

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

3^e édition

DUNOD

Directeur d'ouvrage
Daniel FREDON

© Dunod, 2007, 2014, 2018, 2022 pour la nouvelle présentation

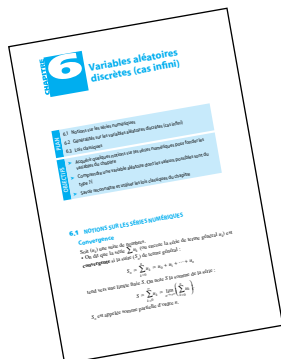
11, rue Paul Bert 92240 Malakoff
www.dunod.com

ISBN : 9782100854677

Le Code de la propriété intellectuelle n'autorisant, aux termes de l'article L. 122-5, 2° et 3° a), d'une part, que les « copies ou reproductions strictement réservées à l'usage privé du copiste et non destinées à une utilisation collective » et, d'autre part, que les analyses et les courtes citations dans un but d'exemple et d'illustration, « toute représentation ou reproduction intégrale ou partielle faite sans le consentement de l'auteur ou de ses ayants droit ou ayants cause est illicite » (art. L. 122-4).

Cette représentation ou reproduction, par quelque procédé que ce soit, constituerait donc une contrefaçon sanctionnée par les articles L. 335-2 et suivants du Code de la propriété intellectuelle.

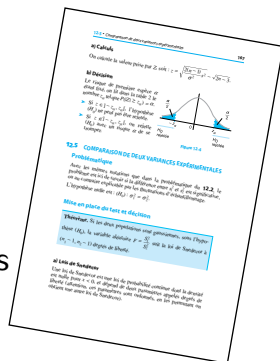
La page d'entrée de chapitre



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



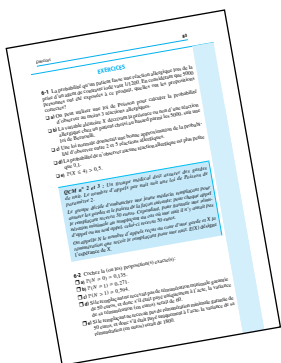
Un peu de méthode



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

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 : $f_i = \frac{n_i}{n}$.

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

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

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 : $g^n = x_1^{n_1} \times \dots \times x_p^{n_p}$.

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 $x_1, \dots, x_k$ affectées de coefficients $n_1, \dots, n_k$ par leur moyenne partielle affectée de la somme des coefficients $n_1 + \dots + n_k$.

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

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

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 :

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

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

La variance peut aussi se calculer par :

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



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 :

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

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 $\frac{\sigma}{\bar{x}}$.

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.

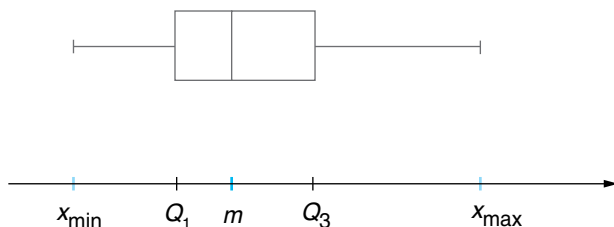


Figure 1.1