risque α est susceptible de se trouver la population f d'individus immunisés dans l'échantillon.

Application numériques :

$$n = 50, n = 200, \alpha = 0.05, \alpha = 0.20.$$

Solution de l'exercice 8

X : variable aléatoire qualitative à deux modalités (immunisé- non immunisé).

Un échantillon avec $n \geq 30$ et T=proportion observée d'individus immunisés dans l'échantillon.

D'où, l'intervalle de confiance (ou de pari) de p :

$$\left[T - u\sqrt{\frac{T(1-T)}{n}}, T + u\sqrt{\frac{T(1-T)}{n}}\right] \text{ au risque } \alpha \text{ (car n} \geq 30).$$

où u est lu dans la table de la loi normale N(0.1) avec $\Phi(u) = 1 - \frac{\alpha}{2}$.

T	n	$\sqrt{\frac{T(1-T)}{n}}$	α	u	$T - u\sqrt{\frac{T(1-T)}{n}}$	$T + u\sqrt{\frac{T(1-T)}{n}}$
0.45	50	0.07035624	0.05	1.96	0.31210178	0.58789822
0.45	200	0.03517812	0.05	1.96	0.38105089	0.51894911
0.45	50	0.07035624	0.20	1.28	0.35983145	0.54016855
0.45	200	0.03517812	0.20	1.28	0.40491572	0.49508428

Exercice 9

Un échantillon de 80 stimulateurs cardiaques étudié a donné les résultats suivants :

La moyenne est : $\bar{x} = 0.31$ et l'écart-type est $\sigma = 0.015$

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

Solution de l'exercice 9

1. On est en présence d'un grand échantillon $n = 80 > 30$. La loi normale sera utilisée.

La formule donnant

$$IC(m) = [x - u\frac{\sigma}{\sqrt{n}}, \bar{x} + u\frac{\sigma}{\sqrt{n}}],$$

et

$$\Phi(u) = 1 - \frac{\alpha}{2} = 1 - \frac{0.05}{2} = 0.975,$$

donc $u = 1.96$ (La table $N(0.1)$), alors l'intervalle de confiance demandé sera

$$IC(m) = [0.31 - 1.96\frac{0.015}{\sqrt{80}}, 0.31 + 1.96\frac{0.015}{\sqrt{80}}],$$

on a

$$IC(m) = [0.307, 0.313].$$

2. L'erreur absolue $u\frac{\sigma}{\sqrt{n}}$ doit être $< 0.001 \Longrightarrow 1.96\frac{0.015}{\sqrt{n}} < 0.001 \Longrightarrow n > 29.4^2 = 864.36$.

On prend un échantillon de $n = 865$ stimulateurs pour avoir 1 erreur absolue qui ne dépasserait pas 0.001.

Exercice 10

Le poids à la naissance, obtenu à partir des accouchements sur une longue période de temps dans un certain hôpital, montre une moyenne $m = 3175g$ et un écart type $\sigma = 584g$.

Calculer la probabilité que le poids moyen m à la naissance d'un échantillon de 35 nourrissons se situe entre 3033 et $3317g$.

Solution de l'exercice 10

L'intervalle de pari (ou de fluctuation) pour la moyenne $\bar{x}$ de l'échantillon de taille $n = 35$ est donné par hypothèse

$$TC(m) = [3033; 3317].$$

On doit calculer le niveau de confiance $1 - \alpha$. Le centre de cet intervalle est son milieu $\frac{3033+3317}{2} = 3175$. Cet intervalle de pari est centré sur la moyenne $m = 3175$. L'erreur absolue donnée par la formule $u\frac{\sigma}{\sqrt{n}}$ est égale à la demilongueur de l'intervalle de pari

$$\begin{aligned} u\frac{\sigma}{\sqrt{n}} &= \frac{3317 - 3033}{2} \\ &= 142 \\ &\Longrightarrow u = 1.44, \end{aligned}$$

la lecture de la table $N(0.1)$ pour

$$\Phi(1.44) = 1 - \frac{\alpha}{2},$$

on a

$$1 - \frac{\alpha}{2} = 0.9251,$$

donc $\alpha = 0.1498 \approx 0.15 = 15\%$. Par conséquent le niveau de confiance est

$$1 - \alpha = 85\%.$$

Exercice 11

Déterminer l'estimation du maximum de vraisemblance pour une variable de Bernoulli.

Solution de l'exercice 11

$$L(X) = B(1, p),$$

d'où

$$P(X = x) = p^x(1-p)^{1-x}, x = 0 \text{ ou } x = 1,$$

$$\begin{aligned} L(x, x, ..., x, p) &= \Pi_i p^{x_i}(1-p)^{1-x_i} \\ &= p^{\sum x_i}(1-p)^{1-\sum x_i} \end{aligned}$$

alors

$$\ln L = \left(\sum x_i\right) \ln p + (n - \sum x_i) \ln(1-p)$$

La solution du maximum de vraisemblance conduit à $\hat{p}$

$$(\ln L)'_p = \left(\sum x_i\right) \frac{1}{p} + (n - \sum x_i) \frac{1}{1-p} = 0,$$

donc

$$\hat{p} = \frac{\sum x_i}{n} = \bar{x}.$$

Exercice 12

Cent patients ont reçu un nouveau traitement contre la migraine. A la fin du traitement, le médecin leur a demandé s'ils avaient perçu une amélioration de leur état migraineux. La proportion de réponses favorables était de 57%. Quel est l'intervalle de confiance, au risque de 5%, du pourcentage de patients satisfaits par ce nouveau traitement ?

Solution de l'exercice 12

X : variable aléatoire qualitative à deux modalités (amélioration - non amélioration)

Un échantillon avec :

$n = 100$ (> 30) et $T = 0.57$ (proportion de réponses favorables).

D'où, l'intervalle de confiance de la proportion p de réponses favorables dans la population de migraineux est :

$$IC(p) = \left[T - 1.96\sqrt{\frac{T(1-T)}{n}}, T + 1.96\sqrt{\frac{T(1-T)}{n}}\right] \text{ au risque 5\% (car } n \geq 30).$$

donc

$$IC(p) = \left[0.57 - 1.96\sqrt{\frac{0.57 \times 0.43}{100}}, 0.57 + 1.96\sqrt{\frac{0.57 \times 0.43}{100}}\right] \text{ au risque } 5\% \text{ (car } n \geq 30).$$

on a

$$IC(p) = [0.473, 0.667]\,.$$

L'intervalle de confiance ne devant jamais etre raccourci, l'arrondi de la borne inférieure de l'intervalle se fait par défaut et celui de la borne supérieure par excès $[0.473, 0.667]$.

Dans la population des migraineux, il y a 95 chances sur 100 que le pourcentage de patients satisfaits soit compris entre 47.3% et 66.7%.

Exercice 13

On veut estimer par sondage la proportion de ménages possédant un ordinateur en Algérie. Sur 200 ménages interrogés, 80 déclarent en posséder un.

Calculer la proportion de ménage possédant un ordinateur dans cet échantillon. Donnez une estimation ponctuelle et par intervalle de confiance au risque 5%, de cette proportion pour toute la population algérienne.

Solution de l'exercice 13

$T = \frac{80}{200} = 0.4$, proportion de ménages dans l'échantillon.

Soit p la proportion de ménages possédant un ordinateur sur toute la population algérienne. f est un estimation ponctuelle de p.

Ici la loi de T peut être à proximitée par une loi normale.

On peut construire un intervalle de confiance

$$\left[T - u\frac{\sqrt{T(1-T)}}{\sqrt{n}}, f - u\frac{\sqrt{T(1-T)}}{\sqrt{n}}\right],$$

pour préciser l'estimation de p, ici $u = 1.96$ donc l'intervalle de confiance au risque 5% est :

$$[0.3321, 0.4679]\,.$$

Exercice 14

Des essais en laboratoire sur 20 lampes miniatures donnent les durées de vie suivantes, en heures :

451	*412*	*412*	*375*	*407*	*454*	*375*	*393*	*355*	*364*
414	*413*	*345*	*432*	*392*	*329*	*439*	*381*	*451*	*413*

On suppose la durée de vie distribuée normalement.

1. Donner l'estimation ponctuelle de la durée de vie moyenne et de sa variance pour l'ensemble de la production.

2. Estimer par un intervalle ayant un niveau de confiance de 95% la durée de vie moyenne.

3. Estimer par un intervalle ayant un niveau de confiance de 95% de la variance.

Solution de l'exercice 14

1. On trouve $\bar{x} = 400.35$, $\hat{S} = 36.01$ et $\hat{S}^2 = 1297$.

2. *Lorsque* σ^2 *est inconnu et* $n = 20 < 30$, *un intervalle de confiance au niveau* 0.95 *de* m *est*

$$IC(m) = \left[\bar{x} - t_{n-1,\frac{0.05}{2}} \frac{\hat{S}}{\sqrt{n}}; \bar{x} + t_{n-1,\frac{0.05}{2}} \frac{\hat{S}}{\sqrt{n}}\right],$$

dans la table de student on a

$$t_{19,\frac{0.05}{2}} = t_{19,0.025} = 2.093,$$

alors

$$IC(m) = \left[400.35 - 2.093\frac{36.01}{\sqrt{20}}; 400.35 + 2.093\frac{36.021}{\sqrt{20}}\right],$$

donc

$$IC(m) = [383.\,50, 417.\,21]\,.$$

3. Estimation de l'écart-type par intervalle de confiance avec $\alpha = 0.05$

$$IC(\sigma^2) = \left[\frac{(n-1)\,\hat{S}^2}{\chi_{n-1,\frac{\alpha}{2}}}; \frac{(n-1)\,\hat{S}^2}{\chi_{n-1,1-\frac{\alpha}{2}}}\right],$$

dans la table khi 2 à 19 degrés de liberté

$$\chi_{n-1,\frac{\alpha}{2}} = \chi_{20-1,\frac{0.05}{2}} = \chi_{19,0.025} = 32.8523,$$

$$\chi_{n-1,1-\frac{\alpha}{2}} = \chi_{19,1-\frac{0.05}{2}} = \chi_{19,0.975} = 8.9065,$$

donc

$$IC(\sigma^2) = \left[\frac{19 \times 1297}{32.8523}, \frac{19 \times 1297}{8.9065}\right],$$

on a

$$IC(\sigma^2) = [750.11, 2766.86],$$

et donc l'écart-type a 95% de chances de vérifier

$$IC(\sigma) = [27.39, 52.6].$$

Exercice 15

En vue de réaliser un programme de rééducation, des chercheurs ont soumis un questionnaire de neuropsychologie cognitive à 150 enfants dyslexiques tirés au sort. Le questionnaire comporte 20 questions et les chercheurs ont recueilli pour chaque enfant dyslexique le nombre xi de bonnes réponses. Les résultats ainsi récoltés sont tels que : X

$$\sum_{i=1}^{150} x_i = 1502, \sum_{i=1}^{150} x_i^2 = 19486,$$

1. Identifier la population, la variable, son type et son/ses paramètre(s).

2. Donner une estimation ponctuelle du nombre moyen de bonnes réponses dans la population étudiée.

3. Donner une estimation ponctuelle de l'écart-type de la variable.

4. Estimer le nombre moyen de bonnes réponses dans la population par un intervalle de confiance au niveau 99%.

5. Quelle est la marge d'erreur dans l'estimation du nombre moyen de bonnes réponses au niveau 99% ?

Solution de l'exercice 15

1. Population P : f Enfants dyslexiques g.

Variable quantitative X = "Nombre de bonnes réponses au questionnaire"

2 paramètres inconnus : moyenne et écart-type .

2. On estime la moyenne par la moyenne observée

$$\bar{x} = \frac{1502}{150} = 10.01.$$

3. Commençons par calculer la variance observée

$$\begin{aligned} S^2 &= \frac{1}{n}\sum_{i=1}^{150} x_i^2 - \bar{x}^2 \\ &= \frac{1}{150}19486 - (10.01)^2 \\ &= 29.7. \end{aligned}$$

La variance corrigée vaut donc

$$\begin{aligned}\hat{S}^2 &= \frac{n}{n-1}S^2 \\ &= \frac{150 \times 29.7}{149} \\ & 29.9.\end{aligned}$$

Finalement, on estime l'écart-type par l'écart-type corrigé $\hat{S} = \sqrt{29.9} = 5.47$

4. Puisque $n = 150 > 30$, l'estimation par intervalle à 99% ($\alpha = 0.01$) est donnée par

$$IC(m) = \left[\bar{x} - u\frac{\hat{S}}{\sqrt{n}}, \bar{x} + u\frac{\hat{S}}{\sqrt{n}}\right],$$

où $\Phi(u) = 1 - \frac{0.01}{2} = 0.995$ Dans la table de $N(0.1)$ on a

$$u = 2.57,$$

alors

$$IC(m) = \left[10.01 - 2.57\frac{5.47}{\sqrt{150}}; 10.01 + 2.57\frac{5.47}{\sqrt{150}}\right],$$

donc

$$IC(m) = [8.86, 11.16],$$

5. La marge d'erreur est la demi-longueur de l'intervalle obtenu à la question précedente, elle vaut donc 1.15.

Exercice 16

Une usine fabrique des câbles. La masse maximale en tonnes supportée par un câble est une variable aléatoire réelle X suivant la loi Normale de moyenne m inconnue. Une étude portant sur un échantillon de 51 câbles a donné une moyenne des charges maximales supportées égales à $\bar{x} = 12.2$ tonnes et d'écart-type empirique $\hat{S} = 0.50$

1. Déterminer un intervalle de confiance pour la moyenne m au niveau 99%

2. Déterminer un intervalle de confiance pour la variance σ^2 au niveau 90%

3. Déterminer la taille minimale de l'échantillon étudié pour que l'amplitude de l'intervalle de confiance pour m au niveau 99% soit inférieure ou égale à 0.3 ?

Solution de l'exercice 16

1. Puisque $n = 51 > 30$, l'estimation par intervalle à 99% ($\alpha = 0.01$) est donnée par

$$IC(m) = \left[\bar{x} - u\frac{\hat{S}}{\sqrt{n}}, \bar{x} + u\frac{\hat{S}}{\sqrt{n}}\right],$$

où $\Phi(u) = 1 - \frac{0.01}{2} = 0.995$ Dans la table de $N(0.1)$ on a

$$u = 2.57,$$

alors

$$IC(m) = \left[12.2 - 2.57\frac{0.5}{\sqrt{51}}, 12.2 + 2.57\frac{0.5}{\sqrt{51}}\right],$$

donc

$$IC(m) = [12.02, 12.38],$$

2. Estimation par intervalle de confiance au risque de 10% ($\alpha = 0.10$) de σ^2

$$IC(\sigma^2) = \left[\frac{(n-1)\hat{S}^2}{\chi_{n-1,\frac{\alpha}{2}}}; \frac{(n-1)\hat{S}^2}{\chi_{n-1,1-\frac{\alpha}{2}}}\right],$$

dans la table khi-deux à 49 degrés de liberté

$$\chi_{n-1,\frac{\alpha}{2}} = \chi_{50,\frac{0.10}{2}} = \chi_{50,0.05} = 67.50,$$

$$\chi_{n-1,1-\frac{\alpha}{2}} = \chi_{50,1-\frac{0.10}{2}} = \chi_{50,0.95} = 34.76,$$

donc

$$IC(\sigma^2) = \left[\frac{50 \times 0.5^2}{67.50}, \frac{50 \times 0.5^2}{34.76}\right],$$

on a

$$IC(\sigma^2) = [0.185, 0.359],$$

3. Soit L : l'amplitude de l'intervalle de confiance de m, donc

$$L = 2u\frac{\hat{S}}{\sqrt{n}},$$

$$\begin{aligned} L &< 0.3 \Rightarrow 2u\frac{\hat{S}}{\sqrt{n}} < 0.3 \\ &\Rightarrow \sqrt{n} > \frac{2u}{0.3}\hat{S} \\ &\Rightarrow n > 73.96. \end{aligned}$$

Donc, la taille minimale de l'échantillon est $n = 74$.

Exercice 17

Un biochimiste étudie un type de moisissure qui attaque les cultures de blé. La toxine continue dans cette moisissure est obtenue sous forme d'une solution organique. La quantité de substance toxique par gramme de solution est une variable aléatoire réelle X suivant une loi Normale. L'unité est le milligramme. On mesure la quantité de substance toxique par gramme de solution. Sur 9 extraits, on a obtenu les mesures suivantes

16.2	*14.1*	*14.6*	*14.8*	*15*	*13.5*	*15.3*	*16.1*	*15.8*

1. Déterminer une estimation ponctuelle de la moyenne et de l'écart-type de la quantité de substance toxique par gramme de solution.

2. Déterminer un intervalle de confiance pour la quantité moyenne de substance toxique par gramme de solution au niveau 95%.

3. Déterminer un intervalle de confiance pour l'carte-type de la quantité de substance toxique par gramme de solution au niveau 90%

4. Déterminer la taille minimum d'échantillon pour que l'amplitude de l'intervalle de confiance de la moyenne soit inférieure à 0.4.

Solution de l'exercice 17

1.

	x_i	n_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
	16.2	1	1.16	1.3456
	14.1	1	-0.94	0.8836
	14.6	1	-0.44	0.1936
	14.8	1	-0.24	0.0576
	15	1	-0.04	0.0016
	13.5	1	-1.54	2.3716
	15.3	1	0.26	0.0676
	16.1	1	1.06	1.1236
	15.8	1	0.76	0.5776
total	135.4	9	/	6.6224

$$\bar{x} = \frac{135.4}{9} = 15.04 \ g$$

$$\hat{S}^2 = \frac{1}{9-1} 6.6224 = 0.8278 \text{ (en } g^2\text{)}$$

donc $\hat{S} \approx 0.91 \ g$ est un estimateur de σ.

2.

· X : La quantité de substance toxique par gramme.

· Un échantillon de taille $n = 9 < 30$.

Eestimation par intervalle de confiance, au risque de 5%, de la moyenne dans la population avec m et σ inconnus (avec n<30) :

$$IC(m) = \left[\bar{x} - t_{n-1,\frac{0.05}{2}} \frac{\hat{S}}{\sqrt{n}}; \bar{x} + t_{n-1,\frac{0.05}{2}} \frac{\hat{S}}{\sqrt{n}}\right],$$

dans la table de student on a

$$t_{9-1,\frac{0.05}{2}} = t_{8,0.025} = 2.306,$$

alors

$$IC(m) = \left[15.04 - 2.306\frac{\sqrt{0.8278}}{\sqrt{9}}, 15.04 + 2.306\frac{\sqrt{0.8278}}{\sqrt{9}}\right],$$

donc

$$IC(m) = [14.341, 15.739] \text{ (en } g).$$

3. Estimation par intervalle de confiance au risque de 10% de σ^2

$$IC(\sigma^2) = \left[\frac{(n-1)\hat{S}^2}{\chi_{n-1,\frac{\alpha}{2}}}; \frac{(n-1)\hat{S}^2}{\chi_{n-1,1-\frac{\alpha}{2}}}\right],$$

dans la table khi-deux à 8 degrés de liberté

$$\chi_{n-1,\frac{\alpha}{2}} = \chi_{9-1,\frac{0.1}{2}} = \chi_{8,0.05} = 15.51,$$

$$\chi_{n-1,1-\frac{\alpha}{2}} = \chi_{9-1,1-\frac{0.1}{2}} = \chi_{8,0.95} = 2.73,$$

donc

$$IC(\sigma^2) = \left[\frac{8 \times 0.8278}{15.51}, \frac{8 \times 0.8278}{2.73}\right],$$

on a

$$IC(\sigma^2) = [0.42698, 2.4258] \text{ (en } g^2)$$

4. On cherche n avec $L < 0.4$, (L : amplitude de l'intervalle)

$$\begin{aligned} L &= 2t_{n-1,\frac{\alpha}{2}}\frac{\hat{S}}{\sqrt{n}} < 0.4 \\ &\Rightarrow \sqrt{n} > \frac{2t_{n-1,\frac{\alpha}{2}}}{0.4}\hat{S} \\ &\Rightarrow \sqrt{n} > \frac{2}{0.4}2.306 \times 0.91 \\ &\Rightarrow \sqrt{n} > 10.492 \times 10.492 \\ &\Rightarrow n > 110.08 \end{aligned}$$

Donc la taille minimum d'échantillon pour l'amplitude soit inférieure 0.4 est $n = 111$.

Exercice 18

On a contrôlé le dosage d'un produit dans un mélange à la sortie d'une chaîne de conditionnement. Pour un échantillon de 10 lots de cinq Kilogramme de mélange, on a obtenu les résultats suivants où x_i représente la masse du produit en grammes

x_i	15.2	13.1	14.5	15.8	15	13.5	16.1

1- Donner une estimation ponctuelle de la moyenne et de l'écart type de la population.

2- Déterminer l'intervalle de confiance de la moyenne de la population au niveau 99%.

3- Déterminer l'intervalle de confiance de la variance de la population au niveau 99%.

4. Déterminer la taille minimum d'échantillon pour que l'amplitude de l'intervalle de confiance de la moyenne soit inférieure 0.5.

5- Un lot de cinq Kilogramme de mélange est dit " de qualité extra" s'il contient entre $13.7g$ et $14.6g$ de produit. On admet que la variable aléatoire x qui, à un lot de cinq Kilogramme de mélange, associe la masse en g du produit dosé, est une variable normale de moyenne $m = 14$ et d'écart-type $\sigma = 3$.

- Calculer le pourcentage de qualité extra dans le mélange.

Solution de l'exercice 18

1.

	x_i	n_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
	15.2	1	0.46	0.2116
	13.1	1	-1.64	2.6896
	14.5	1	0.24	0.0576
	15.8	1	1.06	1.1236
	15	1	0.26	0.0676
	13.5	1	-1.24	1.5376
	16.1	1	1.36	1.8496
total	103.2	7	/	7.5372

$$\begin{aligned} \bar{x} &= \frac{103.2}{7} = 14.74, \\ \hat{S}^2 &= \frac{1}{7-1} 7.5372 = 1.256, \end{aligned}$$

donc $\hat{S} = 1.1207$ est un estimateur de σ.

2.

· X : La masse du produit en garammes.

· Un échantillon de taille $n = 7 < 30$.

Eestimation par intervalle de confiance, au risque de 1%, de la moyenne dans la population avec m et σ inconnus (avec n<30) :

$$IC(m) = \left[\bar{x} - t_{n-1,\frac{0.01}{2}} \frac{\hat{S}}{\sqrt{n}}; \bar{x} + t_{n-1,\frac{0.01}{2}} \frac{\hat{S}}{\sqrt{n}} \right],$$

dans la table de student on a

$$t_{7-1,\frac{0.01}{2}} = t_{6,0.005} = 3.707,$$

alors

$$IC(m) = \left[14.74 - 3.707 \frac{1.1207}{\sqrt{7}}, 14.74 + 3.707 \frac{1.1207}{\sqrt{7}} \right],$$

donc

$$IC(m) = [13.17, 16.31] \text{ (en } g).$$

3. Estimation par intervalle de confiance au risque de 1% de σ^2

$$IC(\sigma^2) = \left[\frac{(n-1)\hat{S}^2}{\chi_{n-1,\frac{\alpha}{2}}}; \frac{(n-1)\hat{S}^2}{\chi_{n-1,1-\frac{\alpha}{2}}} \right],$$

dans la table khi 2 à 19 degrés de liberté

$$\chi_{n-1,\frac{\alpha}{2}} = \chi_{7-1,\frac{0.01}{2}} = \chi_{6,0.005} = 18.55,$$

$$\chi_{n-1,1-\frac{\alpha}{2}} = \chi_{7-1,1-\frac{0.01}{2}} = \chi_{6,0.995} = 0.58,$$

donc

$$IC(\sigma^2) = \left[\frac{6 \times 1.256}{18.55}, \frac{6 \times 1.256}{0.58}\right],$$

on a

$$IC(\sigma^2) = [0.406\,25, 12.\,993] \text{ (en } g^2)$$

4. On cherche n avec $L < 0.5$, (L : amplitude de l'intervalle)

$$\begin{aligned} L &= 2t_{n-1,\frac{\alpha}{2}} \frac{\hat{S}}{\sqrt{n}} < 0.5 \\ &\Rightarrow \sqrt{n} > \frac{2}{0.5} t_{n-1,\frac{\alpha}{2}} \hat{S} \\ &\Rightarrow \sqrt{n} > 16.61 \\ &\Rightarrow n > 276.04. \end{aligned}$$

Donc la taille minimum d'échantillon pour l'amplitude soit inférieure 0.5 est 277.

5. Soit $\Phi(z)$ la fonction de répartition de la loi $N(0.1)$ et

$$\begin{aligned} X &\longrightarrow N(14,3) \\ Z &\longrightarrow N(0.1) \end{aligned}$$

$$\begin{aligned} P(13.7 \leq X \leq 14.6) &= P(\frac{13.7-14}{3} \leq Z \leq \frac{14.6-14}{3}) \\ &= P(-0.1 \leq Z \leq 0.2) \\ &= \Phi(0.2) - \Phi(-0.1) \\ &= \Phi(0.2) - (1 - \Phi(0.1)), \end{aligned}$$

lue dans la table $N(0.1)$

$$\begin{aligned} P(13.7 \leq X \leq 14.6) &= 0.5793 - (1 - 0.5398) \\ &= 0.1191. \end{aligned}$$

Le pourcentage de qualité extra dans le mélange est 11.91%.

Chapitre 4

Tests d'hypothèse et Analyse de régression corrélation

4.1 L'essentiel du cours

Les tests d'hypothèse constituent un autre aspect important de l'inférence statistique. Le principe général d'un test d'hypothèse peut s'énoncer comme suit :

• On étudie une population dont les éléments possèdent un caractère (mesurable ou qualitatif) et dont la valeur du paramètre relative au caractère étudié est inconnue.

• Une hypothèse est formulée sur la valeur du paramètre : cette formulation résulte de considérations théoriques, pratiques ou encore elle est simplement basée sur un pressentiment.

• On veut porter un jugement sur la base des résultats d'un échantillon prélevé de cette population.

Un test d'hypothèse (ou test statistique) est une démarche qui a pour but de fournir une règle de décision permettant, sur la base de résultats d'échantillon, de faire un choix entre deux hypothèses statistiques.

4.1.1 Test de la moyenne

Notations :

X, variable quantitative étudiée sur deux populations.

n_1, n_2, tailles des deux échantillons extraits des populations.

m_1, σ_1 et m_2, σ_2, moyennes et variances de X sur les deux populations.

$\bar{x}$, $\bar{y}$, moyennes de X sur les deux échantillons.

$\hat{S}_1^2, \hat{S}_1^2$, variances estimées de X pour les populations, à partir des deux échantillons respectifs.

Le problème posé

La moyenne m_1 est-elle différente de la moyenne m_2 ?

Auquel cas, on posera :

$$H_0 : m_1 = m_2,\ H_1 : m_1 \neq m_2,$$

La variable de décision dépend de la connaissance ou non des variances σ_1^2 et σ_2^2, à savoir :

Cas1 : σ_1^2 et σ_2^2 connues

La variable de décision notée Z est telle que :

H_0 vraie $(m_1 = m_2) \Rightarrow Z \rightarrow N(0.1)$

à condition que :

$X \rightarrow N(m_1, \sigma_1)$ ou $n_1 \geq 30$ dans population 1

$X \rightarrow N(m_2, \sigma_2)$ ou $n_2 \geq 30$ dans population 2

et une réalisation z de Z est définie par

$$z = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

En fonction des hypothèses posées et à l'aide de la table de la loi normale (chapitre 5, table 1), on peut alors déterminer la zone de non-rejet (et la zone de rejet) de H_0, après avoir fixé le risque α.

-Zone de **non-rejet** de H_0 ou intervalle de confiance de z à $1 - \alpha$:

$$z \in [-u, -u]\,.$$

-Zone de **rejet** de H_0 au risque de α :

$$z \in]-\infty, -u[\cup]u, +\infty[.$$

où u, lue dans la table $N(0,1)$ telle que $\Phi(u) = 1 - \frac{\alpha}{2}$.

Remarque

Si σ_1 et σ_2 sont inconnues, on les remplace par les estimateurs $\hat{S}_1^2$ et $\hat{S}_2^2$ respectivement, c.à.d

$$z = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\hat{S}_1^2}{n_1} + \frac{\hat{S}_2^2}{n_2}}}.$$

Cas 2 : σ_1^2 et σ_2^2 inconnues et $\sigma_1^2 = \sigma_2^2 = \sigma$ et $n_1 < 30, n_2 < 30$.

La variable de décision notée T est elle que :

H_0 vrai $(m_1 = m_2) \Rightarrow T \rightarrow$Student, et une réalisation t de Test définie par :

$$t = \frac{\bar{x} - \bar{y}}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

avec

$$S = \sqrt{\frac{(n_1 - 1)\,\hat{S}_1^2 + (n_2 - 1)\,\hat{S}_2^2}{n_1 + n_2 - 2}},$$

En fonction des hypothèses posées et à l'aide de la table de la loi de *Student* (chapitre 5, table 2), on peut alors déterminer la zone de non-rejet (et la zone de rejet) de H_0, après avoir fixé le risque α.

-Zon de **non-rejet** de H_0 ou intervalle de confiance de t à $1 - \alpha$:

$$t \in]-t_c, t_c[\,.$$

-Zon de **rejet** de H_0 au risque de α :

$$t \in]-\infty, -t_c[\cup]t_c, +\infty[.$$

où t_c, *lue dans* la table de Student à $k = n_1 + n_2 - 2$ dégrées de liberté (ddl) et $\gamma = \frac{\alpha}{2}$.

4.1.2 Comparaison de deux proportions

Notations :

X, variable qualitative (àdeux modalités) étudiée sur deux populations.

n_1, n_2, tailles des deux échantillons extraits des populations.

p_1 et p_2, proportions d'une modalité de X sur les deux populations.

f_1, f_2, proportions de cette modalité sur les deux échantillons.

Le problème posé

La proportion p_1 est-elle différente de la proportion p_2 ?

Auquel cas, on posera :

$$H_0 : p_1 = p_2,\ H_1 : p_1 \neq p_2,$$

La variable de décision notée Z est telle que :

H_0 vraie $(p_1 = p_2) \Rightarrow Z \rightarrow N(0.1)$ à condition que :

$n_1 \geq 30, n_1 f_1 \geq 5, n_1(1-f_1) \geq 5$ et $n_2 \geq 30, n_2 f_2 \geq 5, n_1(1-f_1) \geq 5,$

et une réalisation z de Z est définie par

$$z = \frac{f_1 - f_2}{\sqrt{f(1-f)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}},$$

où

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

En fonction des hypothèses posées et à l'aide de la table de la loi normale (chapitre 5, table 1), on peut alors déterminer la zone de non-rejet (et la zone de rejet) de H_0, après avoir fixé le risque α.

-Zone de **non-rejet** de H_0 ou intervalle de confiance de z à $1-\alpha$:

$$z \in [-u, -u]\,.$$

-Zone de **rejet** de H_0 au risque de α :

$$z \in]-\infty, -u[\cup]u, +\infty[.$$

où u, lue dans la table $N(0,1)$ telle que $\Phi(u) = 1 - \frac{\alpha}{2}$.

4.1.3 Test de Fisher (Comparaison de deux variances)

Notations :

X, variable quantitative étudiée sur deux populations.

n_1 et n_2, tailles des deux échantillons extraits des populations.

σ_1^2 et σ_2^2, variances de X sur les deux populations.

$\hat{S}_1^2$ et $\hat{S}_1^2$, variances estimées de X pour les populations, à partir des deux échantillons respectifs.

Le problème posé

La variance σ_1^2 est-elle différente de la variance σ_2^2 ?

Auquel cas, on posera :

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

La variable de décision est telle que :

H_0 vraie $(\sigma_1^2 = \sigma_2^2) \Rightarrow F \rightarrow$ La loi de Fisher F_{n_1-1,n_2-1} à condition que :
$X \rightarrow N(m_1, \sigma_1)$ ou $n_1 \geq 30$ dans population 1
$X \rightarrow N(m_2, \sigma_2)$ ou $n_2 \geq 30$ dans population 2, et une réalisation f de F est définie par

$$f = \frac{S_1^2}{S_2^2} \text{ si } \hat{S}_1^2 > \hat{S}_2^2, \quad f = \frac{S_2^2}{S_1^2} \text{ si } \hat{S}_1^2 < \hat{S}_2^2,$$

En fonction des hypothèses posées et à l'aide de la table de la loi de Fisher (chapitre 5, table 4), on peut alors déterminer la zone de non-rejet (et la zone de rejet) de H_0, après avoir fixé le risque α.

-Zone de **non-rejet** de H_0 :

$$f < F^{\alpha}_{n_1-1,n_2-1}$$

-Zone de **rejet** de H_0 au risque de α :

$$f > F^{\alpha}_{n_1-1,n_2-1}$$

où $F^{\alpha}_{n_1-1,n_2-1}$, lue dans la table de Fisher au risque d'erreur α et à $n_1 - 1$ et $n_2 - 1$ dégrées de liberté.

4.1.4 Les Tests du Khi 2 (Test d'adéquation)

Ce test permet de juger la qualité de l'ajustement d'une distribution expérimentale à une distribution théorique en d'autres termes il permet de tester l'hypothèse que les fréquences observées pour les différentes catégories (classes) sont en adéquation avec une distribution donnée.

Notations :

O_i : représentent les fréquences observées des résultats
T_i : représentent les fréquences attendues (théoriques) des résultats
k : représente le nombre de classes
N : représente le nombre total d'essais

Conditions d'application du test :

1. Les données sélectionnées aléatoirement.
2. Pour chaque catégorie , la fréquence attendue est supérieure ou égale à 5 $(T_i \geq 5)$

Hypothèses à tester :

H_0 : les observations suivent la distribution théorique

H_1 : les observations ne suivent pas la distribution théorique

Statistique du test :

$$K = \sum_{i=1}^{k} \frac{(O_i - T_i)^2}{T_i},$$

Décision :

On accepte H_0 si

$$K < \chi_{k-1-R,\alpha}$$

où la valeurs $\chi_{k-1-R,\alpha}$ est leu dans la table du Khi-deux avec $(k-1-R)$ dégrées de liberté (ddl)$(\gamma = \alpha)$ avec k est le nombre de classes et R est le nombre de paramétres à estimer éventuellement pour caractériser la distribution théorique

On rejette H_0 si

$$K > \chi_{k-1-R,\alpha}.$$

4.1.5 Analyse de régression corrélation

La régression linéaire : l'ajustement à une droite par la méthode des mondres carrés est toujours possible ; le calcul des paramètres de la droite et le fait qu'elle passe par le point de coordonnées x et y valeurs moyennes de $\bar{x}$ et $\bar{y}$; la pratique des tests associés sur la pente et l'ordonnée à l'origine.

La corrélation : le calcul du coefficient de corrélation et toujours possible ; la signification de son signe et de sa valeur r toujours compris entre -1 et 1 : si r est proche de, 1 il y a une ford dépendance linéaire, si r est voisine de zéro il n'y a pas de dépendance linéaire mais il peut exister une autre forme de liaison (par exemple exponentielle...) ; la nécessité de pratique est test statistique pour conclure à une liaison.

Test sur le coefficient de corrélation r

Sous les hypothèses de normalité des distributions de X et Y, on peut tester l'existence d'une liaison supposée linéaire entre X et Y.

Le coefficient r calculé est une valeur issue de l'échantillon de mesures, estimation du coefficient ρ qui caractérise la liaison dans la population des mesures de la série double de X et Y.

$$\begin{cases} H_0 : \rho = 0, \\ H_1 : \rho \neq 0. \end{cases}$$

La variable de décision T suit sous H_0 un loi de student à $n-1$ ddl. $t_c = \frac{r}{s_r}$ est une réalisation de T où s_r est calculer à partir de

$$s_c^2 = \frac{1-r^2}{n-2}.$$

La valeur $t_{n-2,\frac{\alpha}{2}}$ est lue dans la table de Student à $k = n-2$ dégrés de liberté (ddl) et $\gamma = \frac{\alpha}{2}$.

La valeur limite $t_{n-2,\frac{\alpha}{2}}$ qui permet de définir les intervalles de rejet ou non-rejet de H_0.

Si $t_c \in$ la zone de non-rejet de H_0, alors on conclut au risque α au non-rejet de H_0 : le coefficient de corrélation linéaire ne diffère pas significativement de 0.

Si $t_c \notin$ à la zone de rejet de H_0, alors on conclut au risque α au rejet de H_0 : le coefficient de corrélation linéaire diffère significativement de 0 : on peut dire qu'il y a au risque α, liaison entre X et Y. Plus la valeur de r est proche de 1, plus intense est la liaison.

4.2 Exercices corrigés

Exercice 1

On désire comparer la pression artérielle diastolique d'un groupe de sujets sains et d'un groupe de sujets atteints de drépanocytose (hémoglobinopathie). Une étude donne les résultats suivants :

	Effectif n_i	P.A.D.M (mm Hg)	$\hat{S}_i^2$
Sujets sains(A)	88	70.1	$\hat{S}_1^2 = 10.8$
Sujets Drépanocytaires (B)	85	61.8	$\hat{S}_2^2 = 6.9$

P.A.D.M=Pression artérielle diastolique moyenne.

La pression artérielle est différente chez les sujets drépanocytaires ?

Solution de l'exercice 1

Il s'agit d'un test bilatéral (égale conte inégale)

a. Choix des hypothèses :

$$\begin{aligned} H_0 &: m_1 = m_2, \\ H_1 &: m_1 = m_2, \end{aligned}$$

Il s'agit d'un test bilatéral (égale contre inégale)

b. Calcul de la statistique de test observée
$n_1 \geq 30, n_2 \geq 30$, σ_1 et σ_2 inconnus, donc

$$\begin{aligned} z &= \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\hat{S}_1^2}{n_1} + \frac{\hat{S}_2^2}{n_2}}} \\ &= \frac{70.1 - 61.8}{\sqrt{\frac{10.8}{88} + \frac{6.9}{85}}} = 18.381, \end{aligned}$$

c. Identification de seuil critique :
Alors on cherche le seuil critique dans la table $N(0.1)$ avec $\alpha = 0.01$

$$u = 2.575,$$

d. Décision : comme $z \notin]-2.575, 2.575[$, la statistique de test observée se trouve dans la zone de rejet de H_0.

Décision pratique : La pression artérielle est différente chez les sujets drépanocytaires

Exercice 2

Une machine automatique fabrique des pièces dont on veut contrôle la largeur.

On sait que cette largeur est distribuée selon une loi normale de moyenne m et écart-type $\sigma = 2.50mm$.

L'écart-type étant supposé constant, on prélève régulièrement des échantillons de 35 pièces, afin d'établir une carte de contrôle de la moyenne.

Les moyennes obtenu sur des échantillons prélever en début et en d'une même journée sont respectivement de $60,5mm$ et $61,25mm$.

Peut-on dire que la langue des pièces produits à changé au cours de la journée ?

Solution de l'exercice 2

Variable étudiée : X, la largeur des pièces.

L'ensemble des pièces produites par la machine est caractérisé par :

$$X \to N(m, 1.50)$$

Echantillon 1	Echantillon 2
$n_1 = 35$	$n_2 = 35$
$\bar{x} = 60.50$	$\bar{y} = 61.25$
$\sigma_1 = 2.5$	$\sigma_1 = 2.5$

Problème posé : la largeur des pièces produites au cours de la journée a-t-elle change, c'est-à-dire la largeur moyenne des pièces produites en début de journée est-elle différente de la largeur moyenne des pièces produites en fin de journée ?

On pose donc

$$\begin{aligned} H_0 &: m_1 = m_2, \\ H_1 &: m_1 = m_2, \end{aligned}$$

La zone de non-rejet de H_0, au risque de 5%, est donc : $z \in [-1.96, 1.96]$.

Sur nos données,

$$\begin{aligned} z &= \frac{60.5 - 61.25}{\sqrt{\frac{2.5^2}{35} + \frac{2.5^2}{35}}} \\ &= -1.255\,. \end{aligned}$$

La valeur z calculée sur nos données étant dans l'intervalle $[-1.96, 1.96]$, on ne peut pas rejeter H_0, au risque de 5%.

Conclusion : au risque de 5%, on ne peut pas dire que la largeur des pièces a changé au cours de la journée.

Exercice 3

Le service des études d'un laboratoire de fabrication de régimes amaigrissants a commandé une enquête une vue d'améliorer les performances d'un nouveau régime de durée de 6 semaines. Un échantillon de 8 femmes et 12 hommes a été sélectionné.

On a noté le poids des différentes personnes au début et à la fin du régime.

Les résultat (en kilogramme) sur les suivants :

-Femmes

Début (D_i^f)	79	80	75	74	76	71	69	65
Fin (F_i^f)	75	72	74	71	71	66	65	63

-Hommes

Debut (D_i^h)	78	85	88	84	82	79	96	102	95	83	81	92
Fin (F_i^h)	71	80	84	80	80	73	87	91	90	80	77	88

Sachant que le poids il distribué selon une loi normale, le régime pour les hommes est-il plus efficace que celui pour les femmes ?

Solution de l'exercice 3

Variable étudiée : X le poids, plus précisément la différence D de poids entre le début et la fin du régime.

Deux échantillons :

-Femmes

$x_i = F_i^f - D_i^f$	4	8	1	3	5	5	4	2

-Hommes

$y_i = D_i^h - F_i^h$	7	5	4	4	2	6	9	11	5	3	4	4

Echantillon 1	Echantillon 2
$n_1 = 8$	$n_2 = 12$
$\bar{x} = 4$	$\bar{y} = 5.333$
$\hat{S}_1^2 = 4.5714$	$\hat{S}_2^2 = 6.6061$

Problème posé : le régime est-il plus efficace pour les hommes ou la moyenne de la différence de poids chez les hommes est-elle plus grande que la moyenne de la différence de poids chez les femmes ?

La variable de décision dépend alors de l'égalité ou non de ces variances.

Première étape : y a-t-il égalité des variances ?

On pose donc

$$H_0 : \sigma_1 = \sigma_2,\ H_1 : \sigma_1 \neq \sigma_2,$$

H_0 vraie $(\sigma_1^2 = \sigma_2^2) \Rightarrow F \rightarrow$ La loi de Fisher F_{n_1-1,n_2-1} et une réalisation f de F est définie par :

$$\begin{aligned} f &= \frac{S_2^2}{S_1^2} \\ &= 1.445, \end{aligned}$$

Sur la table du Fisher ,on lit

$$F_{n_1-1,n_2-1}^{\alpha} = F_{11,7}^{0.05} = 4.715,$$

Décision : comme $f < \chi_{2,0.01}$, alors on accepte H_0.

Conclusion : au risque de 5%, on ne peut pas dire que les variances soient différentes.

Seconde étape : comparaison des moyennes

On pose donc

$$\begin{aligned} H_0 &: m_1 = m_2, \\ H_1 &: m_1 = m_2, \end{aligned}$$

on trouve z

$$\begin{aligned} S &= \sqrt{\frac{(n_1-1)\,S_1^2 + (n_2-1)\,S_2^2}{n_1+n_2-2}} \\ &= \sqrt{\frac{(7)\,(4.5714) + (11)\,(6.6061)}{8+12-2}} \\ &= 5.815\,, \end{aligned}$$

donc

$$\begin{aligned} z &= \frac{5.333-4}{5.815\sqrt{\frac{1}{8}+\frac{1}{12}}} \\ &= 1.211. \end{aligned}$$

La table de la fonction de répartition de la variable de Student à 18 ($n_1 + n_2 - 2$) degrés de liberté fournit $t_c = 1.734$, au risque de 5%.

La valeur z calculée sur nos données étant dans l'intervalle $]-1.734, 1.734[$ on ne peut pas rejeter H_0, au risque de 5%.

Conclusion : au risque de 5%, on ne peut pas dire que le régime est plus efficace chez les hommes que chez les femmes.

Exercice 4

On a mesuré un marqueur biologique chez 2 séries de sujets, l'une composée des sujets sains, l'autre de sujets atteints d'hépatite alcoolique.

L'étude a retrouvé les résultats suivants :

	Effectif n_i	Moyenne du marqueur $(g\backslash l)$	$\hat{S}_i$
Sujets sains(A)	15	1.6	0.19
Sujets sains (B)	12	1.4	0.21

On suppose que le marqueur se distribue normalement chez les deux populations et que les deux écarts-types sont égaux.

Le marqueur biologique est-il différent chez les sujets atteints d'hépatite alcoolique ? ($\alpha = 0.05$).

Solution de l'exercice 4

Il s'agit d'un test bilatéral (égale conte inégale)

a. Choix des hypothèses :

$$\begin{aligned} H_0 &: m_1 = m_2, \\ H_1 &: m_1 = m_2, \end{aligned}$$

Il s'agit d'un test bilatéral (égale contre inégale)

b. Calcul de la statistique de test observée

$n_1 < 30, n_2 < 30$, σ_1 et σ_2 inconnus, donc

$$\begin{aligned} S &= \sqrt{\frac{(n_1-1)\,S_1^2 + (n_2-1)\,S_2^2}{n_1+n_2-2}} \\ &= \sqrt{\frac{(15-1)\,(0.19)^2 + (12-1)\,(0.21)^2}{15+12-2}} \\ &= 0.199, \end{aligned}$$

$$\begin{aligned} z &= \frac{1.6-1.4}{0.199\sqrt{\frac{1}{15}+\frac{1}{12}}} \\ &= 2.595 \end{aligned}$$

c. Identification de seuil critique :

Alors on cherche le seuil critique dans la table de Student

$$t_{15+12-2,\frac{0.05}{2}} = 2.06,$$

d. Décision : comme $z \notin]-2.06, 2.06[$, la statistique de test se trouve dans la zone de rejet de H_0

Décision pratique : On conclut que les malades atteints de l'hépatite alcoolique présentent une valeur du manque différente de celle des sujets sains

Exercice 5

On veut comparer la précision de deux méthodes de dosage du menthol dans l'essence de menthe poivrée.

Pour cela, on dose le menthol dans 16 flacons par ces deux méthodes. Les variance des résultats obtenus sont respectivement $0,013g^2/L^2$ (méthode 1) et $0,024g^2/L^2$ (méthode 2).

Peut -on dire, au risque de 5%, que ces deux méthodes n'ont pas la même précision (on fera les hypothèses nécessaires) ?

Solution de l'exercice 5

Variable étudiée : X le taux de menthol.

Méthode 1	Méthode 2
$n_1 = 16$	$n_2 = 16$
$\hat{S}^2_{éch1} = 0.013$	$\hat{S}^2_{éch2} = 0.024$

où $\hat{S}^2_{éch1}$ et $\hat{S}^2_{éch2}$, variances de Xsur les deux méthodes.

Le problème posé : les précisions des deux méthodes sont-elles différentes ou la variance des résultats obtenus par la méthode 1 est-elle différente de la variance des résultats obtenus par la méthode 2 ?

On pose l'hypothèse $H_0 : \sigma_1 = \sigma_2$, donc

$$\begin{aligned} f &= \frac{S_2^2}{S_1^2} = \frac{S^2_{éch2}}{S^2_{éch1}} \\ &= \frac{0.024}{0.013} = 1.846 \end{aligned}$$

La table de la loi F de Fischer-Snedcecor à $(15, 15)$ degrés de liberté donne :

$$F^{0.05}_{16-1,16-1} = 2.86,$$

donc $F < F^{\alpha}_{n_1-1,n_2-1}$, on accepte H_0 (on admet alors l'égalité des variances).

Conclusion : au risque de 5%, on peut pas dire que les précisions des deux méthodes soient différentes.

Exercice 6

Dans une usine sont fabriquées des micropipettes de laboratoire de volume calibrer à $100ml$. En cours de fabrication sur deux machines différentes, on prélève deux lots indépendants de 25 micropipettes et on obtient les valeurs suivantes pour les variances calculées sur les deux échantillons $S^2_{éch1} = 0,02ml^2$ et $S^2_{éch2} = 0,015ml^2$. On suppose que la variable aléatoire X, volume d'une micropépittes suit une loi normale.

Peut-on dire pour les variance des volumes de micropipettes fabriquées par ces deux des machines sont différentes.

Solution de l'exercice 6

Variable étudiée : X = volume d'une micropipette (variable quantitative).

Distribution de X supposé normal pour des populations de mécropipettes produites par les deux machines.

$S^2_{éch1} = 0,02ml^2$ et $S^2_{éch2} = 0,015ml^2$ variances de X sur les deux échantillons.

σ_1^2 et σ_2^2 les variances de X dans les populations de toutes les micropipettes, produites par les deux machines.

Problème posé : comparaison des deux variance observées.

Hypothèse :

$$H_0 : \sigma_1 = \sigma_2,\ H_1 : \sigma_1 \neq \sigma_2,$$

test bilatéral.

Conditions d'applications du test : X normal dans les deux populations (supposer dans l'énoncé)

$$\hat{S}_1^2 = \frac{n_1}{n_1 - 1}\hat{S}^2_{éch1} = \frac{25}{24}0,02 = 0.02083$$

et

$$\hat{S}_2^2 = \frac{n_2}{n_2 - 1}\hat{S}^2_{éch2} = \frac{25}{24}0.015 = 0.015625$$

donc

$$\begin{aligned} f &= \frac{S_2^2}{S_1^2} \\ &= \frac{0.02083}{0.015625} = 1.333. \end{aligned}$$

La table de la loi F de Fischer-Snedcecor à $(24, 24)$ degrés de liberté, au risque 5%, donne :

$$F^{0.05}_{16-1,16-1} = 2.27,$$

donc $F < F^{\alpha}_{n_1-1,n_2-1}$, on accepte H_0 (on admet alors l'égalité des variances).

Conclusion : au risque de 5%, on ne peut pas conclure que les variances des mesures des volumes de micropipettes produites par les deux machines sont différentes.

Exercice 7

Pour un échantillon de 12 sujets sains, on a obtenu les résultats expérimentaux suivants :

· Moyenne des taux sanguins de calcium : $\bar{x} = 100$ mg/m

· Ecart-type des taux sanguins de calcium : $\hat{S}_1 = 5.8$ mg/l

Pour un échantillon de 16 sujets présentant une tumeur ostéolytique, on a obtenu les résultats expérimentaux suivants :

· Moyenne des taux sanguins de calcium : $\bar{y} = 130$ mg/m

· Ecart-type des taux sanguins de calcium : $\hat{S}_2 = 6$ mg/l

En supposant que le taux sanguin se distribue selon la loi normale tester au risque de 5% l'hypothèse selon laquelle les moyennes des taux sanguins de calcium de ces deux groupes d'individus sont significativement différentes.

Solution de l'exercice 7

Test de comparaison de deux moyennes m_1 et m_2 à partir de deux échantillons indépendants, présentation des données :

Echantillon 1	Echantillon 2
$n_1 = 12$	$n_2 = 16$
$\bar{x} = 100$	$\bar{y} = 130$
$\hat{S}_1 = 5.8$	$\hat{S}_2 = 6$

Sous cette hypothèse, on teste l'une contre l'autre les deux hypothèses :

$$\begin{aligned} H_0 &: m_1 = m_2, \\ H_1 &: m_1 \neq m_2, \end{aligned}$$

$$\begin{aligned} S &= \sqrt{\frac{(n_1-1)\,S_1^2 + (n_2-1)\,S_2^2}{n_1+n_2-2}} \\ &= \sqrt{\frac{(12-1)\,(5.8)^2 + (16-1)\,(6)^2}{12+16-2}} \\ &= 5.916\,, \end{aligned}$$

$$\begin{aligned} z &= \frac{100-130}{5.916\sqrt{\frac{1}{12}+\frac{1}{16}}} \\ &= -13.28. \end{aligned}$$

La table de la fonction de répartition de la variable de Student à 26 degrés de liberté fournit $t_{12+16-2,0.025} = 2.056$.

La valeur observée $z = -13.28$ appartient à la région de rejet de H_0 ($z \in]-\infty, -2.056[\cup\ ]2.056, +\infty[$) : on rejette donc H_0 en faveur de H_1 au risque $\alpha = 5\%$. Ces 2 moyennes sont significativement différentes.

Exercice 8

On teste deux traitements anti-cancéreux A et B sur deux populations de patients P_A et P_B (De même taille $n_A = n_B = 50$). L'efficacité d'un traitement est évaluée par l'éventuelle diminution de la taille de la lésion tumorale, estimée par l'imagerie médicale, après un an de traitement. Pour la population soumise au traitement A on observe une diminution de la taille des tumeurs dans 27 cas sur 50, pour le traitement B, dans 18 cas.

Peut-on conclure à une différence d'effet des deux traitements (au seuil de 5%) ?

Solution de l'exercice 8

Il s'agit d'un test bilatéral (égale conte inégale)

a. Choix des hypothèses :

$$\begin{aligned} H_0 &: p_1 = p_2, \\ H_1 &: p_1 \neq p_2, \end{aligned}$$

Il s'agit d'un test bilatéral (égale contre inégale)

$n_1 = 50, n_2 = 50, f_1 = \frac{27}{50}$ et $f_2 = \frac{18}{50}$,

b. Calcul de la statistique de test observée

$$\begin{aligned} f &= \frac{50\left(\frac{27}{50}\right) + 50\left(\frac{18}{50}\right)}{50+50} \\ &= 0.45, \end{aligned}$$

alors

$$\begin{aligned} z &= \frac{\frac{27}{50} - \frac{18}{50}}{\sqrt{0.45(1-0.45)\left(\frac{1}{50}+\frac{1}{50}\right)}} \\ &= 1.809, \end{aligned}$$

c. Identification de seuil critique : le test est valable, les conditions sont remplies

$$n_1 \geq 30, n_1 f_1 \geq 5, n_1(1-f_1) \geq 5 \text{ et } n_2 \geq 30, n_2 f_2 \geq 5, n_1(1-f_1) \geq 5,$$

Alors on cherche le seuil critique dans la table $N(0.1)$, $\alpha = 0.05$

$$u = 1.96,$$

d. Décision statistique : comme $z \in]-1.96, 1.96[$, La statistique de test observée se trouve dans la zone d'acceptation de H_0.

Décision pratique : On conclut qu'il n'y a pas une différence d'effet entre les deux traitements.

Exercice 9

A partir du génotype des parents, on s'attend à ce que les enfants aient des génotypes répartis comme suit :

25% de génotype AA,

50% de génotype Aa,

25% de génotype aa.

Pour une maladie particulière , AA représente un enfant sain , Aa un enfant porteur et aa un enfant malade.

Le tableau suivant donne les fréquences des génotypes pour 90 malades choisis aléatoirement

génotype	AA	Aa	aa
Fréquences observées O_i	22	55	13

Tester au niveau de significativité $\alpha = 1\%$, l'hypothèse que ces fréquences observées peuvent être ajustées aux fréquences attendues de la distribution théorique.

Solution de l'exercice 9

H_0 : la distribution des génotypes des enfants est adéquate avec la distribution donnée

$$p_1 = 0.25, p_2 = 0.50, p_3 = 0.25,$$

H_1 : la distribution des génotypes des enfants n'est pas adéquate avec la distribution donnée

-Vérifions que les conditions du test sont satisfaites

1. Les données sont sélectionnées aléatoirement

2. Les fréquences attendues sont supérieures ou égales à 5 pour cela on doit d'abord les calculer $T_i = Np_i$

Calcul des T_i et les différences entre les O_i et les T_i.

Génotype	AA	Aa	aa
Fréq .obs .O_i	22	55	13
Fréq.the T_i	$90 \times 0.25 = 22.5$	$90 \times 0.50 = 45$	$90 \times 0.25 = 22.5$
$O_i - T_i$	-0.5	10	-9.5
$(O_i - T_i)^2$	0.25	100	90.25
$\frac{(O_i - T_i)^2}{T_i}$	0.0111	2.2222	4.0111

Statistique du test :

$$\begin{aligned} K &= \sum_{i=1}^{3} \frac{(O_i - T_i)^2}{T_i} \\ &= 0.0111 + 2.2222 + 4.0111 \\ &= 6.2444 \end{aligned}$$

Valeur critique :
Sur la table du khi-deux ,on lit

$$\chi_{2,0.01} = 9.210$$

Décision : comme $k < \chi_{2,0.01}$, alors on accepte H_0.
Conclusion :

Avec un risque de 1% on peut dire que la distribution observée est conforme avec la distribution donnée.

Exercice 10

Une analyse de régression linéaire à partir d'un échantillon de taille $n = 15$ a produit ce qui suit :

$$\bar{x} = 13.4, \bar{y} = 56.4, \sum_{i=1}^{15}(x_i - \bar{x})(y_i - \bar{y}) = 156.4,$$

$$\sum_{i=1}^{15}(x_i - \bar{x})^2 = 173.5 \text{ et } \sum_{i=1}^{15}(y_i - \bar{y})^2 = 40.621.$$

1. Déterminez la droite de régression linéaire de Y en X ($y = ax + b$) de la série statistique double (X, Y).

2. Déterminer le coefficient de corrélation linéaire r.

3. Tester l'existence d'une corrélation linéaire entre la variable aléatoire Y et la variable aléatoire X. (On teste les hypothèses $H_0 : \rho = 0$ contre $H_1 : \rho \neq 0$, le risque est de 5%.

Solution de l'exercice 10

A partir des données on aura la droite de régression linéaire et le coefficient de corrélation

1. La droite de régression : $y = ax + b$

$$\begin{aligned} a &= \frac{Cov(x,y)}{S(x)} = \frac{\sum_{i=1}^{15}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{15}(x_i - \bar{x})^2} \\ &= \frac{156.4}{173.5} = 0.901\,44, \end{aligned}$$

$$\begin{aligned} \bar{y} &= a\bar{x} + b \Longrightarrow b = \bar{y} - a\bar{x} \\ &\Longrightarrow b = 56.4 - 0.90144 \times 13.4 \\ &\Longrightarrow b = 44.\,321, \end{aligned}$$

donc

$$y = 0.90144x + 44.321.$$

2. Le coefficient de corrélation

$$r = \frac{Cov(x,y)}{\sigma_x \sigma_y}.$$

Détermination de

$$\begin{aligned} \sigma_x^2 &= \frac{1}{15}\sum_{i=1}^{15}(x_i - \bar{x})^2 = 11.567, \\ \sigma_y^2 &= \frac{1}{15}\sum_{i=1}^{15}(y_i - \bar{y})^2 = 12.107, \\ Cov(x,y) &= \frac{1}{15}\sum_{i=1}^{15}(x_i - \bar{x})(y_i - \bar{y}) = 10.427, \end{aligned}$$

on a

$$\begin{aligned} r &= \frac{10.427}{\sqrt{11.567}\sqrt{12.107}} \\ &= 0.8811. \end{aligned}$$

3. Test d'hypothèse sur la corrélation linéaire

$$\begin{cases} H_0 : \rho = 0 \\ H_1 : \rho \neq 0 \end{cases}$$

La variable de décision T suit sous H_0 un loi de student à $n-1$ ddl et

$$\begin{aligned} s_c^2 &= \frac{1-r^2}{n-2} = \\ &\quad \frac{1-(0.8811)^2}{15-2} \\ &= 1.720\,5 \times 10^{-2}, \end{aligned}$$

on a

$$\begin{aligned} t_c &= \frac{r}{s_r} \\ &= \frac{0.8811}{\sqrt{1.720\,5 \times 10^{-2}}} \\ &= 6.717\,, \end{aligned}$$

Dans la table 2 de student, on trouve $t_{n-2,\frac{\alpha}{2}} = t_{15.-2,\frac{0.05}{2}} = 2.16$. Donc $t_c \notin [-2.16, 2.16]$, t_c étant à l'extérieur de l'intervalle de confiance à 95%, on rejette H_0 au seuil de 5%.

Exercice 11

Dans l'optique d'étudier la liaison entre le poids d'un père et de son fils ainé, on a relevé le poids de 12 pères et leur fils ainé respectif. Les résultats (en Kg) sont dans le tableau suivant :

Poids Père	65	63	67	64	68	62	70	66	68	67	69	71
Poids Fils	68	66	68	65	69	66	68	65	71	67	68	71

Peu-on dire que les poids des pères et fils sont liés ? (On supposera que les hypothèses de normalité sont vérifiées).

Solution de l'exercice 11

Deux variables sont étudiées ici, à savoir : Poids du père et poids du fils (variables quantitatives).

Ces variables ont été observées sur 12 pères et 12 fils.

Problème posé :

Les poids du père et du fils sont-ils corrélés linéairement ?

Il s'agit donc d'un problème de corrélation (aucune variable contrôlée), plus exactement du calcul du coefficient de corrélation r et de sa comparaison à la valeur zéro.

a. Calcul du coefficient de corrélation linéaire r

Après avoir calculé $\bar{y}, \bar{y}, S(x), S(y)$, et $Cov(x, y)$ comme dans le premier chapitre, en obtient

$$\begin{aligned} r &= \frac{Cov(x,y)}{S(x)S(y)} \\ &= 0.726. \end{aligned}$$

b. Comparaison du coefficient de corrélation à la valeur zéro

$$\begin{cases} H_0 : \rho = 0, \\ H_1 : \rho \neq 0, \end{cases}$$

La variable de décision T suit sous H_0 un loi de student à $n-1$ ddl et

$$\begin{aligned} s_c^2 &= \frac{1-r^2}{n-2} = \\ & \frac{1-(0.726)^2}{12-2} \\ &= 4.729\,2 \times 10^{-2}, \end{aligned}$$

on a

$$\begin{aligned} t_c &= \frac{r}{s_r} \\ &= \frac{0.726}{\sqrt{4.729\,2 \times 10^{-2}}} \\ &= 3.34. \end{aligned}$$

Dans la table 2 de student, on trouve $t_{n-2,\frac{\alpha}{2}} = t_{12-2,0.025} = 2.228$. Donc $t_c \notin [-2.228, 2.228]$, t_c étant à l'extérieur de l'intervalle de confiance à 95%, on rejette H_0 au seuil de 5% et même au seul de 1% (hors intervalle $[-3.169, 3.169]$).

Conclusion : au seuil de 1%, on peut dire que le poids du père est corrélé linéairement avec celui de son fils ané

Exercice 12

Sur un échantillon de 12 femmes, on dose l'hormone l'octogène dans le sang circulant et dans le liquide amniotique (en ng/ml). Les résultats obtenus se présenter de la table

Liquide amniotique	11	8	15	13	10	11	14	9	12	13	10	13
Poids Fils	4.8	3.9	7.3	6.7	4.4	5.4	7.1	4.3	5.7	6.3	4.9	6.1

Le taux de l'hormone lactogène dans le sang et dans le liquide amniotique sont-ils corrélés linéairement au seul de 5% ?

Solution de l'exercice 12

Deux variables sont étudiés ici, à savoir :

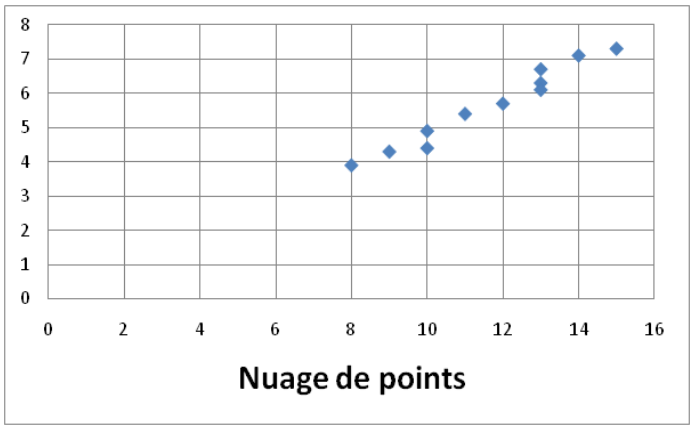
Taux X d'hormones lactogène dans le liquide amniotique et le taux Y d'hormones l'octogènes dans le sang (variables quantitatives)

Ces variables ont été observées sur 12 femmes enceintes (12 couples de données ou $n = 12$)

Problème posé :

Il s'agit donc d'un problème de corrélation, plus exactement : calcul du coefficient de corrélation r et comparaison à la valeur zéro.

a. Représentation du nuage de points associé aux 12 couples de mesures



b. Calcul du coefficient de corrélation linéaire r

Après avoir calculé $\bar{y}, \bar{y}, S(x), S(y)$, et $Cov(x, y)$ comme dans le premier chapitre, en obtient

$$\begin{aligned} r &= \frac{Cov(x,y)}{S(x)S(y)} \\ &= 0.975143 \approx 0.975. \end{aligned}$$

c. Comparaison du coefficient de corrélation à la valeur zéro

$$\begin{cases} H_0 : \rho = 0, \\ H_1 : \rho \neq 0, \end{cases}$$

La variable de décision T suit sous H_0 un loi de student à $n-1$ ddl et

$$\begin{aligned} s_c^2 &= \frac{1-r^2}{n-2} \\ &= \frac{1-(0.975)^2}{12-2} \\ &= 4.937\,5 \times 10^{-3}, \end{aligned}$$

on a

$$\begin{aligned} t_c &= \frac{r}{s_r} \\ &= \frac{0.975}{\sqrt{4.937\,5 \times 10^{-3}}} \\ &= 13\,88, \end{aligned}$$

Dans la table de student, on trouve $t_{n-2,\frac{\alpha}{2}} = t_{12-2,\frac{0.05}{2}} = 2.228$. Donc $t_c \notin [-2.228, 2.228]$, t_c étant à l'extérieur de l'intervalle de confiance à 95%, on rejette H_0 au seuil de 5% et même au seul de 0.1% (hors intervalle $[-4.587, 4.587]$).

Conclusion : au seuil de 0.1%, on peut dire que le taux d'hormone lactogène dans le sang est corrélé linéairement à celui dans le liquide amniotique.

Chapitre 5

Les tables statistiques

Les tables statistiques répertorient les valeurs d'une fonction de répartition ou d'une probabilité individuelle pour une variable aléatoire donnée suivant une loi de probabilité définie. Elles aident à trouver rapidement des valeurs exprimées par certaines fonctions, sans passer par des calculs complexes. En effet, certaines fonctions de répartition ne sont formulables sous une forme mathématique exploitable et dans d'autres cas, la définition d'une primitive de la fonction de densité est trop complexe pour calculer directement la fonction de répartition.

Il existe différentes tables statistiques pour différentes lois de probabilité, y compris la loi normale, la loi de Student, la loi de khi-deux, etc. Ces tables vous permettent de trouver des valeurs critiques à partir de statistiques calculées, ou de trouver des probabilités associées à certaines valeurs.

1-Table de la loi normale centrée réduite N(0.1)

La table qui apparaît à la page suivante nous permet de trouver la surface à gauche d'une valeur donnée sous la densité de la loi normale de moyenne 0 et de variance 1, aussi appelée la loi normale standard ou la loi normale centrée et réduite.

Exemple

1-On suppose que Z suit la loi N (0, 1) et on veut trouver $P(Z \leq 1.26)$.

Puisque 1.26 peut s'´ecrire sous la forme 1.26 = 1.20 + 0.06, on trouve

$P(Z \leq 1.26)$ `a l'intersection de la ligne « 1.2 » et de la colonne « 0.06 » de la table. On obtient $P(Z \leq 1.26) = \Phi(1.26) = 0.8962$.

Bref, la surface à gauche de 1.26 sous la densité de la loi N (0, 1) est égale `a 0.8962.

2- On suppose que Z suit la loi N (0, 1) et on veut trouver $P(Z \leq -0.94)$.

$P(Z \leq -0.94) = \Phi(-0.94)$

$= 1 - \Phi(0.94)$

$= 1 - 0.8264 = 0.1736$

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999

z	0.841	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.091	3.291
$\Phi(z)$	0.8000	0.9000	0.9500	0.9750	0.9800	0.9900	0.9950	0.9975	0.9990	0.9995

Table 1

FONCTION DE RÉPARTITION DE LA LOI NORMALE STANDARD $\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{\frac{-x^2}{2}} dx$

2-Table de la loi de Student

La table qui apparaît à la page suivante (Table2) nous donne certains quantiles de la loi de Student.

Exemple 1

Trouvons le quantile d'ordre **0.975** de la loi de Student avec 18 degrés de liberté. On pose 1-γ=0.975

On a donc $\boldsymbol{\gamma = 1 - 0.975 = 0.025}$. Dans la table, le quantile d'ordre **0.975** de la loi de Student avec 18 degrés de liberté se trouve donc à l'intersection de la ligne $\ll k = 18 \gg$ avec la colonne $\ll \gamma = 0.025 \gg$. On obtient la valeur 2.101. Ce quantile est habituellement dénoté $t_{18,0.025}$. On a donc $\mathbf{t_{18,0.025} = 2.101}$.

Exemple 2

Trouvons centile de la loi de Student avec 23 degrés de libertés, le 20^e

Il s'agit donc du quantile d'ordre 0.20. Ce quantile est souvent dénoté $t_{23,0.80}$. Puisque la loi de Student est symétrique par rapport à l'origine, on a $\mathbf{t_{23,0.80} = -t_{23,0.20}}$. La table nous donne $\mathbf{t_{23,0.20} = 0.858}$. On a donc t23,0.80 = −0.858. Le 20e centile de la loi de Student avec 23 degrés de liberté est donc égal à -0.858.

k	γ										
	0.25	0.20	0.15	0.10	0.05	0.025	0.010	0.005	0.0025	0.0010	0.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	127.3	318.3	636.6
2	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	14.09	22.33	31.60
3	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	7.453	10.21	12.92
4	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850
21	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.135	3.527	3.819
22	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.119	3.505	3.792
23	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.104	3.485	3.767
24	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.091	3.467	3.745
25	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.078	3.450	3.725
26	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.067	3.435	3.707
27	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.057	3.421	3.690
28	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.047	3.408	3.674
29	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.038	3.396	3.659
30	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646
40	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	2.971	3.307	3.551
50	0.679	0.849	1.047	1.299	1.676	2.009	2.403	2.678	2.937	3.261	3.496
60	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	2.915	3.232	3.460
80	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	2.887	3.195	3.416
100	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	2.871	3.174	3.390
120	0.677	0.845	1.041	1.289	1.658	1.980	2.358	2.617	2.860	3.160	3.373
∞	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	2.807	3.090	3.291

Table 1

Loi de Student avec k degrés de liberté quantiles d'ordre 1-γ

3. Table de la loi du khi-deux

La table qui apparaît à la page suivante (table3), nous donne certains quantiles de la loi du khi-deux.

Exemple. Trouvons le 99e centile de la loi du khi-deux avec 15 degrés de liberté. Il s'agit donc du quantile d'ordre 0.99. Ce quantile est souvent dénoté $\chi 2_{15,0.01}$. On trouve à l'intersection de la ligne $k = 15$ avec la colonne $\gamma = 0.01$, on obtient

$$\chi^2_{15,0.01} = 30.58$$

	γ										
k	0.995	0.990	0.975	0.950	0.900	0.500	0.100	0.050	0.025	0.010	0.005
1	0.00	0.00	0.00	0.00	0.02	0.45	2.71	3.84	5.02	6.63	7.88
2	0.01	0.02	0.05	0.10	0.21	1.39	4.61	5.99	7.38	9.21	10.60
3	0.07	0.11	0.22	0.35	0.58	2.37	6.25	7.81	9.35	11.34	12.84
4	0.21	0.30	0.48	0.71	1.06	3.36	7.78	9.94	11.14	13.28	14.86
5	0.41	0.55	0.83	1.15	1.61	4.35	9.24	11.07	12.83	15.09	16.75
6	0.68	0.87	1.24	1.64	2.20	5.35	10.65	12.59	14.45	16.81	18.55
7	0.99	1.24	1.69	2.17	2.83	6.35	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	7.34	13.36	15.51	17.53	20.09	21.96
9	1.73	2.09	2.70	3.33	4.17	8.34	14.68	16.92	19.02	21.67	23.59
10	2.16	2.56	3.25	3.94	4.87	9.34	15.99	18.31	20.48	23.21	25.19
11	2.60	3.05	3.82	4.57	5.58	10.34	17.28	19.68	21.92	24.72	26.76
12	3.07	3.57	4.40	5.23	6.30	11.34	18.55	21.03	23.34	26.22	28.30
13	3.57	4.11	5.01	5.89	7.04	12.34	19.81	22.36	24.74	27.69	29.82
14	4.07	4.66	5.63	6.57	7.79	13.34	21.06	23.68	26.12	29.14	31.32
15	4.60	5.23	6.27	7.26	8.55	14.34	22.31	25.00	27.49	30.58	32.80
16	5.14	5.81	6.91	7.96	9.31	15.34	23.54	26.30	28.85	32.00	34.27
17	5.70	6.41	7.56	8.67	10.09	16.34	24.77	27.59	30.19	33.41	35.72
18	6.26	7.01	8.23	9.39	10.87	17.34	25.99	28.87	31.53	34.81	37.16
19	6.84	7.63	8.81	10.12	11.65	18.34	27.20	30.14	32.85	36.19	38.58
20	7.43	8.26	9.59	10.85	12.44	19.34	28.41	31.41	34.17	37.57	40.00
21	8.03	8.90	10.28	11.59	13.24	20.34	29.62	32.67	35.48	38.93	41.40
22	8.64	9.54	10.98	12.34	14.04	21.34	30.81	33.92	36.78	40.29	42.80
23	9.26	10.20	11.69	13.09	14.85	22.34	32.01	35.17	38.08	41.64	44.18
24	9.89	10.86	12.40	13.85	15.66	23.34	33.20	36.42	39.36	42.98	45.56
25	10.52	11.52	13.12	14.61	16.47	24.34	34.28	37.65	40.65	44.31	46.93
26	11.16	12.20	13.84	15.38	17.29	25.34	35.56	38.89	41.92	45.64	48.29
27	11.81	12.88	14.57	16.15	18.11	26.34	36.74	40.11	43.19	46.96	49.65
28	12.46	13.57	15.31	16.93	18.94	27.34	37.92	41.34	44.46	48.28	50.99
29	13.12	14.26	16.05	17.71	19.77	28.34	39.09	42.56	45.72	49.59	52.34
30	13.79	14.95	16.79	18.49	20.60	29.34	40.26	43.77	46.98	50.89	53.67
40	20.71	22.16	24.43	26.51	29.05	39.34	51.81	55.76	59.34	63.69	66.77
50	27.99	29.71	32.36	34.76	37.69	49.33	63.17	67.50	71.42	76.15	79.49
60	35.53	37.48	40.48	43.19	46.46	59.33	74.40	79.08	83.30	88.38	91.95
70	43.28	45.44	48.76	51.74	55.33	69.33	85.53	90.53	95.02	100.42	104.22
80	51.17	53.54	57.15	60.39	64.28	79.33	96.58	101.88	106.63	112.33	116.32
90	59.20	61.75	65.65	69.13	73.29	89.33	107.57	113.14	118.14	124.12	128.30
100	67.33	70.06	74.22	77.93	82.36	99.33	118.50	124.34	129.56	135.81	140.17

Table 3

Loi du Khi-deux avec degrés de liberté quantiles d'ordre 1- γ

Si k est entre 30 et 100 mais n'est pas un multiple de 10, on utilise la table ci-haut et on fait une interpolation linéaire. Si k > 100 on peut, grâce au théorème limite central, approximer la loi $\chi 2(k)$ par la loi N (k, 2k).

4-Table de la loi de Fisher

La table qui apparaît dans les pages suivantes nous donne le 95^e centile de la loi de Fisher avec k degrés de liberté au numérateur et ℓ degrés de liberté au dénominateur. Ce quantile est dénoté $F^k_{\ell,0.05}$.

Exemple. Quel est le 95^e centile de la loi de Fisher avec 10 degrés de liberté au numérateur et 15 degrés de liberté au dénominateur?

Ce quantile est dénoté $F^{10}_{15,0.05}$. On le trouve à l'intersection de la ligne $\ell = 15$ avec la colonne k = 10. On obtient $F^{10}_{15,0.05} = 2.544$.

	1	2	3	4	5	6	7	8	9	10
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40
3	10.13	9.552	9.277	9.117	9.013	8.941	8.887	8.845	8.812	8.786
4	7.709	6.944	6.591	6.388	6.256	6.163	6.094	6.041	5.999	5.964
5	6.608	5.786	5.409	5.192	5.050	4.950	4.876	4.818	4.772	4.735
6	5.987	5.143	4.757	4.534	4.387	4.284	4.207	4.147	4.099	4.060
7	5.591	4.737	4.347	4.120	3.972	3.866	3.787	3.726	3.677	3.637
8	5.318	4.459	4.066	3.838	3.687	3.581	3.500	3.438	3.388	3.347
9	5.117	4.256	3.863	3.633	3.482	3.374	3.293	3.230	3.179	3.137
10	4.965	4.103	3.708	3.478	3.326	3.217	3.135	3.072	3.020	2.978
11	4.844	3.982	3.587	3.357	3.204	3.095	3.012	2.948	2.896	2.854
12	4.747	3.885	3.490	3.259	3.106	2.996	2.913	2.849	2.796	2.753
13	4.667	3.806	3.411	3.179	3.025	2.915	2.832	2.767	2.714	2.671
14	4.600	3.739	3.344	3.112	2.958	2.848	2.764	2.699	2.646	2.602
15	4.543	3.682	3.287	3.056	2.901	2.790	2.707	2.641	2.588	2.544
16	4.494	3.634	3.239	3.007	2.852	2.741	2.657	2.591	2.538	2.494
17	4.451	3.592	3.197	2.965	2.810	2.699	2.614	2.548	2.494	2.450
18	4.414	3.555	3.160	2.928	2.773	2.661	2.577	2.510	2.456	2.412
19	4.381	3.522	3.127	2.895	2.740	2.628	2.544	2.477	2.423	2.378
20	4.351	3.493	3.098	2.866	2.711	2.599	2.514	2.447	2.393	2.348
21	4.325	3.467	3.072	2.840	2.685	2.573	2.488	2.420	2.366	2.321
22	4.301	3.443	3.049	2.817	2.661	2.549	2.464	2.397	2.342	2.297
23	4.279	3.422	3.028	2.796	2.640	2.528	2.442	2.375	2.320	2.275
24	4.260	3.403	3.009	2.776	2.621	2.508	2.423	2.355	2.300	2.255
25	4.242	3.385	2.991	2.759	2.603	2.490	2.405	2.337	2.282	2.236
26	4.225	3.369	2.975	2.743	2.587	2.474	2.388	2.321	2.265	2.220
27	4.210	3.354	2.960	2.728	2.572	2.459	2.373	2.305	2.250	2.204
28	4.196	3.340	2.947	2.714	2.558	2.445	2.359	2.291	2.236	2.190
29	4.183	3.328	2.934	2.701	2.545	2.432	2.346	2.278	2.223	2.177
30	4.171	3.316	2.922	2.690	2.534	2.421	2.334	2.266	2.211	2.165
40	4.085	3.232	2.839	2.606	2.449	2.336	2.249	2.180	2.124	2.077
50	4.034	3.183	2.790	2.557	2.400	2.286	2.199	2.130	2.073	2.026
60	4.001	3.150	2.758	2.525	2.368	2.254	2.167	2.097	2.040	1.993
70	3.978	3.128	2.736	2.503	2.346	2.231	2.143	2.074	2.017	1.969
80	3.960	3.111	2.719	2.486	2.329	2.214	2.126	2.056	1.999	1.951
90	3.947	3.098	2.706	2.473	2.316	2.201	2.113	2.043	1.986	1.938
100	3.936	3.087	2.696	2.463	2.305	2.191	2.103	2.032	1.975	1.927
150	3.904	3.056	2.665	2.432	2.274	2.160	2.071	2.001	1.943	1.894
200	3.888	3.041	2.650	2.417	2.259	2.144	2.056	1.985	1.927	1.878
400	3.865	3.018	2.627	2.394	2.237	2.121	2.032	1.962	1.903	1.854

Table 4

Quantiles d'ordre 0.95 de la loi de Fisher

Degrés de liberté du numérateur sur la première ligne
Degrés de liberté du dénominateur sur la colonne de gauche

Bibliographie

[1] BENAISSA. A., Biostatistique-informatique, 1ère Année Médecine, *Faculté de Médecine Université Batna 2 Algeria.* 2021.

[2] Bernard.Y., Méthodes Statistiques pour la Biologie ; *Université Joseph Fourier, Grenoble I.*

[3] Carrat.F, Mallet. A, Morice .V., Biostatistique, *Université Pierre et Marie Curie.* $2013 - 2014$.

[4] Claude. B., STT-1920 Méthodes statistiques, Département de mathématiques et de statistique. Université Laval. 2011.

[5] Dagnelie. P., Statistique théorique et appliquée, *Tome* 1 *et* 2. *Ed, Université Larcier et De-Boeck, Belgique.*2009.

[6] Gaetan. M., Biostatistique, *Tome* 1 *et* 2. $2^{ème}$ *Ed. Scherrer, Canada.B.* 2009

[7] Gilbert. D., Probabilités statistique inférentielle fiabilité, *Premier cycle, IUP, Prépa, BTS, IUT.* 2007.

[8] Harvey.J., Biostate, Une approche intuitive. *Ed. Univ. De Boeck et Larcier ; Motulsky. .Belgique.*1995.

[9] Huguier.M ; Biostatistique au quotidien ; *Ed. Elsevier.A.* 2003.

[10] Jean-Christophe. B., Statistiques, *Université de La Rochelle. Octobre-Novembre* 2008.

[11] Jean-Jacques. R., Statistique : Estimation, *Préparation à l'Agrégation Bordeaux 1. Année* $2012 - 2013$

[12] Khalidi. K., Méthodes statistiques ; *Rappels et cours, Office des publication universitaires 1, Place centrale de Ben-Aknoun (Alger).*1998.

[13] Menaceur. A., Polycopié de cours :Biostatistiques, 3ème Année Licence IMM, *Université 8 Mai 1945 – Guelma. Algeria.* 2017.

[14] Nakache.J.P ; Statistique explicative appliquée, *Ed. Technip, France.J.* 2003.

[15] Nemiche. M., Exercices Corrigés Statistique et Probabilités, *Faculté des Sciences d'Agadir (STU 3).* 2015.

[16] Université de Batna2, Faculté de médecine, 1 ère année médecine, Corrigé-type de TD de Bio-statistiques 2019/2020.