

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/313841800>

Cours 2 de Statistique pour STU

Book · January 2017

CITATIONS

0

READS

7,903

1 author:



Sghir Aissa

Université Mohammed Premier

29 PUBLICATIONS 14 CITATIONS

[SEE PROFILE](#)

Some of the authors of this publication are also working on these related projects:



sghir aissa [View project](#)

Cours et exercices de Statistique pour STU-S₍₃₎

SGHIR AISSA
sghir.aissa@gmail.com

Faculté des Sciences Meknès



Plan :

- 1 Introduction
- 2 Statistique descriptive univariée
- 3 Statistique descriptive bivariée
- 4 Notions de probabilités et variables aléatoires
- 5 Échantillonnage et estimation
- 6 Tests des hypothèses
- 7 Exercices de révisions
- 8 Examens (sessions normale et de rattrapage 2016-2017)

Introduction

Statistique descriptive univariée

Statistique descriptive bivariée

Notions de probabilités et variables aléatoires

Échantillonnage et estimation

Tests des hypothèses

Exercices de révisions

Examens (sessions normale et de rattrapage 2016-2017)

Plan :

- 1 Introduction
- 2 Statistique descriptive univariée
- 3 Statistique descriptive bivariée
- 4 Notions de probabilités et variables aléatoires
- 5 Échantillonnage et estimation
- 6 Tests des hypothèses
- 7 Exercices de révisions
- 8 Examens (sessions normale et de rattrapage 2016-2017)

Introduction

Statistique descriptive univariée

Statistique descriptive bivariée

Notions de probabilités et variables aléatoires

Échantillonnage et estimation

Tests des hypothèses

Exercices de révisions

Examens (sessions normale et de rattrapage 2016-2017)

Introduction

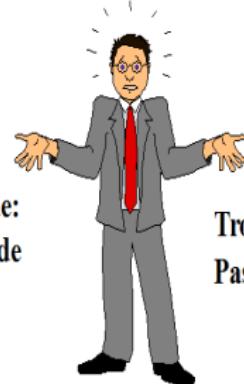
Introduction

Population Échantillon aléatoire



Taille de l'échantillon?

Trop grande:
Exige trop de
ressources



Trop petite:
Pas assez précis

Introduction

Statistique descriptive univariée

Statistique descriptive bivariée

Notions de probabilités et variables aléatoires

Échantillonnage et estimation

Tests des hypothèses

Exercices de révisions

Examens (sessions normale et de rattrapage 2016-2017)

Introduction

Dans cet exemple et dans d'autres domaines, (biologie, géologie, physique, chimie, finance, ...), les managers doivent pouvoir disposer d'outils performants d'aide à la décision et l'analyse des informations recueillis.

Introduction

Statistique descriptive univariée

Statistique descriptive bivariée

Notions de probabilités et variables aléatoires

Échantillonnage et estimation

Tests des hypothèses

Exercices de révisions

Examens (sessions normale et de rattrapage 2016-2017)

Introduction

La statistique s'inscrit dans cette perspective et dont la définition est la suivante :

La statistique est un ensemble de méthodes scientifiques dont l'objectif est d'analyser, structurer et modéliser des informations numériques.

Introduction

Statistique descriptive univariée

Statistique descriptive bivariée

Notions de probabilités et variables aléatoires

Échantillonnage et estimation

Tests des hypothèses

Exercices de révisions

Examens (sessions normale et de rattrapage 2016-2017)

Introduction

Les méthodes statistiques peuvent être classés en deux groupes :

1) Les Statistiques descriptives

Elle regroupe les méthodes dont l'objectif principal est la description des données étudiées. Cette description des données se fait à travers leur représentation graphique, et le calcul de résumés numériques. Dans cette optique, on ne fait pas appel à des outils de type probabiliste.

Introduction

Statistique descriptive univariée

Statistique descriptive bivariée

Notions de probabilités et variables aléatoires

Échantillonnage et estimation

Tests des hypothèses

Exercices de révisions

Examens (sessions normale et de rattrapage 2016-2017)

Introduction

On cite trois types des statistiques descriptives :

Statistique descriptive univariée : étude de la population selon une seule variable.

Statistique descriptive bivariée : étude des corrélations et relations éventuelles entre deux variables de la même population.

Statistique descriptive multivariée : étude des relations éventuelles entre plusieurs variables de la même population.

Introduction

Statistique descriptive univariée

Statistique descriptive bivariée

Notions de probabilités et variables aléatoires

Échantillonnage et estimation

Tests des hypothèses

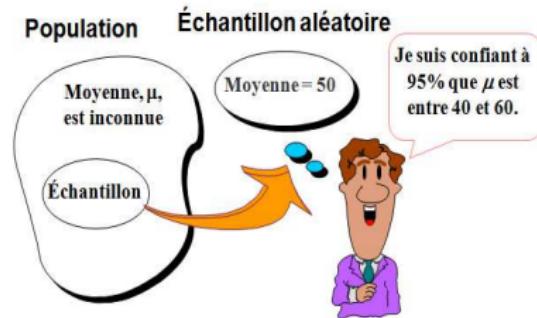
Exercices de révisions

Examens (sessions normale et de rattrapage 2016-2017)

Introduction

2) La statistique inférentielle

Ce terme regroupe les méthodes dont l'objectif principal est de préciser un phénomène sur une population globale, à partir de son observation sur un échantillon de cette population. Ce passage ne se fait que moyennant des hypothèses de type probabiliste.



Introduction

Statistique descriptive univariée

Statistique descriptive bivariée

Notions de probabilités et variables aléatoires

Échantillonnage et estimation

Tests des hypothèses

Exercices de révisions

Examens (sessions normale et de rattrapage 2016-2017)

Introduction

Remarque

La statistique descriptive précède en général la statistique inférentielle dans une démarche de traitement de données : les deux aspects de la statistique se complètent bien plus qu'ils ne s'opposent.

Plan :

- 1 Introduction
- 2 Statistique descriptive univariée
- 3 Statistique descriptive bivariée
- 4 Notions de probabilités et variables aléatoires
- 5 Échantillonnage et estimation
- 6 Tests des hypothèses
- 7 Exercices de révisions
- 8 Examens (sessions normale et de rattrapage 2016-2017)

Vocabulaires

Vocabulaires

Population : ensemble des individus objets de l'étude.
(Étudiants, entreprises, plantes, animaux, produits,...)

Échantillon : sous-ensemble issu de la population.
(Une classe, une ville, hommes, femmes,...)

Unité statistique : chaque individu.
(Un étudiant, une plante, un homme, une femme,...)

Vocabulaires

Variable : caractère ou propriété mesuré sur chaque individu notée X, Y, \dots
(Note, taille, poids, sex, couleur,...)

Modalités : les valeurs possibles de la variable.

Série statistique : suite des valeurs prises par une variable X notées
(x_1, x_2, x_3, \dots).

Vocabulaires

Les variables sont classées en deux types :

Variable quantitative : les modalités sont mesurables ou repérables.

- **Variable quantitative discrète** : l'ensemble des modalités est fini ou dénombrable : (Note, taille, poids, âge, mesure,...)
- **Variable quantitative continue** : l'ensemble des modalités est un intervalle fini ou infini : ([8; 20[, [0; +∞[, ℝ,...])

Vocabulaires

Variable qualitative : les modalités ne sont pas mesurables.

- **Variable qualitative nominale** : les modalités ne peuvent pas être ordonnées : (sex, couleur,...)
- **Variable qualitative ordinale** : les modalités peuvent être ordonnées : (taille d'un vêtement : XXL, XL, L, M, S).

Vocabulaires

Effectif totale n : le nombre de toutes les valeurs prises par la variable.

Effectif n_i : nombre d'apparitions de la valeur x_i dans la population ou dans l'échantillon.

$$\sum_{i=1}^J n_i = n_1 + n_2 + \dots + n_J = n.$$

Vocabulaires

Fréquence f_i associée à la valeur x_i

$$\begin{cases} f_i = \frac{n_i}{n}, \\ \sum_{i=1}^J f_i = f_1 + f_2 + \dots + f_J = 1. \end{cases}$$

Pourcentage p_i associé à la valeur x_i

$$\begin{cases} p_i = 100 \times f_i \%, \\ \sum_{i=1}^J p_i = p_1 + p_2 + \dots + p_J = 100 %. \end{cases}$$

Vocabulaires

Effectif cumulé N_i

$$\left\{ \begin{array}{l} N_1 = n_1, \\ N_2 = n_1 + n_2, \\ \dots \\ N_J = n_1 + n_2 + \dots + n_J = n. \end{array} \right.$$

Fréquence cumulée F_i

$$\left\{ \begin{array}{l} F_1 = f_1, \\ F_2 = f_1 + f_2, \\ \dots \\ F_J = f_1 + f_2 + \dots + f_J = 1. \end{array} \right.$$

Exemples

Remarque

Avant de citer les exemples de cette section, nous présentons un exemple d'un modèle de questionnaire pour la collection des informations sur une population.

Type de logement	appartement	<input type="checkbox"/>
	maison individuelle	<input type="checkbox"/>
	autres	<input type="checkbox"/>
Surface habitable		<input type="text"/>
Nombre de pièces d'habitation		<input type="text"/>
Y a-t-il une cuisine ?	oui, privée	<input type="checkbox"/>
	oui, commune	<input type="checkbox"/>
	non	<input type="checkbox"/>
Salle de bains ou douche ?	oui, privée	<input type="checkbox"/>
	oui, commune	<input type="checkbox"/>
	non	<input type="checkbox"/>

Exemples

Variable qualitative nominale

On s'intéresse à la variable $X = \text{état-civil}$ sur une population de $n = 20$ personnes. Considérons la série statistique suivante avec **C** : célibataire, **M** : marié, **V** : veuf, **D** : divorcé.

M D M C C M C C C M C M V M V D C C M C

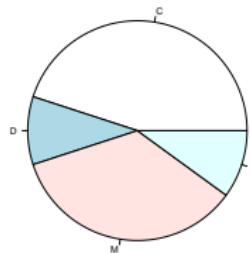
Tableau statistique

x_i	n_i	f_i	$p_i\%$	N_i	F_i
C	9	0.45	45	9	0.45
M	7	0.35	35	16	0.75
V	2	0.10	10	18	0.85
D	2	0.10	10	20	1

Exemples

Diagramme en secteurs

x_i	$p_i \%$	$d_i = p_i \times 3.6^\circ$
C	45	162
M	35	126
V	10	36
D	10	36



Exemples

Variable qualitative ordinale

On interroge une population de $n = 50$ personnes sur leur dernier diplôme :

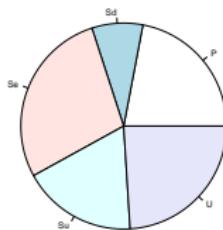
Sd : Sans diplôme, **P** : Primaire, **Se** : Secondaire, **Su** : Supérieur non-universitaire et **U** : Universitaire.

Sd	Sd	Sd	Sd	P	P	P	P	P	P	P	P	P	P	P	P	Se	Se	Se
Se	Su	Su	Su															
Su	Su	Su	Su	U	U	U	U	U	U	U	U	U	U	U	U	U	U	

Exemples

Tableau statistique

x_i	n_i	N_i	f_i	p_i	F_i
Sd	4	4	0.08	8	0.08
P	11	15	0.22	22	0.30
Se	14	29	0.28	28	0.58
Su	9	38	0.18	18	0.76
U	12	50	0.24	24	1



Exemples

Variable quantitative discrète

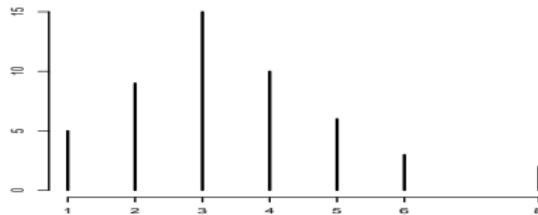
Un quartier est composé d'une population de 50 ménages, et la variable X représente le nombre de personnes par ménage. Les valeurs de la variable sont :

1	1	1	1	1	1	2	2	2	2	2
2	2	2	2	2	3	3	3	3	3	3
3	3	3	3	3	3	3	3	3	3	4
4	4	4	4	4	4	4	4	4	4	5
5	5	5	5	5	5	6	6	6	8	8

Exemples

Diagramme en bâtonnets des effectifs

x_i	n_i	N_i	f_i	F_i
1	5	5	0.10	0.10
2	9	14	0.18	0.28
3	15	29	0.30	0.58
4	10	39	0.20	0.78
5	6	45	0.12	0.90
6	3	48	0.06	0.96
8	2	50	0.04	1

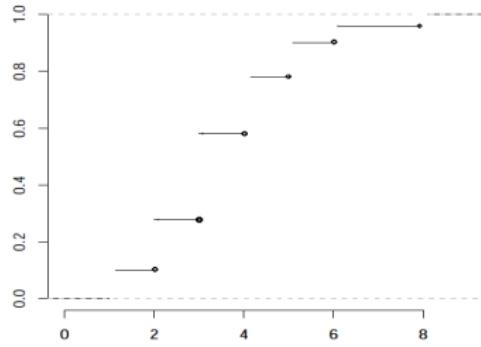


Exemples

Fonction de répartition

Les fréquences cumulées sont représentées au moyen de la fonction de répartition. Cette fonction est définie de \mathbb{R} dans $[0, 1]$ et vaut :

$$F(x) = \begin{cases} 0 & x < x_1 \\ F_j & x_j \leq x < x_{j+1} \\ 1 & x_J \leq x. \end{cases}$$



Exemples

Variable quantitative continue

Très souvent, la prise en compte de toute les valeurs observées ne permet pas de donner une interprétation simple des résultats et conduit à des calculs inutiles. On peut souvent se contenter de regarder des regroupements en classes.

Exemples

Exemple

On mesure la variable $X=\text{taille en centimètre}$ d'une population de 50 élèves d'une classe.

152	152	152	153	153
154	154	154	155	155
156	156	156	156	156
157	157	157	158	158
159	159	160	160	160
161	160	160	161	162
162	162	163	164	164
164	164	165	166	167
168	168	168	169	169
170	171	171	171	171

Exemples

Tableau statistique

On va procéder à des regroupements en classes (intervalles) de même amplitude. En règle générale, on choisit au moins cinq classes, sinon on utilise la règle de Sturge : le nombre de classes est $J = 1 + (3.3 \times \log_{10}(n))$.

La longueur de chaque classe est $l = (x_{\max} - x_{\min})/J$.

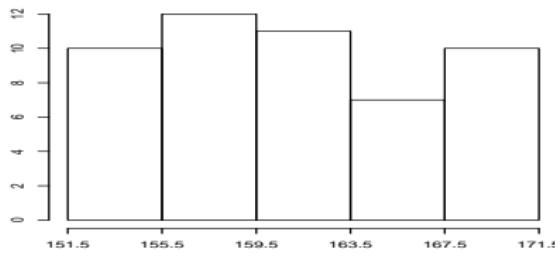
Par exemple pour $J = 5$, $x_{\max} = 171$ et $x_{\min} = 152$, on prend $l \simeq 4$.

classe	n_i	N_i	f_i	F_i
[151.5 ; 155.5[10	10	0.20	0.20
[155.5 ; 159.5[12	22	0.24	0.44
[159.5 ; 163.5[11	33	0.22	0.66
[163.5 ; 167.5[7	40	0.14	0.80
[167.5 ; 171.5[10	50	0.20	1

Variable quantitative continue

Histogramme des effectifs

classe	n_i	N_i	f_i	F_i
[151.5 ; 155.5[10	10	0.20	0.20
[155.5 ; 159.5[12	22	0.24	0.44
[159.5 ; 163.5[11	33	0.22	0.66
[163.5 ; 167.5[7	40	0.14	0.80
[167.5 ; 171.5[10	50	0.20	1



Exemples

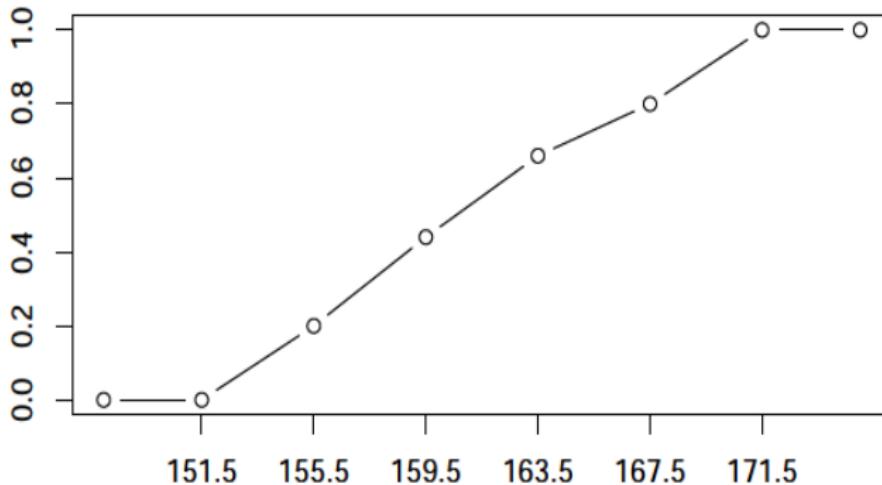
Fonction de répartition

Si $[c_j^-; c_j^+]$ désigne la classe j , on note, de manière générale :

- c_j^- la borne inférieure de la classe j ,
- c_j^+ la borne supérieure de la classe j ,
- $c_j = (c_j^+ + c_j^-)/2$ le centre de la classe j ,
- $a_j = c_j^+ - c_j^-$ l'amplitude de la classe j ,
- n_j l'effectif de la classe j ,
- N_j l'effectif cumulé de la classe j ,
- f_j la fréquence de la classe j ,
- F_j la fréquence cumulée de la classe j .

$$F(x) = \begin{cases} 0 & x < c_1^- \\ F_{j-1} + \frac{f_j}{c_j^+ - c_j^-}(x - c_j^-) & c_j^- \leq x < c_j^+ \\ 1 & c_j^+ \leq x \end{cases}$$

Exemples



Paramètres de position

Le mode ou classe modale

C'est la valeur ou classe correspondant à l'effectif (ou fréquence) le plus élevé.

Exemple 1

x_i	n_i	f_i
C	9	0.45
M	7	0.35
V	2	0.10
D	2	0.10

le mode est $x_1 = C$: célibataire correspondant à l'effectif $n_1 = 9$ ou la fréquence $f_1 = 0.45$.

Paramètres de position

Exemple 2

classe	n_i	N_i	f_i	F_i
[151.5 ; 155.5[10	10	0.20	0.20
[155.5 ; 159.5[12	22	0.24	0.44
[159.5 ; 163.5[11	33	0.22	0.66
[163.5 ; 167.5[7	40	0.14	0.80
[167.5 ; 171.5[10	50	0.20	1

la classe modale est [155.5; 159.5[.

Paramètres de position

La moyenne

La moyenne \bar{x} ne peut être définie que sur une variable quantitative.

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{x_1 + \dots + x_n}{n}.$$

Exemple

Les nombres d'enfants de 8 familles sont les suivants 0, 0, 1, 1, 1, 2, 3, 4. La moyenne est

$$\bar{x} = \frac{0 + 0 + 1 + 1 + 1 + 2 + 3 + 4}{8} = 1.5.$$

Paramètres de position

La moyenne peut être calculée à partir des valeurs distinctes et des effectifs.

$$\bar{x} = \frac{1}{n} \sum_{i=1}^J n_i \times x_i = \frac{n_1 \times x_1 + \dots + n_J \times x_J}{n}.$$

Exemple

Les nombres d'enfants de 8 familles sont les suivants 0, 0, 1, 1, 1, 2, 3, 4. La moyenne est

$$\bar{x} = \frac{2 \times 0 + 3 \times 1 + 1 \times 2 + 1 \times 3 + 1 \times 4}{8} = 1.5.$$

Paramètres de position

La médiane

Cas d'une variable quantitative discrète

La médiane, notée $x_{\frac{n}{2}}$, est une valeur centrale de la série statistique qui la partage en deux groupes de même effectifs. Elle est obtenue de la manière suivante :

On trie la série statistique par ordre croissant des valeurs observées :

Par exemple, avec la série observée :

$$3 \quad 2 \quad 1 \quad 0 \quad 0 \quad 1 \quad 2,$$

on obtient :

$$0 \quad 0 \quad 1 \quad 1 \quad 2 \quad 2 \quad 3.$$

$n = 7$ est impair, alors la médiane est la valeur du rang $(n+1)/2 = 4$.

Donc $x_{\frac{1}{2}} = 1$.

Paramètres de position

Si n est pair, alors la médiane est la moyenne des deux valeurs de rang $n/2$ et $(n/2) + 1$.

Exemple

Pour $n = 8$, si on a :

0 0 1 1 2 2 3 4

alors

$$x_{\frac{1}{2}} = \frac{1+2}{2} = 1.5.$$

Paramètres de position

La médiane

Cas d'une variable quantitative continue

De manière générale, on définira la médiane comme étant la valeur (abscisse) correspondant à la fréquence cumulée $F = 0.5$ ou effectif cumulé $N = \frac{n}{2}$.

On l'obtiendra en général par lecture graphique (valeur approchée

$x_{\frac{1}{2}} = F^{-1}(0.5)$) sur la courbe des fréquences cumulées, ou par une formule d'interpolation linéaire (valeur exacte) sur la courbe des effectifs cumulées.

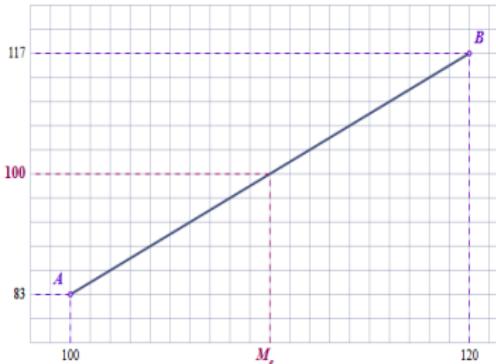
Paramètres de position

Cas d'une variable quantitative continue

Budget (en milliers d'euros)	[0; 50[[50; 70[[70; 100[[100; 120[[120; 150[[150; 200[[200; 240[
Nombre d'entreprises	10	28	45	34	42	33	8
Effectifs cumulés	10	38	83	117	159	192	200

Il y a 200 entreprises, donc la classe médiane est la classe [100; 120[.

Calculons la médiane par interpolation linéaire :



Paramètres de position

La médiane est l'abscisse du point de la droite passant par les points $A(100; 83)$ et $B(120; 117)$ dont l'ordonnée est égale à 100.

La droite (AB) a pour équation, $y = mx + p$ avec

$$m = \frac{117 - 83}{120 - 100} = 1,7$$

Comme $A(100; 83)$ est un point de la droite (AB) , nous obtenons la relation

$$y = 1,7(x - 100) + 83 \Leftrightarrow y = 1,7x - 87$$

La médiane est la valeur associée à un effectif égal à 100 d'où x est solution de l'équation

$$1,7x - 87 = 100 \Leftrightarrow x = \frac{100 + 87}{1,7} = 110$$

Le budget médian est de 110 milliers d'euros.

Paramètres de dispersion

L'étendue

L'étendue est défini par :

$$E = x_{\max} - x_{\min}.$$

Exemple

Pour la série 1 1 2 1 1 3 5 5 5 5 3 2 5
on a

$$E = 5 - 1 = 4.$$

Paramètres de dispersion

La variance et l'écart type

$$s_X^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2, \quad s_X = \sqrt{s_X^2}.$$

Exemple

Soit la série statistique 2 3 4 4 5 6 7 9 de taille 8. On a

$$\bar{x} = \frac{2 + 3 + 4 + 4 + 5 + 6 + 7 + 9}{8} = 5.$$

$$s_X^2 = \frac{(2 - 5)^2 + (3 - 5)^2 + (4 - 5)^2 + \dots + (9 - 5)^2}{8} = 4.5$$

$$s_X = \sqrt{s_X^2} = \sqrt{4.5} = 2.12$$

Paramètres de dispersion

La variance peut aussi s'écrire :

$$s_X^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2.$$

Exemple

Soit la série statistique 2 3 4 4 5 6 7 9 de taille 8. On a

$$\bar{x} = \frac{2 + 3 + 4 + 4 + 5 + 6 + 7 + 9}{8} = 5.$$

$$s_X^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2 = \frac{2^2 + 3^2 + 4^2 + 4^2 + 5^2 + 6^2 + 7^2 + 9^2}{8} - 5^2 = 4.5.$$

Paramètres de dispersion

Remarque

La variance peut aussi s'écrire avec les effectifs :

$$s_x^2 = \frac{1}{n} \sum_{i=1}^J n_i \times (x_i - \bar{x})^2$$

$$s_x^2 = \frac{1}{n} \sum_{i=1}^J n_i \times x_i^2 - \bar{x}^2.$$

Paramètres de dispersion

Remarque

Pour calculer la moyenne et la variance dans le cas d'une variable continue, on calcule les centres des classes qui vont jouer le rôle des valeurs x_i du cas discret.

Exemple

classe	n_i	centre x_i
[0; 10[10	$\frac{0+10}{2} = 5$
[10; 20[4	15
[20; 30[20	25
[30; 40[6	35

$$\bar{x} = \frac{10 \times 5 + 4 \times 15 + 20 \times 25 + 6 \times 35}{40} = 20.5.$$

Utilisation d'une calculatrice Casio fx 82MS et fx 82ES



* On passe en mode statistique (SD s'affiche en haut de l'écran)

* On efface les mémoires statistiques

Shift EXE (Sel : stat clear)

Exemple :

Note (x_i)	Effectif (n_j)
12	1
15	4

→ 12 ; 1 M+ (n = 1)
 → 15 ; 4 M+ (n = 5)

Utilisation d'une calculatrice Casio fx 82ms et fx 92

Exemple :

Note (x_i)	Effectif (n_i)
12	1
15	4

→ (n = 1)
 → (n = 5)

* Pour déterminer l'effectif total

⇒ On trouve : N = 5

SECONDE

* Pour déterminer la moyenne

⇒ On trouve la note moyenne : $\bar{x} = 14,4$

SECONDE

* Pour déterminer l'écart type

⇒ On trouve l'écart type : $\sigma = 1,2$

SECONDE

* Pour déterminer $\Sigma n_i x_i$

⇒ On trouve : $\Sigma n_i x_i = 72$

SECONDE

* Pour déterminer $\Sigma n_i x_i^2$

⇒ On trouve : $\Sigma n_i x_i^2 = 1044$

SECONDE

Utilisation d'une calculatrice Casio fx 82ES et fx 92

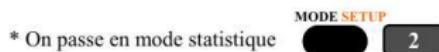


En premier, vérifier le paramétrage de la calculatrice :



(Menu SETUP-STAT-ON : On montre la colonne des fréquences)

Remarque : Pour masquer cette colonne taper « 2 » au lieu de « 1 ».



Les mémoires statistiques sont automatiquement effacées quand on entre dans la mode « STAT »

* On passe en mode statistique 1

* On choisit le type : 1 (1-VAR)
 * On saisit les données. **Normalement l'éditeur STAT est affiché.**
 S'il ne l'est pas faire : SHIFT STAT 1 2 (Stat-Data)

Exemple :

Note (x_i)	Effectif (n_j)
12	1
15	4



	STAT	X	Freq
1	12	1	
2	15	4	

Appuyer sur EXE
 Pour enregistrer
 chaque donnée.
 Se déplacer dans le
 tableau avec le curseur
 Enfin appuyer sur AC



Utilisation d'une calculatrice Casio fx 82ES et fx 92

Exemple :

Note (x_i)	Effectif (n_i)
12	1
15	4



	STAT	X	Freq
1		12	1
2		15	4

Appuyer sur **EXE**
 Pour enregistrer
 chaque donnée.
 Se déplacer dans le
 tableau avec le curseur

- * Pour déterminer l'effectif total : STAT-Var-n
 ↳ On trouve : N = 5



- * Pour déterminer la moyenne : STAT-Var- \bar{x}
 ↳ On trouve la note moyenne : $\bar{x} = 14,4$



- * Pour déterminer l'écart type : STAT-Var-xσn
 ↳ On trouve l'écart type : $\sigma = 1,2$



- * Pour déterminer $\Sigma n_i x_i$: STAT-Sum-Σx
 ↳ On trouve : $\Sigma n_i x_i = 72$



- * Pour déterminer $\Sigma n_i x_i^2$: STAT-Sum-Σx²
 ↳ On trouve : $\Sigma n_i x_i^2 = 1044$



Exercices

Exercice 1 :

On donne les couleurs de $n = 15$ plantes.

V V R N R R V R R R J N N N N

1. De quel type est la variable couleur des plantes ?
2. Construire le tableau statistique des effectifs et pourcentages cumulés.
3. Déterminer le mode.
4. Construire le diagramme en secteurs.

Exercices

Exercice 2 :

Trente éprouvettes d'acier spécial sont soumises à des essais de résistance. Pour chacune, on note le nombre de chocs nécessaires pour obtenir la rupture. Les résultats obtenus sont les suivants :

2	2	3	1	2	1	4	2	3	2
3	2	3	3	4	1	1	4	2	3
2	3	2	2	3	4	3	2	3	2

1. De quel type est cette variable ?
2. Construire le tableau statistique.
3. Construire le diagramme en bâtonnets des fréquences.
4. Déterminer la médiane, la moyenne, la variance et l'écart type de cette variable.
5. Déterminer la fonction de répartition et tracer sa courbe.

Exercice

Exercice 3 :

On pèse les $n = 50$ élèves d'une classe et nous obtenons les résultats résumés dans le tableau suivant :

43	43	43	47	48	48	48	48	49	49
49	50	50	51	51	52	53	53	53	54
54	56	56	56	57	59	59	59	62	62
63	63	65	65	67	67	68	70	70	70
72	72	73	77	77	81	83	86	92	93

1. De quel type est la variable poids ?
2. Construire le tableau statistique en adoptant les classes suivantes : $[40; 45], [45; 50], [50; 55], [55; 60], [60; 65], [65; 70], [70; 80], [80; 100]$.
3. Construire l'histogramme des fréquences.
4. Déterminer la fonction de répartition et tracer sa courbe.
5. Déterminer la médiane directement et par interpolation linéaire.
6. Déterminer la moyenne, la variance et l'écart type de la variable poids.



Plan :

- 1 Introduction
- 2 Statistique descriptive univariée
- 3 Statistique descriptive bivariée
- 4 Notions de probabilités et variables aléatoires
- 5 Échantillonnage et estimation
- 6 Tests des hypothèses
- 7 Exercices de révisions
- 8 Examens (sessions normale et de rattrapage 2016-2017)

Introduction

Introduction

L'objectif de cette partie est d'étudier sur une même population de n individus, deux caractères différents X et Y et de rechercher s'il existe un lien entre ces deux variables. Chacune des deux variables peut être, soit quantitative, soit qualitative.

Introduction

La série statistique est alors une suite de n couples des valeurs prises par les deux variables sur chaque individu : $(x_1, y_1), \dots, (x_n, y_n)$.

**L'effectif associe à l'observation (x_i, y_j) est noté n_{ij} .
et sa fréquence notée :**

$$f_{ij} = \frac{n_{ij}}{n}.$$

Les résultats sont regroupés dans un tableau appelé tableau de contingence.

Tableau de contingence

Tableau de contingence des effectifs

On s'intéresse à une éventuelle relation entre X : le sexe de $n = 200$ personnes et Y : la couleur des yeux.

X/Y	Bleu	Vert	Marron	Total
Homme	$n_{11} = 10$	$n_{12} = 50$	$n_{13} = 20$	$n_{1\bullet} = 80$
Femme	$n_{21} = 20$	$n_{22} = 60$	$n_{23} = 40$	$n_{2\bullet} = 120$
Total	$n_{\bullet 1} = 30$	$n_{\bullet 2} = 110$	$n_{\bullet 3} = 60$	$n = 200$

$n_{1\bullet}, n_{2\bullet}$ et $n_{\bullet 1}, n_{\bullet 2}, n_{\bullet 3}$ sont appelés effectifs marginaux.

$$\left\{ \begin{array}{l} n_{11} + n_{12} + n_{13} = n_{1\bullet}, \\ n_{21} + n_{22} + n_{23} = n_{2\bullet}, \\ n_{11} + n_{21} = n_{\bullet 1}, \\ n_{12} + n_{22} = n_{\bullet 2}, \\ n_{13} + n_{23} = n_{\bullet 3}, \\ n_{11} + n_{12} + n_{13} + n_{21} + n_{22} + n_{23} = n. \end{array} \right.$$

Tableau de contingence

Tableau de contingence des fréquences

X/Y	Bleu	Vert	Marron	Total
Homme	$f_{11} = 0.05$	$f_{12} = 0.25$	$f_{13} = 0.10$	$f_{1\bullet} = 0.40$
Femme	$f_{21} = 0.10$	$f_{22} = 0.30$	$f_{23} = 0.20$	$f_{2\bullet} = 0.60$
Total	$f_{\bullet 1} = 0.15$	$f_{\bullet 2} = 0.55$	$f_{\bullet 3} = 0.30$	1

$f_{1\bullet}, f_{2\bullet}$ et $f_{\bullet 1}, f_{\bullet 2}, f_{\bullet 3}$ sont appelées fréquences marginales.

$$\left\{ \begin{array}{l} f_{ij} = \frac{n_{ij}}{n}, f_{i\bullet} = \frac{n_{i\bullet}}{n}, f_{\bullet j} = \frac{n_{\bullet j}}{n} \\ f_{11} + f_{12} + f_{13} = f_{1\bullet}, \\ f_{21} + f_{22} + f_{23} = f_{2\bullet}, \\ f_{11} + f_{21} = f_{\bullet 1}, \\ f_{12} + f_{22} = f_{\bullet 2}, \\ f_{13} + f_{23} = f_{\bullet 3}, \\ f_{11} + f_{12} + f_{13} + f_{21} + f_{22} + f_{23} = 1. \end{array} \right.$$

Droite de régression linéaire et prédition

Covariance

$$s_{XY} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x}\bar{y}.$$

Coefficient de corrélation

$$r_{XY} = \frac{s_{XY}}{s_X s_Y}.$$

- Le coefficient de corrélation mesure la dépendance linéaire entre les variables X et Y .

Droite de régression linéaire et prédition

- On a $-1 \leq r_{XY} \leq 1$. Si r_{XY} est proche de 1 ou -1, les variables X et Y sont dits : **fortement corrélées**.
- Si le coefficient de corrélation est positif, les points du nuage sont alignés le long d'une droite croissante. Dans ce cas X et Y évoluent dans le même sens.
- Si le coefficient de corrélation est négatif, les points sont alignés le long d'une droite décroissante. Dans ce cas X et Y évoluent dans des sens opposés.
- Si le coefficient de corrélation est nul ou proche de zéro, il n'y a pas dépendance linéaire.

Droite de régression linéaire et prédition

Droite de régression linéaire et nuage des points

Neuf étudiants émettent un avis pédagogique vis-à-vis d'un professeur selon une échelle d'appréciation de 1 à 20. On relève par ailleurs la note obtenue par ces étudiants l'année précédente auprès du professeur.

Y= avis	5	7	16	6	12	14	10	9	8
X= résultat	8	11	10	13	9	17	7	15	16

Droite de régression linéaire et prédition

La droite de régression linéaire est la droite qui ajuste au mieux un nuage de points au sens des moindres carrés. On considère que la variable X est explicative et que la variable Y est dépendante. L'équation de la droite de régression linéaire de Y en X est :

$$y = \hat{a} + \hat{b}x,$$

avec

$$\hat{b} = \frac{s_{XY}}{s_X^2} \quad \hat{a} = \bar{y} - \hat{b}\bar{x}.$$

N.B : la droite de régression passe par le point (\bar{x}, \bar{y}) .

Droite de régression linéaire et prédition

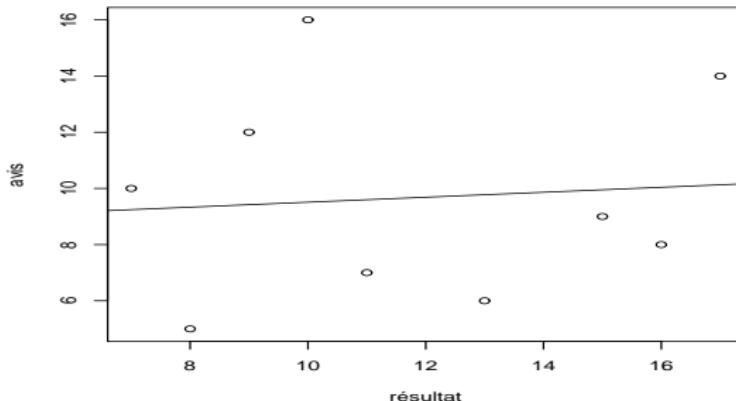
Dans notre exemple, on a :

$$\left\{ \begin{array}{l} \bar{x} = 106/9 = 11.78 \\ \bar{y} = 87/9 = 9.667 \\ s_x^2 = 1354/9 - 11.78^2 = 11.73 \\ s_y^2 = 951/9 - 9.667^2 = 12.22 \\ s_{XY} = 1034/9 - 9.667 \times 11.78 = 1.037 \\ r_{XY} = \frac{1.037}{\sqrt{11.73} \times \sqrt{12.22}} = 0.087 > 0 \end{array} \right.$$

Droite de régression linéaire et prédition

Finalement l'équation de la droite de régression linéaire de Y en X :

$$y = 0.088x + 8.625$$



Droite de régression linéaire et prédition

Prédiction

Y= avis	5	7	16	6	12	14	10	9	8
X= résultat	8	11	10	13	9	17	7	15	16

Dans notre exemple si on veut prédire, sur la base de notre modèle, l'avis pour un étudiant ayant obtenu $x = 12/20$, alors la valeur ajustée est :

$$y = 0.088 \times 12 + 8.625 = 9.681.$$

Droite de régression linéaire et prédition

Résidus ou erreurs de prédition

Les résidus de la régression linéaire sont définis par :

$$e_i = y_i - (\hat{a} + \hat{b}x_i) = y_i - y_i^*.$$

Le résidu e_i est l'erreur que l'on commet en utilisant la droite de régression linéaire pour prédire y_i à partir de x_i .

Les résidus sont les différences entre les valeurs observées y_i et les valeurs ajustées y_i^* de la variable dépendante.

Par exemple pour la valeur $x_3 = 12$, on donne $y_3 = 10$ et on a
 $y_3^* = 0.088 \times 12 + 8.625 = 9.681$, donc $e_3 = y_3 - y_3^* = 0.319$.

Droite de régression linéaire et prédition

moyenne résiduelle

$$\bar{e} = \frac{1}{n} \sum_{i=1}^n e_i = 0.$$

Variance résiduelle

$$s_e^2 = \frac{1}{n} \sum_{i=1}^n e_i^2.$$

La variance résiduelle peut également s'écrire :

$$s_e^2 = s_y^2 \times (1 - r_{XY}^2).$$

Dans notre exemple on a :

$$s_e^2 = 12.22 \times (1 - 0.087^2) = 12.127.$$

Droite de régression linéaire et prédition

Le coefficient de détermination

$$R^2 = r_{XY}^2.$$

Il représente la proportion de variance expliquée par le modèle.

Dans notre modèle, on a $R^2 = 0.087^2 = 0.008$.
(0.8% très faible).

Utilisation d'une calculatrice Casio fx 82ES et fx 82ES

Utilisation d'une calculatrice Casio fx 82MS

Exemple

Y= pression	1003	1005	1010	1011	1014
X= température	10	15	20	25	30

Utilisation d'une calculatrice Casio fx 82ES et fx 82ES

Dans le mode REG:

1 (Lin)

SHIFT CLR 1 (Scl) **=** (Stat clear)

10 **.** 1003 **DT** n= REG 1.

Chaque fois que vous appuyez sur **DT** pour enregistrer la valeur, le nombre de données saisies jusqu'à ce point est indiqué à l'écran (valeur *n*).

15 **.** 1005 **DT**

20 **.** 1010 **DT** 25 **.** 1011 **DT**

30 **.** 1014 **DT**

Coefficient de régression A = 997,4

SHIFT S-VAR **1** **=**

Coefficient de régression B = 0,56

SHIFT S-VAR **2** **=**

Coefficient de corrélation r = 0,982607368

SHIFT S-VAR **3** **=**

Pression atmosphérique à 18°C = 1007,48

18 **SHIFT S-VAR** **2** **=**

Température à 1000 hPa = 4,642857143

1000 **SHIFT S-VAR** **1** **=**

Coefficient de détermination = 0,965517241

SHIFT S-VAR **X²** **=**

Covariance de l'échantillon = 35

SHIFT S-SUM **3** **X** **SHIFT S-VAR** **1** **=**
SHIFT S-VAR **1** **=**

Utilisation d'une calculatrice Casio fx 82ES et fx 82ES

Utilisation d'une calculatrice Casio fx 82ES

Types de calculs statistiques

Touche	Elément du menu	Calcul statistique
1	1-VAR	Une variable
2	A+BX	Régression linéaire
3	_+CX ²	Régression quadratique
4	ln X	Régression logarithmique
5	e ^X	Régression exponentielle e
6	A•B ^X	Régression exponentielle ab
7	A•X ^B	Régression de puissance
8	1/X	Régression inverse

Utilisation d'une calculatrice Casio fx 82ES et fx 82ES

Sous-menu Reg (**SHIFT** **1** (STAT) **7** (Reg))

Sélectionnez cet élément du menu :	Pour obtenir ceci :
1 A	Pente A de la droite de régression
2 B	Constante B de la droite de régression
3 <i>r</i>	Coefficient de corrélation <i>r</i>
4 \hat{x}	Valeur estimée de <i>x</i>
5 \hat{y}	Valeur estimée de <i>y</i>

Exercices

Exercice 4 :

Considérons un échantillon de $n = 10$ fonctionnaires (ayant entre 40 et 50 ans) d'un ministère. Soit X le nombre d'années de service et Y le nombre de jours d'absence pour raison de maladie (au cours de l'année précédente) déterminé pour chaque personne appartenant à cet échantillon.

x_i	2	14	16	8	13	20	24	7	5	11
y_i	3	13	17	12	10	8	20	7	2	8

1. Déterminer les moyennes de X et Y et la covariance entre X et Y .

Exercices

2. Déterminer le coefficient de corrélation entre les variables X et Y . Donnez une interprétation.
3. Déterminer la droite de régression linéaire Y en fonction de X .
4. Tracer le nuage de points (X, Y) .
5. Tracer la droite de régression linéaire de Y en X .
6. Déterminer, sur la base de ce modèle, le nombre de jours d'absence pour un fonctionnaire ayant 22 ans de service.
7. Déterminer la variance résiduelle et le coefficient de détermination.

Interprétez.

Exercices

Exercice 5 :

On étudie un échantillon de taille $n = 100$ sur lequel ont été mesurés deux caractères X et Y . On a observé les résultats suivants :

$$\sum_{i=1}^{100} x_i = 800$$

$$\sum_{i=1}^{100} y_i = 1200$$

$$\sum_{i=1}^{100} x_i^2 = 7200$$

$$\sum_{i=1}^{100} y_i^2 = 1600$$

$$\sum_{i=1}^{100} x_i y_i = 10200$$

1. Déterminer les moyennes, les variances et la covariance de X et Y .
2. En déduire le coefficient de corrélation entre X et Y . Interpréter.
3. Déterminer la droite de régression linéaire de Y en X .
4. Déterminer la droite de régression linéaire de X en Y .
5. Déterminer la variance résiduelle et le coefficient de détermination.

Interpréter.

Plan :

- 1 Introduction
- 2 Statistique descriptive univariée
- 3 Statistique descriptive bivariée
- 4 Notions de probabilités et variables aléatoires
- 5 Échantillonnage et estimation
- 6 Tests des hypothèses
- 7 Exercices de révisions
- 8 Examens (sessions normale et de rattrapage 2016-2017)

Notions de probabilités

L'objectif ici de proposer un modèle mathématique permettant de décrire une situation aléatoire qui est définie à l'aide du langage des événements. La notion de probabilité permet de quantifier la chance qu'ont les événements d'être réalisés ou non.

Notions de probabilités

Vocabulaires

Expérience aléatoire : on ne peut pas prédire a priori son résultat.

Réalisation : un résultat possible de l'expérience aléatoire.

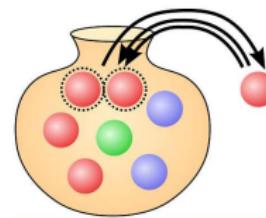
Espace fondamental Ω : l'ensemble de tous les résultats possibles.

Événement : sous ensemble de Ω .

Notions de probabilités

Exemples

1) on tire simultanément deux boules dans une urne qui contient quatres boules rouges, deux boules bleues et une boule verte :



$$\Omega = \{(R_1, R_2), (R_1, R_3), (R_2, R_3), (R_1, V_1), \dots\}$$

$$\implies \text{card } (\Omega) = C_7^2 = 21.$$

Notions de probabilités

2) on lance une pièce de monnaie deux fois :



$$\Omega = \{(P, P), (F, P), (P, F), (F, F)\} \implies \text{card } (\Omega) = 4.$$

L'événement A : (obtenir pile une seule fois) est $A = \{(F, P), (P, F)\}$.

Notions de probabilités

3) on lance un dé une seule fois :



$$\Omega = \{1, 2, 3, 4, 5, 6\} \implies \text{card } (\Omega) = 6.$$

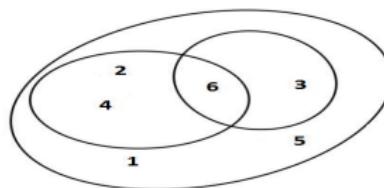
L'évènement B : (obtenir un nombre pair) est $A = \{2, 4, 6\}$.

Notions de probabilités

L'union : $A \cup B$ est réalisé dès que A ou B est réalisé.

L'intersection : $A \cap B$ est réalisé dès que A et B sont réalisés conjointement.

Par exemple, dans un lancer de dé, si $A = (\text{obtenir un nombre pair})$ et $B = (\text{obtenir un multiple de 3})$, alors $A = \{2, 4, 6\}$, $B = \{3, 6\}$, $A \cap B = \{6\}$ et $A \cup B = \{2, 3, 4, 6\}$.



Le complémentaire \bar{A} : c'est l'évènement $\Omega \setminus A$.

Pour $A = \{2, 4, 6\}$, on a $\bar{A} = \{1, 3, 5\} =$ (les nombres impairs).

Notions de probabilités

Définition de probabilité

Une probabilité P sur un ensemble fini $\Omega = \{\omega_1, \dots, \omega_n\}$ est une pondération p_1, p_2, \dots, p_n des éléments de cet ensemble telle que :

- Pour tout $k \in [1 : n]$, $p_k \geq 0$.
- $\sum_{k=1}^n p_k = 1$.

On attribue à toute éventualité ω_i la probabilité $p_i = P(\{\omega_i\})$ et on attribue à tout événement $A \subset \Omega$ la probabilité :

$$P(A) = \sum_{k, \omega_k \in A} p_k.$$

Notions de probabilités

Exemple

Dans un lancer d'un dé équilibré, on a $\Omega = \{1, \dots, 6\}$ et $p_k = P(\{k\}) = 1/6$, pour tout $k \in [1 : 6]$. Si l'événement A est (obtenir un nombre pair strictement supérieur à 3) alors $A = \{4, 6\}$ et $P(A) = P(\{4\}) + P(\{6\}) = 1/6 + 1/6 = 1/3$.

Notions de probabilités

Remarque

Dans le cas où les symétries font que tous les résultats possibles $\omega_1, \dots, \omega_n$ sont équiprobables, on a :

$$P(A) = \frac{\text{card}(A)}{\text{card}(\Omega)}.$$

Dans l'exemple précédent, on a $\Omega = \{1, \dots, 6\}$ et $A = \{4, 6\}$:

$$P(A) = \frac{\text{card}(A)}{\text{card}(\Omega)} = \frac{2}{6} = \frac{1}{3}.$$

Notions de probabilités

Propriétés

- ① Soit A et B deux événements, alors

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

- ② $P(\bar{A}) = 1 - P(A)$.
③ $P(\Omega) = 1$, (événement certain).
④ $P(\emptyset) = 0$, (événement impossible).
⑤ $0 \leq P(A) \leq 1$.

Notions de probabilités

Probabilité conditionnelle et indépendance

- Soient deux événements A et B , si $P(B) > 0$, alors

$$P_B(A) = \frac{P(A \cap B)}{P(B)}.$$

- On dit que deux événements A et B sont indépendants si $P(A \cap B) = P(A).P(B)$ ou encore si $P_B(A) = P(A)$ ou $P_A(B) = P(B)$.

Notions de probabilités

Exemple

Nous jetons un dé équilibré et nous considérons les deux événements :

$A = (\text{le résultat est impair})$ et $B = (\text{le résultat est } \leq 4)$.

$$A = \{1, 3, 5\} \Rightarrow P(A) = \frac{1}{2}$$

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

$$A \cap B = \{1, 3\} \Rightarrow P(A \cap B) = \frac{1}{3}$$

On a $P(A \cap B) = \frac{1}{3} = P(A).P(B)$ donc A et B sont indépendants.

Notions de variables aléatoires

Notions de variables aléatoires

La notion de variable aléatoire formalise l'association d'une valeur au résultat d'une expérience aléatoire. Une variable aléatoire X est une application mesurable de l'ensemble fondamental Ω dans \mathbb{R} .

Notions de variables aléatoires

Variables aléatoires discrètes

- **Une variable aléatoire discrète prend uniquement des valeurs dans un sous ensemble discret**

$$E = \{x_1, \dots, x_n, \dots\}.$$

- **Une distribution de probabilité p_X est une fonction qui associe à chaque valeur x_i la probabilité :**

$$p_i = p_X(x_i) = P(X = x_i).$$

Notions de variables aléatoires

Exemple

On considère une expérience aléatoire consistant à lancer deux pièces de monnaie. L'ensemble des résultats possibles est :

$$\Omega = \{(P, P), (F, P), (P, F), (F, F)\}.$$

Considérons la variable aléatoire X représentant le nombre de faces obtenues.
Les valeurs prises par X son 0,1,2.

$$p_1 = P(X = 0) = 1/4, \quad p_2 = P(X = 1) = 1/2 \text{ et } p_3 = P(X = 2) = 1/4.$$

x_i	0	1	2
$p_i = P(X = x_i)$	1/4	1/2	1/4

Notions de variables aléatoires

Espérance mathématique

$$E(X) = 0 \times 1/4 + 1 \times 1/2 + 2 \times 1/4 = 1.$$

Variance

$$\begin{aligned} V(X) &= E(X - E(X))^2 \\ &= (0 - 1)^2 \times 1/4 + (1 - 1)^2 \times 1/2 + (2 - 1)^2 \times 1/4 = 1/2. \end{aligned}$$

Écart type

$$\sigma(X) = \sqrt{V(X)} = \sqrt{1/2} = 0.70.$$

Notions de variables aléatoires

Quelques lois discrètes usuelles

Nom	Loi	$\mathbb{E}(X)$	$\text{Var}(X)$
Bernoulli $\mathcal{B}(p)$	$\mathbb{P}(X = 1) = p = 1 - \mathbb{P}(X = 0)$	p	$p(1 - p)$
binomiale $\mathcal{B}(n, p)$	$\forall 0 \leq k \leq n, \mathbb{P}(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$	np	$np(1 - p)$
Poisson $\mathcal{P}(\lambda)$	$\forall n \in \mathbb{N}, \mathbb{P}(X = n) = \exp(-\lambda) \frac{\lambda^n}{n!}$	λ	λ
géométrique $\mathcal{Geo}(p)$	$\forall n \in \mathbb{N}^*, \mathbb{P}(X = n) = p(1 - p)^{n-1}$	$\frac{1}{p}$	$\frac{1-p}{p^2}$

Notions de variables aléatoires

Variables aléatoires continues

- **Une variable aléatoire X continue prend ses valeurs dans un intervalle de \mathbb{R} .**
- **X possède une fonction de répartition notée $\Phi(x) = P(X \leq x)$ et une densité notée $f(x)$ telles que :**

$$\Phi(x) = \int_{-\infty}^x f(t)dt \text{ ou encore } \Phi'(x) = f(x).$$

- **La densité $f(x)$ vérifie : $\int_{-\infty}^{+\infty} f(x)dx = 1$.**

Notions de variables aléatoires

- **Espérance mathématique**

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

- **Variance**

$$V(X) = E(X - E(X))^2 = \int_{-\infty}^{+\infty} (x - E(X))^2 f(x)dx.$$

Notions de variables aléatoires

Quelques lois continues usuelles

Nom	Densité	$\mathbb{E}(X)$	$\text{Var}(X)$
uniforme $\mathcal{U}[a, b]$	$\frac{1}{b-a} \mathbf{1}_{[a,b]}(x)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
exponentielle $\mathcal{E}(\lambda)$	$\lambda \exp(-\lambda x) \mathbf{1}_{\{x \geq 0\}}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
normale $\mathcal{N}_1(m, \sigma^2)$	$\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-m)^2}{2\sigma^2}\right)$	m	σ^2
Cauchy $\mathcal{C}(a)$	$\frac{a}{\pi} \frac{1}{x^2+a^2}$	non intégr.	-

Notions de variables aléatoires

Exemple

Si X suit la loi exponentielle $\mathcal{E}(\lambda)$, alors quand $x > 0$, sa fonction de répartition vaut :

$$\Phi(x) = P(X \leq x) = \int_{-\infty}^x f(t)dt = \int_0^x \lambda \exp(-\lambda t)dt = \left[-\exp(-\lambda t) \right]_0^x = 1 - e^{-\lambda x}.$$

On calcule la moyenne par une intégration par partie :

$$E(X) = \int_{-\infty}^{+\infty} xf(x)dx = \int_0^{+\infty} x\lambda \exp(-\lambda x)dx = \left[-\frac{1+x\lambda}{\lambda} \exp(-\lambda x) \right]_0^{+\infty} = \frac{1}{\lambda}.$$

Il est également possible de montrer que la variance vaut : $V(X) = \frac{1}{\lambda^2}$.

Notions de variables aléatoires

Loi normale $\mathcal{N}(0, 1)$ de moyenne 0 et variance 1

La table statistique ci-dessous nous donne pour chaque $z > 0$, la valeur $\Phi(z) = P(Z \leq z) = P(Z < z)$ de la fonction de répartition de Z .

Pour $z < 0$, on a $\Phi(z) = 1 - \Phi(-z)$.

On a encore.

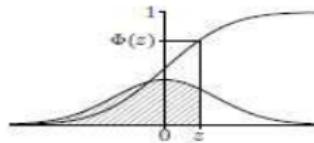
$$P(X \geq z) = P(X > z) = 1 - \Phi(z).$$

$$P(a < X < b) = \Phi(b) - \Phi(a).$$

Notions de variables aléatoires

A.1. LOI NORMALE $\mathcal{N}(0, 1)$

1^e Fonction de répartition de la loi Normale. — La fonction de répartition Φ de la loi Normale $\mathcal{N}(0, 1)$ est définie par $\Phi(z) = \int_{-\infty}^z e^{-u^2/2} du / \sqrt{2\pi}$, $z \in \mathbb{R}$. Pour tout $z \in \mathbb{R}$, on a $\Phi(z) = 1 - \Phi(-z)$.



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



Notions de variables aléatoires

Dans cette table, on trouve par exemple

$$P(Z \leq 0.25) = \Phi(0.25) = \Phi(0.20 + 0.05) = 0.5987.$$

Cette dernière valeur est l'intersection de la ligne 0.2 et la colonne 0.05 dans la table statistique.

$$P(Z > 0.25) = 1 - \Phi(0.25) = 1 - 0.5987 = 0.4013.$$

$$\Phi(-0.25) = 1 - \Phi(0.25) = 1 - 0.5987 = 0.4013.$$

$$P(-0.25 < Z < 0.25) = \Phi(0.25) - \Phi(-0.25) = 2\Phi(0.25) - 1.$$

Notions de variables aléatoires

Loi normale $\mathcal{N}(\mu, \sigma^2)$ de moyenne μ et variance σ^2

Si X suit la loi normale $\mathcal{N}(\mu, \sigma^2)$, alors

$$Z = \frac{X - \mu}{\sigma} \implies Z \sim \mathcal{N}(0, 1).$$

Exemple

Si $X \sim \mathcal{N}(72, 64)$ alors la moyenne est $\mu = 72$ et l'écart type $\sigma = 8$. Dans ce cas la variable $Z = \frac{X-72}{8}$ suit une loi $\mathcal{N}(0, 1)$, par suite

$$\begin{aligned} P(X > 80) &= 1 - P(X \leq 80) = 1 - P\left(\frac{X - 72}{8} \leq \frac{80 - 72}{8}\right) \\ &= 1 - P[Z \leq 1] = 1 - \Phi(1) = 1 - 0.8413 = 0.1587. \end{aligned}$$

Exercices

Exercice 1 :

On tire simultanément deux boules dans une urne qui contient : trois boules vertes, deux boules noires et une boule rouge. Soit X la variable aléatoire qui correspond au nombre de boules vertes tirées.

- ① Déterminer la loi de X . En déduire $E(X)$, $V(X)$ et $\sigma(X)$.
- ② Calculer $P(X \leq 1)$ et $P(X > 1)$.

Exercices

Exercice 2 :

On lance une pièce de monnaie équilibré trois fois successive. Soit X la variable aléatoire qui correspond au nombre de faces obtenues.

- ① Déterminer la loi de X . En déduire $E(X)$, $V(X)$ et $\sigma(X)$.
- ② Calculer $P(X \leq 2)$ et $P(X > 2)$.

Exercices

Exercice 3 :

On lance deux trièdres équilibrés numérotés 0, 1, 2, 3. Soit X la variable aléatoire qui correspond à la somme des deux chiffres obtenus.

- ➊ Déterminer la loi de X . En déduire $E(X)$, $V(X)$ et $\sigma(X)$.

Exercices

Exercice 4 :

Soit X une variable aléatoire qui suit la loi exponentielle de paramètre 1.

- ① Vérifier que $\int_{-\infty}^{+\infty} f(x)dx = 1$.
- ② Calculer $E(X)$, $V(X)$ et $\sigma(X)$.
- ③ Calculer $P(X \leq 1)$ et $P(X > 1)$.

Exercices

Exercice 5 :

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

- ① $P(Z \leq 1.23)$.
- ② $P(Z \leq -1.23)$.
- ③ $P(Z > 1.23)$.
- ④ $P(Z \in [-1.23, 1.23])$.
- ⑤ $P(Z \in [-0.88, 0.36])$.

Exercices

Exercice 6 :

Soit $Z \sim \mathcal{N}(0, 1)$. Déterminer les valeurs z telles que :

- ① $P(Z \leq z) = 0.975$.
- ② $P(Z \leq -z) = 0.3438$.

Exercices

Exercice 7 :

Soit une variable aléatoire $X \sim \mathcal{N}(53, \sigma^2 = 100)$ représentant le résultat d'un examen pour un étudiant d'une section. Déterminer la probabilité pour que le résultat soit compris entre 33.4 et 72.6.

Exercices

Exercice 8 :

Sur une route principale où la vitesse est limitée à 80 km/h, un radar a mesuré la vitesse de toutes les automobiles pendant une journée. En supposant que les vitesses recueillies soient distribuées normalement avec une moyenne de 72 km/h et un écart-type de 8 km/h, quelle est approximativement la proportion d'automobiles ayant commis un excès de vitesse ?

Exercices

Exercice 9 :

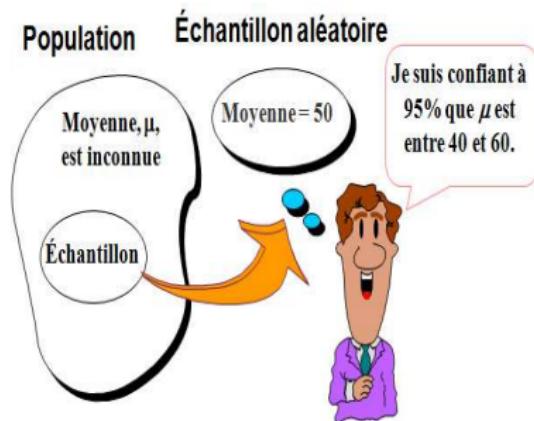
On suppose que la glycémie est distribuée normalement dans la population, avec une moyenne de $1g/l$ et un écart type $0.03g/l$. On mesure la glycémie chez un individu. Calculer la probabilité pour que sa glycémie soit :

- ① inférieure à 1.06 .
- ② supérieure 0.9985 .

Plan :

- 1 Introduction
- 2 Statistique descriptive univariée
- 3 Statistique descriptive bivariée
- 4 Notions de probabilités et variables aléatoires
- 5 Échantillonnage et estimation
- 6 Tests des hypothèses
- 7 Exercices de révisions
- 8 Examens (sessions normale et de rattrapage 2016-2017)

Échantillonnage et estimation



Échantillonnage

Exemple

Une université reçoit 7000 applications par année provenant d'éventuels étudiants. Le formulaire de demande d'admission inclut le score d'un test d'aptitude (SAT) ainsi que l'information sur le lieu de résidence de l'étudiant. Le directeur des admissions aimeraient avoir une idée sur :

- le score moyen SAT des postulants,
- la proportion des postulants qui sont résidents de la province ?

Il y a deux façons d'obtenir cette information :

Échantillonnage

1)- Effectuer un recensement des 7000 postulants

- Scores SAT : x_1, \dots, x_{7000}
- Moyenne de la population

$$\mu = \frac{1}{7000} \sum_{i=1}^{7000} x_i = 990$$

- Variance de la population

$$\sigma^2 = \frac{1}{7000} \sum_{i=1}^{7000} (x_i - \mu)^2 = 6400$$

- Proportion des résidants de la population

$$p = \frac{5040}{7000} = 0.72$$

Échantillonnage

2) Prendre un échantillon de taille $n = 50$ postulants est faire des estimations ponctuelles

No	Postulant	SAT	Résidant
1	Mohammed	1025	Oui
2	Rachid	950	Oui
3	Hassan	1090	Non
.
.
.
.
.
50	Karim	1279	oui
	Total	49850	34 oui

Échantillonnage

Dans ce cas on prend :

- \bar{x} comme **estimateur** de la moyenne μ

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{50} \sum_{i=1}^{50} x_i = 997 \simeq 990 = \mu$$

- s^2 comme **estimateur** de la variance σ^2

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{50-1} \sum_{i=1}^{50} (x_i - \bar{x})^2 = 6301 \simeq 6400 = \sigma^2$$

- \bar{p} comme **estimateur** de la proportion p

$$\bar{p} = \frac{n_{oui}}{n} = \frac{34}{50} = 0.68 \simeq 0.72 = p$$

Échantillonnage

Les erreurs d'échantillonnage sont :

- $|\bar{x} - \mu|$ pour la moyenne échantillonnale.
- $|s^2 - \sigma^2|$ pour la variance échantillonnale.
- $|\bar{p} - p|$ pour la proportion échantillonnale.

Raisons pour faire un échantillonnage au lieu d'un recensement

- Lorsque la population est très grande.
- Par souci d'économie.
- Obtenir de l'information rapidement.

Échantillonnage

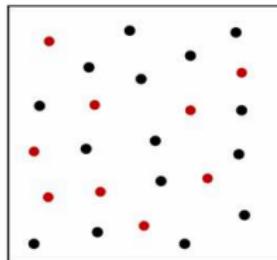
Méthodes d'échantillonnage

- **L'échantillonnage aléatoire et simple** : le choix se fait parmi tous les individus de la population (au sens statistique), qui ne forme qu'un grand ensemble.
- **L'échantillonnage stratifié** : si la population est très hétérogène, elle peut être divisée en sous-ensembles exclusifs (ou strates). Au sein de ces strates l'échantillonnage est ensuite aléatoire et simple. Les strates sont identifiées dans l'analyse statistique comme les niveaux d'un facteur fixe.

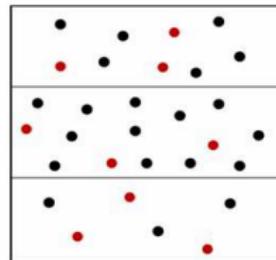
Échantillonnage

- **L'échantillonnage en grappes** : si les strates sont très nombreuses, on en choisit certaines au hasard (les grappes). Au sein de ces grappes l'échantillonnage est ensuite aléatoire et simple. Les grappes sont identifiées dans l'analyse statistique comme les niveaux d'un facteur aléatoire.
- **L'échantillonnage par degrés** : il est une généralisation de l'échantillonnage en grappes. Au sein de la population on choisit des grappes " primaires ", puis à l'intérieur de celles-ci des grappes " secondaires " (toujours au hasard), et ainsi du suite. Au dernier niveau l'échantillonnage est aléatoire et simple.

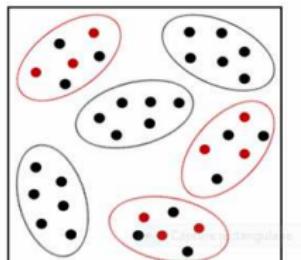
Échantillonnage



Echantillonage aléatoire et simple



Echantillonage stratifié



Echantillonage en grappes

Distribution d'échantillonnage de la moyenne

Distribution d'échantillonnage de \bar{X}

Théorème centrale limite 1

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

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right), \quad \text{i.e.,} \quad Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0, 1).$$

Distribution d'échantillonnage de la moyenne

Remarque

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

$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{s^2}{n}\right), \quad \text{i.e.,} \quad \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} \sim \mathcal{N}(0, 1).$$

Échantillonnage

Exemple

Les statistiques des notes obtenues en mathématiques au BAC STI en France pour l'année 2006 sont : Moyenne nationale =10.44, Écart-type=1.46.

Une classe de BTS comporte 35 élèves en 2006/2007 issus d'un BAC STI en 2006.

Calculer la probabilité que la moyenne de cette classe soit supérieure à 10.

Échantillonnage

Corrigé

$$n = 35 > 30 \Rightarrow \bar{X} \sim \mathcal{N}(10.44, \frac{1.46^2}{n}) \Rightarrow Z = \frac{\bar{X} - 10.44}{\frac{1.46}{\sqrt{35}}} \sim \mathcal{N}(0, 1).$$

$$\begin{aligned} P(\bar{X} > 10) &= 1 - P(\bar{X} \leq 10) = 1 - P\left(\frac{\bar{X} - 10.44}{\frac{1.46}{\sqrt{35}}} \leq \frac{10 - 10.44}{\frac{1.46}{\sqrt{35}}}\right) \\ &= 1 - P(Z \leq -1.78) = 1 - \Phi(-1.78) = \Phi(1.78) = 0.9625. \end{aligned}$$

**La probabilité que la moyenne de cette classe soit supérieure à 10 est
 $\simeq 96.25 \%$.**

Échantillonnage

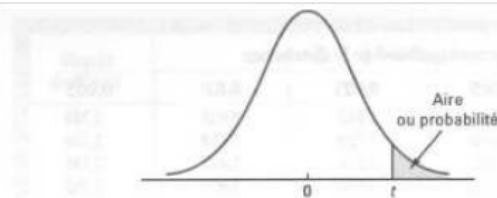
Théorème centrale limite 2

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

$$T = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} \sim t(n - 1) : \text{Loi de Student à } n-1 \text{ degré de liberté ddl.}$$

Échantillonnage

Figure – Table de la loi Student T : $P(T > t) = \alpha$



Les chiffres de la table correspondent aux valeurs t pour différentes aires ou probabilités situées dans la queue supérieure de la distribution de Student. Par exemple, avec 10 degrés de liberté et une aire de 0,05 dans la queue supérieure de la distribution, $t_{0,05} = 1,812$. (pour test unilatéral !)

Degrés de liberté	Aire dans la queue supérieure de la distribution					
	0,20	0,10	0,05	0,025	0,01	0,005
1	1,376	3,078	6,314	12,706	31,821	63,656
2	1,061	1,886	2,920	4,303	6,965	9,925
3	0,978	1,638	2,353	3,182	4,541	5,841
4	0,941	1,533	2,132	2,776	3,747	4,604
5	0,920	1,476	2,015	2,571	3,365	4,032
6	0,906	1,440	1,943	2,447	3,143	3,707
7	0,896	1,415	1,895	2,365	2,998	3,499
8	0,889	1,397	1,860	2,306	2,896	3,355
9	0,883	1,383	1,833	2,262	2,821	3,250

Échantillonnage

Exemple

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

Échantillonnage

Corrigé

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

$$T = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} = \frac{\bar{X} - 120}{\frac{6}{\sqrt{25}}} \sim t(25 - 1) = t(24).$$

Par suite

$$\begin{aligned} P(\bar{X} > 123) &= P\left(\frac{\bar{X} - 120}{\frac{6}{\sqrt{25}}} \geq \frac{123 - 120}{\frac{6}{\sqrt{25}}}\right) \\ &= P(T > 2.5) \simeq P(T > 2.492) = 0.01. \end{aligned}$$

Échantillonnage

Distribution d'échantillonnage de la proportion

Théorème

Si p est la proportion de la population, alors pour un échantillon de taille grande $n > 30$ on a

$$\bar{p} \sim \mathcal{N}\left(p, \frac{p(1-p)}{n}\right) \quad \text{i.e.,} \quad Z = \frac{\bar{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim \mathcal{N}(0, 1).$$

Échantillonnage

Exemple

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

Corrigé

$$Z = \frac{\bar{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim \mathcal{N}(0, 1)$$

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

Échantillonnage

Distribution d'échantillonnage de la variance

Théorème

Si σ^2 est la variance de la population et S^2 est la variance échantillonale d'un échantillon de taille n . Alors $Y^2 = \frac{(n-1)S^2}{\sigma^2}$ suit la loi khi-deux à $\nu = n - 1$ ddl.

Exemple

Dans la table ci-dessous, pour $\alpha = 0.05$ et $n = 10$, on a $\nu = 10 - 1 = 9$ et $x = 16.9190$.

Échantillonnage

Figure – Table de la loi Khi-deux : $P(Y^2 > x) = \alpha$

$\nu \setminus \alpha$	0,990	0,975	0,950	0,900	0,100	0,050	0,025	0,010	0,001
1	0,0002	0,0010	0,0039	0,0158	2,7055	3,8415	5,0239	6,6349	10,8276
2	0,0201	0,0506	0,1026	0,2107	4,6052	5,9915	7,3778	9,2103	13,8155
3	0,1148	0,2158	0,3518	0,5844	6,2514	7,8147	9,3484	11,3449	16,2662
4	0,2971	0,4844	0,7107	1,0636	7,7794	9,4877	11,1433	13,2767	18,4668
5	0,5543	0,8312	1,1455	1,6103	9,2364	11,0705	12,8325	15,0863	20,5150
6	0,8721	1,2373	1,6354	2,2041	10,6446	12,5916	14,4494	16,8119	22,4577
7	1,2390	1,6899	2,1673	2,8331	12,0170	14,0671	16,0128	18,4753	24,3219
8	1,6465	2,1797	2,7326	3,4895	13,3616	15,5073	17,5345	20,0902	26,1245
9	2,0879	2,7004	3,3251	4,1682	14,6837	16,9190	19,0228	21,6660	27,8772
10	2,5582	3,2470	3,9403	4,8652	15,9872	18,3070	20,4832	23,2093	29,5883
11	3,0535	3,8157	4,5748	5,5778	17,2750	19,6751	21,9200	24,7250	31,2641
12	3,5706	4,4038	5,2260	6,3038	18,5493	21,0261	23,3367	26,2170	32,9095
13	4,1069	5,0088	5,8919	7,0415	19,8119	22,3620	24,7356	27,6883	34,5282
14	4,6604	5,6287	6,5706	7,7895	21,0641	23,6848	26,1189	29,1412	36,1233
15	5,2293	6,2621	7,2609	8,5468	22,3071	24,9958	27,4884	30,5779	37,6973
16	5,8189	6,8977	7,8616	9,3192	22,5412	26,2069	28,8454	31,8999	39,2594

Estimation

Estimation ponctuelle

Pour un échantillon de taille n , on prend :

- \bar{X} est un estimateur de la moyenne μ

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

- S^2 est un estimateur de la variance σ^2

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

- \bar{p} est un estimateur de la proportion p .

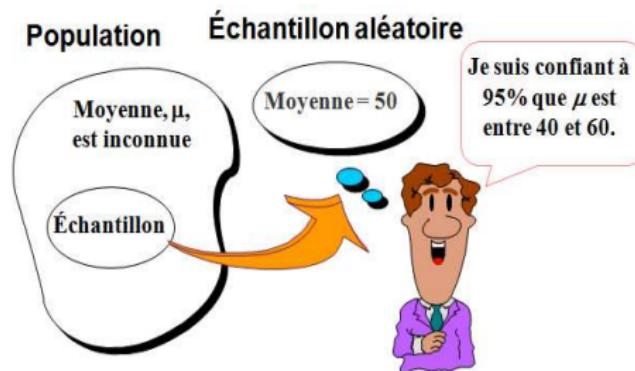
Estimation

Estimation par intervalle de confiance

Les estimations ponctuelles, bien qu'utiles, ne fournissent aucune information concernant la précision des estimations, c'est-à-dire qu'elles ne tiennent pas compte de l'erreur possible dans l'estimation due aux fluctuations d'échantillonnage.

Estimation

La théorie des intervalles de confiance (IC) consiste à construire, autour de l'estimation ponctuelle, un intervalle qui aura une grande probabilité ($1 - \alpha$) de contenir la vraie valeur du paramètre.



Estimation

1) Estimation par IC de la moyenne μ de la population

1^{er} cas : Lorsque la taille de l'échantillon est grande ($n \geq 30$) et la variance de la population de X est connue, on obtient un intervalle de confiance pour μ au seuil de confiance $(1 - \alpha)$ de la forme :

$$IC_{(1-\alpha)\%} = \left[\bar{x} - z_\alpha \frac{\sigma}{\sqrt{n}}; \bar{x} + z_\alpha \frac{\sigma}{\sqrt{n}} \right]$$

i.e., $P(\mu \in IC_{(1-\alpha)\%}) = (1 - \alpha).$

$\Phi(z_\alpha) = 1 - \frac{\alpha}{2}$, où Φ est la fdr de la loi normale $\mathcal{N}(0, 1)$.

Estimation

- Ceci est aussi vrai pour de petits échantillons lorsque la variable aléatoire X suit une loi normale et que la variance de X est connue.
- Lorsque la taille de l'échantillon est grande ($n \geq 30$) et la variance de la population de X est inconnue, on obtient un intervalle de confiance pour μ au seuil de confiance $(1 - \alpha)$ de la forme :

$$IC_{(1-\alpha)\%} = \left[\bar{x} - z_\alpha \frac{s}{\sqrt{n}}; \bar{x} + z_\alpha \frac{s}{\sqrt{n}} \right].$$

Estimation

Exemple

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

Corrigé

Puisque $\alpha = 0.05$ (5%), alors $\Phi(z_\alpha) = 1 - \frac{0.05}{2} = 0.975$, par suite $z_\alpha = 1.96$.

Finalement $IC_{95\%} = [167.86; 168.14]$, i.e., $P(\mu \in [167.86; 168.14]) = 0.95$.

Estimation

2eme cas : Lorsque la taille de l'échantillon est petite ($n < 30$) et X suit une loi normale de variance **inconnue**, on obtient un intervalle de confiance pour μ au seuil de confiance $(1 - \alpha)$ de la forme :

$$IC_{(1-\alpha)\%} = \left[\bar{x} - t_\alpha \frac{s}{\sqrt{n}}; \bar{x} + t_\alpha \frac{s}{\sqrt{n}} \right]$$

t_α se déduit de la table student comme suit : $P(T > t_\alpha) = \frac{\alpha}{2}$.

Estimation

Exemple

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

Estimation

Corrigé

pour un coefficient de confiance de 0,95, on a $\alpha = 0,05$, et $\frac{\alpha}{2} = 0,025$. On a $n - 1 = 10 - 1 = 9$ degrés de liberté, alors la table de la distribution Student nous donne $t_{\alpha} = 2,262$. Finalement $I_{C95\%} = [328.54; 371.46]$. i.e., nous sommes confiants à 95% que la moyenne des loyers mensuels (le vrai paramètre de la population μ), se trouve entre 328.54 et 371.46

Estimation

Détermination de la taille de l'échantillon

Taille de l'échantillon?



Quelle est la taille n de l'échantillon qui permettrait d'affirmer qu'en utilisant un estimateur ponctuel, l'erreur commise pour un coefficient de confiance $(1 - \alpha)$ serait moindre que la marge d'erreur E ?

Estimation

Si par exemple on fixe :

$$E = z_\alpha \frac{\sigma}{\sqrt{n}},$$

l'erreur maximale commise pour un coefficient de confiance $(1 - \alpha)$, alors la taille de l'échantillon sera :

$$n = \left[\frac{z_\alpha \cdot \sigma}{E} \right]^2.$$

Estimation

Exemple

Si on fixe une marge d'erreur $E = 500$, alors pour un écart type de $\sigma = 5000$ et pour $z_\alpha = 1.96$, on trouve $n = 384$. i.e., on a besoin d'un échantillon de taille $n = 384$ pour arriver à une précision de ± 500 à un seuil de confiance de 95%.

Estimation

2) Estimation par IC de la proportion p de la population

Lorsque n est grand ($n \geq 30$) et si \bar{p} est la proportion échantillonnable alors un intervalle de confiance pour la proportion p inconnue de la population au seuil de confiance $(1 - \alpha)$ de la forme :

$$IC_{(1-\alpha)\%} = \left[\bar{p} - z_\alpha \sqrt{\frac{\bar{p}(1 - \bar{p})}{n}}; \bar{p} + z_\alpha \sqrt{\frac{\bar{p}(1 - \bar{p})}{n}} \right],$$

avec $\Phi(z_\alpha) = 1 - \frac{\alpha}{2}$, où Φ est la fonction de répartition de la loi normale $\mathcal{N}(0, 1)$.

Estimation

Exemple

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

Estimation

Corrigé

On a $n = 500$, $\bar{p} = 220/500 = 0.44$ et $z_\alpha = 1.96$ donc

$IC_{95\%} = [0.3965; 0.4835]$, i.e., SPI est confiant à 95% que la proportion des votants qui favoriseront ce candidat est entre 0.3965 et 0.4835.

Estimation

3) Estimation par IC de la variance σ^2 de la population

Soit n la taille de l'échantillon et s^2 la variance échantillonnale. Ici, Ici $Y^2 = \frac{(n-1)s^2}{\sigma^2}$ suit la loi de loi khi-deux à $n - 1$ ddl. On cherche les deux nombres a et b tels que : $P(Y^2 \geq a) = \frac{\alpha}{2}$ et $P(Y^2 \geq b) = 1 - \frac{\alpha}{2}$. L'intervalle de confiance au seuil $(1 - \alpha)$ pour la variance inconnue σ^2 de la population est de la forme :

$$IC_{(1-\alpha)\%} = \left[\frac{(n-1)s^2}{a}; \frac{(n-1)s^2}{b} \right].$$

Estimation

Figure – Table de la loi Khi-deux χ^2 : $P(\chi^2 > x) = \alpha$

$\nu \setminus \alpha$	0,990	0,975	0,950	0,900	0,100	0,050	0,025	0,010	0,001
1	0,0002	0,0010	0,0039	0,0158	2,7055	3,8415	5,0239	6,6349	10,8276
2	0,0201	0,0506	0,1026	0,2107	4,6052	5,9915	7,3778	9,2103	13,8155
3	0,1148	0,2158	0,3518	0,5844	6,2514	7,8147	9,3484	11,3449	16,2662
4	0,2971	0,4844	0,7107	1,0636	7,7794	9,4877	11,1433	13,2767	18,4668
5	0,5543	0,8312	1,1455	1,6103	9,2364	11,0705	12,8325	15,0863	20,5150
6	0,8721	1,2373	1,6354	2,2041	10,6446	12,5916	14,4494	16,8119	22,4577
7	1,2390	1,6899	2,1673	2,8331	12,0170	14,0671	16,0128	18,4753	24,3219
8	1,6465	2,1797	2,7326	3,4895	13,3616	15,5073	17,5345	20,0902	26,1245
9	2,0879	2,7004	3,3251	4,1682	14,6837	16,9190	19,0228	21,6660	27,8772
10	2,5582	3,2470	3,9403	4,8652	15,9872	18,3070	20,4832	23,2093	29,5883
11	3,0535	3,8157	4,5748	5,5778	17,2750	19,6751	21,9200	24,7250	31,2641
12	3,5706	4,4038	5,2260	6,3038	18,5493	21,0261	23,3367	26,2170	32,9095
13	4,1069	5,0088	5,8919	7,0415	19,8119	22,3620	24,7356	27,6883	34,5282
14	4,6604	5,6287	6,5706	7,7895	21,0641	23,6848	26,1189	29,1412	36,1233
15	5,2203	6,2621	7,2600	8,5468	22,3071	24,9958	27,4884	30,5770	37,6973

Estimation

Exemple

Pour $n = 31$, $s^2 = 100$, on a le ddl = $31 - 1 = 30$. $\alpha = 0.05$ donc $\frac{\alpha}{2} = 0.025$ donne $a = 46.98$ et $1 - \frac{\alpha}{2} = 0.975$ donne $b = 16.80$. Finalement $IC_{95\%} = [0.3965; 0.4835]$.

Exercices

Exercice 1 (Échantillonnage) :

Pour estimer l'âge moyen d'une population de 4000 employés, un échantillon aléatoire de 40 employés est sélectionné. Quelle est la probabilité que l'âge moyen des employés de l'échantillon soit compris entre l'âge moyen de la population 2 si l'on sait que l'écart type de la population est de 8,2 ans ?

Exercices

Exercice 2 (Échantillonnage) :

Une élection a eu lieu et un candidat a eu une proportion 40% des voix. On prélève un échantillon de 100 bulletins de vote.

Quelle est la probabilité que, dans l'échantillon, le candidat ait entre 35% et 45% des voix ?

Exercices

Exercice 3 (Estimation) : Dans le cadre d'une étude sur la violence verbale au travail, on a interrogé au hasard 500 salariés de différents secteurs. 145 d'entre eux déclarent avoir déjà subi une violence verbale au travail.

- 1) Identifiez la population, la variable et son type.
- 2) Donnez une estimation ponctuelle de la proportion de salariés ayant déjà subi une violence verbale.
- 3) Donnez une estimation de cette proportion par un intervalle de confiance à 90% et 99%.
- 4) Si avec les mêmes données on calculait un intervalle de confiance à 99%, serait-il plus grand ou plus petit que celui trouvé à la question précédente ?

Exercices

Exercice 4 (Estimation) : On admet que le taux de cholestérol chez une femme suit une loi normale. Sur un échantillon de 10 femmes, on a obtenu les taux de cholestérol (en g/l) suivants :

3	1.8	2.1	2.7	1.4	1.9	2.2	2.5	1.7	2
---	-----	-----	-----	-----	-----	-----	-----	-----	---

- 1)- Déterminer une estimation ponctuelle de la moyenne et de la variance du taux.
- 2)- Déterminer un intervalle de confiance pour la moyenne du taux au seuil 5%.

Exercices

Exercice 5 (Estimation) : Une machine produit des pièces de type X. La masse, exprimée en (g), d'une pièce tiré au hasard dans la production, est distribuée selon une loi normale. On tire un échantillon de 17 pièces de masses suivantes :

250	254	254	253	256	250	257	251
253	255	250	255	252	261	252	251
							255

- 1) Donnez une estimation ponctuelle de la moyenne et la variance de production.
- 2) Déterminer une estimation par intervalle de confiance à 95% et 99% de la masse moyenne.
- 3) On suppose maintenant que la variance de la population est connue $\sigma^2 = 8.51$. Préciser la loi de la moyenne échantillonale et déterminer la taille minimale donner à un échantillon pour obtenir un intervalle de confiance pour la moyenne au niveau 95% d'amplitude inférieure à 2.
Conclure.

Exercices

Exercice 6 (Estimation) :

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 pour chacun d'eux :

Classes	[1610; 1615[[1615; 1620[[1620; 1625[[1625; 1630[[1630; 1635[
Effectifs	7	8	42	75	18

- 1) En convenant que les valeurs mesurées sont regroupées au centre de chaque classe, donner une estimation ponctuelle de la moyenne et la variance de la quantité de bicarbonate de sodium dans la population formée de l'ensemble de tous les comprimés fabriqués et supposée très grande.
- 2) Déterminer une estimation par intervalle de confiance à 95% de la quantité moyenne de bicarbonate de sodium dans la population.



Exercices

Exercice 7 (Estimation) :

On a mesuré le poids de raisin produit par m^2 sur $10\ m^2$ pris au hasard dans une vigne. On suppose que le poids de raisin produits par une souche de cette vigne suit une loi normale $\mathcal{N}(\mu, \sigma^2)$. On a obtenu les résultats suivants exprimés en (Kg) :

24	34	36	41	43	47	54	59	65	69
----	----	----	----	----	----	----	----	----	----

- 1) Déterminer une estimation ponctuelle de la moyenne théorique μ et de la variance théorique σ^2 .
- 2) Déterminer une estimation par intervalle de confiance à 95% de μ .
- 3) Déterminer une estimation par intervalle de confiance à 95% de σ^2 .

Plan :

- 1 Introduction
- 2 Statistique descriptive univariée
- 3 Statistique descriptive bivariée
- 4 Notions de probabilités et variables aléatoires
- 5 Échantillonnage et estimation
- 6 Tests des hypothèses
- 7 Exercices de révisions
- 8 Examens (sessions normale et de rattrapage 2016-2017)

Tests des hypothèses

Introduction

La majorité des tests repose sur le principe suivant :

- On définit une hypothèse nulle notée H_0 contre l'hypothèse alternative H_1 .

Test : $\begin{cases} \text{hypothèse nulle } H_0 \\ \text{hypothèse alternative } H_1 \end{cases}$

Le test a pour objectif d'accepter ou de rejeter H_0 avec un risque connu à partir des données dont on dispose.

Tests des hypothèses

- On détermine alors une statistique qui est une variable aléatoire construite à partir des données.
 - Sous H_0 , cette variable suit une loi de probabilité connue (normale, student, khi-deux,...)
 - On détermine alors l'intervalle de confiance dont lequel doit tomber la statistique avec une probabilité donnée ($1 - \alpha$), (le plus souvent 95% pour $\alpha = 5\%$).

Tests des hypothèses

- On définit alors la règle de décision suivante :
 - Si la statistique tombe dans l'intervalle, on accepte H_0 .
Attention, cela ne veut pas dire que H_0 est vraie mais que le test et les données ne permettent pas de voir un écart significatif à H_0 .
 - Si la statistique ne tombe pas dans l'intervalle, on rejette H_0 avec le risque α de se tromper, (par exemple $\alpha = 5\%$).

Tests des hypothèses

1) Test de conformité de la proportion à une référence

Test bilatéral : $\begin{cases} H_0 : p = p_0, \\ H_1 : p \neq p_0. \end{cases}$

On calcule :

$$u = \frac{\bar{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}.$$

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

- ① Si $u \in]-z_\alpha; z_\alpha[$, on ne peut rejeter H_0 .
- ② Sinon, on rejette H_0 avec une probabilité α de se tromper.

Tests des hypothèses

Exemple

Sur un échantillon de taille $n = 400$ de naissances, on a observé 206 mâles, soit une proportion de mâles de $\bar{p} = 206/400 = 0.515$. On se demande si il y a autant de mâles que de femelles. i.e., si $p_0 = 0.5$. On peut effectuer alors le test :

$$\begin{cases} H_0 : p = p_0 = 0.5, \\ H_1 : p \neq p_0 = 0.5. \end{cases}$$

Tests des hypothèses

On calcule :

$$u = \frac{\bar{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.515 - 0.5}{\sqrt{\frac{0.5(1-0.5)}{400}}} = 0.6.$$

Pour $\alpha = 5\% = 0.05$, on a $z_\alpha = 1.96$.

Comme $u \in]-z_\alpha; z_\alpha[$, alors on ne peut rejeter H_0 : donc il est possible que $p = 0.5$.

Tests des hypothèses

2) Test de conformité de la moyenne à une référence Cas d'un petit échantillon $n < 30$ où σ est inconnue

Test bilatéral : $\begin{cases} H_0 : \mu = \mu_0, \\ H_1 : \mu \neq \mu_0. \end{cases}$

On calcule :

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}.$$

On détermine t_α pour la loi student à $n - 1$ ddl, ($P(T > t_\alpha) = \frac{\alpha}{2}$), et on décide que :

- ① Si $t \in] - t_\alpha; t_\alpha [$, on ne peut rejeter H_0 .
- ② Sinon, on rejette H_0 avec une probabilité α de se tromper.

Tests des hypothèses

Exemple

Soit $n = 25$, $X \sim \mathcal{N}(\mu, \sigma^2)$, $\bar{x} = 15$ et $s = 9$.

Peut-on affirmer que la moyenne inconnue de la population est 12.

Test bilatéral : $\begin{cases} H_0 : \mu = \mu_0 = 12, \\ H_1 : \mu \neq \mu_0 = 12. \end{cases}$

On calcule :

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{15 - 12}{\frac{9}{\sqrt{25}}} = 1.67.$$

Le ddl est 24, donc pour $\alpha = 5\% = 0.05$, on trouve $t_\alpha = 2.07$.

Comme $t \in] - t_\alpha; t_\alpha [$, on ne peut rejeter H_0 .

Tests des hypothèses

Cas d'un grand échantillon $n \geq 30$ où σ est inconnue

Test bilatéral : $\left\{ \begin{array}{l} H_0 : \mu = \mu_0, \\ H_1 : \mu \neq \mu_0. \end{array} \right.$

On calcule :

$$u = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}.$$

On détermine z_α pour la loi normale $\mathcal{N}(0, 1)$, et on décide que :

- ① Si $u \in] -z_\alpha; z_\alpha [$, on ne peut rejeter H_0 .
- ② Sinon, on rejette H_0 avec une probabilité α de se tromper.

Tests des hypothèses

3) Test de conformité de la variance à une référence

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

Ici on travaille avec la loi Khi-deux à $n - 1$ ddl. On cherche les deux nombres a et b tels que :

$P(Y^2 \geq a) = \frac{\alpha}{2}$ et $P(Y^2 \geq b) = 1 - \frac{\alpha}{2}$. On calcule :

$$y^2 = \frac{n-1}{\sigma_0^2} s^2.$$

- ① Si $y^2 \in]a; b[$, on ne peut rejeter H_0 .
- ② Sinon, on rejette H_0 avec une probabilité α de se tromper.

Tests des hypothèses

Exemple

Test bilatéral : $\left\{ \begin{array}{l} H_0 : \sigma^2 = \sigma_0^2 = 90, \\ H_1 : \sigma^2 \neq \sigma_0^2 = 90. \end{array} \right.$

Pour $n = 31$, $s^2 = 100$, on a le ddl = $31 - 1 = 30$. $\alpha = 0.05$ donc $\frac{\alpha}{2} = 0.025$ donne $a = 46.98$ et $1 - \frac{\alpha}{2} = 0.975$ donne $b = 16.80$.

$$y^2 = \frac{n-1}{\sigma_0^2} s^2 = \frac{30}{90} \times 100 = 33.34.$$

Comme $y^2 \in]a; b[$, on ne peut rejeter H_0 .

Tests des hypothèses

4) Test d'indépendance du khi-deux entre deux variables qualitatives

Le test du khi-deux est largement utilisé pour l'étude de l'indépendance entre deux caractères qualitatifs.

La présentation des résultats se fait sous forme d'un tableau de contingence à deux entrées.

Chaque entrée représente les modalités d'une des variables. On détermine alors le tableau attendu sous l'hypothèse d'indépendance.

Tests des hypothèses

Exemple

Un échantillon de $n = 1000$ personnes ont été interrogées sur une question précise. On a demandé à ces personnes de préciser leur sexe. Nous voudrions savoir si la réponse Y à la question est indépendante du sexe X .

X/Y	Favorable	Défavorable	Indécis	Total
Masculin	$n_{11} = 210$	$n_{12} = 194$	$n_{13} = 91$	$n_{1\bullet} = 495$
Féminin	$n_{21} = 292$	$n_{22} = 151$	$n_{23} = 62$	$n_{2\bullet} = 505$
Total	$n_{\bullet 1} = 502$	$n_{\bullet 2} = 345$	$n_{\bullet 3} = 153$	$n = 1000$

Tests des hypothèses

Sous hypothèse d'indépendance, la distribution conjointe est simplement le produit des distributions marginales, i.e., $f_{ij} = f_{i\bullet}f_{\bullet j}$. Si l'on estime f_{ij} par $\frac{n_{ij}}{n}$ et $f_{i\bullet}$ par $\frac{n_{i\bullet}}{n}$ et $f_{\bullet j}$ par $\frac{n_{\bullet j}}{n}$, on devrait donc avoir :

$$n_{ij} \approx \frac{n_{i\bullet}n_{\bullet j}}{n}.$$

L'idée est de calculer l'écart entre les deux termes : le n_{ij} observé et le n_{ij} prédit ou théorique, si cet écart devient trop important, on devra rejeter l'hypothèse que les variables sont indépendantes.

Tests des hypothèses

On calcule :

$$Q = \sum_{i=1}^k \sum_{j=1}^l \frac{\left(n_{ij} - \frac{n_{i\bullet} n_{\bullet j}}{n} \right)^2}{\frac{n_{i\bullet} n_{\bullet j}}{n}}$$

La statistique Q est distribuée suivant une khi-deux à $(k - 1)(l - 1)$ ddl où k et l désignent le nombre de valeurs des deux variables.

Test de validité du test : il faut que la fréquence théorique $\frac{n_{i\bullet} n_{\bullet j}}{n} \geq 5$ pour tous i et j .

Nous acceptons H_0 si $Q \leq q_\alpha$ avec $P(Y^2 > q_\alpha) = \alpha$.

Tests des hypothèses

Dans notre exemple, pour $\alpha = 0.05$, on a le ddl est $(3 - 1) \times (2 - 1) = 2$, $q_\alpha = 5.99$ et la valeur de Q est :

$$Q = \frac{\left(\frac{495 \times 502}{1000} - 210\right)^2}{\frac{495 \times 502}{1000}} + \frac{\left(\frac{495 \times 345}{1000} - 194\right)^2}{\frac{495 \times 345}{1000}} + \frac{\left(\frac{495 \times 153}{1000} - 91\right)^2}{\frac{495 \times 153}{1000}} + \\ \frac{\left(\frac{505 \times 502}{1000} - 292\right)^2}{\frac{505 \times 502}{1000}} + \frac{\left(\frac{505 \times 345}{1000} - 151\right)^2}{\frac{505 \times 345}{1000}} + \frac{\left(\frac{505 \times 153}{1000} - 62\right)^2}{\frac{505 \times 153}{1000}} = 24.15.$$

Finalement on rejette H_0 avec un risque $\alpha = 0.05$ de se tromper.

Tests des hypothèses

5) Tests d'homogénéité dans le cas des échantillons indépendants

Dans deux populations P_1 et P_2 , on étudie un même caractère. On cherche à comparer les deux populations quant à ce caractère pour savoir si elles sont homogènes ou pas.

Tests des hypothèses

Test de comparaison de deux proportions

Soit p_1 (resp. p_2) la proportion d'individus ayant la propriété dans P_1 (resp. dans P_2). On extrait deux échantillons indépendants E_1 et E_2 de tailles resp. n_1 et n_2 .

Test bilatéral : $\begin{cases} H_0 : p_1 = p_2 = p, \\ H_1 : p_1 \neq p_2. \end{cases}$

Dans l'échantillon E_1 de taille n_1 on estime la proportion p_1 par \bar{p}_1 et dans l'échantillon E_2 de taille n_2 on estime la proportion p_2 par \bar{p}_2 et en regroupant les deux échantillons, on peut estimer p par :

$$\bar{p} = \frac{n_1 \bar{p}_1 + n_2 \bar{p}_2}{n_1 + n_2}.$$

Tests des hypothèses

Test de validité du test : $n_1\bar{p}_1 \geq 5$, $n_1(1 - \bar{p}_1) \geq 5$, $n_2\bar{p}_2 \geq 5$ et $n_2(1 - \bar{p}_2) \geq 5$.

On calcule :

$$u = \frac{\bar{p}_1 - \bar{p}_2}{\sqrt{(\frac{1}{n_1} + \frac{1}{n_2})\bar{p}(1 - \bar{p})}}.$$

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

- ① Si $u \in] - z_\alpha; z_\alpha [$, on ne peut rejeter H_0 .
- ② Sinon, on rejette H_0 avec une probabilité α de se tromper.

Tests des hypothèses

Exemple

Dans une même catégorie sociale, un échantillon de 40 hommes a fourni 8 fumeurs et un échantillon de 60 femmes a fourni 18 fumeuses. On se demande si la proportion de fumeurs est la même pour les deux sexes.

$$\text{Test bilatéral : } \begin{cases} H_0 : p_1 = p_2, \\ H_1 : p_1 \neq p_2. \end{cases}$$

$$n_1 = 40, \bar{p}_1 = \frac{8}{40} = 0.2, n_2 = 60 \text{ et } \bar{p}_2 = \frac{18}{60} = 0.3.$$

$$n_1\bar{p}_1 = 8 \geq 5, n_1(1 - \bar{p}_1) = 32 \geq 5, n_2\bar{p}_2 = 18 \geq 5 \text{ et } n_2(1 - \bar{p}_2) = 42 \geq 5.$$

$$\bar{p} = 0.26, u = -1.12 \text{ et pour } \alpha = 5\% = 0.05, z_\alpha = 1.96.$$

Comme $u \in] -z_\alpha; z_\alpha [$, alors on ne peut rejeter H_0 : la proportion de fumeurs ne diffère pas significativement entre les deux sexes.

Tests des hypothèses

Test de comparaison de deux variances

$$\text{Test bilatéral : } \begin{cases} H_0 : \sigma_1^2 = \sigma_2^2, \\ H_1 : \sigma_1^2 \neq \sigma_2^2. \end{cases}$$

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

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

On calcule : $f = \frac{s_1^2}{s_2^2}$ telle que $f \geq 1$ sinon on permute les échantillons de sorte que $f \geq 1$.

Dans la table de la loi F de Fisher Snédécor et avec le ddl $(n_1 - 1, n_2 - 1)$, on cherche f_α tel que $P(F \geq f_\alpha) = \frac{\alpha}{2}$, et on décide que :

- ① Si $f < f_\alpha$, on ne peut rejeter H_0 .
- ② Sinon, on rejette H_0 avec une probabilité α de se tromper.

Tests des hypothèses

Exemple

Dans un article de la revue "Biometrika", le biologiste Latter donne la longueur (en mm) des oeufs de Coucou trouvés dans les nids de deux espèces d'oiseaux :

- dans des nids de petite taille (Roitelet) : $\left\{ \begin{array}{llllllll} 19,8 & 22,1 & 21,5 & 20,9 & 22,0 & 21,0 & 22,3 & 21,0 \\ 20,3 & 20,9 & 22,0 & 22,0 & 20,8 & 21,2 & 21,0 \end{array} \right.$

- dans des nids de taille plus grande (Fauvette) : $\left\{ \begin{array}{llllllll} 22,0 & 23,9 & 20,9 & 23,8 & 25,0 & 24,0 & 23,8 \\ 21,7 & 22,8 & 23,1 & 23,5 & 23,0 & 23,0 & 23,1 \end{array} \right.$

On se demande si le Coucou adapte la taille de ses oeufs à la taille du nid.

Tests des hypothèses

Observation sur l'échantillon E_1 de taille $n_1 = 15$: $\bar{x}_1 = 21.25$ et $s_1^2 = 0.516$.

Observation sur l'échantillon E_2 de taille $n_2 = 14$: $\bar{x}_2 = 23.11$ et $s_2^2 = 1.101$.

Comme $f = \frac{s_1^2}{s_2^2} < 1$, on permute les deux échantillons et on prend

$$f' = \frac{s_2^2}{s_1^2} = 2.14 > 1.$$

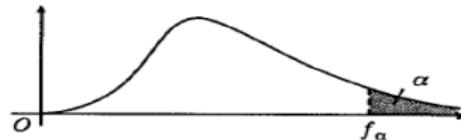
Le ddl de la loi F de Fisher Snédécor est $(n_2 - 1, n_1 - 1) = (13, 14)$ et pour $\alpha = 5\%$ on trouve que la valeur de f_α telle que $P(F \geq f_\alpha) = \frac{\alpha}{2}$ est compris entre 2.95 et 3.15.

Comme $f' < f_\alpha$, on ne peut rejeter H_0 et les variances des deux populations ne sont pas différentes significativement au risque 5%.

Tests des hypothèses

Lois de Fisher ($\alpha = 0,025$)

Si F est une variable aléatoire qui suit la loi de Snédécor à (v_1, v_2) degrés de liberté, la table donne le nombre f_α tel que $P(F \geq f_\alpha) = \alpha = 0,025$.



$v_1 \backslash v_2$	1	2	3	4	5	6	8	10	15	20	30	∞
1	648	800	864	900	922	937	957	969	985	993	1 001	1 018
2	38,5	39,0	39,2	39,2	39,3	39,3	39,4	39,4	39,4	39,4	39,5	39,5
3	17,4	16,0	15,4	15,1	14,9	14,7	14,5	14,4	14,3	14,2	14,1	13,9
4	12,2	10,6	9,98	9,60	9,36	9,20	8,98	8,84	8,66	8,56	8,46	8,26
5	10,0	8,43	7,76	7,39	7,15	6,98	6,76	6,62	6,43	6,33	6,23	6,02
6	8,81	7,26	6,60	6,23	5,99	5,82	5,60	5,46	5,27	5,17	5,07	4,85
7	8,07	6,54	5,89	5,52	5,29	5,12	4,90	4,76	4,57	4,47	4,36	4,14
8	7,57	6,06	5,42	5,05	4,82	4,65	4,43	4,30	4,10	4,00	3,89	3,67
9	7,21	5,71	5,08	4,72	4,48	4,32	4,10	3,96	3,77	3,67	3,56	3,33
10	6,94	5,46	4,83	4,47	4,24	4,07	3,85	3,72	3,52	3,42	3,31	3,08
11	6,72	5,26	4,63	4,28	4,04	3,88	3,66	3,53	3,33	3,23	3,12	2,88
12	6,55	5,10	4,47	4,12	3,89	3,73	3,51	3,37	3,18	3,07	2,96	2,72
13	6,41	4,97	4,35	4,00	3,77	3,60	3,39	3,25	3,05	2,95	2,84	2,60
14	6,30	4,86	4,24	3,89	3,66	3,50	3,29	3,15	2,95	2,84	2,73	2,49
15	6,20	4,76	4,15	3,80	3,58	3,41	3,20	3,06	2,86	2,76	2,64	2,40

Tests des hypothèses

Test de comparaison de deux moyennes

Test bilatéral : $\begin{cases} H_0 : \mu_1 = \mu_2, \\ H_1 : \mu_1 \neq \mu_2. \end{cases}$

- Cas de grands échantillons indépendants $n_1 > 30$ et $n_2 > 30$

On calcule :

$$u = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}}.$$

On détermine z_α pour la loi normale $\mathcal{N}(0, 1)$, et on décide que :

- ① Si $u \in]-z_\alpha; z_\alpha[$, on ne peut rejeter H_0 .
- ② Sinon, on rejette H_0 avec une probabilité α de se tromper.

Tests des hypothèses

- Cas de petits échantillons extraits de populations gaussiennes $n_1 < 30$ et $n_2 < 30$ et tel que $\sigma = \sigma_1 = \sigma_2$

La variance commune σ^2 peut être estimée par :

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}.$$

On calcule :

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}.$$

On détermine t_α dans la table de Student à $n_1 + n_2 - 2$ ddl, et on décide que :

- ➊ Si $t \in] - t_\alpha; t_\alpha [$, on ne peut rejeter H_0 .
- ➋ Sinon, on rejette H_0 avec une probabilité α de se tromper.

Tests des hypothèses

Exemple

Dans un article de la revue "Biometrika", le biologiste Latter donne la longueur (en mm) des oeufs de Coucou trouvés dans les nids de deux espèces d'oiseaux :

- dans des nids de petite taille (Roitelet) :
$$\left\{ \begin{array}{cccccccc} 19,8 & 22,1 & 21,5 & 20,9 & 22,0 & 21,0 & 22,3 & 21,0 \\ 20,3 & 20,9 & 22,0 & 22,0 & 20,8 & 21,2 & 21,0 \end{array} \right.$$
- dans des nids de taille plus grande (Fauvette) :
$$\left\{ \begin{array}{cccccccc} 22,0 & 23,9 & 20,9 & 23,8 & 25,0 & 24,0 & 23,8 \\ 21,7 & 22,8 & 23,1 & 23,5 & 23,0 & 23,0 & 23,1 \end{array} \right.$$

On se demande si le Coucou adapte la taille de ses oeufs à la taille du nid.

Tests des hypothèses

Observation sur l'échantillon E_1 de taille $n_1 = 15$: $\bar{x}_1 = 21.25$ et $s_1^2 = 0.516$.

Observation sur l'échantillon E_2 de taille $n_2 = 14$: $\bar{x}_2 = 23.11$ et $s_2^2 = 1.101$.

On est dans le cas des petits échantillons $n_1 < 30$, $n_2 < 30$.

$s^2 = 0.798$, $t = -5.61$ et le ddl de la loi Student est $n_1 + n_2 - 2 = 27$. Pour pour $\alpha = 0.05$, on trouve $t_\alpha = 2.052$. Comme t n'est pas dans $[-t_\alpha; t_\alpha]$, on rejette H_0 avec une probabilité α de se tromper. La taille moyenne des oeufs de Coucou sont différentes dans les nids de Roitelet et de Fauvettes.

Tests des hypothèses

6) Test de normalité : test de Shapiro et Wilk

La plupart des méthodes de test paramétriques requièrent la normalité des données. Il est donc important de disposer d'une méthode permettant de vérifier cette normalité. Il existe plusieurs tests pour vérifier la normalité des données observées. Nous présentons ici le test de Shapiro et Wilk pour une série d'observations d'une variable quantitative.

NB : Ce test est valable pour des petits échantillons ($n \leq 30$).

Tests des hypothèses

Exemple

Nous avons réalisé $n = 10$ mesures d'un alésage en (mm). On a obtenu les résultats suivants :

12.124; 12.230; 12.327; 12.242; 12.466

12.215; 12.026; 12.359; 12.215; 12.387

On désire tester la normalité de cette série de résultats.

Tests des hypothèses

Démarche de vérification :

- Classer les différentes valeurs de la série par ordre croissant :

12.026; 12.124; 12.215; 12.215; 12.230

12.242; 12.327; 12.359; 12.387; 12.466

- Calculer la moyenne \bar{x} de la série de mesure :

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = 12.259.$$

- Calculer le nombre :

$$T^2 = \sum_{i=1}^n (x_i - \bar{x})^2 = 0.1514.$$

Tests des hypothèses

- **Calculer les différences respectives d_i :**

$$d_1 = x_{10} - x_1 = 12.466 - 12.026 = 0.440$$

$$d_2 = x_9 - x_2 = 12.387 - 12.124 = 0.263$$

$$d_3 = x_8 - x_3 = 12.359 - 12.215 = 0.144$$

$$d_4 = x_7 - x_4 = 12.327 - 12.215 = 0.112$$

$$d_5 = x_6 - x_5 = 12.242 - 12.230 = 0.012$$

Tests des hypothèses

- A chacune de ces différences d_i , on affecte le coefficient a_i , donné par la table ci-dessus et on calcule $d_i * a_i$:

Figure – Valeurs des a_i :

n \ J	2	3	4	5	6	7	8	9	10
1	0.7071	0.7071	0.6872	0.6646	0.6431	0.6233	0.6052	0.5888	0.5739
2		0.0000	0.1677	0.2413	0.2806	0.3031	0.3164	0.3244	0.3291
3				0.0000	0.0875	0.1401	0.1743	0.1976	0.2141
4						0.0000	0.0561	0.0947	0.1224
5								0.0000	0.0399

Tests des hypothèses

- **Calculer les nombres :**

$$d_1 * a_1 = 0.440 * 0.5739 = 0.2525$$

$$d_2 * a_2 = 0.263 * 0.3291 = 0.0865$$

$$d_3 * a_3 = 0.144 * 0.2141 = 0.0308$$

$$d_4 * a_4 = 0.112 * 0.1224 = 0.0137$$

$$d_5 * a_5 = 0.012 * 0.0399 = 0.0005$$

- **Calculer la valeur :**

$$b = \sum_{i=1}^5 d_i * a_i = 0.384$$

- **Calculer le rapport :**

$$W = \frac{b^2}{T^2} = 0.9739$$

Tests des hypothèses

- Par exemple pour $\alpha = 0.05$, on trouve $w_\alpha = 0.842$ dans la table suivante de Shapiro et Wilk.

Table de Shapiro et Wilk

n	Risque 5 %	Risque 1 %
	$W_{0,05}$	$W_{0,01}$
5	0,762	0,686
6	0,788	0,713
7	0,803	0,730
8	0,818	0,749
9	0,829	0,764
10	0,842	0,781
11	0,850	0,792
12	0,859	0,805

Tests des hypothèses

- Si $W > w_\alpha$, on accepte la normalité des données, sinon on rejette cette hypothèse. Dans notre exemple, on a : $W = 0.9739 > w_\alpha = 0.842$, l'hypothèse de normalité est acceptée.

Tests des hypothèses

7) Analyse de la variance : ANOVA à un facteur

Il s'agit d'une généralisation à k populations de tailles $(n_j)_{1 \leq j \leq k}$ des tests de comparaison de moyennes de deux échantillons. Elle permet d'étudier l'effet d'une variable qualitative (facteur) sur une variable quantitative. On utilise des mesures de variance afin de déterminer le caractère significatif, ou non, des différences de moyenne mesurées sur les populations.

L'analyse de variance suppose l'homogénéité des variances et la normalité des données.

$$\left\{ \begin{array}{l} H_0 : \mu_1 = \mu_2 = \dots = \mu_k, \\ H_1 : \text{au moins l'une des moyennes diffère des autres.} \end{array} \right.$$

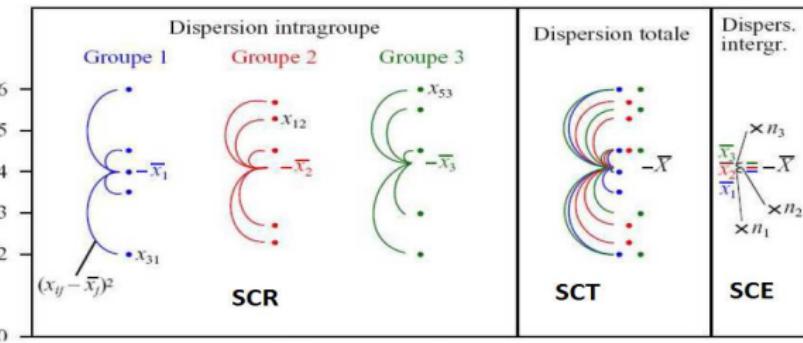
Tests des hypothèses

Exemple

Soit $k = 3$ populations de taille $n_1 = n_2 = n_3 = 5$ et $n = n_1 + n_2 + n_3 = 15$.

Critère de classification →

Observations	Groupe 1	Groupe 2	Groupe 3
	4,0	5,3	2,0
	6,0	2,7	3,0
	2,0	4,5	4,5
	4,5	2,3	5,5
	3,5	5,7	6,0



Tests des hypothèses

Les moyennes :

$$\bar{x}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} x_{ij} \quad \bar{x} = \frac{1}{n} \sum_{j=1}^k n_j \bar{x}_j$$

La variation totale :

$$SCT = \sum_{j=1}^k \sum_{i=1}^{n_j} (x_{ij} - \bar{x})^2.$$

La variation intragroupe résiduelle non expliquée par le facteur :

$$SCR = \sum_{j=1}^k \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2.$$

Tests des hypothèses

La variation intergroupe expliquée par le facteur :

$$SCE = \sum_{j=1}^k \sum_{i=1}^{n_j} (\bar{x}_j - \bar{x})^2 = \sum_{j=1}^k n_j (\bar{x}_j - \bar{x})^2.$$

Équation d'ANOVA :

$$SCT = SCR + SCE.$$

On pose :

$$V_E = \frac{SCE}{k-1} = \frac{SCE}{ddl} \qquad V_I = \frac{SCR}{n-k} = \frac{SCR}{ddl} \qquad F = \frac{V_E}{V_I}.$$

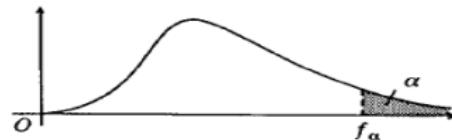
Règle de décision : Pour un seuil α , on cherche dans la table de Fisher

Snédécor la valeur $F_{(\alpha, k-1, n-k)}$. On accepte H_0 si $F < F_{(\alpha, k-1, n-k)}$ ce qui est le cas dans notre exemple.

Tests des hypothèses

Lois de Fisher ($\alpha = 0,05$)

Si F est une variable aléatoire qui suit la loi de Snédécor à (v_1, v_2) degrés de liberté, la table donne le nombre f_α tel que $P(F \geq f_\alpha) = \alpha = 0,05$.



$v_1 \backslash v_2$	1	2	3	4	5	6	8	10	15	20	30	∞
1	161	200	216	225	230	234	239	242	246	248	250	254
2	18,5	19,0	19,2	19,2	19,3	19,3	19,4	19,4	19,4	19,4	19,5	19,5
3	10,1	9,55	9,28	9,12	9,01	8,94	8,85	8,79	8,70	8,66	8,62	8,53
4	7,71	6,94	6,59	6,39	6,26	6,16	6,04	5,96	5,86	5,80	5,75	5,63
5	6,61	5,79	5,41	5,19	5,05	4,95	4,82	4,74	4,62	4,56	4,50	4,36
6	5,99	5,14	4,76	4,53	4,39	4,28	4,15	4,06	3,94	3,87	3,81	3,67
7	5,59	4,74	4,35	4,12	3,97	3,87	3,73	3,64	3,51	3,44	3,38	3,23
8	5,32	4,46	4,07	3,84	3,69	3,58	3,44	3,35	3,22	3,15	3,08	2,93
9	5,12	4,26	3,86	3,63	3,48	3,37	3,23	3,14	3,01	2,94	2,86	2,71
10	4,96	4,10	3,71	3,48	3,33	3,22	3,07	2,98	2,85	2,77	2,70	2,54
11	4,84	3,98	3,59	3,36	3,20	3,09	2,95	2,85	2,72	2,65	2,57	2,40
12	4,75	3,89	3,49	3,26	3,11	3,00	2,85	2,75	2,62	2,54	2,47	2,30
13	4,67	3,81	3,41	3,18	3,03	2,92	2,77	2,67	2,53	2,46	2,38	2,21
14	4,60	3,74	3,34	3,11	2,96	2,85	2,70	2,60	2,46	2,39	2,31	2,13
15	4,54	3,68	3,29	3,06	2,90	2,79	2,64	2,54	2,40	2,33	2,25	2,07

Tests des hypothèses

$$\begin{array}{lllll}
 n_1 = 5 & n_2 = 5 & n_3 = 5 & n = 15 & \sum n_j (\bar{x}_j - \bar{X})^2 = 0,10 \\
 T_1 = 20,0 & T_2 = 20,5 & T_3 = 21,0 & T = 61,5 & \downarrow \\
 \bar{x}_1 = 4,0 & \bar{x}_2 = 4,1 & \bar{x}_3 = 4,2 & \bar{X} = 4,1 & \text{SCE} \\
 \sum (x_{ij} - \bar{x}_i)^2 = 8,50 & & \sum (x_{ij} - \bar{x}_3)^2 = 11,30 & \sum (x_{ij} - \bar{X})^2 = 29,26 & \\
 \sum (x_{ij} - \bar{x}_2)^2 = 9,36 & & & & \downarrow \\
 & & & & \text{SCT} \\
 & & & & \text{SCR} = 29,16 \\
 & & & & \text{SCT} = \text{SCR} + \text{SCE}
 \end{array}$$

Sources de variation	Dispersions	Degrés de liberté	Variances
Totale	$\text{SCT} = 29,26$	$15 - 1 = 14$	$29,26/14 = 2,09$
Intergroupe	$\text{SCE} = 0,10$	$3 - 1 = 2$	$V_E = 0,10/2 = 0,05$
Intragroupe	$\text{SCR} = 29,16$	$15 - 3 = 12$	$V_I = 29,16/12 = 2,43$

$$F = V_E/V_I = 0,0206 \quad P = 0,9797$$

$$F_{(0,05,2,12)} = 3,89$$

Tests des hypothèses

Remarque

Pourquoi ne pas réaliser une série de tests pour comparer la moyenne de toutes les paires de groupes : Par exemple, considérons 7 groupes d'observations tirées indépendamment d'une même population statistique. Il faudrait réaliser $C_7^2 = 21$ tests pour comparer toutes les paires de groupes. Chaque test étant réalisé au niveau $\alpha = 0.05$, on a, dans chaque cas, 5 chances sur 100 de rejeter H_0 même si H_0 est vraie. La probabilité de rejeter H_0 au moins une fois au cours de 21 tests est 0.66 et non 0.05 (calcul basé sur distribution binomiale). Pour être valide, le test global doit avoir une erreur 0.05. La solution de ce problème est la méthode ANOVA.

Exercices

Exercice 1 :

On admet que le taux de cholestérol chez une femme suit une loi normale $\mathcal{N}(\mu, \sigma^2)$. Sur un échantillon de 10 femmes, on a obtenu les résultats suivants :

$$\sum_{i=1}^{10} x_i = 21.3 \quad \sum_{i=1}^{10} x_i^2 = 47.49$$

- 1. Déterminer une estimation ponctuelle de la moyenne et de la variance du taux.
- 2. Déterminer un intervalle de confiance pour la moyenne du taux au seuil 5%.
- Tester au seuil 5% l'hypothèse que la moyenne de la population est 2.

Exercices

Exercice 2 :

Dans une usine du secteur de l'agroalimentaire, une machine à embouteiller est alimentée par un réservoir d'eau et par une file d'approvisionnement en bouteilles vides. Pour contrôler le bon fonctionnement de la machine. Pour une production d'une heure, on suppose que la variable aléatoire X qui à toute bouteille, prise au hasard dans cette production, associe le volume d'eau (en litres) qu'elle contient, est une variable aléatoire despérance μ et d'écart-type σ inconnus. On a prélevé un échantillon de 100 bouteilles, et on a obtenu un volume d'eau moyen : 1,495 l et un écart-type corrigé de 0,01.

- 1)- Déterminer un intervalle de confiance pour la moyenne au seuil 1 %.
- 2)- Tester au seuil 1% si la moyenne vaut 1.5 l.

Exercices

Exercice 3 :

On sait qu'une maladie atteint 10% des individus d'une population P donnée.
Un chercheur a expérimenté un traitement sur un échantillon de n individus : il a alors recensé 5% de malades.

- 1. Déterminer la valeur maximale de n qui permette au chercheur de conclure à l'efficacité du traitement au risque de 5%
- 2. Déterminer un intervalle de confiance pour la proportion au seuil 1%.

Exercices

Exercice 4 :

Le volume d'une pipette d'un type donné suit une loi normale $\mathcal{N}(\mu, \sigma)$. Le fabricant annonce un écart-type $\sigma_0 = 0.2\mu l$. Pour le vérifier, on pipette 20 fois un liquide, on observe une moyenne de $10\mu l$ et un écart-type de $0.4\mu l$.

- 1)- Déterminer un intervalle de confiance pour la variance au seuil 5 %.
- 2)- Tester au seuil 5 % si l'écart type vaut 0.2.

Exercices

Exercice 5 :

On s'intéresse à une éventuelle relation entre X : le sexe de $n = 200$ personnes et Y : la couleur des yeux.

X/Y	Bleu	Vert	Marron
Homme	$n_{11} = 10$	$n_{12} = 50$	$n_{13} = 20$
Femme	$n_{21} = 20$	$n_{22} = 60$	$n_{23} = 40$

1. Appliquer le test du khi-deux pour tester l'indépendance de X et Y .

Exercices

Exercice 6 :

Nous avons réalisé 10 dosages. On a obtenu les résultats suivants :

60 80 55 45 60 65 65 60 70 40

1. Utiliser le test de Shapiro et Wilk pour tester la normalité de ces données.

Exercices

Exercice 7 :

On compare les effets d'un même traitement dans deux hôpitaux différents. Dans le premier hôpital, 70 des 100 malades traités montrent des signes de guérison. Dans le deuxième hôpital, c'est le cas pour 100 des 150 malades traités.

1. Quelle conclusion peut-on en tirer au risque de 5% ? (comparer les proportions).

Exercices

Exercice 8 :

Sur deux groupes de même taille : 10 malades, on expérimente les effets d'un traitement destiné à diminuer la pression artérielle. On observe les résultats suivants (valeurs de la tension artérielle systolique en cm Hg). On supposera les populations gaussiennes.

Groupe 1	15	18	17	20	21	18	17	15	19	16
Groupe 2	12	16	17	18	17	15	18	14	16	18

1. Le traitement a-t-il une action significative, au risque de 5% ?
(comparer les variances puis comparer les moyennes).

Exercices

Exercice 9 (ANOVA) :

Pour définir l'impact de la nature du sol sur la croissance d'une plante X, un botaniste a mesuré la hauteur des plantes pour 4 types de sol. Pour chaque type de sol, il disposait de 3 réplicats.

1. La croissance de plante X est-elle dépendante de la nature du sol ?

Types de sol			
1	2	3	4
15	25	17	10
9	21	23	13
4	19	20	16

Plan :

- 1** Introduction
- 2** Statistique descriptive univariée
- 3** Statistique descriptive bivariée
- 4** Notions de probabilités et variables aléatoires
- 5** Échantillonnage et estimation
- 6** Tests des hypothèses
- 7** Exercices de révisions
- 8** Examens (sessions normale et de rattrapage 2016-2017)

Exercices de révisions

Exercice 1 (Variable continue et construction des classes) :

On a relevé les salaires annuels en (DH) de $n = 30$ personnes :

1860	2010	2110	2380	2600	2770	2770	2810	2920	2950
3180	3250	3250	3280	3360	4310	4320	4960	5430	5670
5710	5850	6230	6250	6960	7470	7880	8710	9440	9590

- 1. De quel type est la variable salaire ? Déterminer sa médiane. Interpréter.
- 2. Construire le tableau statistique en adoptant des classes de même amplitude selon la règle de Sturge.
- 3. Construire l'histogramme des fréquences.
- 4. Déterminer la classe modale et les centres des classes.
- 5. En déduire la moyenne, la variance et l'écart type de la variable salaire.

Exercices de révisions

Exercice 2 (Régression linéaire) :

On considère la série double statistique suivante :

x_i	2	3	5	1	4
y_i	4	9	11	3	8

- 1. De quel type sont les variables X et Y.
- 2. Déterminer les médianes, les moyennes et les variances de X et Y.
- 3. Déterminer la covariance et le coefficient de corrélation entre X et Y. Donner une interprétation.
- 4. Déterminer la droite de régression linéaire de Y en X.

Exercices de révisions

- 5. Tracer le nuage de points et la droite de régression linéaire de Y en X .
- 6. Vérifier que la droite de régression passe par le point (\bar{x}, \bar{y}) .
- 7. Établir, sur la base de ce modèle, la valeur y^* correspond à la valeur $x^* = 3,5$.
- 8. Déterminer la variance résiduelle et le coefficient de détermination.
Interpréter.

Exercices de révisions

Exercice 3 (Loi de probabilité discrète) :

On lance deux trièdres équilibrés numérotés : 1; 2; 3; 4. On note X la variable aléatoire qui donne le plus grand des deux numéros obtenus.

- 1. Donner la loi de X . En déduire : $P(X \leq 2)$, $E(X)$, $V(X)$ et $\sigma(X)$.

Exercices de révisions

Exercice 4 (Loi de probabilité continue) :

Soit X une variable aléatoire continue qui suit la loi de Pareto de densité :

$$f(x) = \frac{1_{(x>1)}}{x^2}.$$

- 1. Vérifier que $\int_{-\infty}^{+\infty} f(x)dx = 1$ puis calculer $E(X)$.

Exercices de révisions

Exercice 5 (Comparaisons des proportions) :

Pour traiter un certain type de tumeur, on a utilisé deux schémas thérapeutiques :

- Sur 40 malades traités avec le schéma A, on a observé la mort de 6 malades,
- Sur 60 malades traités avec le schéma B, on a observé la mort de 15 malades.

- 1. Donner une estimation ponctuelle des proportions dans les deux populations.
- 2. Pour la population 1, donner une estimation par intervalle de confiance de la proportion au risque 1%.
- 3. Tester au risque 1% pour la population 2 si la proportion vaut 20%.
- 4. Comparer au risque 5% les proportions des deux populations. Peut-on dire que les schémas A et B diffèrent significativement au risque 5% ?

Exercices de révisions

Exercice 6 (Comparaisons des moyennes) :

Sur deux groupes de même taille : 9 malades, on expérimente les effets d'un nouveau médicament. On observe les résultats suivants :

Groupe 1	15	18	17	20	21	18	17	15	19
Groupe 2	12	16	17	18	17	15	18	14	16

- 1. Donner une estimation ponctuelle des moyennes et variances dans les deux populations.
- 2. Pour la population 1, donner une estimation par intervalle de confiance de la moyenne au risque 5%.
- 3. Pour la population 1, donner une estimation par intervalle de confiance de la variance au risque 5%.

Exercices de révisions

- 4. Tester au risque 5% pour la population 2 si la moyenne vaut 16.
- 5. Tester au risque 5% pour la population 2 si la variance vaut 4.
- 6. Comparer au risque 5% les variances des deux populations.
- 7. Comparer au risque 5% les moyennes des deux populations.

Exercices de révisions

Exercice 7 (ANOVA et test de Shapiro et Wilk) :

Afin de tester l'hypothèse que la consommation de caféine facilite l'apprentissage, trois groupes d'étudiants se préparent à un examen : le groupe 1 boit une tasse, le groupe 2 boit 2 tasses et le groupe 3 boit 3 tasses de café. Voici leurs scores à l'examen :

Groupe 1	Groupe 2	Groupe 3
50	48	57
42	47	59
53	65	48

- 1. Utiliser le test de Shapiro et Wilk pour tester la normalité des neuf données.
- 2. Construire le tableau d'ANOVA et conclure au risque 5%.

Exercices de révisions

Exercice 8 (Test d'indépendance de Khi-deux) :

Une étude a été réalisée sur le cancer de la gorge. Pour cela, une population de 1000 personnes a été interrogée. les résultats obtenus sont donnés dans le tableau de contingences suivant :

	Atteint du cancer de la gorge	Non atteint du cancer de la gorge
Fumeur	344	258
Non-fumeur	160	238

- 1. Doit-on rejeter au risque 5% l'hypothèse d'indépendance des deux caractères : $X=(\text{être fumeur})$ et $Y=(\text{être atteint du cancer de la gorge})$
- 2. Vérifier la validité du test.

Plan :

- 1 Introduction
- 2 Statistique descriptive univariée
- 3 Statistique descriptive bivariée
- 4 Notions de probabilités et variables aléatoires
- 5 Échantillonnage et estimation
- 6 Tests des hypothèses
- 7 Exercices de révisions
- 8 Examens (sessions normale et de rattrapage 2016-2017)

Examen (session normale 2016-2017)

Exercice 1 (2 pts=1+1) :

Soit $X \sim \mathcal{N}(\mu = 170, \sigma^2 = 64)$.

- 1. Calculer $P(160 < X < 176)$.
- 2. Déterminer la valeur de z tel que $P(170 - z < X < 170 + z) = 0,668$.

Examen (session normale 2016-2017)

Loi Normale N(0,1) : P(Z<z)

Z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,500 0	0,504 0	0,508 0	0,512 0	0,516 0	0,519 9	0,523 9	0,527 9	0,531 9	0,535 9
0,1	0,539 8	0,543 8	0,547 8	0,551 7	0,555 7	0,559 6	0,563 6	0,567 5	0,571 4	0,575 3
0,2	0,579 3	0,583 2	0,587 1	0,591 0	0,594 8	0,598 7	0,602 6	0,606 4	0,610 3	0,614 1
0,3	0,617 9	0,621 7	0,625 5	0,629 3	0,633 1	0,636 8	0,640 6	0,644 3	0,648 0	0,651 7
0,4	0,655 4	0,659 1	0,662 8	0,666 4	0,670 0	0,673 6	0,677 2	0,680 8	0,684 4	0,687 9
0,5	0,691 5	0,695 0	0,698 5	0,701 9	0,705 4	0,708 8	0,712 3	0,715 7	0,719 0	0,722 4
0,6	0,725 7	0,729 1	0,732 4	0,735 7	0,738 9	0,742 2	0,745 4	0,748 6	0,751 7	0,754 9
0,7	0,758 0	0,761 1	0,764 2	0,767 3	0,770 4	0,773 4	0,776 4	0,779 4	0,782 3	0,785 2
0,8	0,788 1	0,791 0	0,793 9	0,796 7	0,799 5	0,802 3	0,805 1	0,807 8	0,810 6	0,813 3
0,9	0,815 9	0,818 6	0,821 2	0,823 8	0,826 4	0,828 9	0,831 5	0,834 0	0,836 5	0,838 9
1,0	0,841 3	0,843 8	0,846 1	0,848 5	0,850 8	0,853 1	0,855 4	0,857 7	0,859 9	0,862 1
1,1	0,864 3	0,866 5	0,868 6	0,870 8	0,872 9	0,874 9	0,877 0	0,879 0	0,881 0	0,883 0
1,2	0,884 9	0,886 9	0,888 8	0,890 7	0,892 5	0,894 4	0,896 2	0,898 0	0,899 7	0,901 5
1,3	0,903 2	0,904 9	0,906 6	0,908 2	0,909 9	0,911 5	0,913 1	0,914 7	0,916 2	0,917 7
1,4	0,919 2	0,920 7	0,922 2	0,923 6	0,925 1	0,926 5	0,927 9	0,929 2	0,930 6	0,931 9
1,5	0,933 2	0,934 5	0,935 7	0,937 0	0,938 2	0,939 4	0,940 6	0,941 8	0,942 9	0,944 1
1,6	0,945 2	0,946 3	0,947 4	0,948 4	0,949 5	0,950 5	0,951 5	0,952 5	0,953 5	0,954 5
1,7	0,955 4	0,956 4	0,957 3	0,958 2	0,959 1	0,959 9	0,960 8	0,961 6	0,962 5	0,963 3
1,8	0,964 1	0,964 9	0,965 6	0,966 4	0,967 1	0,967 8	0,968 6	0,969 3	0,969 9	0,970 6
1,9	0,971 3	0,971 9	0,972 6	0,973 2	0,973 8	0,974 4	0,975 0	0,975 6	0,976 1	0,976 7



Examen (session normale 2016-2017)

Exercice 2 (9 pts=2+2+1+1+1+2) :

On considère la série double statistique suivante :

y_i	5	1	3	2	4
x_i	11	3	9	4	8

- 1. Déterminer les médianes, les moyennes et les variances de X et Y.
- 2. Déterminer la covariance et le coefficient de corrélation entre X et Y. Donner une interprétation.
- 3. Déterminer la droite de régression linéaire de Y en X.
- 4. Tracer le nuage de points et la droite de régression linéaire de Y en X.
- 5. Établir, sur la base de ce modèle, la valeur y^* correspond à la valeur $x^* = 4,5$.
- 6. Déterminer la variance résiduelle et le coefficient de détermination. Donner une interprétation.

Examen (session normale 2016-2017)

Exercice 3 (5 pts=1+1+2+1) :

Dans une université, on veut comparer les proportions de réussite dans la filière STU S3 sur deux années différentes :

- Pour l'année 2014, on note 80 succès pour 120 interrogés,
 - Pour l'année 2015, on note 70 succès pour 110 interrogés.
- 1. Pour l'année 2014, estimer par intervalle de confiance la proportion au risque 5%.
 - 2. Pour l'année 2015, tester au risque 5% si la proportion vaut 60%.
 - 3. Comparer au risque 5% les proportions des deux années. Peut-on dire que les deux années diffèrent significativement au risque 5% ?
 - 4. Vérifier les conditions de validité de ce dernier test.

Examen (session normale 2016-2017)

Exercice 4 (4 pts=2+2) :

Le tableau suivant présente la production laitière en litres par jours de trois races de vaches :

Race (1)	Race (2)	Race (3)
147	153	142
150	148	165
148	159	1 57

- 1. Tester au risque 5% la normalité des données.
- 2. Construire le tableau d'ANOVA et conclure au risque 5%.

Examen (session normale 2016-2017)

Loi Fisher Snédécor à (v_1, v_2) ddl

$$P(F \geq f_a) = \alpha = 0,05.$$

$v_1 \backslash v_2$	1	2	3	4	5	6	8	10	15	20	30	∞
1	161	200	216	225	230	234	239	242	246	248	250	254
2	18,5	19,0	19,2	19,2	19,3	19,3	19,4	19,4	19,4	19,4	19,5	19,5
3	10,1	9,55	9,28	9,12	9,01	8,94	8,85	8,79	8,70	8,66	8,62	8,53
4	7,71	6,94	6,59	6,39	6,26	6,16	6,04	5,96	5,86	5,80	5,75	5,63
5	6,61	5,79	5,41	5,19	5,05	4,95	4,82	4,74	4,62	4,56	4,50	4,36
6	5,99	5,14	4,76	4,53	4,39	4,28	4,15	4,06	3,94	3,87	3,81	3,67
7	5,59	4,74	4,35	4,12	3,97	3,87	3,73	3,64	3,51	3,44	3,38	3,23
8	5,32	4,46	4,07	3,84	3,69	3,58	3,44	3,35	3,22	3,15	3,08	2,93
9	5,12	4,26	3,86	3,63	3,48	3,37	3,23	3,14	3,01	2,94	2,86	2,71
10	4,96	4,10	3,71	3,48	3,33	3,22	3,07	2,98	2,85	2,77	2,70	2,54

Examen (session de rattrapage 2016-2017)

Exercice 1 (2 pts=1+1) :

Soit X une variable aléatoire de densité de probabilité : $f(x) = \frac{1_{(x>1)}}{x^2}$.

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

2. Calculer l'espérance $E(X)$.

Examen (session de rattrapage 2016-2017)

Exercice 2 (9 pts=1+2+2+1+1+1+1) :

On considère la série double statistique suivante :

x_i	16	8	7	5	11	2	14	13	20	24
y_i	17	12	7	2	8	3	13	10	8	20

- ① Déterminer les médianes de X et Y . Donner une interprétation.
- ② Déterminer les moyennes et les variances de X et Y .
- ③ Déterminer la covariance et le coefficient de corrélation entre X et Y .
Donner une interprétation.
- ④ Déterminer la droite de régression linéaire de Y en X .
- ⑤ Tracer le nuage de points et la droite de régression linéaire de Y en X .
- ⑥ Etablir, sur la base de ce modèle, y^* correspond à la valeur $x^* = 10$.
- ⑦ Déterminer la variance résiduelle et le coefficient de détermination.
Donner une interprétation.

Examen (session de rattrapage 2016-2017)

Exercice 3 (9 pts=1+1+1+1+1+2+2) :

Sur deux groupes de deux types de produits, on test un nouveau processus.
On observe les résultats suivants :

Produit A	17	18	17	15	14	16	2	12	16	18
Produit B	17	15	19	1	15	18	21	18	17	20

- ① Donner une estimation ponctuelle des moyennes et variances dans les deux groupes.
- ② Pour le groupe B, estimer par intervalle de confiance la moyenne au risque 5 %.
- ③ Pour le groupe A, estimer par intervalle de confiance la variance au risque 5 %.

Examen (session de rattrapage 2016-2017)

- **4- Pour le groupe B, tester au risque 5 % si la moyenne vaut 16.**
- **5- Pour le groupe B, tester au risque 5 % si la variance vaut 20.**
- **6- Comparer au risque 5 % les variances des deux groupes.**
- **7- Pour le groupe A, tester au risque 5 % la normalité des données.**

Examen (session de rattrapage 2016-2017)

Loi Student : $P(T>t) = \alpha$

Degrés de liberté	alpha : Aire dans la queue supérieure de la distribution					
	0,20	0,10	0,05	0,025	0,01	0,005
1	1,376	3,078	6,314	12,706	31,821	63,656
2	1,061	1,886	2,920	4,303	6,965	9,925
3	0,978	1,638	2,353	3,182	4,541	5,841
4	0,941	1,533	2,132	2,776	3,747	4,604
5	0,920	1,476	2,015	2,571	3,365	4,032
6	0,906	1,440	1,943	2,447	3,143	3,707
7	0,896	1,415	1,895	2,365	2,998	3,499
8	0,889	1,397	1,860	2,306	2,896	3,355
9	0,883	1,383	1,833	2,262	2,821	3,250

Examen (session de rattrapage 2016-2017)

Loi Khi-deux à v ddl : $P(Y^2 > x)$

$v \setminus \alpha$	0,990	0,975	0,950	0,900	0,100	0,050	0,025	0,010
1	0,0002	0,0010	0,0039	0,0158	2,7055	3,8415	5,0239	6,6349
2	0,0201	0,0506	0,1026	0,2107	4,6052	5,9915	7,3778	9,2103
3	0,1148	0,2158	0,3518	0,5844	6,2514	7,8147	9,3484	11,3449
4	0,2971	0,4844	0,7107	1,0636	7,7794	9,4877	11,1433	13,2767
5	0,5543	0,8312	1,1455	1,6103	9,2364	11,0705	12,8325	15,0863
6	0,8721	1,2373	1,6354	2,2041	10,6446	12,5916	14,4494	16,8119
7	1,2390	1,6899	2,1673	2,8331	12,0170	14,0671	16,0128	18,4753
8	1,6465	2,1797	2,7326	3,4895	13,3616	15,5073	17,5345	20,0902
9	2,0879	2,7004	3,3251	4,1682	14,6837	16,9190	19,0228	21,6660
10	2,5582	3,2470	3,9403	4,8652	15,9872	18,3070	20,4832	23,2093

Examen (session de rattrapage 2016-2017)

Loi Fisher Snédécor à (v_1, v_2) ddl

$$P(F \geq f_{\alpha}) = \alpha = 0,025$$

$v_1 \backslash v_2$	1	2	3	4	5	6	8	10	15	20	30	∞
1	648	800	864	900	922	937	957	969	985	993	1 001	1 018
2	38,5	39,0	39,2	39,2	39,3	39,3	39,4	39,4	39,4	39,4	39,5	39,5
3	17,4	16,0	15,4	15,1	14,9	14,7	14,5	14,4	14,3	14,2	14,1	13,9
4	12,2	10,6	9,98	9,60	9,36	9,20	8,98	8,84	8,66	8,56	8,46	8,26
5	10,0	8,43	7,76	7,39	7,15	6,98	6,76	6,62	6,43	6,33	6,23	6,02
6	8,81	7,26	6,60	6,23	5,99	5,82	5,60	5,46	5,27	5,17	5,07	4,85
7	8,07	6,54	5,89	5,52	5,29	5,12	4,90	4,76	4,57	4,47	4,36	4,14
8	7,57	6,06	5,42	5,05	4,82	4,65	4,43	4,30	4,10	4,00	3,89	3,67
9	7,21	5,71	5,08	4,72	4,48	4,32	4,10	3,96	3,77	3,67	3,56	3,33
10	6,94	5,46	4,83	4,47	4,24	4,07	3,85	3,72	3,52	3,42	3,31	3,08

Examen (session de rattrapage 2016-2017)

Table de Shapiro et Wilk

n	Risque 5 %	Risque 1 %
	$W_{0,05}$	$W_{0,01}$
5	0,762	0,686
6	0,788	0,713
7	0,803	0,730
8	0,818	0,749
9	0,829	0,764
10	0,842	0,781
11	0,850	0,792
12	0,859	0,805

Tables des valeurs des a_j

n	2	3	4	5	6	7	8	9	10
j									
1	0,7011	0,7071	0,6872	0,6646	0,6431	0,6233	0,6052	0,5888	0,5739
2		0,0000	0,1677	0,2413	0,2806	0,3031	0,3164	0,3244	0,3391
3			0,0000	0,0875	0,1401	0,1743	0,1976	0,2141	
4				0,0000	0,0561	0,0947	0,1224		
5					0,0000	0,0399			

Bibliographie

-  Emile Amzallag et Norbert Piccioli. *Introduction à la statistique*.
-  Renée Veysseyre. *Aide-mémoire : Statistique et probabilités pour l'ingénieur*.
-  Yves Tillé. *Résumé du cours de statistique descriptive*.