

Benoît Clément

Analyse de données en sciences expérimentales

**Probabilités et Statistiques,
avec exemples en sciences
physique**

DUNOD

Tout le catalogue sur
www.dunod.com



Illustration de couverture : © kentoh - Fotolia.fr

Le pictogramme qui figure ci-contre mérite une explication. Son objet est d'alerter le lecteur sur la menace que représente pour l'avenir de l'écrit, particulièrement dans le domaine de l'édition technique et universitaire, le développement massif du photocopillage.

Le Code de la propriété intellectuelle du 1^{er} juillet 1992 interdit en effet expressément la photocopie à usage collectif sans autorisation des ayants droit. Or, cette pratique s'est généralisée dans les établissements

d'enseignement supérieur, provoquant une baisse brutale des achats de livres et de revues, au point que la possibilité même pour

les auteurs de créer des œuvres nouvelles et de les faire éditer correctement est aujourd'hui menacée. Nous rappelons donc que toute reproduction, partielle ou totale, de la présente publication est interdite sans autorisation de l'auteur, de son éditeur ou du Centre français d'exploitation du droit de copie (CFC, 20, rue des Grands-Augustins, 75006 Paris).



© Dunod, Paris, 2012
ISBN 978-2-10-058039-2

Le Code de la propriété intellectuelle n'autorisant, aux termes de l'article L. 122-5, 2^o et 3^o 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.

TABLE DES MATIÈRES

Avant-propos	VII
---------------------	------------

PARTIE I

ÉCHANTILLONS

Chapitre 1. Distributions et fréquences	3
1.1 Échantillon et population	3
1.2 Mesures de position	4
1.3 Mesures de dispersion	5
1.4 Tableau de fréquences	7
Chapitre 2. Échantillons multidimensionnels	13
2.1 Mesures de forme	13
2.2 Régression linéaire	15
2.3 Matrice de covariance et décorrélation de variables	17
Chapitre 3. Incertitudes de mesures	21
3.1 Le concept d'incertitude	21
3.2 Lien avec l'écart type	21
3.3 Propagation des incertitudes	23
Exercices	27

PARTIE II

PROBABILITÉS

Chapitre 4. Un peu de théorie des probabilités	33
4.1 Définitions de la probabilité	33
4.2 Fonction de probabilité	34
4.3 Combinatoire	37
Chapitre 5. Variables aléatoires	41
5.1 Fonction de répartition et densité de probabilité	41
5.2 Changement de variable	43
5.3 Mesures de forme	44
5.4 Densités usuelles	48

Analyse de données en sciences expérimentales

Chapitre 6. Variables aléatoires multidimensionnelles	55
6.1 Densités marginales et conditionnelles	55
6.2 Covariance	56
6.3 La loi multinormale	57
6.4 Somme de variables aléatoires indépendantes	57
Exercices	61

PARTIE III

STATISTIQUES

Chapitre 7. Estimation paramétrique	67
7.1 Estimateur	67
7.2 Estimateurs des moments	69
7.3 Estimateur du maximum de vraisemblance	72
Chapitre 8. Estimation et incertitude	77
8.1 Estimation par intervalle	77
8.2 Retour sur l'incertitude	81
8.3 Estimation non paramétrique	83
Chapitre 9. Tests statistiques	87
9.1 Le test du χ^2 de Pearson	87
9.2 Le test de Kolmogorov-Smirnov	89
9.3 Exemple	90
Chapitre 10. Ajustements de données	93
10.1 Introduction	93
10.2 Régression dans le cas général	94
10.3 Méthode des moindres carrés ou méthode du χ^2	95
Chapitre 11. Introduction aux méthodes Monte-Carlo	107
11.1 Échantillonnage	107
11.2 Intégration Monte-Carlo	110
11.3 Propagation d'incertitudes	113
11.4 Simulation de systèmes physiques	114
Exercices	118

Annexes

Corrigés	127
Formulaire	161
Tableaux	173
Bibliographie	179
Index	181

AVANT-PROPOS

Il existe traditionnellement un clivage dans la communauté scientifique entre théoriciens et expérimentateurs. D'un côté les théoriciens tentent de décrire les phénomènes au moyen d'équations. Les modèles ainsi développés permettent de décrire l'évolution de systèmes physiques. Ils sont prédictifs s'ils permettent le calcul d'un grand nombre d'observables à partir de quelques paramètres libres. De l'autre les expérimentateurs conçoivent et réalisent des expériences destinées à mesurer telle ou telle propriété physique, avec la meilleure précision possible.

L'analyse des données se situe à l'interface et devient de plus en plus un domaine à part entière de la physique. Elle regroupe toutes les techniques visant à extraire une information utile d'un ensemble de données. Ceci consiste à présenter ces données de manière compréhensible et à les réduire à une information utile, la plus compacte possible avec une perte minimum d'information. Pour l'expérimentateur son rôle premier sera la détermination la plus correcte possible des incertitudes qui entachent toute mesure. Pour le théoricien, elle fournira des outils statistiques pour résoudre numériquement des problèmes complexes, par exemple au moyen de méthodes Monte-Carlo. Mais surtout, l'analyse de données permet de faire le lien entre la mesure expérimentale et les modèles théoriques. Les problèmes auxquels on sera confrontés (estimation de paramètres d'un modèle et de l'incertitude associée à ce résultat, détermination de niveaux de confiance, contrainte d'un modèle théorique, test d'hypothèse) reviennent tous à répondre à la question : « Quelles informations les données expérimentales apportent-elles sur le modèle théorique ? » La réponse à cette question relève de l'inférence statistique, qui elle-même repose sur la théorie des probabilités. Ce sont les bases de ces deux concepts que nous allons développer dans cet ouvrage. Pour les étudiants, leur première application sera la rédaction de comptes rendus de travaux pratiques, c'est pourquoi deux notions centrales serviront de fil conducteur à l'exposé : d'une part, l'estimation et la propagation des incertitudes, d'autre part, l'ajustement de courbes sur des points expérimentaux.

L'ouvrage est découpé en trois grandes parties. La première introduira simplement des outils permettant de caractériser et de représenter un ensemble fini de valeurs tels la moyenne ou l'écart type. La seconde partie formalisera ces outils dans le cadre de la théorie des probabilités. Enfin la dernière partie concernant l'inférence statistique illustrera comment, à partir de la théorie des probabilités, on peut tenter d'extraire des informations d'un échantillon de données. Cette dernière partie se terminera par une brève introduction aux méthodes Monte-Carlo. Chaque partie est accompagnée d'exercices d'application dont les solutions sont données en annexes. Ces annexes

Analyse de données en sciences expérimentales

sont complétées par un formulaire regroupant de manière synthétique les principales formules de l'ouvrage ainsi que par quelques tables de fonctions statistiques utiles.

L'essentiel du contenu de cet ouvrage repose sur des cours donnés en troisième année de licence de physique et en première année de master à l'Université Joseph Fourier. Il n'existerait pas sans la contribution de nombreux collègues. Je tiens particulièrement à remercier Arnaud Lucotte qui a patiemment relu ce document, Frederic Mayet et Eric Liatard qui m'ont, les premiers, proposé d'enseigner l'analyse de données, ainsi que Laurent Derome, Julien Billard et Thomas Delemontex pour les discussions enrichissantes qui m'ont permis d'étendre ma compréhension de ce vaste sujet qu'est l'analyse statistique.

DISTRIBUTIONS ET FRÉQUENCES

1

Nous allons commencer notre étude par la caractérisation d'un ensemble de données brutes (par exemple des mesures) par un petit nombre de grandeurs : il s'agit alors de réduire un grand volume de données à un nombre réduit de valeurs tout en minimisant la perte d'information.

1.1 ÉCHANTILLON ET POPULATION

Un **échantillon** est un ensemble de n réalisations $\{x_i\}_{i=1\dots n}$ d'une même grandeur, par exemple :

- plusieurs mesures de l'énergie de l'électron issu de la désintégration β d'un noyau,
- la date de naissance de la moitié des étudiants d'un amphi,
- le résultat d'une série de lancers de dé,
- ...

Généralement l'échantillon est une sous-partie d'une **population** plus vaste (l'ensemble des énergies que peut prendre l'électron, l'ensemble des dates de naissance, l'ensemble des résultats possibles d'un lancer de dé...). Un échantillon a toujours une taille finie, alors que la population peut être infinie (ce n'est pas le cas pour les résultats d'un lancer de dé). De plus les valeurs de l'échantillon peuvent être soit discrètes (dates, lancer de dé) soit continues (énergies). On se limitera dans ce cours à l'étude d'échantillons à valeurs entières ou réelles. On parlera par la suite de variables aléatoires discrètes (ou dénombrables) ou continues pour décrire une population.

L'analyse statistique des données vise à estimer les propriétés d'une population (potentiellement infinie) à partir de celles d'un échantillon (de taille finie). On va s'intéresser dans ce chapitre à la distribution d'un échantillon, c'est-à-dire étudier et caractériser la répartition de ses valeurs :

- répartition aléatoire,
- regroupement autour d'une ou plusieurs valeurs particulières : **mesure de localisation ou de position**,
- dispersion des valeurs autour de ces valeurs caractéristiques : **mesure de dispersion**,
- autre...

Les données brutes sont le plus souvent inexploitable telles quelles. Voici à titre d'exemple les valeurs du rapport du nombre de neutrons (N) au nombre de protons (Z) pour les noyaux stables de $Z > 55$:

1,321 ; 1,357 ; 1,392 ; 1,410 ; 1,428 ; 1,446 ; 1,464 ; 1,421 ; 1,438 ; 1,344 ; 1,379 ;
1,413 ; 1,448 ; 1,389 ; 1,366 ; 1,383 ; 1,400 ; 1,416 ; 1,433 ; 1,466 ; 1,500 ; 1,322 ;
1,370 ; 1,387 ; 1,403 ; 1,419 ; 1,451 ; 1,483 ; 1,396 ; 1,428 ; 1,375 ; 1,406 ; 1,421 ;
1,437 ; 1,453 ; 1,468 ; 1,500 ; 1,446 ; 1,363 ; 1,393 ; 1,424 ; 1,439 ; 1,454 ; 1,469 ;
1,484 ; 1,462 ; 1,382 ; 1,411 ; 1,441 ; 1,455 ; 1,470 ; 1,500 ; 1,449 ; 1,400 ; 1,428 ;
1,442 ; 1,457 ; 1,471 ; 1,485 ; 1,514 ; 1,464 ; 1,478 ; 1,416 ; 1,444 ; 1,458 ; 1,472 ;
1,486 ; 1,500 ; 1,465 ; 1,479 ; 1,432 ; 1,459 ; 1,472 ; 1,486 ; 1,513 ; 1,466 ; 1,493 ;
1,421 ; 1,447 ; 1,460 ; 1,473 ; 1,486 ; 1,500 ; 1,526 ; 1,480 ; 1,506 ; 1,435 ; 1,461 ;
1,487 ; 1,500 ; 1,512 ; 1,538 ; 1,493 ; 1,450 ; 1,475 ; 1,500 ; 1,512 ; 1,525 ; 1,550 ;
1,506 ; 1,530 ; 1,487 ; 1,512 ; 1,524 ; 1,536 ; 1,518 ; 1,577 ; 1,554 ; 1,586 ; 1,586.

Il y a au total 110 valeurs. Nous allons utiliser cet échantillon tout au long de ce chapitre. Il semble difficile de tirer des conclusions des données présentées sous cette forme. On va donc chercher à réduire l'information contenue dans ces données à un nombre plus faible de grandeurs caractéristiques.

1.2 MESURES DE POSITION

On peut définir plusieurs grandeurs qui caractérisent la position de la distribution, c'est-à-dire une valeur unique qui permet de caractériser l'ensemble des valeurs de l'échantillon. La plus connue est la **moyenne**. Dans certaines situations, on utilise aussi la médiane ou le mode.

1.2.1 La moyenne

La moyenne de l'échantillon est donnée par :

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \dots + x_n) = \frac{1}{n} \sum_{i=1}^N x_i = \frac{1}{n} \sum_i x_i. \quad (1.1)$$

En utilisant les valeurs de N/Z du paragraphe précédent on trouve : $\bar{x} = 1,4564$.

Remarque

Il existe de multiples notations pour la valeur moyenne d'une grandeur x : m_x , μ_x , $\bar{x}$, $\langle x \rangle$.