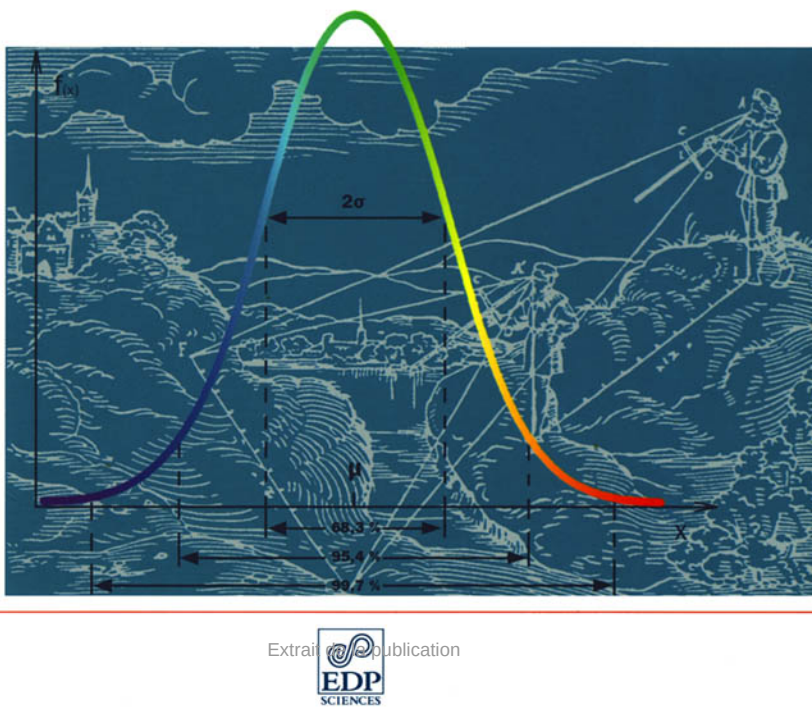
COLLECTION GRENOBLE SCIENCES
DIRIGÉE PAR JEAN BORNAREL

ANALYSE STATISTIQUE DES DONNÉES EXPÉRIMENTALES

Konstantin PROTASSOV



EDP SCIENCES

ANALYSE STATISTIQUE DES DONNÉES EXPÉRIMENTALES

Grenoble Sciences

Grenoble Sciences poursuit un triple objectif :

- réaliser des ouvrages correspondant à un projet clairement défini, sans contrainte de mode ou de programme,
- garantir les qualités scientifique et pédagogique des ouvrages retenus,
- proposer des ouvrages à un prix accessible au public le plus large possible.

Chaque projet est sélectionné au niveau de Grenoble Sciences avec le concours de referees anonymes. Puis les auteurs travaillent pendant une année (en moyenne) avec les membres d'un comité de lecture interactif, dont les noms apparaissent au début de l'ouvrage. Celui-ci est ensuite publié chez l'éditeur le plus adapté.

(Contact : Tél. : (33)4 76 51 46 95 - E-mail : Grenoble.Sciences@ujf-grenoble.fr)

Deux collections existent chez EDP Sciences :

- la ***Collection Grenoble Sciences***, connue pour son originalité de projets et sa qualité
- ***Grenoble Sciences - Rencontres Scientifiques***, collection présentant des thèmes de recherche d'actualité, traités par des scientifiques de premier plan issus de disciplines différentes.

Directeur scientifique de Grenoble Sciences

Jean BORNAREL, Professeur à l'Université Joseph Fourier, Grenoble 1

Comité de lecture pour "Analyse statistique des données expérimentales"

- J.P. BERTRANDIAS, Professeur à l'Université Joseph Fourier, Grenoble 1
- C. FURGET, Maître de conférences à l'Université Joseph Fourier, Grenoble 1
- B. HOUCHMANDZADEH, Directeur de recherches au CNRS, Grenoble
- M. LESIEUR, Professeur à l'Institut National Polytechnique, Grenoble
- C. MISBAH, Directeur de recherches au CNRS, Grenoble
- J.L. PORTESEIL, Professeur à l'Université Joseph Fourier, Grenoble 1
- P. VILLEMAIN, Maître de conférences à l'Université Joseph Fourier, Grenoble 1

Grenoble Sciences reçoit le soutien
du **Ministère de l'Éducation nationale,** du **Ministère de la Recherche,**
de la **Région Rhône-Alpes,** du **Conseil général de l'Isère**
et de la **Ville de Grenoble.**

ISBN 2-86883-456-6
ISBN 2-86883-590-2

ANALYSE STATISTIQUE DES DONNÉES EXPÉRIMENTALES

Konstantin PROTASSOV

EDP SCIENCES

17, avenue du Hoggar
Parc d'Activité de Courtabœuf, BP 112
91944 Les Ulis Cedex A, France

Ouvrages Grenoble Sciences édités par EDP Sciences

Collection Grenoble Sciences

Chimie. Le minimum vital à savoir *(J. Le Coarer)* - Electrochimie des solides *(C. Déportes et al.)* - Thermodynamique chimique *(M. Oturan & M. Robert)* - Chimie organométallique *(D. Astruc)*

Introduction à la mécanique statistique *(E. Belorizky & W. Gorecki)* - Mécanique statistique. Exercices et problèmes corrigés *(E. Belorizky & W. Gorecki)* - La symétrie en mathématiques, physique et chimie *(J. Sivardière)* - La cavitation. Mécanismes physiques et aspects industriels *(J.P. Franc et al.)* - La turbulence *(M. Lesieur)* - Magnétisme : I Fondements, II Matériaux et applications *(sous la direction d'E. du Trémolet de Lacheisserie)* - Du Soleil à la Terre. Aéronomie et météorologie de l'espace *(J. Lilensten & P.L. Blelly)* - Sous les feux du Soleil. Vers une météorologie de l'espace *(J. Lilensten & J. Bornarel)* - Mécanique. De la formulation lagrangienne au chaos hamiltonien *(C. Gignoux & B. Silvestre-Brac)* - La mécanique quantique. Problèmes résolus, Tomes 1 et 2 *(V.M. Galitsky, B.M. Karnakov & V.I. Kogan)*

Exercices corrigés d'analyse, Tomes 1 et 2 *(D. Alibert)* - Introduction aux variétés différentielles *(J. Lafontaine)* - Analyse numérique et équations différentielles *(J.P. Demailly)* - Mathématiques pour les sciences de la vie, de la nature et de la santé *(F. & J.P. Bertrandias)* - Approximation hilbertienne. Splines, ondelettes, fractales *(M. Attéia & J. Gaches)* - Mathématiques pour l'étudiant scientifique, Tomes 1 et 2 *(Ph.J. Haug)*

Bactéries et environnement. Adaptations physiologiques *(J. Pelmont)* - Enzymes. Catalyseurs du monde vivant *(J. Pelmont)* - La plongée sous-marine à l'air. L'adaptation de l'organisme et ses limites *(Ph. Foster)* - L'ergomotricité. Le corps, le travail et la santé *(M. Gendrier)* - Endocrinologie et communications cellulaires *(S. Idelman & J. Verdetti)*

L'Asie, source de sciences et de techniques *(M. Soutif)* - La biologie, des origines à nos jours *(P. Vignais)* - Naissance de la physique. De la Sicile à la Chine *(M. Soutif)*

Minimum Competence in Scientific English *(J. Upjohn, S. Blattes & V. Jans)* - Listening Comprehension for Scientific English *(J. Upjohn)* - Speaking Skills in Scientific English *(J. Upjohn, M.H. Fries & D. Amadis)*

Grenoble Sciences - Rencontres Scientifiques

Radiopharmaceutiques. Chimie des radiotraceurs et applications biologiques *(sous la direction de M. Comet & M. Vidal)* - Turbulence et déterminisme *(sous la direction de M. Lesieur)* - Méthodes et techniques de la chimie organique *(sous la direction de D. Astruc)*

PRÉFACE

Le but de ce petit ouvrage est de répondre aux questions les plus fréquentes que se pose un expérimentateur et de permettre à un étudiant d'analyser, d'une façon autonome, ses résultats et leurs précisions. C'est cet esprit assez "utilitaire" qui a déterminé le style de présentation.

Dans l'analyse des données expérimentales, il existe plusieurs niveaux qui sont conditionnés par notre désir d'obtenir une information plus ou moins riche, mais aussi par le temps que nous sommes prêts à y consacrer. Fréquemment, nous voulons juste obtenir la valeur d'une grandeur physique sans nous préoccuper de vérifier les hypothèses à la base de notre démarche. Parfois, cependant, les résultats obtenus nous paraissent être en contradiction avec nos estimations préliminaires et ainsi nous sommes obligés d'effectuer un travail plus scrupuleux. Ce livre est écrit pour permettre au lecteur de choisir le niveau d'analyse nécessaire.

La partie "indispensable" du texte correspondant au premier niveau est composée avec une police de caractères normale. Les questions qui correspondent à une analyse plus approfondie et qui nécessitent un appareil mathématique plus complexe sont composées avec une police de caractères spéciale. Cette partie du livre peut être sautée lors d'une première lecture.

A la base de toute analyse des données expérimentales, on trouve une approche statistique qui exige des considérations mathématiques rigoureuses et parfois complexes. Néanmoins, l'expérimentateur n'a pas toujours besoin de connaître les détails et les subtilités mathématiques. De plus, rares sont les situations où les conditions expérimentales correspondent exactement aux conditions d'application de tel ou tel théorème. C'est pourquoi l'accent est mis non pas sur la démonstration des résultats mathématiques mais sur leur signification et leur interprétation physique. Parfois, pour alléger la présentation, la rigueur mathématique est volontairement sacrifiée et remplacée par une argumentation "physiquement évidente".

Le plan du livre est simple. Dans l'introduction, on présente les causes d'erreurs et on définit le langage utilisé. Le premier chapitre rappelle les principaux résultats de statistique essentiels à l'analyse des données. Le deuxième chapitre présente des notions plus complexes de statistique, il est consacré aux fonctions de varables aléatoires. Dans le troisième chapitre qui est la partie la plus importante, on s'efforce de répondre aux questions les plus fréquentes qui se posent dans l'analyse des données expérimentales. Le dernier chapitre est consacré aux méthodes les plus fréquemment utilisées pour l'ajustement de paramètres.

Bien que ce livre soit particulièrement adapté au travail d'étudiants de second cycle, il pourra être également utile aux jeunes chercheurs, aux ingénieurs et à tous ceux qui sont amenés à réaliser des mesures.

J'aimerais remercier mes collègues enseignants et chercheurs qui ont lu le manuscrit et qui m'ont fait des propositions pour améliorer son contenu. Je voudrais exprimer ma profonde gratitude à M. Elie Belorizky qui m'a encouragé à écrire ce livre et avec qui j'ai eu des discussions très fructueuses.

POURQUOI LES INCERTITUDES EXISTENT-ELLES ?

Le but de la majorité des expériences en physique consiste à comprendre un phénomène et à le modéliser correctement. Nous effectuons des mesures et nous avons souvent à nous poser la question : "quelle est la valeur de telle ou telle grandeur ?", parfois sans nous demander préalablement si cette formulation est correcte et si nous serons capables de trouver une réponse.

La nécessité de cette interrogation préalable devient évidente dès qu'on mesure la même grandeur plusieurs fois. L'expérimentateur qui le fait est fréquemment confronté à une situation assez intéressante : s'il utilise des appareils suffisamment précis, il s'aperçoit que des mesures répétées de la même grandeur donnent parfois des résultats qui sont un peu différents de celui de la première mesure. Ce phénomène est général, que les mesures soient simples ou sophistiquées. Même les mesures répétées de la longueur d'une tige métallique peuvent donner des valeurs différentes. La répétition de l'expérience montre que, d'une part les résultats sont toujours un peu différents et d'autre part cette différence n'est en général pas très grande. Dans la plupart des cas, on reste proche d'une certaine valeur moyenne, mais de temps en temps on trouve des valeurs qui sont différentes de celle-ci. Plus les résultats sont éloignés de cette moyenne, plus ils sont rares.

Pourquoi cette dispersion existe-t-elle ? D'où vient cette variation ? Une raison de cet effet est évidente : les conditions de déroulement d'une expérience varient toujours légèrement, ce qui modifie la grandeur mesurable. Par exemple, quand on détermine plusieurs fois la longueur d'une tige métallique, c'est la température ambiante qui peut varier et ainsi faire varier la longueur. Cette variation des conditions extérieures (et la variation correspondante de la valeur physique) peut être plus ou moins importante, mais elle est inévitable et, dans les conditions réelles d'une expérience physique, on ne peut pas s'en affranchir.

Nous sommes "condamnés" à effectuer des mesures de grandeurs qui ne sont presque jamais constantes. C'est pourquoi même la question de savoir quelle est la valeur d'un paramètre peut ne pas être absolument correcte. Il faut poser cette question de manière pertinente et trouver des moyens adéquats pour décrire les grandeurs physiques. Il faut trouver une définition qui puisse exprimer cette particularité physique. Cette définition doit refléter le fait que la valeur physique varie toujours, mais que ses variations se regroupent autour d'une valeur moyenne.

La solution est de caractériser une grandeur physique non pas par une valeur, mais plutôt par la probabilité de trouver dans une expérience telle ou telle valeur. Pour cela on introduit une fonction appelée *distribution de probabilité* de détection d'une valeur physique, ou plus simplement la *distribution* d'une valeur physique, qui montre

quelles sont les valeurs les plus fréquentes ou les plus rares. Il faut souligner une fois encore que, dans cette approche, il ne s'agit pas tellement de la valeur concrète d'une grandeur physique, mais surtout de la *probabilité* de trouver différentes valeurs.

On verra par la suite que cette fonction – la distribution d'une valeur physique – est heureusement suffisamment simple (en tout cas, dans la majorité des expériences). Elle a deux caractéristiques. La première est sa valeur moyenne qui est aussi la valeur la plus probable. La deuxième caractéristique de cette fonction de distribution indique, grosso modo, la région autour de cette moyenne dans laquelle se regroupe la majorité des résultats des mesures. Elle caractérise *la largeur* de cette distribution et est appelée *l'incertitude*. Comme nous pourrons le voir par la suite, cette largeur a une interprétation rigoureuse en terme de probabilités. Pour des raisons de simplicité nous appellerons cette incertitude *"l'incertitude naturelle"* ou *"initiale"* de la grandeur physique elle-même. Ce n'est pas tout à fait vrai, puisque cette erreur ou incertitude est souvent due aux conditions expérimentales. Bien que cette définition ne soit pas parfaitement rigoureuse, elle est très utile pour la compréhension.

Le fait que, dans la plupart des expériences, le résultat puisse être caractérisé par seulement deux valeurs, permet de revenir sur la question avec laquelle nous avons commencé notre discussion : "Peut-on se demander quelle est la valeur d'un paramètre physique ?" Il se trouve que dans le cas où deux paramètres sont nécessaires et suffisants pour caractériser une grandeur physique, on peut réconcilier notre envie de poser cette question et la rigueur de l'interprétation d'un résultat en termes de probabilités. La solution existe : on appellera *valeur physique* la valeur moyenne de la distribution et *incertitude* ou *erreur* de la valeur physique la largeur de la distribution[1]. C'est une convention admise de dire que "la grandeur physique a une valeur donnée avec une incertitude donnée". Cela signifie que l'on présente la valeur moyenne et la largeur d'une distribution et que cette réponse a une interprétation précise en termes de probabilités.

Le but des mesures physiques est la détermination de cette fonction de distribution ou, au moins, de ses deux paramètres majeurs : la moyenne et la largeur. Pour déterminer une distribution on doit répéter plusieurs fois une mesure pour connaître la fréquence d'apparition des valeurs. Pour obtenir l'ensemble des valeurs possibles ainsi que leurs probabilités d'apparition, on devrait en fait effectuer un nombre infini de mesures. C'est très long, trop cher, et personne n'en a besoin.

On se limite donc à un nombre fini de mesures. Bien sûr, cela introduit une erreur

1 Pour des raisons historiques, les deux termes "incertitude" et "erreur" sont utilisés en physique pour décrire la largeur d'une distribution. Depuis quelques années, les organismes scientifiques internationaux essaient d'introduire des normes pour utiliser correctement ces deux termes (de la même façon que l'on a introduit le système international d'unités). Aujourd'hui, on appelle une erreur la différence entre le résultat d'une mesure et la vraie valeur de la grandeur mesurée. Tandis que l'incertitude de mesure est un paramètre, associé au résultat d'une mesure, qui caractérise la dispersion des valeurs qui peuvent raisonnablement être attribuées à la grandeur mesurée. Dans ce livre, nous tâcherons de suivre ces normes, mais parfois nous utiliserons des expressions plus habituelles pour un physicien. Par exemple, une formule très connue dans l'analyse des données expérimenatles porte le nom de "la formule de propagation des erreurs". Nous utiliserons toujours ce nom bien connu bien que, selon les normes actuelles, nous aurions dû l'appeller "la formule de propagation des incertitudes". Le lecteur intéressé trouvera dans la bibliographie toutes les réfèrences sur les normes actuelles.

(incertitude) supplémentaire. Cette incertitude, due à l'impossibilité de mesurer avec une précision absolue la distribution initiale (naturelle), s'appelle *l'erreur statistique* ou *l'erreur accidentelle*. Il est assez facile, du moins en théorie, de diminuer cette erreur : il suffit d'augmenter le nombre de mesures. En principe, on peut la rendre négligeable devant l'incertitude initiale de la grandeur physique. Cependant un autre problème plus délicat apparaît.

Il est lié au fait que, dans chaque expérience physique existe un appareil, plus ou moins compliqué, entre l'expérimentateur et l'objet mesurable. Cet appareil apporte inévitablement des modifications de la distribution initiale : il la déforme. Dans le cas le plus simple, ces changements peuvent être de deux types : l'appareil peut *"décaler"* la valeur moyenne et il peut *élargir la distribution*.

Le décalage de la valeur moyenne est un exemple de ce qu'on appelle les *"erreurs systématiques"*. Ce nom exprime que ces erreurs apparaissent dans chaque mesure. L'appareil donne systématiquement une valeur qui est différente (plus grande ou plus petite) de la valeur "réelle". Mesurer avec un appareil dont le zéro est mal réglé est l'exemple le plus fréquent de ce genre d'erreurs. Malheureusement, il est très difficile de combattre ce type d'erreurs : il est à la fois difficile de les déceler et de les corriger. Pour cela, il n'y a pas de méthodes générales et il faut étudier chaque cas.

Par contre, il est plus facile de maîtriser l'élargissement de la distribution introduit par l'appareil. On verra que cette incertitude ayant la même origine que les incertitudes initiales (naturelles) s'ajoute "simplement" à celles-ci. Dans un grand nombre d'expériences, l'élargissement dû à l'appareil permet de simplifier les mesures : supposons que nous connaissions l'incertitude (la largeur) introduite par un appareil et que celle-ci soit nettement plus grande que l'incertitude initiale. Il est possible de négliger l'incertitude naturelle par rapport à l'incertitude d'appareillage. Il suffit donc de faire une seule mesure et de prendre l'incertitude de l'appareil comme incertitude de la mesure. Evidemment, dans ce genre d'expérience, il faut être sûr que l'incertitude de l'appareil domine l'incertitude naturelle, mais on peut toujours le vérifier en faisant des mesures répétitives. L'appareil peu précis ne permettra pas d'obtenir les variations dues à la largeur initiale.

Il faut remarquer que la séparation entre incertitude d'appareillage et incertitude naturelle reste assez conventionnelle : on peut toujours dire que la variation des conditions d'expérience fait partie de l'incertitude d'appareillage. Dans ce livre, on ne parle pas des mesures en mécanique quantique, où existe une incertitude de la valeur physique à cause de la relation d'incertitude de Heisenberg. En mécanique quantique, l'interférence appareil–objet devient plus compliquée et intéressante. Cependant nos conclusions générales ne sont pas modifiées puisque, en mécanique quantique, la notion de probabilité est non seulement utile et naturelle, mais elle est indispensable.

Nous avons compris que pour déterminer expérimentalement une valeur physique il est nécessaire (mais pas toujours suffisant) de trouver la moyenne (la valeur) et la largeur (l'incertitude). Sans la détermination de l'incertitude, l'expérience n'est pas complète : on ne peut la comparer ni avec une théorie ni avec une autre expérience. Nous avons également vu que cette incertitude contient trois contributions possibles. La première est l'incertitude naturelle liée aux changements des conditions d'expérience ou à la nature-même des grandeurs (en statistique ou en mécanique quantique). La

deuxième est l'incertitude statistique due à l'impossibilité de mesurer précisément la distribution initiale. La troisième est l'incertitude d'appareillage due à l'imperfection des outils de travail de l'expérimentateur.

Un expérimentateur se pose toujours deux questions. Premièrement, comment peut-on mesurer une grandeur physique, c'est-à-dire les caractéristiques de sa distribution : la moyenne et la largeur ? Deuxièmement, *comment* et *jusqu'où* faut-il diminuer cette incertitude (largeur) de l'expérience ? C'est pourquoi l'expérimentateur doit comprendre les relations entre les trois composantes de l'incertitude et trouver comment les minimiser : on peut diminuer l'incertitude naturelle en changeant les conditions de l'expérience, l'incertitude statistique en augmentant le nombre de mesures, l'incertitude d'appareillage en utilisant des appareils plus précis.

Cependant, on ne peut pas réduire les incertitudes infiniment. Il existe une limite raisonnable de l'incertitude. L'évaluation de cette limite est non seulement une question de temps et d'argent dépensés, mais c'est aussi une question de physique. Il ne faut pas oublier que, quelle que soit la grandeur à mesurer, nous ne pourrons jamais tenir compte de tous les facteurs physiques qui peuvent influencer sa valeur. De plus, tous nos raisonnements et discussions sont effectués dans le cadre d'un modèle ou, plus généralement, de notre vision du monde. Ce cadre peut ne pas être exact.

C'est pourquoi notre problème est de choisir des méthodes expérimentales et des méthodes d'estimation des incertitudes en adéquation avec la précision souhaitable et possible.

Diverses situations existent selon la précision désirée. Dans la première nous voulons seulement obtenir l'ordre de grandeur de la valeur mesurée ; dans ce cas, l'incertitude doit aussi être évaluée grossièrement. Dans la seconde nous désirons obtenir une précision de l'ordre de un à dix pour cent ; il faut alors faire attention en déterminant les incertitudes, car les méthodes choisies doivent évoluer en fonction de la précision requise. Plus on cherche de précision, plus la méthode doit être élaborée, mais le prix à payer est la lenteur des calculs et leur volume. Dans la troisième nous cherchons à obtenir une précision du même ordre de grandeur que celle de l'étalon correspondant au paramètre physique mesuré ; le problème de l'incertitude peut alors être plus important que celui de la valeur.

Dans cet ouvrage, nous considérons seulement les méthodes d'estimation d'erreurs dans la seconde situation. La plupart des paragraphes apporte réponse à une question concrète : comment calcule-t-on les incertitudes pour une expérience avec un petit nombre de mesures ? comment peut-on ajuster les paramètres d'une courbe ? comment compare-t-on une expérience et une théorie ? quel est le nombre de chiffres significatifs ? etc. Le lecteur qui connaît les bases de la statistique peut omettre sans problème les premiers paragraphes et chercher la réponse à sa question. Dans le cas contraire, l'ouvrage lui apporte l'information nécessaire sur les parties de la statistique utiles au traitement des incertitudes.

CHAPITRE 1

RAPPELS SUR LA THÉORIE DES PROBABILITÉS

Dans ce chapitre, nous avons réuni des notions de base de la théorie des probabilités : la définition d'une probabilité et ses propriétés élémentaires ainsi que l'introduction des distributions les plus fréquemment utilisées dans l'analyse des données expérimentales. Parmi ces distributions, celle de Gauss joue un rôle très particulier, c'est pourquoi la partie esssentielle de ce chapitre (paragraphes 1.2 et 1.4) lui est consacrée car elle et est indispensable à la compréhension du reste du livre.

1.1 PROBABILITÉS

Pour pouvoir décrire une grandeur physique en termes de probabilité il faut rappeler les définitions et les propriétés les plus simples. Pour les mesures les plus fréquentes faites en laboratoire nous n'avons pas besoin de toute la panoplie des méthodes de la statistique mathématique et notre expérience du monde est largement suffisante pour comprendre et assimiler les propriétés fondamentales des probabilités. Logiquement, chaque lecteur de ce livre a déjà eu l'occasion dans sa vie de jouer, au moins aux cartes et ainsi la notion de probabilité ne lui est pas étrangère.

1.1.1 DÉFINITIONS ET PROPRIÉTÉS

Supposons que l'on observe un événement E répété N_e fois (on dit que l'on prend un échantillon de N_e événements). Dans n cas, cet événement est caractérisé par une marque distinctive a (appelée aussi caractère). Si les résultats des événements dans cette suite sont indépendants, alors la probabilité $\mathcal{P}(a)$ que la marque a se manifeste est définie comme

$$\mathcal{P}(a) = \lim_{N_e \to \infty} \frac{n}{N_e}. \tag{1}$$

On voit toute de suite que la probabilité varie de 0 à 1

$$0 \leq \mathcal{P} \leq 1$$

et que la somme sur tous les caractères (de même nature) possibles $\{i\}, i = a, b, c, \ldots$ est égale à 1

$$\sum_i \mathcal{P}(i) = 1. \tag{2}$$

Un exemple d'événement est le tirage d'une carte du jeu. La marque distinctive serait la catégorie de couleur (pique, cœur, carreau ou trèfle). Pour un jeu de 52 cartes, la probabilité d'une catégorie de couleur est égale à $1/4$. On notera par A l'ensemble d'événements où ce signe s'est manifesté.

Introduisons deux opérations très simples avec les probabilités. Définissons par $A+B$ l'ensemble des événements dans lesquels la marque a ou la marque b, ou les deux, sont présentes (ici a et b peuvent être de nature différente). Par exemple, a est une catégorie de couleur, b est la valeur de la carte (le roi, la dame, etc.) De plus, définissons par AB l'ensemble des événements dans lesquels ces deux signes se manifestent simultanément.

Alors,

$$\mathcal{P}(A+B) = \mathcal{P}(A) + \mathcal{P}(B) - \mathcal{P}(AB).$$

C'est-à-dire, pour trouver la probabilité qu'un événement possède au moins une des marques nous devons, d'abord, ajouter deux probabilités $\mathcal{P}(A)$ et $\mathcal{P}(B)$. Cependant, certains événements peuvent avoir les deux signes en même temps et on les a comptés deux fois. C'est pourquoi il faut soustraire la probabilité $\mathcal{P}(AB)$.

Prenons un jeu de 52 cartes avec 13 cartes dans chaque couleur (le roi, la dame, le valet et 10 cartes numérotées de 1 à 10). Pour une carte tirée au hasard, la probabilité d'être soit le roi soit une carte de cœur (a étant le roi, b une carte de cœur) est égale à

$$\begin{aligned}\mathcal{P}(&\text{“soit le roi, soit une carte de cœur”})\\ &= \mathcal{P}(\text{“roi”}) + \mathcal{P}(\text{“cœur”}) - \mathcal{P}(\text{“roi de cœur”})\\ &= \frac{4}{52} + \frac{13}{52} - \frac{1}{52} = \frac{16}{52}.\end{aligned}$$

Introduisons une notion un peu plus compliquée. Supposons que l'événement A puisse se produire de n_a manières différentes, l'événement B de n_b manières et l'événement AB de n_{ab} manières. Si le nombre total de réalisations possibles est égal à N (ne pas confondre avec le nombre N_e d'événements introduit au début du paragraphe), alors

$$\mathcal{P}(A) = \frac{n_a}{N}, \qquad \mathcal{P}(AB) = \frac{n_{ab}}{N}.$$

On peut réécrire $\mathcal{P}(AB)$ comme

$$\mathcal{P}(AB) = \frac{n_a}{N} \cdot \frac{n_{ab}}{n_a} = \mathcal{P}(A) \cdot \frac{n_{ab}}{n_a}.$$

Parmi les n_a cas où l'événement A se produit, il y a une proportion n_{ab}/n_a où l'événement B s'est également produit. On peut introduire la probabilité correspondante qui s'appelle la *probabilité conditionnelle* $\mathcal{P}(A/B)$ de l'événement B, c'est-à-dire la probabilité d'observer B sous réserve que A se soit produit.

Ainsi, la dernière formule prend la forme

$$\mathcal{P}(AB) = \mathcal{P}(A) \cdot \mathcal{P}(B/A).$$

Si l'événement A n'a pas d'influence sur la probabilité d'événement B, on dit alors que les deux événements sont *indépendants* et

$$\mathcal{P}(B/A) = \mathcal{P}(B).$$

Dans ces conditions, on obtient pour la probabilité d'apparition de deux événements à la fois $\mathcal{P}(AB)$ une relation très importante :

$$\mathcal{P}(AB) = \mathcal{P}(A) \cdot \mathcal{P}(B), \tag{3}$$

ce qui montre que les probabilités des événements indépendants se multiplient. On utilisera cette propriété plusieurs fois dans ce livre.

Considérons l'exemple de notre jeu de 52 cartes. Soit A "un roi", B "une carte de cœur". Donc $n_a = 4$, $n_b = 13$, $N = 52$ et les probabilités correspondantes :

$$\mathcal{P}(A) = \frac{4}{52}, \qquad \mathcal{P}(B) = \frac{13}{52}.$$

Vu que $\mathcal{P}(AB) = \mathcal{P}(\text{"roi de cœur"}) = 1/52$, on conclut que

$$\mathcal{P}(AB) = \frac{1}{52} = \frac{4}{52} \cdot \frac{13}{52} = \mathcal{P}(A) \cdot \mathcal{P}(B),$$

et ainsi, dans le jeu de 52 cartes, ces deux événements sont indépendants.

Ajoutons juste une carte à notre jeu – un joker qui n'appartient à aucune catégorie de couleur. n_a, à nouveau, est égal à 4, n_b à 13, mais N est égal à 53. Donc,

$$\mathcal{P}(A) = \frac{4}{53}, \qquad \mathcal{P}(B) = \frac{13}{53}, \qquad \mathcal{P}(AB) = \frac{1}{53}.$$

On s'aperçoit facilement que

$$\mathcal{P}(A) \cdot \mathcal{P}(B) = \frac{4}{53} \cdot \frac{13}{53} = \frac{1}{53} \cdot \frac{52}{53} < \frac{1}{53} = \mathcal{P}(AB), \tag{4}$$

et ainsi ces deux événements ne sont plus indépendants dans le jeu de 53 cartes ! L'explication de cette différence est relativement simple : si nous savons qu'une carte est un roi alors elle ne peut pas être le joker, et ainsi nous avons déjà obtenu une certaine information pour déterminer sa catégorie de couleur.

1.1.2 GRANDEURS DISCRÈTES ET CONTINUES, FONCTIONS DE DISTRIBUTION

Une grandeur physique peut avoir une valeur numérique discrète ou continue. Dans le premier cas, on l'appellera grandeur "discrète", dans le deuxième, "continue". Les exemples de grandeurs discrètes sont la catégorie de couleur, la valeur de la carte, si

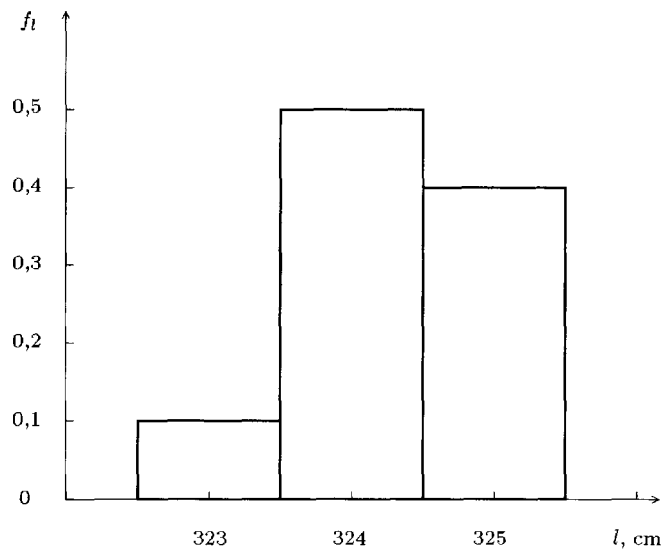


Figure 1.1 : Histogramme de la première série de mesures de la longueur l : sont portées sur l'axe des abscisses la valeur mesurée et sur l'axe des ordonnées la fréquence de son apparition $f_l = n_l/N$

l'on reprend notre exemple, ou le comptage d'un détecteur, si l'on considère des exemples plus physiques. Mais plus fréquemment en physique, on mesure des grandeurs continues, comme la longueur, la durée, le courant, etc.

Cette distinction des valeurs (ou des grandeurs) discrètes et continues est tout à fait justifiée. Néanmoins, en physique, on décrit assez souvent une grandeur continue par une valeur discrète et vice versa. De ce point de vue, cette séparation est, en partie, conventionnelle et les propriétés (ou même l'écriture) valables pour les valeurs discrètes seront utilisées pour les valeurs continues et inversement. On franchira cette frontière régulièrement, même parfois sans se rendre compte de ce que l'on fait. Cette attitude correspond à un parti pris de présentation. Le lecteur ne doit pas en déduire que le passage à la limite s'effectue dans tous les cas sans difficulté.

Pour illustrer le caractère conventionnel de cette distinction, considérons un exemple de mesure de la longueur d'une chambre (il est évident que la longueur est une grandeur continue) à l'aide d'un décimètre qui possède aussi des divisions centimètriques. Le fait même que nous disposions d'un décimètre avec des divisions nous oblige à décrire une grandeur continue à l'aide de valeurs entières donc discrètes (on aura un certain nombre de décimètres ou de centimètres). On peut aller plus loin et dire que la représentation d'une longueur par un nombre fini de chiffres est un passage obligé d'une valeur continue à une valeur discrète.

Bien sûr, il existe des situations où une valeur discrète ne peut pas être remplacée par une valeur continue, par exemple dans le jeu de cartes. Cependant, ces situations sont rares dans les expériences de physique. Nous observerons par la suite des passages des valeurs d'un type à l'autre. Les propriétés de probabilité resteront les mêmes dans

les deux cas. C'est pourquoi nous donnerons les démonstrations générales pour les variables continues et considérerons que les résultats s'appliquent aussi aux variables discrètes.

Continuons notre expérience mentale. Supposons qu'après avoir fait une dizaine de mesures rapides, nous ayons trouvé une fois la longueur de la chambre égale à 323 centimètres, cinq fois — 324 cm et quatre fois — 325 cm. Les résultats sont présentés sur la figure 1.1 qui s'appelle un "histogramme". Sur l'axe des abscisses, on montre la valeur mesurée et, sur l'axe des ordonnées, le nombre relatif $f_l = n_l/N$ (n_l mesures de la valeur l par rapport au nombre total N de mesures) c'est-à-dire la fréquence d'apparition de chaque valeur. Le sol n'était pas plat, notre décimètre n'était pas toujours droit, la longueur était, la plupart du temps, comprise entre 324 et 325 cm et nous ne savions pas dans quel sens il fallait l'arrondir. D'où la dispersion de nos résultats.

Pour clarifier la situation nous avons pris un instrument de mesure gradué en millimètres et en augmentant sensiblement le nombre de mesures nous avons obtenu les nouveaux résultats représentés sur la figure 1.2. Avec une autre échelle on retrouve les mêmes tendances : les résultats sont légèrement différents et se regroupent autour d'une certaine valeur.

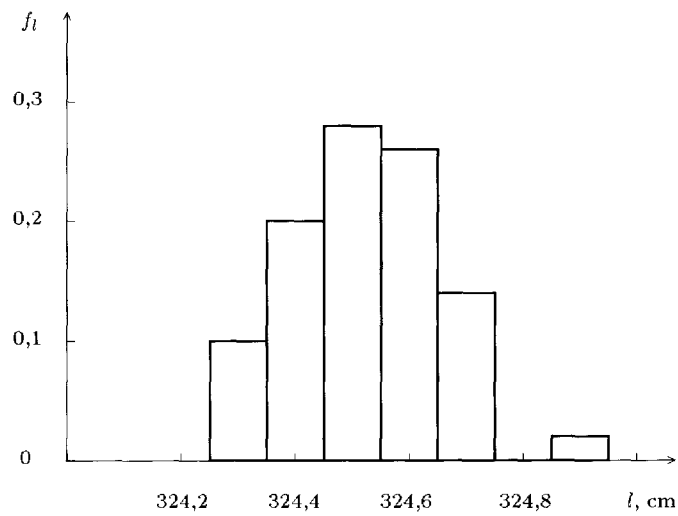


Figure 1.2 : Histogramme de la deuxième série de mesures de la longueur l : sont portées sur l'axe des abscisses la valeur mesurée et sur l'axe des ordonnées la fréquence de son apparition $f_l = n_l/N$

On peut continuer ainsi notre expérience en diminuant l'échelle et en augmentant le nombre de mesures dans chaque série. La forme des histogrammes tendra vers une forme en cloche qui, lorsque le nombre de mesures tend vers l'infini, peut être décrite par une fonction continue $f(x)$ (figure 1.3).

Chaque histogramme donne le nombre relatif de résultats se trouvant dans un inter-

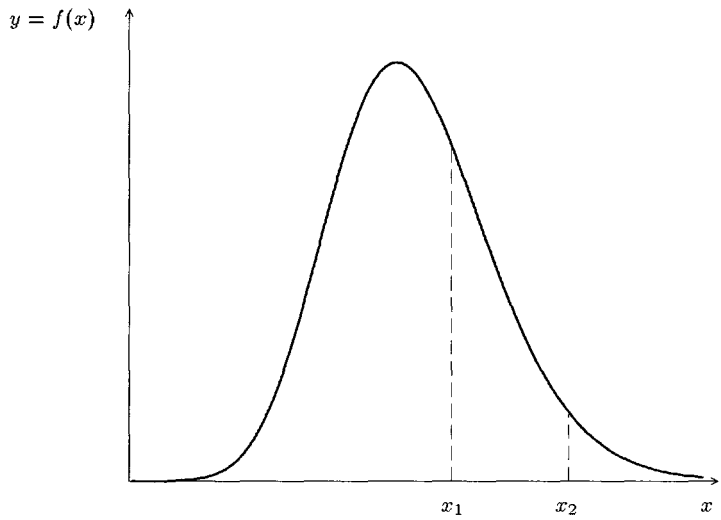


Figure 1.3 : Fonction de la densité de probabilité

valle donné. Ainsi, dans le cas d'un grand nombre de mesures et selon notre définition (1), le produit $f(x)dx$ donne la probabilité que la grandeur mesurée se trouve dans l'intervalle de x à $x + dx$. La fonction $f(x)$ représente la *densité de probabilité*. On l'appellera aussi la fonction de *distribution de probabilité*. x varie au hasard et s'appelle *variable aléatoire*.

D'après notre définition, la probabilité $\mathcal{P}$ de trouver la valeur dans l'intervalle compris entre x_1 et x_2 est égale à

$$\mathcal{P} = \int_{x_1}^{x_2} f(x)dx$$

qui est la somme (l'intégrale) de $f(x)$ pour toutes les valeurs de x entre x_1 et x_2.

Selon (2), $f(x)$ obéit à la condition

$$\int_{-\infty}^{+\infty} f(x)dx = 1, \tag{5}$$

ce qui signifie que la probabilité de trouver une valeur de x quelconque est égale à 1. Par commodité mathématique, nous avons pris ici des limites infinies pour l'intégrale. Mais une grandeur physique, par exemple la longueur, peut ne pas varier dans ces limites (elle ne peut pas être négative). Cela signifie que la fonction $f(x)$ utilisée pour décrire cette grandeur doit devenir très petite en dehors des limites que nous choisissons effectivement.

Pour une grandeur discrète qui prend les valeurs numériques $x_i \equiv \{x_1, x_2, \ldots\}$ nous

avons exactement la même relation de normalisation :

$$\sum_{i=1}^{\infty} \mathcal{P}(x_i) = 1, \tag{5'}$$

où $\mathcal{P}(x_i)$ est la probabilité de trouver la valeur x_i.

On peut souligner que le passage d'un histogramme à une fonction continue est analogue à la notion d'intégrale comme limite de la somme des aires de rectangles élémentaires sous la courbe représentant une fonction quand le nombre de divisions tend vers l'infini.

1.1.3 PROPRIÉTÉS DE LA FONCTION DE DISTRIBUTION

Comment pouvons-nous caractériser la fonction de distribution de probabilité $f(x)$? Théoriquement, il faut la connaître à chaque point x mais il est évident que ceci n'est pas réalisable expérimentalement : nous ne pouvons pas mesurer la probabilité pour chaque valeur x.

A priori, cette fonction $f(x)$ doit être positive, vu sa relation avec la probabilité, tendre vers zéro à plus l'infini et à moins l'infini assez rapidement pour que l'intégrale (5) existe, et avoir la forme de la courbe présentée sur la figure 1.3. Il est logique d'introduire au moins deux paramètres qui décrivent la position de la courbe (c'est-à-dire celle de son maximum) sur l'axe et son étalement.

Ainsi la première caractéristique de la distribution de probabilité $f(x)$ est la *valeur moyenne* de x

$$\overline{x} \equiv \int_{-\infty}^{+\infty} x f(x) dx. \tag{6}$$

Chaque valeur possible de x est multipliée par la probabilité de son apparition $f(x)dx$ et la somme (l'intégrale) est effectuée sur toutes les valeurs possibles.

Pour une variable discrète

$$\overline{x} \equiv \sum_{i=1}^{\infty} x_i \mathcal{P}(x_i). \tag{6'}$$

La barre sur x est la notation standard indiquant la valeur moyenne arithmétique.

Bien évidemment, nous supposons que cette intégrale (cette somme) ainsi que les intégrales (les sommes) que nous allons définir existent. C'est une hypothèse physique naturelle mais nous discuterons aussi d'exemples où elle n'est pas valable.

L'étalement de la distribution peut être décrit par *la variance* ou le carré de *l'écart-type* et défini par

$$D \equiv \sigma^2 \equiv \overline{(x - \overline{x})^2} = \int_{-\infty}^{+\infty} (x - \overline{x})^2 f(x) dx \tag{7}$$

pour une variable continue, et par

$$D \equiv \sigma^2 \equiv \sum_{i=1}^{\infty} (x_i - \overline{x})^2 \mathcal{P}(x_i) \tag{7'}$$

pour une variable discrète.

Pour chaque valeur de x, on considère l'écart par rapport à la valeur moyenne $\overline{x}$ et on calcule la valeur moyenne du carré de cet écart. Pourquoi avoir choisi cette caractéristique plutôt qu'une autre ? Parce que la simple valeur moyenne de l'écart est nulle. Nous aurions pu prendre comme caractéristique $\overline{|x - \overline{x}|}$ mais nous verrons à la fin de ce paragraphe que, sous cette forme, la variance ne présente pas certaines propriétés remarquables et fort utiles.

Il est facile de démontrer qu'avec la définition (7) le carré de l'écart-type s'écrit

$$\sigma^2 \equiv \overline{(x - \overline{x})^2} = \overline{x^2} - \overline{x}^2 \tag{8}$$

Prenons l'exemple le plus simple : une distribution de probabilité constante (voir figure 1.4) d'une grandeur x qui peut varier de a à b

$$f(x) = \begin{cases} 1/(b-a), & \text{si } a \leq x \leq b, \\ 0, & \text{autrement.} \end{cases} \tag{9}$$

La valeur de cette constante est définie par la condition de normalisation (5).

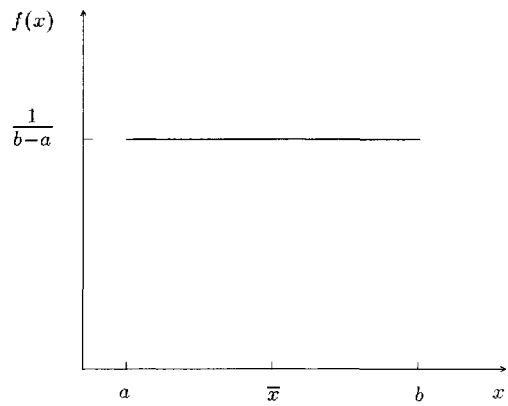


Figure 1.4 : Distribution constante

La valeur moyenne de x pour cette fonction de distribution est

$$\overline{x} = \int_a^b x f(x) dx = \frac{b+a}{2} \tag{10}$$

et sa variance :

$$\sigma^2 = \overline{x^2} - \overline{x}^2 = \frac{1}{3} \frac{b^3 - a^3}{(b-a)} - \left(\frac{b+a}{2}\right)^2 = \frac{(b-a)^2}{12} \tag{11}$$