mini Manuel de probabilités et statistique

Cours + QCM

Françoise Couty-Fredon
Professeure certifiée hors classe

Jean Debord
Praticien hospitalier attaché au CHU de Limoges
Chargé de cours à la faculté des sciences de Limoges

Daniel Fredon
Maître de conférences de mathématiques appliquées

3e édition

DUNOD

Directeur d'ouvrage
Daniel FREDON

11, rue Paul Bert 92240 Malakoff
www.dunod.com

ISBN 978-2-10-085110-2

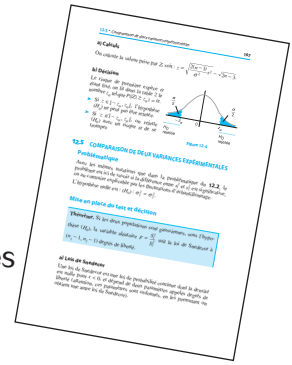
Comment utiliser ce Mini-Manuel ?

La page d'entrée de chapitre


CHAPITRE 6
Variables aléatoires discrètes (cas infini)

Elle donne le plan du cours, ainsi qu'un rappel des objectifs pédagogiques du chapitre.

Le cours



Le cours, concis et structuré, expose les notions importantes du programme.

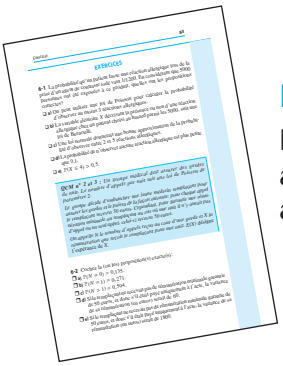
Les rubriques

Une erreur à éviter

Un peu de méthode

Un exemple pour comprendre

Les points clés à retenir



Les exercices, QCM ou QROC

Ils sont proposés en fin de chapitre, avec leur solutions, pour se tester tout au long de l'année.

Table des matières

Avant-propos 1

Statistique descriptive

1 Statistique à une dimension 3

2 Statistique à deux dimensions 17

Probabilités

3 Probabilités (généralités) 33

4 Probabilité conditionnelle 45

5 Variables aléatoires discrètes (cas fini) 59

6 Variables aléatoires discrètes (cas infini) 77

7 Variables aléatoires continues 91

Statistique inférentielle

8 Échantillonnage – Estimation d'un paramètre 107

9 Introduction aux tests statistiques 123

10 Test du khi-deux (χ^2) 129

11 Comparaison de deux proportions 145

12 Comparaison de deux moyennes, de deux variances 159

13 Analyse de la variance 181

14 Régression linéaire **197**

15 Corrélation **211**

16 Tests non paramétriques **225**

Tables

1 – Fonction de répartition de la loi normale réduite **241**
2 – Loi normale réduite (table de l'écart réduit) **242**
3 – Lois de Student **243**
4 – Lois de Pearson ou lois du χ^2 **244**
5 – Lois de Snedecor ($\alpha = 0{,}025$) **245**
6 – Lois de Snedecor ($\alpha = 0{,}05$) **246**
7 – Test de Mann et Whitney ($\alpha = 0{,}05$) **247**
8 – Test de Mann et Whitney ($\alpha = 0{,}01$) **248**
9 – Test de Wilcoxon **248**
10 – Table du coefficient de corrélation linéaire **249**
11 – Coefficient de corrélation de rang de Spearman **250**
12 – Test de Kruskal et Wallis **250**

Glossaire **251**

Index **255**

Avant-propos

Ce livre est destiné à tous les étudiants en sciences de la vie, de la Terre et de la santé : licences, pharmacie, médecine, IUT et BTS à dominante biologique ou agricole. Mais il concerne aussi tous les utilisateurs de statistiques en laboratoire.

Pour satisfaire l'attente d'un tel public, nous avons choisi d'aborder une large étendue de sujets. Comme ce livre est découpé en chapitres autonomes, chacun pourra, d'une année à l'autre, retrouver les sujets qui le concernent. Un index détaillé situé en fin d'ouvrage aidera dans ce choix.

En statistique, les notations ne sont pas toutes universelles, ce qui complique la consultation d'ouvrages variés. Pour notre part, nous nous sommes efforcés de respecter les règles de cohérence suivantes :

– utiliser les lettres grecques pour des valeurs relatives à la population et des lettres latines pour des valeurs relatives à un échantillon ;
– bien distinguer une variable aléatoire (notée par une majuscule) et une valeur numérique prise par cette variable aléatoire (notée par la même lettre, mais en minuscule).

Pour des révisions express à l'approche d'un examen ou d'un concours, nous vous conseillons, chez le même éditeur :
D. Fredon ; Statistique et probabilités en 30 fiches ; collection Express Sciences.
Bien sûr, la charité bien ordonnée commence par soi-même. Mais il y a plus important : vous y trouverez des notations en cohérence avec ce livre, et des exercices différents pour compléter votre entraînement.

Toutes vos remarques, vos commentaires, vos critiques, et même vos encouragements, seront accueillis avec plaisir. Vous pouvez me les communiquer à l'adresse électronique suivante : daniel.fredon@laposte.net

Daniel Fredon

CHAPITRE 1

Statistique à une dimension

PLAN

1.1 Généralités
1.2 Représentations graphiques
1.3 Paramètres de position
1.4 Paramètres de dispersion
1.5 Paramètres de forme

OBJECTIFS

- Savoir représenter graphiquement une série statistique après avoir choisi l'aspect à mettre en évidence
- Résumer certains aspects (position, dispersion, forme) d'une série statistique en calculant un nombre adapté

1.1 GÉNÉRALITÉS

Vocabulaire général

La statistique étudie des ensembles appelés **populations**, dont les éléments sont appelés **individus**. Dans le cas d'une série statistique à une variable, à chaque individu on associe une éventualité d'un **caractère statistique**.

Si les éventualités ne sont pas des nombres, le caractère est dit **qualitatif** et les éventualités s'appellent les modalités du caractère.

Si les éventualités sont des nombres, le caractère est dit **quantitatif** et les éventualités sont les valeurs du caractère.

Un caractère quantitatif est dit **continu** s'il peut prendre toutes les valeurs d'un intervalle. Il est discontinu, ou **discret**, s'il ne peut prendre que des valeurs isolées.

Série statistique

Dans le cas d'un caractère qualitatif ou quantitatif discret, on dispose d'une série statistique quand on connaît pour chaque individu la modalité, ou la valeur, prise par le caractère. L'**effectif** d'une modalité, ou d'une valeur, est le nombre de fois où elle apparaît dans la population.

Quand le caractère quantitatif est continu, ou discret avec beaucoup de valeurs, on considère des intervalles, en général du type $]a, b]$, que l'on appelle des **classes statistiques**. La longueur $b - a$ de l'intervalle est l'**amplitude** de la classe. Sa **densité** est le quotient de l'effectif par l'amplitude.

Effectifs, fréquences

Pour une valeur (ou une modalité) d'un caractère, ou pour une classe statistique, la **fréquence** est le quotient de l'effectif concerné n_i par l'effectif total n, soit : $f_i = \dfrac{n_i}{n}$.

La somme des fréquences est donc égale à 1.

Si on veut obtenir la répartition en pourcentages, il suffit de multiplier les fréquences par 100.

Effectifs cumulés, fréquences cumulées

Lorsque le caractère est quantitatif, on range les valeurs (ou les classes) par ordre croissant.

- L'**effectif cumulé** jusqu'à k est la somme des effectifs associés aux valeurs du caractère qui sont inférieures ou égales à k.
- La **fréquence cumulée** jusqu'à k s'obtient en additionnant les fréquences associées aux valeurs $\leq k$, ou en divisant l'effectif cumulé par l'effectif total.

1.2 REPRÉSENTATIONS GRAPHIQUES

Cas d'un caractère qualitatif ou quantitatif discret

- Si on veut insister sur la comparaison des effectifs, on trace un diagramme à bandes, ou un diagramme en bâtons. Les longueurs doivent être proportionnelles aux effectifs.
- Si on préfère mettre en évidence les pourcentages pour comparer visuellement les structures de plusieurs séries statistiques (c'est-à-dire les répartitions en pourcentages), on représente les données à l'aide :

– de graphiques circulaires (parfois appelés camemberts), où les angles au centre du disque, ou du demi-disque, sont proportionnels aux pourcentages ;

– ou de bandes subdivisées de longueur fixe.

Cas d'un caractère quantitatif continu

On peut utiliser les représentations précédentes. Mais on construit le plus souvent un **histogramme** :

Les intervalles des classes statistiques sont reportés sur un axe. Il servent de bases à des rectangles dont les aires sont proportionnelles aux effectifs. Pour ceci, les côtés des rectangles perpendiculaires à l'axe sont proportionnels aux densités des classes.

1.3 PARAMÈTRES DE POSITION

Moyenne

a) Définition

Notons $x_1,\ldots,x_p$ les valeurs du caractère, $n_1,\ldots,n_p$ les effectifs correspondants et $n = n_1 + \cdots + n_p$ l'effectif total. La moyenne de la série statistique est le nombre :

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{p} n_i x_i = \sum_{i=1}^{p} f_i x_i.$$

Quand les informations sont fournies avec des classes statistiques, on utilise la même formule en retenant comme valeurs x_i les milieux des classes.

La définition précédente est celle de la moyenne arithmétique. On définit aussi d'autres moyennes pour $x_i > 0$ comme :

– la moyenne harmonique h telle que : $\frac{n}{h} = \sum_{i=1}^{p} \frac{n_i}{x_i}$

– la moyenne géométrique g telle que : $g^n = x_1^{n_1} \times \cdots \times x_p^{n_p}$.

b) Propriétés

La moyenne ne change pas si on remplace les effectifs par des effectifs proportionnels.

La moyenne ne change pas si on remplace k valeurs $x_1,\ldots,x_k$ affectées de coefficients $n_1,\ldots,n_k$ par leur moyenne partielle affectée de la somme des coefficients $n_1 + \cdots + n_k$.

Par exemple, si la population est subdivisée en trois sous-populations dont les moyennes partielles sont $\overline{x}_1, \overline{x}_2, \overline{x}_3$ et les effectifs N_1, N_2, N_3, alors la moyenne de la population totale est :

$$\overline{x} = \frac{N_1\overline{x}_1 + N_2\overline{x}_2 + N_3\overline{x}_3}{N_1 + N_2 + N_3}.$$

Médiane

a) Cas d'un caractère quantitatif discret

On ordonne les n valeurs de la série statistique par ordre croissant.

Si n est impair, la médiane est la valeur de rang $\frac{n+1}{2}$.

Si n est pair, les valeurs de rangs $\frac{n}{2}$ et $\frac{n}{2}+1$ déterminent un intervalle médian. On retient souvent comme médiane le milieu de cet intervalle.

b) Cas d'un caractère quantitatif continu

Dans ce cas, la médiane est le nombre m tel que la fréquence cumulée jusqu'à m soit égale à 0,5.

Mode

On appelle **mode**, ou dominante, d'une série statistique toute valeur (ou modalité) correspondant à l'effectif maximal (densité maximale dans le cas de classes statistiques).

1.4 PARAMÈTRES DE DISPERSION

Variance, écart type

Avec les mêmes notations que précédemment, on appelle variance de la série statistique le nombre V :

$$V = \frac{1}{n}\sum_{i=1}^{p} n_i(x_i - \overline{x})^2 = \sum_{i=1}^{p} f_i(x_i - \overline{x})^2.$$

On appelle écart type de la série statistique le nombre $\sigma = \sqrt{V}$.

La variance peut aussi se calculer par :

$$V = \frac{1}{n}\left(\sum_{i=1}^{p} n_i x_i^2\right) - (\overline{x})^2.$$

Les calculatrices et les tableurs fournissent directement, sous des notations diverses, l'écart type σ et la variance σ^2.

Mais ils fournissent aussi un autre nombre *s*, sous des notations variées, qui est tel que :

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{p} n_i (x_i - \overline{x})^2 = \frac{n}{n-1}\sigma^2.$$

s^2 est destiné à estimer la variance d'une population quand on ne dispose que d'un échantillon de taille *n*.

On l'appelle la variance estimée et elle ne doit pas être confondue avec la variance *V* de l'échantillon (*cf.* chap. 8).

Coefficient de variation

Le **coefficient de variation** d'une série statistique est le quotient $\frac{\sigma}{\overline{x}}$.

C'est un nombre sans dimension qui permet de comparer la dispersion de séries statistiques dont les moyennes sont très différentes.

Autres paramètres de dispersion

- L'**étendue** d'une série statistique associée à un caractère quantitatif est la différence entre la plus grande valeur observée et la plus petite.
- En partageant la série ordonnée des résultats en quatre parties de même effectif, on obtient les quartiles Q_1, Q_2, Q_3. Le deuxième quartile Q_2 est la médiane. L'**écart interquartile** est le nombre $Q_3 - Q_1$.

Boîte de dispersion (ou boîte à moustaches)

C'est une représentation graphique d'un caractère quantitatif. Elle sert à comparer visuellement plusieurs séries statistiques.

Pour une série donnée, on trace un rectangle qui s'étend de Q_1 à Q_3 et on marque la médiane par un trait. On ajoute les moustaches qui sont les segments qui vont de la valeur minimale à Q_1, et de Q_3 à la valeur maximale.

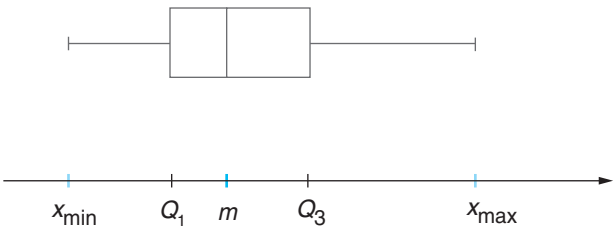


Figure 1.1

1.5 PARAMÈTRES DE FORME

Moments

Pour $r \in \mathbb{N}$, on définit le moment d'ordre r : $m_r = \frac{1}{n}\sum_{i=1}^{p} n_i x_i^r$,

le moment centré d'ordre r : $\mu_r = \frac{1}{n}\sum_{i=1}^{p} n_i (x_i - \overline{x})^r$.

Coefficient γ_1 de Fisher (dissymétrie) : $\gamma_1 = \frac{\mu_3}{\sigma^3}$

Si $\gamma_1 = 0$, la distribution est symétrique.

Si $\gamma_1 < 0$, la distribution est étalée vers la gauche.

Si $\gamma_1 > 0$, la distribution est étalée vers la droite.

Coefficient γ_2 de Fisher (aplatissement) : $\gamma_2 = \frac{\mu_4}{\sigma^4} - 3$

Si $\gamma_1 = 0$, l'applatissement est le même que celui de la loi de Gauss réduite (*cf.* chap. 7).

Si $\gamma_1 < 0$, la distribution est plus aplatie.

Si $\gamma_1 > 0$, la distribution est moins aplatie.

Effets d'un regroupement en classes

Lorsque la série statistique comporte un grand nombre de valeurs, les calculs sont simplifiés en effectuant d'abord un regroupement en classes, puis en remplaçant chaque classe par son milieu. Mais les résultats en sont légèrement modifiés.

Si la distribution des valeurs est uniforme dans chaque classe, la moyenne n'est pas changée (associativité de la moyenne).

Mais la variance, qui mesure la dispersion, est modifiée puisqu'on concentre toutes les valeurs d'une classe en un seul point. Dans le cas où toutes les classes sont de même amplitude d, on peut améliorer le résultat avec la correction de Sheppard, qui consiste à retrancher $\frac{d^2}{12}$ à la valeur de la variance obtenue à partir des valeurs groupées.

MOTS-CLÉS

- Caractère statistique
- Représentation graphique
- Paramètres d'une série statistique

EXERCICES

1-1 Gwenaël, qui a 16 ans, soufre d'une fracture complexe du tibia et du péroné après avoir chuté à ski. Si Y est le type de fracture, quel est le type de la variable décrivant Y ?

❑ **a)** Variable discrète finie.

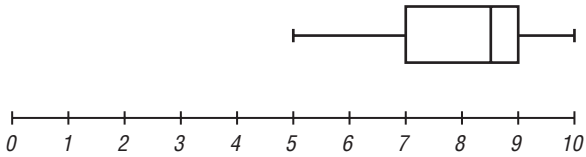
❑ **b)** Variable ordinale.

❑ **c)** Variable qualitative nominale.

❑ **d)** Variable entière finie.

❑ **e)** Variable quantitative non dénombrable.

1-2 Concernant le diagramme en box-plot suivant :



❑ **a)** La moyenne est égale à 8, 5.

❑ **b)** La médiane est égale à 8, 5.

❑ **c)** L'écart interquartile est égal à 2.

❑ **d)** 50 % de l'effectif est contenu dans l'intervalle [7 ; 9].

❑ **e)** Il y a une valeur extrême.

1-3 Une étude sur le sommeil (en heures par jour) des enfants de 4 ans a donné les résultats suivants sur les 591 enfants de l'étude :
minimum : 7; premier quartile : 9 ; médiane : 10 ; troisième quartile : 11 ; maximum : 14.

❑ **a)** La variable étudiée était qualitative ordinale.

❑ **b)** 295 enfants dormaient au moins 10 heures.

❑ **c)** Le deuxième quartile était égal à 8.

❑ **d)** L'étendue était de 3.

❑ **e)** 148 enfants dormaient au moins 11 heures.

1-4 On considère une série statistique de 60 taux d'hémoglobine dans le sang (g/L) mesurés chez des adultes présumés en bonne santé. La série est rangée par valeurs non décroissantes. Les valeurs en gras indique que le taux d'hémoglobine a été mesuré sur une femme.

105 ; 110 ; 112 ; 112 ; 118 ; 119 ; 120 ; 120 ; 125 ; 126 ; 127 ; 128 ; 130 ; 132 ; 133 ; 134 ; 135 ; 138 ; 138 ; 138 ; 138 ; 141 ; **142** ; 144 ; **145** ; 146 ; **148** ; **148** ; 148 ; 149 ; **150** ; 150 ; 150 ; **151** ; 151 ; 153 ; 153 ; 153 ; **154** ; **154** ; 154 ; 155 ; 156 ; 156 ; **158** ; 160 ; 160 ; 160 ; 163 ; 164 ; 164 ; 165 ; 166 ; 168 ; 168 ; 170 ; 172 ; 172 ; 176 ; 179.

Résultats partiels

$$\text{Hommes : } \sum_{i=1}^{30} x_{ih} = 4\,766 \text{ g/L} \quad \sum_{i=1}^{30} x_{ih}^2 = 759\,954 \quad (\text{g/L})^2$$

$$\text{Femmes : } \sum_{i=1}^{30} x_{if} = 3\,988 \text{ g/L} \quad \sum_{i=1}^{30} x_{if}^2 = 536\,176 \quad (\text{g/L})^2$$

a) On considère le groupement en classes :
]104;114] ; [114;124] ; [124;134] ;]134;144] ;]144;154] ;]154;164] ;]164;174] ;]174;184]

Pour chacune des deux séries : hommes, femmes, déterminez les effectifs et les fréquences de chaque classe.

b) Effectuez une représentation graphique adaptée des deux distributions groupées en classes de la question précédente.

c) Calculez les moyennes $\overline{x}$, $\overline{x}_f$, $\overline{x}_h$, des trois distributions initiales : ensemble, femmes, hommes ;

d) Calculez les moyennes $\overline{x}'$, $\overline{x}'_f$, $\overline{x}'_h$, des trois distributions (ensemble, femmes, hommes) après le regroupement en classes de la question **a)**, en remplaçant chaque classe par son milieu.

e) Calculez les médianes m, m_f, m_h des trois distributions initiales : ensemble, femmes, hommes.

f) Calculez l'écart interquartile pour chacune des trois distributions initiales : ensemble, femmes, hommes.

g) Calculez les variances et les écarts type des trois distributions initiales : femmes, hommes, ensemble.

h) Calculez les variances et les écarts type des trois distributions après le regroupement en classes de la question **a)**, en remplaçant chaque classe par son milieu.

i) Pour la distribution des femmes, calculez les moments m_1, m_2, m_3, m_4. Déduisez-en les valeurs des moments centrés $\mu_1, \mu_2, \mu_3, \mu_4$, puis des coefficients de forme γ_1 et γ_2 de Fisher.

1-5 Dans l'étude de la répartition de la végétation en fonction de divers facteurs écologiques, on utilise une carte au 1/200 000 sur laquelle sont représentées les séries de végétation. On superpose une grille dont la maille est de 1 cm. Des renseignements annexes fournissent, pour chaque point de la grille, la température moyenne T en °C, la pluviosité annuelle moyenne P en mm, et la nature du sol.

En étudiant la région de Limoges, on a ainsi obtenu pour la population constituée par les points étudiés :

• Pour le *chêne pédonculé*

P	]700 ; 800]	]800 ; 900]	]900 ; 1000]	]1000 ; 1100]	]1100 ; 1200]
effectifs	10	85	185	122	138

P	]1200 ; 1300]	]1300 ; 1400]	]1400 ; 1500]	]1500 ; 1600]	]1600 ; 1700]
effectifs	43	15	12	13	10

P	]1700 ; 1800]	]1800 ; 1900]	]1900 ; 2000]
effectifs	6	5	1

T	]7 ; 8]	]8 ; 9]	]9 ; 10]	]10 ; 11]	]11 ; 12]	]12 ; 13]
effectifs	4	25	109	250	205	52

sols	acides	calcaires	montagneux
effectifs	502	49	94

• Pour le *chêne pubescent*

P	]700 ; 800]	]800 ; 900]	]900 ; 1000]	]1000 ; 1100]
effectifs	14	103	37	3

T	]11 ; 12]	]12 ; 13]
effectifs	34	123

sols	acides	calcaires
effectifs	23	134

a) En assimilant chaque classe à son milieu, calculez la pluviosité moyenne $\overline{P}_1$ pour les zones où vit le chêne pédonculé, puis $\overline{P}_2$ pour le chêne pubescent. Calculez les écarts type correspondant à ces deux séries statistiques et les coefficients de variation.

b) Calculez de même les températures moyennes $\overline{T}_1$ et $\overline{T}_2$ et les écarts type correspondants.

c) Construisez deux graphiques pour visualiser la comparaison de la nature des sols habités par le chêne pédonculé et le chêne pubescent.

d) Conclusions écologiques ?

SOLUTIONS

1-1 ❑ **a)** ❑ **b)** ☒ **c)** ❑ **d)** ❑ **e)**

Les modalités sont les types de fracture, donc non numériques. La variable est donc qualitative. Il n'y a pas d'ordre entre les types de fracture; la variable n'est donc pas ordinale.

1-2 ❑ **a)** ☒ **b)** ☒ **c)** ☒ **d)** ☒ **e)**

a) faux et **b. vrai :** Le trait dans la boîte correspond à la médiane.

c) vrai : L'écart interquartile est $9 - 7 = 2$.

d) vrai : L'intervalle interquartile contient 50 % de la population.

e) vrai : Les valeurs extrêmes sont 5 et 10.

1-3 ❑ **a)** ☒ **b)** ❑ **c)** ❑ **d)** ☒ **e)**

a) faux : la variable est quantitative.

b) vrai : la médiane est 10, il y a donc la moitié des 591 enfants, soit 295, qui dorment au moins 10 heures.

c) faux: le deuxième quartile est la médiane, soit 10.

d) faux : l'étendue est la différence entre la valeur maximale et la valeur minimale, soit $14 - 7 = 7$.

e) vrai : le troisième quartile est 11, il y a donc un quart des 591 enfants, soit 148, qui dorment au moins 11 heures.

1-4 a)

	femmes		hommes	
classes	**effectifs**	**fréquences**	**effectifs**	**fréquences**
]104 ; 114]	4	0,133	0	0
]114 ; 124]	4	0,133	0	0
]124 ; 134]	8	0,267	0	0
]134 ; 144]	6	0,200	2	0,067
]144 ; 154]	7	0,233	10	0,333
]154 ; 164]	1	0,033	9	0,300
]164 ; 174]	0	0	7	0,233
]174 ; 184]	0	0	2	0,067
total	30	1	30	1

b) On peut dessiner deux histogrammes en portant indifféremment en ordonnées les effectifs, les fréquences ou les densités des classes, car les classes sont de même amplitude.

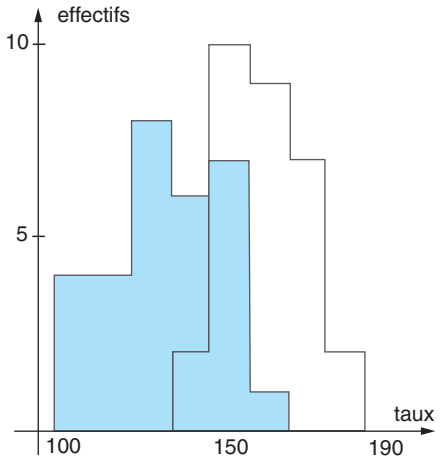


Figure 1-2

c) Pour la série totale : $\overline{x} = \dfrac{4\,766 + 3\,988}{60} = 145{,}9$

Pour la série des femmes : $\overline{x}_f = \dfrac{3\,988}{30} \approx 132{,}9$

Pour la série des hommes : $\overline{x}_h = \dfrac{4\,766}{30} \approx 158{,}9.$

d) $\overline{x}' = \dfrac{1}{60}[4 \times 109 + \cdots + 2 \times 179] \approx 145{,}3$

$$\overline{x}'_f = \frac{1}{30}[4 \times 109 + \cdots + 1 \times 159] \approx 132{,}7$$

$$\overline{x}'_h = \frac{1}{30}[2 \times 139 + \cdots + 2 \times 179] = 158$$

Les différences avec les résultats de la question précédente signifient que la répartition dans chaque classe n'est pas uniforme.

e) Pour la série totale, la valeur de rang 30 est 149 et celle de rang 31 est 150. D'où : $m = \dfrac{149 + 150}{2} = 149{,}5$.

Pour la série des femmes, la valeur de rang 15 est 133 et celle de rang 16 est 134. D'où : $m_f = 133{,}5$.

Pour la série des hommes, la valeur de rang 15 est 156 et celle de rang 16 est 160. D'où : $m_h = 158$.

f) Pour la série totale, la valeur de rang 15 est 133 ; la valeur de rang 16 est 134. D'où : $Q_1 = 133{,}5$.

La valeur de rang 45 est 158 ; la valeur de rang 46 est 160. D'où : $Q_3 = 159$.

Pour la série totale, l'écart interquartile est donc : $Q_3 - Q_1 = 25{,}5$.

Pour la série des femmes, la valeur de rang 8 est 120. D'où : $Q_1 = 120$.

La valeur de rang 23 est 114. D'où : $Q_3 - Q_1 = 25$.

Pour la série des hommes, on obtient de même : $Q_3 - Q_1 = 166 - 151 = 15$.

g) Pour la série des femmes, on a, en utilisant le théorème de Koenigs :

$$V_f = \frac{536\,176}{30} - \left(\frac{3\,988}{30}\right)^2 \approx 201{,}3 \text{ d'où : } \sigma_f = \sqrt{V_f} \approx 14{,}2.$$

Pour la série des hommes :

$$V_h = \frac{759\,954}{30} - \left(\frac{4\,766}{30}\right)^2 \approx 93{,}2 \text{ d'où : } \sigma_h = \sqrt{V_h} \approx 9{,}7.$$

Pour la série complète :

$$V = \frac{759\,954 + 536\,176}{60} - (145{,}9)^2 \approx 315{,}4 \text{ d'où : } \sigma = \sqrt{V} \approx 17{,}8.$$

h) On obtient :

pour la série des femmes : $V'_f \approx 196{,}6$ et $\sigma'_f \approx 14{,}0$

pour la série des hommes : $V'_h = 109$ et $\sigma'_h \approx 10{,}4$

pour la série totale : $V' \approx 313{,}2$ et $\sigma' \approx 17{,}7$

i) Pour la série des femmes :

$$m_1 = \frac{3\,988}{30} \approx 132{,}9\,;\; m_2 = \frac{536\,176}{30} \approx 17\,872{,}5$$

$$m_3 = \frac{1}{30}\sum_{i=1}^{30} x_{if}^3 = \frac{72\,872\,368}{30} \approx 2\,429\,079$$

$$m_4 = \frac{1}{30}\sum_{i=1}^{30} x_{if}^4 = \frac{10\,006\,377\,210}{30} \approx 333\,545\,907$$

$$\mu_2 = m_2 - m_1^2 \approx 201{,}3$$

$$\mu_3 = m_3 - 3m_1m_2 + 2m_1^3 \approx -285{,}4$$

$$\mu_4 = m_4 - 4m_1m_3 + 6m_1^2m_2 - 3m_1^4 \approx 84\,493{,}1$$

Coefficient γ_1 de Fisher : $\gamma_1 = \dfrac{\mu_3}{(\mu_2)^{3/2}} \approx -0{,}100$. Comme $\gamma_1 < 0$, la distribution est étalée vers la gauche.

Coefficient γ_2 de Fisher : $\gamma_2 = \dfrac{\mu_4}{(\mu_2)^2} - 3 \approx -0{,}914$. Comme $\gamma_2 < 0$, la distribution est plus aplatie que celle de la loi de Gauss réduite.

1-5 a) Paramètres des deux séries statistiques de pluviosité

Pour le chêne pédonculé :

$$\overline{P}_1 \approx 1\,073 \text{ mm}\,;\; \sigma_1 \approx 200{,}7 \text{ mm}\,;\; \frac{\sigma_1}{\overline{P}_1} \approx 0{,}19.$$

Pour le chêne pubescent :

$$\overline{P}_2 \approx 868{,}5 \text{ mm}\,;\; \sigma_2 \approx 60{,}6 \text{ mm}\,;\; \frac{\sigma_2}{\overline{P}_2} \approx 0{,}07.$$

b) Paramètres des deux séries statistiques de température

Pour le chêne pédonculé :

$$\overline{T}_1 \approx 10{,}7\ ^\circ\text{C}\,;\; \sigma_3 \approx 0{,}99\ ^\circ\text{C}\,;\; \frac{\sigma_3}{\overline{T}_1} \approx 0{,}09.$$

Pour le chêne pubescent :

$$\overline{T}_2 \approx 12{,}3\ ^\circ\text{C}\,;\; \sigma_4 \approx 0{,}41\ ^\circ\text{C}\,;\; \frac{\sigma_4}{\overline{T}_2} \approx 0{,}03.$$

c) Représentations graphiques de la nature des sols

Si on veut comparer la nature des sols habités par le chêne pédonculé et par le chêne pubescent, ce sont les pourcentages qui interviennent et non les effectifs. Pour ceci, on peut adopter des graphiques circulaires.

Sols	Chêne pédonculé		Chêne pubescent	
	%	angles en °	%	angles en °
acides	77,83	280	14,65	53
calcaires	7,60	27	85,35	307
montagneux	14,57	53	0	0
total	100,00	360	100,00	360

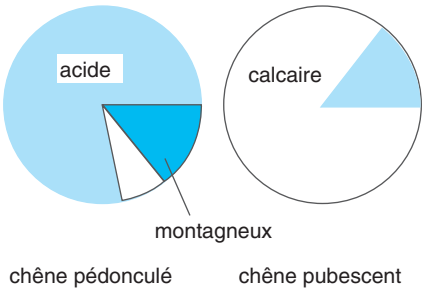


Figure 1-3

d) Conclusions écologiques

On observe que le chêne pubescent préfère un climat plus chaud ($\overline{T}_2 > \overline{T}_1$) et plus sec ($\overline{P}_2 < \overline{P}_1$) que le chêne pédonculé. Par ailleurs, les sols occupés par le chêne pédonculé sont le plus souvent acides et ceux occupés par le chêne pubescent sont souvent calcaires. On peut dire aussi que le chêne pédonculé est une espèce plus résistante car il accepte des températures et des précipitations plus variées (coefficients de variation plus élevés).

En fait, pour la série de végétation étudiée, c'est la nature du sol qui est le facteur primordial.

CHAPITRE 2

Statistique à deux dimensions

PLAN

2.1 Distribution à deux dimensions
2.2 Paramètres d'une série statistique double
2.3 Ajustement

OBJECTIFS

- Savoir déduire des informations d'une étude menée simultanément sur deux caractères X et Y.
- Modéliser une dépendance affine entre un caractère numérique Y et un caractère numérique X susceptible d'expliquer Y.
- Quantifier la qualité du modèle ainsi obtenu.

2.1 DISTRIBUTION À DEUX DIMENSIONS

Généralités

Déterminer une distribution statistique à deux dimensions relative au couple (X, Y), c'est connaître:

– les valeurs possibles $x_1,\ldots,x_p$ pour le caractère statistique X (ou les modalités, ou les classes) ;

– les valeurs possibles $y_1,\ldots,y_q$ pour le caractère statistique Y (ou les modalités, ou les classes) ;

– l'effectif n_{ij} correspondant à chaque observation $(X = x_i$ et $Y = y_j)$.

Si n désigne l'effectif total, la fréquence correspondante est $f_{ij} = \dfrac{n_{ij}}{n}$.

Ces renseignements se présentent souvent avec un tableau à double entrée.

Distributions marginales

À partir de la distribution statistique du couple (X, Y), on peut déduire la distribution statistique concernant le caractère X seul, et celle qui est relative au caractère Y seul :

$(X = x_i)$ a pour effectif : $n_{i\bullet} = \sum_{j=1}^{q} n_{ij}$ et pour fréquence : $f_{i\bullet} = \frac{n_{i\bullet}}{n}$.

$(Y = y_j)$ a pour effectif : $n_{\bullet j} = \sum_{i=1}^{p} n_{ij}$ et pour fréquence : $f_{\bullet j} = \frac{n_{\bullet j}}{n}$.

La détermination des effectifs $n_{i\bullet}$ et $n_{\bullet j}$ se fait à partir du tableau à double entrée par addition suivant les lignes et les colonnes, et en reportant les résultats en marge du tableau.

Distributions conditionnelles

a) Distribution conditionnelle de Y pour $X = x_i$

C'est la distribution des $n_{i\bullet}$ observations vérifiant la condition $X = x_i$ et réparties selon les valeurs prises par Y. Pour ceci, il suffit d'extraire du tableau à double entrée la ligne correspondant à $X = x_i$.

On obtient des fréquences conditionnelles en divisant ces effectifs par $n_{i\bullet}$.

b) Distribution conditionnelle de X pour $Y = y_j$

De la même manière, c'est la distribution des $n_{\bullet j}$ observations vérifiant la condition $Y = y_j$. Et en divisant par le total $n_{\bullet j}$ de la colonne j, on obtient des fréquences conditionnelles.

Indépendance statistique

Deux caractères statistiques X et Y sont dits indépendants si :

$\forall i \quad \forall j \qquad f_{ij} = f_{i\bullet} \times f_{\bullet j}$ ou, ce qui revient au même :

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

Pour qu'il y ait indépendance, il faut que l'égalité ait toujours lieu. Pour démontrer qu'il n'y a pas indépendance, il suffit de fournir un seul cas où l'égalité n'a pas lieu.

L'indépendance statistique de X et de Y correspond au fait que les lignes sont proportionnelles, ainsi que les colonnes.

L'indépendance statistique de X et de Y correspond à la fois :

– à l'indépendance de Y par rapport à X : les fréquences conditionnelles de Y pour $X = x_i$ ne dépendent pas de i ;

– à l'indépendance de X par rapport à Y : les fréquences conditionnelles de X pour $Y = y_j$ ne dépendent pas de j.

2.2 PARAMÈTRES D'UNE SÉRIE STATISTIQUE DOUBLE

Moyennes et variances marginales

$\overline{x} = \frac{1}{n}\sum_{i=1}^{p} n_{i\bullet}\, x_i$; $\overline{y} = \frac{1}{n}\sum_{j=1}^{q} n_{\bullet j}\, y_j$. $G(\overline{x},\overline{y})$ est le point moyen.

$$V(X) = \sigma^2(X) = \frac{1}{n}\sum_{i=1}^{p} n_{i\bullet}(x_i - \overline{x})^2 = \frac{1}{n}\left(\sum_{i=1}^{p} n_{i\bullet} x_i^2\right) - (\overline{x})^2$$

$$V(Y) = \sigma^2(Y) = \frac{1}{n}\sum_{j=1}^{q} n_{\bullet j}(y_j - \overline{y})^2 = \frac{1}{n}\left(\sum_{j=1}^{q} n_{\bullet j} y_j^2\right) - (\overline{y})^2$$

Covariance

Définition

$$\mathrm{Cov}(X,Y) = \frac{1}{n}\sum_{\substack{1\le i\le p\\ 1\le j\le q}} n_{ij}(x_i - \overline{x})(y_j - \overline{y})$$

$$= \frac{1}{n}\left(\sum_{\substack{1\le i\le p\\ 1\le j\le q}} n_{ij}\, x_i\, y_j\right) - \overline{x}\,\overline{y}$$

Propriétés

- $\mathrm{Cov}(aX + b, cY + d) = ac\,\mathrm{Cov}(X,Y)$
- $\mathrm{Cov}(X,X) = V(X)$
- $|\mathrm{Cov}(X,Y)| \le \sigma(X)\sigma(Y)$.
- Si les caractères X et Y sont indépendants, alors Cov $(X, Y) = 0$. Attention, la réciproque est fausse.

Corrélation

Définition. On appelle **coefficient de corrélation linéaire** de X et de Y le réel r défini par :

$$r = \frac{\text{Cov}(X,Y)}{\sigma(X)\sigma(Y)}.$$

Propriétés

- Le nombre r est invariant pour tout changement d'origine et d'échelle.
- On a toujours $-1 \leq r \leq 1$.
- Si X et Y sont indépendants, alors $r = 0$; la réciproque étant fausse.
- Le nuage des points (x_i, y_j) est une droite si, et seulement si : $r = 1$ (droite à pente positive) ou $r = -1$ (droite à pente négative).
  Si $|r|$ est voisin de 1, on dit qu'il existe une forte corrélation linéaire entre X et Y.

Attention, cela ne signifie pas qu'il existe une relation de cause à effet entre X et Y. La confusion entre corrélation et causalité est une erreur courante.

2.3 AJUSTEMENT

La méthode des moindres carrés

a) Généralités

On considère un nuage de points $(x_1, y_1),\ldots,(x_p, y_p)$ avec des coefficients de pondération $n_1,\ldots,n_p$ (le plus souvent ces coefficients sont tous égaux à 1).

L'allure du nuage de points et des considérations sur le phénomène étudié peuvent suggérer une relation fonctionnelle entre x et y, par exemple :

$y = ax + b$; $y = ax^b$; $y = a\ln x + b\ldots$

Après avoir choisi un modèle, une distance entre les points expérimentaux donnés et une courbe du type choisi, on détermine les valeurs des paramètres qui rendent la distance minimum.

b) Ajustement par une droite

Quand les points expérimentaux sont à peu près alignés, on retient comme modèle $y = ax + b$ (ajustement affine), ou $y = ax$ (ajustement linéaire) quand la droite passe obligatoirement par l'origine.

Dans la **méthode des moindres carrés**, on choisit de rendre minimum la distance S, somme des carrés des écarts verticaux.

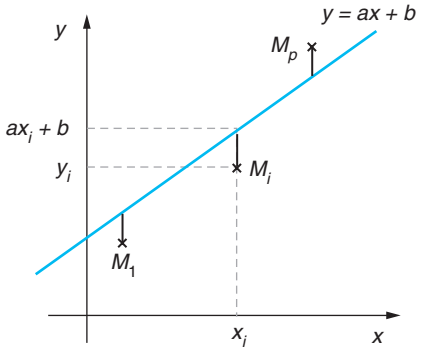


Figure 2-1

Droite de régression de *Y* par rapport à *X*

La droite d'équation $y = ax + b$ qui rend S minimum est celle qui passe par le point moyen $M(\overline{x}, \overline{y})$ et dont la pente est égale à :

$$a = \frac{\text{Cov}(X,Y)}{V(X)}.$$

Cette droite s'appelle la droite de régression de Y par rapport à X.

La covariance de X et de Y est définie par :

$$\text{Cov}(X,Y) = \frac{1}{n}\sum_{i=1}^{p} n_i (x_i - \overline{x})(y_i - \overline{y}) = \frac{1}{n}\left(\sum_{i=1}^{p} n_i x_i y_i\right) - \overline{x}\,\overline{y}$$

La deuxième expression se retient par :

moyenne des produits – produit des moyennes.

- L'écriture de Cov(X, Y) est modifiée car, ici, la série statistique n'est pas donnée par un tableau à double entrée, mais par des points expérimentaux numérotés avec un seul indice.
- De nombreuses calculatrices donnent directement r et les coefficients de la droite de régession. Regardez si votre machine donne $y = ax + b$ ou $y = a + bx$.

Décomposition de la variance

Quand on cherche une droite de régression de Y par rapport à X, les variables X et Y ne jouent pas le même rôle :

Y est la variable à expliquer ;

X est la variable potentiellement explicative.

Après obtention de la droite de régression de Y par rapport à X, on peut écrire :

$$V(Y) = V(aX+b) + \frac{1}{n}\sum_{i=1}^{p} n_i(y_i - ax_i - b)^2$$

égalité que l'on interprète par :

variance de Y = variance expliquée par l'ajustement affine + variance résiduelle

On constate que :

$$\frac{\text{variance expliquée}}{\text{variance totale}} = \frac{V(aX+b)}{V(Y)} = a^2\frac{V(X)}{V(Y)} = \frac{(\text{Cov}(X,Y))^2}{V(X)V(Y)} = r^2.$$

r^2 apparaît donc comme une mesure de la qualité de l'ajustement affine.

Autre modèle : ajustement linéaire

La droite d'équation $y = ax$ qui rend S minimum est définie par :

$$a = \frac{\sum_{i=1}^{p} n_i\, x_i\, y_i}{\sum_{i=1}^{p} n_i x_i^2}.$$

La qualité de l'ajustement linéaire réalisé peut être mesurée par le nombre :

$$d = \frac{\left(\sum_{i=1}^{p} n_i\, x_i\, y_i\right)^2}{\left(\sum_{i=1}^{p} n_i\, x_i^2\right)\left(\sum_{i=1}^{p} n_i\, y_i^2\right)}$$

On a toujours $0 \leq d \leq 1$.

Tous les points sont alignés si, et seulement si, $d = 1$.

MOTS-CLÉS

- Distributions marginales
- Ajustement affine par la méthode des moindres carrés
- Coefficient de corrélation linéaire

EXERCICES

QCM n°1 et 2. *Soit deux variables aléatoires discrètes indépendantes X et Y caractérisées par les lois de probabilité suivantes:*

X	0	1	2
$\mathbb{P}(X=x_i)$	0,2	0,3	0,5

Y	1	2
$\mathbb{P}(Y=x_i)$	0,3	0,7

On s'intéresse à la somme de X et de Y: $S = X + Y$.

2-1 Cochez la (ou les) proposition(s) exacte(s):

❑ **a)** La moyenne de X est de 0,5.
❑ **b)** La moyenne de X est de 1,5.
❑ **c)** La moyenne de X est de 1,3.
❑ **d)** La moyenne de X est de 1.
❑ **e)** La moyenne de S est de 3.

2-2 Cochez la (ou les) proposition(s) exacte(s):

❑ **a)** $\mathbb{P}(S = 2) = 0{,}23$.
❑ **b)** $\mathbb{P}(S = 2) = 0{,}31$.
❑ **c)** $\mathbb{P}(S = 3) = 0{,}36$.
❑ **d)** $\mathbb{P}(S = 3) = 0{,}72$.
❑ **e)** $\mathbb{P}(S = 3) = 0{,}82$.

2-3 On mesure les pressions artérielles systoliques (X) et diastoliques (Y) chez 40 sujets. On observe $m_X = 126{,}2$; $\sum x_i^2 = 646\,464$; $m_Y = 69{,}225$; $\sum y_i^2 = 196\,421$.
La pente de la droite de régression de Y en fonction de X vaut 0,6319.

❑ **a)** La covariance de X et de Y vaut 113,2 (à 0,1 près).

❑ **b)** Le coefficient de corrélation entre X et Y est positif.

❑ **c)** Si X vaut 120, on s'attend à une valeur moyenne de 86 (à 10^0 près) pour Y.

❑ **d)** La part de la variance de Y expliquée par X est de 0,79 (à 0,01 près).

❑ **e)** La somme des produits $x_i y_i$ vaut 35339 (à 10^0 près).

2-4 Dans une population constituée par des ménages ayant des enfants, on a procédé à l'étude simultanée des deux caractères statistiques quantitatifs :

X le nombre d'enfants ;

Y l'âge du premier enfant.

Les effectifs obtenus figurent dans le tableau ci-dessous :

Y / X	de 0 à 4	de 5 à 9	de 10 à 14	de 15 à 19	de 20 à 24
1	30	28	35	43	21
2	26	35	32	31	27
3	20	29	26	23	18
4	2	15	18	16	19
≥ 5	0	3	4	5	10

a) Déterminez les distributions marginales de X et de Y.

b) Les deux caractères X et Y sont-ils indépendants ?

2-5 On veut voir si la tension artérielle Y est corrélée à l'âge X.

Après mesures et calculs, on obtient :

moyennes : $\overline{x} = 35$; $\overline{y} = 13{,}5$

variances : $V(X) = 64$; $X(Y) = 4$

covariance : $\text{Cov}(X, Y) = 10$.

Calculez et commentez le coefficient de corrélation linéaire entre X et Y.

2-6 Dans la série statistique suivante, x représente le nombre de jours d'exposition au soleil d'une feuille et y le nombre de stomates aérifères au mm^2.

x	2	4	8	10	24	40	52
y	6	11	15	20	39	62	85

En admettant que y est une fonction de x, ajustez à la série une droite d'équation $y = ax + b$ par la méthode des moindres carrés.

2-7 On met au point une méthode de dosage d'une vitamine en s'appuyant sur l'existence généralement observée d'une relation affine entre le diamètre d'une colonie bactérienne et le logarithme de la dose de vitamine contenue dans son milieu de culture.
Avec les résultats expérimentaux ci-dessous,

Dose en µg	10	20	40
Diamètre en mm	2 ; 3 ; 2	3 ; 5 ; 4	6 ; 7 ; 6

quelle est la meilleure estimation de la dose contenue dans un milieu où la colonie bactérienne aurait un diamètre de 3 mm ?

2-8 Une étude théorique de l'évolution d'une population en extinction conduit à penser que le nombre d'individus N de cette population varie avec le temps t suivant une loi du type : $N(t) = ae^{-kt}$ où a et k sont des constantes strictement positives.
On veut déterminer expérimentalement la valeur de la constante k. Pour cela, on observe pendant 8 mois un échantillon composé initialement de 200 individus, notant à la fin de chaque mois le nombre de survivants :

t	1	2	3	4	5	6	7	8
$N(t)$	180	154	140	120	112	97	84	76

a) Déduisez-en une valeur approchée de k lorsque t est exprimé en mois, en utilisant un ajustement affine.

b) Quel sera, à votre avis, le nombre de survivants de cet échantillon à la fin de l'année en cours ? puis à la fin de l'année suivante ?

2-9 Pour une personne, on a fait varier l'intensité du travail fourni X exprimée en kilojoules par minute et on a relevé la fréquence cardiaque Y (nombre de battements par minute). On a obtenu les résultats suivants :

x_i	9,6	12,8	18,4	31,2	36,8	47,2	49,6	56,8
y_i	70	86	90	104	120	128	144	154

a) Représentez ces données par un nuage de points.

b) Calculez le coefficient de corrélation linéaire r.

c) Déterminez la droite de régression de Y par rapport à X.

d) Décomposez la variance de Y en variance expliquée par l'ajustement affine et variance résiduelle.

e) Estimez la fréquence cardiaque lorsque l'intensité du travail fourni est de 30 kilojoules par minute ; puis lorsqu'elle est de 75.

SOLUTIONS

2-1 ❑ a) ❑ b) ☒ c) ❑ d) ☒ e)

Les moyennes de X et de Y s'obtiennent avec la définition :

$E(X) = 0 \times 0{,}2 + 1 \times 0{,}3 + 2 \times 0{,}5 = 1{,}3$,

$E(Y) = 1 \times 0{,}3 + 2 \times 0{,}7 = 1{,}7$.

On a toujours $E(X + Y) = E(X) + E(Y)$,

soit ici $E(S) = 1{,}3 + 1{,}7 = 3$.

2-2 ☒ a) ❑ b) ☒ c) ❑ d) ❑ e)

Il faut décomposer les événements :

• Pour avoir $S = 2$, il faut avoir, soit ($X = 0$ et $Y = 2$), soit ($X = 1$ et $Y = 1$).

Ces deux événements étant incompatibles on a donc :

$\mathbb{P}(S = 2) = \mathbb{P}(X = 0 \text{ et } Y = 2) + \mathbb{P}(X = 1 \text{ et } Y = 1)$.

Par ailleurs les variables X et Y étant indépendantes,

$\mathbb{P}(X = 0 \text{ et } Y = 2) = \mathbb{P}(X = 0) \times \mathbb{P}(Y = 2) = 0{,}2 \times 0{,}7 = 0{,}14$

et $\mathbb{P}(X = 1 \text{ et } Y = 1) = \mathbb{P}(X = 1) \times \mathbb{P}(Y = 1) = 0{,}3 \times 0{,}3 = 0{,}09$.

Donc $\mathbb{P}(S = 2) = 0{,}14 + 0{,}09 = 0{,}23$.

• De la même manière,

$\mathbb{P}(S = 3) = \mathbb{P}(X = 1 \text{ et } Y = 2) + \mathbb{P}(X = 2 \text{ et } Y = 1)$

$= 0{,}3 \times 0{,}7 + 0{,}5 \times 0{,}3 = 0{,}36$.

2-3 ❑ a) ☒ b) ❑ c) ☒ d) ❑ e)

Les informations permettent de calculer :

$$\sigma_X^2 = \frac{1}{n}\sum_i x_i^2 - (m_X)^2 = \frac{646\,464}{40} - 126{,}2^2 = 235{,}16$$

$$\sigma_Y^2 = \frac{1}{n}\sum_i y_i^2 - (m_Y)^2 = \frac{196\,421}{40} - 69{,}225^2 = 118{,}42$$

De $a = \dfrac{\text{Cov}(X,Y)}{\sigma_X^2}$ on déduit : $\text{Cov}(X,Y) = 0{,}6319 \times 235{,}16 = 148{,}6$

$$r = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y} = 0{,}89$$

La valeur de r permet de modéliser par la droite de régression de Y en X :

$y = 0{,}6319x + b$ avec $69{,}225 = 0{,}6319 \times 126{,}2 + b$ soit $b = -10{,}52$.
Pour $x = 120$, le modèle $y = 0{,}6319x - 10{,}52$ donnerait $y = 65{,}3$. Mais on ne sait rien du domaine de validité du modèle.
la part de la variance de Y expliquée par le modèle est $a^2 \dfrac{\sigma_X^2}{\sigma_Y^2} = 0{,}79$.
De $\text{Cov}(X,Y) = \dfrac{1}{n}\sum_i x_i y_i - m_X m_Y$ on déduit $\sum_i x_i y_i = 355\,392$.

2-4 a) Les distributions marginales s'obtiennent par addition en lignes et en colonnes, ce qui donne

➤ pour X:

valeurs	1	2	3	4	≥ 5
effectifs	157	151	116	70	22

➤ pour Y :

valeurs	de 0 à 4	de 5 à 9	de 10 à 14	de 15 à 19	de 20 à 24
effectifs	78	110	115	118	95

Dans les deux cas, le total est le même, ce qui permet une vérification. C'est le total général $n = 516$.

b) Les deux caractères ne sont pas indépendants. Pour le prouver, il suffit de fournir un seul contre-exemple. Pour la première case du tableau, la valeur en cas d'indépendance $\dfrac{157 \times 78}{516} \approx 23{,}7$ est différente de la valeur observée 30.

Il est normal qu'il n'y ait pas indépendance statistique, car l'âge du premier enfant est en général plus élevé quand il y a beaucoup d'enfants.

2-5 On a $r = \dfrac{\text{Cov}(X,Y)}{\sigma(X)\sigma(Y)} = \dfrac{10}{2 \times 8} \approx 0{,}625$.
Ce résultat correspond à une corrélation assez faible.

2-6 Si les calculatrices statistiques sont autorisées, on obtient directement :

$$y = 1{,}5247\,x + 3{,}5056.$$

Sinon, vous devez d'abord calculer $\overline{x} = 20$ et $V(X) \approx 323{,}43$, puis $\overline{y} = 34$ et $V(Y) \approx 754{,}29$ puis $\text{Cov}(X,Y) \approx 493{,}14$.

La droite de régression de Y par rapport à X

– passe par le point moyen $(\overline{x}, \overline{y})$,

– a pour coefficient directeur $\dfrac{\text{Cov}(X,Y)}{V(X)} \approx 1{,}5247$.

Elle a donc pour équation :

$y - 34 = 1{,}5247(x - 20)$ soit $y = 1{,}5247x + 3{,}5056$.

On peut juger la qualité de l'ajustement affine en calculant le coefficient de corrélation linéaire :

$$r = \frac{\text{Cov}(X,Y)}{\sigma(X)\sigma(Y)} \approx \frac{493,14}{17,98 \times 27,46} \approx 0,9984.$$

Comme r est très voisin de 1, les observations expérimentales sont très bien modélisées par la relation $y = ax + b$. Mais n'oubliez pas que la validation du modèle a eu lieu pour x entre 2 et 52.

2-7 Soit y le diamètre (en mm) de la colonie bactérienne, et $x = \ln d$ le logarithme népérien de la dose d de vitamine contenue dans le miieu de la culture.
Il s'agit d'ajuster une droite d'équation $y = ax + b$ aux résultats expérimentaux :

x	ln 10	ln 10	ln 20	ln 20	ln 20	ln 40	ln 40
y	2	3	3	4	5	6	7
effectifs	2	1	1	1	1	2	1

Si les calculatrices statistiques sont autorisées, on obtient directement : y = 2,885 x – 4,422.

Sinon, vous devez d'abord calculer $\overline{x} \approx 3{,}00$ et $V(X) \approx 0{,}320$, puis $\overline{y} \approx 4{,}22$ et $V(Y) \approx 3{,}062$, puis $\text{Cov}(X,Y) \approx 0{,}924$.

La droite de régression de Y par rapport à X

– passe par le point moyen $(\overline{x}, \overline{y})$,

– a pour coefficient directeur $\dfrac{\text{Cov}(X,Y)}{V(X)} \approx 2{,}885$,

et on retrouve la même équation.

On peut juger la qualité de l'ajustement affine en calculant le coefficient de corrélation linéaire :

$$r = \frac{\text{Cov}(X,Y)}{\sigma(X)\sigma(Y)} \approx \frac{0,924}{0,566 \times 1,75} \approx 0,933.$$

Comme r est voisin de 1, les observations expérimentales sont bien modélisées par la relation $y = ax + b$.

Si $y = 3$, on obtient en reportant l'estimation : $x = 2{,}572$,
puis $d = e^x \approx 13{,}1$.

Vous pouviez aussi choisir pour x le logarithme décimal de d. La valeur de x obtenue est évidemment différente, mais avec la bonne fonction réciproque $d = 10^x$, la valeur prévue de d est la même.

2-8 a) Évolution d'une population en extinction

Beaucoup de calculatrices ont une fonction qui permet de rentrer les données brutes et d'obtenir l'ajustement par une fonction exponentielle, donc de répondre directement à la question.

Mais ici, l'énoncé impose de détailler en passant par un ajustement affine.

Le modèle théorique peut aussi s'écrire :

$$\ln N = -kt + \ln a$$

ce qui signifie une dépendance affine entre $\ln N$ et t. Pour obtenir la droite de régression de $\ln N$ par rapport à t, il faut transformer les résultats expérimentaux :

t	0	1	2	3	4	5	6	7	8
$\ln N$	ln 200	ln 180	ln 154	ln 140	ln 120	ln 112	ln 97	ln 84	ln 76

En rentrant ces valeurs dans votre calculatrice, vous obtenez :

$$\ln N = -0{,}12t + 5{,}30.$$

Par identification au modèle, on obtient donc :

$k = 0{,}12$ et $\ln a = 5{,}3$ soit $a = e^{\ln a} = 200{,}4$.

On peut juger la qualité de l'ajustement affine en demandant à sa calculatrice le coefficient de corrélation $r \approx -0{,}9998$. La corrélation étant très forte, le modèle théorique s'ajuste très bien aux données expérimentales.

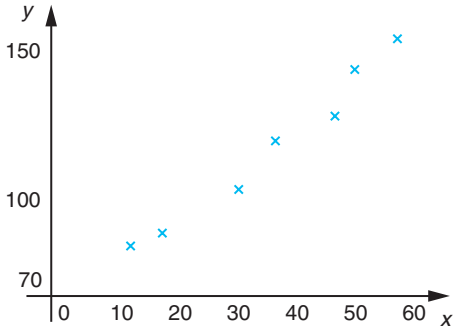
b) Estimation affine

En reportant dans l'équation de la droite de régression de $\ln N$ par rapport à t, on obtient :
pour $t = 12$: $\ln N = 3{,}84$ soit $N \approx 46$
pour $t = 24$: $\ln N = 2{,}38$ soit $N \approx 11$

Mais ce calcul nécessite que le modèle reste valable pour des valeurs de t extérieures à la zone observée. Il s'agit d'extrapolation, et c'est une démarche parfois risquée.

2-9 a) Nuage de points

On observe que les points expérimentaux sont à peu près alignés, ce qui justifie l'hypothèse d'un modèle du type $y = ax + b$.



b) Calculs

$\overline{x} = 32{,}8$; $V(X) = 278{,}72$; $\sigma(X) \approx 16{,}695$

$\overline{y} = 112$; $V(Y) = 762$; $\sigma(Y) \approx 27{,}604$

$\text{Cov}(X,Y) = 454$; $r = \dfrac{\text{Cov}(X,Y)}{\sigma(X)\sigma(Y)} \approx 0{,}985.$

r étant voisin de 1, cela signifie que le modèle $y = ax + b$ traduit bien la réalité expérimentale et que $a > 0$.

c) Droite de régression de *Y* par rapport à *X*

Elle passe par le point moyen $(\overline{x}, \overline{y})$ et a pour coefficient directeur : $a = \dfrac{\text{Cov}(X,Y)}{V(X)} \approx 1{,}629$. La droite de régression de Y par rapport à X a donc pour équation : $y = 1{,}629x + 58{,}573$.

d) Décomposition de la variance de *Y*

Variance expliquée par l'ajustement affine :

$$V(aX + b) = a^2 V(X) \approx 739{,}51.$$

Comme $V(Y) = 762$, la variance résiduelle est :

$$762 - 739{,}51 = 22{,}49.$$

On a : $\dfrac{\text{variance expliquée}}{\text{variance totale}} = r^2 \approx 0{,}970\,484.$

L'ajustement affine permet donc d'expliquer 97 % de la variance totale, ce qui confirme la qualité du modèle affine.

e) Estimation affine

En remplaçant x par 30 dans l'équation de (D_1), on obtient $y = 107$. Cette estimation de la fréquence cardiaque est bonne car la corrélation est forte et on vient de réaliser une interpolation affine.

En remplaçant x par 75, on obtiendrait $y = 181$. Mais il s'agit d'une extrapolation car 75 est en dehors de la zone des valeurs observées et rien ne permet de supposer que le modèle reste valable. En particulier, il est très possible que cette intensité du travail soit insupportable!

Quand on modélise des observations expérimentales, il faut à la fois apprécier la qualité du modèle (ici, c'est le calcul de r) et ne pas oublier qu'un modèle a toujours une zone de validité limitée (dont les bords peuvent être imprécis). On peut avoir une fonction mathématique définie pour des valeurs où elle ne représente plus rien sur le plan expérimental.

CHAPITRE 3

Probabilités (généralités)

PLAN

3.1 Algèbre des événements

3.2 Probabilités : définitions et propriétés

3.3 Construction d'une probabilité sur un univers fini

3.4 Rappels et compléments d'analyse combinatoire

OBJECTIFS

- Comprendre une formalisation élémentaire des premiers concepts : expérience aléatoire, événement, probabilité
- Faire les calculs nécessaires pour construire une probabilité dans le cas d'un nombre fini de possibilités
- Savoir dénombrer diverses situations en comptant tous les cas, une fois, et une seule

3.1 ALGÈBRE DES ÉVÉNEMENTS

Généralités

Une **expérience aléatoire** ε est une expérience qui, répétée dans des conditions apparemment identiques, peut conduire à des résultats différents. L'ensemble de tous les résultats possibles est l'**univers** Ω associé à ε.

On dit qu'un **événement** est lié à ε si, quel que soit le résultat $\omega \in \Omega$, on sait dire si l'événement est réalisé ou non. On convient d'identifier un tel événement à l'ensemble des $\omega \in \Omega$ pour lesquels il est réalisé. Un événement lié à ε est donc identifié à une partie de Ω.

Événements particuliers

Un singleton $\{\omega\}$ est un événement élémentaire.

Ω est l'événement certain car il est toujours réalisé.

$\varnothing$ est l'événement impossible car il n'est jamais réalisé.

Opérations sur les événements

a) Événement contraire $\overline{A}$

$\overline{A}$ est réalisé si, et seulement si, A n'est pas réalisé.

b) Événement $A \cap B$

$A \cap B$ est réalisé si, et seulement si, A et B sont simultanément réalisés. Plus généralement, $\bigcap_{i \in I} A_i$ est réalisé si, et seulement si, tous les événements sont réalisés.

Si $A \cap B = \varnothing$, c'est-à-dire si la réalisation simultanée des événements A et B est impossible, les événements A et B sont **incompatibles**.

c) Événement $A \cup B$

$A \cup B$ est réalisé si, et seulement si, au moins un des événements est réalisé. Plus généralement, $\bigcup_{i \in I} A_i$ est réalisé si, et seulement si, au moins un des événements est réalisé.

d) Système complet d'événements

Une partition de Ω est un système complet d'événements. Autrement dit, des événements $(A_i)_{i \in I}$ forment un système complet s'ils sont différents de $\varnothing$, deux à deux incompatibles et si $\bigcup_{i \in I} A_i = \Omega$.

e) Inclusion

$A \subset B$ signifie que la réalisation de A implique la réalisation de B.

Tribu des événements

a) Dans le cas où Ω est fini, ou dénombrable (en bijection avec $\mathbb{N}$), on retient $\mathcal{P}(\Omega) = \mathcal{T}$ comme ensemble des événements liés à $\mathcal{E}$.

b) Dans le cas où Ω est infini non dénombrable, on retient comme ensemble des événements liés à $\mathcal{E}$, une partie $\mathcal{T}$ de $\mathcal{P}(\Omega)$ qui vérifie les propriétés suivantes :

(1) $\Omega \in \mathcal{T}$

(2) $A \in \mathcal{T} \Rightarrow \overline{A} \in \mathcal{T}$ (stabilité de $\mathcal{T}$ par passage au complémentaire)

(3) Pour toute suite $(A_n)_{n \in \mathbb{N}}$ d'éléments de $\mathcal{T}$, $\bigcup_{n \in \mathbb{N}} A_n = A_0 \cup A_1 \ldots$ est encore un élément de $\mathcal{T}$ (stabilité de $\mathcal{T}$ par réunion dénombrable)

On dit que $\mathcal{T}$ est une **tribu** sur Ω.

Les trois axiomes de définition de $\mathcal{T}$ entraînent les autres propriétés :

(4) $\varnothing \in \mathcal{T}$

(5) Pour toute suite $(A_n)_{n\in\mathbb{N}}$ d'éléments de $\mathcal{T}$, $\bigcap_{n\in\mathbb{N}} A_n = A_0 \cap A_1 \ldots$ est encore un élément de $\mathcal{T}$ (stabilité de $\mathcal{T}$ par intersection dénombrable)

c) Dans tous les cas, on appelle **espace probabilisable** lié à l'expérience aléatoire $\mathcal{E}$, le couple $(\Omega, \mathcal{T})$ où Ω est l'univers des résultats possibles et $\mathcal{T}$ la tribu des événements liés à $\mathcal{E}$.

3.2 PROBABILITÉS : DÉFINITIONS ET PROPRIÉTÉS

Définitions

$(\Omega, \mathcal{T})$ étant un espace probabilisable associé à une expérience aléatoire $\mathcal{E}$, on appelle probabilité sur Ω, toute application P de $\mathcal{T}$ dans $\mathbb{R}$, qui vérifie les axiomes suivants :

(1) $P(\Omega) = 1$

(2) $\forall A \in \mathcal{T} \quad \forall B \in \mathcal{T} \quad A \cap B = \varnothing \Rightarrow P(A \cup B) = P(A) + P(B)$

(3) Pour toute suite $(A_n)_{n\in\mathbb{N}}$ d'événements deux à deux incompatibles, on a : $P\left(\bigcup_{n\in\mathbb{N}} A_n\right) = \sum_{n=0}^{\infty} P(A_n)$ (*cf.* chapitre 6 pour la définition d'une série numérique).

On appelle alors **espace probabilisé** (associé à $\mathcal{E}$) le triplet $(\Omega, \mathcal{T}, P)$.

> Si Ω est fini, l'axiome **(3)** est inutile. Et comme on a alors $\mathcal{T} = \mathcal{P}(E)$, l'espace probabilisé peut se noter (Ω, P).

Propriétés

$$P(\overline{A}) = 1 - P(A) \qquad 0 \leq P(A) \leq 1$$

$$P(\varnothing) = 0 \qquad A \subset B \Rightarrow P(A) \leq P(B)$$

A et B étant des événements quelconques, on a :

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

Système complet d'événements

Si $A_1,\ldots,A_n$ forment un système complet d'événements, on a :

$$\sum_{i=1}^{n} P(A_i) = 1.$$

Dans le cas infini, si $(A_n)_{n\in\mathbb{N}}$ est un système complet d'événements, on a :

$$\sum_{n=0}^{\infty} P(A_n) = 1.$$

3.3 CONSTRUCTION D'UNE PROBABILITÉ SUR UN UNIVERS FINI

Cas général

Soit $\Omega = \{\omega_1,\ldots,\omega_n\}$ un univers fini. Notons A_i l'événement élémentaire $A_i = \{\omega_i\}$.

Théorème. Toute probabilité P sur Ω est entièrement déterminée par la donnée des n nombres réels $p_i = P(A_i)$vérifiant les seules conditions :

$$\forall i \in \{1,\ldots,n\} \quad p_i \geq 0 \quad \text{et} \quad \sum_{i=1}^{n} p_i = 1.$$

Probabilité uniforme sur Ω fini

Dans toutes les situations où aucun événement élémentaire ne doit être distingué des autres, on suppose que tous les événements élémentaires sont équiprobables.

Sur un univers fini Ω, l'hypothèse d'équiprobabilité définit une probabilité P unique, dite **probabilité uniforme** sur Ω, donnée par :

$$\forall A \in \mathcal{P}(\Omega) \qquad P(A) = \frac{\text{card } A}{\text{card } \Omega}.$$

card A (cardinal de A, nombre d'éléments de A) est souvent appelé nombre de cas favorables (sous-entendu à la réalisation de A) et card Ω nombre de cas possibles.

Dans ce cas, le calcul de $P(A)$ se ramène à des problèmes de dénombrement.

3.4 RAPPELS ET COMPLÉMENTS D'ANALYSE COMBINATOIRE

Indications générales

a) Questions à se poser

Pour dénombrer des situations, il est commode de se poser les questions :

- quel est le nombre n d'objets de référence ?
- quel est le nombre p d'objets concernés par une situation ?
- les p objets sont-ils considérés sans ordre (en vrac ; tirage simultané) ou avec ordre (c'est-à-dire que la situation est différente si les mêmes p objets sont classés de façon différente) ?
- les répétitions sont-elles impossibles (les p objets sont tous distincts ; tirage sans remise) ou possibles (tirage avec remise) ?

b) Opérations à effectuer

Quand une situation comporte plusieurs choix :

on effectue un produit quand on doit faire un choix, *puis* un autre …

on effectue une somme quand on considère un cas *ou bien* un autre …

Ne cherchez pas à toujours placer une formule toute faite. Et n'hésitez pas, par exemple, à utiliser un arbre, mais uniquement quand l'ordre compte car un arbre de choix comporte un ordre dans sa structure.

Situations sans répétition

a) Avec ordre

Dans un ensemble à n éléments, il s'agit de choisir p éléments tous distincts (ce qui nécessite $p \leq n$) et avec ordre. Une telle situation est un **arrangement** de n éléments pris p à p. Leur nombre (qu'on peut noter A_n^p) est :

$$n(n-1)\ldots(n-p+1) = \frac{n!}{(n-p)!}.$$

Dans le cas particulier où $p = n$, on dit qu'il s'agit d'une **permutation** d'un ensemble à n éléments ; il y en a $n!$

b) Sans ordre

Dans un ensemble à n éléments, il s'agit de choisir une partie à p éléments (ce qui nécessite $p \leq n$). Leur nombre est le nombre de **combi-**

naisons de n éléments pris p à p (ancienne notation C_n^p). La notation actuelle est :

$$\binom{n}{p} = \frac{n!}{p!(n-p)!}.$$

Propriétés

$$\binom{n}{p} = \binom{n}{n-p} \quad ; \quad \binom{n+1}{p+1} = \binom{n}{p} + \binom{n}{p+1}.$$

Cas particuliers

$$\binom{n}{0} = \binom{n}{n} = 1 \quad ; \quad \binom{n}{1} = \binom{n}{n-1} = n.$$

Situations avec répétition

a) Avec ordre (arrangements avec répétition)

Dans un ensemble à n éléments, il s'agit de choisir p éléments rangés (avec la possibilité de choisir plusieurs fois le même). Il y a n^p possibilités.

b) Sans ordre (combinaisons avec répétition)

Dans un ensemble à n éléments, il s'agit de choisir p éléments sans ordre (avec la possibilité de choisir plusieurs fois le même).

$\binom{n+p-1}{p}$ est le nombre de combinaisons avec répétition.

Il est aussi noté : K_n^p.

c) Permutations avec répétition

Soit un ensemble à n éléments comportant :

n_1 éléments d'un premier type, indiscernables entre eux,

n_2 éléments d'un deuxième type, indiscernables entre eux…

n_q éléments d'un q-ième type, indiscernables entre eux.

Une permutation avec répétition de ces n éléments est une disposition ordonnée de ces éléments. Il y en a $\frac{n!}{n_1!n_2!\ldots n_q!}$.

Tableau récapitulatif des formules de dénombrement

	sans répétition	avec répétition
avec ordre	A_n^p	n^p
sans ordre	$\binom{n}{p}$	$\binom{n+p-1}{p}$

MOTS-CLÉS

- Expérience aléatoire
- Événement
- Probabilité
- Dénombrement

EXERCICES

3-1 On choisit au hasard 3 médecins dans un groupe de 15 médecins dont 5 sont spécialistes. La probabilité qu'aucun médecin ne soit spécialiste parmi ces 3 médecins est égale à:

❑ **a)** $\frac{2}{25}$ ❑ **b)** $\frac{3}{13}$ ❑ **c)** $\frac{24}{91}$ ❑ **d)** $\frac{8}{27}$

❑ **e)** Aucune des propositions précédentes n'est exacte.

3-2 Une boîte contient 11 jetons de poker:
2 jetons de 20 € ; 5 jetons de 100 € ; 1 jeton de 500 € ; 3 jetons de 50 €.
Si l'on choisit dans cette boîte successivement et sans remise 5 jetons, quelle est la probabilité d'avoir exactement la somme de 270 € ?

❑ **a)** $\frac{5}{308}$ soit environ 1,6 %. ❑ **b)** $\frac{2}{11}$ soit environ 18,2 %.

❑ **c)** $\frac{47}{231}$ soit environ 20,3 %. ❑ **d)** $\frac{5}{231}$ soit environ 2,2 %.

❑ **e)** $\frac{1}{924}$ soit environ 0,61 %.

3-3 Dans une population, 45 % des individus sont vaccinés contre la fièvre jaune, 60 % sont vaccinés contre la diphtérie, et 30 % sont vaccinés contre les deux maladies. Quelle est la probabilité, pour un individu choisi au hasard, de n'être vacciné contre aucune de ces deux maladies ?

3-4 Quelle est la probabilité pour que, dans un groupe de n personnes choisies au hasard, deux personnes au moins aient la même date d'anniversaire (on considérera que l'année a 365 jours tous équiprobables) ?

3-5 Un groupe composé de 80 hommes et de 60 femmes doit désigner 10 de ses membres pour être de garde ce soir. Si la désignation se fait au hasard, quelle est la probabilité pour que le groupe de garde

a) ne comporte que des hommes ?

b) ne comporte que des femmes ?

c) comporte un nombre égal d'hommes et de femmes ?

3-6 Une étagère contient 25 livres appartenant à 3 collections différentes, une de 10, une de 8, une de 7 livres. Vus de loin, les ouvrages d'une même collection sont indiscernables. Quel est le nombre d'aspects différents que peut prendre l'étagère vue de loin ?

3-7 18 personnes se sont présentées à une collecte de sang. Parmi celles-ci, on a noté :

11 personnes du groupe O ;

4 personnes du groupe A ;

2 personnes du groupe B ;

1 personne du groupe AB.

À l'issue de la collecte, on prélève au hasard 3 flacons parmi les 18 flacons obtenus. Calculez la probabilité des événements suivants :

a) les sangs des 3 flacons appartiennent au même groupe ;

b) parmi les 3 flacons prélevés, il y a au moins 1 flacon contenant du sang de groupe A ;

c) les sangs des 3 flacons appartiennent à 3 groupes différents.

3-8 On extrait 8 cartes d'un jeu de 52 cartes bien battues. Quelle est la probabilité pour que :

a) 4 cartes soient des as ?

b) 4 cartes soient des as et 2 cartes soient des rois ?

c) l'on ait 3 cartes d'une même couleur et 3 cartes d'une autre couleur (un jeu de 52 cartes comporte 4 couleurs : trèfle, carreau, cœur, pique) ?

d) au moins une carte soit un as ?

SOLUTIONS

3-1 ❑ a) ❑ b) ☒ c) ❑ d) ❑ e)

On choisit 3 médecins parmi 15, sans ordre et sans répétitions. Il y a $\binom{15}{3} = 455$ possibilités.

Pour qu'aucun médecin ne soit spécialiste, il faut choisir 3 médecins parmi 10, soit $\binom{10}{3} = 120$ possibilités.

Le choix étant au hasard, les possibilités sont équiprobables et la probabilité demandée est $\frac{120}{455} = \frac{24}{91}$.

3-2 ❑ a) ❑ b) ❑ c) ☒ d) ❑ e)

Les tirages sont sans remise et l'ordre dans lequel les jetons sont obtenus n'a pas d'importance. Il y a donc $\binom{11}{5} = 462$ tirages possibles.

Après tâtonnements si nécessaire, on s'aperçoit que la seule façon d'obtenir un total de 270 € avec 5 jetons est de choisir 1 jeton de 100 €(5 possibilités), puis 3 jetons de 50 €(1 possibilité) puis 1 jeton de 20 €(5 possibilités), soit $5 \times 1 \times 2 = 10$ possibilités.

La probabilité demandée est donc: $\frac{10}{462} \approx 0{,}0216$ soit 2,2 %.

3-3 Appelons F l'événement « l'individu est vacciné contre la fièvre jaune » et D l'événement « l'individu est vacciné contre la diphtérie ». Le tirage ayant lieu au hasard, on a :

$$P(F) = 0{,}45 \; ; \; P(D) = 0{,}6 \; ; \; P(F \cap D) = 0{,}3,$$

et on demande :

$$\begin{aligned} P(\overline{F} \cap \overline{D}) &= P(\overline{F \cup D}) = 1 - P(F \cup D) \\ &= 1 - [P(F) + P(D) - P(F \cap D)] \\ &= 0{,}25. \end{aligned}$$

3-4

Le mot *au moins*, doit vous faire penser à l'événement contraire $\bar{A}$.

Le nombre de cas possibles est 365^n (arrangements avec répétitions), et le nombre de cas favorables pour $\overline{A}$ est A_{365}^n (arrangements d'ordre n).

Tous les cas sont équiprobables. On a donc :

$$P(A) = 1 - \frac{A_{365}^{n}}{365^{n}} = 1 - \frac{365 \times 364 \times \cdots \times (365 - (n-1))}{365 \times 365 \times \cdots \times 365}$$
$$= 1 - \left(1 - \frac{1}{365}\right)\left(1 - \frac{2}{365}\right)\cdots\left(1 - \frac{n-1}{365}\right)$$

Pour $n = 23$, on obtient $P(A) \approx 0{,}5073$, ce qui signifie que dans un groupe de 23 personnes (et a fortiori s'il y en a plus), il y a plus d'une chance sur deux pour qu'au moins deux personnes aient la même date anniversaire. Comme, en plus, les jours ne sont pas tout à fait équiprobables, la probabilité réelle est encore un peu plus élevée.

3-5 Une équipe de garde étant constituée de 10 personnes prises parmi 140, sans ordre et sans répétition, il y a :

$$\binom{140}{10} = \frac{140 \times 139 \times \cdots \times 131}{1 \times 2 \times \cdots \times 10} \approx 5{,}736\,58 \times 10^{14} \text{ cas possibles.}$$

a) Il y a $\binom{80}{10}$ équipes constituées par 10 hommes. D'où :

$$p_1 = \frac{\binom{80}{10}}{\binom{140}{10}} \approx 0{,}002\,87.$$

b) Il y a $\binom{60}{10}$ équipes constituées par 10 femmes. D'où :

$$p_2 = \frac{\binom{60}{10}}{\binom{140}{10}} \approx 0{,}000\,13.$$

c) Pour constituer une équipe de garde comportant autant d'hommes que de femmes, il faut choisir 5 hommes et 5 femmes. D'où :

$$p_3 = \frac{\binom{80}{5} \times \binom{60}{5}}{\binom{140}{10}} \approx 0{,}228\,87.$$

3-6 Il s'agit de permutations avec répétitions avec $n = 25$, $n_1 = 10$, $n_2 = 8$, $n_3 = 7$. Il y a donc :

$\frac{25!}{10!8!7!} = 21\,034\,470\,600$ situations qui apparaissent différentes.

3-7 Les trois flacons étant distincts et non ordonnés, il y a $\binom{18}{3} = 816$ prélèvements possibles ;

a) Les sangs des trois flacons appartiennent au même groupe s'ils sont :

ou bien du groupe O, soit $\binom{11}{3} = 165$ façons,

ou bien du groupe A, soit $\binom{4}{3} = 4$ façons.

La probabilité de l'événement E_1 est donc :

$$P(E_1) = \frac{165 + 4}{816} \approx 0{,}207.$$

b) L'événement E_2 « au moins un flacon du groupe A » peut se décomposer comme réunion des événements incompatibles : « exactement un flacon du groupe A », « exactement deux flacons du groupe A », « exactement trois flacons du groupe A », ce qui donne :

$$P(E_2) = \frac{\binom{4}{1} \times \binom{14}{2} + \binom{4}{2} \times \binom{14}{1} + \binom{4}{3}}{\binom{18}{3}} = \frac{452}{816} \approx 0{,}554.$$

Mais on peut aller plus vite en considérant l'événement contraire $\overline{E_2}$ « aucun flacon n'est du groupe A ».

$$P(\overline{E_2}) = \frac{\binom{14}{3}}{\binom{18}{3}} = \frac{364}{816} \text{ d'où : } P(E_2) = 1 - P(\overline{E_2}) = \frac{452}{816}.$$

c) L'événement E_3 « les sangs des 3 flacons appartiennent à 3 groupes différents » peut se décomposer comme réunion d'événements deux à deux incompatibles :

– les groupes sont $\{O, A, B\}$, ce qui correspond à $11 \times 4 \times 2 = 88$ prélèvements ;
– les groupes sont $\{O, A, AB\}$, ce qui correspond à $11 \times 4 \times 1 = 44$ prélèvements ;
– les groupes sont $\{O, B, AB\}$, ce qui correspond à $11 \times 2 \times 1 = 22$ prélèvements ;
– les groupes sont $\{A, B, AB\}$, ce qui correspond à $4 \times 2 \times 1 = 8$ prélèvements.

On obtient donc :

$$P(E_3) = \frac{88 + 44 + 22 + 8}{816} = \frac{162}{816} \approx 0{,}199.$$

3-8 Les 8 cartes étant distinctes et non ordonnées, il y a $\binom{52}{8} = 752\,538\,150$ tirages possibles.

a) Pour obtenir un tirage comportant 4 as, il faut choisir 4 as (1 possibilité) et 4 autres cartes, de $\binom{48}{4} = 194\,580$ façons.

La probabilité de E_1 est donc : $P(E_1) = \dfrac{194\,580}{752\,538\,150} \approx 2{,}6 \times 10^{-4}$.

b) Pour réaliser l'événement E_2, il faut choisir 4 as (1 possibilité) et 2 rois de $\binom{4}{2} = 6$ façons, et 2 autres cartes de $\binom{44}{2} = 946$ façons. D'où :

$$P(E_2) = \frac{1 \times 6 \times 946}{752\,538\,150} \approx 7{,}5 \times 10^{-5}.$$

c) Pour réaliser l'événement E_3, il faut choisir les 2 couleurs concernées de $\binom{4}{2} = 6$ façons, puis les 3 cartes de la première couleur choisie de $\binom{13}{3} = 286$ façons, puis 3 cartes de la deuxième couleur choisie de $\binom{13}{3} = 286$ façons.

D'où : $P(E_3) = \dfrac{6 \times 286 \times 286 \times 325}{752\,538\,150} \approx 0{,}212$.

c) L'événement E_4 peut se décomposer en : un as, ou deux as, ou trois as, ou quatre as.
Mais on peut aller plus vite en considérant l'événement contraire $\overline{E_4}$ « aucun as » : $P(\overline{E_4}) = \dfrac{\binom{48}{8}}{\binom{52}{8}} = \dfrac{377\,348\,994}{752\,538\,150}$

d'où : $P(E_4) = 1 - P(\overline{E_4}) = \dfrac{375\,189\,156}{752\,538\,150} \approx 0{,}4986$.

CHAPITRE 4

Probabilité conditionnelle

PLAN

4.1 Probabilité conditionnelle

4.2 Utilisation des probabilités conditionnelles lors d'un test diagnostique

4.3 Événements indépendants

4.4 Formule de Bayes

4.5 Expériences aléatoires successives

OBJECTIFS

- Étudier la modification de probabilité entraînée par une information
- Formaliser la notion d'indépendance entre deux événements, deux expériences aléatoires
- Calculer la probabilité de diverses hypothèses quand un événement vient d'avoir lieu

4.1 PROBABILITÉ CONDITIONNELLE

Définition

Soit (Ω, τ, P) un espace probabilisé et A un événement tel que $P(A) \neq 0$.

Pour un événement quelconque B, on appelle probabilité conditionnelle de B sachant que A est réalisé, le nombre :

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

Utilisation

Il est courant de connaître directement $P(B/A)$. On utilise alors la relation sous la forme, appelée formule des probabilités composées :

$$P(A \cap B) = P(A) \times P(B|A).$$

Généralisation

La formule des probabilités composées se généralise au cas de n événements ($n \geq 2$).

Par exemple, pour trois événements A,B,C tels que $P(A) \neq 0$ et $P(A \cap B) \neq 0$, on peut écrire :

$$P(A \cap B \cap C) = P(A) \times P(B|A) \times P(C|A \cap B).$$

4.2 UTILISATION DES PROBABILITÉS CONDITIONNELLES LORS D'UN TEST DIAGNOSTIQUE

Recueil des données

On recrute des sujets soumis à une évaluation permettant de savoir s'ils ont la maladie étudiée (M) ou non ($\overline{M}$).

On leur applique un test qui donne un résultat positif (+) ou négatif (–).

Les notions qui suivent sont inchangées s'il s'agit d'un symptôme présent (+) ou absent (–).

On regroupe les effectifs observés selon le tableau :

	M	$\overline{M}$	
+	n_1	n_2	
–	n_3	n_4	
			n

- Il y a n_1 individus **vrais positifs** qui sont déclarés positifs alors qu'ils sont malades.
- Il y a n_2 individus **faux positifs** qui sont déclarés positifs alors qu'ils ne sont pas malades.
- Il y a n_3 individus **faux négatifs** qui sont déclarés négatifs alors qu'ils sont malades.
- Il y a n_4 individus **vrais négatifs** qui sont déclarés négatifs alors qu'ils ne sont pas malades.

Évaluation du test diagnostique (population connue)

- La **sensibilité** du test est la probabilité qu'un sujet soit positif au test sachant qu'il est malade :

$$S_e = \mathbb{P}(+|M) = \frac{n_1}{n_1 + n_3}.$$

- La **spécificité** du test est la probabilité qu'un sujet soit négatif au test sachant qu'il n'est pas malade :

$$S_p = \mathbb{P}(-|\bar{M}) = \frac{n_4}{n_2 + n_4}.$$

Utilisation du test diagnostique (population inconnue)

- La **valeur prédictive positive** est la probabilité qu'un sujet soit réellement malade sachant qu'il est positif au test :

$$VPP = \mathbb{P}(M|+).$$

En désignant par $x = \mathbb{P}(M)$ la prévalence de la maladie, on a :

$$VPP = \frac{xS_e}{xS_e + (1-x)(1-S_p)}.$$

- La **valeur prédictive négative** est la probabilité qu'un sujet ne soit pas malade sachant qu'il est négatif au test :

$$VPN = \mathrm{P}(\bar{M}|-).$$

En désignant par $x = \mathbb{P}(M)$ la prévalence de la maladie, on a :

$$VPN = \frac{(1-x)S_p}{x(1-S_e) + (1-x)S_p}.$$

- Le **rapport de vraisemblance positif** est le rapport entre la probabilité d'avoir un test positif lorsque l'individu est malade et la probabilité d'avoir un test positif lorsque l'individu n'est pas malade.

$$RVP = \frac{P(+|M)}{P(+|\bar{M})} = \frac{S_e}{1-S_p}$$

Il quantifie l'apport d'un test positif.

- Le **rapport de vraisemblance négatif** est le rapport entre la probabilité d'avoir un test négatif lorsque l'individu est malade et la probabilité d'avoir un test négatif lorsque l'individu n'est pas malade.

$$RVN = \frac{P(-|M)}{P(-|\bar{M})} = \frac{1-S_e}{S_p}.$$

Il quantifie l'apport d'un test négatif.

4.3 ÉVÉNEMENTS INDÉPENDANTS

Définition

Dans un espace probabilisé (Ω, τ, P), deux événements A et B sont dits indépendants si, et seulement si :

$$P(A \cap B) = P(A) \times P(B).$$

Ne confondez pas événements indépendants (la réalisation de l'un ne modifie pas la probabilité de l'autre) et événements incompatibles (la réalisation de l'un empêche la réalisation de l'autre).

Propriétés

A et B étant deux événements d'un espace probabilisé (Ω, τ, P), on a :

A et B indépendants $\Leftrightarrow$ A et $\overline{B}$ indépendants

$\Leftrightarrow$ $\overline{A}$ et B indépendants

$\Leftrightarrow$ $\overline{A}$ et $\overline{B}$ indépendants

Généralisation

Trois événements A,B,C sont indépendants dans leur ensemble, s'ils sont indépendants deux à deux, soit : $P(A \cap B) = P(A) \times P(B)$

$P(B \cap C) = P(B) \times P(C)$; $P(C \cap A) = P(C) \times P(A)$
et si de plus : $P(A \cap B \cap C) = P(A) \times P(B) \times P(C)$.

4.4 FORMULE DE BAYES

Formule des probabilités totales

Soit $E_1, \ldots, E_n$ un système complet d'événements. Pour tout événement A, on a :

$$P(A) = \sum_{i=1}^{n} P(E_i) \times P(A|E_i),$$

c'est-à-dire :

$$P(A) = P(E_1) \times P(A|E_1) + \ldots + P(E_n) \times P(A|E_n).$$

Formule de Bayes

a) Cas général

$E_1, \ldots, E_n$ étant un système complet d'événements et A un événement tel que $P(A) \neq 0$, on a :

$$\forall j \in \{1,\ldots,n\} \quad P(E_j|A) = \frac{P(E_j) \times P(A|E_j)}{\sum_{i=1}^{n} P(E_i) \times P(A|E_i)}$$

b) Cas particulier le plus utilisé

Comme E_1, $\overline{E}_1$ forment un système complet d'événements, on obtient :

$$P(E_1|A) = \frac{P(E_1) \times P(A|E_1)}{P(E_1) \times P(A|E_1) + P(\overline{E}_1) \times P(A|\overline{E}_1)}$$

$$P(\overline{E}_1|A) = \frac{P(\overline{E}_1) \times P(A|\overline{E}_1)}{P(E_1) \times P(A|E_1) + P(\overline{E}_1) \times P(A|\overline{E}_1)}$$

c) Visualisations

L'usage de la formule de Bayes peut être facilité par des représentations graphiques comme un arbre, ou un tableau d'effectifs après avoir assimilé probabilités et fréquences grâce à la loi des grands nombres.

4.5 EXPÉRIENCES ALÉATOIRES SUCCESSIVES

Expériences indépendantes

Des expériences aléatoires sont dites indépendantes si le résultat de l'une n'influence pas le résultat de l'autre.

Lors de la réalisation d'expériences aléatoires successives $\varepsilon_1, \ldots, \varepsilon_n$, un événement du type « réaliser l'événement A_1 lors de ε_1 et A_2 lors de ε_2 … et A_n lors de ε_n » peut se coder $A_1A_2 \ldots A_n$ ou $A_1 \times A_2 \times \cdots \times A_n$.

Si on a défini les probabilités $P(A_1)$, $P(A_2)$, … , $P(A_n)$, et si les expériences $\varepsilon_1, \ldots, \varepsilon_n$ sont indépendantes, on définit la probabilité de $A_1A_2 \ldots A_n$ en posant :

$$P(A_1A_2 \ \ldots \ A_n) = P(A_1) \times \cdots \times P(A_n).$$

Loi des grands nombres

Supposons que n expériences aléatoires successives, indépendantes, soient décrites par le même espace probabilisé.

On démontre alors que la fréquence d'apparition d'un événement A « tend » vers sa probabilité lorsque n tend vers l'infini.

De ce fait, pour une population de grande taille, un expérimentaliste assimile souvent probabilité et fréquence.

Risque relatif

Considérons une personne tirée au hasard dans une population et notons :

M l'événement « il présente la maladie M » (exemple, un cancer du poumon)

C l'événement « il présente le critère C » (exemple, il fume plus de …).

Le risque relatif d'être atteint de M pour ceux qui présentent le critère C par rapport à ceux qui ne le présente pas, est :

$$\frac{P(M|C)}{P(M|\overline{C})}.$$

Si ce quotient vaut 4 dans l'exemple, cela signifie que ceux qui fument plus de … ont 4 fois plus de risque d'avoir un cancer du poumon que les autres.

MOTS-CLÉS

- Probabilité conditionnelle
- Indépendance
- Formule de Bayes

EXERCICES

4-1 Soit A et B deux événements tels que A $\subset$ B. Les événements A et B sont indépendants si, et seulement si :

- ❑ **a)** $\mathbb{P}(A) = \mathbb{P}(B) = 0$.
- ❑ **b)** $\mathbb{P}(A) \in [0\,;\,1]$.
- ❑ **c)** $\mathbb{P}(A) = 0$ ou $\mathbb{P}(B) = 1$.
- ❑ **d)** $\mathbb{P}(A) + \mathbb{P}(B) = 1$.
- ❑ **e)** Aucune des propositions précédentes n'est exacte.

4-2 Dans une population hétérosexuelle, on veut connaître la performance d'un test de dépistage d'une maladie sexuellement transmissible. Le test peut être positif ou négatif. Dans cette population, la probabilité d'une infection est de 0,001.

Le test de dépistage a les performances suivantes:

$\mathbb{P}(\text{Test positif} \mid \text{malade}) = 0{,}999$;

$\mathbb{P}(\text{Test négatif} \mid \text{non malade}) = 0{,}992$.

Quelle est la probabilité que le sujet soit malade si le test est positif?

❑ **a)** 0,005. ❑ **b)** 0,11. ❑ **c)** 0,23. ❑ **d)** 0,50. ❑ **e)** 0,89.

4-3 Dans une maison de retraite, 80 % des résidents sont des femmes. 20 % d'entre elles souffrent de maladie d'Alzheimer. 40 % des hommes souffrent également de cette maladie.

❑ **a)** La probabilité qu'un résident choisi au hasard souffre de la maladie d'Alzheimer est de 24 %.

❑ **b)** Il y a 5 fois moins d'hommes que de femmes dans cet institut.

❑ **c)** La probabilité d'être un homme et de souffrir de la maladie d'Alzheimer vaut 0, 40.

❑ **d)** Sachant qu'une personne choisie au hasard souffre de la maladie d'Alzheimer, la probabilité que ce soit une femme vaut 2/3.

❑ **e)** Il y a, en nombre de malades, deux fois moins d'hommes que de femmes qui souffrent de la maladie d'Alzheimer dans cet institut.

4-4 Dans une population donnée, une cardiopathie congénitale a une prévalence de 0,6 %. Dans le cas d'un mariage consanguin, le risque qu'un enfant soit atteint de la maladie est de 30 %. Dans le cas d'un mariage non consanguin le risque pour l'enfant est de 0,3 %. **Quelle est la fréquence des mariages consanguins ?**

❑ **a)** Environ 0, 1 %.

❑ **b)** Environ 1 %.

❑ **c)** Environ 1, 8 %.

❑ **d)** Environ 3 %.

❑ **e)** Environ 5 %.

4-5 Dans une certaine population, il y a 45 % de fumeurs et 35 % de personnes atteintes de bronchite.

Sachant que parmi les fumeurs il y a 65 % de bronchiteux, calculez la probabilité pour qu'une personne atteinte de bronchite soit fumeur.

4-6 Dans un espace probabilisé (Ω, τ, P), on considère deux événements A et B tels que : $P(A) = 0{,}5$; $P(B) = 0{,}3$; $P(A \cup B) = 0{,}65$. Les événements A et B sont-ils indépendants ?

4-7 Un laboratoire a mis au point un alcootest. On sait que 2 % des personnes contrôlées par la police sont réellement en état d'ébriété.

Les premiers essais ont conduit aux résultats suivants :

– lorsqu'une personne est réellement en état d'ébriété, 95 fois sur 100 l'alcootest se révèle positif ;

– lorsqu'une personne n'est pas en état d'ébriété, 96 fois sur 100 l'alcootest se révèle négatif.

Quelle est la probabilité pour qu'une personne soit réellement en état d'ébriété lorsque l'alcootest est positif ?

4-8 Dans une population Ω, deux maladies M_1 et M_2 sont présentes respectivement chez 10 % et 20 % des individus (le nombre de ceux qui souffrent des deux maladies est négligeable).

On entreprend un dépistage systématique des maladies M_1 et M_2. Pour cela, on applique un test qui réagit à la maladie sur 90 % des malades de M_1, sur 70 % des malades de M_2, et sur 10 % des individus qui n'ont aucune de ces deux affections.

a) Quand on choisit au hasard un individu ω de Ω, quelle est la probabilité pour que le test réagisse ?

b) Sachant que pour cet individu ω le test a réagi, donnez les probabilités pour que ce soit à cause de la maladie M_1, à cause de la maladie M_2, sans que ω ait l'une des deux maladies.

c) On hospitalise les gens dont le test est positif, pour examens divers et éventuellement traitement. En moyenne le coût pour un malade de M_1 est de 1 500 €, pour un malade de M_2 il est de 1 000 €, et pour un non malade il est de 400 €.

Donnez la moyenne de ce coût sur l'ensemble des individus ayant un test positif.

Si on répartit le coût uniformément sur l'ensemble de la population Ω, combien devra payer chaque individu de Ω ?

4-9 Un scanner peut être utilisé pour détecter des lésions des artères coronaires. Le résultat du scanner est soit « artères coronaires normales », soit « artères coronaires anormales ».

La sensibilité du scanner est de 95 %, tandis que sa spécificité est de 85 %.

a) Calculez la probabilité que le résultat du scanner soit « anormal » si les artères sont saines.

b) On applique ce test à une population à risque moyen où la prévalence de la maladie est de 30 %. Déterminez la valeur prédictive positive et la valeur prédictive négative et du scanner.

SOLUTIONS

4-1 ❑ **a)** ❑ **b)** ☒ **c)** ❑ **d)** ❑ **e)**

Comme $A \subset B$, on a $A \cap B = A$. Les deux événements sont alors indépendants si, et seulement si: $\mathbb{P}(A \cap B) = \mathbb{P}(A) = \mathbb{P}(A) \times \mathbb{P}(B)$.
Pour $\mathbb{P}(A) = 0$, cette égalité est vérifiée.
Pour $\mathbb{P}(A) \neq 0$, cette égalité se ramène à $\mathbb{P}(B) = 1$.

4-2 ❑ **a)** ☒ **b)** ❑ **c)** ❑ **d)** ❑ **e)**

- **Avec la formule de Bayes**

$$\mathbb{P}(M|+) = \frac{\mathbb{P}(M)\mathbb{P}(+|M)}{\mathbb{P}(M)\mathbb{P}(+|M) + \mathbb{P}(\bar{M})\mathbb{P}(+|\bar{M})}$$
$$= \frac{0{,}001 \times 0{,}999}{0{,}001 \times 0{,}999 + 0{,}999 \times (1 - 0{,}992)} = \frac{1}{9} \approx 0{,}11.$$

- **Avec un tableau d'effectifs**

Prenons un échantillon représentatif de grande taille (1 000 000 de personnes pour n'avoir que des nombres entiers d'individus). En assimilant probabilités et fréquences, nous pouvons construire le tableau:

	test positif	test négatif	totaux
malade	999	1	1000
non malade	7992	991 008	999 000
totaux	8991	991 009	1 000 000

La probabilité demandée est $\mathbb{P}(M|+) = \frac{999}{8991} = \frac{1}{9} \approx 0{,}11$.

4-3 ☒ **a)** ❑ **b)** ❑ **c)** ❑ **d)** ❑ **e)**

La solution la plus rapide consiste à considérer un représentatif de grande taille (1000 par exemple) pour pouvoir assimiler probabilités et fréquences. Les informations de l'énoncé conduisent au tableau d'effectifs:

	femmes	hommes	totaux
malade	160	80	240
non malade	640	120	760
totaux	800	200	1000

4-4 ☒ **a)** ❑ **b)** ❑ **c)** ❑ **d)** ❑ **e)**

Désignons par M (resp. C) l'événement : un individu pris au hasard dans la population présente la maladie (resp. est l'enfant d'un mariage consanguin).

On connaît $\mathbb{P}(M) = 0{,}006$, $\mathbb{P}(M|C) = 0{,}3$, $\mathbb{P}(M|\overline{C}) = 0{,}003$ et on demande $\mathbb{P}(C)$.

- **Avec un calcul de probabilités**

$$\begin{aligned}\mathbb{P}(M) &= \mathbb{P}(M \cap C) + \mathbb{P}(M \cap \overline{C}) \\ &= \mathbb{P}(C) \times \mathbb{P}(M|C) + \mathbb{P}(\overline{C}) \times \mathbb{P}(M|\overline{C}) \\ &= 0{,}3\,\mathbb{P}(C) + 0{,}003(1 - \mathbb{P}(C)) = 0{,}003 + 0{,}297\,\mathbb{P}(C)\end{aligned}$$

De $0{,}006 = \mathbb{P}(M) = 0{,}003 + 0{,}297\,\mathbb{P}(C)$, on tire alors: $\mathbb{P}(C) = \dfrac{0{,}003}{0{,}297}$ $\approx 0{,}01$.

- **Avec un tableau d'effectifs**

Considérons un échantillon représentatif de la population de grande taille pour pouvoir assimiler les fréquences et les probabilités, par exemple 1000 individus. Désignons par a le nombre inconnu d'individus correspondants à C et reportons les informations:

	C	$\overline{C}$	totaux
M	$0,3\,a$	$0,003\,(1000 - a)$	6
$\overline{M}$	$0,7\,a$	$0,997\,(1000 - a)$	994
totaux	a	$1000 - a$	1000

En lisant la première ligne, on a:

$$0{,}3a + 0{,}003\,(1000 - a) = 6 \Leftrightarrow 0{,}297a = 3 \Leftrightarrow a = \frac{3}{0{,}297} \approx 10$$

soit une fréquence d'environ $\dfrac{10}{1000} = 0{,}01$, ou encore 1 %.

4-5 Considérons l'espace probabilisé associé au tirage au hasard d'une personne dans la population. Notons les événements :

F : « c'est un fumeur » ; B : « c'est un bronchiteux »

En assimilant fréquences et probabilités avec la loi des grands nombres, les hypothèses s'écrivent :

$P(F) = 0{,}45$; $P(B) = 0{,}35$; $P(B|F) = 0{,}65$.

Et on demande :

$$P(F|B) = \frac{P(F \cap B)}{P(B)} = \frac{P(F)\,P(B|F)}{P(B)} = \frac{0{,}45 \times 0{,}65}{0{,}35}$$
$$= \frac{0{,}2925}{0{,}35} \approx 0{,}8357.$$

Pour les rebelles à l'écriture mathématique, l'assimilation entre fréquences et probabilités permet une version plus visuelle. Prenons une population de référence nombreuse, par exemple 10 000 personnes. Il y a donc 4500 fumeurs et 3500 bronchiteux. Parmi les 4500 fumeurs il y en a 65 % qui sont bronchiteux, soit 2925. Il est alors facile de compléter le tableau :

	B	$\overline{B}$	
F	2925	1575	4500
$\overline{F}$	575	4925	5500
	3500	6500	10 000

pour en déduire qu'il y a 2925 fumeurs parmi les 3500 bronchiteux, ce qui conduit au résultat déjà obtenu.

4-6 On a toujours : $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ ce qui conduit ici à : $P(A \cap B) = 0{,}15$.
Comme, d'autre part, $P(A) \times P(B) = 0{,}15$, les événements A et B sont indépendants en probabilité.

4-7 • Appelons E l'événement « la personne contrôlée est en état d'ébriété » et A « l'alcootest est positif ».
Les indications fournies peuvent s'écrire :

$$P(E) = 0{,}02\,;\, P(A|E) = 0{,}95\,;\, P(\overline{A}|\overline{E}) = 0{,}96$$

et on demande : $P(E/A)$.
D'après la formule de Bayes, on a :

$$P(E|A) = \frac{P(E)\,P(A|E)}{P(E)\,P(A|E) + P(\overline{E})\,P(A|\overline{E})}$$
$$= \frac{0{,}02 \times 0{,}95}{0{,}02 \times 0{,}95 + 0{,}98 \times 0{,}04}$$
$$= \frac{0{,}0190}{0{,}0582} \approx 0{,}3265$$

Avec un alcootest pourtant efficace, la faible valeur obtenue provient de la faible valeur de $P(E)$. En médecine, un problème analogue correspondrait à la difficulté de diagnostic d'une maladie peu fréquente à partir d'un seul symptome.

• L'utilisation de la formule de Bayes peut se visualiser par l'arbre suivant :
• L'assimilation entre fréquences et probabilités sur une population nombreuse permet une version accessible aux non-matheux :

	A	$\overline{A}$	
E	190	10	200
$\overline{E}$	392	9408	9800
	582	9418	10 000

La probabilité conditionnelle demandée $P(E|A)$ est donnée par la fréquence conditionnelle : $\frac{190}{582}$.

4-8 Choisissons au hasard un individu ω dans la population Ω et notons :
M_1 l'événement « ω a la maladie M_1 » ;
M_2 l'événement « ω a la maladie M_2 » ;
N l'événement « ω n'a ni la maladie M_1, ni la maladie M_2 ».
De cette façon $\{M_1, M_2, N\}$ constitue un système complet d'événements.
D'autre part, désignons par R l'événement « le test réagit ».
Les informations fournies peuvent s'écrire :
$P(M_1) = 0{,}1$; $P(M_2) = 0{,}2$; $P(N) = 0{,}7$;
$P(R|M_1) = 0{,}9$; $P(R|M_2) = 0{,}7$; $P(R|N) = 0{,}1$;
et se visualiser par l'arbre (figure 4.2).

a) Probabilité de *R*

D'après la formule des probabilités totales, on a :

$$P(R) = P(M_1) \times P(R|M_1) + P(M_2) \times P(R|M_2) + P(N) \times P(R|N) = 0{,}3$$

La probabilité pour que le test réagisse est donc de 0,3.

Sur l'arbre, cela revient à additionner les probabilités des chemins qui se terminent par R.

b) Probabilités des hypothèses quand le test réagit

$$P(M_1|R) = \frac{P(M_1 \cap R)}{P(R)} = \frac{0{,}09}{0{,}3} = 0{,}3$$
$$P(M_2|R) = \frac{P(M_2 \cap R)}{P(R)} = \frac{0{,}14}{0{,}3} = \frac{7}{15} \approx 0{,}47$$
$$P(N|R) = \frac{P(N \cap R)}{P(R)} = \frac{0{,}07}{0{,}3} = \frac{7}{30} \approx 0{,}23$$

Lorsque le test est positif, il y a donc une probabilité de 0,3 que ce soit à cause de M_1 environ 0,47 que ce soit à cause de M_2, environ 0,23 que ω n'ait ni M_1, ni M_2.

Nous venons en fait d'appliquer la formule de Bayes.

c) Coûts d'hospitalisation

• Sur l'ensemble des individus ayant un test positif, on sera amené à dépenser :
1 500 € avec une probabilité 0,3 ;
1 000 € avec une probabilité $\frac{7}{15}$;
400 € avec une probabilité $\frac{7}{30}$.
Le coût moyen sera donc :

$$1\,500 \times 0{,}3 + 1\,000 \times \frac{7}{15} + 400 \times \frac{7}{30} = 1\,010 \text{ €}$$

• Sur l'ensemble de la population, on est amené à dépenser :
1 500 € avec une probabilité 0,09 ;
1 000 € avec une probabilité 0,14 ;
400 € avec une probabilité 0,07.
Le coût moyen sera donc :

$$1\,500 \times 0{,}09 + 1\,000 \times 0{,}14 + 400 \times 0{,}07 = 303 \text{ €}$$

4-9 **a)** Désignons par A l'événement « les artères sont anormales » et par + « le scanner déclare les artères anormales ».
On connaît la sensibilité $Se = \mathbb{P}(+|A) = 0.95$ et la spécificité $Sp = \mathbb{P}(-|\overline{A}) = 0{,}85$.

On demande : $\mathbb{P}(+|\overline{A}) = 1 - \mathbb{P}(-|\overline{A}) = 0{,}15$.

b) On sait que :

$$VPN = \frac{(1-x)Sp}{x(1-Se)+(1-x)Sp} = \frac{0,7 \times 0,85}{0,3 \times 0,05 + 0,7 \times 0,85}$$
$$\approx 0,98.$$
$$VPP = \frac{xS_e}{xS_e + (1-x)(1-Sp)} = \frac{0,3 \times 0,95}{0,3 \times 0,95 + 0,7 \times 0,15}$$
$$\approx 0,73.$$

Mais on peut faire un tableau d'effectifs avec un échantillon représentatif de la population :

	Lésions	Absence de lésions	Total
Résultat « normal »	15	595	610
Résultat « anormal »	285	105	390
Total	300	700	1000

$$VPN = \frac{595}{610} \approx 0,9754 \quad VPP = \frac{285}{390} \approx 0,7308$$

CHAPITRE 5

Variables aléatoires discrètes (cas fini)

PLAN

5.1 Premières définitions

5.2 Variables aléatoires indépendantes

5.3 Opérations sur les variables aléatoires

5.4 Paramètres d'une variable aléatoire

5.5 Lois classiques

OBJECTIFS

- Comprendre une variable aléatoire quand les valeurs possibles sont en nombre fini
- Définir l'espérance mathématique et la variance d'une variable aléatoire et des variables aléatoires obtenues par des opérations algébriques
- Savoir reconnaître et utiliser les lois classiques du chapitre

5.1 PREMIÉRES DÉFINITIONS

Univers et probabilité images

(Ω, P) étant un espace probabilisé fini, on appelle **variable aléatoire** toute application X de Ω dans $\mathbb{R}$. $\Omega_1 = X(\Omega)$ s'appelle l'univers-image.

On définit une probabilité-image en posant :

$$\forall a \in \Omega_1 \quad P_1(\{a\}) = P(\{\omega \in \Omega; X(\omega) = a\}).$$

En fait on utilise des notations abrégées :

$(\{\omega \in \Omega\,;\, X(\omega) = a\})$ se note $X = a$,

$(\{\omega \in \Omega\,;\, X(\omega) < a\})$ se note $X < a$,

$(\{\omega \in \Omega\,;\, a \leq X(\omega) \leq b\})$ se note, $a \leq X \leq b$,

et on écrit $P(\{\omega \in \Omega\,;\, X(\omega) = a\}) = P(X = a)$.

Distribution de probabilité

Si X est une variable aléatoire dont l'univers-image $\{x_1, \ldots, x_n\}$ est probabilisé par la connaissance des nombres $p_i = P(X = x_i)$, la distribution de probabilité, ou loi de probabilité, de X est l'ensemble des couples (x_i, p_i).

Il est toujours utile de vérifier que l'on a bien $p_1 + \cdots + p_n = 1$.

Fonction de répartition

X étant une variable aléatoire, on appelle fonction de répartition associée à X, la fonction de $\mathbb{R}$ dans [0, 1], notée F, et définie par :

$$\forall x \in \mathbb{R} \qquad F(x) = P(X \leq x).$$

5.2 VARIABLES ALÉATOIRES INDÉPENDANTES

Couple de variables aléatoires

Soit X et Y deux variables aléatoires définies sur le même espace probabilisé fini (Ω, P) dont les univers-images sont respectivement :

$$X(\Omega) = \{x_1, \ldots, x_q\} \quad \text{et} \quad Y(\Omega) = \{y_1, \ldots, y_r\}.$$

La loi du couple (X,Y) est définie par la donnée des nombres :

$$p_{ij} = P(X = x_i \text{ et } Y = y_j) \quad \text{où} \quad 1 \leq i \leq q \text{ et } 1 \leq j \leq r.$$

il est commode de reporter ces nombres dans un tableau à double entrée.

Lois marginales

Si on a reporté les nombres p_{ij} dans un tableau à double entrée, en additionnant suivant les lignes et suivant les colonnes, on aboutit aux lois marginales de X et de Y définies par :

$$P(X = x_i) = p_{i\bullet} = \sum_{j=1}^{r} p_{ij} = p_{i1} + p_{i2} + \cdots + p_{ir}$$

$$P(Y = y_j) = p_{\bullet j} = \sum_{i=1}^{q} p_{ij} = p_{1j} + p_{2j} + \cdots + p_{qj}$$

Indépendance de deux variables aléatoires

Les variables aléatoires X et Y sont dites indépendantes si, et seulement si, les événements $(X = x_i)$ et $(Y = y_j)$ sont indépendants pour i et j quelconques, c'est-à-dire :

$$\forall i \quad \forall j \qquad p_{ij} = p_{i\bullet} \times p_{\bullet j}$$

5.3 OPÉRATIONS SUR LES VARIABLES ALÉATOIRES

Addition ou produit par un nombre

Soit X une variable aléatoire définie sur (Ω, P) et a et λ des réels. Les variables aléatoires $X + a$ et λX sont définies sur Ω par :

$$\forall \omega \in \Omega \qquad (X + a)(\omega) = X(\omega) + a \qquad (\lambda X)(\omega) = \lambda X(\omega).$$

Si $\{x_1, \ldots, x_q\}$ est l'univers-image de X, l'univers-image de $X + a$ est : $\{x_1 + a, \ldots, x_q + a\}$ et celui de λX : $\{\lambda x_1, \ldots, \lambda x_q\}$.
Les probabilités-images sont définies par :

$$P(X + a = x_i + a) = P(X = x_i) = P(\lambda X = \lambda x_i).$$

Somme

Soit X et Y deux variables aléatoires définies sur le même espace probabilisé fini (Ω, P). La somme $X + Y$ est la variable aléatoire définie sur Ω par :

$$\forall \omega \in \Omega \qquad (X + Y)(\omega) = X(\omega) + Y(\omega).$$

L'univers-image de $Z = X + Y$ est constitué par les réels z_k du type $z_k = x_i + y_j$. Et on a $P(Z = z_k) = \sum p_{ij}$, la somme étant étendue à tous les couples (i,j) tels que $z_k = x_i + y_j$.

Produit

Le produit XY est la variable aléatoire définie sur Ω par :

$$\forall \omega \in \Omega \qquad (XY)(\omega) = X(\omega)\, Y(\omega).$$

L'univers-image de $T = XY$ est constitué par les réels t_k du type $t_k = x_i y_j$. Et on a $P(T = t_k) = \sum p_{ij}$, la somme étant étendue à tous les couples (i,j) tels que $t_k = x_i\, y_j$.

5.4 PARAMÉTRES D'UNE VARIABLE ALÉATOIRE

Espérance mathématique, variance, écart type

Soit X une variable aléatoire définie sur Ω fini, dont la loi de probabilité est $(x_1, p_1), \ldots, (x_n, p_n)$ où $p_i = P(X = x_i)$.

- L'**espérance mathématique** de X est le réel : $E(X) = \sum_{i=1}^{n} p_i\, x_i$.
- La **variance** de X est le réel : $V(X) = \sum_{i=1}^{n} p_i\, [x_i - E(X)]^2$.
- L'**écart type** de X est le réel $\sigma(X) = \sqrt{V(X)}$.
- **Théorème de Koenigs**

$$V(X) = E(X^2) - (E(X))^2 = \left(\sum_{i=1}^{n} p_i x_i^2\right) - (E(X))^2.$$

Covariance, corrélation

Comme en statistiques, on définit :

- la **covariance** de deux variables aléatoires X et Y :

$$\text{Cov}\,(X,Y) = E[(X - E(X))\,(Y - E(Y))] = E(XY) - E(X)\,E(Y),$$

- le **coefficient de corrélation** de X et de Y : $r = \dfrac{\text{Cov}\,(X,Y)}{\sigma(X)\,\sigma(Y)}$.

Théorèmes

a) Dans le cas général

On a :

$$E(X + a) = E(X) + a\,;\; V(X + a) = V(X)$$
$$E(\lambda X) = \lambda E(X)\,;\, V(\lambda X) = \lambda^2 V(X)$$
$$E(X + Y) = E(X) + E(Y)\,;$$
$$V(X + Y) = V(X) + V(Y) + 2\,\text{Cov}\,(X,Y).$$

b) Dans le cas où X et Y sont indépendantes

On a : $\text{Cov}(X,Y) = 0\,; V(X + Y) = X(X) + V(Y)$;

$$E(XY) = E(X)\,E(Y).$$

Mais ces relations peuvent être vérifiées sans que X et Y soient indépendantes.

Variable centrée réduite

Si X est une variable aléatoire telle que $E(X) = \mu$ et $V(X) = \sigma^2$, on appelle variable centrée réduite associée à X la variable aléatoire $Y = \dfrac{X - \mu}{\sigma}$.

Elle vérifie $E(Y) = 0$ et $V(Y) = 1$.

5.5 LOIS CLASSIQUES

Loi discrète uniforme

a) Loi de probabilité

L'univers-image de X est $\Omega_1 = \{1, \ldots, n\}$ et les probabilités :

$$\forall k \in \Omega_1 \qquad P(X = k) = \frac{1}{n}.$$

b) Paramètres

$$E(X) = \frac{n+1}{2} \qquad ; \qquad V(X) = \frac{n^2 - 1}{12}.$$

Loi binomiale

a) Conditions du modèle

On obtient une loi binomiale quand :

- on répète n fois la même expérience aléatoire, les n répétitions étant indépendantes entre elles ;
- on s'intéresse seulement à la réalisation, ou non, d'un événement fixé A de probabilité p, et on pose $q = 1 - p$;
- on considère la variable aléatoire X égale au nombre de fois où l'événement A a été réalisé au cours des n épreuves.

Dans ces conditions, on dit que X suit la loi binomiale de paramètres n et p. Cette loi se note $\mathcal{B}(n, p)$.

b) Loi de probabilité

L'univers-image de X est $\Omega_1 = \{1, \ldots, n\}$ et les probabilités :

$$\forall k \in \Omega_1 \qquad P(X = k) = \binom{n}{k} p^k q^{n-k}$$

c) Paramètres

$$E(X) = np \qquad ; \qquad V(X) = npq.$$

Loi hypergéométrique

Une loi hypergéométrique dépend de trois paramètres entiers positifs : $N, K \leq N$ et $n \leq N$. En notant l le plus petit des deux entiers K et n, l'univers-image est $\{0,\ldots,l\}$ et on a :

$$\forall k \in \{0,\ldots,l\} \qquad P(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}.$$

En langage imagé, on dispose d'une urne constituée de N boules dont K présentent un type A. On prélève n boules sans remise et on compte le nombre X de boules de type A obtenues.

En fait la loi hypergéométrique est peu utilisée : on l'approxime par une loi binomiale dès que la taille N de la population est grande par rapport à la taille n de l'échantillon. Cela signifie qu'un tirage sans remise est alors assimilé à un tirage avec remise.

En posant $p = \dfrac{K}{N}$ et $q = 1 - p$, on obtient : $E(X) = np$ et $V(X) = npq\,\dfrac{N-n}{N-1}$ où $\dfrac{N-n}{N-1}$ est le facteur d'exhaustivité.

MOTS-CLÉS

- Variable aléatoire
- Distribution de probabilité
- Lois marginales
- Espérance mathématique
- Variance
- Loi uniforme
- Loi binomiale

EXERCICES

5-1 On admet que dans la population 8 % des personnes sont des gauchers. On s'intéresse au nombre de gauchers dans un groupe de 6 personnes prises au hasard.

❑ **a)** La loi du nombre de gauchers dans un groupe de 6 personnes est une loi binomiale.

❑ **b)** La probabilité pour que les 6 personnes soient gauchères est $(0, 08)^6$.

❑ **c)** La probabilité pour que les 6 personnes soient gauchères est 0, 48.

❑ **d)** La probabilité pour que les 6 personnes soient gauchères est $1 - (0, 92)^6$.

❑ **e)** La probabilité pour que 3 des personnes sur 6 soient gauchères est 0, 5.

5-2 Une première dent de lait (l'incisive inférieure) tombe avant l'âge de 5 ans chez 10 % des enfants. Dans une famille de 4 enfants, tous âgés de plus de 5 ans, on veut savoir combien ont perdu leur première dent de lait avant 5 ans.
Si on suppose qu'il n'y a aucun facteur génétique ou environnemental dans la perte des dents de lait, et donc que les cas sont indépendants, on peut dire pour cette famille :

❑ **a)** La loi du nombre d'enfants ayant perdu une dent de lait avant 5 ans est la loi binomiale.

❑ **b)** La probabilité pour que les 4 enfants aient perdu une dent de lait avant 5 ans est $(0, 1)^4$.

❑ **c)** La probabilité pour que les 4 enfants aient perdu une dent de lait avant 5 ans est (0, 4).

❑ **d)** La probabilité pour que les 4 enfants aient perdu une dent de lait avant 5 ans est $1 - (0, 9)^4$.

❑ **e)** La probabilité pour que 2 enfants sur 4 aient perdu une dent de lait avant 5 ans est $\frac{1}{2}$

5-3 Dans une population, la fréquence de la maladie M est de 2 %. **Quelle est la taille minimale d'échantillon à constituer pour que la probabilité d'y observer au moins un malade de M soit supérieure à** 0,75**?**

❑ **a)** 14. ❑ **b)** 25. ❑ **c)** 48. ❑ **d)** 69. ❑ **e)** 95.

5-4 Un clochard suit une route indéfiniment bordée d'arbres alignés, distants les uns des autres de 10 mètres. Il décide, au cours de sa promenade, de jouer au jeu suivant :

Devant chaque arbre, il lance son unique pièce de monnaie. Si la pièce retombe sur pile, il continue dans la même direction. Si elle retombe sur face, il rebrousse chemin jusqu'à l'arbre voisin.

Au bout de six déplacements, il s'endort au pied de l'arbre où il se trouve.

On appelle X la distance arithmétique, en mètres, entre l'arbre devant lequel il commence son jeu et l'arbre d'arrivée.

a) Déterminez la loi de probabilité de cette variable aléatoire sachant que la pièce n'est pas truquée. Quelle est la distance ayant la plus grande probabilité ?

b) Calculez l'espérance mathématique et la variance de X.

5-5 Neuf accidentés passent, un par un, un examen radiologique. Quatre ont une fracture au niveau des membres et cinq au niveau du bassin (aucun ne présente les deux types de fracture). L'ordre de passage est constitué au hasard.

On appelle X la variable aléatoire « nombre d'accidentés des membres précédant le premier accidenté du bassin ».

a) Déterminez la loi de probabilité de X.

b) Calculez l'espérance mathématique et la variance de X.

5-6 Dans une population très nombreuse, des études régulières ont montré qu'il y avait 2 % d'individus de type A.

Calculez la probabilité, dans un échantillon de 100 individus tirés au hasard, d'obtenir :

a) aucun individu du type A ;

b) au moins deux individus du type A.

5-7 L'infarctus du myocarde a six facteurs de risque. La présence ou l'absence de chacun de ces facteurs est modélisé par une loi de Bernoulli. On admet, pour simplifier, que chaque facteur de risque a une probabilité d'apparaître de 0,4 et que chaque facteur de risque est indépendant des autres.

On note N le nombre de facteurs de risque présents simultanément chez un individu pris au hasard.

1) Quelle loi suit N ?

2) Calculez $\mathbb{P}(N > 0)$ et $\mathbb{P}(N = 3)$.

5-8 Une machine à embouteiller peut tomber en panne. La probabilité d'une panne est de 0,01 à chaque emploi de la machine. La machine doit être utilisée 100 fois.

a) Le nombre de pannes observées est une variable aléatoire X. Calculez les probabilités d'obtenir : $X = 0$, $X = 1$, $X = 2$, $X = 3$, $X \geq 4$.

b) On estime le coût d'une réparation à 500 €. La dépense, exprimée en euros, pour les réparations de la machine est une variable aléatoire Y.
Calculez l'espérance mathématique de Y et son écart type.

5-9 Un veilleur de nuit doit ouvrir 12 portes avec 12 clés différentes mais non discernables.

a) Quelle est la probabilité pour qu'il ouvre la première porte au k-ième essai sachant qu'à chaque fois qu'il choisit une clé, il ne la remet pas dans le trousseau si elle ne convient pas.

b) Le nombre total d'essais effectués définit une variable aléatoire X dont on demande de déterminer la distribution de probabilité, l'espérance mathématique et l'écart type.
Pour chaque porte, le processus recommence comme pour la première porte, mais avec seulement les clés restantes.

5-10 En terminant d'effeuiller la marguerite, on compte :
1 point pour *un peu*, 3 points pour *beaucoup*, 5 points pour *passionnément*, 10 points pour *à la folie*, 0 point pour *pas du tout*.

On effeuille successivement deux marguerites. Soit X la variable aléatoire égale au nombre de points obtenu avec la première marguerite.

Soit Y la variable aléatoire égale au plus grand des deux nombres obtenus.

a) Déterminez la loi du couple (X,Y).

b) Précisez les lois marginales de X et de Y. Les variables aléatoires X et Y sont-elles indépendantes ?

c) Déterminez la distribution de probabilité de $Z = X + Y$.

d) Déterminez la distribution de probabilité de $T = XY$.

e) Calculez $E(X)$, $V(X)$, $E(Y)$, $V(Y)$, $E(X + Y)$, $V(X + Y)$, $E(XY)$, $V(XY)$, $\text{Cov}(X,Y)$, r.

SOLUTIONS

5-1 ☒ **a)** ☒ **b)** ☐ **c)** ☐ **d)** ☐ **e)**

a) vrai : Soit X le nombre de gauchers dans un groupe de 6 personnes prises au hasard. Pour chacune des $n = 6$ personnes, la probabilité qu'il s'agisse d'un gaucher est $p = 0{,}08$ et on peut admettre que les personnes sont indépendantes entre elles pour ce caractère. Dans ce cas, X suit la loi binomiale $\mathcal{B}(6; 0, 08)$.

b) vrai, **c) faux** et **d) faux :** $\mathbb{P}(X = 6) = (0{,}08)^6$.

e) faux :

$$\mathbb{P}(X = 3) = \binom{6}{3}(0{,}08)^3(0{,}92)^3 = 20(0{,}08)^3(0{,}92)^3 \approx 0{,}009\,.$$

5-2 ☒ **a)** ☒ **b)** ❑ **c)** ❑ **d)** ❑ **e)**

a) vrai : Soit X le nombre d'enfants d'une famille de 4 enfants qui ont perdu leur première dent de lait avant 5 ans. dans les hypothèses énoncées, X uit la loi binomiale $\mathcal{B}(4; 0, 10)$

b) vrai, **c) faux** et **d) Faux :** $\mathbb{P}(X = 4) = (0{,}1)^4$.

e) faux :

$$\mathbb{P}(X = 2) = \binom{4}{2}(0{,}1)^2(0{,}9)^2 = 6(0{,}1)^2(0{,}9)^2 = 0{,}0486\,.$$

5-3 ❑ **a)** ❑ **b)** ❑ **c)** ☒ **d)** ❑ **e)**

Soit X la variable aléatoire égale au nombre de patients atteints de M sur un échantillon de taille n. Supposons $n \geq 30$ pour que le tirage des individus puisse être considéré comme avec remise. Dans ce cas, X suit la loi binomiale $\mathcal{B}(n; 0, 02)$. On veut :

$\mathbb{P}(X \geq 1) > 0{,}75$

$\Leftrightarrow 1 - \mathbb{P}(X < 1) > 0{,}75$ en considérant l'événement contraire de $X \geq 1$

$\Leftrightarrow \mathbb{P}(X < 1) < 1 - 0{,}75 \Leftrightarrow \mathbb{P}(X = 0) < 0{,}25$

$\Leftrightarrow (0{,}98)^n < 0{,}25 \Leftrightarrow n \ln 0{,}98 < \ln 0{,}25$ car ln est une fonction croissante

$\Leftrightarrow n > \dfrac{\ln 0{,}25}{\ln 0{,}98}$ il faut inverser l'inégalité car on divise par un nombre négatif

$\Leftrightarrow n \geq 69$. les calculs sont justifiés car on a bien $n \geq 30$

5-4 **a)** Les valeurs possibles pour X (l'univers-image) sont ; {0 ; 20 ; 40 ; 60}.

➤ Pour aboutir à $X = 0$, il faut avoir obtenu 3 fois pile et 3 fois face. Chacun des événements élémentaires (6 lancers successifs de la pièce) a pour probabilité $\left(\dfrac{1}{2}\right)^6$ si la pièce est bien équilibrée. Et il y a $\binom{6}{3} = 20$ cas possibles.

D'où : $P(X = 0) = 20(0{,}5)^6 = 0{,}3125$.

➤ Pour aboutir à $X = 20$, il faut avoir obtenu 4 fois pile et 2 fois face, ou bien 2 fois pile et 4 fois face.

$$\text{D'où : } P(X = 20) = \binom{6}{4}\left(\frac{1}{2}\right)^6 + \binom{6}{2}\left(\frac{1}{2}\right)^6 = 30(0{,}5)^6$$
$$= 0{,}468\,75.$$

➤ Pour aboutir à $X = 40$, il faut avoir obtenu 5 fois pile et 1 fois face, ou bien 1 fois pile et 5 fois face.

$$\text{D'où : } P(X = 40) = \binom{6}{5}\left(\frac{1}{2}\right)^6 + \binom{6}{1}\left(\frac{1}{2}\right)^6 = 12(0{,}5)^6$$
$$= 0{,}1875.$$

➤ Pour aboutir à $X = 60$, il faut avoir obtenu 6 fois pile, ou bien 6 fois face.

D'où : $P(X = 60) = 2(0{,}5)^6 = 0{,}031\,25$.

Vérifiez que $P(X = 0) + P(X = 20) + P(X = 40) + P(X = 60) = 1$.

La distance la plus probable est 20 mètres.

Autre solution possible en remarquant que le nombre Y de lancers donnant pile suit la loi $\mathcal{B}\left(6; \frac{1}{2}\right)$ et que $X = |10Y - 10(6 - Y)| = |20Y - 60|$.

b) $E(X) = (0{,}5)^6(600 + 480 + 120) = 18{,}75$

$E(X^2) = (0{,}5)^6(12\,000 + 19\,200 + 7\,200) = 600$

$V(X) = 600 - (18{,}75)^2 = 248{,}4375$.

Il est possible d'obtenir ces résultats avec votre calculatrice, en statistique à une dimension en rentrant les valeurs et les effectifs associés : (0; 20), (20; 30), (40; 12), (60; 2).

5-5 **a)** Les valeurs possibles pour X sont : $\{0; 1; 2; 3; 4\}$.

➤ L'événement $(X = 0)$ s'écrit aussi « le premier accidenté a une fracture du bassin ». On a :

$$P(X = 0) = \frac{5}{9} = \frac{70}{126} \approx 0{,}5556.$$

➤ L'événement $(X = 1)$ signifie « le premier accidenté a une fracture des membres et le deuxième une fracture du bassin ».

On a : $P(X = 1) = \frac{4}{9} \times \frac{5}{8} = \frac{5}{18} = \frac{35}{126} \approx 0{,}2778$.

- L'événement $(X = 2)$ correspond à un ordre de passage qui commence par MMB. En utilisant à nouveau les probabilités conditionnelles, on a :

$$P(X = 2) = \frac{4}{9} \times \frac{3}{8} \times \frac{5}{7} = \frac{5}{42} = \frac{15}{126} \approx 0{,}1190.$$

- L'événement $(X = 3)$ correspond à un ordre de passage commençant par $MMMB$. On a :

$$P(X = 3) = \frac{4}{9} \times \frac{3}{8} \times \frac{2}{7} \times \frac{5}{6} = \frac{5}{126} \approx 0{,}0397.$$

- L'événement $(X = 4)$ correspond à un ordre de passage commençant par $MMMMB$. On a :

$$P(X = 4) = \frac{4}{9} \times \frac{3}{8} \times \frac{2}{7} \times \frac{1}{6} = \frac{1}{126} \approx 0{,}0079.$$

Les réductions au même dénominateur 126 ont été faites pour vérifier facilement que la somme des probabilités élémentaires est bien égale à 1.

b) $E(X) = 0 \times \frac{70}{126} + 1 \times \frac{35}{126} + \cdots = \frac{84}{126} = \frac{2}{3} \approx 0{,}667$

$$E(X^2) = 0^2 \times \frac{70}{126} + 1^2 \times \frac{35}{126} + \cdots = \frac{156}{126} = \frac{26}{21} \approx 1{,}238$$

$$V(X) = \frac{26}{21} - \left(\frac{2}{3}\right)^2 = \frac{50}{63} \approx 0{,}794$$

5-6 Soit X le nombre d'individus du type A figurant dans l'échantillon. Chacun des $n = 100$ individus a la probabilité $p = 0{,}02$ d'être du type A. Et les tirages des individus peuvent être considérés comme indépendants car la population est très nombreuse.

Dans ce cas, X suit la loi binomiale $\mathcal{B}(100; 0{,}02)$.

Le tirage réel des 100 individus est sans remise. Il a été assimilé à un tirage avec remise à cause de l'hypothèse « population nombreuse ». Cela revient à remplacer une loi hypergéométrique par une loi binomiale car l'effectif total N est grand.

a) $P(X = 0) = (0{,}98)^{100} \approx 0{,}1326.$

b) $P(X \geq 2) = 1 - [P(X = 0) + P(X = 1)]$

$$= 1 - (0{,}98)^{100} - 100 \times 0{,}02 \times (0{,}98)^{99} \approx 0{,}5967.$$

5-7 **1)** La loi suivie par N est la loi binomiale de paramètres $n = 6$ et $p = 0{,}4$ puisque N est la somme de 6 variables aléatoires indépendantes qui suivent toutes la loi de Bernoulli de paramètre $p = 0{,}4$.

2) $\mathbb{P}(N > 0) = 1 - \mathbb{P}(N = 0) = 1 - (0{,}6)^6 \approx 0{,}953.$

$$\mathbb{P}(N = 3) = \binom{6}{3}(0,4)^3(0,6)^3 = 20 \times (0,4)^3 \times (0,6)^3 \approx 0,276.$$

5-8 a) À chacune des $n = 100$ utilisations de la machine, la probabilité de panne est toujours égale à $p = 0,01$.
En supposant de plus que les pannes sont indépendantes entre elles, le nombre total de pannes X suit la loi binomiale $\mathcal{B}(100;0,01)$.

> Dans la réalité industrielle, les hypothèses de l'invariance de la probabilité de panne et de l'indépendance entre les pannes ne sont pas toujours réalisées. Il vous reste alors à étudier les modèles de la fiabilité, ce qui correspond à une orientation différente de celles de ce livre.

$$P(X = 0) = (0,99)^{100} \approx 0,366$$
$$P(X = 1) = 100 \times 0,01 \times (0,99)^{99} \approx 0,370$$
$$P(X = 2) = \frac{100 \times 99}{2} \times (0,01)^2 \times (0,99)^{98} \approx 0,185$$
$$P(X = 3) = \frac{100 \times 99 \times 98}{6} \times (0,01)^3 \times (0,99)^{97} \approx 0,061$$
$$P(X \geq 4) = 1 - [P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3)] \approx 0,018.$$

b) Le coût de chaque réparation étant de 500 €, on a : $Y = 500X$.
Comme X suit la loi binomiale $\mathcal{B}(100;0,01)$, on sait que :
$E(X) = 100 \times 0,01 = 1$ et $V(X) = 100 \times 0,01 \times 0,99 = 0,99$
D'où :
$E(Y) = 500\,E(X) = 500$ € et $\sigma(Y) = 500\,\sigma(X) \approx 497,5$ €.

5-9 a) Probabilité d'ouvrir la première porte au k-ième essai

➤ $k = 1$

La probabilité d'ouvrir au premier essai est $p_1 = \frac{1}{12}$.

➤ $k = 2$

Pour ouvrir au deuxième essai, il faut : ne pas ouvrir au premier (probabilité $\frac{11}{12}$), puis ouvrir au deuxième essai sachant que la bonne clé est parmi les 11 clés restantes (probabilité $\frac{1}{11}$).
D'où : $p_2 = \frac{11}{12} \times \frac{1}{11} = \frac{1}{12}$.

➤ D'une façon générale, pour $1 \leq k \leq 12$, pour ouvrir au k-ième essai, il faut : ne pas ouvrir lors des $k-1$ premiers essais (probabilité $1 - \frac{k-1}{12}$), puis ouvrir au k-ième essai sachant que la bonne clé est parmi les $12-(k-1)$ clés restantes (probabilité $\frac{1}{12-(k-1)}$).

D'où : $p_k = \frac{12-k+1}{12} \times \frac{1}{12-k+1} = \frac{1}{12}$.

b) Loi de *X* nombre total d'essais

Pour k de 1 à 12, notons X_k le nombre d'essais pour ouvrir la k-ième porte.

On a : $X = X_1 + X_2 + \cdots + X_{12}$.

Et il s'agit de variables aléatoires indépendantes car la façon d'ouvrir une porte n'a pas d'influence sur l'ouverture de la porte suivante.

D'après la question précédente, X_1 suit la loi uniforme sur $\{1,\ldots,12\}$

D'où : $E(X_1) = \frac{12+1}{2} = 6{,}5$ et $V(X_1) = \frac{12^2-1}{12} = \frac{143}{12}$.

De même, pour $1 \leq k \leq 12$, X_k suit la loi uniforme sur $\{1,\ldots,12-(k-1)\}$.

D'où : $E(X_k) = \frac{12-(k-1)+1}{2} = \frac{14-k}{2}$

et $V(X_k) = \frac{(12-(k-1))^2-1}{12} = \frac{(13-k)^2-1}{12}$.

D'après les théorèmes sur l'espérance mathématique et la variance de la somme de variables aléatoires indépendantes, on en déduit :

$$E(X) = \sum_{k=1}^{12} E(X_k) = \sum_{k=1}^{12} \frac{14-k}{2} = 12 \times 7 - \frac{1}{2}\sum_{k=1}^{12} k = 45$$

$$V(X) = \sum_{k=1}^{12} V(X_k) = \sum_{k=1}^{12} \frac{(13-k)^2}{12} - \frac{1}{12} \times 12 = \frac{1}{12}\sum_{k=1}^{12} k^2 - 1$$

$$= \frac{319}{6}$$

puis : $\sigma(X) = \sqrt{V(X)} \approx 7{,}29$.

Rappelons que : $\sum_{k=1}^{n} k = \frac{n(n+1)}{2}$ et $\sum_{k=1}^{n} k^2 = \frac{n(n+1)(2n+1)}{6}$

5-10 a) Loi du couple

L'expérience aléatoire est représentée par l'espace probabilisé $(\Omega, \mathcal{P}(\Omega), P)$ où $\Omega = \{0;1;3;5;10\}^2$ et où P est la probabilité uniforme sur Ω (qui comporte 25 éléments).

Chaque événement du type ($X = i$ et $Y = j$) avec $i \in \{0; 1; 3; 5; 10\}$ et $j \in \{0; 1; 3; 5; 10\}$ se ramène à un événement de $\mathcal{P}(\Omega)$. Par exemple :

$$P(X = 10 \text{ et } Y = 10) = P(\{(10; 0), (10; 1), (10; 3), (10; 5), (10; 10)\}) = \frac{5}{25}$$

L'ensemble des résultats déterminant la loi du couple (X, Y) figure dans le tableau ci-dessous :

X \ Y	0	1	3	5	10
0	$\frac{1}{25}$	$\frac{1}{25}$	$\frac{1}{25}$	$\frac{1}{25}$	$\frac{1}{25}$
1	0	$\frac{2}{25}$	$\frac{1}{25}$	$\frac{1}{25}$	$\frac{1}{25}$
3	0	0	$\frac{3}{25}$	$\frac{1}{25}$	$\frac{1}{25}$
5	0	0	0	$\frac{4}{25}$	$\frac{1}{25}$
10	0	0	0	0	$\frac{5}{25}$

b) Lois marginales

Par addition, on obtient les lois marginales :

de X : $\left(0;\frac{1}{5}\right)$, $\left(1;\frac{1}{5}\right)$, $\left(3;\frac{1}{5}\right)$, $\left(5;\frac{1}{5}\right)$, $\left(10;\frac{1}{5}\right)$

de Y : $\left(0;\frac{1}{25}\right)$, $\left(1;\frac{3}{25}\right)$, $\left(3;\frac{5}{25}\right)$, $\left(5;\frac{7}{25}\right)$, $\left(10;\frac{9}{25}\right)$

Comme, par exemple, $P(X = 0 \text{ et } Y = 0) \neq P(X = 0) \times P(Y = 0)$, les variables aléatoires X et Y ne sont pas indépendantes.

La définition mathématique va dans le même sens que l'intuition : X est associée à la première marguerite et Y aux deux marguerites. Il doit donc y avoir un lien entre X et Y.

c) Loi de la somme

Les valeurs possibles pour $Z = X + Y$ sont :

$$\{0; 1; 2; 3; 4; 5; 6; 8; 10; 11; 13; 15; 20\}$$

et les probabilités correspondantes :

$$P(Z=0) = P(X=0 \text{ et } Y=0) = \frac{1}{25}$$

$$P(Z=1) = P(X=0 \text{ et } Y=1) + P(X=1 \text{ et } Y=0) = \frac{1}{25}$$

$$P(Z=2) = P(X=1 \text{ et } Y=1) = \frac{2}{25}$$

$$P(Z=3) = P(X=0 \text{ et } Y=3) + P(X=3 \text{ et } Y=0) = \frac{1}{25}$$

$$P(Z=4) = P(X=1 \text{ et } Y=3) + P(X=3 \text{ et } Y=1) = \frac{1}{25}$$

$$P(Z=5) = P(X=5 \text{ et } Y=0) + P(X=0 \text{ et } Y=5) = \frac{1}{25}$$

$$P(Z=6) = P(X=5 \text{ et } Y=1) + P(X=3 \text{ et } Y=3)$$
$$P(X=1 \text{ et } Y=5) = \frac{4}{25}$$

$$P(Z=8) = P(X=5 \text{ et } Y=3) + P(X=3 \text{ et } Y=5) = \frac{1}{25}$$

$$P(Z=10) = P(X=10 \text{ et } Y=0) + P(X=5 \text{ et } Y=5)$$
$$P(X=0 \text{ et } Y=10) = \frac{5}{25}$$

$$P(Z=11) = P(X=10 \text{ et } Y=1) + P(X=1 \text{ et } Y=10) = \frac{1}{25}$$

$$P(Z=13) = P(X=10 \text{ et } Y=3) + P(X=3 \text{ et } Y=10) = \frac{1}{25}$$

$$P(Z=15) = P(X=10 \text{ et } Y=5) + P(X=5 \text{ et } Y=10) = \frac{1}{25}$$

$$P(Z=20) = P(X=10 \text{ et } Y=10) = \frac{5}{25}$$

d) Loi du produit

Les valeurs possibles pour $T = XY$ sont :

$$\{0; 1; 3; 5; 9; 10; 15; 25; 30; 50; 100\}$$

et les probabilités correspondantes :

$$P(T=0) = P(X=0 \text{ et } Y=0) + P(X=0 \text{ et } Y=0)$$
$$+P(X=0 \text{ et } Y=3) + P(X=0 \text{ et } Y=5) + P(X=0 \text{ et } Y=10)$$

$$+P(X = 1 \text{ et } Y = 0) + P(X = 3 \text{ et } Y = 0) + P(X = 5 \text{ et } Y = 0)$$
$$+P(X = 10 \text{ et } Y = 0) = \frac{5}{25}$$

$$P(T = 1) = P(X = 1 \text{ et } Y = 1) = \frac{2}{25}$$

$$P(T = 3) = P(X = 1 \text{ et } Y = 3) + P(X = 3 \text{ et } Y = 1) = \frac{1}{25}$$

$$P(T = 5) = P(X = 1 \text{ et } Y = 5) + P(X = 5 \text{ et } Y = 1) = \frac{1}{25}$$

$$P(T = 9) = P(X = 3 \text{ et } Y = 3) = \frac{3}{25}$$

$$P(T = 10) = P(X = 1 \text{ et } Y = 10) + P(X = 10 \text{ et } Y = 1) = \frac{1}{25}$$

$$P(T = 15) = P(X = 3 \text{ et } Y = 5) + P(X = 5 \text{ et } Y = 3) = \frac{1}{25}$$

$$P(T = 25) = P(X = 5 \text{ et } Y = 5) = \frac{4}{25}$$

$$P(T = 30) = P(X = 3 \text{ et } Y = 10) + P(X = 10 \text{ et } Y = 3) = \frac{1}{25}$$

$$P(T = 50) = P(X = 5 \text{ et } Y = 10) + P(X = 10 \text{ et } Y = 5) = \frac{1}{25}$$

$$P(T = 100) = P(X = 10 \text{ et } Y = 10) = \frac{5}{25}$$

e) Calculs de paramètres

$$E(X) = \frac{19}{5} = 3{,}8 \; ; E(X^2) = \frac{135}{5} = 27 \; ; V(X) = 12{,}56$$

$$E(Y) = \frac{143}{25} = 5{,}72 \; ; E(Y^2) = \frac{1123}{25} = 44{,}92 \; ; V(Y) = 12{,}2016$$

$$E(X + Y) = \frac{238}{25} = 9{,}52 \; ; V(X + Y) = 40{,}6496$$

On observe que l'on a bien $E(X + Y) = E(X) + E(Y)$, ce qui est un résultat général, mais que $V(X + Y) \neq V(X) + V(Y)$ ce qui confirme que X et Y ne sont pas indépendantes.

$$E(XY) = 29{,}68 \; ; \; V(XY) = 1379{,}2576$$

On observe que $E(XY) \neq E(X)\,E(Y)$, ce qui confirme que X et Y ne sont pas indépendantes.

$$\mathrm{Cov}(X,Y) = E(XY) - E(X)\,E(Y) = 7{,}944$$

$$r = \frac{\mathrm{Cov}(X,Y)}{\sigma(X)\sigma(Y)} \approx 0{,}6417$$

CHAPITRE 6

Variables aléatoires discrètes (cas infini)

PLAN

6.1 Notions sur les séries numériques

6.2 Généralités sur les variables aléatoires discrètes (cas infini)

6.3 Lois classiques

OBJECTIFS

- Acquérir quelques notions sur les séries numériques pour fonder les variables du chapitre
- Comprendre une variable aléatoire dont les valeurs possibles sont du type $\mathbb{N}$
- Savoir reconnaître et utiliser les lois classiques du chapitre

6.1 NOTIONS SUR LES SÉRIES NUMÉRIQUES

Convergence

Soit (u_k) une suite de nombres.

• On dit que la série $\sum u_k$ (ou encore la série de terme général u_k) est **convergente** si la suite (S_n) de terme général :

$$S_n = \sum_{k=0}^{n} u_k = u_0 + u_1 + \cdots + u_n$$

tend vers une limite finie S. On note S la somme de la série :

$$S = \sum_{k=0}^{\infty} u_k = \lim_{n \to +\infty} \left(\sum_{k=0}^{n} u_k \right).$$

S_n est appelée somme partielle d'ordre n.

La différence $R_n = S - S_n = \sum_{k=n+1}^{\infty} u_k$ est le reste d'ordre n. C'est l'erreur commise en remplaçant S par sa valeur approchée S_n.

• Si la série $\sum u_n$ n'est pas convergente, on dit qu'elle est **divergente**.

Convergence absolue

La série $\sum u_k$ est dite **absolument convergente**, si la série $\sum |u_k|$ est convergente.

Si une série est absolument convergente, alors elle est convergente. mais la réciproque est fausse.

Dans le cas d'une série absolument convergente, la somme ne dépend pas de l'ordre des termes. Alors que si une série est convergente sans être absolument convergente, en modifiant l'ordre des termes, on peut obtenir une série qui converge vers n'importe quel réel choisi à l'avance.

Séries classiques

a) Les séries de Riemann

$$\sum_k \frac{1}{k^\alpha} \text{ converge} \Leftrightarrow \alpha > 1.$$

En particulier, la série divergente $\sum \frac{1}{k}$ est appelée **série harmonique**.

b) La série exponentielle

$$\forall x \in \mathbb{R} \qquad e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}$$

c) La série géométrique

$$\text{si } |x| < 1 \qquad \frac{1}{1-x} = \sum_{k=0}^{\infty} x^k$$

et celles que l'on obtient en dérivant, comme :

$$\text{si } |x| < 1 \qquad \frac{1}{(1-x)^2} = \sum_{k=1}^{\infty} k\, x^{k-1}$$

$$\text{si } |x| < 1 \qquad \frac{2}{(1-x)^3} = \sum_{k=2}^{\infty} k(k-1)x^{k-2}$$

6.2 GÉNÉRALITÉS SUR LES VARIABLES ALÉATOIRES DISCRÉTES (CAS INFINI)

Définition

Une variable aléatoire X est dite **discrète** lorsque l'ensemble des valeurs prises par X est dénombrable, c'est-à-dire assimilable à l'ensemble des entiers naturels $\mathbb{N}$ (ou $\mathbb{N}^*$ si on enlève 0).

Dans ce cas, la **distribution de probabilité** (ou **loi de probabilité**) de X correspond à la connaissance des nombres $p_k = P(X = k)$, ces nombres étant soumis aux conditions :

$$\forall k \in \mathbb{N} \quad p_k \geq 0 \quad \text{et} \quad \sum_{k=0}^{\infty} p_k = 1.$$

Par rapport au cas fini, ce qui change, c'est que la somme $\sum_{k=0}^{\infty} p_k$ est une série et non plus une somme comportant un nombre fini de termes. Il peut donc arriver qu'il y ait un problème de convergence.

Espérance mathématique, variance

À condition que les séries écrites soient absolument convergentes, on définit :

- **l'espérance mathématique** de X par :

$$E(X) = \sum_{k=0}^{\infty} k\, p_k$$

- **la variance** de X par :

$$V(X) = E[X - E(X)]^2 = \sum_{k=0}^{\infty} (k - E(x))^2 p_k$$

$$= E(X)^2 - (E(X))^2 = \left(\sum_{k=0}^{\infty} k^2 p_k \right) - (E(X))^2$$

- **l'écart type** par : $\sigma(X) = \sqrt{V(X)}$.

- **L'indépendance** de deux variables aléatoires se définit à partir de la loi du couple, de manière analogue au cas fini. La seule différence, c'est que les sommes qui conduisent aux lois marginales sont des séries.
- **Les théorèmes relatifs aux opérations** sur les variables aléatoires sont les mêmes, aussi bien les théorèmes généraux que les théorèmes vérifiés dans le cas de variables aléatoires indépendantes, dont les principaux sont :

  – on a toujours :

  $$E(X+Y)=E(X)+E(Y)\ ;$$

  – si X et Y sont indépendantes, on a :

  $$V(X+Y)=V(X)+V(Y) \quad \text{et} \quad E(XY)=E(X)\,E(Y).$$

6.3 LOIS CLASSIQUES

Loi géométrique

a) Conditions du modèle

Dans les mêmes hypothèses qui conduisent à la loi binomiale, on obtient la loi géométrique quand la variable aléatoire X désigne le temps d'attente de l'événement A, c'est-à-dire le rang de la première réalisation de A.

b) Loi de probabilité

L'univers-image de X est $\mathbb{N}^*$:

$$\forall k \in \mathbb{N}^* \qquad P(X=k)=pq^{k-1}$$

c) Paramètres

$$E(X)=\frac{1}{p} \quad ; \quad V(X)=\frac{q}{p^2}.$$

Loi de Poisson

a) Conditions du modèle

La loi de Poisson est utilisée pour modéliser le nombre d'apparitions d'un événement rare, par exemple dans la désintégration atomique.

En écologie, on l'utilise pour modéliser la distribution d'une espèce végétale sans intervention humaine (ce qui est à l'opposé d'une plantation de peupliers où on a une loi uniforme, comme le paysage!).

b) Loi de probabilité

X suit la loi de Poisson de paramètre λ (avec $\lambda > 0$), loi notée $\mathcal{P}(\lambda)$, si son univers-image est $\mathbb{N}$ et si :

$$\forall k \in \mathbb{N} \qquad P(X = k) = \mathrm{e}^{-\lambda} \frac{\lambda^k}{k!}$$

c) Paramètres

$$E(X) = \lambda \qquad ; \qquad V(X) = \lambda.$$

d) Somme

Si X suit la loi $\mathcal{P}(\lambda_1)$ et Y la loi $\mathcal{P}(\lambda_2)$, et si X et Y sont indépendantes, alors $X + Y$ suit la loi $\mathcal{P}(\lambda_1 + \lambda_2)$.

Approximation d'une loi binomiale par une loi de Poisson

Théorème. Soit (X_n) une suite de variables aléatoires discrètes telles que, pour tout n, X_n suive la loi binomiale $\mathcal{B}(n, p_n)$ avec $\lim\limits_{n\to\infty} n\,p_n = \lambda$ (avec $\lambda > 0$).
Alors (X_n) converge en loi vers une variable aléatoire discrète X qui suit la loi de Poisson $\mathcal{P}(\lambda)$, ce qui signifie :

$$\lim_{n\to\infty} P(X_n = k) = \mathrm{e}^{-\lambda} \frac{\lambda^k}{k!}.$$

Utilisation pratique

Si n est grand et p assez petit, on peut remplacer la loi binomiale $\mathcal{B}(n, p)$ par la loi de Poisson de même espérance mathématique $\mathcal{P}(np)$.

Dans la pratique, on admet souvent que cette approximation est satisfaisante lorsque $n \geq 30$ et $p \leq 0{,}1$ avec $np \leq 10$.

Mais il ne s'agit que d'une convention, qui peut donc varier selon les auteurs. L'intérêt d'une telle approximation apparaît quand les calculs sont plus simples. Par exemple, avec la loi binomiale $\mathcal{B}(100; 0{,}05)$, on a :

$$P(X = 4) = \binom{100}{4} (0{,}05)^4 (0{,}95)^{96} \approx 0{,}178$$

et avec la loi de Poisson approchée $\mathcal{P}(5)$:

$$P(X = 4) = \mathrm{e}^{-5} \frac{5^4}{4!} \approx 0{,}175.$$

Avec un temps de calcul réduit, on obtient une valeur numérique très proche.

Bizarrerie d'une série semi-convergente

Une série semi-convergente est une série qui converge, sans être absolument convergente. Dans ce cas, la convergence dépend de l'ordre des termes.

Considérons, comme exemple, la série harmonique alternée :

$$S = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \cdots + \frac{1}{2n+1} - \frac{1}{2n+2} + \cdots$$

On peut aussi l'écrire :

$$S = 1 - \frac{1}{2} - \frac{1}{4} + \frac{1}{3} - \frac{1}{6} - \frac{1}{8} + \cdots + \frac{1}{2n+1} - \frac{1}{2(2n+1)} - \frac{1}{2(2n+2)} + \cdots$$

$$= \frac{1}{2} - \frac{1}{4} + \frac{1}{6} - \frac{1}{8} + \cdots + \frac{1}{2(2n+1)} - \frac{1}{2(2n+2)} + \cdots$$

On a divisé sa somme par 2 ! Bizarre !

En fait, pour qu'une série ne dépende pas de l'ordre des termes, il faut, et il suffit, qu'elle soit absolument convergente.

C'est pourquoi, on a eu cette exigence dans la définition de $E(X)$.

MOTS CLÉS

- Série numérique
- Loi géométrique
- Loi de Poisson

EXERCICES

6-1 La probabilité qu'un patient fasse une réaction allergique lors de la prise d'un agent de contraste iodé vaut 1/1200. En considérant que 5000 personnes ont été exposées à ce produit, quelles ont les propositions correctes?

❑ **a)** On peut utiliser une loi de Poisson pour calculer la probabilité d'observer au moins 3 réactions allergiques.

❑ **b)** La variable aléatoire X décrivant la présence ou non d'une réaction allergique chez un patient choisi au hasard parmi les 5000, suit une loi de Bernoulli.

❑ **c)** Une loi normale donnerait une bonne approximation de la probabilité d'observer entre 2 et 5 réactions allergiques.

❑ **d)** La probabilité de n'observer aucune réaction allergique est plus petite que 0,1.

❑ **e)** $\mathbb{P}(X \leq 4) > 0{,}5$.

QCM n° 2 et 3 : *Un groupe médical doit assurer des gardes de nuit. Le nombre d'appels par nuit suit une loi de Poisson de paramètre 2.*

Le groupe décide d'embaucher une jeune médecin remplaçant pour assurer les gardes et le paiera de la façon suivante: pour chaque appel le remplaçant recevra 30 euros. Cependant, pour garantir une rémunération minimale au remplaçant au cas où une nuit il n'y aurait pas d'appel ou un seul appel, celui-ci recevra 50 euros.

On appelle N le nombre d'appels reçus au cour d'une garde et X la rémunération que reçoit le remplaçant pour une nuit. E(X) désigne l'espérance de *X*.

6-2 Cochez la (ou les) proposition(s) exacte(s):

❒ **a)** $\mathbb{P}(N = 0) = 0{,}135$.

❒ **b)** $\mathbb{P}(N = 1) = 0{,}271$.

❒ **c)** $\mathbb{P}(N > 1) = 0{,}594$.

❒ **d)** Si le remplaçant ne recevait pas de rémunération minimale garantie de 50 euros, et donc s'il était payé uniquement à l'acte, la variance de sa rémunération (en euros) serait de 60.

❒ **e)** Si le remplaçant ne recevait pas de rémunération minimale garantie de 50 euros, et donc s'il était payé uniquement à l'acte, la variance de sa rémunération (en euros) serait de 1800.

6-3 X étant la rémunération en euros que reçoit le remplaçant en tenant compte de toutes les règles établies avec le cabinet médical, on a:

❒ **a)** $E(X) = 72{,}17$.

❒ **b)** $E(X) = 60$.

❒ **c)** $E(X) = 12{,}17$.

❒ **d)** $\text{Var}(X) = 1962{,}6$.

❒ **e)** $\text{Var}(X) = 962{,}6$.

6-4 **a)** Déterminez les réels a et b tels que

$$\forall k \in \mathbb{N}^* \qquad \frac{1}{k(k+1)} = \frac{a}{k} + \frac{b}{k+1}$$

b) Montrez qu'en posant, pour tout $k \in \mathbb{N}^*$, $P(X = k) = \dfrac{\alpha}{k(k+1)}$ on peut définir une distribution de probabilité sur $\mathbb{N}^*$ en choisissant bien α.

c) Dans ce cas, déterminez, si elles existent, $E(X)$ et $V(X)$.

6-5 Le nombre mensuel X d'apparition d'un événement rare suit une loi de Poisson. La probabilité d'observer 2 cas en un mois est de 0,201; celle d'observer 3 cas est de 0,074.
Estimez le nombre moyen de cas pour un mois.

6-6 Soit X et Y deux variables aléatoires indépendantes. X suit la loi de Poisson de paramètre λ_1, Y suit la loi de Poisson de paramètre λ_2.
Étudiez la loi de probabilité de la variable aléatoire $Z = X + Y$.

6-7 Un bureau de réservation reçoit, entre 10 h et 12 h, en moyenne, 1,2 appels téléphoniques par minute. On modélise ce phénomène par une variable aléatoire de Poisson. Déterminez :

a) la probabilité pour qu'entre 11 h et 11 h 01 on ait :

1) aucun appel ;

2) un appel ;

3) deux appels ;

b) la probabilité de recevoir 4 appels entre 11 h et 11 h 02.

6-8 Un liquide contient 10^5 bactéries par litre, réparties au hasard.

On en prélève 1 mm^3.

a) Quelle est la probabilité que ce prélèvement ne contienne aucune bactérie ?

b) Quelle est la probabilité qu'il contienne au moins 3 bactéries ?

6-9 Dans une station de ski, on peut se rendre aux départs respectifs des pistes A et B par deux remontées mécaniques qui partent du même point D de la station.

Le nombre de skieurs qui se présentent en D pendant une heure est une variable aléatoire N qui suit une loi de Poisson de paramètre λ.

On admet d'autre part qu'on a atteint un régime stable tel que chacun des skieurs choisit, indépendamment des précédents, A ou B avec des probabilités fixes p et $q = 1 - p$.

On note X la variable aléatoire : nombre de skieurs qui choisissent A pendant une heure.

a) Déterminez la loi conjointe du couple (X, N) en calculant pour k et n entiers : $P(X = k \text{ et } N = n)$.

b) Déterminez la loi marginale de X en calculant, pour tout k entier, $P(X = k)$. De quelle loi s'agit-il ?

c) Calculez le nombre moyen de skieurs se présentant pendant une heure au départ de la piste A.

SOLUTIONS

6-1 ☒ **a)** ☐ **b)** ☐ **c)** ☒ **d)** ☒ **e)**

- Soit X le nombre de patients présentant une réaction allergique dans un échantillon de 5000 personnes. X suit la loi binomiale $\mathcal{B}(5000; 1/1200)$.
- Les valeurs de n et de p permettent d'approximer cette loi par la loi de Poisson de paramètre $\lambda = np = 5000 \times \dfrac{1}{1200} = \dfrac{25}{6}$.
- Les valeurs de n et de p ne permettent d'approximer par une loi normale.
- $\mathbb{P}(X = 0) = \mathrm{e}^{-\lambda} \approx 0{,}016$.
- $\mathbb{P}(X = 1) = \mathrm{e}^{-\lambda}\lambda \approx 0{,}065$; $\mathbb{P}(X = 2) = \mathrm{e}^{-\lambda}\dfrac{\lambda^2}{2} \approx 0{,}135$.
- $\mathbb{P}(X = 3) = \mathrm{e}^{-\lambda}\dfrac{\lambda^3}{6} \approx 0{,}187$. ; $\mathbb{P}(X = 4) = \mathrm{e}^{-\lambda}\dfrac{\lambda^4}{24} \approx 0{,}195$.

On a donc $\mathbb{P}(X \leq 4) \approx 0{,}598$.

6-2 ☒ **a)** ☒ **b)** ☒ **c)** ☐ **d)** ☒ **e)**

- N suit la loi de Poisson $\mathcal{P}(2)$. On a donc:

$\mathbb{P}(N = 0) = \mathrm{e}^{-2} \approx 0{,}135$; $\mathbb{P}(N = 1) = \mathrm{e}^{-2}2 \approx 0.271$

$\mathbb{P}(N > 1) = 1 - [\mathbb{P}(N = 0) + \mathbb{P}(N = 1)] \approx 0{,}594$.

• Sans rémunération minimale garantie, le remplaçant recevrait $30N$ et on a $\text{Var}(30N) = 30^2\,\text{Var}\,(N) = 900 \times 2 = 1800$.

6-3 ☒ **a)** ❑ **b)** ❑ **c)** ❒ **d)** ☒ **e)**

• N est définie par:

$$\begin{cases} \text{si } N = 0 \text{ ou } 1 & X = 50 \\ \text{si } N \geq 2 & X = 30N \end{cases}$$

• En notant $p_k = \mathbb{P}(N = k)$, on a:

$$E(X) = 50 \times 0{,}406 + 30\left(\sum_{k=2}^{+\infty} kp_k\right)$$

D'autre part, de : $2 = E(N) = \sum_{k=0}^{+\infty} kp_k = p_1 + \sum_{k=2}^{+\infty} kp_k$

on déduit: $\sum_{k=2}^{+\infty} kp_k = 1{,}729$ puis $E(X) = 72{,}17$.

• De la même manière $E(X^2) = 50^2 \times 0{,}406 + 30^2\left(\sum_{k=2}^{+\infty} k^2 p_k\right)$

et $2 = \text{Var}(N) = \sum_{k=0}^{+\infty} k^2 p_k - (E(N))^2 = p_1 + \sum_{k=2}^{+\infty} k^2 p_k - 4$

d'où $\sum_{k=2}^{+\infty} k^2 p_k = 5{,}729$ puis $E(X^2) = 6171{,}1$

On en déduit $\text{Var}\,(X) = E(X^2) - (E(X))^2 = 962{,}6$

6-4 **a)** En réduisant au même dénominateur et en utilisant l'unicité de l'écriture d'un polynôme, ou par des méthodes plus rapides, on obtient :

$$\forall k \in \mathbb{N}^* \qquad \frac{1}{k(k+1)} = \frac{1}{k} - \frac{1}{k+1}$$

b) Il faut $\alpha > 0$ et que la série des p_k converge vers 1.
Écrivons une somme partielle et transformons la, avec la question précédente et en renumérotant les indices de sommation :

$$\frac{1}{\alpha} S_n = \sum_{k=1}^{n} \frac{1}{k} - \sum_{k=1}^{n} \frac{1}{k+1} = \sum_{k=1}^{n} \frac{1}{k} - \sum_{k=2}^{n+1} \frac{1}{k}$$

Les valeurs de k toutes présentes s'éliminent et il reste après simplification : $\frac{1}{\alpha} S_n = 1 - \frac{1}{n+1}$ dont la limite est 1 quand n tend vers l'infini.

En choisissant $\alpha = 1$, on a donc une distribution de probabilité sur $\mathbb{N}^*$.

c) $E(X) = \sum_{k=1}^{\infty} k\, p_k = \sum_{k=1}^{\infty} \frac{1}{k+1}$

Comme $k\, p_k$ est équivalent à $\frac{1}{k}$ terme général d'une série divergente, X n'a pas d'espérance mathématique.

6-5 Si X suit la loi de Poisson de paramètre λ, avec les informations fournies, on a :

$$P(X = 2) = e^{-\lambda} \frac{\lambda^2}{2} = 0{,}201 \text{ et } P(X = 3) = e^{-\lambda} \frac{\lambda^3}{6} = 0{,}074.$$

On en déduit :

$$\frac{P(X = 3)}{P(X = 2)} = \frac{\lambda}{3} = \frac{0{,}074}{0{,}201} \text{ d'où } \lambda \approx 1{,}1.$$

Comme $E(X) = \lambda$ pour une loi de Poisson, il y a donc en moyenne environ 1,1 cas par mois.

6-6 Si Z suit une loi de Poisson (ce qui n'est pas sûr), son paramètre sera nécessairement $\lambda_1 + \lambda_2$ puisque, dans le cas d'une loi de Poisson, l'espérance mathématique est égale au paramètre et que les espérances mathématiques s'ajoutent toujours.

Les valeurs possibles pour X et Y étant $\mathbb{N}$, il est en de même pour Z.

Soit k un élément quelconque de $\mathbb{N}$. On veut calculer $P(Z = k)$.

L'événement $X + Y = k$ se décompose comme réunion des événements deux à deux incompatibles :

$$(X = 0 \text{ et } Y = k),\ (X = 1 \text{ et } Y = k - 1),\ \ldots\ (X = k \text{ et } Y = 0).$$

D'où :

$$P(Z = k) = \sum_{i=0}^{k} P(X = i \text{ et } Y = k - i).$$

X et Y étant des variables aléatoires indépendants, on a :

$$P(X = i \text{ et } Y = k - i) = P(X = i) \times P(Y = k - i).$$

D'autre part, X et Y suivant des lois de Poisson :

$$P(X = i) = e^{-\lambda_1} \frac{\lambda_1^i}{i!} \quad ; \quad P(Y = k - i) = e^{-\lambda_2} \frac{\lambda_2^{k-i}}{(k - i)!}$$

D'où :

$$P(Z = k) = \sum_{i=0}^{k} e^{-\lambda_1} \frac{\lambda_1^i}{i!} e^{-\lambda_2} \frac{\lambda_2^{k-i}}{(k-i)!}$$

$$= e^{-(\lambda_1+\lambda_2)} \sum_{i=0}^{k} \frac{1}{i!(k-i)!} \lambda_1^i \lambda_2^{k-i}$$

$$= \frac{e^{-(\lambda_1+\lambda_2)}}{k!} \sum_{i=0}^{k} \binom{k}{i} \lambda_1^i \lambda_2^{k-i}$$

$$= \frac{e^{-(\lambda_1+\lambda_2)}}{k!} (\lambda_1 + \lambda_2)^k \text{ d'après la formule du binôme.}$$

On observe donc que Z suit la loi de Poisson de paramètre $\lambda_1 + \lambda_2$, mais il a fallu l'hypothèse d'indépendance entre X et Y.

6-7 **a)** Soit X le nombre d'appels téléphoniques reçus pendant 1 minute. X suit une loi de Poisson de paramètre λ.

On sait que, pour une loi de Poisson, $E(X) = \lambda$. Comme l'information fournie peut s'écrire $E(X) = 1{,}2$, on en déduit que $\lambda = 1{,}2$.

1) $P(X = 0) = e^{-1,2} \approx 0{,}301$.

2) $P(X = 1) = e^{-1,2} \times 1{,}2 \approx 0{,}361$.

3) $P(X = 2) = e^{-1,2} \times \frac{(1,2)^2}{2} \approx 0{,}217$.

b) Notons X le nombre d'appels reçus entre 11 h et 11 h 01, Y le nombre d'appels reçus entre 11 h 01 et 11 h 02, Z le nombre d'appels reçus entre 11 h et 11 h 02.

On a $Z = X + Y$ et on demande $P(Z = 4)$.

X et Y suivent la loi de Poisson de paramètre $\lambda = 1{,}2$. D'autre part, on va les supposer indépendantes, c'est-à-dire que le nombre d'appels reçus entre 11 h et 11 h 01 n'a pas d'influence sur le nombre d'appels reçus entre 11 h 01 et 11 h 02, ce qui suppose que le central n'a pas été saturé.

Dans ce cas, on sait que (*cf.* exercice précédent), Z suit la loi de Poisson de paramètre $\lambda + \lambda = 2{,}4$.

D'où : $P(Z = 4) = e^{-2,4} \times \frac{(2,4)^4}{24} \approx 0{,}125$.

6-8 L'exercice suppose les bactéries réparties au hasard dans le liquide. C'est le comportement de la majorité des bactéries, mais pas de toutes. Par exemple, certaines sont attirées par les parois et dans ce cas là, l'exercice ne s'applique plus.

Dans un litre de liquide, chacune des 10^5 bactéries peut :

– être présente dans le mm^3 prélevé, avec une probabilité $p = 10^{-6}$ car il y a 10^6 mm^3 dans un litre et la répartition est supposée équiprobable ;

– être absente du mm^3 prélevé, avec une probabilité $q = 1 - p$.

Soit X le nombre de bactéries présentes dans le mm^3 prélevé.

X suit la loi binomiale $\mathcal{B}(10^5; 10^{-6})$ et les calculs directs sont possibles (et pénibles).

Avec $n = 10^5$ et $p = 10^{-6}$, nous pouvons aussi approximer la loi binomiale par la loi de Poisson de paramètre $\lambda = E(X) = np = 0{,}1$, ce qui conduit à des calculs beaucoup plus faciles.

a) On demande $P(X = 0)$.

➤ *Calcul direct*

$P(X = 0) = (1 - 10^{-6})^{10^5} \approx 0{,}904\ 837\ 463$

➤ *Approximation par la loi de Poisson*

$P(X = 0) = \mathrm{e}^{-0{,}1} \approx 0{,}904\ 837\ 418.$

b) On demande

$P(X \geq 3) = 1 - [P(X = 0) + P(X = 1) + P(X = 2)].$

➤ *Calcul direct*

$$P(X = 1) = 10^5 \times 10^{-6}(1 - 10^6)^{10^5 - 1} \approx 0{,}090\ 483\ 837$$

$$P(X = 2) = \frac{10^5 \times (10^5 - 1)}{2}(10^{-6})^2(1 - 10^{-6})^{10^5 - 2} \approx 0{,}004\ 524\ 151$$

D'où : $P(X \geq 3) \approx 0{,}000\ 154\ 549.$

➤ *Approximation par la loi de Poisson*

$$P(X \geq 3) = 1 - \mathrm{e}^{-0{,}1}\left(1 + 0{,}1 + \frac{(0{,}1)^2}{2}\right) \approx 0{,}000\ 154\ 653.$$

Pour chacune des deux questions, la précision retenue est uniquement destinée à montrer la qualité de l'approximation par la loi de Poisson. Elle serait évidemment illusoire en situation expérimentale.

6-9 a) Soit k et n des entiers naturels. Pour que l'événement ($X = k$ et $N = n$) ne soit pas impossible, il est nécessaire que $k \leq n$.
Dans ce cas, on peut écrire :

$P(X = k \text{ et } N = n) = P(N = n) \times P(X = k/N = n).$

N suivant une loi de Poisson de paramètre λ, on a : $P(N = n) = \mathrm{e}^{-\lambda}\dfrac{\lambda^n}{n!}$.

Si on sait que $N = n$, comme chacun des skieurs qui se présentent en D a une probabilité p de choisir A et qu'il y a indépendance entre les skieurs, on a une loi binomiale, soit :

$P(X = k \ / \ N = n) = \binom{n}{k} p^k q^{n-k}$. Donc :

$$P(X = k \text{ et } N = n) = \mathrm{e}^{-\lambda} \frac{\lambda^n}{n!} \frac{n!}{k!(n-k)!} p^k q^{n-k} \text{ avec } k \leq n.$$

b) Soit k entier naturel, et cherchons $P(X = k)$. Dans ce cas, n peut prendre toutes les valeurs entières telles que $n \geq k$. Donc :

$$\begin{aligned} P(X = k) &= \sum_{n=k}^{\infty} P(X = k \text{ et } N = n) = \frac{\mathrm{e}^{-\lambda}}{k!}(\lambda p)^k \sum_{n=k}^{\infty} \frac{(\lambda q)^{n-k}}{(n-k)!} \\ &= \frac{\mathrm{e}^{-\lambda}}{k!}(\lambda p)^k \sum_{n'=0}^{\infty} \frac{(\lambda q)^{n'}}{n'!} \text{ en posant } n' = n - k \\ &= \frac{\mathrm{e}^{-\lambda}}{k!}(\lambda p)^k \mathrm{e}^{\lambda q} \\ &= \mathrm{e}^{-\lambda p} \frac{(\lambda p)^k}{k!} \end{aligned}$$

La loi de X est donc la loi de Poisson de paramètre λ_p.

c) Le nombre moyen de skieurs qui se présentent pendant une heure au départ de la piste A est $E(X)$.
Comme X suit une loi de Poisson de paramètre λ_p, on a : $E(X) = \lambda_p$.

CHAPITRE 7

Variables aléatoires continues

PLAN

7.1 Notions sur les intégrales généralisées
7.2 Généralités sur les variables aléatoires continues
7.3 Lois classiques

OBJECTIFS

- Acquérir quelques notions sur les intégrales généralisées pour fonder les variables du chapitre
- Comprendre une variable dont les valeurs possibles sont du type $\mathbb{R}$
- Savoir reconnaître et utiliser les lois classiques du chapitre

7.1 NOTIONS SUR LES INTÉGRALES GÉNÉRALISÉES

- Soit f une fonction définie sur $[a, +\infty[$ et intégrable sur tout segment $[a, x]$. On dit que f est d'intégrale convergente sur $[a, +\infty[$, ou que l'intégrale $\int_a^{+\infty} f(t)\,\mathrm{d}t$ **converge**, ou existe, si la fonction :

$$x \mapsto \int_a^x f(t)\,\mathrm{d}t$$

possède une limite finie lorsque x tend vers $+\infty$. On note alors :

$$\lim_{x\to\infty} \int_a^x f(t)\,\mathrm{d}t = \int_a^{+\infty} f(t)\,\mathrm{d}t.$$

Dans le cas contraire, on dit que l'intégrale **diverge**.

- On définit de manière analogue : $\int_{-\infty}^a f(t)\,\mathrm{d}t = \lim_{x\to-\infty} \int_x^a f(t)\,\mathrm{d}t$ puis, avec a quelconque :

$$\int_{-\infty}^{+\infty} f(t)\,\mathrm{d}t = \int_{-\infty}^a f(t)\,\mathrm{d}t + \int_a^{+\infty} f(t)\,\mathrm{d}t.$$

7.2 GÉNÉRALITÉS SUR LES VARIABLES ALÉATOIRES CONTINUES

Si Ω n'est pas dénombrable, il n'est plus possible de choisir $\mathcal{T} = \mathcal{P}(\Omega)$. La tribu retenue se construit à partir des intervalles. Pour définir la distribution de probabilité d'une telle variable aléatoire X, il faut connaître la probabilité des événements :

$$X \leq a \; ; \; a \leq X \leq b \; ; \; X = a.$$

Définition

Soit X une variable aléatoire et F_X sa fonction de répartition, c'est-à-dire $F_X(x) = P(X \leq x)$. On dit que X est une variable continue s'il existe une fonction f de $\mathbb{R}$ dans $\mathbb{R}$, dite **densité de probabilité** de X, telle que :

(1) $\forall x \in \mathbb{R} \quad f(x) \geq 0$;

(2) f est continue sur $\mathbb{R}$ sauf peut-être en un nombre fini de points où elle admet une limite à gauche et une limite à droite finies ;

(3) $\int_{-\infty}^{+\infty} f(t)\,\mathrm{d}t$ existe et vaut 1 ;

(4) F_X est liée à f par : $F_X(x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t$ pour tout x.

Propriété

Si $a < b$, on a :

$$P(a \leq X \leq b) = F_X(b) - F_X(a) = \int_a^b f(t)\mathrm{d}t.$$

On en déduit que $P(X = a) = 0$ pour tout a.

On a donc un événement qui n'est pas impossible et dont la probabilité est nulle. Si cela vous trouble beaucoup, pensez qu'un point n'est pas vide et qu'il a pourtant une longueur nulle.

Espérance mathématique, variance

Soit X une variable aléatoire continue dont f est une densité. À condition que les intégrales convergent, on définit :

➤ **l'espérance mathématique** de X par :

$$E(X) = \int_{-\infty}^{+\infty} tf(t)\mathrm{d}t$$

➤ la **variance** de X par :

$$V(X) = E[X - E(X)]^2 = \int_{-\infty}^{+\infty} (t - E(X))^2 f(t)\,\mathrm{d}t$$
$$= E(X^2) - (E(X))^2 = \left(\int_{-\infty}^{+\infty} t^2 f(t)\,\mathrm{d}t\right) - (E(X))^2$$

➤ l'**écart type** par : $\sigma(X) = \sqrt{V(X)}$.

L'indépendance de deux variables aléatoires se définit là encore à partir de la loi du couple.

7.3 LOIS CLASSIQUES

Loi uniforme sur [*a*, *b*]

a) Densité

X suit la loi uniforme sur le segment $[a,b]$, notée $\mathcal{U}[a,b]$, si elle admet pour densité de probabilité la fonction f définie par :

$$\begin{cases} f(x) = 0 & \text{si } x \notin [a,b] \\ f(x) = \dfrac{1}{b-a} & \text{si } x \in [a,b] \end{cases}$$

b) Paramètres

$$E(X) = \frac{a+b}{2} \quad ; \quad V(X) = \frac{(b-a)^2}{12}.$$

En statistique descriptive, quand les données sont groupées en classes, le réflexe qui consiste à remplacer chaque intervalle par son milieu est fondé sur l'hypothèse de répartition uniforme. Comme $E(X)$ est le milieu de l'intervalle, la moyenne est inchangée. Mais la variance est modifiée. On peut alors utiliser la correction de Sheppard (retrancher $\frac{h^2}{12}$ où h est l'amplitude des classes) fondée sur le calcul de $V(X)$.

Loi exponentielle

a) Densité

X suit la loi exponentielle de paramètre $\lambda > 0$, notée $\mathcal{E}(\lambda)$, si elle admet pour densité de probabilité la fonction f définie par :

$$\begin{cases} f(x) = 0 & \text{si } x < 0 \\ f(x) = \lambda \mathrm{e}^{-\lambda x} & \text{si } x \geq 0 \end{cases}$$

b) Paramètres

$$E(X) = \frac{1}{\lambda} \quad ; \quad V(X) = \frac{1}{\lambda^2}.$$

La loi exponentielle est utilisée, par exemple, pour modéliser la durée de vie d'un appareil qui fonctionne sans usure, les seules causes de panne étant externes.

Loi normale, ou loi de Gauss, ou loi de Laplace-Gauss

a) Densité

X suit la loi de Gauss de paramètres μ et σ, notée $\mathcal{N}(\mu,\sigma)$, si elle admet pour densité de probabilité la fonction f définie par :

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

b) Paramètres

$$E(X) = \mu \quad ; \quad V(X) = \sigma^2.$$

c) Loi normale centrée réduite

La variable centrée réduite $Z = \dfrac{X-\mu}{\sigma}$ suit la loi $\mathcal{N}(0,1)$.

Tout problème relatif à X se ramène à Z et on dispose (support papier ou électronique) de plusieurs tables concernant Z.

- La **table de la fonction de répartition** (*cf.* annexe table 1) notée dans ce cas particulier $\Phi(x)$ (ou $\Pi(x)$) pour la distinguer de la notation générale $F(x)$.
  Un réel x étant donné (arrondi à 10^{-2} sur support papier), la table donne $\Phi(x)$ pour $x \geq 0$.
  Pour $x < 0$, on a :
  $$\Phi(-x) = 1 - \Phi(x).$$

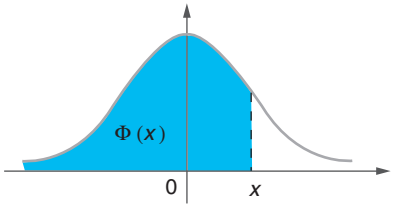


Figure 7-1

- La **table des écarts réduits** (*cf.* annexe table 2)

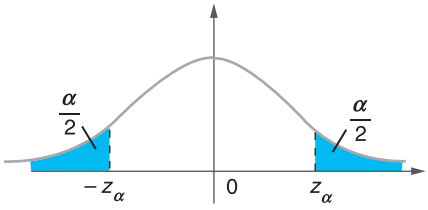


Figure 7-2

Une probabilité α étant donnée (arrondie à 10^{-2} sur support papier), la table donne la valeur $z_\alpha > 0$ telle que

$$P(|Z| \geq z_\alpha) = \alpha.$$

d) Somme

Si X suit la loi $\mathcal{N}(\mu_1, \sigma_1)$ et Y la loi $\mathcal{N}(\mu_2, \sigma_2)$, et si X et Y sont indépendantes, alors $X + Y$ suit $\mathcal{N}\left(\mu_1 + \mu_2, \sqrt{\sigma_1^2 + \sigma_2^2}\right)$.

Approximation d'une loi binomiale par une loi normale

Théorème. Soit X une variable aléatoire qui suit la loi binomiale $\mathcal{B}(n, p)$. Pour n assez grand et p pas trop voisin de 0 et de 1, X suit à peu près la loi normale $\mathcal{N}(np, \sqrt{npq})$ de même espérance mathématique et de même écart type.

En pratique on utilise souvent cette approximation lorsque $n \geq 30$, $np \geq 5$ et $nq \geq 5$. Mais d'autres conventions existent.

Correction de continuité

Si k_1 et k_2 sont deux entiers compris entre 0 et n, les intervalles $]k_1, k_2[$ et $[k_1, k_2]$ n'ont pas la même probabilité pour la loi binomiale, alors qu'ils ont la même probabilité pour la loi normale. Cela est dû au fait qu'on approche une loi discrète par une loi continue.

On peut corriger cette différence en remplaçant $]k_1, k_2[$ par $]k_1 + 0{,}5 ; k_2 - 0{,}5]$ et en remplaçant $[k_1, k_2]$ par $[k_1 - 0{,}5 ; k_2 + 0{,}5]$.

Pour rassurer les non-matheux

On ne fera du calcul intégral que dans des cas très simples. La plupart du temps, représentez-vous graphiquement la situation.

Une densité est une fonction positive, avec parfois quelques sauts verticaux, et dont la surface entre l'axe des abscisses et la courbe a un sens et vaut 1.

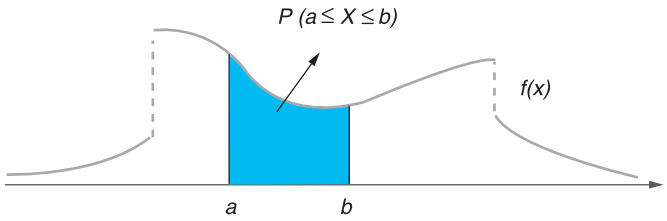


Figure 7-3

La probabilité d'un intervalle $[a,b]$ est alors visualisée par la surface appuyée sur $[a,b]$ et limitée par la courbe de f.

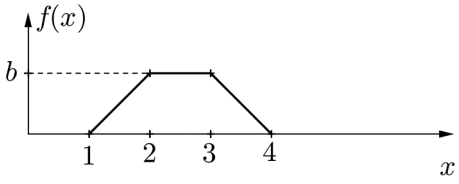
Dans la plupart des cas pratiques, il vous restera à apprendre à lire dans des tables toutes prêtes … sauf si vous avez le droit d'utiliser un ordinateur.

MOTS-CLÉS

- Intégrale généralisée
- Densité de probabilité
- Loi uniforme
- Loi exponentielle
- Loi normale

EXERCICES

7-1 Soit une fonction *fx*) nulle en dehors de l'intervalle [1; 4], augmentant linéairement dans l'intervalle |1; 2], puis égale à une constante b dans l'intervalle [2; 3], puis décroissant linéairement dans l'intervalle [3; 4].



Quelle doit être la valeur de b pour que cette fonction soit une densité de probabilité?

❑ **a)** $b = 0,2$. ❑ **b)** $b = 0,5$. ❑ **c)** $b = 1$. ❑ **d)** $b = 2$.
❑ **e)** Toutes les propositions précédentes sont inexactes.

7-2 Concernant la loi de Gauss centrée réduite :

❑ **a)** $\mathbb{P}(Z > 1,5) = 2$.

❑ **b)** $\mathbb{P}(0 < Z < 1) = 0,3413$.

❑ **c)** $\mathbb{P}(Z < 1,96) = 0,867$.

❑ **d)** $\mathbb{P}(Z < -1,96) < 0,03$.

❑ **e)** $\mathbb{P}(Z < 0) = 0,4$.

7-3 La cholestérolémie (en g/L) est distribuée selon une loi normale (de Gauss). On sait que 2, 5 % des hommes de plus de 60 ans ont moins de 1, 40 g/L et que 2, 5 % ont plus de 2, 60 g/L.

❑ **a)** La moyenne de la cholestérolémie est 2, 0 g/L.

❑ **b)** L'écart-type est de 0, 3 g/L.

❑ **c)** La variance est 0, 09 g/L.

❑ **d)** Il y a 16 chances sur 100 pour que la cholestérolémie d'un sujet soit supérieure à 2, 3 g/L.

❑ **e)** Il y a 32 chances sur 100 pour que la cholestérolémie d'un sujet soit supérieure à 2, 3 g/L.

7-4 Soit f la fonction définie sur $\mathbb{R}$ par :

$$f(x) = ax\,(1-x) \text{ si } x \in [0;1] \quad ; \quad f(x) = 0 \text{ si } x \notin [0;1].$$

a) Pour quelle valeur de a, f est-elle une densité de probabilité ?

b) Calculez alors $E(X)$ et $V(X)$ pour une variable aléatoire X admettant cette densité.

7-5 Pour un certain type d'ampoules électriques, la durée de vie en heures d'une ampoule est une variable aléatoire dont la loi de probabilité admet une densité de probabilité f définie par :

$$f(t) = 0 \text{ si } t < 0 \quad ; \quad f(t) = at\mathrm{e}^{-\lambda t} \quad \text{si } t \geq 0.$$

où a et λ sont des constantes strictement positives.
Sachant que la durée de vie moyenne de ces ampoules est de 1 000 heures, déterminez la valeur des constantes a et λ.

7-6 *Exercices de lecture des tables de la loi normale centrée réduite*

a) Si X suit la loi $\mathcal{N}(4; 2)$, déterminez $P(X \leq 6)$.

b) Si X suit la loi $\mathcal{N}(3; 1,5)$, déterminez y pour que $P(X \leq y) = 0,4218$.

c) Si X suit la loi $\mathcal{N}(5; 2)$, déterminez $P(2{,}5 \leq X \leq 6{,}5)$.

d) Si X suit la loi $\mathcal{N}(6; 2)$, déterminez un intervalle, centré sur la moyenne, de probabilité 0,9.

7-7 Dans une population de veaux, la masse d'un animal pris au hasard est une variable aléatoire X qui suit une loi normale d'espérance mathématique 500 kg et d'écart type 40 kg. On prélève un échantillon de 80 veaux.

a) Combien de veaux pèsent plus de 560 kg ?

b) Combien de veaux pèsent moins de 480 kg ?

c) Combien de veaux ont une masse comprise entre 450 et 550 kg ?

d) On sélectionne pour la reproduction les 15% supérieurs de l'échantillon. À partir de quelle masse un animal sera-t-il sélectionné ?

7-8 On suppose, dans cet exercice, que toutes les durées de trajet suivent des lois normales.

a) Une directrice quitte son domicile à 8 h 45 pour aller à son bureau qui ouvre à 9 h. Quelle est la probabilité pour qu'elle arrive en retard sachant que la durée moyenne du trajet est de 13 min avec un écart type égal à 3 min ?

b) Le secrétaire se rend au même bureau en utilisant le train puis l'autobus. Le train part à 8 h 32, le trajet durant en moyenne 16 min avec un écart type de 2 min. L'autobus part à 8 h 50 (sans attendre l'arrivée du train), le trajet durant en moyenne 9 min avec un écart type de 1 min. Quelle est la probabilité pour que le secrétaire arrive à l'heure ?

c) Quelle est la probabilité pour que la directrice ou le secrétaire arrive à l'heure ?

7-9 Dans une population homogène de 20 000 habitants, la probabilité pour qu'une personne quelconque demande à être vaccinée contre la grippe est de 0,4.
De combien de vaccins doit-on disposer pour que la probabilité qu'on vienne à en manquer soit inférieure à 0,1 ?

7-10 Dans un certain type de graine, la probabilité de germination est $p = 0{,}8$. Une personne sème 400 graines. Calculez la probabilité pour que 300, au moins, germent.

7-11 La longueur des tiges de chrysanthèmes en fleurs coupées intervient dans le classement par catégorie. Pour simplifier, on supposera par la suite que cette longueur sera le seul critère de classement. Un chrysanthème sera classé en catégorie extra si la longueur de sa tige est supérieure ou égale à 80 cm.
Au 1er décembre, on évalue la production d'une certaine serre à 6 000 chrysanthèmes pour le mois. À cette époque, les chrysanthèmes classés

en catégorie extra sont payé au producteur 10 € les dix, et les autres 6 € les dix seulement.
La qualité de la production ayant été étudiée sur un échantillon de 100 tiges coupées de chrysanthèmes, on en conclut que la longueur des tiges coupées est une variable aléatoire qui suit une loi normale de moyenne 92 cm et d'écart type 8 cm.

a) Quelle est la probabilité pour qu'une fleur soit classée en catégorie extra ?

b) Quelle est l'espérance mathématique du nombre de fleurs qui seront classées en catégorie extra sur les 6 000 fleurs de la production de décembre ?

c) Déduisez-en l'espérance mathématique de la recette pour le total de la production de la serre pendant ce mois.

7-12 Albert et Bernard décident de faire n parties de pile ou face, avec un enjeu de 1 € par partie.
Chacun d'eux dispose de la somme de 20 €. Le règlement aura lieu à la fin de la n-ième partie.

a) Soit X le nombre de parties que gagnera Albert. À quelle double inégalité doit satisfaire X pour que le règlement puisse s'effectuer sans dette de l'un ou l'autre joueur ?

b) Déterminez une valeur de n pour que la probabilité d'un règlement sans dette soit au moins égale à 0,68.

SOLUTIONS

7-1 ❑ **a)** ☒ **b)** ❑ **c)** ❑ **d)** ❑ **e)**
La fonction f est positive, continue. Il reste à choisir b pour que la surface au-dessus de l'axe des abscisses soit égale à 1. Cette surface se ramène à deux carrés d'aire b. Il faut choisir $b = 0,5$.

7-2 ❑ **a)** ☒ **b)** ❑ **c)** ☒ **d)** ❑ **e)**
Il s'agit de lectures dans la table de la fonction de répartition de $\mathcal{N}(0 ; 1)$.
$\mathbb{P}(Z > 1,5) = 1 - \Phi(1,5) = 1 - 0,9332 = 0,0668$.
$\mathbb{P}(0 < Z < 1) = \Phi(1) - \Phi(0) = 0,8413 - 0,5 = 0,3413$.
$\mathbb{P}(Z < 1,96) = \Phi(1,96) = 0,9750$.
$\mathbb{P}(Z < -1,96) = \Phi(-1,96) = 1 - \Phi(1,96) = 1 - 0,9750 = 0,0250$.
$\mathbb{P}(Z < 0) = 0,5$.

7-3 ☒ **a)** ☒ **b)** ❑ **c)** ☒ **d)** ❑ **e)**
• La cholestérolémie X(en g/L) des hommes de plus 60 ans suit $\mathcal{N}(\mu;\sigma)$.
On donne $\mathbb{P}(X < 1,4) = 0,025$ et $\mathbb{P}(X > 2,6) = 0,025$.

Sans particularisme des données, on ramène les informations à la loi centrée réduite et on obtient deux équations pour déterminer μ et σ. Mais ici les probabilités avant 1, 4 et après 2, 6 sont égales ce qui entraîne que $\mu = \dfrac{1{,}4 + 2{,}6}{2} = 2{,}0$. Et l'intervalle centré sur la moyenne qui contient 95 % de la population est $[\mu - 2\sigma ; \mu + 2\sigma]$, soit $\sigma = 0{,}3$.

• La proposition **c.** est fausse à cause de l'unité qui devrait être au carré.

$$\bullet\ \mathbb{P}(X > 2{,}3) = \mathbb{P}\left(\frac{X-2}{0{,}3} > \frac{2{,}3-2}{0{,}3}\right) = \mathbb{P}(Z > 1)$$
$$= 1 - \Phi(1) = 1 - 0{,}8413 = 0{,}1587 \approx 0{,}16.$$

7-4 **a)** La fonction f est positive ou nulle si $a \geq 0$ et continue sur $\mathbb{R}$.
Pour que f soit une densité de probabilité, il reste donc la condition :

$$\int_{-\infty}^{+\infty} f(x)\,\mathrm{d}x = 1.$$

$$\int_{-\infty}^{+\infty} f(x)\,\mathrm{d}x = a\int_0^1 x(1-x)\,\mathrm{d}x = a\left[\frac{x^2}{2} - \frac{x^3}{3}\right]_0^1 = \frac{a}{6}.$$

Il faut donc choisir $a = 6$.

b) $E(X) = \displaystyle\int_{-\infty}^{+\infty} x\, f(x)\mathrm{d}x = 6\int_0^1 x^2(1-x)\mathrm{d}x$

$$= 6\left[\frac{x^3}{3} - \frac{x^4}{3}\right]_0^1 = \frac{1}{2}.$$

$$E(X^2) = \int_{-\infty}^{+\infty} x^2 f(x)\mathrm{d}x = 6\int_0^1 x^3(1-x)\mathrm{d}x$$

$$= 6\left[\frac{x^4}{4} - \frac{x^5}{5}\right]_0^1 = 0{,}3.$$

$$V(X) = E(X^2) - (E(X))^2 = 0{,}05.$$

7-5 • Comme f est positive et continue, pour qu'elle soit une densité de probabilité, il faut que : $\displaystyle\int_0^{+\infty} a\, t\, \mathrm{e}^{-\lambda t}\,\mathrm{d}t = 1$.

À l'aide d'une intégration par parties, on obtient :

$$\int_0^x t\,\mathrm{e}^{-\lambda t}\mathrm{d}t = \left[-\frac{a}{\lambda} t\,\mathrm{e}^{-\lambda t}\right]_0^x + \frac{a}{\lambda}\int_0^x \mathrm{e}^{-\lambda t}\mathrm{d}t$$

$$= -\frac{a}{\lambda} x\,\mathrm{e}^{-\lambda x} - \frac{a}{\lambda^2}\mathrm{e}^{-\lambda x} + \frac{a}{\lambda^2}$$

Comme $\lim\limits_{x\to+\infty} \mathrm{e}^{-\lambda x} = 0$ et $\lim\limits_{x\to 0} x\,\mathrm{e}^{-\lambda x} = 0$, on obtient :

$$\int_0^{+\infty} a\,t\,\mathrm{e}^{-\lambda t}\mathrm{d}t = \frac{a}{\lambda^2} = 1.$$

• Calculons $E(X) = \int_0^{+\infty} a\,t^2\,\mathrm{e}^{-\lambda t}\mathrm{d}t$.

À l'aide d'une intégration par parties, on obtient :

$$\int_0^x a\,t^2\,\mathrm{e}^{-\lambda t}\mathrm{d}t = \left[-\frac{a}{\lambda}t^2\,\mathrm{e}^{-\lambda t}\right]_0^x + \frac{2}{\lambda}\int_0^x a\,t\,\mathrm{e}^{-\lambda t}\mathrm{d}t$$

puis en faisant tendre x vers $+\infty$:

$$E(X) = \frac{2}{\lambda}\int_0^{+\infty} a\,t\,\mathrm{e}^{-\lambda t}\mathrm{d}t = \frac{2}{\lambda} \text{ car } \lim_{x\to+\infty} x^2\mathrm{e}^{-\lambda x} = 0.$$

• Comme, par hypothèse, $E(X) = 1000$, on en déduit successivement :

$$\lambda = 2\times 10^{-3} \qquad \text{et} \qquad a = 4\times 10^{-6}.$$

7-6

Ayez le réflexe de ramener tout problème relatif à une loi normale à la loi normale centrée réduite.

Vous disposez de deux tables (1 et 2). Vous pouvez répondre à toutes les questions avec une seule table. Mais il est plus simple d'utiliser la table 1 quand vous connaissez des bornes et cherchez une probabilité, et la table 2 quand vous connaissez une probabilité et cherchez des bornes.

a) $P(X \leq 6) = P\left(\dfrac{X-4}{2} \leq 1\right) = P(Z \leq 1)$ où Z suit $\mathcal{N}(0,1)$

$= \Phi(1) = 0{,}8413$

b) $P(X \leq y) = P\left(\dfrac{X-3}{1{,}5} \leq \dfrac{y-3}{1{,}5}\right)$

$= P(Z \leq a)$ où Z suit $\mathcal{N}(0,1)$ et

$$a = \frac{y-3}{1{,}5}$$

$P(Z \leq a) = \Phi(\mathrm{a}) = 0{,}4218$ entraîne $a < 0$ puisque la probabilité est $< 0{,}5$. Avec la table 2, on a : $\dfrac{\alpha}{2} = 0{,}4218$, soit $\alpha \approx 0{,}84$, d'où, par lecture, $-a = 0{,}202$. On en déduit $x \approx 2{,}7$.

b) $P(2{,}5 \leq X \leq 6{,}5) = P\left(-1{,}25 \leq \dfrac{X-5}{2} \leq 0{,}75\right)$
$= P(-1{,}25 \leq Z \leq 0{,}75)$
$= \Phi(0{,}75) - \Phi(-1{,}25) = 0{,}7734 - (1 - 0{,}8944)$
$= 0{,}6678.$

c) On cherche un intervalle $[6 - y; 6 + y]$ tel que :

$$0{,}9 = P(6 - y \leq X \leq 6 + y) = P\left(-\frac{y}{2} \leq \frac{X-6}{2} \leq \frac{y}{2}\right)$$

soit $P\left(-\dfrac{y}{2} \leq Z \leq \dfrac{y}{2}\right) = 0{,}90.$

On est dans la situation de la table 2 avec $\alpha = 0{,}10$ et on lit $\dfrac{y}{2} = 1{,}645$.
L'intervalle cherché est donc [2,71; 9,29].

7-7 Si X suit $\mathcal{N}(500;40)$ alors $Z = \dfrac{X - 500}{40}$ suit $\mathcal{N}(0;1)$.

a) $P(X > 560) = P(Z > 1{,}5) = 1 - P(Z \leq 1{,}5) = 1 - 0{,}9332 = 0{,}0668$.
En assimilant les fréquences expérimentales aux probabilités (car $n = 80$ est « grand »), le nombre de veaux est donc : $0{,}0668 \times 80 \approx 5$.

b) $P(0 < X < 480) = P(-12{,}5 < Z < -0{,}5) \approx P(Z < -0{,}5)$
$P(Z < -0{,}5) = \Phi(-0{,}5) = 1 - \Phi(0,\ 5) = 1 - 0{,}6915 = 0{,}3085$
ce qui conduit à : $0{,}3085 \times 80 \approx 25$ veaux.

c) $P(450 < X < 550) = P(-1{,}25 < Z < 1{,}25) = 2\Phi(1{,}25) - 1 = 0{,}7888$, ce qui conduit à : $0{,}7888 \times 80 \approx 63$ veaux.

d) On cherche k tel que $P(X > k) = 0{,}15 = P\left(Z > \dfrac{k - 500}{40}\right)$.

Cette situation correspond au graphique de la table 2 avec $\dfrac{\alpha}{2} = 0{,}15$ et $\dfrac{k - 500}{40} > 0$.
Avec $\alpha = 0{,}30$, on lit donc $\dfrac{k - 500}{40} = 1{,}036$, d'où $k \approx 541{,}4$. Un animal sera donc sélectionné à partir de 541,4 kg.

7-8 a) Soit X la durée (en min) du trajet de la directrice.
X suit la loi $\mathcal{N}(13;3)$. La directrice arrive en retard si $X > 15$.

$P(X > 15) = P\left(Z > \dfrac{2}{3}\right)$ où Z suit $\mathcal{N}(0;\ 1)$

$\approx 1 - \Phi(0{,}67) = 1 - 0{,}7486 = 0{,}2514$.

b) Le secrétaire arrive à l'heure si le train arrive avant 8 h 50 et l'autobus arrive devant le bureau avant 9 h.
Soit Y la durée (en min) du trajet en train ; Y suit $\mathcal{N}(16;\ 2)$.

Soit Y' la durée (en min) du trajet en autobus ; Y' suit $\mathcal{N}(9; 1)$.
La probabilité d'arriver à l'heure est :

$$P(Y < 18) \times P(Y' < 10) \text{ car } Y \text{ et } Y' \text{ sont indépendantes}$$
$$= P(Z < 1) \times P(Z < 1) \text{ où } Z \text{ suit } \mathcal{N}(0; 1)$$
$$= (\Phi(1))^2 = (0{,}8413)^2 = 0{,}7078.$$

c)

« au moins un individu arrive à l'heure » : pensez à l'événement contraire.

La probabilité que la directrice soit en retard est 0,2514.
La probabilité que le secrétaire soit en retard est $1 - 0{,}7078 = 0{,}2922$.
Comme ces deux événements sont indépendants, la probabilité pour que les deux soient en retard est : $0{,}2514 \times 0{,}2922 = 0{,}0735$.
On en déduit la probabilité de l'événement contraire : la directrice ou le secrétaire arrive à l'heure avec la probabilité : $1 - 0{,}0735 = 0{,}9265$.

7-9

Le premier réflexe à avoir, c'est de distinguer ce qui est aléatoire (on notera avec une majuscule) et ce qui est inconnu mais fixé après la résolution du problème (on notera avec une minuscule).

Le problème posé est un exemple de gestion de stock. C'est la demande qui est aléatoire alors que le stock à constituer est inconnu mais fixe.

Soit D la variable aléatoire égale au nombre de vaccins demandés. En supposant la population homogène par rapport à une telle demande, et les demandes individuelles indépendantes, D suit la loi binomiale de paramètres $n = 20\,000$ et $p = 0{,}4$. Le problème consiste à déterminer le nombre x de vaccins à stocker pour que :

$$P(D > x) \leq 0{,}1.$$

Avec la loi binomiale, le problème est quasi impossible. Mais avec $n = 20\,000$, $np = 8\,000$, $nq = 12\,000$, nous pouvons approximer la loi de Dpar la loi de Gauss de paramètres :

$$E(D) = np = 8\,000 \text{ et } \sigma(D) = \sqrt{npq} = \sqrt{4\,800}.$$

On a donc :

$$P(D > x) \leq 0{,}1 \Rightarrow P\left(Z > \frac{x - 8\,000}{\sqrt{4\,800}}\right) \leq 0{,}1 \text{ où } Z \text{ suit } \mathcal{N}(0; 1).$$

On est dans la configuration graphique de la table 2 avec $\frac{\alpha}{2} = 0{,}10$ et $\frac{x - 8\,000}{\sqrt{4\,800}} > 0$.

Avec $\alpha = 0{,}20$, on lit donc $\dfrac{x - 8\,000}{\sqrt{4\,800}} = 1{,}282$, d'où $x \approx 8089$.

On stockera donc 8 090 vaccins et peut-être 8 100, suivant des critères extérieurs comme l'emballage …

7-10 Soit X la variable aléatoire représentant le nombre de graines germées sur un total de 400 graines. Si l'on admet que la germination d'une graine n'a pas d'influence sur la germination des graines voisines, X suit la loi binomiale $\mathcal{B}$ (400; 0,8). Le calcul direct de :

$$P(X \geq 300) = P(X = 300) + P(X = 301) + \ldots + P(X = 400)$$

est pratiquement impossible. Mais nous sommes dans le cas où nous pouvons approximer la loi de X par une loi normale.
Les paramètres de la loi de X sont :

$$E(X) = np = 320 \quad \text{et} \quad \sigma(X) = \sqrt{npq} = 8.$$

X suit donc approximativement la loi normale $\mathcal{N}$ (320; 8).
Avec la correction de continuité, on obtient :

$$\begin{aligned} P(X \geq 300) &= P(X \geq 299{,}5) = P\left(\frac{X - 320}{8} \geq -2{,}5625\right) \\ &= 1 - \Phi(-2{,}5625) = \Phi(2{,}5625) \approx \Phi(2{,}56) \end{aligned}$$

Où Φ désigne la fonction de répartition de la loi normale centrée réduite.
On obtient donc : $P(X \geq 300) \approx 0{,}9948$.

7-11 **a)** Si X est la variable aléatoire égale à la longueur en cm de la tige d'un chrysanthème pris au hasard, la fleur est classée en catégorie extra si $X \geq 80$.
Comme X suit la loi normale $\mathcal{N}(92; 8)$, on a donc :

$$\begin{aligned} P(X \geq 80) &= P\left(\frac{X - 92}{8} \geq \frac{80 - 92}{8}\right) \\ &= 1 - \Phi(-1{,}5) = \Phi(-1{,}5) = 0{,}9332. \end{aligned}$$

b) Chacune des $n = 6000$ fleurs a une probabilité $p = 0{,}9332$ d'être classée extra et il y a indépendance entre les fleurs.
Le nombre N de fleurs classées extra suit donc la loi binomiale $\mathcal{B}$(6000; 0,9332).
L'espérance mathématique de N est :
$E(N) = 6000 \times 0{,}9332 \approx 5600$.

c) Parmi les 6000 fleurs de la production mensuelle, il y a donc en moyenne :
5600 fleurs payées 10 € les 10, soit un total de 5600 € ;

400 fleurs payées 6 € les 10, soit un total de 240 €.
L'espérance mathématique de la recette totale pour la production mensuelle est donc : 5600 + 240 = 5840 €.

7-12 **a)** Si X désigne le nombre de parties gagnées par Albert au cours des n parties, Albert a donc perdu $n - X$ parties.
À raison de 1 € par partie, son gain algébrique est donc de $X - (n - X) = 2X - n$.
Le règlement s'effectuera donc sans dette si :

$$-20 \leq 2X - n \leq 20 \Rightarrow 0{,}5n - 10 \leq X \leq 0{,}5n + 10.$$

b) Pour chaque partie, la probabilité pour qu'Albert gagne est de 0,5. Et il y a indépendance entre les parties.
X suit donc la loi binomiale $\mathcal{B}(n; 0{,}5)$.
Si $n \leq 20$, la probabilité d'un règlement sans dette est égale à 1.
Entre 20 et 30, les calculs seraient pénibles. Mais si on cherche n avec $n \geq 30$,on peut approximer la loi de X par la loi normale $\mathcal{N}(0{,}5n; 0{,}5\sqrt{n})$.
Avec la correction de continuité, l'hypothèse s'écrit :

$$\begin{aligned} P(0{,}5n - 10 \leq X \leq 0{,}5n + 10) &= P(0{,}5n - 10{,}5 \leq X \leq 0{,}5n + 10{,}5) \\ &= P\left(-\frac{10{,}5}{0{,}5\sqrt{n}} \leq \frac{X - 0{,}5n}{0{,}5\sqrt{n}} \leq \frac{10{,}5}{0{,}5\sqrt{n}}\right) \geq 0{,}68 \end{aligned}$$

On est dans la situation de la table 2 avec $\alpha = 0{,}32$.

On lit : $\dfrac{10{,}5}{0{,}5\sqrt{n}} = 0{,}994$, d'où $n = 447$.

On peut aussi utiliser la table 1, ce qui conduit à $n = 441$. En situation expérimentale, cela serait sans importance.

CHAPITRE 8

Échantillonnage Estimation d'un paramètre

PLAN

8.1 Échantillonnage
8.2 Estimation ponctuelle non biaisée
8.3 Estimation ponctuelle d'une moyenne et d'une variance
8.4 Estimation ponctuelle d'un pourcentage
8.5 Estimation d'un pourcentage par intervalle de confiance
8.6 Estimation d'une moyenne par intervalle de confiance
8.7 Estimation d'une variance par intervalle de confiance

OBJECTIFS

- Estimer par un nombre une fréquence, une moyenne, une variance, à partir d'une information incomplète
- Situer dans un intervalle une fréquence, une moyenne, une variance, à partir d'une information incomplète, avec un risque choisi ou à déterminer

8.1 ÉCHANTILLONNAGE

Nécessité des échantillons

On s'intéresse souvent à l'étude d'un caractère dans une population à laquelle on n'a pas accès (l'ensemble des poissons d'un océan …). Mais si on y avait accès, un recensement pourrait être trop cher, ou même produire des valeurs douteuses comme quand on interroge une certaine tranche de personnes sur leur âge.

Si on extrait plusieurs échantillons de taille n fixée, les résultats obtenus sont variables, ce qu'on appelle des fluctuations d'échantillonnage. À partir d'un échantillon, on n'a donc pas de certitudes, mais des estimations de paramètres.

L'échantillonnage est dit **non-exhaustif** si le tirage des n individus constituant l'échantillon a lieu avec remise.

Il est **exhaustif** si le tirage est réalisé sans remise. En fait, le plus souvent la taille d'un échantillon est faible par rapport à celle de la population et on assimile alors l'échantillonnage au cas non-exhaustif.

Constitution d'un échantillon

L'échantillon utilisé doit être représentatif de la population, c'est-à-dire reproduire les catégories pertinentes pour l'étude effectuée. Pour un sondage d'opinion on reproduit les tranches d'âge, mais on ne tient pas compte de la couleur des cheveux.

L'échantillon doit être constitué de manière aléatoire et non par volontariat (ce sont les râleurs qui téléphonent), ou par commodité (prélever des épis de blé seulement en bordure du champ).

Deux échantillons

Pour étudier l'effet d'un traitement pouvant agir sur une maladie, d'un protocole pouvant agir sur la croissance… on est amené à constituer et à comparer deux échantillons (ou parfois plus).

Si les échantillons sont constitués par des individus différents, il s'agit d'échantillons **indépendants**.

S'il s'agit des *mêmes* individus soumis, dans un ordre tiré au sort et avec un délai suffisant, au principe actif et à un placebo, à chacun des deux protocoles … les individus sont associés deux à deux. On dit que les échantillons sont **appariés**.

8.2 ESTIMATION PONCTUELLE NON BIAISÉE

On dit qu'une variable aléatoire T_n, associée à un échantillon de taille n, est un **estimateur sans biais**, ou **non biaisé**, d'un paramètre θ si $E(T_n) = \theta$. Dans le cas contraire, l'estimateur est dit biaisé.

Si l'estimateur est non biaisé, cela signifie que, si, pour un grand nombre d'échantillons de taille n, on calcule les diverses estimations t_n obtenues et qu'on en fait la moyenne, on obtient à peu près θ.

Si en plus, on a $\lim_{n\to\infty} V(T_n)$, l'estimateur est dit **convergent**.

La variance étant un indicateur de dispersion, on préfère un estimateur sans biais, convergent, dont la variance soit aussi faible que possible.

8.3 ESTIMATION PONCTUELLE D'UNE MOYENNE ET D'UNE VARIANCE

Notations

On étudie sur la population un caractère quantitatif X dont la moyenne μ et la variance σ^2 sont à estimer.

Pour un échantillon de taille n, on note $x_1, \ldots, x_n$ les valeurs observées et $X_1, \ldots, X_n$ les variables aléatoires associées. La variable aléatoire X_i prend pour valeurs les i-èmes mesures d'un grand nombre d'échantillons. Le réel x_i est la i-ème mesure de l'unique échantillon disponible.

On définit deux variables aléatoires :

$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ qui prend pour valeurs les moyennes des échantillons de taille n ;

$V_e = \frac{1}{n}\sum_{i=1}^{n} (X_i - \overline{X})^2$ qui prend pour valeurs les variances des échantillons de taille n.

Estimation ponctuelle non biaisée d'une moyenne

Théorème

$$E(\overline{X}) = \mu \quad ; \quad V(\overline{X}) = \frac{\sigma^2}{n}.$$

$\overline{X}$ est donc un estimateur sans biais, convergent, de μ.

Dans la pratique, on dispose d'un seul échantillon et on retient comme estimation de la moyenne théorique μ la moyenne $\overline{x}$ de l'échantillon.

Estimation ponctuelle non biaisée d'une variance

Théorème

$$E(V_e) = \frac{n-1}{n}\sigma^2.$$

V_e est donc un estimateur biaisé de σ^2. Pour obtenir un estimateur non biaisé on considère $S^2 = \frac{n}{n-1}V_e$. Ces deux estimateurs sont convergents.

Dans la pratique, on dispose d'un seul échantillon et on retient comme estimation de la variance théorique σ^2 la variance estimée :

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{n}{n-1}\left[\frac{1}{n}\sum_{i=1}^{n}x_i^2 - (\overline{x})^2\right].$$

On estime aussi σ par s, bien que cette estimation soit biaisée.

Attention aux notations choisies selon les critères : en grec ce qui concerne la population, en latin ce qui concerne l'échantillon ; et l'écriture la plus simple pour la notion la plus importante.

Certains auteurs notent s^2 la variance de l'échantillon au lieu de v_e ici. Remarquez que $s^2 = \frac{n}{n-1}v_e$.

Avec une calculatrice élémentaire, on obtient directement s avec la touche S_x ou σ_{n-1}. Dans ce cas, n'oubliez pas d'élever au carré pour avoir s^2.

8.4 ESTIMATION PONCTUELLE D'UN POURCENTAGE

Notations

Si la population est formée d'individus ayant ou non un caractère A, on définit la variable aléatoire F qui prend pour valeurs les fréquences observées de A sur des échantillons de taille n, supposés tirés avec remise.

Soit π la probabilité pour qu'un individu, pris au hasard dans la population, présente le caractère A. C'est π qu'il s'agit d'estimer.

Estimation ponctuelle non biaisée d'un pourcentage

Théorème

$$E(F) = \pi \quad ; \quad V(F) = \frac{\pi(1-\pi)}{n}.$$

F est donc un estimateur sans biais, convergent, de π.

Dans la pratique, on dispose d'un seul échantillon et on retient comme estimation de la proportion théorique π la fréquence f de l'échantillon.

8.5 ESTIMATION D'UN POURCENTAGE PAR INTERVALLE DE CONFIANCE

Principe d'un intervalle de confiance

Soit π la fréquence d'apparition d'un caractère A dans une population et f la fréquence d'apparition du même caractère dans un échantillon de taille n. On sait que f est une estimation ponctuelle non biaisée de π. Mais quelle confiance peut-on accorder à cette estimation ?

On répond à cette question en choisissant un nombre $\alpha \in]0,1[$ et en déterminant un intervalle $]a,b[$ tel que l'on ait la probabilité α de se tromper en affirmant que π appartient à cet intervalle.

L'intervalle obtenu est dit intervalle de confiance de π au coefficient de risque α, ou au coefficient de sécurité $1 - \alpha$.

La construction d'un intervalle de confiance consiste à introduire une variable aléatoire dont on connaît la distribution de probabilité.

Intervalle de confiance de *p*

nF, nombre d'individus ayant le caractère A dans un échantillon de taille n, suit la loi binomiale $\mathcal{B}(n, \pi)$. On peut en déduire un intervalle de confiance de π par cumul de probabilités élémentaires (avec un ordinateur).

Si on peut approximer $\mathcal{B}(n, \pi)$ par une loi normale, alors $Z = \dfrac{F - \pi}{\sqrt{\dfrac{\pi(1-\pi)}{n}}}$ suit approximativement la loi $\mathcal{N}(0,1)$.

On peut donc dire (au risque α) que :

$$f - z_\alpha \sqrt{\frac{\pi(1-\pi)}{n}} \leq \pi \leq f + z_\alpha \sqrt{\frac{\pi(1-\pi)}{n}}.$$

Pour expliciter les bornes, deux points de vue sont possibles.

- Remplacer π par f, ce qui donne l'intervalle de confiance :

$$]f - z_\alpha \sqrt{\frac{f(1-f)}{n}}, f + z_\alpha \sqrt{\frac{f(1-f)}{n}}[.$$

- Estimer sans biais $\pi(1-\pi)$, ce qui donne l'intervalle de confiance :

$$]f - z_\alpha \sqrt{\frac{f(1-f)}{n-1}}, f + z_\alpha \sqrt{\frac{f(1-f)}{n-1}}[.$$

Mais comme n est supposé grand, il s'agit d'une différence sans conséquence.

8.6 ESTIMATION D'UNE MOYENNE PAR INTERVALLE DE CONFIANCE

Cas d'une population gaussienne (σ connu)

Si X suit une loi normale, alors $\overline{X}$ suit la loi normale $\mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}})$.

Pour un risque α donné, on lit l'écart réduit z_α dans la table 2 et on peut affirmer, avec un risque α de se tromper, que :

$$-z_\alpha < \frac{\overline{x} - \mu}{\frac{\sigma}{\sqrt{n}}} < z_\alpha$$

soit :

$$\mu \in \left]\overline{x} - z_\alpha \frac{\sigma}{\sqrt{n}}, \overline{x} + z_\alpha \frac{\sigma}{\sqrt{n}}\right[.$$

Cas d'une population gaussienne (σ inconnu)

La variable aléatoire $T = \dfrac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}}$ suit la loi de Student à $v = n - 1$ degrés de liberté.

Pour un risque α donné, on lit le nombre t_α dans la table 3 (en ligne le degré de liberté et en colonne α) et on peut affirmer (au risque α) que :

$$\mu \in \left]\overline{x} - t_\alpha \frac{s}{\sqrt{n}}, \overline{x} + t_\alpha \frac{s}{\sqrt{n}}\right[.$$

Lorsque le nombre de degrés de liberté tend vers l'infini, la fonction de répartition de la loi de Student tend vers celle de la loi normale centrée réduite. Pour α donné, t_α tend donc vers Z_α. C'est ce qui explique la présence de la ligne donnant Z_α en bas de la table qui donne t_α.

Cas d'une loi quelconque et d'un grand échantillon

Si $n \geq 30$, la variable aléatoire $U = \dfrac{\overline{X} - \mu}{\dfrac{s}{\sqrt{n}}}$ suit approximativement la loi normale centrée réduite. L'intervalle de confiance de μ (au risque α) s'écrit donc :

$$\left]\overline{x} - z_\alpha \frac{s}{\sqrt{n}}, \overline{x} + z_\alpha \frac{s}{\sqrt{n}}\right[.$$

8.7 ESTIMATION D'UNE VARIANCE PAR INTERVALLE DE CONFIANCE

Théorème. Si X suit une loi normale, la variable aléatoire $Y = \dfrac{n-1}{\sigma^2} S^2$ suit la loi du χ^2 (lire khi-deux) à $v = n - 1$ degrés de liberté.

Une loi du χ^2 est une loi de probabilité continue dont la densité est nulle pour $x < 0$, et dépend d'un paramètre appelé nombre de degrés de liberté (ou degré de liberté, ou d.d.l.) ; voir chap. 10 pour l'allure des graphiques des densités.

Utilisation si $n \leq 31$

Pour α donné, on détermine les nombres a et b tels que

$$P(Y \leq a) = \frac{\alpha}{2} \text{ et } P(Y \geq b) = \frac{\alpha}{2}.$$

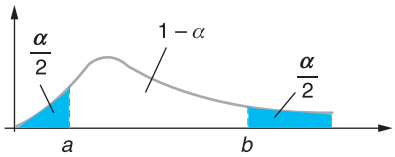


Figure 8-1

Les nombres a et b se lisent dans la table 4. On prend la ligne correspondant au degré de liberté, soit ici $n - 1$. Pour la colonne, regardez bien la légende graphique de la table : il s'agit de la surface à droite. Donc pour $\alpha = 0,05$ (choix le plus classique), vous lisez b dans la colonne 0,025 et a dans la colonne 0,975 (puisque la surface totale est égale à 1).

La seule valeur connue de S^2 étant s^2, on obtient comme intervalle de confiance de σ^2 au risque α :

$$\left]\frac{(n-1)s^2}{b}, \frac{(n-1)s^2}{a}\right[.$$

Utilisation si $n > 31$

Le théorème cité précédemment est vrai quel que soit n. Mais les tables du χ^2 s'arrêtent habituellement au degré de liberté $v = 30$. On ne peut donc pas les utiliser si $n > 31$. Mais, en l'absence d'ordinateur, on dispose du théorème d'approximation qui suit.

Théorème. Si Y est une variable aléatoire qui suit une loi du χ^2 à v degrés de liberté et si $v > 30$, alors la variable aléatoire $Z = \sqrt{2Y} - \sqrt{2v-1}$ suit à peu près la loi réduite $\mathcal{N}(0,1)$.

Utilisation

Ici on a : $Z = \sqrt{\dfrac{2(n-1)S^2}{\sigma^2}} - \sqrt{2n-3}$.

Après avoir choisi le risque α, on lit dans la table 2 la borne z_α tel que $P(-z_\alpha < Z < z_\alpha) = 1 - \alpha$ et on en déduit l'intervalle de confiance de σ^2 :

$$\sigma^2 \in \left]\frac{2(n-1)s^2}{(\sqrt{2n-3}+z_\alpha)^2}, \frac{2(n-1)s^2}{(\sqrt{2n-3}-z_\alpha)^2}\right[.$$

Estimation d'un paramètre par la méthode du maximum de vraisemblance

• Principe

Soit X une variable aléatoire, définie sur la population, de densité f dépendant d'un paramètre θ à estimer.

On dispose d'un échantillon de taille n dont les valeurs observées sont : $x_1, \ldots, x_n$.

La fonction L :

$\theta \mapsto L(\theta) = \prod_{i=1}^{n} f(x_i,\theta) = f(x_1,\theta) \times \cdots \times f(x_n,\theta)$ est dite fonction du maximum de vraisemblance.

La méthode du maximum de vraisemblance consiste à choisir comme estimation de θ la valeur θ_0 qui rend L maximale, c'est-à-dire qui vérifie (en supposant L deux fois continûment dérivable) :

$$\frac{\mathrm{d}L}{\mathrm{d}\theta}(\theta_0) = 0 \text{ et } \frac{\mathrm{d}^2L}{\mathrm{d}\theta^2}(\theta_0) < 0.$$

• **Remarques**

– En situation expérimentale, cette méthode nécessite l'emploi d'un ordinateur.

– Pour certains paramètres, l'estimation par le maximum de vraisemblance peut conduire à des résultats différents que l'estimation non biaisée.

– La méthode peut se généraliser à l'estimation simultanée de plusieurs paramètres.

– La fonction L étant en général strictement positive, on peut maximiser $\ln L$, ce qui est équivalent à maximiser L.

• **Exemple**

– *Énoncé*

On considère une population sur laquelle est définie une variable aléatoire X qui suit une loi de Poisson de paramètre λ. Les valeurs prises par X sur un échantillon de taille n sont $x_1, \ldots, x_n$ et ces nombres appartiennent à $\mathbb{N}$ puisqu'il s'agit d'une loi de Poisson.

Comme il s'agit d'une loi discrète dépendant d'un seul paramètre λ, la fonction du maximum de vraisemblance s'écrit :

$$L(\lambda) = P(X = x_1) \times \cdots \times P(X = x_n).$$

Déterminer par la méthode du maximum de vraisemblance une estimation de λ.

– *Solution*

$L(\lambda) = \dfrac{\lambda^{x_1}}{x_1!}\mathrm{e}^{-\lambda} \times \cdots \times \dfrac{\lambda^{x_n}}{x_n!}\mathrm{e}^{-\lambda} = \mathrm{e}^{-n\lambda}\dfrac{\lambda^{x_1+\ldots+x_n}}{x_1!\ldots x_n!}$. D'où :

$$\ln L(\lambda) = -n\lambda + (x_1 + \cdots + x_n)\ln\lambda - \sum_{i=1}^{n}\ln(x_i!)$$

$$\frac{\mathrm{d}\ln L}{\mathrm{d}\lambda}(\lambda) = -n + (x_1 + \cdots + x_n)\frac{1}{\lambda}$$

$$\frac{\mathrm{d}^2\ln L}{\mathrm{d}\lambda^2}(\lambda) = -\frac{x_1 + \cdots + x_n}{\lambda^2}$$

Pour $\lambda = \lambda_0 = \dfrac{x_1 + \cdots + x_n}{n}$

on a $\dfrac{\mathrm{d}\ln L}{\mathrm{d}\lambda}(\lambda_0) = 0$ et $\dfrac{\mathrm{d}^2\ln L}{\mathrm{d}\lambda^2}(\lambda_0) < 0$.

L'estimation de λ par la méthode du maximum de vraisemblance est donc la moyenne de l'échantillon étudié.

MOTS-CLÉS

- Estimation ponctuelle
- Estimation non biaisée
- Intervalle de confiance

EXERCICES

8-1 Sur un échantillon de 100 individus pris au hasard dans la population générale, la moitié sont porteurs de la maladie M (pour simplifier les calculs, $z_{0,05} = 1{,}96$ sera arrondi à 2).

❑ **a)** L'intervalle de confiance à 95 % de la prévalence de M est [40 % ; 60 %].

❑ **b)** Il y a 95 % de chances de ne pas se tromper en disant que la prévalence de M est comprise entre 40 % et 60 %.

❑ **c)** Il y a 95 % de chances de se tromper en disant que la prévalence de M est comprise entre 40 % et 60 %.

❑ **d)** Il y a 5% de chances de ne pas se tromper en disant que la prévalence de M n'est pas comprise entre 40 % et 60 %.

❑ **e)** Toutes choses égales par ailleurs, l'intervalle de confiance de la prévalence sera [30 % ; 70 %] sur un échantillon représentatif de 1000 individus.

8-2 Un échantillon est composé de 6 mesures d'un dosage. Les valeurs mesurées sont {54;50;55;55;56;45}. La valeur moyenne dans l'échantillon vaut 52,5.

❑ **a)** On peut estimer la moyenne de cette mesure dans la population dont est issu l'échantillon comme étant égale à 52,5.

❑ **b)** On peut estimer la variance de cette mesure dans la population dont est issu l'échantillon comme étant égale à 17,9.

❑ **c)** Le premier quartile de cet échantillon vaut 55.

❑ **d)** L'intervalle de confiance à 95 % de la moyenne vaut [49,1;55,9] (à 0,01 près).

❑ **e)** On ne peut pas calculer un intervalle de confiance avec la formule du cours.

D'après concours Strasbourg

8-3 On considère un échantillon de 169 brebis de race Ile-de-France. Ces brebis ont été mises en lutte. On a obtenu 108 brebis pleines.
Donnez un intervalle de confiance à 95 % du taux t de fertilité de cette race. t désigne le rapport du nombre de brebis pleines au nombre total de brebis.

8-4 Pour une certaine vaccination, on sait, par des études antérieures, que le pourcentage d'échecs est compris entre 10 et 15 pour cent.
On prépare une expérience pour connaître à ±1 (en %) le pourcentage de sujets non immunisés, en acceptant un coefficient de risque $\alpha = 0{,}05$.
Sur combien de sujets, au minimum, l'observation doit-elle porter ?

8-5 À la veille d'une consultation électorale comportant deux candidats, on a interrogé 100 électeurs constituant un échantillon représentatif. 58 d'entre eux ont déclaré avoir l'intention de voter pour le candidat Dupont.

a) Indiquez, avec une probabilité de 0,95, entre quelles limites se situe la proportion du corps électoral favorable à Dupont au moment du sondage. Peut-on en déduire, avec la même probabilité de 0,95, que Dupont serait élu si les opinions ne se modifiaient pas.

b) Avec une même fréquence observée d'électeurs favorables à Dupont, quelle devrait être la taille minimum de l'échantillon pour pouvoir affirmer, avec un risque de 5 %, que Dupont serait élu ?

8-6 Les données suivantes ont été obtenues sur des échantillons d'individus d'une région d'Europe. Le caractère étudié est la masse du cerveau (en g) pour des sujets de 20 à 49 ans.

Hommes

centres des classes	1170	1220	1270	1320	1370	1420	1470	total
effectifs	5	36	45	50	61	49	19	265

Femmes

centres des classes	1070	1120	1170	1220	1270	1320	1370	total
effectifs	12	22	45	54	52	20	10	215

Déterminez un intervalle de confiance au risque de 1 % :

a) pour la moyenne de la population des hommes ;

b) pour la moyenne de la population des femmes.

8-7 On a mesuré le poids de raisin par souche sur 10 souches prises au hasard dans une vigne. On a obtenu les résultats suivants (en kg) :

2,7; 3,2; 3,6; 4,1; 4,3; 4,7; 5,4; 5,9; 6,5; 6,9.

On suppose que le poids de raisin par souche suit une loi normale au niveau de la vigne.

a) Donnez un intervalle de confiance de la moyenne de la population au risque de 0,05.

b) Donnez au risque de 5 % un intervalle de confiance de la variance, puis de l'écart type, de la population.

8-8 Sur une parcelle de soja, on a mesuré la hauteur en cm de 100 plantes à l'âge de 6 semaines. Les résultats obtenus sont les suivants :

hauteurs	36	37	38	39	40	41
effectifs	6	11	26	32	14	11

Dans l'hypothèse d'une population gaussienne, déterminez un intervalle de confiance de la variance de la population, au coefficient de sécurité 0,95.

SOLUTIONS

8-1

☒ **a)** ☒ **b)** ☐ **c)** ☒ **d)** ☐ **e)**

• Il s'agit d'un intervalle de confiance d'une proportion dans le cas d'un échantillon de grande taille, soit:

$$\left[f - z_\alpha \sqrt{\frac{f(1-f)}{n}} ; f + z_\alpha \sqrt{\frac{f(1-f)}{n}} \right].$$

On donne $f = 0{,}5$, $\alpha = 0{,}05$, $z_\alpha = 2$, $n = 100$.

L'intervalle de confiance de la prévalence au risque 5 %, ou au niveau de confiance 95 %, est donc bien [0,4;0,6].

• Quand la taille de l'échantillon augmente, l'amplitude de l'intervalle de confiance diminue: voir la formule. Cela signifie que l'information expérimentale étant augmentée, la conclusion peut être plus précise.

8-2

☒ **a)** ☒ **b)** ☐ **c)** ☐ **d)** ☒ **e)**

• Avec une calculatrice, on obtient: $\overline{x} = 52{,}5$ (estimation non biaisée de μ) et $s^2 = 17{,}9$ (estimation non biaisée de σ^2).

• Pour déterminer un quartile, il faut commencer par ordonner les valeurs de l'échantillon: {45;50;54;55;55;56}. Le premier quartile est 50.

• S'agissant d'un échantillon de petite taille, la formule du cours suppose la population gaussienne, ce qui n'est pas dit.

8-3 Le taux de fertilité t correspond au pourcentage théorique p de l'événement A « la brebis est pleine » (fécondée si vous hésitez sur le vocabulaire de la zootechnie).
On peut supposer l'échantillon non exhaustif car la population des brebis Ile de France est très importante.

De plus, les conditions

$$n = 169 \geq 30; k = 108 \geq 5\,; n - k = 61 \geq 5$$

permettent d'utiliser une approximation par une loi normale, on obtient comme intervalle de confiance de t au risque α :

$$\left]f - z_\alpha \sqrt{\frac{f(1-f)}{n-1}}, f + z_\alpha \sqrt{\frac{f(1-f)}{n-1}}\right[,$$

soit $I =]0{,}56;\, 0{,}72[$ avec $\alpha = 0{,}05, z_\alpha = 1{,}96, f = \dfrac{108}{169}$.

8-4 Si π désigne le pourcentage de sujets non immunisés après vaccination dans la population, on veut connaître un intervalle de confiance de π, au risque $\alpha = 0{,}05$, dont la demi-amplitude soit au maximum de 0,01.
Si $n \geq 30$, nous sommes dans les conditions d'approximation de $\mathcal{B}(n,\pi)$ par une loi normale.
La condition s'écrit donc : $z_{0,05}\sqrt{\dfrac{\pi(1-\pi)}{n}} \leq 0{,}01$ (ici, il est inutile d'estimer π).

Sachant que $z_{0,05} = 1{,}96$, cette condition est équivalente à :

$$n \geq 38\,416\pi(1-\pi).$$

Pour $\pi \in [0{,}1;0{,}15]$ la fonction $\pi \mapsto f(\pi) - \pi^2 + \pi$ est croissante et on a : $f(\pi) \leq f(0{,}15)$ où $f(0{,}15) = 0{,}1275$.

Il suffit donc que n vérifie : $n \geq 38\,416 \times 0{,}1275$ soit $n \geq 4899$.

Nous retiendrons $n = 4900$ en supposant que la population est de taille suffisante pour que le tirage puisse être assimilé à un tirage avec remise.

Dans beaucoup de domaines, la recherche préalable du nombre d'individus sur lesquels doit porter l'expérience est un problème important, surtout quand la durée de l'observation est longue.

8-5 a) Échantillon de taille $n = 100$

Soit π la proportion du corps électoral favorable à Dupont au moment du sondage. L'estimation ponctuelle non biaisée de π est la fréquence $f = 0{,}58$ observée sur l'échantillon. Les valeurs de n et de π permettent d'approximer la loi binomiale $\mathcal{B}(n,\pi)$ par une loi normale.

L'intervalle de confiance de π, au risque α peut donc s'écrire :

$$I = \left]f - z_\alpha\sqrt{\frac{f(1-f)}{n-1}}, f + z_\alpha\sqrt{\frac{f(1-f)}{n-1}}\right[.$$

Comme $\alpha = 0{,}05$, on a $z_{0,05} = 1{,}96$, d'où $I =]0{,}482;\ 0{,}677[$.

La convention où l'on choisit n (voir le cours) ne modifie que très peu I.

On a donc au moins 95 chances sur 100 pour que π soit situé entre 0,482 et 0,677. Mais comme la borne inférieure de cet intervalle est inférieure à 0,5, on ne peut pas affirmer, au niveau de risque choisi, que Dupont serait élu.

Ce que les journalistes appellent fourchette un soir d'élection est un intervalle de confiance.

D'autre part, les instituts de sondage considèrent qu'un échantillon doit être de l'ordre de 1500 personnes pour donner un résultat fiable. On en est loin ici.

b) Échantillon de taille n à déterminer

Pour pouvoir affirmer que Dupont serait élu, il faut que la borne inférieure de l'intervalle de confiance soit supérieure à 0,5. On aura donc :

$$\textbf{(1)} \quad 0{,}5 < f - z_\alpha\sqrt{\frac{f(1-f)}{n-1}} \text{ où } f = 0{,}58.$$

Attention, le risque α d'un intervalle de confiance I de π se décompose en deux risques de se tromper quand π est à l'extérieur de I :

– la valeur de π est à gauche de I (probabilité $\frac{\alpha}{2}$) ;

– la valeur de π est à droite de I (probabilité $\frac{\alpha}{2}$).

Ici, seul le premier cas conduit à une erreur de prévision. Et comme le risque accepté est de 5 %, il faut retenir $\alpha = 0{,}10$ soit $z_\alpha = 1{,}645$.

L'inéquation **(1)** conduit alors à retenir = 104.

Si vous avez trouvé $n = 103$, c'est que vous avez utilisé l'autre point de vue présenté en cours ; pas de problème !

Si vous avez trouvé $n = 148$ ou $n = 147$, c'est que vous considérez que si Dupont obtient 80 % des voix, il n'est pas élu !

Vous avez fait une erreur en ne distinguant pas les deux côtés de l'extérieur de *I*.

Avec $n = 104$, on ne peut pas obtenir exactement $f = 0,58$ mais si 60 électeurs sur 104 se déclarent favorables à Dupont, on a $f \approx 0,58$. Il en est de même avec $n = 103$ ce qui confirme que les deux points de vue présentés en cours donnent des résultats très proches.

8-6 **a)** L'échantillon constitué par les 265 hommes étudiés a pour moyenne $\overline{x}_h \approx 1335,8$ g et pour écart type estimé $s_h \approx 77,57$ g.
S'agissant d'un grand échantillon, la moyenne de la population des hommes a pour intervalle de confiance :

$$\left]\overline{x}_h - z_\alpha \frac{s_h}{\sqrt{n_h}} \ ; \ \overline{x}_h + z_\alpha \frac{s_h}{\sqrt{n_h}}\right[$$

Comme $\alpha = 0,01$ on lit, dans la table 2, $z_{0,01} = 2,576$ et on obtient l'intervalle de confiance : $]1323;1349[$.

b) L'échantillon constitué par les 215 femmes étudiées a pour moyenne $\overline{x}_f \approx 1219,3$ g et pour écart type estimé $s_f \approx 73,54$ g.
De même que précédemment, on en déduit l'intervalle de confiance au risque 1 % pour la moyenne de la population des femmes : $]1206;1233[$.

Mesdames, si cet exercice provoquent des ricanements désobligeants de la part de certains garçons, répondez que les veaux ont une grosse tête …

8-7 L'échantillon est de taille $n = 10$. Il a pour moyenne $\overline{x} = 4,7$ kg et pour écart type estimé $s \approx 1,46$ g.
a) La population étant supposée gaussienne, la moyenne μ au niveau de la vigne a pour intervalle de confiance au risque α :

$$\left]\overline{x} - t_\alpha \frac{s}{\sqrt{n}} \ ; \ \overline{x} + t_\alpha \frac{s}{\sqrt{n}}\right[$$

où t_α est une borne associée à une loi de Student à $n - 1 = 9$ degrés de liberté.
Comme $\alpha = 0,05$, on lit, dans la table 3, $t_{0,05} = 2,262$ et on obtient pour intervalle de confiance de μ : $]3,65 ; 5,75[$.

Si cet intervalle vous paraît bien grand, et donc la conclusion peu précise, c'est parce que l'information expérimentale disponible est faible puisqu'elle ne porte que sur 10 mesures.

b) Dans les hypothèses énoncées, la variable aléatoire $Y = \dfrac{n-1}{\sigma^2}S^2$ suit la loi du χ^2 à $n - 1 = 9$ degrés de liberté.
On a $\alpha = 0{,}05$. On lit dans la table 4 :

$P(Y \geq a) = 0{,}975 \Rightarrow a = 2{,}70$; $P(Y \geq b) = 0{,}025 \Rightarrow b = 19{,}02$.

$P(a < Y < b) = 0{,}95$ devient :

$$P\left(\frac{(n-1)s^2}{b} < \sigma^2 < \frac{(n-1)s^2}{a}\right) = 0{,}95$$

ce qui donne des encadrements au risque 5 % :

$$1{,}01 < \sigma^2 < 7{,}15 \text{ puis } 1{,}01 < \sigma < 2{,}68.$$

8-8 L'échantillon de taille $n = 100$ a pour moyenne $\overline{x} = 38{,}7$ cm et sa variance estimée est $s^2 \approx 1{,}75$ cm^2.
Dans les hypothèses énoncées, la variable aléatoire $Y = \dfrac{n-1}{\sigma^2}S^2$ suit la loi du χ^2 à $n - 1 = 99$ degrés de liberté. Mais les tables disponibles ne permettent pas de lire les nombres a et b.
Cependant, dans ce cas, la variable aléatoire $U = \sqrt{2Y} - \sqrt{2n-3}$ suit à peu près la loi $\mathcal{N}(0,1)$ Comme $\alpha = 0{,}05$ (puisque le coefficient de sécurité est $1 - \alpha$), on peut dire, au risque 5 %, que :

$$\begin{aligned} -1{,}96 < u < 1{,}96 &\Leftrightarrow \sqrt{197} - 1{,}96 < \sqrt{2y} < \sqrt{197} + 1{,}96 \\ &\Leftrightarrow \frac{(\sqrt{197} - 1{,}96)^2}{2} < \frac{99}{\sigma^2}s^2 < \frac{(\sqrt{197} + 1{,}96)^2}{2} \\ &\Leftrightarrow 1.35 < \sigma^2 < 2{,}38. \end{aligned}$$

CHAPITRE 9

Introduction aux tests statistiques

PLAN

9.1 Test d'une hypothèse simple contre une hypothèse simple
9.2 Exemples d'utilisation

OBJECTIFS

- Connaître le fonctionnement général d'un test
- Comprendre la notion de risque associé à une décision

9.1 TEST D'UNE HYPOTHÉSE SIMPLE CONTRE UNE HYPOTHÉSE SIMPLE

Objectifs

Il s'agit de faire un choix entre plusieurs hypothèses possibles sans disposer d'informations suffisantes pour que le choix soit sûr.

On met en avant une hypothèse, dite **hypothèse nulle** et notée (H_0).

On souhaite vérifier si (H_0) est vraie, alors que deux hypothèses seulement sont possibles : (H_0) et une **hypothèse alternative** (H_1).

Si on ne précise pas, (H_1) est le contraire de (H_0). Le test est alors **bilatéral**.

Mais ce n'est pas toujours le cas, par exemple dans le cas d'un **test unilatéral**. L'exemple typique est le test de l'efficacité d'un médicament. On a alors pour (H_0) : « le médicament n'a pas d'effet » et pour (H_1) : « le médicament a de l'effet ». Mais l'effet ne doit pas être négatif. Un côté est donc interdit.

Risques

L'information étant incomplète, toute décision est associée à un risque.

Si on décide que (H_0) est fausse, le risque de se tromper est noté α et s'appelle **risque de première espèce**.

Si on décide que (H_0) est vraie, le risque de se tromper est noté β et s'appelle **risque de deuxième espèce**.

Le concepteur d'un test s'intéresse à la **puissance** du test qui est $1 - \beta$.

L'utilisateur d'un test s'intéresse au risque α et ses conclusions sont donc : (H_0) rejetée au risque α ; (H_0) non rejetée (ou acceptée) au risque α.

Fonctionnement

Dans chaque situation, on dispose d'un théorème dont le schéma est le suivant : Si (H_0) est vraie, et si on a des hypothèses de fonctionnement, alors une variable de décision X suit une loi théorique connue.

On repère la valeur idéale que devrait prendre X. On choisit un risque α (souvent 0,05) et on détermine une zone (en deux morceaux pour un test bilatéral comme dans le cas de la figure ci-dessous, en un morceau pour un test unilatéral), de probabilité α, éloignée de cette valeur idéale.

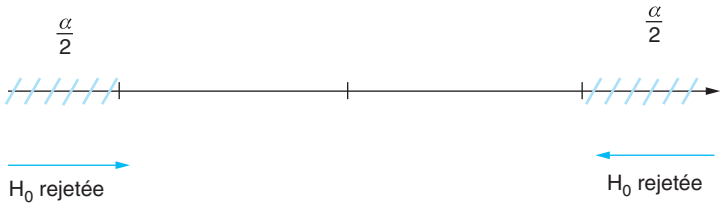


Figure 9-1

Si la valeur prise par X appartient à cette zone critique, on décide de rejeter (H_0) au risque α ; sinon on accepte (H_0).

Les logiciels déterminent souvent le risque minimum α pour lequel on rejette (H_0). Il vous reste alors à apprécier si ce risque est acceptable ou non.

9.2 EXEMPLES D'UTILISATION

Comparer un échantillon à une référence théorique

L'hypothèse (H_0) consiste à supposer que les différences observées sont suffisamment faibles pour être explicables par les hasards du tirage au sort.

Il s'agit d'un **test de conformité**.

Comparer plusieurs échantillons

L'hypothèse (H_0) consiste à supposer qu'ils proviennent d'une même population, c'est-à-dire que les différences observées sont explicables par les fluctuations d'échantillonnage.

Il s'agit d'un **test d'homogénéité**.

Risque du vendeur ; risque de l'acheteur

En économie, le risque de première espèce α s'appelle le risque du vendeur, et le risque de deuxième espèce β le risque de l'acheteur. Pourquoi ?

Un acheteur passe une commande très importante, avec des spécifications à respecter.

À la livraison, l'acheteur ne peut pas tout contrôler. Il analyse un échantillon de produits, en faisant attention au caractère aléatoire du prélèvement.

En général, il n'y a pas de problème. Mais deux types de décision erronée peuvent apparaître :

- Le prélèvement ne respecte pas les spécifications et la commande est refusée, alors qu'elle était globalement bonne [(H_0) est déclarée fausse, alors qu'elle est vraie]. C'est le risque α ; il est supporté par le vendeur.
- Le prélèvement respecte les spécifications et la commande est acceptée, alors qu'elle était globalement mauvaise [(H_0) est déclarée vraie, alors qu'elle est fausse]. C'est le risque β ; il est supporté par l'acheteur.

MOTS-CLÉS

- Hypothèse nulle
- Hypothèse alternative
- Risque de première espèce
- Risque de deuxième espèce

EXERCICES

9-1 L'erreur de première espèce α associée à un test est égale à :

❑ **a)** 1 + l'erreur de seconde espèce.

❑ **b)** La probabilité de ne pas rejeter H_0 sachant que H_0 est fausse.

❑ **c)** La probabilité de rejeter $H0$ sachant que $H0$ est vraie.

❑ **d)** La probabilité de rejeter H_0 sachant que H_0 est fausse.

❑ **e)** Autre réponse.

9-2 Pour évaluer la résistance d'un parasite à un nouveau traitement, on calcule le ratio entre la croissance parasitaire en présence et en absence de traitement. Ce ratio est un nombre réel variant de 0 (traitement totalement efficace) et 1 (aucune action du médicament, c'est-à-dire résistance du parasite).

On souhaite tester la résistance du parasite à partir de la valeur observée x pour ce ratio. L'hypothèse nulle est l'absence de résistance. Sous cette hypothèse, le ratio suit une loi uniforme entre 0 et 0, 5. Sous l'hypothèse alternative de résistance au traitement le ratio suit une loi uniforme entre 0, 4 et 0, 9.

On définit la zone de rejet comme les valeurs de x supérieures à un seuil s. Cochez la (ou les) proposition(s) exacte(s).

❑ **a)** Si le seuil est $s = 0, 4$, le risque de première espèce est de 5 %.

❑ **b)** Si le seuil est $s = 0, 4$, la puissance est de 0 %.

❑ **c)** Si le seuil est $s = 0, 4$, la puissance est de 100 %.

❑ **d)** Si le seuil est $s = 0, 45$, le risque de première espèce est de 5 %.

❑ **e)** Si le seuil est s = 0, 45, la puissance est de 0 %.

SOLUTIONS

9-1 ❑ **a)** ❑ **b)** ☒ **c)** ❑ **d)** ❑ **e)**

Le risque d'erreur α apparaît lors de la conclusion d'un test statistique. Tout test commence par « supposons que H_0 soit vraie ». S'il y a erreur, c'est que la conclusion est fausse, c'est-à-dire qu'on conclut que H_0 est fausse.

9-2 ❑ **a)** ❑ **b)** ☒ **c)** ❑ **d)** ❑ **e)**

• Le risque α est la probabilité de décider H_1 vraie alors que H_0 est vraie. Dans ce cas, on sait que le ratio suit la loi $\mathcal{U}([0; 0, 5]$ et ce ratio doit être supérieur à s.

➤ si $s = 0,4$, la probabilité est donc $\dfrac{0,5 - 0,4}{0,5 - 0} = 0,2$ soit 20 %.

➤ si $s = 0,45$, la probabilité est donc $\dfrac{0,5 - 0,45}{0,5 - 0} = 0,1$ soit 10 %.

• La puissance $1 - \beta$ est la probabilité de décider H_1 vraie alors que H_1 est vraie.

Dans ce cas, on sait que le ratio suit la loi $\mathcal{U}([0,4; 0,9]$ et ce ratio doit être supérieur à s.

➤ si $s = 0,4$, la probabilité est donc 1 soit 100 %.

➤ si $s = 0,45$, la probabilité est donc $\dfrac{0,9 - 0,45}{0,9 - 0,4} = 0,9$ soit 90 %.

CHAPITRE 10

Test du khi-deux (χ^2)

PLAN

10.1 Test de conformité : ajustement à une loi théorique
10.2 Test d'homogénéité : comparaison de plusieurs distributions
10.3 Test d'indépendance de deux caractères

OBJECTIFS

- Tester l'adéquation entre une distribution observée et une loi théorique provenant des lois mathématiques classiques, des lois de la génétique...
- Comparer les distributions observées sur divers échantillons, souvent associés aux modalités d'un facteur étudié
- Savoir si deux caractères qualitatifs peuvent être considérés comme indépendants

10.1 TEST DE CONFORMITÉ : AJUSTEMENT À UNE LOI THÉORIQUE

Problématique

Il s'agit de comparer une loi théorique et une distribution expérimentale.

On définit sur la population étudiée k événements $E_1,\ldots,E_k$ formant un système complet d'événements. Dans le modèle théorique, les probabilités de ces événements sont $p_1,\ldots,p_k$.

Sur un échantillon de taille n, les effectifs observés de ces événements sont $O_1,\ldots,O_k$.

Pour pouvoir confronter les observations et le modèle théorique, on calcule les effectifs théoriques, dits effectifs calculés : $C_i = np_i$ (qui ne sont pas nécessairement des entiers) de façon à avoir le même effectif total dans la théorie et dans l'observation.

Il est souvent commode de présenter ces informations à l'aide d'un tableau.

Pour comparer les O_i et les C_i, on calcule leur distance de façon un peu spéciale, la distance du χ^2.

Mise en place du test

Hypothèse (H_0)

La distribution observée dans l'échantillon est conforme à la distribution théorique choisie.

Théorème. Sous (H_0), la variable aléatoire Y prenant sur tout échantillon de taille n la valeur :

$$\chi_c^2 = \sum_{i=1}^{k} \frac{(O_i - C_i)^2}{C_i}$$

suit la loi du χ^2 à $v = k - 1 - p$ degrés de liberté où p est le nombre de paramètres qu'il faut éventuellement estimer pour connaître la loi théorique.

Remarques

- La plupart des utilisateurs exigent que l'on ait $C_i \geq 5$ pour tout i. Si ne n'est pas le cas, il faut regrouper de façon cohérente des événements jusqu'à ce que la condition soit réalisée.
- En faisant intervenir les fréquences observées $f_i = \frac{O_i}{n}$ on peut aussi écrire :

$$\chi_c^2 = n \sum_{i=1}^{k} \frac{(f_i - p_i)^2}{p_i}.$$

Décision

Le risque α de première espèce est fixé. À chaque valeur de v correspond un type de courbe pour la densité de la loi du χ^2.

Pour v donné, la table 4 permet de lire la borne χ_α^2 telle que : $P(Y \geq \chi_\alpha^2) = \alpha$.

- Si $\chi_c^2 \geq \chi_\alpha^2$, on rejette l'hypothèse (H_0) avec un risque α de se tromper.
- Si $\chi_c^2 < \chi_\alpha^2$, on ne peut pas rejeter l'hypothèse (H_0).

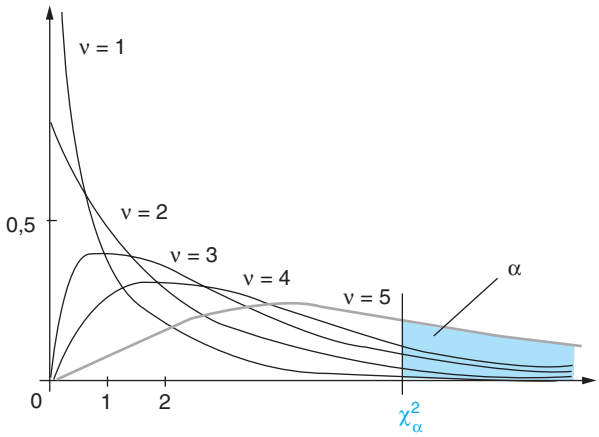


Figure 10-1

10.2 TEST D'HOMOGÉNÉITÉ : COMPARAISON DE PLUSIEURS DISTRIBUTIONS

Problématique

Sur la population P, on considère un caractère qui peut prendre k valeurs $A_1,\ldots,A_k$ (ou k modalités, ou k classes). On dispose de l échantillons $E_1,\ldots,E_l$ pouvant provenir de la population.

On peut donc dire que l'on a l distributions expérimentales dont on souhaite tester l'homogénéité.

Pour tout $i \in \{1,\ldots,k\}$ et pour tout $j \in \{1,\ldots,l\}$, on connaît O_{ij} effectif observé de la valeur A_i dans l'échantillon E_j.

On note $N = \sum_{j=1}^{l}\sum_{i=1}^{k} O_{ij}$ l'effectif total des échantillons.

Mise en place du test

a) Hypothèse (H_0)

Les différences observées entre les différents échantillons ne sont pas significatives. Les échantillons sont extraits d'une même population.

b) Calcul des effectifs théoriques sous (H_0)

Les l échantillons sont réunis en un seul échantillon de taille N, et la probabilité de A_i peut alors être estimée par la fréquence sur la réunion

des échantillons : $p_i = \dfrac{\sum_{j=1}^{l} O_{ij}}{N} = \dfrac{S_i}{N}$.

Si (H_0) est vraie, cette probabilité (toujours assimilée à une fréquence) doit se retrouver dans chaque échantillon.

L'effectif calculé de la classe A_i pour l'échantillon E_j est alors :

$$C_{ij} = p_i \left(\sum_{i=1}^{k} O_{ij} \right) = p_i T_j = \frac{S_i T_j}{N}.$$

Les calculs sont facilités par l'utilisation d'un tableau du genre :

	A_1	. . .	A_i	. . .	A_k	totaux
E_1	O_{11} **(C_{11})**		O_{i1} **(C_{i1})**		O_{k1} **(C_{k1})**	T_1
⋮						
E_j	O_{1j} **(C_{1j})**		O_{ij} **(C_{ij})**		O_{kj} **(C_{kj})**	T_j
⋮						
E_l	O_{1l} **(C_{1l})**		O_{il} **(C_{il})**		O_{kl} **(C_{kl})**	T_l
totaux	S_1		S_i		S_k	N

Quand le tableau des effectifs est construit correctement, les calculs sont mécaniques (mais comme ils sont fondés sur (H_0), rédigez toujours l'hypothèse nulle avant) : faire les totaux des lignes, des colonnes, le total général ; puis pour chaque case, calculez l'effectif théorique :

$$\frac{\text{total au bout de la ligne} \times \text{total au bout de la colonne}}{\text{total général}}$$

Vous reportez ce résultat dans chaque case, en le distinguant de O_{ij} qui y est déjà par tous les moyens à votre convenance, y compris en utilisant deux couleurs.

Vous pouvez vérifier que, sur chaque ligne et chaque colonne, les totaux des effectifs observés et des effectifs calculés sont les mêmes.

Théorème. Sous l'hypothèse (H_0), la variable aléatoire Y prenant sur chaque échantillon de taille N la valeur :

$$\chi_c^2 = \sum_{i=1}^{k}\sum_{j=1}^{l}\frac{(O_{ij} - C_{ij})^2}{C_{ij}}$$

suit la loi du χ^2 à $v = (k-1)(l-1)$ degrés de liberté.

On exige en général que $C_{ij} \geq 5$ pour tout i et pour tout j. Si ce n'est pas le cas, on fait des regroupements.

Décision

Le risque de première espèce α étant fixé et v étant connu, on lit dans la table 4 la valeur χ_α^2 telle que $P(Y \geq \chi_\alpha^2) = \alpha$.

- Si $\chi_c^2 \geq \chi_\alpha^2$, l'hypothèse (H_0) est rejetée au risque α.
- Si $\chi_c^2 < \chi_\alpha^2$, l'hypothèse (H_0) ne peut pas être rejetée.

10.3 TEST D'INDÉPENDANCE DE DEUX CARACTÈRES

Problématique

Dans une population P, chaque individu possède deux caractères qualitatifs A et B ayant les modalités respectives $A_1,\ldots,A_k$ et $B_1,\ldots,B_l$.

Pour tout $i \in \{1,\ldots,k\}$ et pour tout $j \in \{1,\ldots,l\}$, on connaît le nombre O_{ij} d'individus présentant les modalités A_i et B_j.

On note $N = \sum_{j=1}^{l}\sum_{i=1}^{k} O_{ij}$ l'effectif total de l'échantillon étudié.

Mise en place du test

a) Hypothèse (H_0)

Les deux caractères A et B sont indépendants.

b) Calcul des effectifs théoriques sous (H_0)

C_{ij} est l'effectif des individus présentant les modalités A_i et B_j si l'hypothèse (H_0) était vérifiée.

On note les effectifs marginaux :

$$T_j = \sum_{i=1}^{k} O_{ij} \quad \text{et} \quad S_i = \sum_{j=1}^{l} O_{ij}.$$

Sous (H_0), les événements A_i et B_j sont indépendants et on a :

$$P(A_i \cap B_j) = P(A_i) \times P(B_j) \text{ soit : } \frac{C_{ij}}{N} = \frac{S_i}{N} \times \frac{T_j}{N}.$$

On a donc $C_{ij} = \dfrac{S_i T_j}{N}$ et les calculs se présentent comme dans le cas précédent (test d'homogénéité) bien qu'il s'agisse d'un problème différent.

Théorème. Sous l'hypothèse (H_0), la variable aléatoire Y prenant sur chaque échantillon de taille N la valeur :

$$\chi_c^2 = \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{(O_{ij} - C_{ij})^2}{C_{ij}}$$

suit la loi du χ^2 à $v = (k-1)(l-1)$ degrés de liberté.

On exige en général que $C_{ij} \geq 5$ pour tout i et pour tout j. Si ce n'est pas le cas, on fait des regroupements.

Décision

Le risque de première espèce α étant fixé et v étant connu, on lit dans la table 4 la valeur χ_α^2 telle que $P(Y \geq \chi_\alpha^2) = \alpha$.

- Si $\chi_c^2 \geq \chi_\alpha^2$, l'hypothèse (H_0) est rejetée au risque α.
- Si $\chi_c^2 < \chi_\alpha^2$, l'hypothèse (H_0) ne peut pas être rejetée.

Définition mathématique d'une loi du χ^2

X suit la loi de Pearson, ou loi du χ^2 (lire khi-deux), à v degrés de liberté s'il existe v variables $Z_1, \ldots, Z_v$, indépendantes, qui suivent chacune la loi normale centrée réduite $\mathcal{N}(0,1)$, et telles que :

$$X = Z_1^2 + \cdots + Z_v^2.$$

On a :

$$E(X) = v \quad ; \quad V(X) = 2v.$$

MOTS-CLÉS

- Distance du χ^2
- Homogénéité
- Conformité
- Indépendance de deux caractères qualitatifs

EXERCICES

10-1 Dans un test du χ^2 pour un tableau à 3 lignes et 5 colonnes:

❑ **a)** Les variables doivent être quantitatives.

❑ **b)** Il y a 15 degrés de liberté.

❑ **c)** Au moins 8 cases doivent avoir des effectifs calculés sous H_0 supérieurs à 5.

❑ **d)** Les effectifs observés de chaque case sont comparés aux effectifs calculés sous H_0.

❑ **e)** Tous les effectifs observés dans les cases doivent être supérieurs à 1.

10-2 Un épidémiologiste veut savoir si le mois de naissance est associé à la survenue d'une maladie *M*.

Il tire au sort un échantillon de 1200 patients atteints de la maladie *M* et note leur mois de naissance. Il obtient un tableau indiquant le nombre de patients nés chaque mois.

Dans la population, les naissances se répartissent de façon égale entre les différents mois de l'année.

Nous considérons que l'on peut négliger les différences de nombre de jours entre les mois et que tous les mois ont la même durée.

Pour répondre à la question, l'épidémiologiste devra:

❑ **a)** Réaliser un test du χ^2 d'indépendance.

❑ **b)** Réaliser un test du χ^2 d'ajustement.

❑ **c)** Le test du χ^2 réalisé aura un degré de liberté égal à 1.

❑ **d)** Le test du χ^2 réalisé aura un degré de liberté égal à 12.

❑ **e)** Sous l'hypothèse nulle pour le test utilisé, les effectifs théoriques sont tous égaux à 100.

10-3 On a effectué le croisement de balsamines blanches avec des balsamines pourpres. En première génération les fleurs sont toutes pourpres. On obtient en deuxième génération quatre catégories avec des effectifs suivants :

Couleurs	pourpre	rose	blanc lavande	blanc
Effectifs	1790	547	548	213

Peut-on accepter l'hypothèse de répartition mendélienne $\left(\frac{9}{16};\frac{3}{16};\frac{3}{16};\frac{1}{16}\right)$ avec un risque $\alpha = 0{,}05$?

10-4 On cherche à savoir si la fréquence d'une maladie est liée au groupe sanguin. Sur 200 malades observés, on a dénombré 104 personnes du groupe *O*, 76 du groupe *A*, 18 du groupe *B* et 2 du groupe *AB*.

On admettra que dans la population générale la répartition entre les groupes est : groupe *O* : 47 %, groupe *A* : 43 %, groupe *B* : 7 %, groupe *AB* : 3 %. Que concluez-vous ?

10-5 Des cellules vivantes sont incubées en présence d'un composé radioactif. La technique d'autoradiographie permet de mesurer le taux de radioactivité absorbé par chaque organite cellulaire. Les résultats de ces mesures sont alors comparés à une distribution théorique aléatoire simulée par ordinateur en utilisant les surfaces des organites.

Une expérimentation conduit aux résultats suivants :

Organite cellulaire	**Taux de radioactivité (nombre de désintégrations enregistrées)**	
	Expérimental	**Simulé**
membrane plasmique	30	10
vésicules hyaloplasmiques	10	10
zone de Golgi	30	20
reticulum granulaire	20	40
lysosomes	10	5
noyau	0	15

La répartition observée pour la radioactivité est-elle le fait du hasard ?

10-6 Une enquête effectuée auprès du comptoir de 150 coopératives agricoles a permis d'étudier l'arrivée dans le temps des usagers de ces coopératives.

Pendant l'unité de temps, soit une heure, on a noté :

Nombre d'usagers arrivés	0	1	2	3	4	5	6
Nombre de coopératives	37	46	39	19	5	3	1

Peut-on admettre, au risque de 5 %, que la population suit une loi de Poisson ?

10-7 Dans la comparaison du taux d'occupation d'un matériel coûteux pour un mois d'hiver (janvier) et pour un mois d'été (juillet), on dispose de deux échantillons, l'un de 300 observations instantanées en janvier, l'autre de 200 observations instantanées en juillet.

	janvier	**juillet**
Occupation	240	150
Inoccupation	60	50

Peut-on considérer que le taux d'occupation de ce matériel est le même en janvier et en juillet ($\alpha = 0{,}05$) ?

10-8 Les résultats de l'évolution d'une maladie *M*, à la suite de l'emploi de l'un ou l'autre des traitements *A* et *B*, figurent dans le tableau ci-dessous, qui donne le nombre de malades appartenant à chacune des catégories :

	Guérison	**Amélioration**	**État stationnaire**	**Totaux**
A	280	210	110	600
B	220	90	90	400
Totaux	500	300	200	1000

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

10-9 À la suite du même traitement, on a observé 40 bons résultats chez 70 malades jeunes et 50 bons résultats chez 100 malades âgés.
Peut-on dire, au risque 10 %, qu'il existe une liaison entre l'âge du malade et l'effet du traitement ?

10-10 Lors d'une étude biologique portant sur une certaine espèce de mollusques, on a mesuré le taux de protéines *X* en mg de 36 individus appartenant à cette espèce. On a obtenu les résultats suivants :

X	]0 ; 1,5]	]1,5 ; 3]	]3 ; 4,5]	]4,5 ; 6]	]6 ; 7,5]	]7,5 ; 9]	]9 ; 10,5]
Nb d'individus	8	7	4	9	2	3	3

a) Estimez la moyenne et l'écart type de la population.

b) Peut-on admettre que le taux de protéines se distribue de façon gaussienne ?

SOLUTIONS

10-1 ❑ a) ❑ b) ❑ c) ☒ d) ❑ e)

Les variables peuvent être qualitatives ou quantitatives avec regroupements en classes.

Il y aura $(3 - 1)(5 - 1) = 8$ degrés de liberté.

Ce sont les effectifs attendus (calculés sous H_0) qui doivent être supérieurs à 5.

10-2 ❑ a) ☒ b) ❑ c) ❑ d) ☒ e)

On veut comparer la distribution observée et la distribution théorique obtenue en supposant une répartition uniforme entre les mois, soit un effectif de 100 pour chaque mois. Il s'agit donc d'un test d'ajustement.

Le degré de liberté sera $12 - 1 = 11$.

10-3 Il s'agit d'ajuster une répartition observée à une répartition théorique. C'est un test de conformité et on utilise un test du χ^2.

(H_0) : En deuxième génération, on a une répartition mendelienne des couleurs.

Calculs si (H_0) est vérifiée :

Couleurs	pourpre	rose	blanc lavande	blanc	totaux
O_i	1790	547	548	213	$3098 = n$
p_i	$\frac{9}{16}$	$\frac{3}{16}$	$\frac{3}{16}$	$\frac{1}{16}$	1
$C_i = np_i$	1742,625	580,875	580,875	193,875	3098

On a : $\chi_c^2 = \dfrac{(O_1 - C_1)^2}{C_1} + \cdots + \dfrac{(O_4 - C_4)^2}{C_4} \approx 7{,}06.$

Le nombre de degrés de liberté est $v = 4 - 1 = 3$.

Si $\alpha = 0{,}05$, on lit dans la table : $\chi^2_{0,05} = 7{,}81$.

Comme $7{,}06 < 7{,}81$, l'hypothèse (H_0)ne peut pas être rejetée au risque de 5 %.

10-4 On peut comparer la répartition des groupes sanguins sur la population malade et la population saine. Il s'agit de la comparaison d'une loi théorique (répartition sur la population saine) et d'une loi observée (répartition sur la population malade). On utilise un test du χ^2.

(H_0) : La répartition des groupes sanguins est la même dans les deux populations.

Calculs si (H_0) est vérifiée :

Groupes sanguins	*O*	*A*	*B*	*AB*	Totaux
O_i	104	76	18	2	$200 = n$
p_i	0,47	0,43	0,07	0,03	1
$C_i = np_i$	94	86	14	6	200

Comme tous les C_i sont supérieurs à 5, on calcule :

$$\chi_c^2 = \frac{(104-94)^2}{94} + \frac{(76-86)^2}{86} + \frac{(18-14)^2}{14} + \frac{(2-6)^2}{6} \approx 6{,}04.$$

Le nombre de degrés de liberté est $v = 4 - 1 = 3$.

Si $\alpha = 0{,}05$, on lit dans la table : $\chi^2_{0,05} = 7{,}81$.

Comme $7{,}06 < 7{,}81$, l'hypothèse (H_0) ne peut pas être rejetée au risque de 5 %. Donc, sur l'étude de cet échantillon, on ne peut pas dire que la présence de la maladie soit liée au groupe sanguin.

10-5 On va tester l'hypothèse (H_0) : la répartition de la radioactivité est due au hasard. Autrement dit, le taux de radioactivité ne dépend que de la surface de l'organite.

Si (H_0) est vérifiée, il y a donc conformité entre la distribution expérimentale (observée) et la distribution simulée (calculée). Comme tous les effectifs calculés sont supérieurs à 5, on peut calculer la distance :

$$\chi_c^2 = \frac{(30-10)^2}{10} + \cdots + \frac{(0-15)^2}{15} = 75.$$

Le nombre de degrés de liberté est $v = 6 - 1 = 5$.

Avec diverses valeurs de α, on lit : $\chi^2_{0,05} = 11{,}07$, $\chi^2_{0,01} = 15{,}09$, $\chi^2_{0,001} = 20{,}52$. Dans tous les cas, on a $\chi_c^2 > \chi_\alpha^2$.

Même au risque minime de 0,001, l'hypothèse (H_0) est rejetée. Le taux de radioactivité d'un organite n'est donc pas uniquement lié à sa surface.

10-6 Il s'agit d'un test de conformité entre une distribution expérimentale et une distribution théorique, que l'on réalise à l'aide d'un test du χ^2. Si X désigne la variable aléatoire étudiée, l'hypothèse nulle s'écrit : (H_0) : X suit une loi de Poisson $\mathcal{P}(\lambda)$.

Dans ce cas, $E(X) = V(X) = \lambda$.

$\mu = E(X)$ est estimée sans biais par $\overline{x} = 1{,}48$.

$V(X) = \sigma^2$ est estimée sans biais par $s^2 \approx 1{,}58$.

Comme $\overline{x}$ et s^2 sont proches, l'hypothèse (H_0) n'est pas stupide et on peut choisir 1,5 comme estimation de λ, ce qui conduit aux probabilités élémentaires et aux efffectifs théoriques :

E_i	0	1	2	3	4	5	≥ 6
O_i	37	46	39	19	5	3	1
p_i	0,2231	0,3347	0,2510	0,1255	0,0471	0,0141	0,0045
C_i	33,47	50,20	37,65	18,83	7,06	2,12	0,67

On a $p_i = P(X = i) = e^{-1,5}\dfrac{(1,5)^i}{i!}$ et $C_i = np_i$ avec $n = 150$.
Les deux dernières classes ont un effectif calculé inférieur à 5. Il faut donc regrouper les trois dernières classes et l'événement $X \geq 4$ a pour effectif obervé 9 et pour effectif calculé 9,85. D'où :

$$\chi_c^2 = \frac{(37 - 33,47)^2}{33,47} + \cdots + \frac{(9 - 9,85)^2}{9,85} \approx 0,85.$$

Il reste 5 événements et un paramètre a été estimé. Le nombre de degrés de liberté est donc $v = 3$.

Si $\alpha = 0,05$, on lit : $\chi^2_{0,05} = 7,81$.

Comme $0,85 < 7,81$, l'hypothèse (H_0)ne peut pas être rejetée.

10.7 Il s'agit de comparer les distributions observées sur deux échantillons.

(H_0) : le taux d'occupation est le même en janvier et en juillet. Les différences observées sont explicables par les fluctuations d'échantillonnage.

Calculs

Événements / Échantillons	Occupation		Inoccupation		Totaux
	observés	calculés	observés	calculés	
janvier	240	234	60	66	300
juillet	150	156	50	44	200
Totaux	390		110		500

Comme tous les effectifs calculés sont supérieurs à 5, on peut calculer la distance :

$$\chi_c^2 = \frac{(240 - 234)^2}{234} + \frac{(150 - 156)^2}{156} + \frac{(60 - 66)^2}{66} + \frac{(50 - 44)^2}{44}$$
$$\approx 1,75.$$

Le nombre de degrés de liberté est $v = (2 - 1)(2 - 1) = 1$. Si $\alpha = 0,05$, on lit $\chi^2_{0,05} = 3,84$.

Comme $1,75 < 3,84$, l'hypothèse (H_0)ne peut pas être rejetée.

10-8 Avec un test d'homogénéité du χ^2, on va tester l'hypothèse nulle :
(H_0) : les traitements *A* et *B* ont des effets identiques.

Calculs

Pour varier, les effectifs calculés dans chaque case seront mis en rouge.

Événements	Guérison	Amélioration	État stationnaire	Totaux
Traitement				
A	280 **300**	210 **180**	110 **120**	600
B	220 **200**	90 **120**	90 **80**	400
Totaux	500	300	200	1000

Comme tous les effectifs calculés sont supérieurs à 5, on peut calculer la distance :

$$\chi_c^2 = \frac{(280-300)^2}{300} + \cdots + \frac{(90-80)^2}{80} \approx 17{,}92.$$

Le nombre de degrés de liberté est $v = (3-1)(2-1) = 2$.

Si $\alpha = 0{,}05$, on lit $\chi^2_{0,05} = 5{,}99$.

Si $\alpha = 0{,}001$, on lit $\chi^2_{0,001} = 13{,}82$.

Dans tous les cas, on a $\chi_c > \chi^2_\alpha$ et (H_0) est rejetée même au risque très faible de 0,001.

On peut donc être persuadé que les traitements ont des effets différents.

10-9 On va tester l'hypothèse nulle :
(H_0) : les effets du traitement sont indépendants de l'âge du malade.

Il s'agit alors d'un test du χ^2 comme test d'indépendance de deux caractères qualitatifs. Mais on peut aussi considérer que les malades jeunes et les malades âgés conduisent à deux échantillons dont on teste leur homogénéité. Le point de vue est légèrement différent, mais les calculs sont les mêmes.

Les effectifs calculés dans chaque case seront mis en rouge.

Comme tous les effectifs calculés sont supérieurs à 5, on peut calculer la distance :

$$\chi_c^2 = \frac{(40-37{,}06)^2}{37{,}06} + \cdots + \frac{(50-47{,}06)^2}{47{,}06} \approx 0{,}84.$$

Résultats	Bons	Mauvais	Totaux
âge			
jeunes	40 **37,06**	30 **32,94**	70
âgés	50 **52,94**	50 **47,06**	100
Totaux	90	80	170

Le nombre de degrés de liberté est $v = (2 - 1)(2 - 1) = 1$.

Pour $\alpha = 0{,}10$, on lit $\chi^2_{0{,}10} = 2{,}71$.

Comme $0{,}84 < 2{,}71$, on ne peut pas rejeter (H_0)au risque de 10 %.

On accepte donc l'hypothèse qu'il n'existe pas de liaison entre l'âge du malade et l'effet du traitement.

Cet exercice peut aussi se faire en comparant deux fréquences observées (voir ex. 11.4).

10-10 a) Moyenne et écart type estimés

En assimilant chaque clase à son milieu, on obtient comme estimation de la moyenne μ de la population : $\overline{x} \approx 4{,}21$, et comme estimation de l'écart type σ de la population : $s \approx 2{,}86$.

b) Ajustement à une loi de Gauss

(H_0) : Le taux de protéines X se distribue de façon gaussienne.

En utilisant les estimations précédentes, on va utiliser un test du χ^2 pour juger la conformité entre la distribution observée et la loi normale $\mathcal{N}(4{,}21; 2{,}86)$.

Comme l'univers de la loi théorique est $\mathbb{R}$, les classes sont légèrement modifiées. D'autre part, comme d'habitude on ramène tous les calculs concernant X à la variable centrée réduite $U = \dfrac{X - \overline{x}}{s}$ qui suit $\mathcal{N}(0; 1)$.

Classes de X	Classes de U	p_i	C_i	O_i
]–∞ : 1,5]	]–∞ ; –0,95]	0,1711	6,16	8
]1,5 ; 3]	]–0,95 ; –0,42]	0,1661	5,98	7
]3 ; 4,5]	]–0,42 ; 0,10]	0,2026	7,29	4
]4,5 ; 6]	]0,10 ; 0,63]	0,1959	7,05	9
]6 ; 7,5]	]0,63 ; 1,15]	0,1392	5,01	2
]7,5 ; 9]	]1,15 ; 1,67]	0,0776	2,79	3
]9 ; +∞[	]1,67 ; +∞[	0,0475	1,71	3

Pour le calcul des p_i, revoyez si nécessaire le chapitre 7 ; par exemple :

$$P(3 < X \leq 4,5) = P(-0,42 < U \leq 0,10) = \Phi(0,10) - \Phi(-0,42)$$
$$= 0,5398 - 0,3372 = 0,2026.$$

Les deux dernières classes ayant des effectifs calculés inférieurs à 5, on regroupe les trois dernières classes, ce qui donne :

]6 ; +∞ [	]0,63 ; +∞ [	0,2643	9,51	8

On peut alors calculer la distance :

$$\chi^2_c = \frac{(8 - 6,16)^2}{6,16} + \cdots + \frac{(8 - 9,51)^2}{9,51} \approx 2,99.$$

Le nombre de degrés de liberté est $\nu = 5 - 1 - 2 = 2$ car il a fallu estimer deux paramètres.

Pour $\alpha = 0,05$, on lit $\chi^2_{0,05} = 5,99$.

On constate que $\chi^2_c < \chi^2_{0,05}$. Donc, au risque %, on ne rejette pas (H_0) et on peut admettre que le taux de protéines se distribue de façon gaussienne.

CHAPITRE 11

Comparaison de deux proportions

PLAN

11.1 Comparaison d'une proportion expérimentale et d'une proportion théorique

11.2 Comparaison de deux proportions expérimentales

11.3 Comparaison de deux proportions expérimentales (échantillons appariés)

OBJECTIFS

- Comparer la fréquence observée d'un événement bien précis à sa probabilité théorique
- Comparer les pourcentages observés d'un événement bien précis dans deux situations expérimentales
- Choisir entre un test bilatéral et un test unilatéral

11.1 COMPARAISON D'UNE PROPORTION EXPÉRIMENTALE ET D'UNE PROPORTION THÉORIQUE

Problématique

Dans une population, on étudie un caractère statistique à deux modalités A et $\overline{A}$. Chaque individu présente, ou non, la modalité A.

Soit π la proportion (ou la fréquence, ou le pourcentage) d'apparition de A dans la population, et p le pourcentage d'apparition de A observée dans un échantillon de taille n.

Le problème est de savoir si l'on peut considérer l'échantillon comme représentatif de la population, c'est-à-dire si la différence entre les valeurs numériques p et π est explicable par les aléas dus à l'échantillonnage.

Notons P la variable aléatoire qui prend la valeur p sur chaque échantillon de taille n (n est fixé et p varie d'un échantillon à l'autre).

L'hypothèse nulle (H_0) peut s'écrire :

La fréquence observée p est conforme à la fréquence théorique π.

Cas d'un grand échantillon

Théorème. Supposons que l'on puisse approximer la loi binomiale $\mathcal{B}(n, \pi)$ par une loi de Gauss, soit selon la convention retenue ici $n \geq 30$, $n\pi \geq 5$ et $n(1 - \pi) \geq 5$.

Alors, sous l'hypothèse (H_0), la variable aléatoire $Z = \dfrac{P - \pi}{\sqrt{\dfrac{\pi(1-\pi)}{n}}}$

suit à peu près la loi normale centrée réduite $\mathcal{N}(0, 1)$.

a) Calculs

On calcule la valeur prise par la variable aléatoire du théorème, soit le nombre $z = \dfrac{p - \pi}{\sqrt{\dfrac{\pi(1-\pi)}{n}}}$.

b) Décision dans le cas d'un test bilatéral

Dans ce cas (le plus courant), l'hypothèse alternative (H_1) est le contraire de (H_0), c'est-à-dire que la différence entre p et π est trop importante pour être explicable par les fluctuations d'échantillonnage.

On lit dans la table 2 le nombre z_α tel que $P(|Z| \geq z_\alpha) = \alpha$.

- Si $z \in]-z_\alpha, z_\alpha[$, l'hypothèse ($H_0$) ne peut pas être rejetée.
- Si $z \notin]-z_\alpha, z_\alpha[$, on écarte ($H_0$) avec une probabilité α de se tromper.

c) Décision dans le cas d'un test unilatéral

Supposons que la fréquence p observée sur l'échantillon soit a priori supérieure (ou inférieure) à la fréquence théorique π (par exemple, un médicament peut avoir une influence bénéfique ou être sans effet, mais il ne peut pas avoir un effet néfaste). Le signe de z est donc connu a priori. La zone de rejet de (H_0) est alors un intervalle situé d'un seul côté par rapport à 0.

Si, par exemple, $z > 0$, on lit dans la table 2 le nombre z_α tel que $P(Z \geq \upsilon_\alpha) = \alpha$, soit $P(|Z| \geq z_{2\alpha}) = \alpha$.

- Si $z < z_{2\alpha}$, l'hypothèse (H_0) ne peut pas être rejetée.
- Si $z \geq z_{2\alpha}$, on écarte (H_0) avec une probabilité α de se tromper.

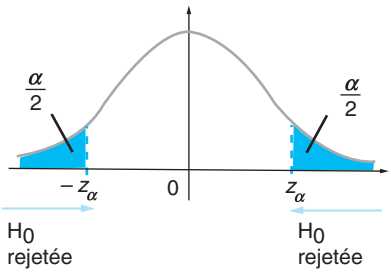


Figure 11-1

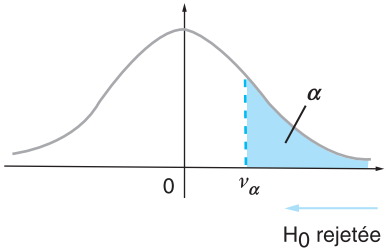


Figure 11-2

Vous pouvez aussi faire un test du χ^2 avec les événements A et $\overline{A}$, les effectifs observés $k = nf$ et $n - k$, et les effectifs théoriques np et $n - np$.

Mais les calculs sont plus longs, et, sans ordinateur, vous ne pouvez pas chercher la valeur frontière de α qui permet de rejeter (H_0), ni réaliser un test unilatéral.

11.2 COMPARAISON DE DEUX PROPORTIONS EXPÉRIMENTALES (échantillons indépendants)

Problématique

Dans deux populations $\mathcal{P}_1$ et $\mathcal{P}_2$, on étudie un caractère statistique à deux modalités A et $\overline{A}$. Chaque individu présente, ou non, la modalité A.

Les fréquences d'apparition de A dans les populations $\mathcal{P}_1$ et $\mathcal{P}_2$ sont les nombres (inconnus) π_1 et π_2.

De $\mathcal{P}_1$ et $\mathcal{P}_2$ on extrait deux échantillons E_1 et E_2, de tailles respectives n_1 et n_2, dans lesquels les fréquences d'apparition observées de A sont, respectivement, $p_1 = \dfrac{k_1}{n_1}$ et $p_2 = \dfrac{k_2}{n_2}$.

Le problème est de savoir si la différence entre p_1 et p_2 est significative, ou au contraire explicable par les hasards du tirage au sort.

Notons P_1 et P_2 les variables aléatoires qui prennent les valeurs p_1 et p_2 sur chaque échantillon de tailles n_1 et n_2.

L'hypothèse nulle peut s'écrire :

(H_0) : la différence entre p_1 et p_2 n'est pas significative ;

les valeurs théoriques sont égales, soit : $\pi_1 = \pi_2 = \pi$.

Cas de grands échantillons

Théorème. Supposons que l'on puisse approximer les lois binomiales par des lois normales, les conventions choisies étant :
$n_1 \geq 30$; $n_2 \geq 30$; $n_1 p_1 \geq 5$; $n_1(1 - p_1) \geq 5$; $n_2 p_2 \geq 5$; $n_2(1 - p_2) \geq 5$:
alors, sous l'hypothèse (H_0), la variable aléatoire :

$$Z = \frac{P_1 - P_2}{\sqrt{\frac{\pi(1-\pi)}{n_1} + \frac{\pi(1-\pi)}{n_2}}}$$

suit à peu près la loi normale centrée réduite.

a) Estimation de π

Sous l'hypothèse (H_0), on peut réunir les deux échantillons. On peut estimer π par la fréquence observée sur cette réunion :

$$\hat{\pi} = \frac{k_1 + k_2}{n_1 + n_2} = \frac{n_1 f_1 + n_2 f_2}{n_1 + n_2}$$

et le théorème reste inchangé en remplaçant π par $\hat{\pi}$.

b) Calculs

On calcule la valeur prise par la variable aléatoire du théorème :

$$z = \frac{p_1 - p_2}{\sqrt{\hat{\pi}(1-\hat{\pi})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}.$$

c) Décision dans le cas d'un test bilatéral

Dans ce cas (le plus courant), l'hypothèse alternative (H_1) est le contraire de (H_0), c'est-à-dire que la différence entre p_1 et p_2 est trop importante pour être explicable par les fluctuations d'échantillonnage.

On lit dans la table 2 le nombre z_α tel que $P(|Z| \geq z_\alpha) = \alpha$.

➤ Si $z \in]-z_\alpha, z_\alpha[$, l'hypothèse (H_0) ne peut pas être rejetée.

➤ Si $z \notin]-z_\alpha, z_\alpha[$, on écarte (H_0) avec une probabilité α de se tromper.

d) Décision dans le cas d'un test unilatéral

Si a priori $\pi_1 < \pi_2$ est impossible, l'hypothèse alternative est $\pi_1 > \pi_2$. Dans ce cas, on a toujours $z > 0$.

On détermine alors v_α tel que $P(Z \geq v_\alpha) = \alpha$, ce qui correspond à $v_\alpha = z_{2\alpha}$.

➤ Si $z < z_{2\alpha}$, l'hypothèse (H_0) ne peut pas être rejetée.

➤ Si $z > z_{2\alpha}$, on rejette (H_0) avec un risque d'erreur α.

Pour comparer deux fréquences expérimentales, on peut aussi utiliser un test d'homogénéité du χ^2.

11.3 COMPARAISON DE DEUX PROPORTIONS EXPÉRIMENTALES (échantillons appariés)

Problématique

Il s'agit encore de comparer deux proportions relatives à une modalité A. Mais ici, les modalités A et $\overline{A}$ sont appariées.

Les paires concordantes AA et $\overline{A}\overline{A}$ ne fournissent aucune information sur la différence des populations. On s'intéresse aux paires discordantes :

$A\overline{A}$ observée a fois ;

$\overline{A}A$ observée b fois.

Le test revient à comparer la proportion observée de $A\overline{A}$ (par exemple), soit $p = \dfrac{a}{a+b}$ à la proportion théorique $\pi = \dfrac{1}{2}$ qui découle de H_0.

Calculs

Dans les hypothèses d'approximation d'une loi binomiale par une loi normale, on calcule :

$$z = \frac{\dfrac{a}{a+b} - \dfrac{1}{2}}{\sqrt{\dfrac{1}{2} \times \dfrac{1}{2} \times \dfrac{1}{a+b}}} = \frac{a-b}{\sqrt{a+b}}.$$

Décision

Comme d'habitude quand on utilise la loi normale centrée réduite.

Remarque

Comme $Y = Z^2$ suit la loi du χ^2 à 1 degré de liberté, on peut aussi utiliser un test du χ^2 avec $y = z^2 = \dfrac{(a-b)^2}{a+b}$.

Comparaison de deux fréquences observées dans le cas de petits échantillons : test de Fisher

Avec les notations déjà utilisées, les effectifs connus peuvent se présenter en tableau :

	A	$\overline{A}$	**totaux**
E_1	k_1	$n_1 - k_1$	n_1
E_2	k_2	$n_2 - k_2$	n_2
totaux	k	$n-k$	n

avec $n = n_1 + n_2$ et $k = k_1 + k_2$.

On teste (H_0) : pas de différence significative entre $p_1 = \dfrac{k_1}{n_1}$ et $p_2 = \dfrac{k_2}{n_2}$.

Sous (H_0), en supposant les totaux fixes, la configuration du tableau précédent a pour probabilité :

$$\frac{n_1!\, n_2!\, k!\, (n-k)!}{k_1!\, k_2!\, (n_1-k_1)!\, (n_2-k_2)!\, n!}.$$

Pour réaliser le test, on cumule les probabilités des configurations (à totaux inchangés) au moins aussi défavorables à (H_0) que l'observation et on compare ce cumul au risque α.

Si la probabilité cumulée est inférieure à α, on rejette (H_0).

MOTS-CLÉS

- Comparaison à une probabilité
- Comparaison de deux fréquences dans la situation tout ou rien (l'événement étudié a lieu ou n'a pas lieu)

EXERCICES

11-1 Une nouvelle maladie est étudiée dans une région donnée où la population est également répartie entre hommes et femmes.
Lors d'une première étude, il a été constaté une prévalence de cette maladie plus élevée chez les femmes.
Une seconde étude est ensuite menée sur 225 personnes atteintes par cette maladie, choisies au hasard, parmi lesquelles 100 sont de sexe masculin.
On décide de tester, sur la base des données de cette étude, le caractère plus féminin de cette maladie.

❑ **a)** C'est un test de comparaison d'une proportion observée à une proposition théorique.

❑ **b)** C'est un test de comparaison de deux proportions observées.

❑ **c)** C'est un test unilatéral.

❑ **d)** On peut conclure au caractère plus féminin de cette maladie au seuil de 5% mais pas de 1%.

❑ **e)** Le seuil de signification du test est de 3 % environ.

11-2 La réaction du greffon contre l'hôte est la complication majeure des greffes de cellules souches hématopoiétiques. Une étude a pour objectif de comparer son incidence lors de greffes de moelle osseuse et lors de greffes de sang de cordon ombilical, deux sources différentes de cellules souches hématopoïétiques.
Sur 467 sujets ayant reçu une greffe de sang de cordon, 149 ont développé une réaction de rejet contre 360 sur 700 sujets ayant reçu une greffe de moelle osseuse.

❑ **a)** C'est un test de comparaison d'une proportion observée à une proposition théorique.

❑ **b)** C'est un test de comparaison de deux proportions observées.

❑ **c)** C'est un test unilatéral.

❑ **d)** Au risque $\alpha = 0{,}05$, une différence significative a été mise en évidence lors de greffes de sang de cordon et de moelle osseuse.

❑ **e)** Le seuil de signification du test est de 3 % environ.

11-3 Dans la population française, le pourcentage d'individus dont le sang est de rhésus négatif est de 15 %.
Dans un échantillon représentatif de 200 Basques français on observe que 44 personnes sont de rhésus négatif. Peut-on dire, au risque $\alpha = 0{,}05$, que les Basques diffèrent du reste de la France en ce qui concerne le caractère rhésus ?

11-4 Dans une population, le pourcentage d'individus présentant des rides est de 25 %. Sur 200 personnes ayant suivi un traitement anti-rides, on a observé que 40 personnes avaient des rides.
Au risque $\alpha = 0{,}05$, peut-on dire que le traitement est efficace ?

11-5 On sait qu'une maladie atteint 10 % des jeunes ovins d'une région donnée. Un chercheur a expérimenté un traitement sur un échantillon de n agneaux. Il a recensé alors 5 % de malades.
Déterminez la valeur minimale de n qui permette au chercheur de conclure à l'efficacité du traitement au risque $\alpha = 0{,}05$.

11-6 Reprenez les données de l'exercice **10-9** (test du χ^2 ; âge du malade et effet du traitement) et fournissez une deuxième solution.

11-7 Dans des services de maladies infectieuses, on observe des contaminations parmi les 2 100 employés qui constituent le personnel infirmier. On impose à 50 de ces personnes, tirées au hasard, des mesures de protection particulières et l'on observe alors chez elles, pendant une certaine période, 7 contaminations.
On choisit au hasard 50 employés non protégés. Pendant la même période, on note dans ce groupe 11 contaminations.
À quel risque α peut-on conclure à l'efficacité du dispositif de protection ? Commentez le résultat.

11-8 Une année, le taux de réussite nationale au baccalauréat dans une série donnée a été de 67 %.
Tous les tests qui suivent seront réalisés avec $\alpha = 0{,}05$.
a) Dans un centre d'examen A, il y a eu 216 reçus sur 300 candidats présentés. Les résultats de ce centre sont-ils conformes aux résultats nationaux ?
b) Dans un centre d'examen B de la même ville, il y a eu 128 reçus sur 200 candidats. Les résultats des centres A et B sont-ils significativement différents ?

11-9 On a réalisé une étude de la pratique du sport avant et après l'accouchement. On a obtenu les résultats suivants :
Nombre de femmes faisant du sport avant et faisant du sport après : 25.
Nombre de femmes ne faisant pas de sport avant et ne faisant pas de sport après : 35.
Nombre de femmes faisant du sport avant ne faisant pas de sport après : 25.
Nombre de femmes ne faisant pas de sport avant et faisant du sport après : 15.
On désire savoir si l'accouchement modifie la pratique du sport.

SOLUTIONS

11-1 ☒ **a)** ❑ **b)** ☒ **c)** ☒ **d)** ❑ **e)**

Il s'agit de comparer une proportion théorique de femmes $\pi = 0,5$ et une proportion observée $p = \frac{125}{225}$.

H_0: pas de différence significative entre π et p;
H_1: π et p significativement différents avec $f > p$.

L'échantillon étant de grande taille, on calcule $z = \frac{p - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \approx 1,667$.

Le test étant unilatéral :
pour un risque de 5 % on lit $z_{0,10} = 1,645$;
pour un risque de 1 % on lit $z_{0,02} = 2,326$.
La proposition **d.** est donc exacte.
Comme $z_{0,09} = 1,695$, le seuil de signification du test est compris entre 4,5% et 5 %.

11-2 ❑ **a)** ☒ **b)** ❑ **c)** ☒ **d)** ❑ **e)**

On teste H_0: pas différence entre les taux de rejet entre les deux sources.

Il s'agit de comparer deux proportions expérimentales $p_1 = \frac{149}{467} \approx 0,319$ et $p_2 = \frac{360}{700} \approx 0,514$ dans le cas d'échantillons indépendants de grandes tailles.
Le test est bilatéral car on ne sait rien a priori.
On calcule d'abord

la proportion de rejet sous H_0, soit : $\hat{p} = \frac{149 + 360}{467 + 700} = \frac{509}{1167} \approx 0,480$,

puis $z = \frac{p_1 - p_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \approx -6,581$.

On lit $z_{0,05} = 1,96$. Comme $|z| > z_{0,05}$, H_0 est rejetée au risque $\alpha = 0,05$.
Comme $z_{0,001} = 3,291$, H_0 est encore rejetée au risque $\alpha = 0,001$. Le seuil de signification est inférieur à 0, 1 %.

11-3 Il s'agit d'un test de conformité entre le pourcentage observé $p = 0,22$ et le pourcentage théorique $\pi = 0,15$.

(H_0): la différence observée entre p et π n'est pas significative ; elle est explicable par les aléas de l'échantillonnage.
On choisit un test bilatéral car il n'y a pas de raison a priori d'avoir $p > \pi$.
Comme $n = 200$ et $\pi = 0{,}15$, on peut approximer la loi binomiale $\mathcal{B}(n,\pi)$ par une loi de Gauss.

Dans ce cas, on calcule : $z = \dfrac{0{,}22 - 0{,}15}{\sqrt{\dfrac{0{,}15 \times 0{,}85}{200}}} \approx 2{,}77.$

Si $\alpha = 0{,}05$, on lit $z_\alpha = 1{,}96$. Comme $z \notin]-z_\alpha, z_\alpha[$, on écarte ($H_0$) et on conclut, au risque 5 %, que les Basques diffèrent du reste de la France en ce qui concerne le caractère rhésus.

11-4 Il s'agit d'un test de conformité entre la fréquence observée $p = 0{,}20$ et la fréquence théorique $\pi = 0{,}25$.
(H_0) : la différence observée entre p et π n'est pas significative.
Comme a priori on doit avoir $p < \pi$ (sinon il ne s'agit plus d'un traitement antirides), on choisit un test unilatéral.
Comme les conditions d'approximation de $\mathcal{B}(n,\pi)$par une loi normale sont satisfaites, on calcule : $z = \dfrac{0{,}20 - 0{,}25}{\sqrt{\dfrac{0{,}25 \times 0{,}75}{200}}} \approx -1{,}63.$

Pour $\alpha = 0{,}05$, le nombre υ_α tel que $P(Z \leq -\upsilon_\alpha) = 0{,}05$ correspond à $P(|Z| \geq \upsilon_\alpha) = 0{,}10$. On lit donc $\upsilon_{0{,}05} = z_{0{,}10} = 1{,}645$.

On constate que $z > -\upsilon_{0{,}05}$, donc l'hypothèse (H_0) ne peut pas être rejetée.
Au risque 5 %, on ne peut pas dire que le traitement est efficace.

Cette expérience n'a pas permis de mettre en évidence l'efficacité du traitement. Mais on ne sait pas si c'est à cause du traitement, ou si c'est l'expérience qui a été conduite sur un nombre trop limité de personnes.

11-5 Il s'agit d'un test de conformité entre la fréquence théorique $\pi = 0{,}10$ et la fréquence expérimentale $p = 0{,}05$.
(H_0) : le traitement n'est pas efficace.
En supposant que $n \geq 30$, sous (H_0), la variable aléatoire $Z = \dfrac{P - \pi}{\sqrt{\dfrac{\pi(1-\pi)}{n}}}$ suit la loi normale réduite.

D'autre part, on a nécessairement $p < \pi$ (si le traitement augmentait le pourcentage des malades, ce serait un curieux chercheur !). Il s'agit donc d'un test unilatéral.

Le nombre υ_α tel que $P(Z \leq -\upsilon_\alpha) = \alpha$ est $\upsilon_\alpha = z_{2\alpha}$, soit ici $\upsilon_{0,05} = z_{0,10} = 1{,}645$.

Le chercheur peut conclure à l'efficacité du traitement au risque 5 %, c'est-à-dire rejeter (H_0) si $z \leq -\upsilon_{0,05}$, soit :

$$\frac{0{,}05 - 0{,}10}{\sqrt{\dfrac{0{,}1 \times 0{,}9}{n}}} \leq -1{,}645 \Leftrightarrow \sqrt{\frac{0{,}1 \times 0{,}9}{n}} \leq \frac{0{,}05}{1{,}645} \Leftrightarrow n \geq 97{,}4.$$

La valeur minimum de n est de 98.

La valeur minimum obtenue concerne le nombre d'observations disponibles après expérience. Pour se prémunir de pertes pendant l'expérience, le chercheur lancera son expérience avec un nombre un peu plus élevé, par exemple 100.

11-6 On peut comparer les deux fréquences observées $p_1 = \dfrac{40}{70}$ et $p_2 = \dfrac{50}{100}$.

Pour ceci, on va tester l'hypothèse nulle : (H_0) $\pi_1 = \pi_2 = \pi$, c'est-à-dire : les effets du traitement sont les mêmes dans la population des malades jeunes et dans la population des malades âgés.

Comme on n'a pas de raison a priori de privilégier une population, on choisit un test bilatéral.

Sous (H_0), π est estimé par $\hat{\pi} = \dfrac{40 + 50}{70 + 100} = \dfrac{9}{17}$.

Les conditions d'approximation des lois binomiales par des lois normales sont vérifiées. On calcule donc :

$$z = \frac{p_1 - p_2}{\sqrt{\hat{\pi}(1 - \hat{\pi})\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}} = \frac{\dfrac{4}{7} - 0{,}5}{\sqrt{\dfrac{9}{17} \times \dfrac{8}{17}\left(\dfrac{1}{70} + \dfrac{1}{100}\right)}} \approx 0{,}92.$$

Pour $\alpha = 0{,}10$, on lit $z_{0,10} = 1{,}645$. Comme $z \in]-z_{0,10}, z_{0,10}[$, on ne rejette pas (H_0) au risque choisi de 10 %.

La conclusion est la même que dans la solution de l'exercice 10-7.

Mais si la question était : « avec quel risque minimum peut-on dire qu'il existe une liaison entre l'âge du malade et l'effet du traitement ? », alors, en l'absence d'ordinateur, il fallait choisir la comparaison de fréquences. En effet, comme $z_{0,35} = 0,935$ et $z_{0,36} = 0,915$, la réponse est 36 % et il vous reste à trouver ce risque beaucoup trop élevé pour une telle affirmation.

11-7 Il s'agit d'une comparaison de deux pourcentages observés de l'événement A « être contaminé ». On dispose des informations : échantillon E_1 (avec le nouveau dispositif de protection)

$$n_1 = 50 \; ; \; k_1 = 7 \; ; \; p_1 = 0{,}14$$

échantillon E_2 (sans le nouveau dispositif de protection)

$$n_2 = 50 \; ; \; k_2 = 11 \; ; \; p_2 = 0{,}22$$

La taille de la population est telle que l'on peut assimiler les tirages à des tirages avec remise. La situation est unilatérale a priori et on va tester :

(H_0) : $\pi_1 = \pi_2 = \pi$; le dispositif de protection n'est pas efficace.

(H_1) : $\pi_1 < \pi_2$; le dispositif de protection est efficace.

On cherche à quel risque minimum α on peut rejeter (H_0) au bénéfice de (H_1).

Sous (H_0), π est estimé par $\hat{\pi} = \dfrac{7 + 11}{50 + 50} = 0{,}18$.

Les conditions d'approximation de lois binomiales par des lois normales étant réunies, la valeur :

$$z = \frac{p_1 - p_2}{\sqrt{\hat{\pi}(1 - \hat{\pi})\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}} = \frac{0{,}14 - 0{,}22}{\sqrt{0{,}18(1 - 0{,}18)\left(\dfrac{1}{50} + \dfrac{1}{50}\right)}}$$

$$\approx -1{,}041$$

est une réalisation d'une loi normale centrée réduite.

Avec la table 2, on lit $z_{0{,}29} = 1{,}058$ et $z_{0{,}30} = 1{,}036$.

Comme il s'agit d'un test unilatéral, on obtient ainsi $2\alpha \approx 0{,}30$, soit un risque de 15 %.

On peut aussi utiliser la table 1 de la fonction de répartition de $\mathcal{N}(0, 1)$ (avec un dessin pour mieux comprendre) : $\Phi(1, 04) = 0,8508 = 1 - \alpha$.

Le risque obtenu étant élevé, on peut penser que l'information disponible ne permet pas de conclure à l'efficacité du dispositif de protection.

11-8 **a)** Il s'agit d'un test de conformité entre la fréquence observée sur l'échantillon $p = \dfrac{216}{300} = 0{,}72$ et la fréquence observée sur la population $\pi = 0{,}67$. On va tester l'hypothèse nulle :

(H_0) : la différence observée entre p et π n'est pas significative.

Comme les conditions d'approximation de $\mathcal{B}(n,\pi)$ par une loi normale sont satisfaites, on calcule : $z = \dfrac{0{,}72 - 0{,}67}{\sqrt{\dfrac{0{,}67 \times 0{,}33}{300}}} \approx 1{,}84$.

Pour $\alpha = 0{,}05$, on lit $z_\alpha = 1{,}96$.

Comme $z \in]-z_\alpha, z_\alpha[$, on ne peut pas rejeter (H_0).

Les résultats du centre *A* ne diffèrent pas significativement des résultats nationaux.

Et pourtant, que de cocoricos dans la presse locale quand il y a 72 % de succès et seulement 67 % dans le pays.

b) Il s'agit de comparer les deux fréquences expérimentales :

$$p_A = \frac{216}{300} = 0{,}72 \quad \text{et} \quad p_B = \frac{128}{200} = 0{,}64.$$

On va tester l'hypothèse nulle :

(H_0) : la différence observée entre p_A et p_B n'est pas significative ; elle est explicable par les aléas dus à l'échantillonnage.

Si (H_0) est vérifiée, on peut réunir les deux échantillons *A* et *B*, ce qui conduit à obtenir $216 + 128 = 344$ reçus sur $300 + 200 = 500$ candidats.

Donc $\hat{\pi} = \frac{344}{500} = 0{,}688$.

Les conditions d'approximation de lois binomiales par des lois normales étant réunies, on sait que, sous (H_0) la variable aléatoire :

$$Z = \frac{P_A - P_B}{\sqrt{\hat{\pi}(1-\hat{\pi})\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}} \text{ suit à peu près la loi normale centrée réduite.}$$

La valeur prise par cette variable aléatoire est : $z \approx 1{,}89$.

Pour $\alpha = 0{,}05$, on lit dans la table 2 : $z_\alpha = 1{,}96$.

Comme $z \in]-z_\alpha, z_\alpha[$, l'hypothèse (H_0) ne peut pas être rejetée. La différence entre les centres *A* et *B* n'est pas significative au risque 5 %.

Et pourtant, que de protestations à prévoir quand il y a 72 % de succès dans un centre et 64 % dans l'autre.

11-9 Il s'agit de séries appariées. On s'intéresse aux paires discordantes en nombres 25 et 15.

• Avec un test de l'écart réduit

$p = \frac{25}{40} = 0{,}625$; $\pi = 0{,}5$; $z = \frac{p - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \approx 1{,}58$. Comme $|z| < 1{,}96$,

H_0 est non rejetée au seuil de 5 %.

• Avec un test du χ^2

	O_i	p_i	C_i
Sport-pas sport	25	0,5	20
Pas sport-sport	15	0,5	20
Totaux	40	1	40

$\chi^2_c = \dfrac{(25-20)^2}{20} + \dfrac{(15-20)^2}{20} = 2{,}5$. Comme $\chi^2_{0,05} = 3{,}84$, H_0 est non rejetée au seuil de 5 %.

Si vous avez fait le calcul de z, on a directement $\chi^2_c = z^2 = 2{,}5$.

CHAPITRE 12

Comparaison de deux moyennes, de deux variances

PLAN

12.1 Comparaison d'une moyenne expérimentale et d'une moyenne théorique

12.2 Comparaison de deux moyennes expérimentales dans le cas d'échantillons indépendants

12.3 Comparaison de deux moyennes expérimentale dans le cas d'échantillons appariés

12.4 Comparaison d'une variance expérimentale et d'une variance théorique

12.5 Comparaison de deux variances expérimentales

OBJECTIFS

- Décider, à partir d'une prélévement limité, si une production respecte une norme
- Distinguer les cas de deux échantillons indépendants et de deux échantillons appariés
- Savoir si un traitement est actif en comparant son effet moyen à l'effet moyen observé sans traitement
- Savoir choisir entre un test bilatéral et un test unilatéral
- Savoir si une méthode de dosage est fiable par la régularité des résultats qu'elle donne

12.1 COMPARAISON D'UNE MOYENNE EXPÉRIMENTALE ET D'UNE MOYENNE THÉORIQUE

Problématique

Soit X une variable aléatoire avec $E(X) = \mu$ et $V(X) = \sigma^2$.

Le caractère quantitatif X est observé sur un échantillon de taille n. Les mesures obtenues ont pour moyenne $\overline{x}$ et pour variance estimée s^2.

Le problème est de savoir si la différence constatée entre μ et $\overline{x}$ est explicable par les fluctuations d'échantillonnage. Les hypothèses à testersont :

(H_0) l'échantillon est extrait au hasard de la population ; sa moyenne $\overline{x}$ est conforme à la moyenne μ de la population.

Pour l'hypothèse alternative, on choisira parmi les deux possibilités :

- (H_1) $\overline{x}$ n'est pas conforme à μ (test bilatéral) ;
- (H_2) $\overline{x}$ n'est pas conforme à μ avec à priori $\overline{x}$ supérieur (ou inférieur) à μ (test unilatéral).

Cas d'un grand échantillon ($n > 30$)

Théorème. Dans le cas $n > 30$, $Z = \dfrac{\overline{X} - \mu}{\dfrac{s}{\sqrt{n}}}$ suit à peu près la loi normale centrée réduite.

Soit z la valeur prise par la variable aléatoire Z.

a) Décision dans le cas d'un test bilatéral

Le risque de première espèce α étant fixé, on lit dans la table 2 la borne z_α telle que

$P(|Z| \geq z_\alpha) = \alpha$.

- Si z appartient à la zone en blanc, l'hypothèse (H_0) ne peut pas être rejetée.
- Si z appartient à la zone tramée, on écarte (H_0) avec une probabilité α de se tromper.

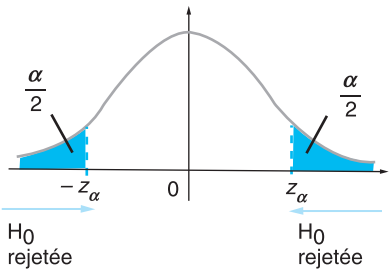


Figure 12-1

b) Décision dans le cas d'un test unilatéral ($z > 0$ par exemple)

Le risque de première espèce α étant fixé, on détermine v_α tel que $P(Z \geq v_\alpha) = \alpha$, ce qui correspond à $P(|Z| \geq v_\alpha) = 2\alpha$, c'est-à-dire $v_\alpha = z_{2\alpha}$

- Si z appartient à la zone en blanc, l'hypothèse (H_0) ne peut pas être rejetée.

➤ Si z appartient à la zone tramée, on écarte (H_0) avec une probabilité α de se tromper.

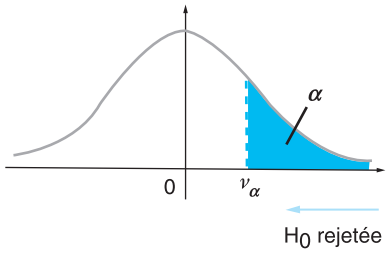


Figure 12-2

Cas d'un petit échantillon et d'une population gaussienne

Théorème 1. Si X suit une loi normale, sous l'hypothèse (H_0), $\dfrac{\overline{X} - \mu}{\dfrac{\sigma}{\sqrt{n}}}$ suit à peu près la loi normale centrée réduite.

Si σ est connu, le test se construit comme dans le cas d'un grand échantillon. Mais en général σ est inconnu et estimé par s. Dans le cas d'un petit échantillon, en remplaçant σ par s, on modifie la loi suivie par $\overline{X}$.

Théorème 2. Sous (H_0), la variable aléatoire $T = \dfrac{\overline{X} - \mu}{\dfrac{s}{\sqrt{n}}}$ suit la loi de Student à $n - 1$ degrés de liberté.

On calcule la valeur prise par T, soit $t = \dfrac{\overline{x} - \mu}{\dfrac{s}{\sqrt{n}}}$.

a) Décision dans le cas d'un test bilatéral

Le risque de première espèce α étant fixé, et le nombre de degrés de liberté étant connu, on lit dans la table 3 le nombre t_α tel que $P(|T| \geq t_\alpha) = \alpha$.

➤ Si t appartient à la zone en blanc, l'hypothèse (H_0) ne peut pas être rejetée.

➤ Si t appartient à la zone tramée, on écarte (H_0) avec un risque α de se tromper.

b) Décision dans le cas d'un test unilatéral (cas $t > 0$)

Le risque de première espèce α étant fixé, et le nombre de degrés de liberté étant connu, on lit dans la table 3 le nombre r_α tel que $P(T \geq r_\alpha) = \alpha$, ce qui correspond à $r_\alpha = t_{2\alpha}$.

- Si $t < r_\alpha$, l'hypothèse (H_0) ne peut pas être rejetée.
- Si $t \geq r_\alpha$, on rejette (H_0) avec un risque α de se tromper.

Ce cas est donc très proche du cas précédent. Il suppose une hypothèse supplémentaire (population gaussienne) et la lecture de la borne de décision se fait dans la table 3 au lieu de la table 2.

12.2 COMPARAISON DE DEUX MOYENNES EXPÉRIMENTALES DANS LE CAS D'ÉCHANTILLONS INDÉPENDANTS

Problématique

Dans deux populations P_1 et P_2, on étudie une variable aléatoire X. On note :

– μ_1 et σ_1 la moyenne et l'écart type de X dans P_1,

– μ_2 et σ_2 la moyenne et l'écart type de X dans P_2.

Tous ces nombres sont inconnus.

De P_1, on extrait un échantillon E_1, de taille n_1, pour lequel on calcule sa moyenne $\overline{x}_1$ et son écart type estimé s_1.

De P_2, on extrait un échantillon E_2, de taille n_2, pour lequel on calcule sa moyenne $\overline{x}_2$ et son écart type estimé s_2.

Les échantillons sont supposés indépendants. Le problème est de savoir si la différence entre les moyennes expérimentales $\overline{x}_1$ et $\overline{x}_2$ est significative, ou au contraire explicable par les fluctuations d'échantillonnage.

Les hypothèses à tester sont :

(H_0) : $\mu_1 = \mu_2$, c'est-à-dire P_1 et P_2 sont homogènes, ou encore la différence entre $\overline{x}_1$ et $\overline{x}_2$ n'est pas significative.

Pour l'hypothèse alternative, on choisira entre les deux possibilités :

– (H_1) : $\mu_1 \neq \mu_2$ (test bilatéral),

– (H_2) : $\mu_1 > \mu_2$ (ou $\mu_1 < \mu_2$) si le signe de $\mu_1 - \mu_2$ est connu a priori (test unilatéral).

Cas de deux grands échantillons ($n_1 \geq 30$ et $n_2 \geq 30$)

Théorème. Sous (H_0), la variable aléatoire $Z = \dfrac{\overline{X}_1 - \overline{X}_2}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$ suit à peu près $\mathcal{N}(0,1)$.

Ici la conclusion est inchangée quand on remplace les valeurs inconnues σ_1^2 et σ_2^2 par les valeurs estimées s_1^2 et s_2^2. On calcule donc $z = \dfrac{\overline{x}_1 - \overline{x}_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$.

a) Décision dans le cas d'un test bilatéral

α étant fixé, on lit z_α dans la table 2.

- Si z appartient à la zone en blanc, l'hypothèse (H_0) ne peut pas être rejetée.
- Si z appartient à la zone tramée, on écarte (H_0) avec un risque α de se tromper.

b) Décision dans le cas d'un test unilatéral (cas $u > 0$)

α étant fixé, la table 2 nous donne $v_\alpha = z_{2\alpha}$.

- Si $z < v_\alpha$, l'hypothèse (H_0) ne peut pas être rejetée.
- Si $z \geq v_\alpha$, on rejette (H_0) avec un risque α de se tromper.

Cas de petits échantillons ($n_1 < 30$ ou $n_2 < 30$) extraits de populations gaussiennes

Théorème. Sous (H_0), si X suit une loi normale dans P_1 et P_2, et si $\sigma_1^2 = \sigma_2^2 = \sigma^2$, alors $T = \dfrac{\overline{X}_1 - \overline{X}_2}{\sqrt{\sigma^2\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$ suit la loi de Student à $n_1 + n_2 - 2$ degrés de liberté.

a) Calculs intermédiaires

Pour utiliser ce théorème, il faut d'abord tester l'égalité des deux variances (*cf.* paragraphe **12.5**).

Si l'hypothèse $\sigma_1^2 = \sigma_2^2 = \sigma^2$ est retenue, cette valeur commune σ^2 est alors estimée par $\hat{\sigma}^2 = \dfrac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$.

Le théorème continue à être à peu près vrai et on calcule :

$$t = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\hat{\sigma}^2\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}.$$

b) Décision dans le cas d'un test bilatéral

α étant fixé et le nombre de degrés de liberté étant connu, on lit t_α dans la table 3.

- Si t appartient à la zone en blanc, l'hypothèse (H_0) ne peut pas être rejetée.
- Si t appartient à la zone tramée, on écarte (H_0) avec un risque α de se tromper.

c) Décision dans le cas d'un test unilatéral (cas $t > 0$)

α étant fixé et le nombre de degrés de liberté étant connu, la table 3 nous donne $r_\alpha = t_{2\alpha}$

- Si $t < r_\alpha$, l'hypothèse (H_0) ne peut pas être rejetée.
- Si $t \geq r_\alpha$, on rejette (H_0) avec un risque α de se tromper.

12.3 COMPARAISON DE DEUX MOYENNES EXPÉRIMENTALES DANS LE CAS D'ÉCHANTILLONS APPARIÉS

Problématique

Deux échantillons sont dits appariés lorsque chaque valeur $x_{i,1}$ de E_1 est associée à une valeur $x_{i,2}$ de E_2 (appariés = associés par paires), par exemple E_1 peut être un groupe de malades avant un traitement et E_2 le groupe des *mêmes* malades après traitement.

Deux échantillons appariés ont donc la même taille *n*, ce qui est une condition nécessaire mais non suffisante.

Le problème est de savoir si la différence entre les moyennes $\overline{x}_1$ et $\overline{x}_2$ des échantillons est explicable par les fluctuations d'échantillonnage.

Les hypothèses à tester sont :

$(H_0) : \mu_1 = \mu_2$

$(H_1) : \mu_1 \neq \mu_2$ si le test est bilatéral ;

$\mu_1 > \mu_2$ (ou $\mu_1 < \mu_2$) si le test est unilatéral.

Mise en place du test

On calcule les n différences $d_i = x_{i1} - x_{i2}$.

L'échantillon $\{d_1, \ldots, d_n\}$ a pour moyenne $\overline{d}$ et pour écart type estimé s_d.

Sous l'hypothèse (H_0), la variable aléatoire $\overline{D} = \overline{X}_1 - \overline{X}_2$ doit avoir une moyenne nulle. On est ainsi ramené à la comparaison d'une moyenne expérimentale $\overline{d}$ et d'une moyenne théorique $\mu = 0$.

- Si $n > 30$, on sait que $Z = \dfrac{\overline{D}}{s_d}\sqrt{n}$ suit $\mathcal{N}(0,1)$.

  On calcule donc la valeur z prise par Z et on la compare à la borne z_α lue dans la table 2.

- Si $n \leq 30$ et si la population des différences est gaussienne, on sait que $T = \dfrac{\overline{D}}{s_d}\sqrt{n}$ suit la loi de Student avec $v = n - 1$.

  On calcule donc la valeur t prise par T et on la compare à la borne t_α lue dans la table 3.

- Si $n \leq 30$ et si les lois ne sont pas connues, on utilise le test de Wilcoxon (*cf.* chapitre 16).

12.4 COMPARAISON D'UNE VARIANCE EXPÉRIMENTALE ET D'UNE VARIANCE THÉORIQUE

Problématique

Avec les mêmes notations que dans la problématique du **12.1**, le problème est ici de savoir si l'échantillon est représentatif de la population en ce qui concerne la régularité des mesures, c'est-à-dire si la différence constatée entre σ^2 et s^2 est explicable par les aléas dus à l'échantillonnage.

L'hypothèse nulle peut s'écrire :

(H_0) : l'échantillon est extrait au hasard de la population ; sa variance estimée s^2 est conforme à la variance σ^2 de la population, c'est-à-dire que la différence des valeurs numériques n'est pas significative.

Théorème. Si X suit une loi normale, sous l'hypothèse (H_0), la variable aléatoire $Y = \dfrac{n-1}{\sigma^2} S^2$ suit la loi du χ^2 à $n-1$ degrés de liberté.

Utilisation dans le cas $n \leq 31$

a) Calculs

On calcule la valeur prise par la variable aléatoire du théorème $y = \dfrac{n-1}{\sigma^2} s^2$

b) Décision

Le risque de première espèce α étant fixé, et le nombre de degrés de liberté étant connu, la table 4 permet de déterminer les nombres a et b tels que :

$$P(Y \geq b) = \frac{\alpha}{2} \text{ et } P(Y \leq a) = \frac{\alpha}{2} \text{ soit } P(Y \geq a) = 1 - \frac{\alpha}{2}$$

- Si $y \in]a,b[$, l'hypothèse (H_0) ne peut pas être rejetée.
- Si $y \notin]a,b[$, on rejette (H_0) avec un risque α de se tromper.

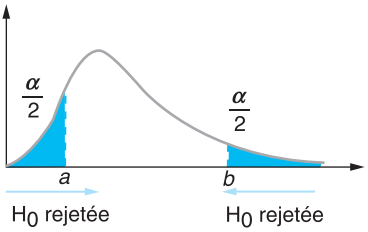


Figure 12-3

Utilisation dans le cas $n > 31$

Ce cas disparaît si vous êtes hors situation de contrôle scolaire et si vous disposez d'un ordinateur et d'un logiciel de statistiques.

Théorème. Si Y est une variable aléatoire qui suit une loi du χ^2 à ν degrés de liberté et si $\nu > 30$, alors la variable aléatoire $Z = \sqrt{2Y} - \sqrt{2\nu - 1}$ suit à peu près la loi $\mathcal{N}(0, 1)$.

a) Calculs

On calcule la valeur prise par Z, soit : $z = \sqrt{\dfrac{2(n-1)}{\sigma^2}s^2} - \sqrt{2n-3}$.

b) Décision

Le risque de première espèce α étant fixé, on lit dans la table 2 le nombre z_α tel que $P(|Z| \geq z_\alpha) = \alpha$.

- Si $z \in]-z_\alpha, z_\alpha[$, l'hypothèse (H_0) ne peut pas être rejetée.
- Si $z \notin]-z_\alpha, z_\alpha[$, on rejette (H_0) avec un risque α de se tromper.

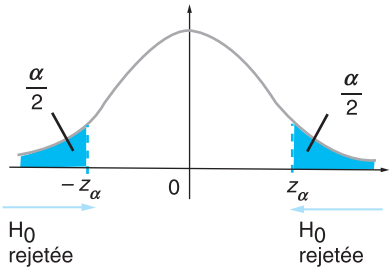


Figure 12-4

12.5 COMPARAISON DE DEUX VARIANCES EXPÉRIMENTALES

Problématique

Avec les mêmes notations que dans la problématique du **12.2**, le problème est ici de savoir si la différence entre s_1^2 et s_2^2 est significative, ou au contraire explicable par les fluctuations d'échantillonnage.

L'hypothèse nulle est : $(H_0) : \sigma_1^2 = \sigma_2^2$.

Mise en place du test et décision

Théorème. Si les deux populations sont gaussiennes, sous l'hypothèse (H_0), la variable aléatoire $F = \dfrac{S_1^2}{S_2^2}$ suit la loi de Snedecor à $(n_1 - 1, n_2 - 1)$ degrés de liberté.

a) Lois de Snedecor

Une loi de Snedecor est une loi de probabilité continue dont la densité est nulle pour $x < 0$, et dépend de deux paramètres appelés degrés de liberté (attention, ces paramètres sont ordonnés, en les permutant on obtient une autre loi de Snedecor).

Le risque de première espèce α étant fixé, les tables permettent de déterminer f'_α et f_α tels que :

$P(f'_\alpha < F < f_\alpha) = 1 - \alpha$

avec $P(F \leq f'_\alpha) = \dfrac{\alpha}{2}$

et $P(F \geq f_\alpha) = \dfrac{\alpha}{2}$.

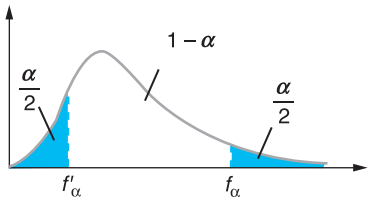


Figure 12-4

En fait, les tables fournissent seulement f_α. On peut obtenir f'_α en sachant que si F suit la loi de Snedecor à $(n_1 - 1, n_2 - 1)$ degrés de liberté, alors $\dfrac{1}{F}$ suit la loi de Snedecor à $(n_2 - 1, n_1 - 1)$ degrés de liberté. Mais la connaissance de f'_α n'est pas nécessaire avec la règle de décision qui suit.

b) Règle de décision

Comme les tables de F ne comportent que des valeurs supérieures à 1, on permute si nécessaire les deux échantillons de sorte que $\dfrac{s_1^2}{s_2^2} \geq 1$.

Attention à permuter les degrés de liberté si vous avez permuté les variances estimées.

Puis on compare $\dfrac{s_1^2}{s_2^2}$ à f_α.

- Si $\dfrac{s_1^2}{s_2^2} < f_\alpha$, on ne peut pas rejeter (H_0).
- Si $\dfrac{s_1^2}{s_2^2} \geq f_\alpha$, on rejette (H_0) avec une probabilité α de se tromper.

Attention, comparez les graphiques du test et la légende graphique des tables disponibles.
Si le risque du test est $\alpha = 0{,}05$, vous devez lire dans la table 5 où la surface à droite de la borne de décision est 0,025.
La loi de Snedecor a d'autres usages (*cf.* chap. 13), ce qui explique la distorsion apparente entre le risque de ce test et le titre de la table.

Choisir a priori des échantillons indépendants ou appariés

Problématique de l'expérimentaliste

Pour étudier l'influence d'un traitement sur la moyenne d'un caractère numérique, on utilise un échantillon témoin (malades avec placebo) et un échantillon traité (malades recevant le principe actif).

Quelles raisons peuvent conduire à choisir a priori des échantillons indépendants ou des échantillons appariés ? et quelles sont les précautions expérimentales recommandées ?

Éléments de réponse

On choisit les mêmes individus (échantillons appariés) si l'on pense qu'il peut y avoir une variabilité des réactions individuelles qui perturberait l'étude du traitement.

Si les individus sont considérés comme interchangeables, on prend des échantillons indépendants.

Dans le premier cas, l'ordre de passage (médicament actif, placebo) doit être tiré au sort pour chaque individu.

L'administration se fait à l'aveugle (infirmier(e) non informé(e)) et avec un délai suffisant.

Dans le second cas, les échantillons sont constitués de façon aléatoire, c'est-à-dire que les individus sont tirés au sort.

MOTS-CLÉS

- Comparaison à une norme
- Échantillons indépendants
- Échantillons appariés
- Lois de Snedecor

EXERCICES

12-1 On a mesuré les dimensions d'une tumeur cérébrale chez des patients traités ou non avec une substance antitumorale. On a obtenu les résultats suivants:
Patients témoins : $n_1 = 20$; $\overline{x}_1 = 7{,}075$ cm^2 ; $s_1 = 0{,}576$ cm^2.

Patients traités : $n_2 = 18$; $\overline{x}_2 = 5{,}850$ cm^2 ; $s_2 = 0{,}614$ cm^2.
On aimerait savoir si le traitement est efficace.

❒ **a)** On utilise un test de comparaison de moyennes sur séries appariées.

❒ **b)** On utilise un test de Student.

❒ **c)** Au risque de 5%, on ne montre pas de différence significative entre les variances.

❒ **d)** Le paramètre statistique calculé s'élève à 6,345 (à 0,001près).

❒ **e)** L'hypothèse alternative qui doit être posée est ici unilatérale car on s'attend à une efficacité a priori du traitement antitumoral.

QCM n°2 et 3. *On souhaite vérifier si deux méthodes de dosage d'une substance dans un laboratoire donnent en moyenne des résultats différents. Pour cela, 10 solutions sont dosées successivement par chacune des méthodes. On admet que le dosage par l'une des méthodes n'altère pas la solution. On obtient les mesures suivantes (en mol/L):*

Solution	1	2	3	4	5	6	7	8	9	10
Méthode 1	1, 9	2, 0	1, 5	1, 4	1, 7	1, 7	1, 5	1, 8	1, 9	1, 3
Méthode 2	1, 8	1, 6	1, 2	1, 6	1, 7	1, 5	1, 0	1, 6	1, 6	1, 1

12-2 Pour comparer les mesures moyennes obtenues par les deux méthodes, on souhaite utiliser un test paramétrique. Parmi les propositions suivantes relatives à la condition nécessaire à l'utilisation d'un tel test, quelle est celle qui est correcte?

❒ **a)** Cette comparaison ne nécessite aucune condition particulière.

❒ **b)** Cette comparaison nécessite la normalité de la mesure pour chacune des méthodes.

❒ **c)** Cette comparaison nécessite l'égalité des variances des deux mesures.

❒ **d)** Cette comparaison nécessite la normalité de la différence entre la mesure obtenue par la méthode 1 et celle obtenue par la méthode 2.

❒ **e)** Cette comparaison nécessite la normalité de la moyenne des deux mesures.

12-3 La condition précédente étant supposée remplie, le test paramétrique conduit à:

❒ **a)** Ne pas rejeter l'égalité des mesures moyennes, au seuil de 10 %.

❒ **b)** Rejeter l'égalité des mesures moyennes, au seuil de 10 % mais pas de 5 %.

❒ **c)** Rejeter l'égalité des mesures moyennes, au seuil de 5 % mais pas de 1 %.

❒ **d)** Rejeter l'égalité des mesures moyennes, au seuil de 1 % mais pas de 0,1 %.

❒ **e)** Rejeter l'égalité des mesures moyennes, au seuil de 0,1 %.

12-4 Les spécifications d'un certain médicament indiquent que chaque comprimé doit contenir 2,5 g de substance active.

100 comprimés sont choisis au hasard dans la production, puis analysés.

Ils contiennent en moyenne 2,6 g de substance active, avec un écart type estimé $s = 0{,}4$ g.

Peut-on dire que le médicament respecte les spécifications ($\alpha = 0{,}05$) ?

12-5 À la suite d'un traitement sur une variété de rongeurs, on prélève un échantillon de 5 animaux et on les pèse. On obtient les poids en g :

$$83 ; 81 ; 84 ; 80 ; 85.$$

À la même époque un grand nombre de mesures a permis d'établir que les rongeurs non traités avaient un poids moyen de 87,6 g.

Le poids moyen des rongeurs traités diffère-t-il significativement de cette norme au seuil 5 % ? On suppose que le poids des rongeurs suit une loi normale.

12-6 On a prélevé deux échantillons de pommes pour en étudier le poids.

Le premier, en début de récolte, a pour taille 100, pour moyenne 120 g et pour écart type estimé 20 g.

Le second, en fin de récolte, a pour taille 150, pour moyenne 150 g et pour écart type estimé 10 g.

La différence entre les poids moyens à ces deux époques différentes de la récolte est-elle significative,

12-7 Pour déterminer le poids moyen d'épis de blé appartenant à deux variétés, on a procédé à dix pesées pour chaque variété. Les moyennes obtenues ont été : $\overline{x}_1 = 170{,}7$cg et $\overline{x}_2 = 168{,}5$ cg.

On admet que le poids de ces graines est distribué dans chaque variété suivant une loi de Gauss et que les variances de deux distributions peuvent être considérées comme égales. Les estimations obtenues sur chaque échantillon sont : $s_1^2 = 432{,}9$ et $s_2^2 = 182{,}7$.

La différence des moyennes est-elle significative au risque $\alpha = 0{,}05$?

12-8 Chez un groupe de 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).

Sujet n°	1	2	3	4	5	6	7	8	9	10
Avant traitement	15	18	17	20	21	18	17	15	19	16
Après traitement	12	16	17	18	17	15	18	14	16	18

Le traitement a-t-il une action significative, au risque 5 % ? On supposera que la variable aléatoire égale à la différence des tensions artérielles suit une loi normale.

12-9 On se demande si la densité de l'écorce d'un chêne-liège est la même sur le côté nord et le côté sud d'un arbre. Pour cela on découpe des cubes de liège de même dimension sur chaque côté nord et chaque côté sud de 20 arbres. Les masses obtenues sont les suivantes :

Arbre	1	2	3	4	5	6	7	8	9	10
Nord	68,3	60,1	52,2	41,7	32,0	30,9	39,3	42,0	37,7	33,5
Sud	72,5	56,0	55,8	39,2	31,4	35,5	39,2	41,1	43,3	31,7

Arbre	11	12	13	14	15	16	17	18	19	20
Nord	32,2	63,3	54,2	47,0	91,9	56,1	79,6	81,2	78,4	46,6
Sud	31,9	58,1	52,7	46,2	90,2	55,4	75,1	86,6	75,3	43,8

Effectuez le test avec un risque $\alpha = 0{,}05$ dans le cas où l'on peut supposer les populations gaussiennes.

a) Y a-t-il une variation significative (au risque 5 %) des durées d'endormissement entre ces deux expériences ?

b) En admettant que les différences des durées d'endormissement conservent la même moyenne et le même écart type estimé, sur quel nombre minimum n (avec $n \geq 30$) d'individus doit porter l'expérience pour conclure, au même risque, à une différence significatice ?

12-10 On désire comparer la régularité du travail d'une nouvelle doseuse pour boîte de haricots verts à la norme habituelle de l'usine pour laquelle l'écart type est $\sigma = 4$ g. On suppose que la variable aléatoire donnant le poids d'une boîte prise au hasard dans la production suit une loi normale.

a) On prélève un échantillon de taille 10 sur lequel on obtient un écart type estimé $s = 4{,}84$ g. Au risque $\alpha = 0{,}05$, peut-on considérer que ce résultat est conforme à la norme souhaitée ?

b) Même question en supposant que les mêmes valeurs numériques ont été obtenues à partir d'un échantillon de taille 50.

12-11 On a étudié l'homogénéité des rendements fouragers de deux types de prairie. Chaque type de prairie a été partagé en plusieurs parcelles. Les résultats sont les suivants (en kg/are) :

	Prairie n° 1	Prairie n° 2
Parcelle 1	19,8	15,9
Parcelle 2	20,6	19,8
Parcelle 3	27,0	20,9
Parcelle 4	29,5	22,5
Parcelle 5	29,9	26,3

On suppose que la variable aléatoire donnantles rendements suit une loi normale.
Peut-on dire, au seuil de 5 %, que les deux populations ont la même variance ?
Si oui, peut-on conclure, en comparant les moyennes, que les rendements sont homogènes dans les deux types de prairie ?

12-12 Dans un article de la revue *Biometrica*, le biologiste Latter donne la longueur L (en mm) des oeufs de coucou trouvés dans les nids de deux espèces d'oiseaux :
– dans des nids de petite taille (roitelet) :
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
– dans des nids de taille plus grande (fauvette) :
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
On suppose que L suit une loi normale dans chacune des deux populations.
Peut-on dire, au seuil de 5 %, que les deux populations ont la même variance ?
Si oui, testez l'hypothèse que le coucou adapte la taille de ses oeufs à la taille du nid dans lequel il pond.

SOLUTIONS

12-1 ❒ **a)** ☒ **b)** ☒ **c)** ☒ **d)** ☒ **e)**

- Il s'agit d'une comparaison de deux moyennes observées dans le cas d'échantillons indépendants de petites tailles. On doit supposer les populations gaussiennes, sinon on pourrait utiliser un test non paramétrique (voir fiche 16) à condition de disposer des mesures.
- On teste d'abord l'égalité des variances, soit $H_0 : \sigma_1^2 = \sigma_2^2$. On calcule $f = \dfrac{s_2^2}{s_1^2} \approx 1{,}136$ que l'on compare au seuil à 5 % lu dans la

table de Snedecor titrée $\alpha = 0{,}025$ avec ddl = (17,19), soit $f_{0{,}05} \approx 2,6$. Comme $f < f_{0{,}05}$ l'égalité des variances est acceptée.

- Pour tester $H_0 : \mu_1 = \mu_2$, on calcule :

$$\hat{\sigma}^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \approx 0{,}353$$

$$\text{puis } t = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\hat{\sigma}^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \approx 6{,}345$$

Le test est bien unilatéral car on attend une amélioration (significative ou non) mais pas une détérioration.
Pour finir (ce qui n'est pas demandé) on choisit $\alpha = 0{,}05$; on lit le seuil $t_{2\alpha} = t_{0{,}10} \approx 1{,}69$, pour ddl = 36.
Comme $6{,}345 > 1,69$, l'hypothèse d'égalité des moyennes est rejetée et on conclut que le traitement est efficace.

12-2 ❒ **a)** ❒ **b)** ❒ **c)** ☒ **d)** ❒ **e)**
Il s'agit d'une comparaison de deux moyennes observées issues d'échantillons appariés. Sous $H_0 : \mu_1 = \mu_2$, on se ramène à comparer la moyenne des différences avec la valeur $\mu = 0$.

12-3 ❒ **a)** ❒ **b)** ☒ **c)** ❒ **d)** ❒ **e)**
Pour réaliser le test on calcule les différences :

$$\{0,1\ ;\ 0,4\ ;\ 0,3\ ;\ -0,2\ ;\ 0\ ;\ 0,2\ ;\ 0,5\ ;\ 0,2\ ;\ 0,3\ ;\ 0,2\}$$

Avec ces valeurs on obtient :
$n = 10$, $\overline{d} = 0,2$, $s_d = 0,2$ puis $t \approx 3,162$.
Avec ddl = 9, on lit :
$t_{0,10} = 1,833$; $t_{0,05} = 2,262$; $t_{0,01} = 3,250$; $t_{0,001} = 4,781$
On rejette donc H_0 à $\alpha = 0,05$ mais pas à $\alpha = 0,01$.

12-4 Il s'agit d'un test de comparaison d'une moyenne expérimentale $\overline{x} = 2,6$ et d'une moyenne théorique $\mu = 2,5$.
L'hypothèse nulle (H_0) est que la différence entre $\overline{x}$ et μ n'est pas significative, et le test est bilatéral car on ne sait rien a priori sur le bienfait qu'il y ait trop, ou pas assez, de substance active.
Comme il s'agit d'un grand échantillon, si (H_0) est vraie, alors la variable aléatoire $Z = \dfrac{\overline{X} - \mu}{\dfrac{s}{\sqrt{n}}}$ suit à peu près la loi $\mathcal{N}(0,1)$.

U prend la valeur $z = \dfrac{2,6 - 2,5}{\dfrac{0,4}{\sqrt{100}}} = 2,5$.

Si $\alpha = 0{,}05$, on a $z_{0,05} = 1{,}96$.

Comme $z \notin]-z_{0,05}; z_{0,05}[$, on rejette (H_0) et on conclut, au risque 5 %, que la production ne respecte pas les spécifications.

Comme $u_{0,01} = 2,576$, on ne rejetterait pas (H_0) si on limitait le risque de première espèce à 1 %.
Ce n'est pas surprenant : quand on diminue α, on rejette moins souvent (H_0).

12-5 Il s'agit de comparer une moyenne expérimentale $\overline{x}$ et une moyenne théorique $\mu = 87{,}6$.

(H_0) : le poids moyen des rongeurs ne diffère pas significativement de la moyenne théorique.

Sous l'hypothèse (H_0), comme on dispose d'un petit échantillon et d'une population gaussienne, la variable aléatoire $T = \dfrac{\overline{X} - \mu}{\dfrac{s}{\sqrt{n}}}$ suit la loi de Student à $n - 1 = 4$ degrés de liberté.

Avec les mesures de l'échantillon, on calcule : $\overline{x} = 82{,}6$ et $s \approx 2{,}07$.

La valeur prise par T est donc : $t \approx -5{,}39$

Avec $\nu = 4$ et $\alpha = 0{,}05$, on lit dans la table 3 : $t_{0,05} = 2{,}776$.

Comme $t \notin]-t_{0,05}; t_{0,05}[$, (H_0) est rejetée au risque 5 %. Le poids moyen des rongeurs traités est significativement différent de la norme. Le traitement a donc un effet sur le poids.

12-6 Il s'agit d'une comparaison de deux moyennes expérimentales provenant de deux grands échantillons.

L'hypothèse nulle (H_0) à tester est que les poids moyens ne sont pas significativement différents aux deux époques de la récolte.

Si (H_0) est vraie, alors $Z = \dfrac{\overline{X}_1 - \overline{X}_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$ suit à peu près la loi $\mathcal{N}(0,1)$.

On a $s_1^2 = (20)^2$ et $s_2^2 = (10)^2$; d'où $z \approx -13{,}88$.

On lit dans la table 2, ou dans le bas de la table 3 :

$$z_{0,05} = 1{,}96 \ ; \ z_{0,01} = 2{,}576 \ ; \ z_{0,001} = 3{,}291.$$

Dans tous les cas $|u| > z_\alpha$ et on rejette (H_0).

Les deux moyennes sont donc significativement différentes, même avec seulement $\alpha = 0{,}001$.

Si on supposait que les pommes ne peuvent que grossir en cours de récolte, on choisirait un test unilatéral. Mais ce n'est pas sûr ; le plus probable est que la cueillette ne concerne pas les mêmes espèces en début et en fin de récolte.

12-7 Il s'agit d'une comparaison de deux moyennes expérimentales provenant de deux petits échantillons indépendants. On va tester : (H_0) : $\mu_1 = \mu_2$; la différence des moyennes n'est pas significative.
Les populations sont supposées gaussiennes et de même variance. Cette variance commune est estimée par :

$$\hat{\sigma}^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = 307{,}8.$$

Sous (H_0), la variable aléatoire $T = \dfrac{\overline{X}_1 - \overline{X}_2}{\sqrt{\hat{\sigma}^2\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$ suit la loi de Student à $n_1 + n_2 - 2$ degrés de liberté.
La valeur prise par T est $t \approx 0{,}28$.
Avec $\alpha = 0{,}05$ et $v = 18$, on lit dans la table 3 : $t_{0,05} = 2{,}101$.
Comme $t \in \,]-t_{0,05}\,;\, t_{0,05}[$, on ne peut pas écarter (H_0).
Les deux moyennes ne sont donc pas significativement différentes.

12-8 Nous sommes en présence de deux échantillons appariés car il s'agit des *mêmes* malades, avant et après le traitement. Et le traitement est, a priori, destiné à diminuer la tension artérielle, ce qui conduit à effectuer un test unilatéral.
Soit μ_1 (avant traitement) et μ_2 (après traitement) les moyennes des populations correspondantes. Les hypothèses à tester sont :
$(H_0) : \mu_1 = \mu_2$;
$(H_1) : \mu_1 > \mu_2$.
Calculons les différences entre les valeurs de la tension artérielle avant et après traitement :

$$\{3\,;\,2\,;\,0\,;\,2\,;\,4\,;\,3\,;\,-1\,;\,1\,;\,3\,;\,-2\}.$$

Cet échantillon provient, par hypothèse d'une population gaussienne. Il a pour moyenne $\overline{d} = 1{,}5$ et pour écart type estimé $s_d \approx 1{,}96$.

Sous (H_0), la variable aléatoire $T = \dfrac{\overline{D} - 0}{\dfrac{s_d}{\sqrt{n}}}$ suit la loi de Student à $n - 1$ degrés de liberté.

La valeur prise par T est $t = \dfrac{\overline{d}}{s_d}\sqrt{10} \approx 2{,}42$.

Pour $\alpha = 0{,}05$ et $v = 9$, comme le test est unilatéral, la valeur frontière est $t_{0,10} = 1{,}833$.

Comme $t > t_{0,10}$, on rejette l'hypothèse nulle au risque 5 %.
On conclut donc que le traitement a une action significative.

12-9 Les échantillons sont appariés car il s'agit de la face nord et de la face sud des *mêmes* arbres. En fait, on veut étudier l'influence des vents dominants sans faire intervenir la variabilité due aux arbres.
Le test est bilatéral et on teste :
$(H_0) : \mu_1 = \mu_2$; la densité de l'écorce est la même sur le côté nord et sur le côté sud.
Considérons l'échantillon constitué par les différences « Nord-Sud » :
{–4,2 ; 4,1 ; –3,6 ; 2,5 ; 0,6 ; –4,6 ; 0,1 ; 0,9 ; –5,6 ; 1,8 ; 0,3 ; 5,2 ; 1,5 ; 0,8 ; 1,7 ; 0,7 ; 4,5 ; –5,4 ; 3,1 ; 2,8}
Le test est équivalent à la comparaison de la moyenne $\overline{d}$ à la moyenne théorique $\mu = 0$.
Les populations étant supposées gaussiennes (en fait, la bonne hypothèse est que la différence des valeurs suit une loi de Gauss), la variable aléatoire $T = \dfrac{\overline{D} - 0}{\dfrac{s_d}{\sqrt{n}}}$ suit la loi de Student à $\nu = n - 1$ degrés de liberté.
Ici on a : $\overline{d} = 0{,}36$; $s_d \approx 3{,}32$; $n = 20$
d'où l'on déduit la valeur prise par T, soit $t \approx 0{,}49$.
Avec $\alpha = 0{,}05$, on lit dans la table 3 la borne $t_{0,05} = 2{,}093$.
Comme $|t| < t_{0,05}$, (H_0) est non rejetée à $\alpha = 0{,}05$. Cette observation ne met pas en évidence de différence significative entre les densités.

12-10 On va tester (H_0) : la variance estimée s^2 est conforme à σ^2.
Comme la population est gaussienne, sous (H_0), la variable aléatoire $Y = \dfrac{n-1}{\sigma^2} S^2$ suit la loi du χ^2 à $\nu = n - 1$ degrés de liberté.

a) Cas $n = 10$

La valeur prise par Y est $y = \dfrac{9 \times 4{,}84^2}{4^2} = 13{,}1769$.

La table du χ^2 permet de déterminer les nombres a et b tels que $P(Y \geq b) = \dfrac{\alpha}{2}$ et $P(Y \leq a) = 1 - \dfrac{\alpha}{2}$.

Avec $\alpha = 0{,}05$ et $\nu = 9$, on lit : $a = 2{,}70$ et $b = 19{,}02$.
Comme $y \in]a, b[$, l'hypothèse (H_0) ne peut pas être rejetée.

a) Cas $n = 50$

Le degré de liberté $\nu = 49$ ne permet pas d'utiliser les tables (sur papier) du χ^2.

Mais, dans ce cas, $Z = \sqrt{2Y} - \sqrt{2v-1}$ suit sensiblement la loi normale centrée réduite. La valeur prise par U est :

$$z = \sqrt{\frac{2 \times 49 \times 4{,}84^2}{4^2}} - \sqrt{97} \approx 2{,}13.$$

Si $\alpha = 0{,}05$, on a $z_{0,05} = 1{,}96$.
Comme $z \notin]-z_{0,05}, z_{0,05}[$, l'hypothèse (H_0) est rejetée avec un risque d'erreur inférieur à 5 %.

Si la conclusion a changé avec les mêmes valeurs numériques, c'est parce que l'information apportée par une observation sur un échantillon de plus grande taille est plus riche qu'avant.

12-11 a) Comparaison des variances

Le premier échantillon (prairie n° 1) a pour taille $n_1 = 5$, pour moyenne $\overline{x}_1 = 25{,}36$ et pour variance estimée $s_1^2 = 23{,}503$.
Le deuxième échantillon (prairie n° 2) a pour taille $n_2 = 5$, pour moyenne $\overline{x}_2 = 21{,}08$ et pour variance estimée $s_2^2 = 14{,}442$.
Nous allons tester l'hypothèse nulle (H_0) : les deux populations ont la même variance.
Comme on suppose que les rendements suivent des lois normales, si (H_0) est vraie, $F = \dfrac{S_1^2}{S_2^2}$ suit la loi de Snedecor à (4; 4) degrés deliberté.
En faisant le quotient dans l'ordre où le résultat est > 1, on a $f = \dfrac{s_1^2}{s_2^2} \approx 1{,}63$.
Pour $\alpha = 0{,}05$, la table 5 indique $f_{0,05} = 9{,}60$.
Comme $f < f_{0,05}$, l'hypothèse (H_0) ne peut pas être rejetée au risque 5 %.

b) Comparaison des moyennes

Ici l'hypothèse nulle est (H_0) : $\mu_1 = \mu_2$, c'est-à-dire les rendements sont homogènes dans les deux types de prairie.
En admettant que $\sigma_1^2 = \sigma_2^2$ d'après la question précédente, comme les populations sont supposées gaussiennes, on sait que, sous (H_0), la variable aléatoire $T = \dfrac{\overline{X}_1 - \overline{X}_2}{\sqrt{\hat{\sigma}^2\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$ suit à peu près la loi de Student à $n_1 + n_2 - 2$ degrés de liberté.
On a : $\hat{\sigma}^2 = \dfrac{4s_1^2 + 4s_2^2}{8} = 18{,}9725$ et T prend la valeur $t \approx 1{,}55$.
Pour $\alpha = 0{,}05$ et $v = 8$, on a $t_{0,05} = 2{,}306$.

Comme $t \in]-t_{0,05}, t_{0,05}[$, l'hypothèse (H_0) ne peut pas être rejetée. On peut donc considérer que les rendements ne sont pas significativement différents dans les deux types de prairie.

12-12 • Estimations ponctuelles

À partir du premier échantillon de taille $n_1 = 15$, on peut estimer la moyenne μ_1 de L dans la première population par $\overline{x}_1 \approx 21{,}25$ mm et la variance σ_1^2 par $s_1^2 \approx 0{,}516$ mm^2.

À partir du deuxième échantillon de taille $n_2 = 14$, on peut estimer la moyenne μ_2 de L dans la deuxième population par $\overline{x}_2 \approx 23{,}11$ mm et la variance σ_2^2 par $s_2^2 \approx 1{,}101$ mm^2.

• Comparaison des variances

On va tester (H_0) : les deux populations ont la même variance.

Les populations étant supposées gaussiennes, si (H_0) est vraie, $F = \dfrac{S_1^2}{S_2^2}$ suit la loi de Snedecor à (14; 13) degrés de liberté.

En faisant le quotient dans l'ordre où le résultat est > 1, on a $f = \dfrac{s_2^2}{s_1^2} \approx 2{,}13$.

Les degrés de liberté sont devenus (13 ;14) à la suite de la permutation des termes du quotient.

Pour $\alpha = 0{,}05$, la table 5 indique $f_{0,05} \approx 3{,}07$.

Comme $f < f_{0,05}$, l'hypothèse (H_0) ne peut pas être rejetée au risque 5 %.

• Comparaison des moyennes

On va tester (H_0) : le coucou n'adapte pas la taille de ses oeufs à celle du nid dans lequel il pond, c'est-à-dire $\mu_1 = \mu_2$.

D'après le test précédent, les variances des deux populations ne sont pas significativement différentes. Leur variance commune peut être estimée par : $\widehat{\sigma}^2 = \dfrac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \approx 0{,}798$.

Sous l'hypothèse (H_0), comme il s'agit de petits échantillons extraits de populations gaussiennes de même variance, la variable aléatoire $T = \dfrac{\overline{X}_1 - \overline{X}_2}{\sqrt{\widehat{\sigma}^2\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$ suit à peu près la loi de Student à $n_1 + n_2 - 2$ degrés de liberté.

La valeur prise par T est $t \approx -5{,}60$.
Pour $\alpha = 0{,}01$ et $v = 27$, on lit : $t_{0{,}01} = 2{,}771$; pour $\alpha = 0{,}001$, on lit : $t_{0{,}001} = 3{,}69$.
Dans tous les cas, on a : $t \notin]-t_\alpha, t_\alpha[$ et on peut affirmer, avec un risque d'erreur inférieur à 0,001, que le coucou adapte la grosseur de ses oeufs à la taille du nid.

Il s'agit d'un phénomène de mimétisme qui permet aux oeufs de coucou de passer plus facilement inaperçus.

CHAPITRE 13

Analyse de la variance

PLAN

13.1 Généralités
13.2 Analyse de la variance à un facteur
13.3 Analyse de la variance à deux facteurs (échantillons de plusieurs observations)
13.4 Analyse de la variance à deux facteurs (échantillons d'une seule observation)

OBJECTIFS

- Comparer simultanément plusieurs moyennes pour étudier l'influence des diverses modalités d'un facteur sur une grandeur mesurable
- Étudier l'influence de deux facteurs sur une grandeur mesurable, et leur interaction

13.1 GÉNÉRALITÉS

L'analyse de variance (comme son nom ne l'indique pas) permet de comparer les moyennes de plusieurs échantillons indépendants afin de tester l'influence d'un ou plusieurs facteurs.

L'analyse de variance n'est valable en toute rigueur que pour des échantillons tirés de populations gaussiennes et de même variance. En général, le non-respect de ces conditions n'a pas trop d'influence sur la validité du test (on dit que l'analyse de variance est une méthode robuste). L'erreur introduite est cependant d'autant plus forte que les effectifs des échantillons sont faibles et inégaux.

13.2 ANALYSE DE LA VARIANCE À UN FACTEUR

Problématique

On dispose de k échantillons indépendants $E_1,\ldots,E_k$, extraits de k populations $P_1,\ldots,P_k$ supposées gaussiennes et de même variance σ^2. Les moyennes respectives des populations sont notées $\mu_1,\ldots,\mu_k$.

L'**analyse de variance** (ou **ANOVA** : ANalysis Of VAriance) permet de comparer globalement les moyennes des populations.

L'hypothèse nulle est donc :

$(H_0) : \mu_1 = \cdots = \mu_k$

En général, les k échantillons correspondent à k modalités d'un facteur contrôlé. Par exemple, il peut s'agir de k groupes de malades, chaque groupe recevant un traitement différent. Le facteur contrôlé est alors le facteur traitement. Il est donc équivalent de formuler l'hypothèse nulle sous la forme :

(H_0) : la moyenne des populations est indépendante du facteur étudié.

Variance résiduelle ; variance factorielle

- Pour chaque échantillon E_i, de taille n_i, on calcule la moyenne $\overline{x}_i$ et la variance estimée s_i^2.
- La réunion de tous les échantillons a pour taille n, pour moyenne $\overline{x}$ et pour variance estimée s^2. On a : $n = \sum_{i=1}^{k} n_i$ et $\overline{x} = \frac{1}{n}\sum_{i=1}^{k} n_i\,\overline{x}_i$.

  s^2 caractérise la dispersion de l'ensemble des données par rapport à la moyenne générale $\overline{x}$.
- Avec les hypothèses de départ, on dispose d'une première estimation de σ^2 appelée **variance résiduelle** (ou **variance intragroupe**) et définie par :

$$s_R^2 = \frac{1}{n-k}\sum_{i=1}^{k}(n_i - 1)s_i^2.$$

s_R^2 est la moyenne des variances estimées s_i^2 affectées des coefficients $n_i - 1$. Elle caractérise la dispersion des valeurs à l'intérieur des échantillons.

- Sous l'hypothèse (H_0), on dispose d'une deuxième estimation de σ^2 appelée **variance factorielle** (ou **variance intergroupe**) et définie par :

$$s_F^2 = \frac{1}{k-1}\sum_{i=1}^{k} n_i(\overline{x}_i - \overline{x})^2.$$

s_F^2 caractérise la dispersion des valeurs d'un échantillon à l'autre, c'est-à-dire la variation due à l'influence du facteur étudié.

Théorème d'analyse de la variance

$$(n-1)s^2 = (n-k)s_R^2 + (k-1)s_F^2.$$

s^2 est donc une moyenne pondérée de s_R^2 et de s_F^2. Ce théorème permet d'obtenir s_F^2 après avoir calculé s_R^2 et s^2, ce qui est plus rapide qu'avec la définition.

Variante des calculs (avec tableur)

– Écrire en colonnes C_i les mesures x_{ij} de chaque échantillon E_i.

– Déterminer l'effectif n_i de chaque échantillon et l'effectif total $n = \sum_i n_i$.

– Pour chaque colonne, additionner les valeurs, élever au carré la somme obtenue et diviser par l'efffectif de l'échantillon, soit $\frac{1}{n_i}\left(\sum_j x_{ij}\right)^2$.

– Additionner tous ces résultats, ce qui donne $A = \sum_i \frac{1}{n_i}\left(\sum_j x_{ij}\right)^2$.

– Additionner toutes les mesures, ce qui donne $B = \sum_{i,j} x_{ij}$.

– Additionner tous les carrés de toutes les mesures : $C = \sum_{i,j} (x_{ij})^2$.

On a alors : $s_F^2 = \frac{1}{k-1}\left[A - \frac{B^2}{n}\right]$; $s_R^2 = \frac{1}{n-k}[C - A]$.

Test de l'hypothèse nulle

Théorème. Sous (H_0), la variable aléatoire $F = \frac{S_F^2}{S_R^2}$ suit la loi de Snedecor à $(k-1, n-k)$ degrés de liberté.

Décision

Soit α le risque de première espèce choisi. On lit dans la table de Snedecor la valeur f_α telle que $P(F \geq f_\alpha) = \alpha$.

- Si $f < f_\alpha$, on ne peut pas écarter (H_0).
- Si $f \geq f_\alpha$, on rejette (H_0) au risque α, c'est-à-dire que l'on attribue une influence significative au facteur étudié.

À la différence de l'utilisation des tables de Snedecor pour comparer deux variances observées (*cf.* chap. 12), le quotient est à effectuer dans un ordre imposé, les degrés de liberté ne sont pas les mêmes, et le risque α du test est le même que celui de la légende de la table.

13.3 ANALYSE DE LA VARIANCE À DEUX FACTEURS (ÉCHANTILLONS DE PLUSIEURS OBSERVATIONS)

Problématique

On étudie simultanément deux facteurs : un facteur A à p modalités et un facteur B à q modalités. Pour chacune des pq modalités du couple (A,B), on dispose d'un échantillon E_{ij} avec $1 \leq i \leq p$ et $1 \leq j \leq q$. Ces échantillons sont supposés extraits de populations gaussiennes ayant la même variance. Ils sont aussi tous de même taille n (avec $n > 1$).

L'analyse de variance à deux facteurs permet de comparer les moyennes de ces pq échantillons et de tester :

– l'influence du facteur A seul ;
– l'influence du facteur B seul ;
– l'influence de l'interaction des deux facteurs : on dit qu'il y a interaction lorsque l'influence d'un facteur sur la moyenne des populations est différente en l'absence ou en présence de l'autre facteur.

Il y a donc trois hypothèses nulles, et par conséquent trois tests :

$(H_0)_A$: le facteur A n'a pas d'influence sur la moyenne des populations ;

$(H_0)_B$: le facteur B n'a pas d'influence sur la moyenne des populations ;

$(H_0)_{AB}$: il n'y a pas d'interaction entre les facteurs A et B.

Variance résiduelle ; variance factorielle

- Pour chaque échantillon E_{ij}, on calcule la moyenne $\overline{x}_{ij}$ et la variance estimée s_{ij}^2.
- La réunion de tous les échantillons a pour taille npq, pour moyenne $\overline{x}$ et pour variance estimée s^2. On a : $\overline{x} = \dfrac{1}{npq} \displaystyle\sum_{\substack{1 \leq i \leq p \\ 1 \leq j \leq q}} n\,\overline{x}_{ij}$.

➤ De façon analogue au cas de l'analyse de variance à un facteur, on définit la **variance résiduelle** par la moyenne des variances estimées :

$$s_R^2 = \frac{1}{(n-1)pq} \sum_{\substack{1 \le i \le p \\ 1 \le j \le q}} (n-1)s_{ij}^2.$$

Si $n > 1$, on a $s_R^2 > 0$.

➤ On définit de même la **variance factorielle** par :

$$s_F^2 = \frac{1}{pq-1} \sum_{\substack{1 \le i \le p \\ 1 \le j \le q}} n(\overline{x}_{ij} - \overline{x})^2.$$

Théorème d'analyse de la variance

$$(npq-1)s^2 = (n-1)pq\, s_R^2 + (pq-1)s_F^2.$$

s^2 est donc une moyenne pondérée de s_R^2 et de s_F^2.

Décomposition de la variance factorielle

Pour étudier l'influence de chacun des deux facteurs A et B et celle de leur interaction, on définit :

– les moyennes conditionnelles

$$\overline{x}_{i\bullet} = \frac{1}{q}\sum_{j=1}^{q} \overline{x}_{ij} \quad \text{et} \quad \overline{x}_{\bullet j} = \frac{1}{p}\sum_{i=1}^{p} \overline{x}_{ij}$$

$\overline{x}_{i\bullet}$ est la moyenne de la i-ième ligne ; $\overline{x}_{\bullet j}$ la moyenne de la j-ième colonne.

– la variance conditionnelle due au facteur A seul

$$s_A^2 = \frac{1}{p-1}\sum_{i=1}^{p} qn(\overline{x}_{i\bullet} - \overline{x})^2$$

– la variance conditionnelle due au facteur B seul

$$s_B^2 = \frac{1}{q-1}\sum_{j=1}^{q} pn(\overline{x}_{\bullet j} - \overline{x})^2$$

– la variance conditionnelle due à l'interaction de A et de B

$$s_{AB}^2 = \frac{1}{(p-1)(q-1)} \sum_{\substack{1 \le i \le p \\ 1 \le j \le q}} n(\overline{x}_{ij} - \overline{x}_{i\bullet} - \overline{x}_{\bullet j} + \overline{x})^2$$

Théorème de décomposition de la variance factorielle

$$(pq-1)s_F^2 = (p-1)s_A^2 + (q-1)s_B^2 + (p-1)(q-1)s_{AB}^2.$$

s_F^2 est donc une moyenne pondérée de s_A^2, s_B^2 et de s_{AB}^2.

Tests des hypothèses nulles

- Sous $(H_0)_A$, la variable aléatoire $F_A = \dfrac{S_A^2}{S_R^2}$ suit la loi de Snedecor à $(p-1,(n-1)pq)$ degrés de liberté.
- Sous $(H_0)_B$, la variable aléatoire $F_B = \dfrac{S_B^2}{S_R^2}$ suit la loi de Snedecor à $(q-1,(n-1)pq)$ degrés de liberté.
- Sous $(H_0)_{AB}$, la variable aléatoire $F_{AB} = \dfrac{S_{AB}^2}{S_R^2}$ suit la loi de Snedecor à $((p-1)(q-1),(n-1)pq)$ degrés de liberté.

Pour mémoriser les degrés de liberté : le premier est associé au numérateur et le second au dénominateur, et ce sont les coefficients qui figurent dans les théorèmes de décomposition.

Le test de chaque hypothèse nulle s'en déduit comme d'habitude.

13.4 ANALYSE DE LA VARIANCE À DEUX FACTEURS (ÉCHANTILLONS D'UNE SEULE OBSERVATION)

Problématique

Si chaque échantillon ne comporte qu'une seule observation (soit $n = 1$), les s_{ij}^2 sont nulles et on a $s_R^2 = 0$. Les quotients effectués précédemment n'ont donc plus de sens.

Mise en place du test et décision

- Le théorème d'analyse de la variance devient :

$$(pq-1)s^2 = (p-1)s_A^2 + (q-1)s_B^2 + (p-1)(q-1)s_{AB}^2$$

et permet d'obtenir s_{AB}^2 après avoir calculé s^2, s_A^2, s_B^2.

- Sous $(H_0)_A$, la variable aléatoire $F_A = \dfrac{S_A^2}{S_{AB}^2}$ suit la loi de Snedecor à $(p-1,(p-1)(q-1))$ degrés de liberté.
- Sous $(H_0)_B$, la variable aléatoire $F_B = \dfrac{S_B^2}{S_{AB}^2}$ suit la loi de Snedecor à $(q-1,(p-1)(q-1))$ degrés de liberté.
- Le test de $(H_0)_A$ et $(H_0)_B$ s'en déduit comme d'habitude, mais on ne peut pas tester $(H_0)_{AB}$.

Comparaison de plusieurs variances expérimentales : test de Bartlett

Dans l'analyse de variance qui précède, les populations (gaussiennes) sont supposées de même variance. En toute rigueur, il faut tester cette hypothèse au préalable, même si c'est une étape souvent omise. On peut en particulier le faire avec le test de Bartlett.

Les notations sont inchangées et l'hypothèse nulle s'écrit :

$(H_0) : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$.

• Théorème

Sous (H_0), la variable aléatoire définie par :

$$B = \frac{1}{\lambda}\left[(n-k)\ln S_R^2 - \sum_{i=1}^{k}(n_i-1)\ln S_i^2\right]$$

avec $\lambda = 1 + \dfrac{1}{3(k-1)}\left[\left(\sum_{i=1}^{k}\dfrac{1}{n_i-1}\right) - \dfrac{1}{n-k}\right]$

suit à peu près une loi du χ^2 à $\nu = k-1$ degrés de liberté.

• Décision

On calcule la valeur b prise par la variable aléatoire B. Le risque α étant choisi, on lit dans la table 4 la borne χ_α^2 telle que $P(B > \chi_\alpha^2) = \alpha$.

- Si $b < \chi_\alpha^2$,(H_0) ne peut pas être rejetée et on peut considérer que les populations ont la même variance ;
- si $b \geq \chi_\alpha^2$, (H_0) est rejetée avec un risque d'erreur égal à α.

MOTS-CLÉS

- Décomposition de la variance
- Variance factorielle
- Variance résiduelle

EXERCICES

13-1 On veut savoir si l'addition de substances adjuvantes à un vaccin modifie la production d'anticorps. Pour cela, on mesure les quantités d'anticorps produites par des sujets après administration de quantités égales du vaccin, additionné ou non d'une substance adjuvante. On a obtenu les taux :

– sans substance adjuvante : 1,3,3,0,1 ;
– avec de l'alumine : 2,4,5,4,3,6 ;
– avec des sels de calcium : 3,3,4,5 ;
– avec des phosphates : 1,4,2,3,3 .

a) Quelle(s) hypothèse(s) faut-il faire pour pouvoir appliquer la technique d'analyse de la variance à la résolution du problème posé, La validité de ces hypothèses est-elle importante dans le cas présent ?

b) Ces hypothèses étant satisfaites, l'efficacité du vaccin dépend-elle :

1) de la présence de substances adjuvantes ?

2) de leur nature ?

c) Si les hypothèses précédentes n'avaient pas été satisfaites, quelle technique statistique aurait-on pu appliquer ?

13-2 On a étudié la durée de développement (en jours) d'un parasite à l'intérieur d'un organisme hôte, en fonction de la température d'élevage (en degrés *C*).

Les résultats obtenus sont groupés dans le tableau qui suit.

La température a-t-elle une influence sur la durée de développement du parasite ?

Température	Nombre d'animaux	Durée de développement	
		Moyenne	Écart type estimé
16	32	81	6,8
20	33	52	5,2
23	31	46	6,7

13-3 On étudie l'activité d'un enzyme sérique, la 5'-nucléotide-phosphodiéstérase (PDE), en fonction de différents facteurs dans l'espèce humaine. Les résultats sont exprimés en unités internationales par litre de sérum. On admettra l'hypothèse de normalité et d'égalité des variances des populations parents.

a) Chez deux groupes de femmes, enceintes ou non, on obtient les résultats suivants :

femmes non enceintes

1,5 ; 1,6 ; 1,4 ; 2,9 ; 2,2 ; 1,8 ; 2,7 ; 1,9 ; 2,2 ; 2,8 ; 2,1 ; 1,8 ; 3,7 ; 1,8 ; 2,1

femmes enceintes

4,2 ; 5,5 ; 4,6 ; 5,4 ; 3,9 ; 5,4 ; 2,7 ; 3,9 ; 4,1 ; 4,1 ; 4,6 ; 3,9 ; 3,5

La grossesse a-t-elle une influence significative sur l'activité de la PDE ?

b) Afin d'évaluer la précocité de l'augmentation d'activité enzymatique lors de la grossesse, on pratique des dosages chez des femmes enceintes à différentes semaines d'aménorrhée.

On obtient les résultats suivants (les échantillons sont indépendants) :

4 sem.	5 sem.	6 sem.	7 sem.	8 sem.
7,2	4,9	10,4	4,6	6,1
4,3	4,8	4,6	5,6	11,4
5,5	4,7	8,4	8,3	8,2
4,6	5,4	6,1	6,9	5,7
4,7	4,7	8,1	4,5	6,6
5,5	4,7	5,4	4,7	6,6
6,6	6,2	6,7	6,7	6,3
5,3	5,6	7,5	4,8	5,9
5,4	3,2	6,4	5,0	5,8
3,9	6,1	5,6	5,0	4,8
5,5	6,7	6,3	5,3	9,1
2,7	5,5	7,7	7,8	13,2

L'âge de la grossesse a-t-il une influence sur l'activité de l'enzyme ?

13-4 On étudie l'activité d'un enzyme chez des sujets jeunes en fonction de l'âge et du sexe. Les résultats sont les suivants :

âge / sexe	moins de 12 ans	plus de 12 ans
garçons	4,9 ; 2,9 ; 2,7 ; 3,9 4,6 ; 3,3 ; 5,9 ; 4,8 4,1 ; 3,5 ; 7,2 ; 6,1	2,1 ; 2,2 ; 1,1 ; 2,9 5,0 ; 3,5 ; 2,4 ; 4,4 2,1 ; 3,0 ; 3,9 ; 5,6
filles	4,5 ; 6,9 ; 4,0 ; 5,4 1,9 ; 3,6 ; 4,8 ; 3,3 7,5 ; 5,8 ; 4,4 ; 6,0	2,4 ; 3,6 ; 4,8 ; 3,9 5,5 ; 5,0 ; 6,8 ; 2,2 3,1 ; 5,0 ; 4,1 ; 4,7

L'activité enzymatique moyenne dépend-t-elle de l'âge, du sexe ?

13-5 Cherchant à réaliser une émulsion la plus stable possible, un expérimentateur associe les émulsionnants a, b, c, d aux corps gras α, β, γ. La stabilité des émulsions obtenues avec chacune des 12 associations est notée de 0 à 10 :

	a	*b*	*c*	*d*
α	2	1	3	1
β	3	2	3	2
γ	3	4	5	3

La stabilité est-elle significativement différente, au risque 2,5 % :

– en fonction du choix du corps gras ?

– en fonction du choix de l'émultionnant ?

SOLUTIONS

13-1 **a)** Il faut supposer que les résultats appartiennent à des populations gaussiennes et de même variance.

Ces hypothèses sont importantes ici car il s'agit de petits échantillons d'effectifs inégaux.

b) 1) Étudions d'abord l'effet de la présence d'une substance adjuvante, quelle qu'en soit la nature. Pour cela, on regroupe tous les résultats obtenus en présence d'adjuvants :

sans adjuvant : $n_1 = 5$; $\overline{x}_1 = 1{,}6$; $s_1^2 = 1{,}8$

avec adjuvant : $n_2 = 15$; $\overline{x}_2 \approx 3{,}47$; $s_2^2 \approx 1{,}695$

total : $n = 20$; $\overline{x} = 3$; $s^2 \approx 2{,}316$

La variance résiduelle vaut donc : $s_R^2 = \frac{1}{18}(4s_1^2 + 14s_2^2) \approx 1{,}72$

La variance factorielle peut se calculer :

- soit avec sa définition :

$$s_F^2 = \frac{1}{1}[5(\overline{x}_1 - \overline{x})^2 + 15(\overline{x}_2 - \overline{x})^2] \approx 13{,}07$$

- soit à partir du théorème d'analyse de la variance :

$$19s^2 = 18s_R^2 + s_F^2.$$

D'où : $f = \frac{s_F^2}{s_R^2} = 7{,}60$

On teste (H_0) : l'efficacité du vaccin ne dépend pas de la présence de substances adjuvantes.

Sous (H_0), on sait que $F = \frac{S_F^2}{S_R^2}$ suit la loi de Snedecor à (1 ; 18) d.d.l.

Le nombre f_α tel que $P(F \geq f_\alpha) = \alpha$ est $f_{0,05} = 4{,}41$ pour $\alpha = 0{,}05$.

Comme $f > f_{0,05}$, l'influence de la présence d'adjuvants est significative au risque 5 %.

Comme il n'y a que deux échantillons, on pouvait aussi comparer les deux moyennes expérimentales (*cf.* chap. 12).

2) Pour tester la nouvelle hypothèse nulle :
(H_0) : l'efficacité du vaccin ne dépend pas de la nature de l'adjuvant, on dispose de trois échantillons :

avec alumine :	$n_1 = 6$	;	$\overline{x}_1 = 4$	;	$s_1^2 = 2$
avec calcium :	$n_2 = 4$	;	$\overline{x}_2 = 3{,}75$	;	$s_2^2 \approx 0{,}917$
avec phosphates :	$n_3 = 5$	;	$\overline{x}_3 = 2{,}6$	;	$s_2^3 = 1{,}3$
total :	$n = 15$	;	$\overline{x} \approx 3{,}47$	;	$s^2 \approx 1{,}695$

La variance résiduelle vaut donc : $s_R^2 = \frac{1}{12}(5s_1^2 + 3s_2^2 + 4s_3^2) \approx 1{,}50.$

La variance factorielle peut se calculer

- soit avec sa définition :

$$s_F^2 = \frac{1}{2}[6(\overline{x}_1 - \overline{x})^2 + 4(\overline{x}_2 - \overline{x})^2 + 5(\overline{x}_3 - \overline{x})^2] \approx 2{,}89$$

- soit à partir du théorème d'analyse de la variance :

$$14s^2 = 12s_R^2 + 2s_F^2.$$

D'où : $f = \dfrac{s_F^2}{s_R^2} \approx 1{,}93$

Sous (H_0), on sait que $F = \dfrac{S_F^2}{S_R^2}$ suit la loi de Snedecor à (2 ; 12) d.d.l.

Le nombre f_α tel que $P(F \geq f_\alpha) = \alpha$ est $f_{0,05} = 3{,}89$ pour $\alpha = 0{,}05$.

Comme $f < f_{0,05}$, on ne rejette pas (H_0) : l'efficacité du vaccin ne dépend pas de façon significative de la nature de l'adjuvant.

c) Si les hypothèses de normalité et d'égalité des variances n'avaient pas été satisfaites, on aurait pu appliquer un test non paramétrique (*cf.* chap. 16), soit ici le test de Kuskall et Wallis.

13-2 Si l'on admet que les distributions des durées de développement sont gaussiennes et de même variance, on peut appliquer l'analyse de la variance à un facteur et tester l'hypothèse nulle :

(H_0) : la température n'a pas d'influence sur la durée de dévelopement du parasite.

On a immédiatement la variance résiduelle :

$$s_R^2 = \frac{1}{93}[31 \times 6{,}8^2 + 32 \times 5{,}2^2 + 30 \times 6{,}7^2] \approx 39{,}20.$$

La moyenne $\overline{x}$ de la réunion des trois échantillons s'obtient à partir des moyennes $\overline{x}_i$ des échantillons :

$$\overline{x} = \frac{32 \times 81 + 33 \times 52 + 31 \times 46}{96} \approx 59{,}73.$$

D'où la variance factorielle :

$$s_F^2 = \frac{1}{2}[32(81 - \overline{x})^2 + 33(52 - \overline{x})^2 + 31(46 - \overline{x})^2] \approx 11\,146{,}48,$$

puis : $f = \dfrac{s_F^2}{s_R^2} \approx 284{,}36.$

Sous (H_0), on sait que $F = \dfrac{S_F^2}{S_R^2}$ suit la loi de Snedecor à (2 ; 93) degrés de liberté. Le nombre f_α tel que $P(F \geq f_\alpha) = \alpha$ est :

$f_{0,05} \approx 3{,}1$ pour $\alpha = 0{,}05$; $f_{0,025} \approx 3{,}8$ pour $\alpha = 0{,}025$.

Comme $f > f_{0,025}$, l'influence de la température est significative au risque 2,5 % (et même sans doute à des risques beaucoup plus faibles !).

13-3 **a)** On dispose de deux échantillons :

femmes non enceintes :	$n_1 = 15$	;	$\overline{x}_1 \approx 2{,}17$	;	$s_1^2 \approx 0{,}387$
femmes enceintes :	$n_2 = 13$	;	$\overline{x}_2 \approx 4{,}29$	;	$s_2^2 \approx 0{,}651$
total :	$n = 28$	;	$\overline{x} \approx 3{,}15$	;	$s^2 \approx 1{,}655$

La variance résiduelle vaut donc : $s_R^2 = \dfrac{1}{26}(14s_1^2 + 12s_2^2) \approx 0{,}51$.

La variance factorielle peut se calculer

- soit avec sa définition :

$$s_F^2 = \frac{1}{1}[15(\overline{x}_1 - \overline{x})^2 + 13(\overline{x}_2 - \overline{x})^2] \approx 31{,}47$$

- soit à partir du théorème d'analyse de la variance :

$$27s^2 = 26s_R^2 + s_F^2.$$

D'où : $f = \dfrac{s_F^2}{s_R^2} \approx 61{,}9$.

On teste (H_0) : la grossesse n'a pas d'influence significative sur l'activité de la PDE.

Sous (H_0), on sait que $F = \dfrac{S_F^2}{S_R^2}$ suit la loi de Snedecor à (1 ; 26) degrés de liberté. Le nombre f_α tel que $P(F \geq f_\alpha) = \alpha$ est :

$f_{0{,}05} \approx 4{,}23$ pour $\alpha = 0{,}05$; $f_{0{,}025} \approx 5{,}66$ pour $\alpha = 0{,}025$.

Comme $f > f_{0{,}025}$, l'influence de la grossesse est significative au risque 2,5 %.

Comme il n'y a que deux échantillons, on pouvait aussi comparer les deux moyennes expérimentales (*cf.* chap. 12).

b) On dispose de cinq échantillons :

à 4 semaines :	$n_1 = 12$	;	$\overline{x}_1 \approx 5{,}10$	;	$s_1^2 \approx 1{,}411$
à 5 semaines :	$n_2 = 12$	;	$\overline{x}_2 \approx 5{,}21$	;	$s_2^2 \approx 0{,}850$
à 6 semaines :	$n_3 = 12$	;	$\overline{x}_3 \approx 6{,}93$	;	$s_3^2 \approx 2{,}495$
à 7 semaines :	$n_4 = 12$	;	$\overline{x}_4 \approx 5{,}77$	;	$s_4^2 \approx 1{,}742$
à 8 semaines :	$n_5 = 12$	;	$\overline{x}_5 \approx 7{,}475$	;	$s_5^2 \approx 6{,}522$
total :	$n = 60$	;	$\overline{x} \approx 6{,}10$	;	$s^2 \approx 3{,}341$

La variance résiduelle vaut donc :

$$s_R^2 = \frac{1}{55}(11s_1^2 + 11s_2^2 + 11s_3^2 + 11s_4^2 + 11s_5^2) \approx 2{,}60.$$

La variance factorielle peut se calculer

➤ soit avec sa définition :

$$s_F^2 = \frac{1}{4}[12(\overline{x}_1 - \overline{x})^2 + \cdots + 12(\overline{x}_5 - \overline{x})^2] \approx 13{,}47$$

➤ soit à partir du théorème d'analyse de la variance :

$$59s^2 = 55s_R^2 + 4s_F^2.$$

D'où : $f = \dfrac{s_F^2}{s_R^2} \approx 5{,}17$.

On teste (H_0) : l'âge de la grossesse n'a pas d'influence significative sur l'activité de la PDE (égalité des cinq moyennes théoriques).

Sous (H_0), on sait que $F = \dfrac{S_F^2}{S_R^2}$ suit la loi de Snedecor à (4 ; 55) degrés de liberté. Le nombre f_α tel que $P(F \geq f_\alpha) = \alpha$ est :

$f_{0,05} \approx 2{,}5$ pour $\alpha = 0{,}05$; $f_{0,025} \approx 3{,}0$ pour $\alpha = 0{,}025$.

Comme $f > f_{0,025}$, l'influence de l'âge de la grossesse est significative au risque 2,5 %.

13-4 Il s'agit d'analyse de la variance à deux facteurs (sexe et âge).

• Calculs

Chaque échantillon a pour taille $n = 12$. On obtient pour les échantillons :

Sexe \ Âge	Moins de 12 ans	Plus de 12 ans
Garçons	$\overline{x}_{1,1} \approx 4{,}49$ $s_{1,1}^2 \approx 1{,}894$	$\overline{x}_{1,2} \approx 3{,}18$ $s_{1,2}^2 \approx 1{,}783$
Filles	$\overline{x}_{2,1} \approx 4{,}84$ $s_{2,1}^2 \approx 2{,}497$	$\overline{x}_{2,2} \approx 4{,}26$ $s_{2,2}^2 \approx 1{,}746$

et pour la réunion des 4 échantillons :

$$\overline{x} \approx 4{,}19 \quad ; \quad s^2 \approx 2{,}245.$$

La variance résiduelle vaut donc :

$$s_R^2 = \frac{1}{4}\left(s_{1,1}^2 + s_{1,2}^2 + s_{2,1}^2 + s_{2,2}^2\right) \approx 1{,}98.$$

D'après le théorème d'analyse de la variance : $47s^2 = 44s_R^2 + 3s_F^2$, on déduit : $s_F^2 \approx 6{,}13$.

Pour décomposer cette variance factorielle, on calcule :

- les moyennes conditionnelles :

$$\overline{x}_{1\bullet} = \frac{\overline{x}_{1,1} + \overline{x}_{1,2}}{2} \approx 3{,}84 \quad ; \quad \overline{x}_{2\bullet} = \frac{\overline{x}_{2,1} + \overline{x}_{2,2}}{2} \approx 4{,}55$$
$$\overline{x}_{\bullet 1} = \frac{\overline{x}_{1,1} + \overline{x}_{2,1}}{2} \approx 4{,}67 \quad ; \quad \overline{x}_{\bullet 2} = \frac{\overline{x}_{1,2} + \overline{x}_{2,2}}{2} \approx 3{,}72$$

- la variance conditionnelle due au facteur A seul (sexe) :

$$s_A^2 = \frac{2 \times 12}{1}[(\overline{x}_{1\bullet} - \overline{x})^2 + (\overline{x}_{2\bullet} - \overline{x})^2] \approx 6{,}09.$$

- la variance conditionnelle due au facteur B seul (âge) :

$$s_B^2 = \frac{2 \times 12}{1}[(\overline{x}_{\bullet 1} - \overline{x})^2 + (\overline{x}_{\bullet 2} - \overline{x})^2] \approx 10{,}74.$$

- la variance conditionnelle due à l'interaction de A et B à partir de la décomposition de la variance factorielle : $3s_F^2 = s_A^2 + s_B^2 + s_{AB}^2$ d'où : $s_{AB}^2 \approx 1{,}58$.

• Tests

- Sous l'hypothèse $(H_0)_A$ « le sexe n'a pas d'influence sur l'activité enzymatique moyenne », la variable aléatoire $F_A = \dfrac{S_A^2}{S_R^2}$ suit la loi de Snedecor à (1 ; 44) degrés de liberté.

  Le nombre f_α tel que $P(F \geq f_\alpha) = \alpha$ est $f_{0{,}05} = 4{,}1$ pour $\alpha = 0{,}05$.

  Comme $f_A \approx 3{,}08$, l'influence du sexe n'est pas significative au risque 5 %.

- Sous l'hypothèse $(H_0)_B$ « l'âge n'a pas d'influence sur l'activité enzymatique moyenne », la variable aléatoire $F_B = \dfrac{S_B^2}{S_R^2}$ suit la loi de Snedecor à (1 ; 44) degrés de liberté.

  Comme $f_B \approx 5{,}42$, l'influence de l'âge est significative au risque 5 %.

- Sous l'hypothèse $(H_0)_{AB}$ « il n'y a pas d'interaction entre l'influence du sexe et celle de l'âge », la variable aléatoire $F_{AB} = \dfrac{S_{AB}^2}{S_R^2}$ suit la loi de Snedecor à (1 ; 44) degrés de liberté.

  Comme $f_{AB} < 1$, l'hypothèse nulle est acceptée.

13-5 Nous pouvons appliquer l'analyse de la variance à deux facteurs en tenant compte du fait que tous les effectifs sont égaux à 1.

• Calculs

On obtient immédiatement les moyennes conditionnelles :

$\overline{x}_{1\bullet} = 1{,}75 \quad ; \quad \overline{x}_{2\bullet} = 2{,}5 \quad ; \quad \overline{x}_{3\bullet} = 3{,}75$

$\overline{x}_{\bullet 1} \approx 2{,}67 \quad ; \quad \overline{x}_{\bullet 2} \approx 2{,}33; \quad \overline{x}_{\bullet 3} \approx 3{,}67 \quad ; \quad \overline{x}_{\bullet 4} = 2$

et pour l'ensemble des observations :

$$\overline{x} \approx 2{,}67 \quad \text{et} \quad s^2 \approx 1{,}33$$

La variance factorielle due au choix du corps gras est :

$$s_A^2 = \frac{4}{2}[(\overline{x}_{1\bullet} - \overline{x})^2 + (\overline{x}_{2\bullet} - \overline{x})^2 + (\overline{x}_{3\bullet} - \overline{x})^2] \approx 4{,}08.$$

La variance factorielle due au choix de l'émulsionnant est :

$$s_B^2 = \frac{3}{3}[(\overline{x}_{\bullet 1} - \overline{x})^2 + (\overline{x}_{\bullet 2} - \overline{x})^2 + (\overline{x}_{\bullet 3} - \overline{x})^2 + (\overline{x}_{\bullet 4} - \overline{x})^2] \approx 1{,}56.$$

La variance factorielle due à l'interaction des deux facteurs se calcule à partir de la décomposition de la variance factorielle :

$11s^2 = 2s_A^2 + 3s_B^2 + 6s_{AB}^2$; d'où : $s_{AB}^2 \approx 0{,}31$.

• Tests

➤ Sous l'hypothèse $(H_0)_A$ « le choix du corps gras n'a pas d'influence sur la stabilité », la variable aléatoire $F_A = \dfrac{S_A^2}{S_{AB}^2}$ suit la loi de Snedecor à (2 ; 6) degrés de liberté.

Le nombre f_α tel que $P(F \geq f_\alpha) = \alpha$ est $f_{0{,}025} = 7{,}26$ pour $\alpha = 0{,}025$.

Comme $f_A \approx 13{,}36 > 7{,}26$, l'influence du choix du corps gras est significative au risque 2,5 %.

➤ Sous l'hypothèse $(H_0)_B$ « le choix de l'émulsionnant n'a pas d'influence sur la stabilité », la variable aléatoire $F_B = \dfrac{S_B^2}{S_{AB}^2}$ suit la loi de Snedecor à (3 ; 6) degrés de liberté.

Le nombre f_α tel que $P(F \geq f_\alpha) = \alpha$ est $f_{0{,}025} = 6{,}60$ pour $\alpha = 0{,}025$.

Comme $f_B \approx 5{,}09 < 6{,}60$, l'influence du choix de l'émulsionnant n'est pas significative au risque 2,5 %.

Mais elle significative au risque 5 % car $f_{0{,}05} = 4{,}76$ et $f_B > f_{0{,}05}$.

CHAPITRE 14

Régression linéaire

PLAN

14.1 Estimation ponctuelle des paramètres d'une droite de régression

14.2 Intervalles de confiance

14.3 Comparaison des paramètres d'une droite de régression expérimentale à des valeurs théoriques

14.4 Comparaison de deux droites de régression expérimentales

OBJECTIFS

- Estimer, par un nombre ou un intervalle, les coefficients d'une droite de régression, et les comparer à des valeurs de référence
- Apprécier par un intervalle la fiabilité d'une estimation de Y obtenue avec une droite de régression
- Comparer les vitesses de réaction de deux grandeurs Y et Y' aux variations d'une même grandeur contrôlée X

14.1 ESTIMATION PONCTUELLE DES PARAMÉTRES D'UNE DROITE DE RÉGRESSION

Problématique

Certaines expériences conduisent à considérer en même temps deux variables X et Y. Deux cas sont possibles :

- X et Y sont deux variables aléatoires dont les valeurs sont déterminées simultanément ;
- X est une variable contrôlée par l'expérimentateur, c'est-à-dire que ses valeurs x_i sont supposées connues sans erreur, et donc reproductibles à l'identique.

Et Y est une variable aléatoire qui est liée à X, et donc dont les valeurs fluctuent quand on reproduit le même x_i.

Nous allons nous limiter à ce seul cas, avec l'objectif de réaliser un ajustement affine entre Y et X. Pour ceci, nous formulons les hypothèses qui suivent.

Pour toute valeur $X = x_i$ fixée, les diverses valeurs de Y définissent une variable aléatoire Y_i. On suppose que les Y_i suivent des lois normales et que : $E(Y_i) = \alpha x_i + \beta$ et $V(Y_i) = \sigma^2$ (valeur indépendante de x_i)

Les valeurs α, β et σ ne sont en général pas connues et vont être estimées.

Il existe une autre présentation. On suppose que $Y = \alpha X + \beta + \varepsilon$ et que pour toute valeur fixée de X, ε suit une loi normale avec $E(\varepsilon) = 0$ et $V(\varepsilon) = \sigma^2$. Ici, ε s'appelle un résidu. C'est pourquoi σ^2 s'appelle la **variance résiduelle** de Y.

Estimations ponctuelles de α et de β

À n valeurs $x_1,\ldots,x_n$ de X, l'expérience a associé n valeurs $y_1,\ldots,y_n$ de Y. À partir de l'échantillon constitué par les n couples $(x_1, y_1),\ldots,(x_n, y_n)$, on peut calculer (*cf.* chapitre 2) la droite de régression $y = ax + b$ et le coefficient de corrélation r.

Soit A et B les variables aléatoires qui prennent les valeurs a et b quand on répète les échantillons de taille n.

Théorème

$$E(A) = \alpha \quad ; \quad E(B) = \beta$$

a et b sont donc des estimations ponctuelles sans biais de α et de β.

Estimation ponctuelle de la variance résiduelle σ^2

Dans le chapitre 2, après détermination de la droite de régression $y = ax + b$ à partir d'un échantillon de taille n, nous avons déjà écrit la décomposition de la variance :

$$V(Y) = \text{variance expliquée} + \text{variance résiduelle}$$

Théorème. L'estimation de σ^2 peut se faire sans biais par :

$$s_R^2 = \frac{n}{n-2} \times \text{variance résiduelle de l'échantillon}$$

$$= \frac{1}{n-2}\sum_{i=1}^{n}(y_i - ax_i - b)^2 = \frac{n}{n-2}(1-r^2)s_e^2(y)$$

Dans cette expression, $s_e^2(y) = \frac{1}{n}\left(\sum_{i=1}^{n} y_i^2\right) - (\overline{y})^2$ est la variance de l'échantillon des y_i.

Dans la suite, $s_e^2(x) = \frac{1}{n}\left(\sum_{i=1}^{n} x_i^2\right) - (\overline{x})^2$ est la variance de l'échantillon des x_i.

Rappelons qu'il ne faut pas confondre les variances s_e^2 des échantillons et les variances estimées s^2, et qu'on a : $ns_e^2 = (n-1)s^2$

14.2 INTERVALLES DE CONFIANCE

Intervalle de confiance de la pente α

Théorèmes. La variance de A est égale à $\frac{\sigma^2}{n s_e^2(x)}$. Elle peut être estimée par $s_A^2 = \frac{s_R^2}{n s_e^2(x)}$.

La variable aléatoire $T = \frac{A - \alpha}{s_A}$ suit la loi de Student à $n-2$ degrés de liberté.

Un risque α_1 étant choisi, on lit dans la table 3 la valeur t_{α_1} telle que $P(|T| \geq t_{\alpha_1}) = \alpha_1$. Et on peut dire, au risque α_1, que la pente théorique α appartient à l'intervalle de confiance :

$$]a - t_{\alpha_1} s_A, a + t_{\alpha_1} s_A[.$$

Intervalle de confiance de l'ordonnée à l'origine β

Théorèmes. La variance de B est égale à $\sigma^2\left(\frac{1}{n} + \frac{\overline{x}^2}{n s_e^2(x)}\right)$. Elle peut être estimée par $s_B^2 = s_R^2\left(\frac{1}{n} + \frac{\overline{x}^2}{n s_e^2(x)}\right) = \frac{s_R^2}{n^2 s_e^2(x)}\left(\sum_{i=1}^{n} x_i^2\right)$. La variable aléatoire $T = \frac{B - \beta}{s_B}$ suit la loi de Student à $n-2$ degrés de liberté.

De même que précédemment, on peut dire, au risque α_1 choisi, que l'ordonnée à l'origine théorique β appartient à l'intervalle de confiance :

$$]b - t_{\alpha_1} s_B, b + t_{\alpha_1} s_B[.$$

Intervalle de confiance d'une valeur individuelle estimée

L'ajustement affine étant réalisé peut servir à prévoir la valeur attendue pour Y quand l'expérimentateur fixe $X = x_0$. L'estimation ponctuelle de cette valeur est $\hat{y}_0 = ax_0 + b$.

Attention, l'utilisation d'une valeur estimée $\hat{y}_0$ n'est justifiée que si r^2 est voisin de 1 (bon modèle) et si x_0 se situe dans la zone où le modèle a été validé.

Au risque α_1, l'intervalle de confi ance de la valeur prise par Y est :

$$\left]\hat{y}_0 - t_{\alpha_1}\sqrt{s_R^2\left(1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{n s_e^2(x)}\right)}, \hat{y}_0 + t_{\alpha_1}\sqrt{s_R^2\left(1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{n s_e^2(x)}\right)}\right[$$

où t_{α_1} vérifie $P(|T| \geq t_{\alpha_1}) = \alpha_1$, lorsque T suit une loi de Student à $n - 2$ degrés de liberté.

14.3 COMPARAISON DES PARAMÉTRES D'UNE DROITE DE RÉGRESSION EXPÉRIMENTALE À DES VALEURS THÉORIQUES

Problématique

Les notations sont les mêmes que précédemment. On connaît les valeurs théoriques α et β relatives à la population et les valeurs a et b d'une droite de régression obtenue à partir d'un échantillon de taille n. On va comparer successivement a et α, puis b et β.

Comparaison des pentes

(H_0) : la différence entre la pente théorique α et la pente expérimentale a est explicable par les fluctuations d'échantillonnage.

Comme $T = \dfrac{A - \alpha}{s_A}$ suit la loi de Student à $n-2$ d.d.l., on calcule $t = \dfrac{a - \alpha}{s_A}$.

D'autre part, le risque de première espèce α_1 étant choisi et le d.d.l. connu, on détermine avec la table 3 le nombre t_{α_1} tel que : $P(|T| \geq t_{\alpha_1}) = \alpha_1$.

- Si $t \in]-t_{\alpha_1}, t_{\alpha_1}[$, (H_0) ne peut pas être rejetée.
- Si $t \notin]-t_{\alpha_1}, t_{\alpha_1}[$, (H_0) est rejetée au risque α_1.

On peut tester $\alpha = 0$. Cela revient à dire que $E(Y)$ ne dépend pas de X. On dit parfois que la régression est significative si l'hypothèse $\alpha = 0$ est rejetée.

Comparaison des ordonnées à l'origine

(H_0) : la différence entre β et b est explicable par les aléas de l'échantillonnage.

Comme $T = \dfrac{B - \beta}{s_B}$ suit la loi de Student à $n - 2$ d.d.l., le test est analogue au cas précédent.

On peut tester $\beta = 0$. Cela revient à dire que, au niveau de la population, la droite de régression passe par l'origine.

14.4 COMPARAISON DE DEUX DROITES DE RÉGRESSION EXPÉRIMENTALES

On se limitera à la comparaison des pentes.

Problématique

Sur une population P, des variables X et Y vérifient les hypothèses formulées en début de chapitre.

Sur une population P' une variable aléatoire Y' est liée à la même variable contrôlée X, avec les mêmes hypothèses.

De chaque population, on extrait un échantillon de tailles respectives n et n'. Les pentes des droites de régression obtenues sont respectivement : a et a', et les estimations ponctuelles de la variance résiduelle : s_R^2 et $s_R'^2$.

On désire comparer les pentes α et α' des droites de régression théoriques.

On suppose que les variances résiduelles σ^2 et σ'^2 sont égales, ce qui peut faire l'objet d'un test préalable (*cf.* en fin de chapitre).

La variance résiduelle commune aux deux populations est alors estimée par la moyenne pondérée :

$$\widehat{\sigma}^2 = \frac{(n-2)s_R^2 + (n'-2)s_R'^2}{n + n' - 4}.$$

Comparaison des pentes de deux droites de régression

$(H_0) : \alpha = \alpha'$.

Théorème. Sous (H_0) et les hypothèses indiquées précédemment, la variable aléatoire $T = \dfrac{A - A'}{\sqrt{\hat{\sigma}^2\left(\dfrac{1}{n\,s_e^2(x)} + \dfrac{1}{n'\,s_e'^2(x)}\right)}}$ suit la loi de Student à $n + n' - 4$ d.d.l.

Dans cette expression, $s_e^2(x)$ est la variance des x_i de l'échantillon extrait de P, c'est-à-dire concernant X et Y ; $s_e'^2(x)$ est la variance des x_i' de l'échantillon extrait de P', c'est-à-dire concernant X et Y'.

On calcule t, valeur prise par T.

Le risque de première espèce α_1 étant choisi et le d.d.l. connu, on détermine avec la table 3 le nombre t_{α_1} tel que $P(|T| \geq t_{\alpha_1}) = \alpha_1$.

- Si $t \in]-t_{\alpha_1}, t_{\alpha_1}[$, (H_0) ne peut pas être rejetée.
- Si $t \notin]-t_{\alpha_1}, t_{\alpha_1}[$, (H_0) est rejetée au risque α_1.

Comparaison de deux variances résiduelles

Dans le paragraphe 4, avant de comparer les pentes des deux droites de régresion, il faut au préalable tester l'égalité des variances résiduelles σ^2 et σ'^2. On a alors :

$(H_0) : \sigma^2 = \sigma'^2$.

Théorème. Sous (H_0) et les hypothèses déjà indiquées, la variable aléatoire $F = \dfrac{S_R^2}{S_R'^2}$ suit la loi de Snedecor à $(n-2, n'-2)$ degrés de liberté.

Le fonctionnement du test est analogue à la comparaison de deux variances expérimentales (*cf.* chapitre 12).

MOTS-CLÉS

- Variance expliquée
- Estimation des coefficients d'une droite de régression
- Comparaison de deux droites de régression de Y en X et de Y' en X

EXERCICES

14-1 On a mesuré l'absorption de la lumière par des solutions de 4-nitrophénol, de concentrations croissantes. On a obtenu les résultats suivants (pour une lumière de longueur d'onde 400 nm) :

Concentration C (en mol/L)	1×10^{-5}	2×10^{-5}	3×10^{-5}	4×10^{-5}	5×10^{-5}
Absorbance *A*	0,1865	0,3616	0,5370	0,7359	0,9238

a) Vérifiez graphiquement qu'on peut admettre l'existence d'une relation affine entre l'absorbance et la concentration.

b) En supposant que les hypothèses du cours sont satisfaites, estimez les paramètres de la droite de régression de A par rapport à C

1) ponctuellement,

2) par des intervalles de confiance au risque 5 %.

14-2 Le produit ionique d'un solvant (pK_s) est lié à sa constante diélectrique (ε) par une relation du type :

$$pK_s = \frac{\alpha}{\varepsilon} + \beta \qquad \textbf{(1)}$$

On connaît les résultats suivants :

Solvants	ε	pK_s
eau	78,5	14
éthanol	24,3	19,1
isopropanol	18,3	20,8
méthanol	32,6	16,7

a) Vérifier graphiquement la validité de la relation **(1)** pour ces solvants.

b) Estimez les valeurs de α et de β

1) ponctuellement,

2) par des intervalles de confiance au coefficient de sécurité 0,95.

c) Pour le *n*-propanol, on a : $\epsilon = 20{,}1$. Estimez son pK_s

1) ponctuellement,

2) par un intervalle de confiance au risque 0,05.

14-3 Un corps chimique se décompose selon une cinétique du premier ordre caractérisée par l'équation : $Q = Q_0\, e^{-kt}$ où :
Q désigne la quantité de corps restant à l'instant t,

Q_0 la quantité initiale,

k la constante de vitesse de la décomposition.

On dispose des données expérimentales suivantes :

t (min)	1	2	3	4	5	6	7	8	9	10
Q (nanomoles)	416	319	244	188	144	113	85	66	50	41

En se ramenant à une régression affine, estimez la valeur de k ponctuellement et par un intervalle de confiance au risque 5 %.

14-4 Reprenez les données de l'exercice **14-1**. Peut-on admettre que la relation entre l'absorbance et la concentration est linéaire, c'est-à-dire que, au niveau de la population, la droite de régression passe par l'origine (risque 5 %) ?

14-5 Reprenez les données de l'exercice **14-1**. Comparez la valeur de la pente a obtenue à la valeur $\alpha = 18\,100$ L/mol fournie par les ouvrages de référence sur le sujet (risque 5 %).

14-6 Pour une série de 9 composés organophosphorés, on a étudié la relation entre la constante d'inhibition de la cholinestérase (K_i) et un paramètre B_a caractérisant la basicité des composés.
La relation a été exprimée sous forme d'une droite de régression :

$$K_i = (8{,}1 \pm 3{,}1)B_a + (-13{,}0 \pm 5{,}7)$$

(les paramètres sont donnés sous la forme : valeur estimée ± écart type).

La régression est-elle significative ?

14-7 Reprenez les données de l'exercice **14-1**. Une autre expérience a donné les résultats suivants :

Concentration *C* (en mol/L)	$2{,}5 \times 10^{-5}$	5×10^{-5}	10×10^{-5}
Absorbance *A*	0,396	0,812	1,608

Comparez les pentes des deux droites de régression.

SOLUTIONS

14-1 **a)** En reportant sur un graphique les points (C_i, A_i), on observe qu'ils sont très bien alignés. On peut donc admettre l'existence d'une relation affine entre A et C.

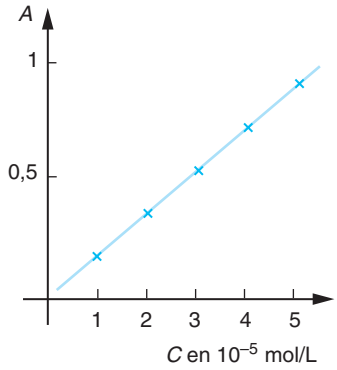


Figure 14-1

b) 1) Soit $A = aC + b$ l'équation de la droite de régression (expérimentale) de A par rapport à C.
On peut obtenir directement avec une calculatrice :

$$a = 18\,489 \quad ; \quad b = -0{,}005\,71.$$

Mais pour la suite, divers résultats intermédiaires seront nécessaires :

$$\overline{C} = 3 \times 10^{-5} \quad ; \quad s_e^2(C) = 2 \times 10^{-10}$$

$$\overline{A} = 0{,}54896 \quad ; \quad s_e^2(A) = 0{,}0684$$

$$r \approx 0{,}99966 \quad ; \quad s_R^2 = \frac{n}{n-2}(1 - r^2)s_e^2(A) \approx 7{,}7457 \times 10^{-5}$$

2) • Soit A la variable aléatoire qui prend la valeur a. Sa variance peut

être estimée par $s_A^2 = \dfrac{s_R^2}{n\,s_e^2(C)} \approx 774\,570$. D'où : $s_A \approx 278{,}3$.

Comme $T = \dfrac{A - \alpha}{s_A}$ suit la loi de Student à 3 d.d.l., le nombre $t_{0,05}$ tel que $P(|T| \geq t_{0,05}) = 0{,}05$ est $t_{0,05} = 3{,}182$.

Au risque 5 %, la pente théorique α appartient donc à l'intervalle de confiance :

$$]a - t_{0,05}\, s_A\, ,\, a + t_{0,05}\, s_A[=]17\,603\ ;\ 19\,375[.$$

• Soit B la variable aléatoire qui prend la valeur b. Sa variance peut être estimée par $s_B^2 = s_R^2\left(\dfrac{1}{n} + \dfrac{\overline{C}^2}{n\, s_e^2(C)}\right) = 1{,}1 s_R^2 \approx 8{,}52 \times 10^{-5}$.

D'où : $s_A \approx 0{,}009\,23$.

Comme $T = \dfrac{B - \beta}{s_B}$ suit la loi de Student à 3 d.d.l., on peut dire, au risque 0,05, que l'ordonnée à l'origine théorique β appartient à l'intervalle de confiance :

$$]b - t_{0,05}\, s_B\, ,\, b + t_{0,05}\, s_B[=]-0{,}0236\ ;\ 0{,}0351[.$$

14-2 **a)** D'après la formule (**1**), pK_s est une fonction affine de $\dfrac{1}{\varepsilon}$. On reporte donc sur un graphique les points de coordonnées :

$X = \dfrac{1}{\varepsilon}$	$y = pK_s$
0,0127	14
0,0412	19,1
0,0546	20,8
0,0307	16,7

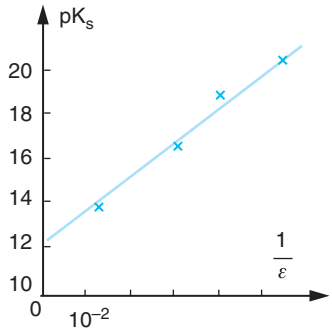


Figure 14-2

On observe que les points sont bien alignés, ce qui confirme la relation (**1**).

b) 1) Avec une calculatrice, on peut obtenir, à partir des points du tableau ci-dessus, la droite de régression qui donne les estimations ponctuelles de α et de β :

$$a \approx 166{,}51 \quad ; \quad b \approx 11{,}86.$$

Mais pour la suite, divers résultats intermédiaires seront nécessaires :

$$\overline{x} \approx 0{,}0348 \quad ; \quad s_e^2(x) \approx 2{,}34 \times 10^{-4}$$

$$\overline{y} = 17{,}65 \quad ; \quad s_e^2(y) \approx 6{,}5625$$

$$r \approx 0{,}9953 \quad ; \quad s_R^2 = \frac{n}{n-2}(1 - r^2)s_e^2(y) \approx 0{,}1237$$

2) • Soit A la variable aléatoire qui prend la valeur a. Sa variance peut être estimée par : $s_A^2 = \dfrac{s_R^2}{n s_e^2(x)} \approx 131{,}94$, d'où : $s_A \approx 11{,}49$.

Comme $T = \dfrac{A - \alpha}{s_A}$ suit la loi de Student à 2 d.d.l., pour $\alpha = 0{,}05$, on a : $t_{0,05} = 4{,}303$. Et l'intervalle de confiance, au risque 5 %, de la pente théorique α est :]117,08 ; 215,93[.

• Soit B la variable aléatoire qui prend la valeur b. Sa variance peut être estimée par : $s_B^2 = s_R^2\left(\dfrac{1}{n} + \dfrac{\overline{x}^2}{n s_e^2(x)}\right) \approx 0{,}19$, d'où : $sB \approx 0{,}44$.

Comme $T = \dfrac{B - \beta}{s_B}$ suit la loi de Student à 2 d.d.l., on peut dire, au coefficient de sécurité 0,95, que l'ordonnée à l'origine théorique β appartient à l'intervalle de confiance :]9,98; 13,73[.

c) 1) À partir de la droite de régression estimée : $y \approx 166{,}51x + 11{,}86$, si on sait que $x_0 = \dfrac{1}{20{,}1}$, on obtient l'estimation ponctuelle de y : $\hat{y}_0 \approx 20{,}1$.

2) Au risque 5 %, l'intervalle de confiance pour le pK_s attendu pour le *n*-propanol est :

$$\left]\hat{y}_0 - t_{0,05}\sqrt{s_R^2\left(1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{n s_e^2(x)}\right)}, \hat{y}_0 + t_{0,05}\sqrt{s_R^2\left(1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{n s_e^2(x)}\right)}\right[$$

où $t_{0,05} = 4{,}303$ correspond à la loi de Student à 2 d.d.l.

Tous calculs faits, on obtient l'intervalle :]18,3; 22,0[.

Le pK_s du *n*-propanol est égal à 19,4. Cette valeur appartient bien à l'intervalle de confiance obtenu.

14-3 Le modèle proposé $Q = Q_0\, e^{-kt}$ peut aussi s'écrire : $\ln Q = \ln Q_0 - kt$.

La relation entre t et $\ln Q$ est donc affine, et $-k$ est la pente de la droite de régression théorique.

➤ Avec $x = t$ et $y = \ln Q$, à partir des 10 points $(t_i, \ln Q_i)$, on obtient successivement :

$\overline{x} = 5{,}5$; $s_e(x) \approx 2{,}87$; $\overline{y} \approx 4{,}85$; $s_e(y) \approx 0{,}75$; $r \approx -0{,}9998$ et de $a \approx -0{,}2605$ on déduit l'estimation ponctuelle :

$k = 0{,}2605\ \text{min}^{-1}$.

➤ On obtient ensuite :

$s_R^2 = \dfrac{n}{n-2}(1-r^2)s_e^2(y) \approx 3{,}24 \times 10^{-4}$, puis :

$s_A^2 = \dfrac{s_R^2}{n\,s_e^2(x)} \approx 3{,}9 \times 10^{-6}$ et $s_A \approx 1{,}98 \times 10^{-3}$.

Au risque 5 %, l'intervalle de confiance de la pente théorique s'écrit :

$$]a - t_{0,05}\, s_A\, ,\, a + t_{0,05} s_A[$$

où $t_{0,05} = 2{,}306$ correspond à la loi de Student à 8 d.d.l.

On obtient ainsi pour intervalle de confiance de k :]0,2559 ; 0,2651[.

14-4 On suppose que les hypothèses du cours sont satisfaites, et on écrit $y = \alpha x + \beta$ l'équation de la droite de régression théorique.

On teste (H_0) : la différence entre la valeur obtenue $b = -0{,}005\,71$ et la valeur théorique $\beta = 0$ est explicable par les aléas dus à l'échantillonnage.

On sait que $T = \dfrac{B - \beta}{s_B}$ suit la loi de Student à $n - 2 = 3$ degrés de liberté.

On a : $t = \dfrac{b - 0}{s_B} \approx -0{,}62$.

Par ailleurs : $t_{0,05} = 3{,}182$.

Comme $t \in]-t_{0,05}, t_{0,05}[$, (H_0) n'est pas rejetée au risque 5 %. On peut donc admettre que la droite de régression théorique passe par l'origine, c'est-à-dire que la relation entre l'absorbance et la concentration est linéaire.

14-5 L'hypothèse nulle s'écrit :

(H_0) : la différence entre la valeur théorique $\alpha = 18\,100$ et la valeur expérimentale $a = 18\,489$ est explicable par les fluctuations d'échantillonnage.

On sait que $T = \dfrac{A - \alpha}{s_A}$ suit la loi de Student à $n - 2 = 3$ d.d.l.

On calcule la valeur prise par T : $t = \dfrac{a - \alpha}{s_A} \approx 1{,}40$.

Par ailleurs : $t_{0,05} = 3{,}182$.

Comme $t \in]-t_{0,05}, t_{0,05}[$, (H_0) est acceptée.

La valeur obtenue ne diffère pas significativement de la valeur de référence.

14-6 Les informations fournies s'écrivent :

$$b = -13{,}0 \quad ; \quad s_B = 5{,}7 \quad ; \quad a = 8{,}1 \quad ; \quad s_A = 3{,}1$$

L'hypothèse nulle à tester s'écrit $(H_0) : \alpha = 0$.

On sait que $T = \dfrac{A - \alpha}{s_A}$ suit la loi de Student à $n - 2 = 7$ d.d.l.

La valeur prise par T est $t = \dfrac{8{,}1 - 0}{3{,}1} \approx 2{,}61$.

Par ailleurs, en choisissant un risque de 5 %, on a : $t_{0{,}05} = 2{,}365$.

Comme $t \notin]-t_{0{,}05}, t_{0{,}05}[$, (H_0) est rejetée au risque 5 %. La régression est significative au risque 5 %.

14-7 Les hypothèses du cours sont supposées vérifiées.

• Calculs

Soit $A = a'C + b'$ l'équation de la droite de régression associée aux résultats de la deuxième expérience. Les calculs habituels conduisent à :

$n' = 3 \quad ; \quad \overline{C}' \approx 5{,}83 \times 10^{-5} \quad ; \quad s'_e(C) \approx 3{,}1180 \times 10^{-5} \quad ;$

$\overline{A}' \approx 0{,}9387 \quad ; \quad s'_e(A) \approx 0{,}5028 \quad ; \quad r' \approx 0{,}9999 \quad ;$

$a' \approx 16\,126 \quad ; \quad b' = -0{,}002 \quad ;$

$$s'^2_R = \frac{n'}{n' - 2}(1 - r'^2)s'^2_e(A) \approx 9{,}257 \times 10^{-5}.$$

• Test préalable : comparaison des variances résiduelles (*cf.* annexe du cours)

On teste $(H_0) : \sigma^2 = \sigma'^2$.

On sait qu'alors $F = \dfrac{S'^2_R}{S^2_R}$ suit la loi de Snedecor à $(n' - 2, n - 2)$ degrés de liberté.

On a : $f = \dfrac{s'^2_R}{s^2_R} \approx 1{,}20$ (les deux variances résiduelles estimées ont été permutées de sorte que leur quotient soit supérieur à 1).

Pour (1 ; 3) degrés de liberté, et un risque de 5 %, on lit dans la table 5 : $f_{0{,}05} = 17{,}4$.

Comme $f < f_{0{,}05}$, (H_0) est non rejetée. On peut accepter l'hypothèse d'égalité des variances résiduelles.

La variance résiduelle commune aux deux populations est alors estimée par :

$$\hat{\sigma}^2 = \frac{3s^2_R + s'^2_R}{4} \approx 8{,}1236 \times 10^{-5}$$

• Comparaison des pentes des deux droites de régression

On teste $(H_0) : \alpha = \alpha'$.

Dans ce cas, la variable aléatoire : $T = \dfrac{A - A'}{\sqrt{\hat{\sigma}^2\left(\dfrac{1}{n s_e^2(C)} + \dfrac{1}{n' s_e'^2(C)}\right)}}$ suit la loi de Student à $n + n' - 4 = 4$ degrés de liberté.

La valeur prise par T est $t \approx 7{,}15$.

D'après la table 3, on a : $t_{0,05} = 2{,}776$; $t_{0,01} = 4{,}604$.

Comme $t > t_{0,01}$, la différence des pentes est significative au risque 1 %.

Il est possible qu'une impureté ait contaminé le deuxième échantillon de 4-nitrophénol.

CHAPITRE 15

Corrélation

PLAN

15.1 Estimation d'un coefficient de corrélation
15.2 Comparaison d'un coefficient de corrélation expérimental à une valeur théorique
15.3 Comparaison de deux coefficients de corrélation expérimentaux
15.4 Comparaison de plusieurs coefficients de corrélation expérimentaux

OBJECTIFS

- Estimer, par un nombre ou un intervalle, la force de la liaison entre deux caractères numériques
- Étudier si une force de liaison observée diffère d'une valeur de référence
- Savoir si une grandeur X est plus liée à Y qu'à Z

15.1 ESTIMATION D'UN COEFFICIENT DE CORRÉLATION

Problématique

Sur une population, on considère deux variables aléatoires X et Y telles que :

- ou bien X est une variable contrôlée, Y une variable dépendante vérifiant les hypothèses du chapitre 14, et la régression de Y par rapport à X est affine ;
- ou bien le couple (X, Y) suit une loi normale à deux dimensions.

Soit ρ le coefficient de corrélation entre X et Y dans la population. Le problème consiste à estimer ρ.

Estimation ponctuelle de ρ

Notations

On tire de la population un échantillon de n couples (x_i, y_i) et on lui associe son coefficient de corrélation : $r = \dfrac{\text{Cov}(x,y)}{s_e(x)s_e(y)}$

où $\text{Cov}(x,y) = \dfrac{1}{n}\left(\sum_{i=1}^{n} x_i\, y_i\right) - \overline{x}\,\overline{y}$ est la covariance de l'échantillon, et où $s_e(x)$ et $s_e(y)$ sont les écarts type des échantillons respectifs $\{x_1,...,x_n\}$ et $\{y_1,...,y_n\}$.

Soit R la variable aléatoire qui prend la valeur r quand on répète les échantillons de taille n.

Théorème

$$E(R) \approx \rho + \frac{\rho(1-\rho^2)}{2(n-1)}$$

Estimation ponctuelle de ρ

En général, on retient r comme estimation ponctuelle de ρ.

Parfois, on utilise une estimation plus précise : $r\left(1 + \dfrac{1-r^2}{2(n-3)}\right)$.

Estimation de ρ par un intervalle de confiance

Notations

Soit z' le nombre défini par $z' = \dfrac{1}{2}\ln\left(\dfrac{1+r}{1-r}\right) = \text{arg th } r$ (lire : *argument tangente hyperbolique de r*), et Z' la variable aléatoire qui prend la valeur z quand on répète les échantillons de taille n.

Soit ζ le nombre défini par $\zeta = \dfrac{1}{2}\ln\left(\dfrac{1+\rho}{1-\rho}\right) = \text{arg th } \rho$.

Théorème. Avec les hypothèses déjà indiquées, pour n assez grand, Z' suit à peu près la loi normale $\mathcal{N}\left(\zeta; \dfrac{1}{\sqrt{n-3}}\right)$.

Cette approximation est convenable pour $n \geq 20$.

Intervalle de confiance de ρ

On déduit du théorème l'intervalle de confiance de ζ, au risque α :

$$\left]z' - \frac{z_\alpha}{\sqrt{n-3}} ; z' + \frac{z_\alpha}{\sqrt{n-3}}\right[=]z_1, z_2[$$

où z_α se lit dans la table 2.

On peut en déduire un intervalle contenant ρ avec une probabilité $1 - \alpha$:

$$]r_1, r_2[=]\text{th}\, z_1, \text{th}\, z_2[$$

où la fonction th (lire : *argument tangente hyperbolique*) est définie par : $\text{th}\, z = \dfrac{1 - e^{-2z}}{1 + e^{-2z}}$.

15.2 COMPARAISON D'UN COEFFICIENT DE CORRÉLATION EXPÉRIMENTAL À UNE VALEUR THÉORIQUE

Problématique

Les hypothèses étant les mêmes que précédemment, on dispose d'un échantillon de n couples (x_i , y_i) dont le coefficient de corrélation est r.

Peut-on considérer que cet échantillon est tiré d'une population où le coefficient de corrélation est ρ ?

L'hypothèse nulle est donc :

(H_0) : l'échantillon est extrait de la population ; la différence entre ρ et r n'est pas significative.

Cas $\rho = 0$

Dans ce cas, si (H_0) est vraie, $T = \dfrac{R\sqrt{n-2}}{\sqrt{1-R^2}}$ suit la loi de Student à $n - 2$ degrés de liberté.

Ce théorème permet d'établir une table, numérotée 10 dans ce livre, qui donne directement la borne r_α telle que $P(|R| \geq r_\alpha) = \alpha$.

- Si $r \in]-r_\alpha, r_\alpha[$, (H_0) ne peut pas être rejetée.
- Si $r \notin]-r_\alpha, r_\alpha[$, (H_0) est rejetée au risque α.

Si (H_0) est rejetée, cela entraîne que les variables aléatoires X et Y ne sont pas indépendantes.

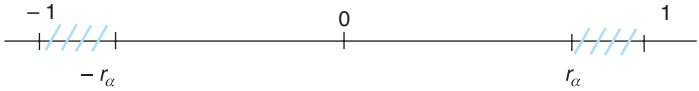


Figure 15-1

Cas $\rho \neq 0$

Sous (H_0) et les hypothèses déjà indiquées, $Z = (Z' - \zeta)\sqrt{n-3}$ suit à peu près $\mathcal{N}(0, 1)$.

On calcule donc : $z = (z' - \zeta)\sqrt{n-3}$.

Par ailleurs, la table 2 fournit le nombre z_α, tel que $P(|Z| \geq z_\alpha) = \alpha$.

- Si $z \in]-z_\alpha, z_\alpha[$, (H_0) ne peut pas être rejetée.
- Si $z \notin]-z_\alpha, z_\alpha[$, (H_0) est rejetée au risque α.

15.3 COMPARAISON DE DEUX COEFFICIENTS DE CORRÉLATION EXPÉRIMENTAUX

Problématique

On considère deux populations, vérifiant les hypothèses déjà indiquées, où les coefficients de corrélation (inconnus) sont ρ_1 et ρ_2.

On dispose de deux échantillons de tailles n_1 et n_2, et de coefficients de corrélation respectifs r_1 et r_2.

On teste $(H_0) : \rho_1 = \rho_2$.

Exécution du test

Notations

Soit $z'_1 = \frac{1}{2}\ln\left(\frac{1+r_1}{1-r_1}\right)$, $z'_2 = \frac{1}{2}\ln\left(\frac{1+r_2}{1-r_2}\right)$ et Z'_1 et Z'_2 les les variables aléatoires correspondantes.

Théorème. Si (H_0) est vraie et si n_1 et n_2 sont assez grands (≥ 20), alors

$$U = \frac{Z'_1 - Z'_2}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}} \text{ suit } \mathcal{N}(0, 1).$$

Le test de (H_0) en résulte comme d'habitude quand une variable aléatoire suit la loi normale centrée réduite.

15.4 COMPARAISON DE PLUSIEURS COEFFICIENTS DE CORRÉLATION EXPÉRIMENTAUX

Problématique

On dispose de k populations ($k > 2$), vérifiant les hypothèses déjà indiquées, où les coefficients de corrélation (inconnus) sont $\rho_1,...,\rho_k$.

On en extrait k échantillons de tailles respectives $n_1,...,n_k$ et de coefficients de corrélation $r_1,...,r_k$.

On désire comparer globalement les coefficients de corrélation, ce qui conduit à tester $(H_0) : \rho_1 = \rho_2 = \cdots = \rho_k$.

Exécution du test

Notations

Pour i variant de 1 à k, on détermine les nombres $z'_i = \frac{1}{2}\ln\left(\frac{1+r_i}{1-r_i}\right)$, puis leur moyenne pondérée :

$$\overline{z}' = \frac{\sum_{i=1}^{k}(n_i - 3)z'_i}{\sum_{i=1}^{k}(n_i - 3)}.$$

On note Z'_i et $\overline{Z}'$ les variables aléatoires correspondantes.

Théorème. Si (H_0) est vraie, et si $n_1,\ldots,n_k$ sont assez grands (≥ 20), alors $Y = \sum_{i=1}^{k}(n_i - 3)(Z'_i - \overline{Z}')^2$ suit la loi du χ^2 à $k - 1$ degrés de liberté.

On calcule donc :

$$y = \sum_{i=1}^{k}(n_i - 3)(z'_i - \overline{z}')^2 = \sum_{i=1}^{k}(n_i - 3)z'^2_i - \overline{z}'^2\sum_{i=1}^{k}(n_i - 3)$$

Par ailleurs, le degré de liberté étant connu et le risque α étant choisi, la table 4 donne le nombre χ^2_α tel que $P(Y \geq \chi^2_\alpha) = \alpha$.

- Si $y \leq \chi^2_\alpha$, on ne peut pas rejeter (H_0).
- Si $y > \chi^2_\alpha$, on rejette (H_0) au risque α.

Fonctions th et argth

• La fonction **tangente hyperbolique** est définie de $\mathbb{R}$ dans $]-1; 1[$ par :

$$\text{th}\, x = \frac{e^x - e^{-x}}{e^x + e^{-x}} = \frac{e^{2x} - 1}{e^{2x} + 1}$$

Elle est impaire, dérivable et

$\forall x \in \mathbb{R}\ (\text{th}\, x)' = 1 - \text{th}^2 x$

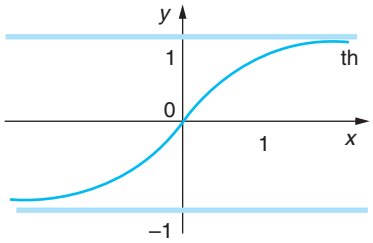


Figure 15-2

• La fonction **argument tangente hyperbolique** est la réciproque de th.

Elle est définie de $]-1\ ;\ 1[$ dans $\mathbb{R}$, impaire, dérivable et :

$$\forall x \in\,]-1\ ;\ 1[\quad (\text{argth})'(x) = \frac{1}{1 - x^2}$$

Elle a pour expression logarithmique :

$$\forall x \in\,]-1\ ;\ 1[\quad \text{argth}\, x = \frac{1}{2} \ln\left(\frac{1+x}{1-x}\right)$$

MOTS-CLÉS

- Estimation d'un coefficient de corrélation
- Indépendance de deux caractères quantitatifs
- Comparaison de k $(k > 2)$ coefficients de corrélation expérimentaux

EXERCICES

15-1 Le coefficient de corrélation d'un échantillon de 100 valeurs est 0,53. Calculez l'intervalle de confiance pour le coefficient de corrélation de la population correspondante, aux risques 5 %, puis 1 %, puis 0,1 %.

15-2 Afin d'estimer le coefficient de corrélation entre deux variables dans une population donnée, on tire une série d'échantillons indépendants, d'effectifs croissants. On obtient les coefficients de corrélation expérimentaux suivants :

Échantillon	n° 1	n° 2	n° 3	n° 4
Taille	10	20	50	100
r	0,80	0,52	0,75	0,68

À partir de chaque échantillon, estimez ρ par un intervalle de confiance au risque 5 %. Représentez graphiquement chaque intervalle en y faisant figurer la valeur de r. Quelles réflexions vous inspirent les résultats ?

15-3 On cherche à estimer le coefficient de corrélation entre deux variables dans une population humaine, au risque 5 %. Une première estimation, portant sur un échantillon de 100 personnes, fournit la valeur 0,60.

Quel est le nombre minimal de personnes qu'il faudrait examiner pour pouvoir estimer la valeur de ρ à ± 10 % près ?

On admet que, dans ces conditions, l'intervalle de confiance de ρ est centré sur r, et on rappelle que la dérivée de argth x est $\dfrac{1}{1-x^2}$.

15-4 D'une population caractérisée par un coefficient de corrélation de 0,75, on extrait un échantillon de 30 individus. Le coefficient de corrélation de l'échantillon est 0,82.

L'échantillon peut-il être considéré comme représentatif de la population ?

15-5 On étudie la corrélation entre les activités de deux enzymes sériques. On a obtenu :

– dans l'espèce humaine, $r = -0{,}296$ pour un échantillon de 30 individus,

– dans l'espèce bovine, $r = 0{,}452$ pour un échantillon de taille 21.

Pour chacune des deux espèces, les corrélations observées sont-elles significativement différentes de $\rho = 0$?

15-6 Une expérience a été faite sur 20 grenouilles mâles, choisies pour leur extrême noirceur ou leur extrême pâleur, pour essayer de voir s'il existe une relation entre la teneur en mélanine de la peau de ces grenouilles et leurs poids. On désigne par X la densité de la mélanine et par Y le poids de la grenouille exprimé en grammes. On a obtenu :

X	0,11	0,15	0,32	0,68	0,64	0,29	0,45	0,51	0,05	0,71
Y	11	19	20	18	17	22	25	24	21	26

X	0,37	0,56	0,97	0,75	0,77	0,86	1,04	0,74	0,32	0,64
Y	28	30	31	23	25	27	29	17	15	25

a) 1) Calculez le coefficient de corrélation r de X et Y.

2) Testez l'hypothèse (H_0) : « la valeur trouvée n'est pas significative. Elle est due au simple hasard » ; autrement dit : il n'y a pas de différence significative entre la valeur r obtenue et la valeur théorique $\rho = 0$.

On prendra $\alpha = 0{,}05$.

b) Déterminez l'intervalle de confiance, au seuil de sécurité 0,95, du coefficient de corrélation de la population-mère.

15-7 Deux laboratoires hospitaliers indépendants étudient la corrélation entre le résultat d'un certain test biologique et l'âge des malades.

Le premier laboratoire obtient $r_1 = 0{,}80$ pour un échantillon de 30 malades.

Le deuxième laboratoire obtient $r_2 = 0{,}95$ sur 50 malades.

La différence entre les deux laboratoires est-elle significative ?

15-8 Deux lots de porcs, A et B, contenant respectivement 26 et 34 porcs, ont été extraits au hasard d'une population de porcs dont on a suivi l'évolution du gain de poids (variable notée Y) et la quantité de nourriture absorbée (variable notée X) pendant une période de 20 jours consécutifs.

On a calculé le coefficient de corrélation entre X et Y dans chacun des deux échantillons, et on a obtenu :

$r_1 = 0{,}85$ pour l'échantillon A ; $r_2 = 0{,}63$ pour l'échantillon B.

Comparez ces deux coefficients de corrélation.

15-9 Reprenez les données de l'exercice **15-9**. Montrez que l'on peut admettre, au risque 5 %, que les 4 échantillons considérés sont bien tirés de la même population.

Quelle est alors la meilleure estimation du coefficient de corrélation de cette population ?

15-10 Lors d'une étude écologique portant sur la répartition géographique d'une certaine espèce d'escargots, on a mesuré le coefficient de

corrélation entre la hauteur et la largeur des coquilles, pour des échantillons d'origines géographiques différentes. Les résultats obtenus sont les suivants :

taille de l'échantillon	125	125	30	200	200
r	0,96	0,89	0,98	0,98	0,97

Peut-on dire que les cinq échantillons sont tirés de la même population ?

SOLUTIONS

15-1 On a : $z = \operatorname{argth} 0{,}53 = \frac{1}{2}\ln\left(\frac{1+0{,}53}{1-0{,}53}\right) \approx 0{,}59.$

En supposant que la population vérifie les hypothèses du cours, et après avoir lu z_α dans la table 2, on obtient successivement :
– l'intervalle de confiance de $\zeta = \operatorname{argth} \rho$:

$$\left]z - \frac{z_\alpha}{\sqrt{97}}, z + \frac{z_\alpha}{\sqrt{97}}\right[=]z_1, z_2[$$

– puis l'intervalle de confiance de ρ au risque α :
$]r_1\ ,\ r_2[=]\text{th}\ z_1\ ,\ \text{th}\ z_2[$.
On peut regrouper les résultats dans un tableau :

α	u_α	$]z_1, z_2[$	$]r_1, r_2[$
0,05	1,960	]0,39 ; 0,79[	]0,37 ; 0,66[
0,01	2,576	]0,33 ; 0,85[	]0,32 ; 0,69[
0,001	3,291	]0,25 ; 0,93|	]0,25 ; 0,73[

15-2 Comme dans l'exercice précédent, on calcule successivement : $z = \operatorname{argth} r = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$ puis l'intervalle de confiance de ζ : $]z_1, z_2[= \left]z - \frac{z_\alpha}{\sqrt{n-3}}, z + \frac{z_\alpha}{\sqrt{n-3}}\right[$

Échantillon	***n***	***r***	***z***	$]z_1, z_2[$	$]r_1, r_2[$
n° 1	10	0,80	1,10	]0,36 ; 1,84[	]0,34 ; 0,95[
n° 2	20	0,52	0,58	]0,10 ; 1,05[	]0,10 ; 0,78[
n° 3	50	0,75	0,97	]0,69 ; 1,26[	]0,60 ; 0,85[
n° 4	100	0,68	0,83	]0,63 ; 1,03[	]0,56 ; 0,77[

puis l'intervalle de confiance de ρ : $]r_1, r_2[=]$ th z_1, th $z_2[$
La représentation graphique suggère les remarques :
– lorsque la taille de l'échantillon augmente, pour un niveau de risque fixé, l'intervalle de confiance se rétrécit, c'est-à-dire que l'estimation devient plus précise ;
– les intervalles de confiance de ρ ne sont pas centrés sur l'estimation ponctuelle r.

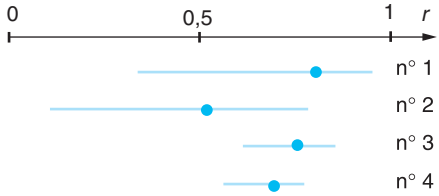


Figure 15-3

Toutefois, lorsque la taille de l'échantillon augmente, l'intervalle de confiance a tendance à devenir symétrique par rapport à r.

15-3 Si l'on estime ρ par la valeur 0,60, l'intervalle de confiance doit avoir une demi-amplitude $\Delta r = 0{,}06$ (estimation à ± 10 % près).
On en déduit la demi-largeur de l'intervalle de confiance de ζ :

$$\Delta z = \Delta(\operatorname{argth} r) \approx \frac{\mathrm{d}}{\mathrm{d}x}(\operatorname{argth} x) \cdot \Delta r = \frac{\Delta r}{1 - r^2}$$

soit :

$$\Delta z \approx \frac{0{,}06}{1 - 0{,}6^2} \approx 0{,}094.$$

On sait que :

$$\Delta z = \frac{u_\alpha}{\sqrt{n-3}} \quad \text{soit} \quad \Delta z = \frac{1{,}96}{\sqrt{n-3}} \quad \text{pour } \alpha = 0{,}05.$$

De $\dfrac{1{,}96}{\sqrt{n-3}} = 0{,}094$ on tire $n \approx 440$.
Il faut donc examiner environ 440 personnes.

15-4 On teste (H_0) : l'échantillon est extrait de la population ; la différence entre $\rho = 0{,}75$ et $r = 0{,}82$ n'est pas significative.
En supposant que la population vérifie les hypothèses du cours, si (H_0) est vraie, on sait que $Z = (Z - \zeta)\sqrt{n-3}$ suit $\mathcal{N}(0; 1)$.
On a :
$z = \operatorname{argth} r = \operatorname{argth} 0{,}82 \approx 1{,}16$; $\zeta = \operatorname{argth} \rho = \operatorname{argth} 0{,}75 \approx 0{,}97$.
D'où : $u = (z - \zeta)\sqrt{n-3} \approx 0{,}96$. Par ailleurs, on a $u_{0,05} = 1{,}96$.
Comme $u \in]-u_{0,05}, u_{0,05}[$, on ne peut pas rejeter (H_0) au risque 5 %.
L'échantillon peut être considéré comme représentatif de la population.

15-5 Pour chacune des deux espèces, on teste (H_0) : la différence entre la valeur observée r et la valeur théorique $\rho = 0$ n'est pas significative.

Dans ce cas, on sait que $T = \dfrac{R\sqrt{n-2}}{\sqrt{1-R^2}}$ suit la loi de Student à $n-2$ degrés de liberté.

➤ Pour l'espèce humaine, on a : $t = \dfrac{-0{,}296\sqrt{28}}{\sqrt{1-(-0{,}296)^2}} \approx -1{,}64$.

Pour $\alpha = 0{,}05$ et 28 d.d.l. on a : $t_\alpha = 2{,}048$. Comme $t \in]-t_\alpha, t_\alpha[$, on ne peut pas rejeter (H_0). On dit parfois que la corrélation n'est pas significative.

➤ Pour l'espèce bovine, on a : $t = \dfrac{0{,}452\sqrt{19}}{\sqrt{1-0{,}452^2}} \approx 2{,}21$.

Pour $\alpha = 0{,}05$ et 19 d.d.l. on a : $t_\alpha = 2{,}093$. Comme $t \notin]-t_\alpha, t_\alpha[$, on rejette (H_0) au risque 5 %. Cela entraîne donc que les activités des deux enzymes sériques étudiés ne sont pas indépendantes dans le cas de l'espèce bovine.

Il est plus rapide d'utiliser la table 10, construite à partir du théorème qui vient d'être rappelé.

- Pour l'espèce humaine, on a $r = -0{,}296$ et on lit, pour $n-2 = 28$ la borne $r_{0,05} \approx 0{,}4$. Comme $|r| < r_{0,05}$, on ne rejette pas (H_0).
- Pour l'espèce bovine, on a $r = 0{,}452$ et on lit, pour $n-2 = 19$ la borne $r_{0,05} = 0{,}4329$. Comme $|r| > r_{0,05}$, on rejette (H_0) au risque 5 %.

15-6 **a) 1)** On obtient directement avec une calculatrice $r \approx 0{,}55$.

2) Pour $\alpha = 0{,}05$ et $n-2 = 18$, on lit, dans la table 10, la borne $r_{0,05} = 0{,}4438$.
Comme $|r| > r_{0,05}$, on rejette (H_0) au risque 5 %. On peut dire que la corrélation observée est significative.

b) On a : $z = \operatorname{argth} r \approx 0{,}62$. En supposant que la population vérifie les hypothèses du cours, on obtient successivement :

➤ l'intervalle de confiance de $\zeta = \operatorname{argth} \rho$:

$$]z_1, z_2[= \left]z - \frac{z_\alpha}{\sqrt{17}} ; z + \frac{z_\alpha}{\sqrt{17}}\right[\approx]0{,}15 ; 1{,}10[\quad (\text{car } z_\alpha = 1{,}96)$$

➤ l'intervalle de confiance de ρ :

$$]r_1 , r_2[=]\operatorname{th} z_1 , \operatorname{th} z_2[\approx]0,\ 15 ;\ 0{,}80[.$$

On peut donc dire, au risque 5 %, que ρ, coefficient de corrélation de la population, appartient à]0,15 ; 0, 80[. On retrouve la question précédente en remarquant que cet intervalle ne contient pas 0.

15-7 On teste $(H_0) : \rho_1 = \rho_2$; la différence entre les deux laboratoires n'est pas significative.

On suppose que les populations vérifient les hypothèses du cours et on observe que n_1 et n_2 sont assez grands.

Dans ce cas, on sait que $Z = \dfrac{Z_1 - Z_2}{\sqrt{\dfrac{1}{n_1 - 3} + \dfrac{1}{n_2 - 3}}}$ suit $\mathcal{N}(0, 1)$.

On obtient successivement :

$z_1 = \text{argth } r_1 = \dfrac{1}{2}\ln\left(\dfrac{1{,}8}{0{,}2}\right) \approx 1{,}10$; $z_2 = \text{argth } 0{,}95 \approx 1{,}83$;

$u \approx -3{,}04$.

Pour $\alpha = 0{,}05$, on a $z_{0{,}05} = 1{,}96$.

Pour $\alpha = 0{,}01$, on a $z_{0{,}01} = 2{,}576$.

Dans tous les cas, $z \notin]-z_\alpha, z_\alpha[$ et (H_0) est rejetée au risque 1 %

La différence entre les deux laboratoires est donc significative au risque 1 %.

15-8 On teste (H_0) : $\rho_1 = \rho_2$; la différence entre r_1 et r_2 est explicable par les fluctuations d'échantillonnage. On a :

$n_1 = 26$; $n_2 = 34$; $z_1 = \text{argth } r_1 \approx 1{,}26$; $z_2 = \text{argth } r_2 \approx 0{,}74$.

$$z = \frac{z_1 - z_2}{\sqrt{\dfrac{1}{23} + \dfrac{1}{31}}} \approx 1{,}87.$$

Pour $\alpha = 0{,}05$, on sait que $z_{0{,}05} = 1{,}96$.

Comme $z \in]-z_\alpha, z_\alpha[$, (H_0), n'est pas rejetée. La différence entre r_1 et r_2 n'est donc pas significative au risque 5 %

15-9 On teste (H_0) : $\rho_1 = \rho_2 = \rho_3 = \rho_4$; les 4 échantillons sont extraits de la même population.

On suppose que les populations vérifient les hypothèses du cours, et l'on va accepter d'utiliser le théorème énoncé bien que $n_1 = 10$ soit faible.

Nous avons déjà calculé les valeurs des z_i. Nous pouvons donc calculer $\overline{z}$:

$$\overline{z} = \frac{7 \times 1{,}10 + 17 \times 0{,}58 + 47 \times 0{,}97 + 97 \times 0{,}83}{168} \approx 0{,}86.$$

On sait que $Y = \sum_{i=1}^{k} (n_i - 3)(Z_i - \overline{Z})^2$ suit la loi de χ^2 à $k - 1 = 3$ degrés de liberté.

Cette variable aléatoire prend la valeur :

$$y = 7(z_1 - \overline{z})^2 + 17(z_2 - \overline{z})^2 + 47(z_3 - \overline{z})^2 + 97(z_4 - \overline{z})^2 \approx 2{,}45.$$

Pour $\alpha = 0{,}05$ et 3 d.d.l., la borne est $\chi^2_{0,05} = 7{,}81$.
Comme $2{,}45 < 7{,}81$, on ne peut pas rejeter (H_0) au risque 5 %.
Les 4 échantillons peuvent être considérés comme issus de la même population.
La meilleure estimation du coefficient de corrélation de cette population s'obtient alors à partir de $\overline{z}$, soit : $r = \text{th}\,\overline{z} \approx 0{,}69$.

15-10 $(H_0) : \rho_1 = \cdots = \rho_5$, les échantillons sont extraits de la même population. On calcule successivement :

$z_1 = \text{argth}\, r_1 \approx 1{,}95$; $z_2 = \text{argth}\, r_2 \approx 1{,}42$; $z_3 = z_4 \approx 2{,}30$; $z_5 \approx 2{,}09$

$$\text{puis : } \overline{z} = \frac{122\, z_1 + 122\, z_2 + 27\, z_3 + 197\, z_4 + 197\, z_5}{665} \approx 2{,}01.$$

La valeur prise par $Y = \sum_{i=1}^{k} (n_i - 3)(Z_i - \overline{Z})^2$ est :

$$y = 122(z_i - \overline{z})^2 + 122(z_2 - \overline{z})^2 + 27(z_3 - \overline{z})^2 + 197(z_4 - \overline{z})^2 + 197(z_5 - \overline{z})^2$$
$$\approx 62{,}55.$$

Pour $v = 4$, on lit : $\chi^2_{0,05} = 9{,}49$; $\chi^2_{0,01} = 13{,}28$; $\chi^2_{0,001} = 18{,}47$.

Dans tous les cas, on a $y > \chi^2_\alpha$ et (H_0) est rejetée.

La différence entre les cinq échantillons est donc significative, même au risque minime de 0,1 %.

CHAPITRE 16

Tests non paramétriques

PLAN

16.1 Introduction
16.2 Test de Mann et Whitney
16.3 Test de Wilcoxon
16.4 Test de Kruskal et Wallis
16.5 Coefficient de corrélation de rang de Spearman

OBJECTIFS

- Savoir comparer deux, ou plus, moyennes d'échantillons dans le cas où ils sont de petites tailles et extraits de populations inconnues
- Conclure sur l'indépendance de deux caractères quantitatifs dans le cas où les populations sont nouvelles et les observations en petit nombre

16.1 INTRODUCTION

Les tests classiques de comparaison de moyennes et de variances, ainsi que l'analyse de la variance, ne s'appliquent en toute rigueur qu'à des échantillons issus de populations normales. En général, le non-respect de cette condition n'a pas trop d'influence sur la validité du test (sauf en ce qui concerne la comparaison des variances). Lorsque l'effectif des échantillons est faible, l'erreur commise peut toutefois être importante.

On préfère alors utiliser un autre type de tests, valables quelle que soit la nature des populations dont sont tirés les échantillons. Ces tests sont dits **non-paramétriques** car ils ne nécessitent pas l'estimation des paramètres (moyenne et écart type) des populations.

Nous étudierons dans ce chapitre quatre tests non-paramétriques :

- Le **test de Mann et Whitney**, qui permet de comparer les moyennes de deux échantillons indépendants (c'est l'analogue non-paramétrique du test de Student).
- Le **test de Wilcoxon**, qui permet de comparer les moyennes de deux échantillons appariés.
- Le **test de Kruskal et Wallis**, qui permet de comparer les moyennes de plusieurs échantillons (c'est l'analogue non-paramétrique de l'analyse de la variance à un facteur).
- Un test non-paramétrique de corrélation: le **test de Spearman**.

Ces quatre tests ont en commun le fait que les valeurs observées sont remplacées par leurs rangs au sein des échantillons : ce sont donc des **tests de rangs**.

16.2 TEST DE MANN ET WHITNEY

Problématique

On dispose de deux échantillons, indépendants et non-exhaustifs, E_1 et E_2, de tailles respectives n_1 et n_2. On veut comparer les deux moyennes expérimentales, c'est-à-dire tester l'hypothèse nulle $(H_0) : \mu_1 = \mu_2$.

Mise en place du test

- On classe par ordre croissant l'ensemble des valeurs des deux échantillons en repérant l'origine de chaque valeur.
- On affecte à chaque valeur de $E_1 \cup E_2$, son rang dans ce classement. S'il y a des ex-aequo, on attribue à chacun un rang égal à la moyenne des rangs qu'ils occupent (par exemple, s'il y a deux quatrièmes ex-aequo, on attribue à chacun d'eux le rang 4,5).
- Pour tout élément x_i de E_1, on compte le nombre d'éléments de E_2 situés après x_i, (en comptant pour 0,5 tout élément de E_2 ex-aequo avec x_i).
- On note m_1 la somme de toutes les valeurs ainsi associées à tous les éléments de E_1.
- On définit de même m_2 en permutant les rôles de E_1 et de E_2.
- Puis on pose $m = \min(m_1, m_2)$, c'est-à-dire que m est la plus petite des deux valeurs m_1 et m_2 obtenues.

On vérifie que $m_1 + m_2 = n_1 n_2$, ce qui permet un contrôle des résultats.

On peut aussi obtenir m_1 et m_2 de la façon suivante : soit r_1 et r_2 la somme des rangs des valeurs de chacun des deux échantillons. En cas d'ex-aequo les rangs sont déterminés comme indiqué ci-dessus. On a :

$$m_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - r_1 \quad \text{et} \quad m_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - r_2$$

Décision

Soit M la variable aléatoire qui prend la valeur m à l'issue de l'expérience aléatoire.

- Les tables 7 et 8 donnent, en fonction de n_1, n_2 et α la valeur m_α, telle que, sous (H_0), $P(M \leq m_\alpha) = \alpha$, dans les cas $\alpha = 0{,}05$ et $\alpha = 0{,}01$. On rejette donc l'hypothèse nulle si $m \leq m_\alpha$.
- Si n_1 et n_2 sont hors des tables, alors, si (H_0) est vraie, M suit approximativement la loi normale $\mathcal{N}(\mu, \sigma)$ avec :

$$\mu = \frac{n_1 n_2}{2} \quad \text{et} \quad \sigma = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}.$$

On calcule donc la valeur de la variable normale réduite : $z = \dfrac{m - \mu}{\sigma}$ et on conclut, comme d'habitude, avec la table 2, c'est-à-dire qu'on rejette (H_0) si $|z| > z_\alpha$.

16.3 TEST DE WILCOXON

Problématique

On dispose de deux échantillons appariés, c'est-à-dire que chaque valeur d'un échantillon est associée à une valeur de l'autre échantillon. Ils sont par conséquent de même taille. L'hypothèse nulle (H_0) est l'égalité des moyennes des deux populations soit $\mu_1 = \mu_2$.

Mise en place du test

- On calcule les différences entre les valeurs appariées. On supprime les différences nulles et on note N le nombre de différences non nulles.

Ici, on supprime les différences nulles, ce qu'il ne faut pas faire dans le test analogue du chapitre 12.

- On classe ces différences par ordre croissant des valeurs absolues.

On ne tient pas compte du signe dans le classement ; mais le signe n'est pas perdu, il va servir après.

- On affecte à chaque différence son rang dans ce classement. S'il y a des ex-aequo, on attribue à chacun un rang égal à la moyenne des rangs qu'ils occupent.
- On calcule: w_+ somme des rangs des différences positives et w_- somme des rangs des différences négatives.

On vérifie que $w_+ + w_- = \frac{N(N+1)}{2}$, ce qui permet un contrôle des résultats.

On note : $w = \min(w_+, w_-)$ la plus petite des deux valeurs w_+ et w_-.

Décision

Soit W variable aléatoire qui prend la valeur w à l'issue de l'expérience aléatoire.

- Si $N \leq 25$, la table 9 donne, en fonction de N, la valeur w_α, telle que, sous (H_0), $P(W \leq w_\alpha) = \alpha$ dans les cas $\alpha = 0{,}05$ et $\alpha = 0{,}01$. On rejette l'hypothèse nulle si $w \leq w_\alpha$.
- Si $N > 25$, lorsque (H_0) est vraie, W suit approximativement la loi normale $\mathcal{N}(\mu, \sigma)$ avec:

$$\mu = \frac{N(N+1)}{4} \quad \text{et} \quad \sigma = \sqrt{\frac{N(N+1)(2N+1)}{24}}.$$

On calcule donc la valeur de la variable normale réduite : $z = \frac{w - \mu}{\sigma}$ et on conclut, comme d'habitude, avec la table 2, c'est-à-dire qu'on rejette (H_0) si $|z| > z_\alpha$.

16.4 TEST DE KRUSKAL ET WALLIS

Problématique

On dispose de k échantillons, indépendants et non exhaustifs, $E_1,\ldots,E_k$, de tailles respectives $n_1,\ldots,n_k$. On veut comparer globalement les k moyennes expérimentales, c'est-à-dire tester l'hypothèse nulle (H_0) : $\mu_1 = \cdots = \mu_k$.

Mise en place du test

On classe par ordre croissant l'ensemble des valeurs de ces k échantillons. Puis on détermine le rang de chaque valeur, de la même manière que dans les tests précédents s'il y a des ex-aequo.

Pour chaque échantillon E_i, on note r_i la somme des rangs des valeurs de cet échantillon.

On calcule alors la quantité : $h = \dfrac{12}{n(n+1)}\left(\sum_{i=1}^{k} \dfrac{r_i^2}{n_i}\right) - 3(n+1)$ où $n = \sum_{i=1}^{k} n_i$ désigne l'effectif total.

Décision

Soit H la variable aléatoire qui prend la valeur h à l'issue de l'expérience aléatoire.

- Si les n_i sont assez grands (borne classique: $n_i > 5$ pour tout i), alors, si (H_0) est vraie, H suit à peu près la loi du χ^2 à $k-1$ degrés de liberté.
  Dans la table 4 on lit la valeur χ^2_α telle que $P(H \geq \chi^2_\alpha) = \alpha$ et on rejette (H_0) si $h \geq \chi^2_\alpha$.
- Si les n_i ne sont pas assez grands, on dispose de tables qui donnent la valeur h_α, telle que $P(H \geq h_\alpha) = \alpha$.
  On rejette donc (H_0) si on obtient $h \geq h_\alpha$.
  La table 12 donne h_α, pour $\alpha = 0{,}05$ et $\alpha = 0{,}01$, dans le cas de trois échantillons de tailles inférieures ou égales à 5.

16.5 COEFFICIENT DE CORRÉLATION DE RANG DE SPEARMAN

Problématique

Sur une population, on considère deux variables aléatoires X et Y, et on veut tester (H_0) : absence de corrélation entre X et Y.

Pour ceci, on dispose généralement de n couples (x_i, y_i) de valeurs de X et de Y déterminées simultanément. Si on ne sait rien sur les lois de X et de Y, on ne peut pas utiliser les résultats du chapitre 15.

Dans ce cas, on range par ordre croissant, séparément, les valeurs $x_1,\cdots,x_n$ et $y_1,\cdots,y_n$.

On remplace alors chaque valeur x_i par son rang x'_i, et chaque valeur y_i par son rang y'_i. En cas d'ex-aequo, on procède comme dans les tests précédents.

Lorsque les observations consistent en un simple classement des individus en fonction des deux critères X et Y, on a dès le départ les couples (x'_i, y'_i).

Coefficient de corrélation de rang de Spearman

C'est le nombre r_S égal au coefficient de corrélation calculé à partir des couples de rangs (x'_i, y'_i). La méthode la plus rapide pour le calcul est d'utiliser une calculatrice avec ces couples de rangs.

Hypothèse nulle

Si ρ_S désigne le coefficient de corrélation de rang de Spearman au niveau des populations, l'hypothèse nulle (H_0) que l'on va tester s'écrit : $\rho_S = 0$.

Décision

• **Dans le cas** $n \leq 13$

Pour $n \in \{4,...,13\}$ et les risques $\alpha = 0{,}10$, $\alpha = 0{,}05$, $\alpha = 0{,}02$ et $\alpha = 0{,}01$, la table 11 donne la borne r_α telle que $P(|R_S| > r_\alpha) = \alpha$.

- Si $|r_S| > r_\alpha$, on rejette (H_0) avec un risque α de se tromper.
- Si $|r_S| \leq r_\alpha$, on ne rejette pas (H_0).

• **Dans le cas** $n > 13$

Dans ce cas, si (H_0) est vraie, $T = \dfrac{R_S\sqrt{n-2}}{\sqrt{1-R_S^2}}$ suit à peu près la loi de Student à $n-2$ degrés de liberté.

On peut en déduire une règle de décision. mais il est plus rapide d'utiliser la table 10 qui fournit directement une borne r_α déduite de la loi de Student qui précède.

- Si $|r_S| > r_\alpha$, on rejette (H_0) avec un risque α de se tromper.
- Si $|r_S| \leq r_\alpha$, on ne rejette pas (H_0).

Autre formule pour le coefficient de corrélation de rang de Spearman dans le cas où il n'y a pas d'ex-aequo

Dans ce cas, si $d_i = x'_i - y'_i$ désignent les différences des rangs, on a aussi : $r_S = 1 - \dfrac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$. Démontrons :

- On a toujours $\sum_{i=1}^{n} x'_i = \dfrac{n(n+1)}{2} = \sum_{i=1}^{n} y'_i$ puisqu'on se ramène à la somme des n premiers nombres entiers.

➤ *S'il n'y a pas d'ex-aequo*, alors $\sum_{i=1}^{n} x_i'^2 = \sum_{i=1}^{n} y_i'^2$ est la somme des carrés des n premiers nombres entiers, soit :

$$\sum_{k=1}^{n} k^2 = \frac{n(n+1)(2n+1)}{6}.$$

➤ D'autre part, on a :

$$\sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} (x_i' - y_i')^2 = \sum_{i=1}^{n} x_i'^2 - 2\sum_{i=1}^{n} x_i' y_i' + \sum_{i=1}^{n} y_i'^2$$

$$\text{d'où : } \sum_{i=1}^{n} x_i' y_i' = \frac{1}{2}\left[\frac{n(n+1)(2n+1)}{3} - \sum_{i=1}^{n} d_i^2\right]$$

➤ r_S peut donc s'écrire :

$$r_S = \frac{\text{Cov}(X',Y')}{\sqrt{V(X')V(Y')}} = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i' y_i' - \left(\frac{n+1}{2}\right)^2}{\frac{(n+1)(2n+1)}{6} - \left(\frac{n+1}{2}\right)^2}$$

$$= \frac{-\frac{1}{2n}\sum_{i=1}^{n} d_i^2 + \frac{(n+1)(2n+1)}{6} - \frac{(n+1)^2}{4}}{\frac{n+1}{12}[4n+2-3n-3]}$$

$$= \frac{-\frac{1}{2n}\sum_{i=1}^{n} d_i^2 + \frac{1}{12}(n^2-1)}{\frac{1}{12}(n^2-1)} = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}.$$

MOTS-CLÉS

- Test de Mann et Whitney
- Test de Wilcoxon
- Test de Kruskal et Wallis
- Test de Spearman

EXERCICES

16-1 Pour tester simultanément l'homogénéité des moyennes de quatre échantillons, nous pouvons utiliser :

❑ **a)** Un test de Mann et Whitney.
❑ **b)** Un test du χ^2.
❑ **c)** Une analyse de variance (ANOVA).
❑ **d)** Un test paramétrique de Kruskal et Wallis.
❑ **e)** Aucune des propositions précédentes n'est exacte.

16-2 Deux groupes *A* et *B* de 10 étudiants, formés à des méthodes pédagogiques différentes, ont subi le même examen. À l'issue de cet examen, le classement des étudiants était le suivant :

A	1		3	4	5		7	8	8 ex			12			15	17			
B		2				6				10	11		13	14	15 ex		18	19	20

On désire savoir si les deux méthodes pédagogiques conduisent à des résultats statistiquement différents.

a) Montrez qu'il faut utiliser un test non-paramétrique.

b) Appliquez le test de Mann et Whitney pour résoudre le problème posé.

16-3 Comparez les moyennes des échantillons :
E_1 : 3,0 ; 9,8 ; 2,0 ; 5,2 ; 3,6 ; 5,9 ; 8,5 ; 9,4
E_2 : 9,3 ; 12,5 ; 11,3 ; 7,6 ; 3,2 ; 8,6 ; 7,2 ; 14,2 ; 9,6 ; 3,8
On ne sait rien de la loi suivie par la variable aléatoire étudiée au niveau des populations.

16-4 Dans le cadre d'une expertise clinique de validation d'un médicament *M*, on administre à 10 malades, successivement à chacun et dans un ordre tiré au sort, le médicament *M* et une même dose d'un médicament de référence *R*.
Les effets de ces deux substances sur chacun des 10 malades sont :

M	5	4	2	3	4	3	8	5	4	5
R	6	3	3	1	1	3	4	2	5	7

Peut-on dire que les médicaments M et R ont des effets significativement différents au risque 5 % ?

16-5 Un chimiste a mis au point une méthode de dosage du principe actif contenu dans des comprimés pharmaceutiques. Il décide de la comparer à une méthode de référence. Pour cela, il dose 12 comprimés par les deux méthodes, avec les résultats suivants :

Comprimé n°	Quantité de principe actif (en mg)	
	Méthode de référence	Méthode testée
1	9,2	9,5
2	10,0	9,0
3	9,0	8,8
4	9,4	9,5
5	10,1	9,1
6	9,5	10,0
7	10,0	10,1
8	10,3	9,3
9	10,2	9,0
10	10,2	9,7
11	9,8	9,1
12	10,1	9,3

Y a-t-il une différence significative entre les résultats des deux méthodes ?

16-6 On a dosé la teneur en calcium de trois types d'eaux issues d'origines géographiques différentes. Chaque type d'eau a fait l'objet de quatre prélèvements. Les résultats des dosages (en mg de calcium par litre d'eau) sont :

Eau 1 : 18 ; 20 ; 22 ; 25

Eau 2 : 15 ; 16 ; 17 ; 21

Eau 3 : 15 ; 20 ; 21 ; 25

L'origine géographique a-t-elle une influence significative sur la teneur en calcium des eaux considérées ?

16-7 On a étudié l'activité d'une enzyme, l'acétylcholinestérase, chez des animaux soumis à l'action d'un insecticide organophosphoré. L'activité enzymatique est exprimée en micromoles de substrat hydrolysé par minute et par mg de protéines. Les résultats obtenus en fonction du temps d'exposition au pesticide sont donnés ci-après (les échantillons sont indépendants).

L'insecticide entraîne-t-il une diminution significative de l'activité de l'enzyme ? (on comparera globalement les quatre échantillons)

Animaux témoins	Animaux traités		
	1 jour	2 jours	3 jours
15,0	15,0	2,0	0,5
8,5	9,0	2,2	3,0
10,0	8,0	4,0	2,3
10,0	2,0	2,4	0,6
7,6	5,0	1,1	0,9
5,0	3,0	0,7	0,5

16-8 On a étudié l'inhibition de la cholinestérase par une série de composés organophosphorés. Pour chaque composé on a déterminé :

– le pouvoir inhibiteur, exprimé par la constante de formation K du complexe enzyme-composé ;

– la lipophilie, exprimée par le coefficient de partage P du composé entre l'eau et l'octanol.

Les valeurs obtenues pour 9 composés sont les suivantes :

log *K*	2,27	2,44	2,46	2,56	3,08	3,23	3,27	3,32	3,71
log *P*	0,089	–0,67	0,021	0,66	0,82	1,88	2,53	2,39	1,67

Y a-t-il une corrélation significative entre l'action inhibitrice et la lipophilie?

16-9 On considère les classements, en mathématiques et en français, d'un groupe de 12 élèves :

Élève	A	B	C	D	E	F	G	H	I	J	K	L
Maths	6	4	12	1	10	5	8	2	11	7	3	9
Français	3	9	11	2	12	4	10	5	8	1	6	7

Y a-t-il une corrélation significative entre les résultats obtenus dans les deux matières ?

SOLUTIONS

16-1 ☒ **a)** ❑ **b)** ☒ **c)** ❑ **d)** ❑ **e)**

Si les échantillons sont de grandes tailles, ou si les populations sont gaussiennes et de même variance on utilise une ANOVA.

Sinon, on utilise un test non paramétrique : le test de Mann et Whitney.

16-2 **a)** Comme on ne connaît que le rang des étudiants, on doit utiliser un test non-paramétrique. Si l'on avait connu leurs notes, on aurait pu hésiter entre un test paramétrique (test de Student) et un test non-paramétrique, bien que le faible effectif des échantillons soit plutôt en faveur de ce dernier.

b) Comme l'hypothèse nulle (H_0) est l'égalité des moyennes des deux classements, on utilise le test de Mann et Whitney. On a :

$m_1 = 10 + 9 + 9 + 9 + 8 + 8 + 8 + 6 + 3{,}5 + 3 = 73{,}5$

$m_2 = 9 + 6 + 3 + 3 + 2 + 2 + 1{,}5 = 26{,}5$

$m = \min(m_1, m_2) = 26{,}5$

On vérifie que $m_1 + m_2 = n_1 n_2$ car $n_1 = 10$ et $n_2 = 10$.

D'après les tables 7 et 8, la valeur m_α telle que $P(M \leq m_\alpha) = \alpha$ est : $m_{0,05} = 23$ pour $\alpha = 0{,}05$ et $m_{0,01} = 16$ pour $\alpha = 0{,}01$.

Dans ces deux cas, on a $m > m_\alpha$. On ne peut donc pas rejeter (H_0).

La différence des deux échantillons n'est pas significative au risque 5 %.

16-3 On teste l'hypothèse nulle (H_0) : $\mu_1 = \mu_2$, soit : les deux échantillons sont extraits de populations ayant la même moyenne.

On va utiliser le test de Mann et Whitney car les échantillons sont de petites tailles, et il n'y a aucune raison de supposer les populations gaussiennes.

Classons l'ensemble des valeurs de $E_1 \cup E_2$ par ordre croissant, en repérant en gras les valeurs de E_2.

2,0 ; 3,0 ; **3,2** ; 3,6 ; **3,8** ; 5,2 ; 5,9 ; **7,2** ; **7,6** ; 8,5 ; **8,6** ; **9,3** ; 9,4 ; **9,6** ; 9,8 ; **11,3** ; **12,5** ; **14,2**

On a :

$m_1 = 10 + 10 + 9 + 8 + 8 + 6 + 4 + 3 = 58$

$m_2 = 6 + 5 + 3 + 3 + 2 + 2 + 1 = 22$

$m = \min(m_1, m_2) = 22$

On vérifie que $m_1 + m_2 = n_1 n_2$ car $n_1 = 8$ et $n_2 = 10$.

D'après les tables 7 et 8, la valeur m_α telle que $P(M \leq m_\alpha) = \alpha$ est : $m_{0,05} = 17$ pour $\alpha = 0{,}05$ et $m_{0,01} = 11$ pour $\alpha = 0{,}01$. Dans ces deux cas, on a $m > m_\alpha$. On ne peut donc pas rejeter (H_0). La différence des moyennes n'est pas significative au risque 5 %.

16-4 On teste l'hypothèse nulle (H_0) : il n'y a pas de différence significative entre les effets moyens des deux médicaments.

Les deux échantillons sont appariés puisqu'il s'agit des mêmes malades.

Comme les échantillons sont de petite taille et qu'il n'y a aucune raison de supposer les populations gaussiennes, on ne peut pas utiliser un test paramétrique. On utilise le test de Wilcoxon.

Calculons les différences entre les résultats des deux traitements, dans l'ordre $M - R$ par exemple. On obtient :

$$\{-1\ ;\ 1\ ;\ -1\ ;\ 2\ ;\ ;\ 3\ ;\ 0\ ;\ 4\ ;\ 3\ ;\ -1\ ;\ -2\}$$

Rangeons ces différences par ordre croissant de valeurs absolues, en éliminant la valeur nulle :

Valeurs	–1	1	–1	–1	2	–2	3	3	4
Rangs provisoires	1	2	3	4	5	6	7	8	9
Rangs moyens	2,5	2,5	2,5	2,5	5,5	5,5	7,5	7,5	9

On en déduit :

$w_+ = 2{,}5 + 5{,}5 + 7{,}5 + 7{,}5 + 9 = 32$

$w_- = 2{,}5 + 2{,}5 + 2{,}5 + 5{,}5 = 13$

On obtient donc : $w = \min(w_+, w_-) = 13$.

On vérifie que $w_+ + w_- = \dfrac{N(N+1)}{2} = 45$ car $N = 9$.

D'après la table 9, les valeurs w_α telles que $P(W \leq w_\alpha) = \alpha$ sont : $w_{0,05} = 6$ pour $\alpha = 0{,}05$ et $w_{0,01} = 2$ pour $= 0{,}01$.

On a : $w > w_{0,05}$. On ne peut donc pas rejeter (H_0) au risque 5 %.

Il n'y a pas de différence significative entre l'action du médicament testé et celle du médicament de référence.

16-5 On teste l'hypothèse nulle (H_0) : il n'y a pas de différence significative entre les moyennes des résultats des deux dosages. Les deux échantillons sont appariés puisqu'il s'agit des mêmes comprimés. On va utiliser le test de Wilcoxon car les échantillons sont de petite taille et il n'y a aucune raison de supposer les populations gaussiennes.

Les différences entre les résultats des deux méthodes sont :

$$\{-0{,}3\ ;\ 1\ ;\ 0{,}2\ ;\ -0{,}1\ ;\ 1\ ;\ -0{,}5\ ;\ -0{,}1\ ;\ 1\ ;\ 1{,}2\ ;\ 0{,}5\ ;\ 0{,}7\ ;\ 0{,}8\}$$

Ces différences sont toutes non nulles. Rangeons-les par ordre croissant des valeurs absolues, et déterminons leurs rangs.

Différences	–0,1	–0,1	0,2	–0,3	–0,5	0,5
Rangs provisoires	1	2	3	4	5	6
Rangs moyens	1,5	1,5	3	4	5,5	5,5

Différences	0,7	0,8	1	1	1	1,2
Rangs provisoires	7	8	9	10	11	12
Rangs moyens	7	8	10	10	10	12

On en déduit :

$w_+ = 3 + 5{,}5 + 7 + 8 + 10 + 10 + 10 + 12 = 65{,}5$

$w_- = 1{,}5 + 1{,}5 + 4 + 5{,}5 = 12{,}5$

$w = \min(w_+, w_-) = 12{,}5$

On vérifie que $w_+ + w_- = \dfrac{N(N+1)}{2} = 78$ car $N = 12$.

D'après la table 9, les valeurs w_α telles que $P(W \leq w_\alpha) = \alpha$ sont : $w_{0,05} = 14$ pour $\alpha = 0{,}05$ et $w_{0,01} = 7$ pour $\alpha = 0{,}01$.

Comme $w < w_{0,05}$, (H_0) est rejetée au risque 5 %.

Mais comme $w > w_{0,01}$, (H_0) n'est pas rejetée au risque 1 %.

Il y a donc entre les deux méthodes une différence significative au risque 5 % (mais pas au risque 1 %).

16-6 On va tester l'hypothèse nulle (H_0) : $\mu_1 = \mu_2 = \mu_3$, c'est-à-dire : il n'y a pas de différence significative entre les teneurs moyennes en calcium des trois types d'eau.

Comme il n'y a aucune raison de supposer les populations gaussiennes et de même variance et comme les échantillons sont de petites tailles, on ne peut pas utiliser l'analyse de la variance. On va utiliser le test de Kruskal et Wallis.

Rangeons par ordre croissant l'ensemble des valeurs des trois échantillons, puis déterminons leurs rangs :

Valeurs			**Rangs**	**Rangs moyens**		
eau 1	**eau 2**	**eau 3**		**eau 1**	**eau 2**	**eau 3**
	15		1		1,5	
		15	2			1,5
	16		3		3	
	17		4		4	
18			5	5		
20			6	6,5		
		20	7			6,5
	21		8		8,5	

Valeurs			Rangs	Rangs moyens		
eau 1	eau 2	eau 3		eau 1	eau 2	eau 3
		21	9			8,5
22			10	10		
25			11	11,5		
		25	12			11,5
Totaux				33	17	28

On en déduit :

$$h = \frac{12}{n(n+1)}\left(\sum_{i=1}^{3}\frac{r_i^2}{n_i}\right) - 3(n+1)$$
$$= \frac{12}{12\times 13}\left(\frac{33^2}{4}+\frac{17^2}{4}+\frac{28^2}{4}\right) - 3\times 13 \approx 2{,}58.$$

Pour 3 groupes de 4 valeurs, la table 12 donne la valeur h_α telle que $P(H \geq h_\alpha) = \alpha$, soit : $h_{0,05} = 5{,}70$ pour $\alpha = 0{,}05$ et $h_{0,01} = 7{,}60$ pour $\alpha = 0{,}01$.

On a : $h < h_{0,05}$. On ne peut donc pas rejeter (H_0) au risque 5 % ; la différence des teneurs en calcium des trois eaux considérées n'est pas significative au risque 5 %.

16-7

Valeurs				*Rangs*	Rangs moyens			
témoins	1 jour	2 jours	3 jours		témoins	1 jour	2 jours	3 jours
			0,5	1				1,5
			0,5	2				1,5
			0,6	3				3
		0,7		4			4	
			0,9	5				5
		1,1		6			6	
		2,0		7			7,5	
	2,0			8		7,5		
		2,2		9			9	
			2,3	10				10
		2,4		11			11	
			3,0	12				12,5

Valeurs				*Rangs*	Rangs moyens			
témoins	**1 jour**	**2 jours**	**3 jours**		**témoins**	**1 jour**	**2 jours**	**3 jours**
	3,0			13		12,5		
		4,0		14			14	
	5,0			15		15,5		
5,0				16	15,5			
7,6				17	17			
	8,0			18		18		
8,5				19	19			
	9,0			20		20		
10,0				21	21,5			
10,0				22	21,5			
	15,0			23		23,5		
15,0				24	23,5			
Totaux					118	97	51,5	33,5

Hypothèse nulle : pas de différence significative entre les activités moyennes des quatre échantillons. En l'absence d'informations concernant la distribution statistique des valeurs de l'activité enzymatique, utilisons le test de Kruskal-Wallis. En appliquant la même méthode que pour l'exercice précédent, on obtient le tableau de la page précédente.
On en déduit :

$$h = \frac{12}{24 \times 25}\left(\frac{118^2}{6} + \frac{97^2}{6} + \frac{51{,}5^2}{6} + \frac{33{,}5^2}{6}\right) - 3 \times 25 \approx 15{,}36.$$

D'autre part, comme $n_1 > 5$, $n_2 > 5$, $n_3 > 5$, $n_4 > 5$, on sait que la variable aléatoire H suit à peu près la loi du χ^2 à $k - 1 = 3$ degrés de liberté.
D'après la table 4, la valeur h_α telle que $P(H \geq h_\alpha) = \alpha$ est donc :
$h_{0,05} = 7{,}81$ pour $\alpha = 0{,}05$;
$h_{0,01} = 11{,}34$ pour $\alpha = 0{,}01$;
$h_{0,001} = 16{,}27$ pour $\alpha = 0{,}001$.
Comme $h > h_{0,01}$, on rejette (H_0) au risque 1 %, et donc aussi au risque 5 %.
La différence des activités enzymatiques moyennes des quatre échantillons est donc significative au risque 1 % (mais pas au risque 0,1 %).

16-8 On teste (H_0) : absence de corrélation entre log K et log P. En l'absence d'informations concernant la distribution statistique des valeurs, utilisons le test non-paramétrique de Spearman.

Déterminons les rangs des valeurs après les avoir rangées par ordre croissant :

Rang (log *K*)	1	2	3	4	5	6	7	8	9
Rang (log *P*)	3	1	2	4	5	7	9	8	6

Si votre calculatrice fournit directement le coefficient de corrélation, à partir de ces couples de rangs vous obtenez $r_S \approx 0{,}83$.
Sinon, comme il n'y a pas d'ex-aequo, vous obtenez r_S par l'expression :

$$r_S = 1 - \frac{6}{9(81-1)}\left[(-2)^2 + 1^2 + 1^2 + 0^2 + 0^2 + (-1)^2 + (-2)^2 + 0^2 + 3^2\right]$$

D'après la table 11, la valeur r_α telle que $P(|R_S| > r_\alpha) = \alpha$ est :
$r_{0,05} = 0{,}68$ pour $\alpha = 0{,}05$; $r_{0,01} = 0{,}82$ pour $\alpha = 0{,}01$.
Comme $r_S \notin]-r_\alpha, r_\alpha[$, on rejette donc (H_0) au risque 5 %, et même au risque 1 %. La corrélation est donc significative au risque 1 %.

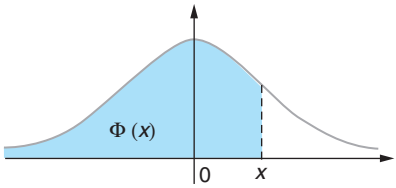
16-9 On teste (H_0) : absence de corrélation entre les résultats des deux matières.
Puisqu'on ne dispose que des rangs des élèves, nous sommes conduits à utiliser le test de Spearman.
Avec une calculatrice, vous obtenez directement à partir des couples de rangs : $r_S \approx 0{,}61$.
Pour $n = 12$, la table 11 donne $r_{0,05} = 0{,}59$.
Comme $r_S \notin]-r_{0,05}, r_{0,05}[$, (H_0) est rejetée au risque 5 %.
La corrélation entre les classements est donc significative au risque 5 %.

Tables

TABLE 1

Fonction de répartition de la loi normale réduite

Si Z suit la loi normale réduite, pour $x \geq 0$, la table donne la valeur $\phi(x) = P(Z \leq x)$.
La valeur x s'obtient par addition des nombres inscrits en marge.
Pour $x < 0$, on a :
$\phi(x) = 1 - \phi(-x)$.



x	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
2,0	0,977 2	0,977 8	0,978 3	0,978 8	0,979 3	0,979 8	0,980 3	0,980 8	0,981 2	0,981 7
2,1	0,982 1	0,982 6	0,983 0	0,983 4	0,983 8	0,984 2	0,984 6	0,985 0	0,985 4	0,985 7
2,2	0,986 1	0,986 4	0,986 8	0,987 1	0,987 5	0,987 8	0,988 1	0,988 4	0,988 7	0,989 0
2,3	0,989 3	0,989 6	0,989 8	0,990 1	0,990 4	0,990 6	0,990 9	0,991 1	0,991 3	0,991 6
2,4	0,991 8	0,992 0	0,992 2	0,992 5	0,992 7	0,992 9	0,993 1	0,993 2	0,993 4	0,993 6
2,5	0,993 8	0,994 0	0,994 1	0,994 3	0,994 5	0,994 6	0,994 8	0,994 9	0,995 1	0,995 2
2,6	0,995 3	0,995 5	0,995 6	0,995 7	0,995 9	0,996 0	0,996 1	0,996 2	0,996 3	0,996 4
2,7	0,996 5	0,996 6	0,996 7	0,996 8	0,996 9	0,997 0	0,997 1	0,997 2	0,997 3	0,997 4
2,8	0,997 4	0,997 5	0,997 6	0,997 7	0,997 7	0,997 8	0,997 9	0,997 9	0,998 0	0,998 1
2,9	0,998 1	0,998 2	0,998 2	0,998 3	0,998 4	0,998 4	0,998 5	0,998 5	0,998 6	0,998 6

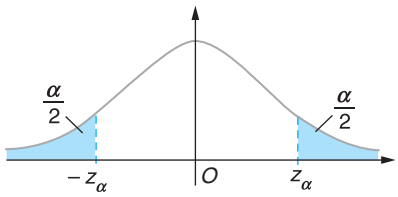
TABLE 2

Loi normale réduite (table de l'écart réduit)

Si Z est une variable aléatoire qui suit la loi normale réduite, la table donne pour α choisi, la valeur z_α telle que :

$P(|Z| \geq z_\alpha) = \alpha$

La valeur α s'obtient par addition des nombres inscrits en marge.



α	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	∞	2,576	2,326	2,170	2,054	1,960	1,881	1,812	1,751	1,695
0,1	1,645	1,598	1,555	1,514	1,476	1,440	1,405	1,372	1,341	1,311
0,2	1,282	1,254	1,227	1,200	1,175	1,150	1,126	1,103	1,080	1,058
0,3	1,036	1,015	0,994	0,974	0,954	0,935	0,915	0,896	0,878	0,860
0,4	0,842	0,824	0,806	0,789	0,772	0,755	0,739	0,722	0,706	0,690
0,5	0,674	0,659	0,643	0,628	0,613	0,598	0,583	0,568	0,553	0,539
0,6	0,524	0,510	0,496	0,482	0,468	0,454	0,440	0,426	0,412	0,399
0,7	0,385	0,372	0,358	0,345	0,332	0,319	0,305	0,292	0,279	0,266
0,8	0,253	0,240	0,228	0,215	0,202	0,189	0,176	0,164	0,151	0,138
0,9	0,126	0,113	0,100	0,088	0,075	0,063	0,050	0,038	0,025	0,013

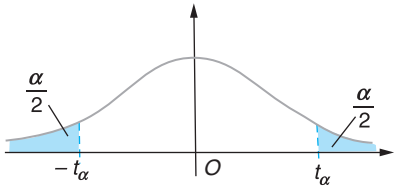
Table pour les petites valeurs de α

α	10^{-3}	10^{-4}	10^{-5}	10^{-6}	10^{-7}
z_α	3,291	3,891	4,417	4,892	5,527

TABLE 3

Lois de Student

Si T est une variable aléatoire qui suit la loi de Student à ν degrés de liberté, la table donne pour α choisi, le nombre t_α tel que $P(|T| \geq t_\alpha = \alpha)$ ·



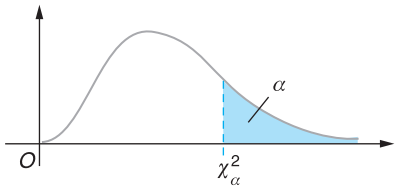
ν \ α	0,90	0,50	0,30	0,20	0,10	0,05	0,02	0,01	0,001
1	0,158	1,000	1,963	3,078	6,314	12,706	31,821	63,657	636,619
2	0,142	0,816	1,386	1,886	2,920	4,303	6,965	9,925	31,598
3	0,137	0,765	1,250	1,638	2,353	3,182	4,541	5,841	12,924
4	0,134	0,741	1,190	1,533	2,132	2,776	3,747	4,604	8,610
5	0,132	0,727	1,156	1,476	2,015	2,571	3,365	4,032	6,869
6	0,131	0,718	1,134	1,440	1,943	2,447	3,143	3,707	5,959
7	0,130	0,711	1,119	1,415	1,895	2,365	2,998	3,499	5,408
8	0,130	0,706	1,108	1,397	1,860	2,306	2,896	3,355	5,041
9	0,129	0,703	1,100	1,383	1,833	2,262	2,821	3,250	4,781
10	0,129	0,700	1,093	1,372	1,812	2,228	2,764	3,169	4,587
11	0,129	0,697	1,088	1,363	1,796	2,201	2,718	3,106	4,437
12	0,128	0,695	1,083	1,356	1,782	2,179	2,681	3,055	4,318
13	0,128	0,694	1,079	1,350	1,771	2,160	2,650	3,012	4,221
14	0,128	0,692	1,076	1,345	1,761	2,145	2,624	2,977	4,140
15	0,128	0,691	1,074	1,341	1,753	2,131	2,602	2,947	4,073
16	0,128	0,690	1,071	1,337	1,746	2,120	2,583	2,921	4,015
17	0,128	0,689	1,069	1,333	1,740	2,110	2,567	2,898	3,965
18	0,127	0,688	1,067	1,330	1,734	2,101	2,552	2,878	3,922
19	0,127	0,688	1,066	1,328	1,729	2,093	2,539	2,861	3,883
20	0,127	0,687	1,064	1,325	1,725	2,086	2,528	2,845	3,850
21	0,127	0,686	1,063	1,323	1,721	2,080	2,518	2,831	3,819
22	0,127	0,686	1,061	1,321	1,717	2,074	2,508	2,819	3,792
23	0,127	0,685	1,060	1,319	1,714	2,069	2,500	2,807	3,767
24	0,127	0,685	1,059	1,318	1,711	2,064	2,492	2,797	3,745
25	0,127	0,684	1,058	1,316	1,708	2,060	2,485	2,787	3,725
26	0,127	0,684	1,058	1,315	1,706	2,056	2,479	2,779	3,707
27	0,127	0,684	1,057	1,314	1,703	2,052	2,473	2,771	3,690
28	0,127	0,683	1,056	1,313	1,701	2,048	2,467	2,763	3,674
29	0,127	0,683	1,055	1,311	1,699	2,045	2,462	2,756	3,659
30	0,127	0,683	1,055	1,310	1,697	2,042	2,457	2,750	3,646
40	0,126	0,681	1,050	1,303	1,684	2,021	2,423	2,704	3,551
80	0,126	0,679	1,046	1,296	1,671	2,000	2,390	2,660	3,460
120	0,126	0,677	1,041	1,289	1,658	1,980	2,358	2,617	3,373
∞	0,126	0,674	1,036	1,282	1,645	1,960	2,326	2,576	3,291

Lorsque le degré de liberté est infini, il s'agit du nombre z_α correspondant à la loi normale centrée réduite (*cf.* table 2).

TABLE 4

Lois de Pearson ou lois du χ^2

Si Y est une variable aléatoire qui suit la loi du χ^2 à ν degrés de liberté, la table donne pour α choisi, le nombre χ^2_α tel que $P(Y \geq \chi^2_\alpha) = \alpha$.



ν \ α	0,99	0,975	0,95	0,90	0,10	0,05	0,025	0,01	0,001
1	0,000 2	0,001	0,004	0,016	2,71	3,84	5,02	6,63	10,83
2	0,02	0,05	0,10	0,21	4,61	5,99	7,38	9,21	13,82
3	0,12	0,22	0,35	0,58	6,25	7,81	9,35	11,34	16,27
4	0,30	0,48	0,71	1,06	7,78	9,49	11,14	13,28	18,47
5	0,55	0,83	1,15	1,61	9,24	11,07	12,83	15,09	20,52
6	0,87	1,24	1,64	2,20	10,64	12,59	14,45	16,81	22,46
7	1,24	1,69	2,17	2,83	12,02	14,07	16,01	18,47	24,32
8	1,65	2,18	2,73	3,49	13,36	15,51	17,53	20,09	26,13
9	2,09	2,70	3,33	4,17	14,68	16,92	19,02	21,67	27,88
10	2,56	3,25	3,94	4,87	15,99	18,31	20,48	23,21	29,59
11	3,05	3,82	4,57	5,58	17,27	19,67	21,92	24,72	31,26
12	3,57	4,40	5,23	6,30	18,55	21,03	23,34	26,22	32,91
13	4,11	5,01	5,89	7,04	19,81	22,36	24,74	27,69	34,53
14	4,66	5,63	6,57	7,79	21,06	23,68	26,12	29,14	36,12
15	5,23	6,26	7,26	8,55	22,31	25,00	27,49	30,58	37,70
16	5,81	6,91	7,96	9,31	23,54	26,30	28,84	32,00	39,25
17	6,41	7,56	8,67	10,08	24,77	27,59	30,19	33,41	40,79
18	7,01	8,23	9,39	10,86	25,99	28,87	31,53	34,80	42,31
19	7,63	8,91	10,12	11,65	27,20	30,14	32,85	36,19	43,82
20	8,26	9,59	10,85	12,44	28,41	31,41	34,17	37,57	45,32
21	8,90	10,28	11,59	13,24	29,61	32,67	35,48	38,93	46,80
22	9,54	10,98	12,34	14,04	30,81	33,92	36,78	40,29	48,27
23	10,20	11,69	13,09	14,85	32,01	35,17	38,08	41,64	49,73
24	10,86	12,40	13,85	15,66	33,20	36,41	39,37	42,98	51,18
25	11,52	13,12	14,61	16,47	34,38	37,65	40,65	44,31	52,62
26	12,20	13,84	15,38	17,29	35,56	38,88	41,92	45,64	54,05
27	12,88	14,57	16,15	18,11	36,74	40,11	43,19	46,96	55,48
28	13,57	15,31	16,93	18,94	37,92	41,34	44,46	48,28	56,89
29	14,26	16,05	17,71	19,77	39,09	42,56	45,72	49,59	58,30
30	14,95	16,79	18,49	20,60	40,26	43,77	46,98	50,89	59,70

Lorsque le degré de liberté ν est tel que $\nu > 30$, la variable aléatoire :

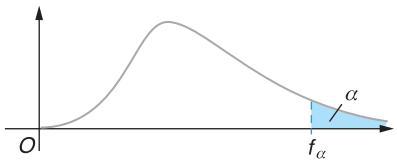
$Z = \sqrt{Y} - \sqrt{2\nu - 1}$

suit à peu près la loi normale réduite.

TABLE 5

Lois de Snedecor ($\alpha = 0{,}025$)

Si F est une variable aléatoire qui suit la loi de Snedecor à (ν_1, ν_2) degrés de liberté, la table donne le nombre f_α tel que $P(F \geq f_\alpha) = \alpha = 0{,}025$.

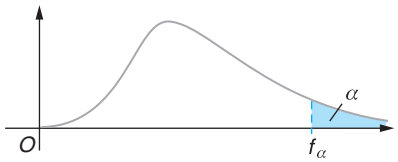


ν_2 \ ν_1	1	2	3	4	5	6	8	10	15	20	30	∞
1	648	800	864	900	922	937	957	969	985	993	1 001	1 018
2	38,5	39,0	39,2	39,2	39,3	39,3	39,4	39,4	39,4	39,4	39,5	39,5
3	17,4	16,0	15,4	15,1	14,9	14,7	14,5	14,4	14,3	14,2	14,1	13,9
4	12,2	10,6	9,98	9,60	9,36	9,20	8,98	8,84	8,66	8,56	8,46	8,26
5	10,0	8,43	7,76	7,39	7,15	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
16	6,12	4,69	4,08	3,73	3,50	3,34	3,12	2,99	2,79	2,68	2,57	2,32
17	6,04	4,62	4,01	3,66	3,44	3,28	3,06	2,92	2,72	2,62	2,50	2,25
18	5,98	4,56	3,95	3,61	3,38	3,22	3,01	2,87	2,67	2,56	2,44	2,19
19	5,92	4,51	3,90	3,56	3,33	3,17	2,96	2,82	2,62	2,51	2,39	2,13
20	5,87	4,46	3,86	3,51	3,29	3,13	2,91	2,77	2,57	2,46	2,35	2,09
22	5,79	4,38	3,78	3,44	3,22	3,05	2,84	2,70	2,50	2,39	2,27	2,00
24	5,72	4,32	3,72	3,38	3,15	2,99	2,78	2,64	2,44	2,33	2,21	1,94
26	5,66	4,27	3,67	3,33	3,10	2,94	2,73	2,59	2,39	2,28	2,16	1,88
28	5,61	4,22	3,63	3,29	3,06	2,90	2,69	2,55	2,34	2,23	2,11	1,83
30	5,57	4,18	3,59	3,25	3,03	2,87	2,65	2,51	2,31	2,20	2,07	1,79
40	5,42	4,05	3,46	3,13	2,90	2,74	2,53	2,39	2,18	2,07	1,94	1,64
50	5,34	3,98	3,39	3,06	2,83	2,67	2,46	2,32	2,11	1,99	1,87	1,55
60	5,29	3,93	3,34	3,01	2,79	2,63	2,41	2,27	2,06	1,94	1,82	1,48
80	5,22	3,86	3,28	2,95	2,73	2,57	2,36	2,21	2,00	1,88	1,75	1,40
100	5,18	3,83	3,25	2,92	2,70	2,54	2,32	2,18	1,97	1,85	1,71	1,35
∞	5,02	3,69	3,12	2,79	2,57	2,41	2,19	2,05	1,83	1,71	1,57	1,00

TABLE 6

Lois de Snedecor ($\alpha = 0{,}05$)

Si F est une variable aléatoire qui suit la loi de Snedecor à (ν_1, ν_2) degrés de liberté, la table donne le nombre f_α tel que $P(F \geq f_\alpha) = \alpha = 0{,}05$.



ν_2 \ ν_1	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
16	4,49	3,63	3,24	3,01	2,85	2,74	2,59	2,49	2,35	2,28	2,19	2,01
17	4,45	3,59	3,20	2,96	2,81	2,70	2,55	2,45	2,31	2,23	2,15	1,96
18	4,41	3,55	3,16	2,93	2,77	2,66	2,51	2,41	2,27	2,19	2,11	1,92
19	4,38	3,52	3,13	2,90	2,74	2,63	2,48	2,38	2,23	2,16	2,07	1,88
20	4,35	3,49	3,10	2,87	2,71	2,60	2,45	2,35	2,20	2,12	2,04	1,84
22	4,30	3,44	3,05	2,82	2,66	2,55	2,40	2,30	2,15	2,07	1,98	1,78
24	4,26	3,40	3,01	2,78	2,62	2,51	2,36	2,25	2,11	2,03	1,94	1,73
26	4,23	3,37	2,98	2,74	2,59	2,47	2,32	2,22	2,07	1,99	1,90	1,69
28	4,20	3,34	2,95	2,71	2,56	2,45	2,29	2,19	2,04	1,96	1,87	1,65
30	4,17	3,32	2,92	2,69	2,53	2,42	2,27	2,16	2,01	1,93	1,84	1,62
40	4,08	3,23	2,84	2,61	2,45	2,34	2,18	2,08	1,92	1,84	1,74	1,51
50	4,03	3,18	2,79	2,56	2,40	2,29	2,13	2,03	1,87	1,78	1,69	1,44
60	4,00	3,15	2,76	2,53	2,37	2,25	2,10	1,99	1,84	1,75	1,65	1,39
80	3,96	3,11	2,72	2,49	2,33	2,21	2,06	1,95	1,79	1,70	1,60	1,32
100	3,94	3,09	2,70	2,46	2,31	2,19	2,03	1,93	1,77	1,68	1,57	1,28
∞	3,84	3,00	2,60	2,37	2,21	2,10	1,94	1,83	1,67	1,57	1,46	1,00

TABLE 7

Test de Mann et Whitney ($\alpha = 0{,}05$)

La table donne la valeur m_α tel que $P(M \leq m_\alpha) = \alpha = 0{,}05$ pour deux échantillons d'effectifs n_1 et n_2 avec $n_1 \leq n_2$.

n_1 \ n_2	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
2	–	–	–	–	0	0	0	0	1	1	1	1	1	2	2	2	2
3	–	0	1	1	2	2	3	3	4	4	5	5	6	6	7	7	8
4	0	1	2	3	4	4	5	6	7	8	9	10	11	11	12	13	14
5		2	3	5	6	7	8	9	11	12	13	14	15	17	18	19	20
6			5	6	8	10	11	13	14	16	17	19	21	22	24	25	27
7				8	10	12	14	16	18	20	22	24	26	28	30	32	34
8					13	15	17	19	22	24	26	29	31	34	36	38	41
9						17	20	23	26	28	31	34	37	39	42	45	48
10							23	26	29	33	36	39	42	45	48	52	55
11								30	33	37	40	44	47	51	55	58	62
12									37	41	45	49	53	57	61	65	69
13										45	50	54	59	63	67	72	76
14											55	59	64	69	74	78	83
15												64	70	75	80	85	90
16													75	81	86	92	98
17														87	93	99	105
18															99	106	112
19																113	119
20																	127

TABLE 8

Test de Mann et Whitney ($\alpha = 0{,}01$)

La table donne la valeur m_α tel que $P(M \leq m_\alpha) = \alpha = 0{,}01$ pour deux échantillons d'effectifs n_1 et n_2 avec $n_1 \leq n_2$.

n_1 \ n_2	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
2	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	0	0
3	–	–	–	–	–	0	0	0	1	1	1	2	2	2	2	3	3
4	–	–	0	0	1	1	2	2	3	3	4	5	5	6	6	7	8
5		0	1	1	2	3	4	5	6	7	7	8	9	10	11	12	13
6			2	3	4	5	6	7	9	10	11	12	13	15	16	17	18
7				4	6	7	9	10	12	13	15	16	18	19	21	22	24
8					7	9	11	13	15	17	18	20	22	24	26	28	30
9						11	13	16	18	20	22	24	27	29	31	33	36
10							16	18	21	24	26	29	31	34	37	39	42
11								21	24	27	30	33	36	39	42	45	48
12									27	31	34	37	41	44	47	51	54
13										34	38	42	45	49	53	57	60
14											42	46	50	54	58	63	67
15												51	55	60	64	68	73
16													60	65	70	74	79
17														70	75	81	86
18															81	87	92
19																93	99
20																	105

TABLE 9

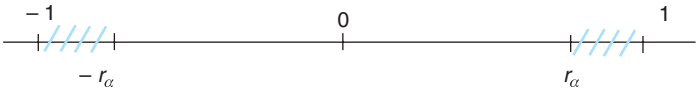
Test de Wilcoxon

La table donne la valeur w_α tel que $P(W \leq w_\alpha) = \alpha$, dans les cas $\alpha = 0{,}05$ et $\alpha = 0{,}01$.

α \ N	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
0,05	2	4	6	8	11	14	17	21	25	30	35	40	46	52	59	66	73	81	89
0,01	–	0	2	3	5	7	10	13	16	20	23	28	32	38	43	49	55	61	68

TABLE 10

Table du coefficient de corrélation linéaire



d.d.l. \ α	0,10	0,05	0,01
1	0,9877	0,9969	0,9999
2	0,9000	0,9500	0,9900
3	0,8054	0,8783	0,9587
4	0,7293	0,8114	0,9172
5	0,6694	0,7545	0,8745
6	0,6215	0,7067	0,8343
7	0,5822	0,6664	0,7977
8	0,5494	0,6319	0,7646
9	0,5214	0,6021	0,7348
10	0,4973	0,5760	0,7079
11	0,4762	0,5529	0,6835
12	0,4575	0,5324	0,6614
13	0,4409	0,5139	0,6411
14	0,4259	0,4973	0,6226
15	0,4124	0,4821	0,6055
16	0,4000	0,4683	0,5897
17	0,3887	0,4555	0,5751
18	0,3783	0,4438	0,5614
19	0,3687	0,4329	0,5487
20	0,3598	0,4227	0,5368
25	0,3233	0,3809	0,4869
30	0,2960	0,3494	0,4487
35	0,2746	0,3246	0,4182
40	0,2573	0,3044	0,3932
45	0,2428	0,2875	0,3721
50	0,2306	0,2732	0,3541
60	0,2108	0,2500	0,3248
70	0,1954	0,2319	0,3017
80	0,1829	0,2172	0,2830
90	0,1726	0,2050	0,2673
100	0,1638	0,1946	0,2540

TABLE 11

Coefficient de corrélation de rang de Spearman

La table donne la valeur r_α tel que $P(|R_s| > r_\alpha) = \alpha$.

α \ n	4	5	6	7	8	9	10	11	12	13
0,10	0,99	0,87	0,77	0,69	0,64	0,59	0,56	0,53	0,51	0,49
0,05	—	0,95	0,85	0,78	0,73	0,68	0,64	0,61	0,59	0,56
0,02	—	0,99	0,93	0,87	0,82	0,77	0,73	0,70	0,67	0,64
0,01	—	—	0,97	0,91	0,86	0,82	0,79	0,75	0,72	0,70

TABLE 12

Test de Kruskal et Wallis

La table donne la valeur h_α tel que $P(H \geq h_\alpha) = \alpha$.

Taille des échantillons	$\alpha = 0{,}05$	$\alpha = 0{,}01$
3 2 2	4,71	
3 3 1	5,10	
3 3 2	5,22	6,26
3 3 3	5,60	6,50
4 2 1	4,94	
4 2 2	5,15	6,30
4 3 1	5,21	
4 3 2	5,42	6,35
4 3 3	5,73	6,75
4 4 1	4,93	6,67
4 4 2	5,45	6,90
4 4 3	5,60	7,14
4 4 4	5,70	7,60
5 2 1	5,00	
5 2 2	5,10	6,40
5 3 1	4,91	6,42
5 3 2	5,25	6,82
5 3 3	5,66	7,03
5 4 1	4,92	6,90
5 4 2	5,27	7,12
5 4 3	5,63	7,44
5 4 4	5,62	7,75
5 5 1	5,00	7,08
5 5 2	5,27	7,30
5 5 3	5,64	7,55
5 5 4	5,64	7,80
5 5 5	5,72	7,98

Glossaire

Échantillon représentatif : la représentativité d'un échantillon dépend de l'observation effectuée. Il s'agit de reproduire les répartitions, connues dans la population, qui ont de l'influence sur l'étude. Pour un sondage relatif à des intentions de vote, on va reproduire, en pourcentages, les tranches d'âge, le sexe, les zones d'habitat, les catégories socio-professionnelles, les revenus … mais il est inutile de tenir compte des cheveux ou de la taille des individus, sauf si un parti des chauves se créait !

Il faut aussi veiller à prélever les individus de façon aléatoire et non par commodité ou par volontariat. Par exemple, en 1989, Europe 1 a demandé à ses auditeurs de téléphoner leur opinion sur le permis à points. Le nombre d'appels a été très important, mais le résultat était très biaisé car ce sont surtout les opposants qui téléphonent !

Échantillons indépendants ou appariés : lorsque les individus sont considérés comme interchangeables par rapport à l'étude en cours, l'expérimentateur prend des échantillons séparés, indépendants.

Lorsque la variabilité entre les individus est forte et doit être gommée, on considère les mêmes individus dans deux situations différentes. Les mesures obtenues constituent alors des échantillons appariés, car les valeurs sont associées par paires.

À ne pas confondre avec échantillons avariés (agro-alimentaire) ou appareillés (orthopédie), termes rencontrés sur des prises de notes d'étudiants !

Loi de Poisson : c'est une loi qui modélise une situation aléatoire où les possibilités sont des entiers naturels. C'est souvent le nombre d'apparitions d'un événement rare.

Par exemple, pour étudier la répartition de la rouille sur un bâteau de pêche, on peut diviser la coque en petites surfaces et compter le nombre de taches dans une surface donnée. Ce nombre est aléatoire et un bon modèle est une loi de Poisson ! Sourire autorisé.

Si vous ajoutez, ce qui n'a rien à voir, la loi de Fisher, décidemment la statistique a des rencontres amusantes !

Moyenne : vous savez depuis longtemps calculer votre moyenne scolaire : vous additionnez vos notes et vous divisez par le nombre de notes. Il s'agit de la moyenne arithmétique.

Vous avez remarqué que le résultat n'est pas toujours une note observée. Figurez-vous que des journalistes ne le savent pas : ils ricanent sur des femmes qui mettent au monde en moyenne 1,87 enfants en se demandant comment accoucher d'une fraction d'enfant !

Mais si je vous posais la question suivante :

un cycliste monte un col à 20 km.h^{-1} *et le redescend à* 60 km.h^{-1} *; quelle est sa vitesse moyenne ?,*

certains répondraient 40, qui est la moyenne arithmétique.

Alors que, si d désigne la longueur du col, les durées sont de $\frac{d}{20}$ puis $\frac{d}{60}$ pour une distance $2d$, ce qui donne comme moyenne : $\frac{2d}{\frac{d}{20}+\frac{d}{60}} = \frac{2}{\frac{1}{20}+\frac{1}{60}} = 30$ qui est la moyenne harmonique des deux vitesses.

Donc le mot moyenne ne doit pas vous faire perdre vos moyens !

Population normale : ce terme curieux laisse penser qu'il y a des populations anormales. Il signifie seulement qu'on s'intéresse à une variable aléatoire X définie sur cette population, et que X suit une loi dite de Gauss, ou de Laplace-Gauss, ou normale. Le terme synonyme est population gaussienne.

La loi normale est souvent utilisée car elle permet de modéliser une mesure qui est le cumul d'un grand nombre de petits phénomènes aléatoires indépendants.

Pour accepter l'hypothèse qu'une population est normale, à partir d'observations nombreuses, on peut :

- tracer l'histogramme des mesures et contrôler visuellement qu'il ressemble à une courbe en cloche ;
- faire une vérification graphique en utilisant un papier, dit gausso-arithmétique, quadrillé de sorte que les points que l'on reporte soient alignés pour une population normale ;
- faire un test de conformité du χ^2 ;
- utiliser un ordinateur pour un test plus évolué, comme le test de Kolmogorov-Smirnov.

Risque : en statistique inférentielle, on est amené à prendre des décisions à partir d'informations incomplètes. Si vous rajoutez la variabilité du vivant, toute affirmation est donc liée à un risque de se tromper.

Dans le cas d'un intervalle de confiance, on souhaite affirmer qu'un paramètre appartient à un intervalle I avec un risque α. Les deux objectifs : précision (intervalle réduit), sécurité (α réduit) sont contradictoires. Il faut donc choisir un compromis. Par exemple, lors d'une soirée électorale, les instituts de sondage donnent dès 20 heures un intervalle où devrait se situer le résultat définitif d'un candidat, en oubliant que cette affirmation se fait avec un certain risque.

Cet intervalle de confiance s'appelle une fourchette, bien que les téléspectateurs ne soient pas toujours à table !

Risque de première espèce, risque de deuxième espèce : quand on teste une hypothèse simple, hypothèse nulle (H_0), contre une hypothèse simple, hypothèse alternative (H_1), chaque affirmation a lieu avec un certain risque.

La décision de rejeter (H_0) se prend avec un risque α de première espèce.

La décision de rejeter (H_1) se prend avec un risque β de deuxième espèce.

Le concepteur d'un nouveau test s'intéresse à sa puissance $1 - \beta$.

L'utilisateur s'intéresse seulement à α et ses conclusions sont :

- je rejette (H_0) au risque α ;
- je ne rejette pas (H_0), ou j'accepte (H_0), avec l'unique expérience disponible ; mais gardez un vocabulaire prudent !

Tests de rangs : dans certains domaines, comme la comparaison de goûts en agro-alimentaire, ou dans les sciences humaines, on dispose de classements. Dans d'autres cas, on a des mesures peu nombreuses issues d'une population inconnue. On les remplace alors par des rangs.

Les tests qui traitent ces situations sont les tests non-paramétriques. Contrairement aux autres tests, la valeur de la variable de décision ne résulte pas d'une formule, mais d'un processus de comptage. Mais la prise de décision est la même : la valeur de la variable de décision appartient à une zone de probabilité α (dont les bornes se lisent dans des tables adaptées) et l'hypothèse nulle (H_0) est rejetée ; sinon (H_0) est acceptée.

Variance : aussi bien dans la pratique (caractère statistique quantitatif) que dans la modélisation (variable aléatoire), la variance est une mesure de la dispersion des valeurs, observées ou possibles, par rapport à la moyenne. La présence des carrés dans la définition empêche que les écarts positifs et négatifs puissent se compenser.

Si les mesures sont en cm, la moyenne est en cm et la variance en cm^2. Il est donc intéressant d'introduire sa racine carrée, appelée écart type, qui est aussi en cm.

Index

A

amplitude 4
ANOVA 182
arrangement 37

B

Bartlett (test de) 187
Bayes 48

C

caractère
  continu 3
  discret 3
  qualitatif 3
  quantitatif 3
  statistique 3
classe statistique 4
coefficient de corrélation 62
coefficient de variation 7
combinaisons 38
converge 91
correction de continuité 95
couple de variables aléatoires 60
covariance 21, 62

D

densité 4
densité de probabilité 92
distribution à deux dimensions 17
distribution de probabilité 60
distributions conditionnelles 18
distributions marginales 18
droite de régression 21

E

écart interquartile 7
écart type 6, 62, 93
échantillon appariés 108
échantillonnage 107
échantillons indépendants 108
effectif cumulé 4
effectif 4
espace probabilisable 35
espace probabilisé 35
espérance mathématique 62, 92
estimateur convergent 108
estimateur sans biais 108
étendue 7
événement 33
événements indépendants 48
exhaustif (échantillon) 108
expérience aléatoire 33
expériences indépendantes 49

F

Fisher (test de) 150
fonction de répartition 60
formule des probabilités totales 48
fréquence 4
fréquence cumulée 4

H

histogramme 5
hypothèse alternative 123
hypothèse nulle 123

I

indépendance statistique 18
individu 3

intégrale convergente 77
intégrale divergente 91
intervalle de confiance 111

K

Koenigs 62
Kruskal et Wallis (test de) 228

L

loi binomiale 63
loi continue uniforme 93
loi de Poisson 80
loi des grands nombres 49
loi discrète uniforme 63
loi exponentielle 93
loi géométrique 80
loi normale 94
lois de Snedecor 167

M

Mann et Whitney (test de) 226
médiane 6
méthode des moindres carrés 20
méthode du maximum de vraisemblance 114
mode 6
moments 8
moyenne 5

N

non-exhaustif (échantillon) 108
non-paramétriques (tests) 225

P

permutation 37
populations 3
probabilité conditionnelle 45
probabilité uniforme 36

R

risque de deuxième espèce 124
risque de première espèce 124
risque relatif 50

S

série absolument convergente 78
série convergente 77
série divergente 78
série exponentielle 78
série géométrique 78
séries de Riemann 78
Spearman (test de) 225
système complet d'événements 36

T

test de conformité du χ^2 129
test d'homogénéité du χ^2 131
tribu 34

V

variable aléatoire 59
Variable centrée réduite 63
variables aléatoires indépendantes 60
variance 6, 62, 92
variance factorielle 182
variance résiduelle 182, 198

W

Wilcoxon (test de) 227

85110 - (I) - OSB 80 - LUM - CMU
Dépôt légal : décembre 2022
Achevé d'imprimer par Dupli-Print

www.dupli-print.fr

Imprimé en France