

# Statistiques en L1 de psychologie

Sébastien LEURENT

Année 2020-2021



Notes de cours sous licence CC BY-SA 3.0

# Introduction

La psychologie est une discipline scientifique, dont la légitimité s'appuie notamment sur le recours à l'expérience pour confronter des hypothèses théoriques à une réalité expérimentale.

L'expérimentation pouvant être un processus long, coûteux et difficile, on est souvent contraint de travailler sur un nombre limité de sujets, considérés comme formant un échantillon d'une plus grande population. Dès lors une question qui survient rapidement est de savoir avec quelle précision et quel degré de certitude il est possible de tirer des conclusions à partir de tels échantillons.

Ce semestre de cours aboutira précisément à introduire des outils d'estimation, permettant de déterminer un intervalle de confiance, c'est à dire de connaître, en fonction de la confiance souhaitée, la *marge d'erreur* que l'on a quand on tire des conclusion à partir d'un petit échantillon.

Afin de suivre une progression logique aboutissant à ces outils, le cours s'articulera en chapitres :

Rappels et compléments de mathématiques } **Calcul** : Outil fondamental

Chapitre 1 : Statistique descriptives univariées }  
Chapitre 2 : Statistique descriptives bivariées } **Statistique descriptive** :  
Résumer des données expérimentales

Chapitre 3 : Introduction aux probabilités } **Probabilité** :  
Chapitre 4 : Loi Normale } Outil de prédiction du résultat d'une expérience,  
sous certaines hypothèse

Chapitre 5 : Estimation } **Statistique inférentielle** :  
Déductions à partir de mesures sur un échantillon

Comme indiqué ci-dessus, ils s'inscrivent dans des branches des mathématiques appelées respectivement « statistique descriptive », « probabilité » et « statistique inférentielle ».

## Déroulement du semestre

Les présentes notes de cours<sup>1</sup>, ainsi que les principaux documents pédagogiques (formulaire, feuilles d'exercices, examens d'années antérieures, etc) sont disponibles sur la page plubel du cours, à l'adresse <https://plubel-prod.u-bourgogne.fr/course/view.php?id=889>. Pour ceux qui n'auraient pas (encore) de compte à l'université de Bourgogne, ils sont aussi accessibles à l'adresse [http://leurent.perso.math.cnrs.fr/stats\\_ps1/](http://leurent.perso.math.cnrs.fr/stats_ps1/).

À chaque séance, il vous est demandé d'avoir le formulaire distribué en début de semestre, ainsi qu'une calculatrice scientifique (par exemple, la calculatrice libre NUMWORKS est parfaitement adaptée, de même que les modèles «Graph 35+ USB» de Casio et «TI-83» de Texas Instruments). La calculatrice (réinitialisée ou en mode examen) et le formulaire (vierge de toute annotation) sont autorisés aux contrôles et examens.

## Contrôle des connaissances (sous réserve de modifications par exemple pour raison sanitaire)

- Un contrôle terminal (CT) a lieu en fin de semestre et donne la moitié de la note de l'UE
- Une note de contrôle continue (CC) est obtenue au cours du semestre, et constitue l'autre moitié de la note de l'UE. Elle n'est pas rattrapable en deuxième session et est constituée
  - pour moitié d'un contrôle commun à tous les groupes de TD, autour du le 10 mars 2021.
  - pour moitié d'une note attribuée au sein de chaque groupe de TD, à partir de note(s) de contrôle(s) mais qui peut aussi prendre en compte la participation de chaque étudiant.

---

1. Ces notes de cours sont largement inspirées de notes de cours écrites par A. Jebrane, qu'il convient de remercier ici.

# Rappels et compléments de mathématiques

Ce chapitre introductif contient d'une part des « rappels » (première partie du chapitre) et d'autre part des compléments. Les rappels s'adressent uniquement aux étudiants les moins à l'aise avec les mathématiques de collège et de lycée, tandis que la partie « compléments » est indispensable pour l'ensemble des étudiants.

## Rappels

### Règles de calcul

On rappelle ici les principales règles de calcul concernant les opérations usuelles : à titre d'exemple on pourra considérer l'expression suivante :

$$-x + \frac{18}{3+6} + 3(-8+9) - 5^2$$

### Notation

- Le symbole «  $x$  » désigne un nombre arbitraire, avec lequel on peut procéder à des calculs même si on ne connaît pas sa valeur.
- Le symbole «  $+$  » désigne l'addition
- Le symbole «  $-$  » peut soit désigner une soustraction, soit « L'opposé » d'un nombre :
  - par exemple le nombre négatif  $-8$  est l'opposé de  $8$
  - de même  $-x$  désigne l'opposé de «  $x$  »
- La multiplication est notée par le symbole  $\times$ , ou par le symbole «  $*$  » sur les calculettes. Parfois, on n'écrit pas du tout la multiplication et le lecteur doit « deviner » sa présence, afin de donner un sens à une formule mathématique.  
Par exemple l'expression «  $3(-8+9)$  » n'aurait aucun sens si on n'ajoute pas une multiplication, de sorte que l'on comprends qu'elle signifie en fait  $3 \times (-8+9)$ .
- Les divisions sont notés indifféremment par le symbole «  $\div$  » (souvent utilisé par les calculatrices Casio), le symbole «  $/$  » (souvent utilisé par les calculatrices TI) ou un trait de fraction.  
Le trait de fraction présente l'avantage d'une plus grande lisibilité en évitant d'avoir à écrire explicitement certaines parenthèses. Ainsi, l'expression  $\frac{18}{3+6}$  signifie  $18 \div (3+6)$  (ou  $18/(3+6)$  ce qui est la même chose).  
Ce qui est « en haut » d'un trait de fraction s'appelle le *numérateur*, et ce qui est « en bas » s'appelle le *dénominateur*.
- $5^2$  désigne le carré de  $5$ . Sur certaines calculatrices ce carré est noté  $5^{\wedge}2$ .
- Lorsqu'on écrit des nombres qui ont beaucoup de chiffres, il est fréquent d'ajouter des espaces pour plus de lisibilité. Par exemple, on écrira  $12\,345,678\,9$  pour désigner le nombre  $12345,6789$ .

### Priorités de calcul

Pour calculer une expression comme  $-x + \frac{18}{3+6} + 3(-8+9) - 5^2$ , on effectue les étapes suivantes :

1. On calcule tout ce qui est dans des parenthèses, ou au numérateur ou dénominateur d'une fraction.  
Pour l'expression  $-x + \frac{18}{3+6} + 3(-8+9) - 5^2$ , on calcule donc  $3+6=9$  et  $-8+9=1$ . On obtient donc

$$-x + \frac{18}{3+6} + 3(-8+9) - 5^2 = -x + \frac{18}{9} + 3 \times 1 - 5^2$$

2. On calcule ensuite les puissances (ici le carré). On calcule donc  $5^2=25$  et on obtient

$$-x + \frac{18}{9} + 3 \times 1 - 5^2 = -x + \frac{18}{9} + 3 \times 1 - 25$$

3. On effectue ensuite les multiplications et les divisions : on calcule donc  $\frac{18}{9}=2$  et  $3 \times (1)=3$ . On obtient donc

$$-x + \frac{18}{9} + 3 \times 1 - 25 = -x + 2 + 3 - 25$$

4. Enfin, on termine par l'addition :

$$-x + 2 + 3 - 25 = -x - 20$$

Pour cet exemple on obtient donc, à l'issue des calculs, que  $-x + \frac{18}{3+6} + 3(-8+9) - 5^2 = -x - 20$ .

Ces étapes s'effectuent toujours dans cet ordre (parenthèses, puissances, multiplications et enfin additions). On peut en pratique s'en remettre aux calculatrices ou ordinateurs, qui connaissent cet ordre, mais lorsqu'on veut manipuler soi même ce type d'expressions sans erreurs de parenthèses, il est utile de connaître ces règles de priorité, qui signifient par exemple que  $8 \times 3 + 9$  est égal à  $(8 \times 3) + 9$  et pas à  $8 \times (3 + 9)$ .

### Autres règles de calcul

- Lorsqu'on multiplie quelque chose par une somme (c'est à dire une addition ou une soustraction) on peut « développer ». Par exemple cela signifie que

$$2 \times (9 - 8) = 2 \times 9 - 2 \times 8$$

- On peut changer l'ordre des termes dans une addition ou une soustraction (mais dans une soustraction, il faut faire attention aux signes). De même on peut changer l'ordre des facteurs dans une multiplication ou une division. Par exemple, cela signifie que  $9 - 6 + 7 = 9 + 7 - 6$ , et que  $\frac{8}{7} \times 4 = \frac{4 \times 8}{7}$ .
- Lorsqu'on multiplie une fraction par une expression, on multiplie simplement le numérateur par cette expression. Lorsqu'on divise une fraction par une expression, on multiplie simplement le dénominateur par cette expression.
- On peut ajouter et soustraire la même quantité sans modifier la valeur d'une expression. On peut de même, sans changer la valeur d'une expression, la multiplier et la diviser par la même quantité. Ainsi, on a par exemple  $4 + x - 4 = x$ , et  $\frac{4 \times 8}{4} = 8$ .

### Calculatrice

Si vous n'êtes pas habitués à utiliser la calculatrice, vous êtes très vivement encouragés à faire les premiers exercices du chapitre de la feuille d'exercice. Un corrigé de ces exercices (et d'une partie des autres exercices) est disponible à l'adresse [http://leurent.perso.math.cnrs.fr/stats\\_ps1/2020-2021/Exercices\\_corrigees.pdf](http://leurent.perso.math.cnrs.fr/stats_ps1/2020-2021/Exercices_corrigees.pdf).

Ayez bien en tête que

- Sur certaines calculatrices (en particulier les calculatrices TI), il y a deux touches « - » : Une touche  $\boxed{-}$  pour la soustraction, et une touche  $\boxed{(-)}$  pour l'opposé d'un nombre.
- Lorsqu'un nombre à virgule est très grand ou très petit, la calculatrice l'affiche en utilisant la notation **E** dont voici deux exemples :
  - **7.2E 12** signifie  $7,2 \times 10^{12}$  (c'est à dire  $7,2 \times 1000000000000$ ). Quand on multiplie 7,2 par  $10^{12}$ , on décale la virgule 12 fois vers la droite, aboutissant à 7200000000000.
  - **7.2E-7** signifie  $7,2 \times 10^{-7}$  (c'est à dire  $7,2 \times 0,0000001$ ). Quand on multiplie 7,2 par  $10^{-7}$ , on décale la virgule 7 fois vers la gauche, aboutissant à 0,00000072.
- La calculette peut garder en mémoire le résultat de certains calculs, pour éviter des erreurs d'arrondis en recopiant des valeurs arrondies. D'une part, « **Ans** » désigne le résultat du dernier calcul. D'autre part, « **→** » (tapé avec  $\boxed{\rightarrow}$  ou  $\boxed{\text{Sto}}$  selon les calculatrices) permet d'enregistrer le résultat d'un calcul. Les exercices 2 et 5 en donnent des exemples d'utilisation.
- La calculatrice peut aussi garder en mémoire plusieurs valeurs à la fois, sous la forme d'une liste, afin d'effectuer des opérations simultanées sur chaque valeur de la liste et/ou d'en calculer (entre autre la somme).
  - On peut créer des listes soit avec « **→** » (comme ci-contre), soit avec l'éditeur de liste (accessible depuis  $\boxed{\text{MENU}}$  sur Casio, ou depuis  $\boxed{\text{STAT}}$  sur TI).
  - Lorsqu'on fait des opérations sur les listes (comme les multiplier par des nombres ou même par d'autres listes), on réalise en fait l'opération sur chaque élément de la liste. Par exemple, la capture d'écran ci-contre, l'expression **List 1**×(**List 2**)<sup>2</sup> désigne la liste dont les éléments sont  $7 \times 3^2$ ,  $1 \times 6^2$  et  $9 \times 8^2$  (c'est à dire que pour calculer **List 1**×(**List 2**)<sup>2</sup>, on a multiplié à chaque fois un élément de **List 1** par le carré d'un élément de **List 2**)

```
(7,1,9)→List 1
      (7,1,9)
(3,6,8)→List 2
      (3,6,8)
List 1×(List 2)2
      (63,36,576)
```

## Complément : Notation $\sum$

La notation  $\sum$  sert à abréger certaines formules mathématiques, par exemple la somme ci-dessous :

$$1^2 + 2^2 + 3^2 + 4^2 + 5^2 + 6^2 + 7^2 \quad (1)$$

Cette somme se décrit simplement par une longue phrase : « on prend tous les nombres entre 1 et 7, on les met au carré, et on les ajoute ». Mathématiquement, cette longue phrase s'écrit simplement :

$$\sum_{i=1}^7 i^2 \quad (2)$$

Plus généralement, la notation  $\sum_{i=1}^r$  <formule compliquée> signifie que

- on calcule plusieurs fois de suite la "formule compliquée" en changeant juste à chaque fois un petit détail :
  - le  $i = 1$  en dessous du symbole  $\sum$  signifie que la première fois que l'on calcule la formule, on y remplace le symbole  $i$  par 1
  - la deuxième fois on remplace  $i$  par 2
  - etc

- Le  $r$  au dessus du symbole  $\sum$  signifie que l'on continue jusqu'à la fois où on remplace  $i$  par  $r$
- Enfin, on additionne les résultats ainsi obtenus

On aurait tout aussi bien pu utiliser un autre nom que «  $i$  » pour désigner ce nombre. Par exemple, la notation  $\sum_{k=5}^8 k^2$  signifie  $5^2 + 6^2 + 7^2 + 8^2$  ce qui vaut 174.

# Chapitre 1

## Statistique descriptive univariée

### 1.1 Introduction : types de variables

#### 1.1.1 Introduction

Les données statistiques que l'on est amené à analyser peuvent être de nature très différente. Cela se traduit par le fait qu'on ne peut pas effectuer les mêmes analyses sur des données de nature trop différente : par exemple, on peut calculer l'âge moyen au sein d'un groupe d'étudiants, alors que la moyenne de la couleur des yeux ne fait aucun sens.

On peut considérer un groupe d'étudiants, que l'on interroge sur leur nombre de frères et sœurs, leur taille, la couleur de leurs yeux, et leur humeur (on leur demande d'indiquer s'ils se sentent "de très bonne humeur", "de bonne humeur", "de relativement bonne humeur" ou "de mauvaise humeur"). On se rend alors compte que

- Cela ne fait pas tellement de sens de calculer la proportion d'étudiants qui mesurent 1m72 (par exemple), parce qu'il n'y a absolument aucune raison qu'un étudiant mesure 1,72m et pas 1,72036194762594 m. D'un point de vue mathématique, on considère que la valeur 1m72 n'a aucune chance de tomber (on tombe sur des valeurs avec beaucoup plus de chiffres après la virgule), donc « la proportion d'étudiants qui mesurent 1m72 » n'a pas tellement de sens (et il serait de même pour n'importe quelle autre taille que 1m72).
- Pour l'humeur des individus, on peut parler d'intervalles (par exemple « se sentir au moins "de bonne humeur" »), et on verra qu'on peut grâce à cela parler de médiane. Par contre, on ne peut pas parler de moyenne (nul ne sait définir la moyenne entre "relativement bonne humeur" et "bonne humeur").
- Pour la couleur des yeux, on ne peut parler ni de moyenne, ni de médiane et d'intervalle (car il n'y a pas de notion que certaines couleurs se situent entre d'autres couleurs).

Le tableau ci-dessous résume alors quels calculs feront sens pour chacun de ces types de données (ces calculs seront définis dans la suite du chapitre) :

	proportion d'une valeur	intervalle de valeurs	médiane	moyenne et écart type
Nb de frères/sœurs	✓	✓	✓	✓
Taille	✗	✓	✓	✓
Humeur	✓	✓	✓	✗
Couleur des yeux	✓	✗	✗	✗

### 1.1.2 Définitions

**Variable statistique** : Propriété qui varie d'un individu à l'autre.

- Exemples**
- le nombre de frères et soeurs
  - la taille
  - la couleur des yeux

**Remarque** Les variables statistiques seront généralement notées par des lettres majuscules comme  $X$ ,  $Y$ ,  $Z$ , etc.

**Individus** : Éléments dont on étudie les propriétés : on étudie la valeur que prend, pour chaque individu la variable statistique.

**Population** : Ensemble des individus que l'on considère.

**Remarque** Les *individus* ne sont pas nécessairement des personnes, ce peut être des lieux, des objets, etc. Dans ce cas le terme *Population* ne désignera pas un groupe de personnes.

- Exemples**
- Si on étudie la marque des téléphones présents dans un amphi, les *individus* sont des téléphones, la *population* est l'ensemble des téléphones présents dans cet amphi et la *variable statistique* est la marque. Dans ce cas, les étudiants présent dans l'amphi ne sont pas qualifiés d'*individus*, et les téléphones qui auraient été oubliés chez soi ou dans la voiture ne le sont pas non plus.
  - Si on s'intéresse au nombre d'habitants des pays membres de l'ONU, alors la France et l'Italie sont des individus, mais pas la Palestine, Taïwan, et le Vatican qui ne sont pas des États membres de l'ONU donc n'appartiennent pas à la *population*.

**Échantillon** : Sous-ensemble d'individus choisis au sein de la population.

**Exemple** Les étudiants inscrits en  $L_1$  de psychologie à l'université de Bourgogne pour l'année 2020/2021 forment un échantillon de l'ensemble des étudiants inscrits cette année en  $L_1$  à l'uB. Si on étudie la répartition homme/femme, cet échantillon est peu représentatif. Si en revanche on considère l'âge des étudiants (comme variable statistique) alors c'est un échantillon assez représentatif

**Proportion** : La proportion d'individus qui ont une certaine propriété est un nombre entre 0 et 1, qui s'obtient en divisant le nombre d'individus qui ont cette propriété par le nombre total d'individus considérés.

**Exemple** Si parmi 5 étudiant·e·s il y en a 2 qui ont une calculatrice TI et 3 qui ont une calculatrice Casio, alors la proportion qui ont une calculatrice TI est  $\frac{2}{5} = 0,4$ .

- Remarques**
- En pratique le « nombre total d'individus considérés » sera souvent la taille d'un échantillon (quand on n'a pas assez d'informations sur l'ensemble de la population, mais qu'on dispose d'informations sur l'échantillon).
  - Il arrive de multiplier par cent ce nombre entre 0 et 1, afin de l'exprimer en pourcentage. Ainsi la proportion 0,3 s'écrit parfois « 30% », tandis que la proportion 0,1 s'écrit parfois « 10% ». Dans ce cas, la proportion est toujours entre 0% et 100%.

**Modalités** : Valeurs que peut prendre la variable.

- Exemples**
- Les modalités de la variable « nombre de frères et soeurs » sont 0, 1, 2, 3, ...
  - “1m72” et “1m60” sont des modalités (parmi d'autres) de la variable “Taille”

**Variable quantitative** : Variable dont les modalités sont des nombres (éventuellement munis d'une unité), pour lesquels l'addition a un sens

- Exemples**
- La taille est une variable quantitative
  - Le numéro de sécurité sociale n'est pas une variable quantitative (ajouter des numéros de sécurité sociale ne fait aucun sens)

Au sein des variables quantitatives, on distingue deux types :



**Variable quantitative discrète** : Variable quantitative dont les modalités sont séparées par de nombreuses valeurs “interdites”

**Exemple** Le nombre de frère et soeur peut être égal à 1 ou 2, mais pas 1,5 ni 1,0356. C’est donc une variable quantitative discrète.

**Variable quantitative continue** : Variable dont les modalités ne sont séparées par aucune valeur interdite (elles forment un intervalle).

**Exemple** La taille.

**Variable qualitative** : Variable qui n’est pas quantitative

- Exemples**
- La couleur des yeux
  - Un numéro de téléphone

Au sein de variables qualitatives, on distingue deux types :

**Variable qualitative ordinale** : Variable qualitative dont les modalités sont ordonnées de manière claire et consensuelle.

- Exemples**
- L’humeur d’une personne : Si on demande à des personnes d’indiquer s’ils sont “de très bonne humeur”, “de bonne humeur”, “de relativement bonne humeur” ou “de mauvaise humeur”, alors cela forme une variable qualitative ordinale.
  - Au contraire, la nationalité n’est pas une variable ordinale, car il n’y a pas d’ordre bien défini, pas de consensus pour savoir si « français » se situe entre « suisse » et « italien » ou si c’est au contraire « suisse » qui se situe entre « français » et « italien ».

**Variable qualitative nominale** : Variable qualitative qui n’est pas ordinale.

**Exemples** La nationalité, la couleur des yeux, etc.

## 1.2 Regroupement de données

Considérons par exemple, que l’on demande à un groupe d’étudiants leur nombre de frères et soeurs, leur humeur, et leur taille. On peut obtenir les données suivantes :

Anais	Nicolas	Catherine	
de très bonne humeur 1 frère/soeur 1m69	de très bonne humeur 4 frères/soeurs 1m82	de très bonne humeur 0 frère/soeur 1m65	
Olivier	Enzo	Loïc	
de relativement bonne humeur 2 frères/soeurs 1m74	de très bonne humeur 3 frères/soeurs 1m91	de très bonne humeur 2 frères/soeurs 1m74	
Isabelle	Bernard	Rémi	Mélanie
de très bonne humeur 1 frère/soeur 1m62	de mauvaise humeur 0 frère/soeur 1m78	de mauvaise humeur 1 frère/soeur 1m81	de bonne humeur 4 frères/soeurs 1m60
Corinne			
de très bonne humeur 1 frère/soeur 1m69			

Une fois obtenues ces données, on va chercher à les mettre sous une forme plus synthétique et permettant de mieux les analyser. Dans ce chapitre, dédié aux statistiques descriptives univariées, on n’étudiera qu’une variable à la fois (sans considérer le lien entre les variables).

### 1.2.1 Regroupement par modalités

Pour une variable fixée, on calcule l'effectif de chaque modalité : c'est le nombre d'individus chez qui la variable prend cette valeur précise. On présente les effectifs sous forme de tableau : par exemple, pour l'humeur et le nombre de frères et soeurs on obtient les tableaux suivants

Modalité	de mauvaise humeur	de relativement bonne humeur	de bonne humeur	de très bonne humeur
Effectif	2	1	1	7

(a) Humeur

Modalité	0	1	2	3	4
Effectif	2	4	2	1	2

(b) Nombre de frères et soeurs

TABLE 1.1: Tableau présentant les effectifs des différentes modalités, pour l'humeur et le nombre de frères et soeurs d'un échantillon d'étudiants

**Remarque :** Nous n'utiliserons pas cette terminologie mais il est utile de savoir que certaines personnes (et certaines calculatrices) utilisent le terme « fréquence absolue » pour désigner les effectifs.

### Notation

On désigne par  $x_1$  la modalité de la première colonne,  
 par  $x_2$  la modalité de la deuxième colonne,  
 ...  
 par  $x_i$  la modalité de la  $i^{\text{ème}}$  colonne.  
 De même on désigne par  $n_1$  l'effectif de la première colonne,  
 par  $n_2$  l'effectif de la deuxième colonne,  
 ...  
 par  $n_i$  l'effectif de la  $i^{\text{ème}}$  colonne.

#### Exemple

Dans la table 1.1b (indiquant le nombre de frères et soeurs), on a  $x_1 = 0$ ,  $x_2 = 1$ ,  $x_3 = 2$ ,  $x_4 = 3$ ,  $x_5 = 4$ ;  $n_1 = 2$ ,  $n_2 = 4$ ,  $n_3 = 2$ ,  $n_4 = 1$ , et  $n_5 = 2$ .  
 Il y a 5 colonnes, d'où  $r = 5$ .  
 L'effectif total est  $n = 2 + 4 + 2 + 1 + 2 = 11$

Enfin on désigne par  $r$  le nombre total de colonnes, et par  $n$  l'effectif total, c'est à dire la "taille de l'échantillon". On a donc

$$n = n_1 + n_2 + \dots + n_r. \quad (1.1)$$

On prendra l'habitude d'écrire la formule (1.2) de manière plus compacte, sous la forme

$$n = \sum_{i=1}^r n_i, \quad (1.2)$$

qui signifie exactement la même chose.

### 1.2.2 Regroupement en classes

Il est parfois pertinent, en particulier pour des variables quantitatives continues, de regrouper les modalités au sein d'intervalles. Ces intervalles s'appellent des *classes*, et on calcule les effectifs comme précédemment :

Classe	[1,60 ; 1,65[	[1,65 ; 1,70[	[1,70 ; 1,75[	[1,75 ; 1,80[	[1,80 ; 1,85[	[1,85 ; 1,90[	[1,90 ; 1,95[
Effectif	2	3	2	1	2	0	1

TABLE 1.2: Regroupement en classes de la taille des étudiants interrogés

**Rappel :** l'intervalle  $[1,75 ; 1,80[$  désigne l'ensemble des nombres qui valent au moins 1,75 mais sont plus petits que 1,80. Il contient par exemple 1,75, 1,768921, et 1,79, mais pas 1,80, ni 2,3, ni  $-12,157$ .

**Remarque :** Regrouper ainsi les données en classes fait perdre une partie de l'information : pour chaque individu au lieu de retenir la valeur précise de la variable, on note juste à quelle classe elle appartient.

## Notation

On désigne par  $r$  le nombre de colonnes.

On désigne par  $a_1$  le minimum de la classe de la première colonne,  
par  $a_2$  le minimum de la classe de la deuxième colonne,  
...  
par  $a_r$  le minimum de la classe de la dernière colonne,  
par  $a_{r+1}$  le maximum de la classe de la dernière colonne.

De même on désigne par  $n_1$  l'effectif de la première colonne,  
par  $n_2$  l'effectif de la deuxième colonne,  
...  
par  $n_r$  l'effectif de la dernière colonne.

### Exemple

Dans la table 1.2, on a  $a_1 = 1,60$ ,  
 $a_2 = 1,65$ ,  $a_3 = 1,70$ ,  $a_4 = 1,75$ ,  
 $a_5 = 1,80$ ,  $a_6 = 1,85$ ,  $a_7 = 1,90$ ,  
 $a_8 = 1,95$ ;  $n_1 = 2$ ,  $n_2 = 3$ ,  $n_3 = 2$ ,  
 $n_4 = 1$ ,  $n_5 = 2$ ,  $n_6 = 0$ , et  
 $n_7 = 1$ .

Il y a 7 colonnes, d'où  $r = 7$ .

L'effectif total est  $n = 2 + 3 + 2 + 1 + 2 + 1 = 11$

### 1.2.3 Fréquences et fréquences cumulées

Pour chaque modalité (ou chaque classe) on calcule une « fréquence relative », c'est à dire la proportion d'individus chez qui la variable prend cette valeur. En notant  $f_i$  la fréquence relative de la  $i^{\text{ème}}$  colonne, on a

$$f_i = \frac{n_i}{n} \quad (1.3)$$

La table 1.3 reprend les tables 1.1 et 1.2 en ajoutant une ligne avec les fréquences relatives (la dernière ligne de ces tables sera décrite au paragraphe suivant). Dans cette table la ligne "Fréquence" s'obtient en divisant la ligne "Effectif" par l'effectif total.

**Exemple** Pour le nombre de frères et soeurs, la fréquence de la troisième colonne est  $f_3 = \frac{n_3}{n} = \frac{2}{11} \simeq 0,182$ . Cela signifie que  $\mathbb{P}_r[X = 2] \simeq 0,182$ , c'est à dire que parmi cet échantillon, il y a environ 18,2% des étudiants qui ont exactement 2 frères et soeurs.

- Remarques**
- Dans ce cours, on utilisera simplement le terme « fréquence » pour désigner les fréquences relatives.
  - Pour les fréquences (et les fréquences cumulées ci-dessous), on gardera de préférence au moins trois chiffres après la virgule.

**Fréquences cumulées** Après avoir calculé ces fréquences, on peut calculer les fréquences cumulées, c'est à dire les sommes des fréquences des premières colonnes. Ces fréquences cumulées constituent la dernière ligne des tables 1.3.

**Exemple** Pour le nombre de frères et soeurs, la fréquence cumulée de la troisième colonne est  $f_1 + f_2 + f_3 = 0,182 + 0,364 + 0,182 = 0,728$ . Cela signifie que  $\mathbb{P}_r[X \leq 2] \simeq 0,728$ , c'est à dire que parmi cet échantillon, il y a environ 72,8 % des étudiants qui ont au maximum 2 frères et soeurs.

**Remarque importante** Pour une variable regroupée en classe, comme la taille, la fréquence relative d'une colonne est associée au maximum de l'intervalle correspondant.

Humeur	mauvaise	relativement bonne	bonne	très bonne
Effectif	2	1	1	7
Fréquence	0,182	0,091	0,091	0,636
Fréquence cumulée	0,182	0,273	0,364	1,000

(a) Humeur

Nombre de frères/soeurs	0	1	2	3	4
Effectif	2	4	2	1	2
Fréquence	0,182	0,364	0,182	0,091	0,182
Fréquence cumulée	0,182	0,546	0,728	0,819	1,001

(b) Nombre de frères et soeurs

Taille	[1,60 ; 1,65[	[1,65 ; 1,70[	[1,70 ; 1,75[	[1,75 ; 1,80[	[1,80 ; 1,85[	[1,85 ; 1,90[	[1,90 ; 1,95[
Effectif	2	3	2	1	2	0	1
Fréquence	0,182	0,273	0,182	0,091	0,182	0,000	0,091
Fréquence cumulée	0,182	0,455	0,637	0,728	0,910	0,910	1,001

(c) Taille

TABLE 1.3: Humeur, taille et nombre de frères et soeurs : calcul des fréquences et des fréquences cumulées

Par exemple pour la table 1.3c, la fréquence cumulée de la deuxième colonne est  $f_1 + f_2 = 0,182 + 0,273 = 0,455$ . Cela signifie que  $\mathbb{P}_r[Y < 1,70] \simeq 0,455$ , où l'on note bien que cette fréquence cumulée correspond à la taille "1m70".

**Remarque** La dernière fréquence cumulée est toujours égale à 1. Toutefois, quand on la calcule on commet souvent des erreurs d'arrondis qui peuvent aboutir à trouver par exemple 0,999 ou 1,001 au lieu de 1.

Il est possible de l'éviter en gardant (sur la calculatrice) plus de chiffres après la virgule, à défaut de quoi il vaut mieux garder 0,999 ou 1,001 plutôt que de « tricher » pour obtenir exactement 1.

## 1.3 Représentations graphiques

### 1.3.1 Représentation des fréquences

Un premier type de graphique consiste à représenter les fréquences. Il y a principalement trois situations :

- Pour des variables qualitatives on utilise fréquemment un diagramme en camembert comme celui de la figure 1.1. Les surfaces des différents quartiers sont proportionnelles aux fréquences. Pour obtenir cela,

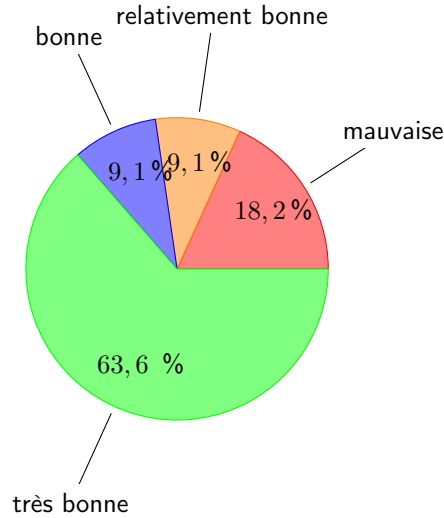


FIGURE 1.1: Humeur des étudiants : représentation en diagramme circulaire ("camembert")

il suffit de donner à chaque quartier l'angle  $f \times 360^\circ$ , où  $f$  est la fréquence.

- Pour des variables numériques discrètes, on peut utiliser un diagramme en bâton, où la hauteur de chaque bâton donne la fréquence d'une modalité. La figure 1.2 montre un diagramme de ce type pour illustrer le nombre de frères et soeurs de l'échantillon d'étudiants considéré.
- Pour des données regroupées en classes, on peut utiliser des histogrammes, constitués de rectangles dont la largeur indique la taille de chaque intervalle, et la surface est proportionnelle à la fréquence. La figure 1.3 représente par un tel histogramme la taille des étudiants de l'échantillon.

Dans le cas présent, l'histogramme met en évidence que la taille de ces étudiants se concentre d'une part entre 1m65 et 1m70 et d'autre part entre 1m80 et 1m90. En analysant plus ces données, on se rend aisément compte que pour l'essentiel, les femmes ont des tailles entre 1m65 et 1m70 alors qu'une partie importante des hommes mesurent entre 1m80 et 1m90. L'histogramme présente deux pics car on a regroupé deux types d'individus aux caractéristiques distinctes.

### 1.3.2 Représentation des fréquences cumulées

Dans le cas de données regroupées en classes, une autre représentation graphique sera utile : le **polygone des fréquences cumulées**. Il s'agit d'une représentation graphique approchée de la fonction  $F_X(a) = \mathbb{P}_r[X \leq a]$ .

**Définition** Étant donnée une variable statistique  $X$ , la fonction  $F_X$  définie par  $F_X(a) = \mathbb{P}_r[X \leq a]$  s'appelle la fonction de répartition de  $X$ .

**Construction du polygone des fréquences cumulées** La figure 1.4 représente ce polygone des fréquences cumulées pour la taille des étudiants de l'échantillon.

Décrivons la façon dont il est construit :

- La taille des étudiants est notée  $Y$ , donc on considère la fonction de répartition  $F_Y$  définie par  $F_Y(a) = \mathbb{P}_r[Y \leq a]$ .
- Les fréquences cumulées calculées indiquent que  $\mathbb{P}_r[Y \leq 1,65] \simeq 0,182$ ,  $\mathbb{P}_r[Y \leq 1,7] \simeq 0,455$ , etc,  
c'est à dire que  $F_Y(1,65) \simeq 0,182$ ,  $F_Y(1,7) \simeq 0,455$ , etc,  
donc on place des points de coordonnées  $(1,65; 0,182)$ ,  $(1,7; 0,455)$ , etc.

On a ainsi un point pour chaque colonne du tableau.

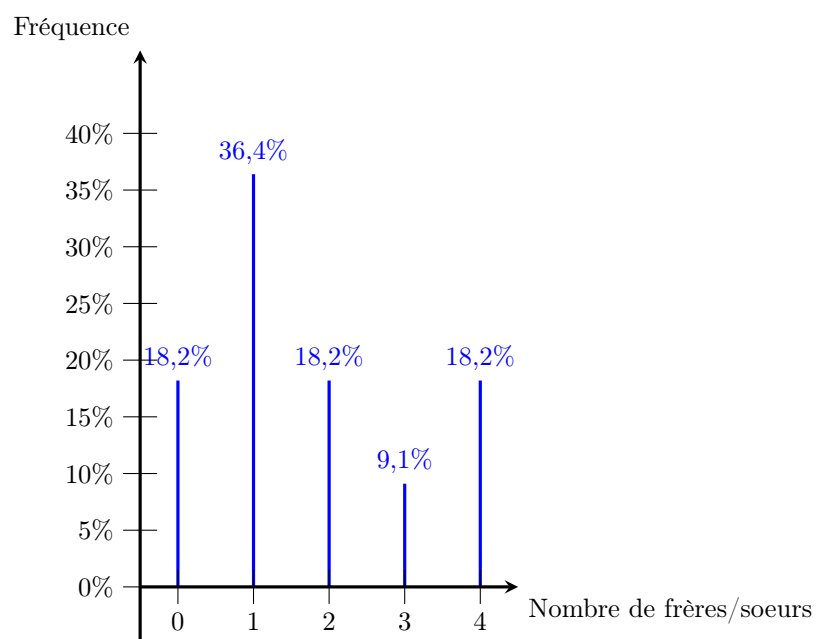


FIGURE 1.2: Nombre de frères et soeurs : représentation en bâtons

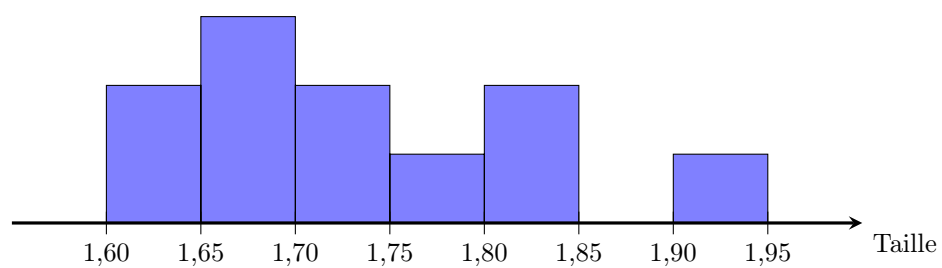


FIGURE 1.3: Histogramme des fréquences, pour la taille des étudiants de l'échantillon

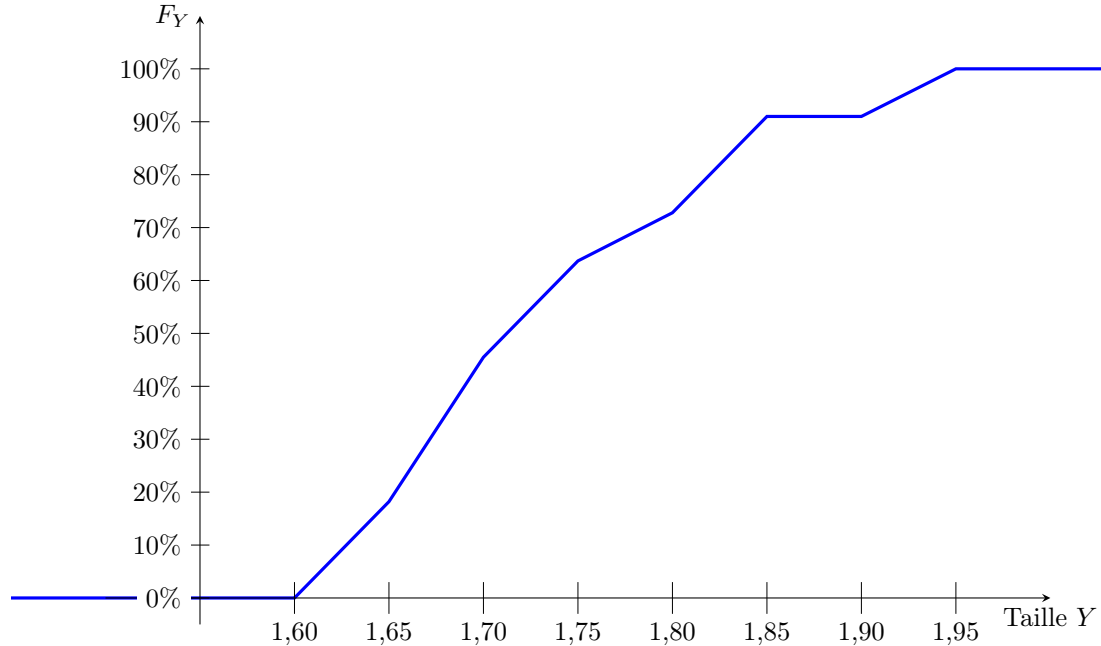


FIGURE 1.4: Polygone des fréquences cumulées, pour la taille des étudiants de l'échantillon

**Remarque :** On peut choisir d'écrire ces fréquences en pourcentage sur le dessin, pour plus de lisibilité.

- De plus  $\mathbb{P}_r[Y \leq 1,6] = 0$ , donc on ajoute un point de coordonnées (1,6; 0)
- Entre ces points on manque d'information pour approximer la fonction  $F_Y$ . On choisit de relier ces points par des segments de droites.
- Lorsque  $t \leq 1,6$ , on a  $\mathbb{P}_r[Y \leq t] = 0$ . En conséquence on trace une demi-droite horizontale à gauche du point de coordonnées (1,6; 0).
- De même, lorsque  $t > 1,95$ , on a  $\mathbb{P}_r[Y \leq t] = 1$ . En conséquence on trace une demi-droite horizontale à droite du point de coordonnées (1,95; 1).

## 1.4 Calcul d'indicateurs

### 1.4.1 Médiane

**Définition** La médiane est définie si  $X$  est une variable numérique ou ordinale. C'est une modalité notée  $\text{Med}$  telle que  $\mathbb{P}_r[X \geq \text{Med}] \geq 0,5$  et que  $\mathbb{P}_r[X \leq \text{Med}] \geq 0,5$ .

C'est à dire qu'il y a une moitié des individus chez qui la variable est inférieure (ou égale) à la médiane, et l'autre moitié des individus chez qui la variable est supérieure (ou égale) à la médiane.

**Mode de calcul et convention** Pour la calculer on ordonne les valeurs, et on choisit la  $\left(\frac{n+1}{2}\right)^{\text{ème}}$  valeur. Si  $\frac{n+1}{2}$  n'est pas entier, on choisit le milieu entre la  $\left(\frac{n}{2}\right)^{\text{ème}}$  et la  $\left(\frac{n}{2} + 1\right)^{\text{ème}}$ .

**Exemple** Pour l'humeur des étudiants : on a  $n = 11$  donc  $\frac{n+1}{2} = 6$ . Si on note "M" pour « Mauvaise humeur », "R" pour « relativement bonne », "B" pour « bonne » et "T" pour « Très Bonne », alors en ordonnant les valeurs on a : M M R B T T T T T T T, et la 6<sup>ème</sup> valeur est "T". La médiane est donc *de très bonne humeur*.

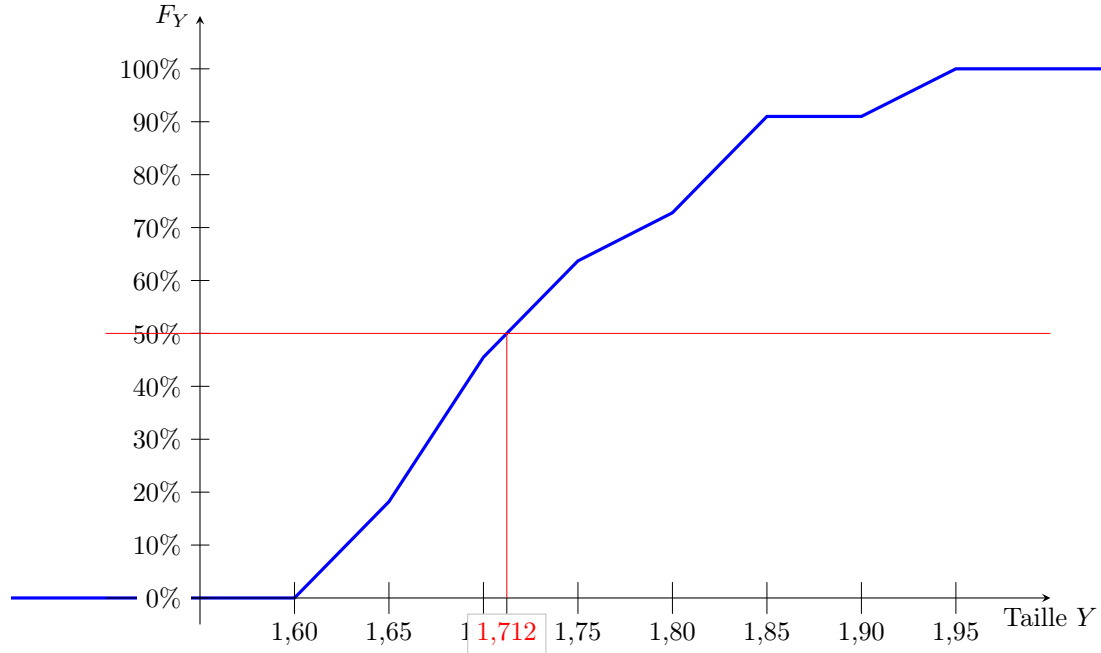


FIGURE 1.5: Lecture graphique de la taille médiane des étudiants de l'échantillon

**Cas de données regroupées en classes** Pour des données regroupées en classe on résout de manière approchée l'équation  $\mathbb{P}_r[x \leq \text{Med}] = 0,5$ , c'est à dire  $F_X(\text{Med}) = 0,5$ . Cela peut soit se résoudre par lecture graphique (plus intuitif mais moins précis), soit en utilisant une formule du formulaire.

1. **Lecture graphique** On cherche à déterminer la taille médiane des étudiants de l'échantillon. On doit résoudre  $F_Y(\text{Med}) = 50\%$ , c'est à dire qu'on lit l'abscisse du point où le polygone des fréquences cumulées croise la droite d'ordonnée 50%. Comme indiqué en figure 1.5, on lit que la taille médiane est environ 1,712m.
2. **Formule du formulaire** On peut démontrer (en utilisant le théorème de Thalès), la formule suivante pour calculer la valeur approchée de la médiane :
  - on appelle  $a_i$  et  $a_{i+1}$  le minimum et le maximum de la première classe dont la fréquence cumulée est supérieure à 0,5.
  - la médiane est alors donnée par la formule  $\text{Med} \simeq a_i + \frac{a_{i+1} - a_i}{F_X(a_{i+1}) - F_X(a_i)} (0,5 - F_X(a_i))$ .

**Exemple** Dans le cas présent (taille médiane des étudiants de l'échantillon), la première classe à avoir une fréquence cumulée supérieure à 0,5 est  $[1,7 ; 1,75[$ . On note donc  $a_i = 1,7$  et  $a_{i+1} = 1,75$ . On a donc  $F_Y(a_i) = 0,455$  et  $F_Y(a_{i+1}) = 0,637$ , d'où  $\text{Med} \simeq 1,7 + \frac{1,75 - 1,7}{0,637 - 0,455} (0,5 - 0,455) \simeq 1,712$ .

**Quartiles** On a vu que la médiane consiste à séparer les données en deux moitiés égales de part et d'autre de la médiane, de sorte que la médiane est « la valeur du milieu ».

On aurait aussi pu par exemple les séparer en "trois quart" d'un côté, "un quart de l'autre". Dans ce cas on parle non pas de médiane mais de quartile.

Dans ce cours, on ne calculera de quartile que pour des données regroupées en classes, et il faut alors résoudre  $F_X(Q_1) = 25\%$  pour définir le premier quartile ( $Q_1$ ) ou  $F_X(Q_3) = 75\%$  pour définir le troisième quartile ( $Q_3$ ). On peut soit faire une résolution graphique, soit utiliser la même formule que pour la médiane en remplaçant 0,5 par 0,25 pour le premier quartile ( $Q_1$ ) et par 0,75 pour le troisième ( $Q_3$ ).

**Attention** la classe  $[a_i, a_{i+1}[$  à considérer change aussi.



### 1.4.2 Moyenne

**Définition** Si  $X$  est une variable quantitative, et si sur un échantillon de taille  $n$  elle prends les valeurs  $x_1, x_2, \dots, x_n$ , alors sa moyenne sur l'échantillon est

$$m(X) = \frac{1}{n} \sum_{i=1}^n x_i.$$

C'est à dire que l'on additionne les valeurs obtenues pour chaque individu et on divise par le nombre d'individus. La valeur que l'on obtient indique une valeur typique de  $X$  sur cet échantillon.

**Exemple** Pour le nombre de frères et soeurs dans notre échantillon d'étudiants : on pose  $x_1 = 1, x_2 = 4, x_3 = 0, x_4 = 2, x_5 = 3, x_6 = 2, x_7 = 1, x_8 = 0, x_9 = 1, x_{10} = 4$  et  $x_{11} = 1$ . On a donc  $m(X) = \frac{1+4+0+2+3+2+1+0+1+4+1}{11} \simeq 1,727$ .

**Mode de calcul pour des données regroupées par modalités** Pour des données regroupées par modalités, la moyenne se calcule selon la formule ci-dessous (en utilisant les notations introduites à la fin de la section 1.2.1 :

$$m(X) = \frac{1}{n} \sum_{i=1}^r n_i x_i$$

**Exemple** Pour le nombre de frères et soeurs dans notre échantillon d'étudiants : on a cette fois-ci a  $x_1 = 0, x_2 = 1, x_3 = 2, x_4 = 3, x_5 = 4, x_6 = 5, x_7 = 6$ ;  $n_1 = 2, n_2 = 4, n_3 = 2, n_4 = 1, n_5 = 2, n_6 = 0$ , et  $n_7 = 0$ .

La formule s'écrit donc  $m(X) = \frac{2 \times 0 + 4 \times 1 + 2 \times 2 + 1 \times 3 + 2 \times 4}{11} \simeq 1,727$ .

Bien entendu, cela revient au même que la formule précédente, la seule différence est la façon de présenter les données.

**Mode de calcul pour des données regroupées en classes** Pour les données regroupées en classes, on appelle  $c_i$  le milieu de la  $i^{\text{ème}}$  classe, c'est à dire

$$c_i = \frac{a_i + a_{i+1}}{2}$$

On peut approximer la moyenne par l'expression ci-dessous

$$m(X) \approx \frac{1}{n} \sum_{i=1}^r n_i c_i$$

**Exemple** Pour la taille des étudiants de l'échantillon, on a  $c_1 = \frac{1,6+1,65}{2} = 1,625, c_2 = \frac{1,65+1,7}{2} = 1,675$ , etc. En conséquence la formule devient  $m(Y) \approx \frac{2 \times 1,625 + 3 \times 1,675 + 2 \times 1,725 + 1,775 + 2 \times 1,825 + 0 \times 1,875 + 1,925}{11} \simeq 1,734$ .

**Remarque** Dans le cas des données regroupées en classes, on manque d'information précise (on sait juste combien de valeurs appartiennent à chaque intervalle, mais pas où précisément elles se situent au sein de l'intervalle). Pour l'exemple ci-dessus, on peut comparer avec la valeur exacte de la moyenne, à savoir  $m(Y) = \frac{1,69+1,82+1,65+1,74+1,91+1,74+1,62+1,78+1,81+1,6+1,69}{11} \simeq 1,732$ .

### 1.4.3 Écart type

**Définition** Si  $X$  est une variable quantitative, on définit son écart-type  $s(X)$  par

$$s(X) = \sqrt{\text{Var}(X)}$$

où la variance  $\sqrt{\text{Var}(X)}$  est définie par

$$\text{Var}(X) = m\left(X - m(X)\right)^2 = m(X^2) - \left(m(X)\right)^2$$

L'écart type donne une valeur typique de la différence entre deux valeurs de  $X$  au sein cet échantillon.

**Mode de calcul** On calcule  $m(X^2)$  avec la même formule que  $m(X)$ , mais en mettant au carré les modalités  $x_i$  (ou les centres de classes  $c_i$ ) :

- Pour des données brutes on a  $m(X^2) = \frac{1}{n} \sum_{i=1}^n (x_i^2)$ .
- Pour des données regroupées par modalités, on a  $m(X^2) = \frac{1}{n} \sum_{i=1}^r n_i (x_i^2)$ .
- Pour des données regroupées en classes, on a  $m(X^2) = \frac{1}{n} \sum_{i=1}^r n_i (c_i^2)$ .

**Exemple** Pour le nombre de frères et soeurs, on obtient

$$m(X^2) = \frac{2 \times (0^2) + 4 \times (1^2) + 2 \times (2^2) + 1 \times (3^2) + 2 \times (4^2)}{11} \simeq 4,818.$$

D'où  $\text{Var}(X) \simeq 4,818 - (1,727)^2 \simeq 1,835$  et  $s(X) \simeq \sqrt{1,835} \simeq 1,355$ .

#### 1.4.4 « Écart type corrigé »

Dans ce cours nous utiliserons très peu l'écart-type corrigé  $\hat{s}(X)$ , bien qu'il soit plus fréquemment utilisé pour estimer l'écart-type d'une grande population à partir de celui d'un échantillon. Sa définition est

$$\hat{s}(X) = \sqrt{\frac{n}{n-1}} s(X).$$

Les calculatrices calculent généralement deux écart-types (voir formulaire). Le plus grand des deux est l'écart type corrigé  $\hat{s}$ , tandis que le plus petit des deux est l'écart type  $s$  considéré dans ce cours.

## Chapitre 2

# Statistique descriptive bivariée

### 2.1 Exemple et introduction

On considère un échantillon de 16 personnes de 35 à 80 ans, auxquels on attribue une note indiquant leurs performances mémorielles. On note leur âge  $X$  et leur performances mémorielles  $Y$ .

Les données mesurées sont les suivantes :

X	80	56	67	72	67	51	80	42	56	47	53	37	48	46	44	35
Y	22	38	33	36	22	47	30	45	48	55	38	81	46	33	51	60

Au cours du chapitre précédent, on a vu comment étudier d'une part l'âge des individus de l'échantillon, et d'autre part leur performances mémorielles, de manière complètement dissociée. Mais l'intérêt de cet échantillon réside précisément dans la possibilité de mettre en évidence un lien entre l'âge et la mémoire, ce qui nécessite l'étude simultanée de ces deux propriétés.

L'objet de ce chapitre sera précisément d'étudier le lien entre deux variables définies sur les mêmes individus.

#### 2.1.1 Définitions

**Variable appariées** : Si deux variables sont définies pour les mêmes individus, on dit que ce sont des variables appariées.

- Exemples**
- La note de *statistiques* et la note de *psychologie du développement* des étudiants de L1 de psychologie sont appariées.
  - Au sein des couples dijonnais, on peut considérer le revenu du mari et le revenu de la femme. On considère alors que les *individus* sont les couples, ces deux variables sont définies pour les mêmes « individus », elles sont appariées.
  - En revanche, le revenu des hommes dijonnais et le revenu des femmes dijonnaises ne sont pas des variables appariées, elles portent sur des individus différents.

**Remarque** En pratique, « appariées » signifie que chaque valeur d'une variable est associée à une valeur de l'autre variable (correspondant au même individu). Cette condition est nécessaire pour étudier le lien entre deux variables.

**variables dépendante et indépendante** Pour des variables appariées, si l'une des deux variables est “manipulable” par l'expérimentateur, on l'appelle en sciences humaines *variable indépendante* : il peut par exemple s'agir du dosage d'un traitement que l'on administre, ou du sexe des personnes que l'on choisit d'interroger. Dans ce cas, l'autre variable est appelée *variable dépendante*.

<b>Exemple</b>	variable indépendante	variable dépendante
	dosage d'un traitement	intensité de la douleur manifestée par les malade
	sexe des personnes interrogées	taille
	alimentation de rats de laboratoires	état de santé
	revenu des parents	niveau d'étude des enfants

**Remarque** On constate sur ces exemples qu'une variable peut être manipulable soit en fixant artificiellement sa valeur (dosage d'un traitement, alimentation de rats, etc), soit en choisissant la composition de l'échantillon (sexe, revenu, etc).

**Mise en garde** On peut être tenté de considérer la variable dépendante comme une *conséquence* dont la variable indépendante serait la *cause*. Cette interprétation peut sembler raisonnable quand on considère l'état de santé de rats selon l'alimentation qu'on leur administre. Mais elle peut être trompeuse dans deux nombreuses situations : par exemple pour le lien entre le revenu des parents et le niveau d'étude des enfants, on peut tout à fait imaginer que ce n'est pas directement le revenu des parents qui impacte les études de leurs enfants mais que ce sont simplement les mêmes causes qui impactent ces deux variables et contribuent à ce que l'on détecte un lien entre les deux variables.

**Notation** Lorsqu'il y a une variable indépendante et une variable dépendante, on appelle  $X$  la variable indépendante et  $Y$  la variable dépendante.

Dans le cas présent, l'âge est la variable indépendante tandis que les performances mémorielles sont la variable dépendante.

**Objectifs** Dans ce chapitre on cherchera à répondre principalement à deux questions :

- Y a-t'il un fort lien entre les deux variables ? On appellera **corrélacion** l'intensité de ce lien, que l'on mesurera à l'aide de *coefficients de corrélation*.
- Peut-on prédire la variable d'une variable en fonction de l'autre variable ? On le fera dans ce chapitre au moyen d'une **régression linéaire**.

## 2.2 Nuage statistique

Une première façon de synthétiser efficacement les données et de se faire une idée du lien entre les deux variables consiste à réaliser un nuage de points : pour chaque individu, on place un point qui est positionné horizontalement en la valeur  $x_i$  que prend la variable  $X$  chez cet individu, et verticalement en la valeur  $y_i$  de la variable  $Y$ , obtenant donc la figure 2.1.

Sur cet exemple, on constate qu'il n'y a aucun point où  $X$  et  $Y$  sont simultanément élevés. Cela montre déjà un lien entre les variables.

## 2.3 Coefficients de corrélation

Dans ce cours, on abordera deux coefficients de corrélacions différents :

Le coefficient de corrélation linéaire traduit le fait que deux variables soient liées par une relation linéaire (ou affine), c'est-à-dire le fait que les points du nuage statistique soient concentrés autour d'une droite.

Le coefficient de corrélation des rangs de Spearman traduit le fait qu'une des variables augmente (ou diminue) quand l'autre augmente. Dans l'exemple précédent, ce coefficient permettrait de confirmer que les performances mémorielles diminuent avec l'âge.

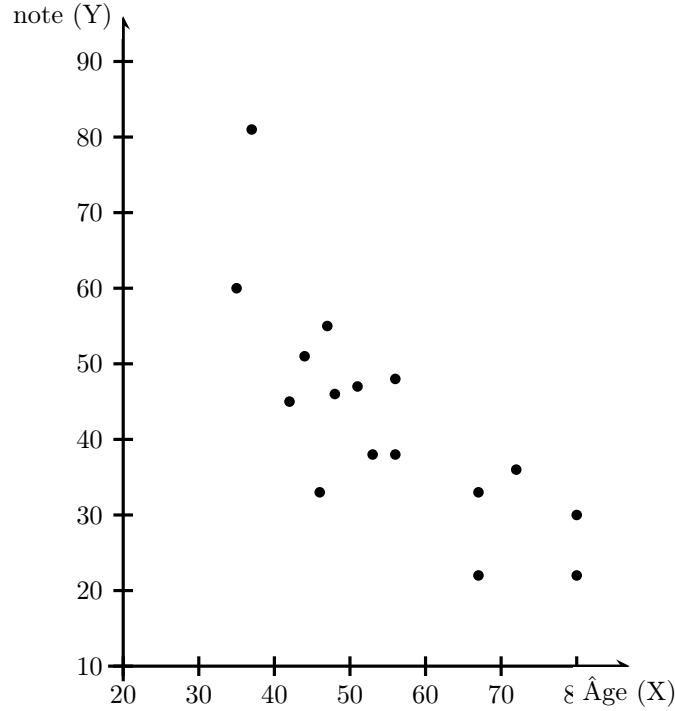


FIGURE 2.1: Nuage de points indiquant la note en fonction de l'âge

### 2.3.1 Coefficient de corrélation linéaire

#### Définition

- On appelle covariance de deux variables quantitatives  $X$  et  $Y$  la quantité  $\text{Cov}(X; Y) = m((X - m(X)) \times (Y - m(Y)))$  qui est aussi égale à  $m(XY) - m(X)m(Y)$ . Cette seconde expression est plus efficace pour les calculs.
- Le coefficient de corrélation linéaire des variables  $X$  et  $Y$  est :  $r(X; Y) = \frac{\text{Cov}(X; Y)}{s(X) \cdot s(Y)}$

**Mode de calcul** Dans ce chapitre les données ne seront pas regroupées par modalité (ni par classe), de sorte que la moyenne de  $X \times Y$  se calcule simplement par la formule :  $m(XY) = \frac{\sum x_i y_i}{n}$ .

**Exemple** Pour les données présentées en début de chapitre, on obtient :

$$\text{moyenne: } m(X) = \frac{\sum x_i}{n} = \frac{80+56+67+\dots+35}{16} = \frac{881}{16} \simeq 55,06$$

$$m(X^2) = \frac{\sum x_i^2}{n} = \frac{80^2+56^2+67^2+\dots+35^2}{16} = \frac{51\,567}{16}$$

$$\text{Var}(X) = m(X^2) - m(X)^2 = \frac{51\,567}{16} - \left(\frac{881}{16}\right)^2 \simeq 191,06$$

$$\text{Écart-type: } s(X) = \sqrt{\text{Var}(X)} \simeq 13,82$$

$$\text{moyenne: } m(Y) = \frac{\sum y_i}{n} = \frac{22+38+33+\dots+60}{16} = \frac{685}{16} \simeq 42,81$$

$$m(Y^2) = \frac{\sum y_i^2}{n} = \frac{22^2+38^2+33^2+\dots+60^2}{16} = \frac{32\,671}{16}$$

$$\text{Var}(Y) = m(Y^2) - m(Y)^2 = \frac{32\,671}{16} - \left(\frac{685}{16}\right)^2 \simeq 209,03$$

$$\text{Écart-type: } s(Y) = \sqrt{\text{Var}(Y)} \simeq 14,46$$

$$m(XY) = \frac{\sum x_i y_i}{n} = \frac{80 \times 22 + 56 \times 38 + \dots + 35 \times 60}{16} = \frac{35\,206}{16} = 2\,200,375$$

$$\text{Cov}(X, Y) = m(XY) - m(X)m(Y) = \frac{35\,206}{16} - \frac{881}{16} \cdot \frac{685}{16} \simeq -156,988$$

$$r(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} = \frac{-156,988}{\sqrt{191,059 \times 209,027}} \simeq -0,786$$

et le coefficient de corrélation linéaire vaut donc environ -0,786.

### Interpretation

- Le coefficient de corrélation linéaire est toujours entre  $-1$  et  $+1$ . Plus il est proche de zéro moins cela traduit de lien entre les variables. Plus il est proche de  $1$  ou  $-1$ , plus cela indique un fort lien linéaire entre les deux variables.
  - Si le coefficient de corrélation linéaire est supérieur à  $0,75$  ou inférieur à  $-0,75$ , cela traduit un fort lien linéaire entre les deux variables, c'est à dire que le nuage de points est presque aligné le long d'une droite.
  - S'il est entre  $0,5$  et  $0,75$ , ou entre  $-0,75$  et  $-0,5$ , cela traduit déjà un lien entre les deux variables, même s'il n'est pas très fort ou bien ne correspond pas très précisément à une droite.
- De plus, lorsqu'il indique un lien entre les variables,
  - si le coefficient de corrélation linéaire est positif, il indique que  $Y$  tend à augmenter quand  $X$  augmente.
  - S'il est négatif au contraire, cela indique que  $Y$  tend à diminuer quand  $X$  augmente.

## 2.3.2 Coefficient de corrélation des rangs de Spearman

### Calcul des rangs

Lorsqu'on mesure une variable quantitative (ou ordinale) sur plusieurs individus, on peut calculer des rangs pour chaque individu.

Considérons par exemple la variable  $X$ , pour les données indiquées en début de ce chapitre :

- Chez le 16<sup>ème</sup> individu,  $X$  prend la valeur 35, qui est la plus petite valeur. On y associe le rang 1, et on pose  $x'_{16} = 1$ .
- Chez le 12<sup>ème</sup> individu,  $X$  prend la valeur 37, qui est la deuxième plus petite valeur. On y associe le rang 2 (pour « deuxième plus petite valeur »), et on pose  $x'_{12} = 2$ .
- De même la troisième plus petite valeur est 42, qui correspond au 8<sup>ème</sup> individu, de sorte que l'on pose  $x'_8 = 3$ .
- On continue en posant  $x'_{15} = 4$ ,  $x'_{14} = 5$ ,  $x'_{10} = 6$ ,  $x'_{13} = 7$ ,  $x'_6 = 8$  et  $x'_{11} = 9$ .
- La valeur suivante est 56 qui correspond aux individus numérotés 2 et 9. On devrait attribuer le rang 10 à l'un et 11 à l'autre, mais pour éviter de choisir arbitrairement auquel des deux attribuer le rang 10, on décide de leur attribuer à tous les deux le rang 10,5. On pose donc  $x'_2 = 10,5$  et  $x'_9 = 10,5$ .
- Se souvenant que les rangs que l'on vient d'attribuer correspondent à 10 et à 11, on reprend au rang 12, qui correspond à la valeur 67. Or deux individus ont cette valeur : le 3<sup>ème</sup> et le 5<sup>ème</sup> individu. On devrait leur attribuer les rangs 12 et 13, mais comme précédemment, on leur attribue le rang 12,5. On pose donc  $x'_3 = 12,5$  et  $x'_5 = 12,5$ .
- On continue ainsi jusqu'à avoir attribué à chaque individu un rang  $x'_i$  qui ordonne les individus en fonction de la valeur de  $X$ . On obtient ainsi la table 2.1.

Individu	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Âge X	80	56	67	72	67	51	80	42	56	47	53	37	48	46	44	35
Rang $x'_i$	15,5	10,5	12,5	14	12,5	8	15,5	3	10,5	6	9	2	7	5	4	1,0

TABLE 2.1: Calcul des rangs pour la variable  $X$

De même, on calcule ensuite les rangs  $y'_i$  qui ordonnent les individus en fonction de la valeur de  $Y$ . Pour cet exemple, quand on les ajoute à la table 2.1 on obtient la table 2.2.

Âge X	80	56	67	72	67	51	80	42	56	47	53	37	48	46	44	35
Note Y	22	38	33	36	22	47	30	45	48	55	38	81	46	33	51	60
Rang $x'_i$	15,5	10,5	12,5	14	12,5	8	15,5	3	10,5	6	9	2	7	5	4	1,0
Rang $y'_i$	1,5	7,5	4,5	6	1,5	11	3	9	12	14	7,5	16	10	4,5	13	15

TABLE 2.2: Rangs des deux variables  $X$  et  $Y$

### Coefficient de corrélation des rangs de Spearman

On pourrait considérer les rangs que l'on vient de calculer comme les valeurs de deux variables aléatoires  $X'$  et  $Y'$ . Dès lors on pourrait calculer leurs moyennes, leurs écart types, leur covariance, et leur coefficient de corrélation linéaire. Le coefficient ainsi obtenu est ce que l'on appelle le « **Coefficient de corrélation des rangs de Spearman** », que l'on note  $r_s(X; Y)$ .

En pratique, une formule plus courte permet de calculer le coefficient de corrélation de rangs de Spearman sans passer par le calcul de moyennes, écart types, etc des variables  $X'$  et  $Y'$  :

$$r_s(X; Y) \approx 1 - \left( 6 \frac{\sum_{i=1}^n (x'_i - y'_i)^2}{n(n^2 - 1)} \right)$$

### Exemple

Dans le cas présent, cette formule donne  $r_s(X; Y) \approx 1 - \left( 6 \frac{(15.5-1.5)^2 + (10.5-7.5)^2 + (12.5-4.5)^2 + \dots + (1.0-15)^2}{16(16^2-1)} \right) \simeq -0,774$ .

### Organisation du calcul

Pour simplifier le calcul, on peut rajouter à la table 2.2 un ligne qui permet le calcul de  $\sum_{i=1}^n (x'_i - y'_i)^2$  : on y reporte pour chaque individu la valeur de  $(x'_i - y'_i)^2$  ainsi que la somme de ces valeurs, qui vaut donc  $\sum_{i=1}^n (x'_i - y'_i)^2$ . C'est ce qui est fait dans la table 2.3

Âge (X)	80	56	67	72	67	51	80	42	56	47	53	37	48	46	44	35
Note (Y)	22	38	33	36	22	47	30	45	48	55	38	81	46	33	51	60
Rang $x'_i$	15,5	10,5	12,5	14	12,5	8	15,5	3	10,5	6	9	2	7	5	4	1,0
Rang $y'_i$	1,5	7,5	4,5	6	1,5	11	3	9	12	14	7,5	16	10	4,5	13	15
$(x'_i - y'_i)^2$	196	9	64	64	121	9	156,25	36	2,25	64	2,25	196	9	0,25	81	196
																Total
																1206

TABLE 2.3: Table résumant les calculs permettant de calculer efficacement  $r_s(X; Y)$

### Interpretation

- Le coefficient de corrélation des rangs de Spearman est toujours entre  $-1$  et  $+1$ . Plus il est proche de zéro moins cela traduit de lien entre les variables. Plus il est proche de  $1$  ou  $-1$ , plus cela indique un fort lien entre les deux variables.
- Si le coefficient de corrélation de Spearman est supérieur à  $0,75$  ou inférieur à  $-0,75$ , cela traduit un fort lien entre les deux variables, c'est à dire que le nuage de points est concentré le long d'une courbe, qui peut être une droite ou bien avoir une autre forme.
- S'il est entre  $0,5$  et  $0,75$ , ou entre  $-0,75$  et  $-0,5$ , cela traduit déjà un lien entre les deux variables, même s'il n'est pas très fort.

- De plus, lorsqu'il indique un lien entre les variables,
  - si le coefficient de corrélation de Spearman est positif, il indique que  $Y$  tend à augmenter quand  $X$  augmente.
  - S'il est négatif au contraire, cela indique que  $Y$  tend à diminuer quand  $X$  augmente.

### Remarque

On pourra noter que contrairement aux autres indicateurs que l'on a calculé, les coefficients de corrélation n'ont pas d'unité. Cela le distingue par exemple de la moyenne et de l'écart type : si l'âge avait été mesuré en mois, la moyenne et l'écart type de l'âge auraient été 12 fois plus grands, alors que les coefficients de corrélation n'auraient pas changé.

## 2.4 Droites de régression

La régression consiste à prédire une variable à partir de la valeur de l'autre variable. Dans ce chapitre on procédera uniquement par **régression linéaire**, c'est à dire que l'on approchera le nuage de points par une droite.

### 2.4.1 Existence de deux droites distinctes

Sur l'exemple du nuage de points de la figure 2.1 page 21, on pourrait se poser les questions suivantes :

- Si une personne a 80 ans, quelle note s'attend-on à ce qu'elle ait ?  
*En regardant le nuage de points, on se dit qu'à 80 ans, la note est entre 20 et 30 environ, soit autour de 25 en moyenne.*
- Si une personne obtient la note 25, quel âge s'attend-on à ce qu'elle ait ? *En regardant, au sein du nuage de points, la portion qui correspond à la note 25, on s'attend à un âge pouvant aller de 35 à 85 ou 90 ans environ. En moyenne on s'attendrait à un âge autour de 60 ans.*

Bien qu'énoncées « à la louche », ces réponses font bien sentir une propriété importante : Si les gens de 80 ans en général une note proche de 25, ça ne signifie par pour autant que les gens ayant la note 25 ait près de 80 ans.

Mathématiquement, cela se traduit par le fait qu'il y ait deux équations légèrement différentes (deux droites distinctes) selon que l'on veuille estimer la note en connaissant l'âge, ou bien qu'on veuille au contraire estimer l'âge en connaissant la note.

### 2.4.2 Équation des deux droites

#### Détermination de $Y$ à partir de $X$

Pour estimer la valeur de  $Y$  chez un individu pour lequel on connaît la valeur de  $X$ , on utilise la droite  $D_{Y|X}$  d'équation :

$$D_{Y|X} : Y = aX + b \quad \text{où} \quad a = \frac{\text{Cov}(X; Y)}{\text{Var}(X)} = r(X; Y) \times \frac{s(Y)}{s(X)}, \quad \text{et} \quad b = m(Y) - a \cdot m(X)$$

#### Exemple

Pour estimer la note d'une personne de 80 ans, on calcule cette droite :

$$\text{on pose } a = \frac{\text{Cov}(X, Y)}{\text{Var}(X)} \simeq \frac{-156,988}{191,06} \simeq -0,822 \text{ et } b = m(Y) - a m(X) \simeq 42,812 - (-0,822) \times 55,062 \simeq 88,073$$

D'où l'équation de la droite  $D_{Y|X} : Y = -0,822 X + 88,073$

Donc pour  $x = 80$ , on s'attend à  $y = -0,822 \times 80 + 88,073 = 22,313$ .



### Détermination de $X$ à partir de $Y$

Pour estimer la valeur de  $X$  chez un individu pour lequel on connaît la valeur de  $Y$ , on utilise la droite  $D_{X|Y}$  d'équation :

$$D_{X|Y} : X = a'Y + b' \quad \text{où} \quad a' = \frac{\text{Cov}(X; Y)}{\text{Var}(Y)} = r(X; Y) \times \frac{s(X)}{s(Y)}, \quad \text{et} \quad b' = m(X) - a' \cdot m(Y)$$

#### Exemple

Pour estimer l'âge d'une personne qui a la note 25, on calcule cette droite :

$$\text{on pose } a' = \frac{\text{Cov}(X, Y)}{\text{Var}(Y)} \simeq \frac{-156,988}{209,03} \simeq -0,751 \text{ et } b' = m(X) - a' m(Y) \simeq 55,062 - (-0,751) \times 42,812 \simeq 87,214$$

D'où l'équation de la droite  $D_{X|Y} : X = -0,751 Y + 87,214$

Donc pour  $y = 25$ , on s'attend à  $x = -0,751 \times 25 + 87,214 = 68,439$ .

### 2.4.3 Remarque

Une telle régression linéaire est d'autant plus pertinente que le coefficient de corrélation linéaire est proche de 1 (ou de -1).



## Chapitre 3

# Introduction aux probabilités

### 3.1 Introduction : loi uniforme

#### 3.1.1 Définitions

**Évènement aléatoire** On appelle “évènement” les différents résultats d’une expérience aléatoire.

**Loi de probabilité** On appelle loi de probabilité une règle de calcul permettant de déterminer la probabilité des différents évènements possibles dans un contexte précis.

**Loi uniforme** On parle de loi uniforme si tous les évènements élémentaires ont la même probabilité.

**Variable aléatoire** Quantité qui varie d’un évènement à l’autre.

#### 3.1.2 Exemple

On dispose de deux dés à 4 faces ; l’un de couleur bleu et l’autre de couleur rouge. On lance ces deux dés et on note par  $X$  la somme des chiffres indiqués par les dés ( $X$  est donc toujours entre 2 et 8).

On liste ci-dessous les résultats possibles :

$1 + 1$  ;  $1 + 2$  ;  $1 + 3$  ;  $1 + 4$   
 $2 + 1$  ;  $2 + 2$  ;  $2 + 3$  ;  $2 + 4$   
 $3 + 1$  ;  $3 + 2$  ;  $3 + 3$  ;  $3 + 4$   
 $4 + 1$  ;  $4 + 2$  ;  $4 + 3$  ;  $4 + 4$

On appelle chacun de ces 16 cas “évènement élémentaire”, et comme ils ont chacun la même probabilité on parle de “loi uniforme”.

Comme  $X$  varie d’un cas à l’autre, c’est une “variable aléatoire”.

Les affirmations “ $X=2$ ”, “ $X=3$ ”, etc sont elles aussi des “évènements” dont on peut calculer la probabilité : par exemple  $\mathbb{P}[X = 3] = \frac{2}{16} = 0,125$ , car deux cas sur 16 (en l’occurrence  $1 + 2$  et  $2 + 1$ ) correspondent à  $X = 3$ . On calcule ainsi les probabilités suivantes :

Évènement	$X=2$	$X=3$	$X=4$	$X=5$	$X=6$	$X=7$	$X=8$
Probabilité	$\frac{1}{16} \simeq 0,062$	$\frac{2}{16} = 0,125$	$\frac{3}{16} \simeq 0,188$	$\frac{4}{16} = 0,25$	$\frac{3}{16} \simeq 0,188$	$\frac{2}{16} = 0,125$	$\frac{1}{16} \simeq 0,062$

TABLE 3.1: Probabilité de chaque valeur de  $X$

Si on se focalise sur la variable  $X$ , alors les évènements élémentaires sont “ $X = 2$ ”, “ $X = 3$ ”, “ $X = 4$ ”, “ $X = 5$ ”, “ $X = 6$ ”, “ $X = 7$ ” et “ $X = 8$ ”, à partir desquels on peut constituer d’autres évènements comme “ $X \leq 3$ ” (constitué de “ $X = 2$ ” et “ $X = 3$ ”), “ $X \geq 5$ ”, etc. Dans ce cadre, la table 3.1 permet de calculer toutes les probabilités, elle donne donc la **loi** de  $X$ . Cette loi n’est pas la loi uniforme car les probabilités  $\mathbb{P}[X = 2]$ ,  $\mathbb{P}[X = 3]$ , etc. ne sont pas toutes égales entre elles.

### 3.1.3 Moyenne et écart type

On a vu que sur un total de 16 cas,  $X$  vaut 2 dans 1 cas, trois dans 2 cas, 4 dans 3 cas, etc.

On peut alors calculer la moyenne, l'écart type, etc :

$$\text{moyenne: } m(X) = \frac{\sum_i x_i n_i}{n} = \frac{2 \times 1 + 3 \times 2 + 4 \times 3 + \dots + 8 \times 1}{16} = \frac{80}{16} = 5$$

$$m(X^2) = \frac{\sum_i x_i^2 n_i}{n} = \frac{2^2 \times 1 + 3^2 \times 2 + 4^2 \times 3 + \dots + 8^2 \times 1}{16} = \frac{440}{16}$$

$$Var(X) = m(X^2) - m(X)^2 = \frac{440}{16} - \left(\frac{80}{16}\right)^2 = 2,5$$

$$\text{Écart-type: } s(X) = \sqrt{Var(X)} \simeq 1,58$$

## 3.2 Énumération des cas

En présence d'une loi uniforme, le calcul de probabilité d'un évènement se réalise en comptant le nombre de cas correspondant à cet évènement, et en divisant leur nombre par le nombre total de cas.

Nous allons donc désormais étudier sur quelques exemples les façons de compter le nombre de cas correspondant à des évènements précis.

### 3.2.1 Permutations

#### Exemple d'expérience

Une psychologue étudie les dessins d'enfants, auxquels on demande de dessiner leur famille. Elle se concentre sur des enfants qui ont un seul frère ou soeur, et cherche à savoir s'ils dessinent leur famille dans un ordre complètement aléatoire ou s'ils ont plutôt tendance à commencer par dessiner leur parents et eux même, avant de dessiner leur frère/soeur en dernier.

Afin d'analyser les données recueillies, il faut déterminer quelle serait la probabilité de terminer par le frère/soeur si l'ordre était complètement aléatoire.

On peut lister les cas, en notant "E" pour l'enfant qui dessine, "F" pour son frère ou sa soeur, "M" pour sa mère et "P" pour son père. On obtient la liste ci dessous :

EFMP	EFPM	FEMP	FEPF	MEFP	MEPF	PEFM	PEMF
EMFP	EMPF	FMEP	FMPE	MFEP	MFPE	PFEM	PFME
EPFM	EPMF	FPEM	FPME	MPEF	MPFE	PMEF	PMFE

Parmi ces 24 cas, seuls six terminent par "F", la probabilité de terminer par le frère/soeur est donc (si chaque ordre a la même probabilité) de  $\frac{6}{24} = 0,25$ .

#### Nombre de cas

Sur cet exemple, on constate que la liste des cas comporte exactement  $4 \times 3 \times 2 = 24$  cas. En effet, il y a 4 choix possibles pour la première lettre, et pour chaque choix de première lettre il y a trois possibilités pour la deuxième lettre, puis chaque choix des deux premières lettres contient deux possibilité pour la troisième lettre, et enfin pour la dernière lettre il ne reste plus qu'un possibilité.

Plus généralement, s'il y a avait eu  $n$  lettres disponibles, le nombres d'ordres possibles aurait été  $n \times (n-1) \times (n-2) \times \dots \times 2 \times 1$  (que l'on peut aussi écrire  $1 \times 2 \times 3 \times \dots \times n$ ).

#### Définition

Étant donné un entier  $n$ , le nombre  $n \times (n-1) \times (n-2) \times \dots \times 2 \times 1$  s'appelle la factorielle de  $n$ , on le note «  $n!$  ». Il est égal au nombre de permutations d'un ensemble contenant  $n$  éléments.

#### Exemples

On a par exemple  $7! = 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1 = 5\,040$ .

De plus on considère que  $0! = 1$  et que  $1! = 1$ .

## Application

Considérons maintenant des enfants qui ont trois frères et soeurs, et considérons l'ordre dans lequel ils dessinent les 6 membres de leur famille. On demande quelle est la probabilité que les trois premiers membres dessinés soient l'enfant lui-même et ses parents (auquel cas les trois derniers dessinés sont les trois frères et soeurs de l'enfant).

- On note tout d'abord, comme la famille compte six membres, le nombre d'ordres possibles est  $6! = 6 \times 5 \times 4 \times 3 \times 2 \times 1 = 720$ . Il serait donc déraisonnable d'en faire la liste comme précédemment.
- Parmi ces 720 ordres possibles, on cherche à compter ceux où les trois premiers sont les parents et l'enfant lui-même, et les trois derniers sont les frères et soeurs. Pour les cas qui nous intéressent il y a donc  $3! = 3 \times 2 \times 1 = 6$  façon d'ordonner les trois premiers éléments, et pour chacun de ces ordres, il y a  $3! = 3 \times 2 \times 1 = 6$  possibilités pour les trois derniers. Ainsi le nombre de cas qui nous intéresse est  $3! \times 3! = 36$ .
- Ainsi cette probabilité correspond à 36 cas sur 720. La probabilité est donc de  $\frac{36}{720} = 0,05$ .

### 3.2.2 Combinaisons

#### Exemple

On revient à l'exemple des enfants qui n'ont qu'un frère/soeur (pour être à nouveau en mesure de lister tous les cas possibles). On se demande désormais quelle est la probabilité que les deux premières personnes dessinées soient les parents de l'enfant (sans se soucier que ce soit la mère qui soit dessinée avant le père ou bien l'inverse).

1. Une première méthode consiste à reprendre la liste de cas considérée en début de section 3.2.1 : parmi les 24 cas, on compte qu'il y en a 4 où les deux premières personnes dessinées soient les parents de l'enfant. La probabilité est donc  $\frac{4}{24} \simeq 0,167$ .
2. Une autre méthode consisterait à changer de point de vue sur les « cas » qu'il faut lister. On pourrait considérer qu'un « cas » est donnée par l'ensemble des deux premières personnes dessinées (sans se soucier de l'ordre). Si l'on procède ainsi, on obtient la liste ci-dessous :

EF	EM	EP	FM	FP	MP
----	----	----	----	----	----

Cette liste compte 6 cas, parmi lesquels un seul (le dernier) correspond à avoir commencé par les deux parents. On déduit que la probabilité recherchée est  $\frac{1}{6} \simeq 0,167$ .

On notera bien que dans la deuxième méthode on n'a pas ajouté par exemple « PM », car on considère que c'est le même cas que « MP ».

On peut constater que chacun des cas listés pour la deuxième méthode correspond à 4 cas distincts dans la première méthode, mais que la liste dressée dans la seconde méthode permet de répondre plus simplement à la question posée (à savoir la probabilité que les deux premières personnes dessinées soient les parents de l'enfant).

#### Nombre de cas

Lorsqu'on choisit parmi  $n$  éléments un groupe de  $k$  éléments, le nombre de possibilités est le nombre noté  $\binom{n}{k}$  qui vaut :

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

(en supposant que  $0 \leq k \leq n$ ).

**Remarque :** la notation  $\binom{n}{k}$  se lit «  $k$  parmi  $n$  ». Certains auteurs la notent aussi  $C_n^k$ .

**Origine de la formule :** Quand on compte l'ensemble des permutations de  $n$  éléments, il y en a plusieurs qui correspondent au même  $k$  premiers éléments. Il y en a précisément  $k!(n-k)!$ , car il y a  $k!$  façon

d'ordonner les  $k$  premiers éléments, et pour chacune d'entre elle, il y a  $(n - k)!$  façons d'ordonner les  $(n - k)$  derniers éléments. En conséquence, pour compenser le fait que chaque combinaison correspond à plusieurs permutations, on divise le nombre de permutations (ce qui vaut  $n!$ ) par le nombre de fois qu'apparaît chaque combinaison (c'est à dire  $k!(n - k)!$ ).

### Exemples

$$\begin{aligned} \bullet \quad \binom{4}{2} &= \frac{4!}{2! \times 2!} = \frac{4 \times 3 \times 2}{(2 \times 1) \times (2 \times 1)} = (4 \times 3) \times \frac{1}{2 \times 1} \\ &= \frac{2 \times 2 \times 3}{2} = 2 \times 3 = 6 \end{aligned}$$

c'est pourquoi on a obtenu 6 cas avec la deuxième méthode dans l'exemple en début de section 3.2.2.

$$\begin{aligned} \bullet \quad \binom{6}{3} &= \frac{6!}{3! \times 3!} = \frac{6 \times 5 \times \dots \times 1}{(3 \times 2 \times 1) \times (3 \times 2 \times 1)} = (6 \times 5 \times 4) \times \frac{1}{3 \times 2 \times 1} \\ &= \frac{3 \times 2 \times 5 \times 4}{3 \times 2} = 5 \times 4 = 20 \end{aligned}$$

c'est pourquoi on a obtenu une probabilité  $\frac{1}{20} = 0,05$  dans l'exemple d'application à la fin de la section 3.2.1.

### Application

On considère un groupe de 25 personnes parmi lesquelles 6 fument quotidiennement. Si on choisit au hasard 3 personnes parmi ce groupe de 25, quelle est la probabilité de choisir un fumeur et deux non-fumeurs ?

- On commence par compter le nombre de cas possibles : il y a  $\binom{25}{3}$  choix possible de 3 personnes au sein de l'échantillon. La calculatrice nous indique que  $\binom{25}{3} = 2300$  donc il y a 2300 cas.
- On compte ensuite le nombre de cas où on a un fumeur et deux non-fumeurs, sachant que le groupe de 25 personnes comptait 6 fumeurs et 19 non-fumeurs. Il y a donc  $\binom{19}{2}$  choix possibles pour choisir deux non-fumeurs, et pour chacun de ces choix, il y a 6 choix d'un fumeur (pour compléter et obtenir trois personnes). Ainsi le nombre de cas qui nous intéresse est  $6 \times \binom{19}{2}$ . Une calculatrice nous indique que cela vaut 1026.
- On conclue que la probabilité recherchée vaut  $\frac{1026}{2300} \simeq 0,446$ .

## 3.3 Loi binomiale : répétition d'une expérience

### 3.3.1 Exemple

On considère une personne atteinte de troubles de la mémoire. Chaque jour, il y a 40% de chances qu'il y ait au moins une fois, au cours de la journée, où elle éprouve une gêne à cause de ses troubles de la mémoire.

Pendant trois jours consécutifs, on lui demande chaque soir si elle a éprouvé une telle gêne au cours de la journée. On note  $X$  le nombre de jours (parmi les trois) où elle a éprouvé une gêne et on cherche à déterminer la loi de  $X$ .

- On commence par lister tous les cas possibles, et leur probabilité :

Cas	AAA	AAG	AGA	AGG	GAA	GAG	GGA	GGG
Probabilité	$0,6 \times 0,6 \times 0,6$	$0,6 \times 0,6 \times 0,4$	$0,6 \times 0,4 \times 0,6$	$0,6 \times 0,4 \times 0,4$	$0,4 \times 0,6 \times 0,6$	$0,4 \times 0,6 \times 0,4$	$0,4 \times 0,4 \times 0,6$	$0,4 \times 0,4 \times 0,4$

Dans cette liste de cas, on a noté « G » pour les jours où elle ressentait une gêne, et « A » pour les jours se déroulant en l'absence de gêne.

On a de plus calculé la probabilité de chaque cas : par exemple pour être dans le cas « AGA » il faut à la fois n'avoir ressenti aucun gêne le premier jour (celà a 60% de chances d'arriver) puis en avoir ressenti le second jour (celà a 40% de chances d'arriver) et enfin ne pas en ressentir le troisième jour (celà a 60% de chances d'arriver). Ainsi l'évènement noté « AGA » a la probabilité  $0,6 \times 0,4 \times 0,6$ .

- On calcule  $\mathbb{P}[X = 0]$  : comme l'évènement «  $X = 0$  » correspond précisément au cas « AAA », sa probabilité est  $0,6 \times 0,6 \times 0,6 = 0,216$ .

- On calcule  $\mathbb{P}[X = 1]$  : comme l'évènement «  $X = 1$  » correspond aux trois cas « AAG », « AGA » et « GAA », sa probabilité est  $0,6 \times 0,6 \times 0,4 + 0,6 \times 0,4 \times 0,6 + 0,4 \times 0,6 \times 0,6 \simeq 0,432$ .  
On pourra noter qu'en fait ces trois évènements ont la même probabilité  $0,6^2 \times 0,4 = 0,144$ .
- De même on calcule  $\mathbb{P}[X = 2]$  : l'évènement «  $X = 2$  » correspond aux trois cas « AGG », « GAG » et « GGA », qui ont tous trois la probabilité  $0,4^2 \times 0,6 \simeq 0,096$ . En conséquence  $\mathbb{P}[X = 2] = 3 \times 0,4^2 \times 0,6 = 0,288$ .
- Enfin, l'évènement «  $X = 3$  » correspond uniquement au cas « GGG ». Sa probabilité est donc  $\mathbb{P}[X = 3] = 0,4^3 \simeq 0,064$ .

En conséquence on obtient que la loi de  $X$  est :

Évènement	$X = 0$	$X = 1$	$X = 2$	$X = 3$
Probabilité	0,216	0,432	0,288	0,064

### 3.3.2 Formule générale et définition

Le calcul effectué ici sur un exemple peut se généraliser à toute situations où l'on répète plusieurs fois une expérience qui a à chaque fois la même probabilité de succès (indépendamment des résultats des précédentes répétitions de l'expérience).

Dans ce cadre, si l'on compte juste le nombre  $X$  de succès obtenus en répétant  $n$  fois l'expérience, alors pour chaque valeur de  $k$  on trouve

$$\mathbb{P}[X = k] = \binom{n}{k} p^k (1 - p)^{n-k},$$

où  $p$  désigne la probabilité de succès à chaque répétition de l'expérience.

#### Définition

Si la loi d'une variable aléatoire  $X$  est donnée par

$$\mathbb{P}[X = k] = \binom{n}{k} p^k (1 - p)^{n-k},$$

alors on dit que  $X$  suit la loi binomiale de paramètres  $n$  et  $p$ .

#### Notation

L'écriture  $X \sim \mathcal{B}(n; p)$  signifie que «  $X$  suit la loi binomiale de paramètres  $n$  et  $p$  ».

#### Propriétés

Si  $X \sim \mathcal{B}(n; p)$ , alors

- la moyenne de  $X$  est  $m(X) = np$ .
- sa variance est  $\text{Var}(X) = np(1 - p)$ .
- son écart type est  $s(X) = \sqrt{np(1 - p)}$ .

#### Procédure de calcul

On sera souvent amené à calculer la probabilité d'intervalles. Pour notre exemple, on peut par exemple calculer  $\mathbb{P}[1 \leq X \leq 2]$ .

### 1<sup>ère</sup> méthode : énumération des cas

On additionne les probabilités de toutes les valeurs de l'intervalle.

Dans le cas présent on obtient

$$\begin{aligned}\mathbb{P}[1 \leq X \leq 2] &= \mathbb{P}[X = 1] + \mathbb{P}[X = 2] \\ &= \binom{3}{1} (0.4)^1 (1 - 0.4)^{3-1} + \binom{3}{2} (0.4)^2 (1 - 0.4)^{3-2} \\ &\simeq 0.432 + 0.288 \simeq 0.72\end{aligned}$$

### 2<sup>ème</sup> méthode : utilisation de la calculatrice

La calculatrice sait calculer les probabilités de la forme  $\mathbb{P}[X \leq \dots]$ . On se ramène donc à une soustraction de probabilités de ce type :

$$\mathbb{P}[1 \leq X \leq 2] = \mathbb{P}[X \leq 2] - \mathbb{P}[X \leq 0] = 0.936 - 0.216 \simeq 0.72$$

Avec cette méthode il faut prendre garde que pour des entiers  $a$  et  $b$  arbitraires, on a  $\mathbb{P}[a \leq X \leq b] = \mathbb{P}[X \leq b] - \mathbb{P}[X \leq a - 1]$ . En effet, pour compter toutes les valeurs de  $a$  à  $b$ , on compte toutes les valeurs jusqu'à  $b$  et on soustrait celles qui sont strictement plus petites que  $a$  (de sorte qu'il reste celle qui sont entre  $a$  et  $b$ , y compris la valeur  $a$ ).

Quand on « soustrait celles qui sont strictement plus petites que  $a$  », on soustrait  $\mathbb{P}[X \leq a - 1]$  (car les valeurs strictement plus petites que  $a$  sont toutes les valeurs jusqu'à  $a - 1$ ).

## 3.3.3 Échantillonnage

Dans les chapitres suivants, on utilisera la loi binomiale pour décrire le choix aléatoire d'un échantillon.

Par exemple, si on suppose qu'il y a en France environ 24% de fumeurs, et que l'on choisit trois français(e)s au hasard, on peut noter  $X$  le nombre de fumeurs au sein de l'échantillon. Alors  $X \sim \mathcal{B}(3; 0,24)$ , car on a répété trois fois une expérience (choisir une personne au hasard et constater si elle fume ou pas) qui avait à chaque fois 24% de chances de « succès ».

En conséquence on a les probabilités suivantes :

Évènement	$X = 0$	$X = 1$	$X = 2$	$X = 3$
Probabilité	0,439	0,416	0,131	0,014

On peut constater que  $\mathbb{P}[X = 1] = 0,416$  diffère sensiblement de la probabilité  $(\frac{1026}{2300} \simeq 0,446)$  obtenue dans l'exemple page 30 à la fin de la section 3.2.2. Pourtant, on choisissait bien trois personnes au hasard parmi un échantillon où la proportion de fumeurs était  $\frac{6}{25} = 0,24$ .

La raison pour laquelle les probabilités diffèrent est que dans l'exemple page 30 on choisissait parmi 25 individus, de sorte qu'une fois le premier individu choisi, il n'y avait plus que 24 possibilités pour le second, et la probabilité de choisir un fumeur n'est plus la même que lors du choix du premier individu. En revanche si on choisit trois individus parmi l'ensemble des Français (donc parmi des dizaines de millions d'individus), le choix du premier individu n'a pas d'impact significatif sur la probabilité que le deuxième soit fumeur.

## 3.3.4 Tirage avec ou sans "remise"

Lorsque l'on étudie le choix d'un échantillon au hasard, il arrive d'utiliser les termes « avec » ou « sans remise » pour préciser le mode de tirage. On parle de tirage « avec remise » si on autorise le même individu à être choisi plusieurs fois (donc à compter plusieurs fois dans l'échantillon). En toute rigueur, la loi binomiale décrit les tirages « avec remise », mais comme on vient de le dire, elle est en fait aussi très proche de la vérité lorsque l'échantillon est choisi dans une grande population (au moins dix fois plus grande que la taille de l'échantillon). C'est pourquoi on utilisera beaucoup cette loi dans les chapitres suivants.



# Chapitre 4

## La loi normale

### 4.1 Introduction

Dans le chapitre précédent, les probabilités rencontrées se ramenaient à lister tous les cas possibles, leur attribuer la même probabilité, et diviser le nombre de cas favorables par le nombre total de cas. Bien sûr, cette méthode était en général trop longue, et on a montré comment, dans le cas des modèles de tirages, calculer les probabilités sans avoir à lister tous les cas.

Dans ce chapitre, on étudiera des probabilités apparaissant dans un cadre conceptuellement légèrement différent dans lequel les probabilités s'expriment comme l'aire sous une courbe, c'est-à-dire l'intégrale d'une fonction. Dans ce chapitre la courbe en question sera une gaussienne (une « courbe en cloche ») qui définit la « loi normale ». Cette loi est très importante en statistiques, et fournit notamment une bonne approximation de la loi binomiale vue au chapitre précédent.

### Motivation à partir de la loi binomiale

Commençons tout d'abord par un cas traité dans le chapitre précédent : on lance 20 fois une pièce non biaisée et on note  $X$  le nombre de fois qu'on a obtenu pile. On sait alors que  $X$  suit la loi binomiale  $\mathcal{B}(20; 0,5)$ . On peut calculer les probabilités  $\mathbb{P}[X = 0]$ ,  $\mathbb{P}[X = 1]$ , etc. On peut aussi les représenter par un *diagramme en bâtons* : les hauteurs des bâtons sont les probabilités  $\mathbb{P}[X = 0]$ ,  $\mathbb{P}[X = 1]$ , etc. On obtient ainsi la figure 4.1 ci-après.

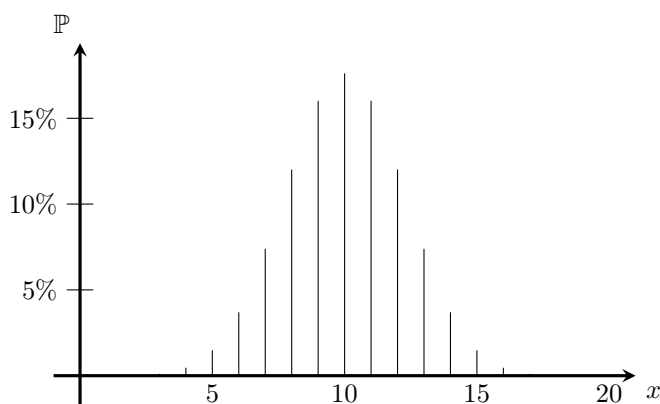
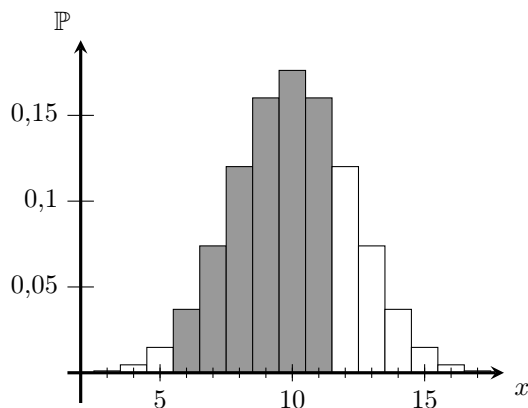


FIGURE 4.1: Diagramme en bâtons pour la loi  $\mathcal{B}(20; 0,5)$ .

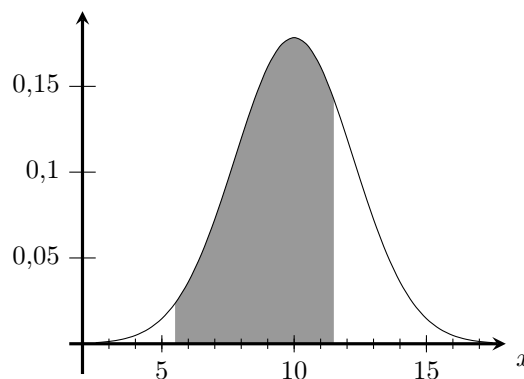
Pour interpréter comme des surface les calculs de probabilités, une idée importante est de remplacer artificiellement les bâtons du diagramme par des rectangles qui se touchent : par exemple à la place du bâton

en «  $x = 7$  » on met un rectangle de  $x = 6,5$  à  $x = 7,5$  (la valeur 7 se trouve ainsi au milieu du rectangle), qui touche le rectangle suivant (rectangle de  $x = 7,5$  à  $x = 8,5$ , qui se substitue au bâton en  $x = 8$ ) et les précédent. Ainsi, chaque probabilité (qui était la hauteur du bâton) devient la surface du rectangle.

On peut alors calculer par exemple la probabilité  $\mathbb{P}[6 \leq X < 12] = \mathbb{P}[X = 6] + \mathbb{P}[X = 7] + \dots + \mathbb{P}[X = 11]$ , c'est-à-dire la somme des aires des rectangles grisés sur la figure 4.2a ci-après. L'idée importante de ce chapitre est d'approcher cette surface par l'aire sous une courbe en forme de cloche, comme en figure 4.2b ci-après.



(a)  $\mathbb{P}[6 \leq X < 12]$  pour  $X \sim \mathcal{B}(20; 0,5)$ .



(b) Aire approchant  $\mathbb{P}[6 \leq X < 12]$ .

FIGURE 4.2: Approximation : loi binomiale et loi normale.

Dans la suite de ce chapitre, on va définir la « loi normale » qui correspond à la surface sous une courbe en cloche, puis indiquer qu'elle permet d'approcher efficacement la loi binomiale.

## 4.2 Loi normale

Loi normale centrée réduite

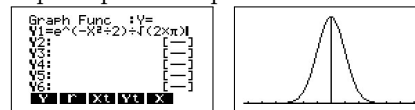
Si la loi d'une variable aléatoire  $Z$  est donnée par

$$\mathbb{P}[a \leq Z \leq b] = \int_a^b \frac{\exp(-x^2/2)}{\sqrt{2\pi}} dx,$$

(cette égalité devant être vraie pour toutes les valeurs de  $a$  et  $b$ ), alors on dit que «  $Z$  suit la loi normale centrée réduite ».

### Remarques

- La formule  $\frac{\exp(-x^2/2)}{\sqrt{2\pi}}$  est la notation mathématique d'une certaine « courbe en cloche ». Dans ce cours, ça n'a aucune importance de connaître et comprendre cette formule, mais cette formule peut par exemple vous servir à tracer cette courbe, comme ci-contre :



- La notation  $\int_a^b$  désigne la surface sous la courbe entre les valeurs  $x = a$  et  $x = b$ . Par exemple en figure 4.2b, la surface en gris correspondrait à  $\int_{5,5}^{11,5} f(x)$  (où  $f$  serait la formule d'une courbe en cloche), car la zone grisée va de  $x = 5,5$  à  $x = 11,5$ .

Encore une fois il n'est pas important dans ce cours de bien comprendre et utiliser cette notation, mais il pourra être utile de faire des dessins (comme ci-dessous) pour guider un raisonnement.

Fonction de répartition

On désigne par  $F(z)$  la probabilité  $\mathbb{P}[Z \leq z]$  (où  $Z$  suit la loi normale centrée réduite), c'est à dire la surface sous la courbe en cloche à gauche de la valeur  $b$ . Comme on le voit sur

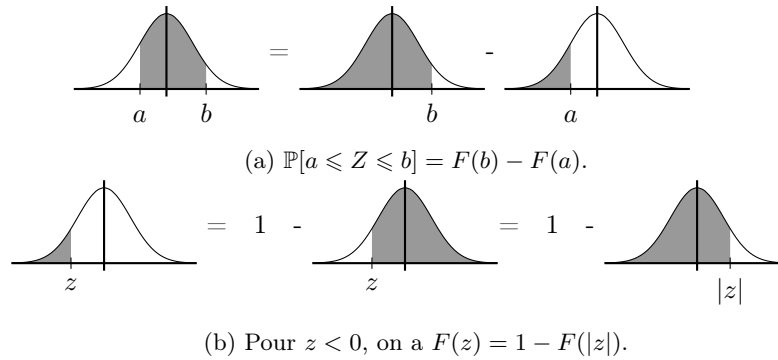


FIGURE 4.3: Représentation graphique des principales formules

la figure 4.3, on a

- $\mathbb{P}[a \leq Z \leq b] = F(b) - F(a)$  (lorsque  $a < b$ ).
- Si  $z < 0$ , alors  $F(z) = 1 - F(|z|)$   
 Dans cette formule  $|z|$  désigne la *valeur absolue* de  $z$ , c'est à dire le nombre  $z$  auquel on a enlevé le signe *moins* s'il était négatif. Par exemple  $|2,7|$  désigne 2,7 et  $|-2,7|$  désigne aussi 2,7. Donc cette formule indique par exemple que  $F(-2,7) = 1 - F(2,7)$ .

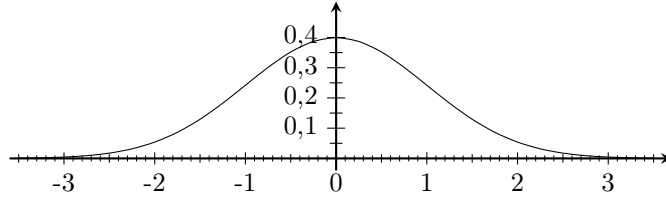
#### Remarques

- Dans ce cours, les lettres  $X$ ,  $Y$ , etc, désignent des variables aléatoires, ou des variables statistiques qui dépendent du contexte. En revanche  $F$  désigne toujours la fonction de répartition de la loi normale centrée réduite. Et en général on tachera de l'utiliser la lettre  $Z$  que pour des variables qui suivent la loi normale centrée réduite.
- Pour la loi normale, les probabilités sont les mêmes pour des inégalités larges ou strictes, c'est à dire que les quatre probabilités  $\mathbb{P}[a \leq Z \leq b]$ ,  $\mathbb{P}[a \leq Z < b]$ ,  $\mathbb{P}[a < Z \leq b]$ , et  $\mathbb{P}[a < Z < b]$  sont égales. Cette propriété importante correspond au fait qu'une loi normale prend des valeurs avec énormément de chiffres après la virgule, qui n'ont aucune chance d'être exactement égales à  $a$  ou  $b$ . Cela constitue une importante différence avec la loi binomiale (pour laquelle il faut bien distinguer entre inégalités strictes et larges).
- la table qui se trouve page 10 du formulaire indique des valeurs de  $F$ , afin de calculer diverses probabilités. Quand  $z$  est positif et a au maximum deux chiffres après la virgule, on trouve  $F(z)$  dans la colonne constituée par le deuxième chiffre après la virgule (dans le nombre  $z$ ), et les chiffres précédents donnent la ligne à considérer.

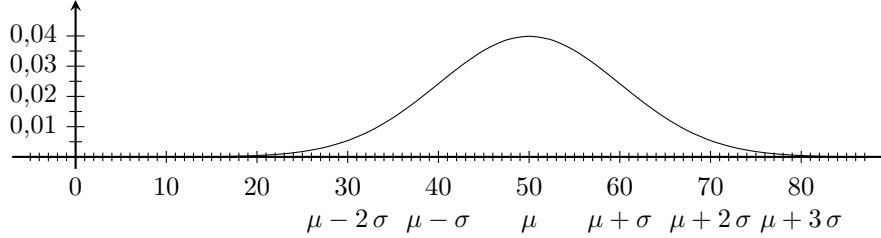
#### Exemples

- On lit  $F(1,73) \simeq 0,9582$  à la ligne 1,7 ; colonne 3.
- On sait que  $F(-2) = 1 - F(2)$ , d'où, en lisant la table,  $F(-2) \simeq 1 - 0,9772 \simeq 0,0228$ .
- On déduit que si  $Z \sim \mathcal{N}(0,1)$ , alors  $\mathbb{P}[-2 \leq Z \leq 1,73] = F(1,73) - F(-2) \simeq 0,9582 - 0,0228 \simeq 0,9354$ .
- Si  $z \geq 4$ , alors  $F(z)$  n'est pas indiqué dans la table, mais on a  $F(z) \simeq 1,0000$ .

**Loi normale** Si  $X$  est une variable aléatoire et  $\mu$  et  $\sigma$  sont deux nombres tels que la variable aléatoire  $Z = \frac{X-\mu}{\sigma}$  suive la loi normale centrée réduite, on dit alors que «  $X$  suit la loi normale de moyenne  $\mu$  et d'écart type  $\sigma$  », et on note  $X \sim \mathcal{N}(\mu; \sigma)$ .



(a) Densité de la loi  $\mathcal{N}(0; 1)$ .



(b) Densité de la loi  $\mathcal{N}(\sigma; \mu)$  (la figure est réalisée pour  $\mu = 50$ ,  $\sigma = 10$ ).

FIGURE 4.4: Comparaison entre les densités de  $\mathcal{N}(0; 1)$  et  $\mathcal{N}(\mu; \sigma)$ .

- Remarques**
- En particulier, la loi normale centrée réduite se note  $\mathcal{N}(0; 1)$ . En effet,  $\mathcal{N}(0; 1)$  signifie qu'on pose  $Z = \frac{X-0}{1}$ , c'est à dire  $Z = X$ , et que  $Z$  doit suivre la loi normale centrée réduite.
  - Cette loi normale (avec une moyenne  $\mu$  et un écart type  $\sigma$ ) s'exprime aussi comme l'aire sous une courbe. Cette fois-ci la courbe (que l'on appelle « densité ») est donnée par la formule  $\frac{\exp(-\frac{(x-\mu)^2}{2\sigma^2})}{\sqrt{2\pi\sigma^2}}$ . La figure 4.4 montre la différence entre cette densité et celle de la loi normale centrée réduite : le sommet de la cloche est décalé (en  $x = \mu$  au lieu de  $x = 0$ ) et la largeur de la cloche est modifiée (en fonction de  $\sigma$ ).

## 4.3 Exemples de calculs

### 4.3.1 Probabilité d'un intervalle

On considère une variable  $X$  qui suit la loi  $\mathcal{N}(20; 5)$ . On cherche à déterminer la probabilité  $\mathbb{P}[19,5 \leq X \leq 26,6]$ .

- On commence par poser  $Z = \frac{X-20}{5}$ , qui suit la loi  $\mathcal{N}(0; 1)$  (car  $X \sim \mathcal{N}(20; 5)$ ). On remarque que la condition  $19,5 \leq X \leq 26,6$  revient rigoureusement au même que  $19,5 - 20 \leq X - 20 \leq 26,6 - 20$ , et que ça revient aussi au même que la condition  $\frac{19,5-20}{5} \leq \frac{X-20}{5} \leq \frac{26,6-20}{5}$ . Donc

$$\mathbb{P}[19,5 \leq X \leq 26,6] = \mathbb{P}\left[\underbrace{\frac{19,5 - 20}{5}}_{-0,1} \leq \underbrace{\frac{X - 20}{5}}_Z \leq \underbrace{\frac{26,6 - 20}{5}}_{1,32}\right] = \mathbb{P}[-0,1 \leq Z \leq 1,32]$$

- Comme  $Z$  suit la loi normale centrée réduite, on peut utiliser la table du formulaire :

$$\mathbb{P}[-0,1 \leq Z \leq 1,32] = \underbrace{F(1,32)}_{\simeq 0,9066} - \underbrace{F(-0,1)}_{1 - F(0,1)} \simeq 0,9066 - (1 - \underbrace{F(0,1)}_{0,5398}) \simeq 0,4464$$

- On a donc ainsi obtenu que  $\mathbb{P}[19,5 \leq X \leq 26,6] \simeq 0,4464$

### Résultat direct avec une calculatrice Casio

Dans **MENU**, choisir **STAT**, puis dans **DIST**, choisir **NORM** puis **Ncd**. Dans le menu qui s'ouvre, entrer 19.5 dans la ligne **Lower** et 26.6 dans la ligne **Upper** (car on veut calculer  $\mathbb{P}[19,5 \leq X \leq 26,6]$ ). Entrer enfin 5 dans la ligne  $\sigma$  et 20 dans la ligne  $\mu$  (car on considère  $X \sim \mathcal{N}(20; 5)$ ). Entrer **CALC** (dans la ligne **Execute**) pour calculer et afficher la probabilité.

D.C.normale	
Data	:Variable
Lower	:19.5
Upper	:26.6
$\sigma$	:5
$\mu$	:20

### Résultat direct avec une calculatrice TI

On utilise la fonction **normalFRep** (ou **normalcdf** sur les calculatrices anglophones). On trouve cette fonction dans le menu **DISTR** accessible par **2nd** **VARS**. On entre ensuite les valeurs 20, 5, 19.5, et 26.6 (séparées par des virgules), pour indiquer qu'on calcule  $\mathbb{P}[19,5 \leq X \leq 26,6]$  et que  $X \sim \mathcal{N}(20; 5)$ . On finit par **)** **ENTER**. La valeur qui s'affiche alors est la probabilité  $\mathbb{P}[19,5 \leq X \leq 26,6]$ .

```
normalFRep(19.5,26.6,20,5)
0.4464103283
```

**Remarque** La plupart des calculatrices calculent la probabilité d'un intervalle dont se seraient indiqués le minimum et le maximum. Lorsque l'on doit calculer par exemple  $\mathbb{P}[X \geq 19,5]$  ou  $\mathbb{P}[X \leq 26,6]$ , on ne dispose pas à la fois d'un maximum et d'un minimum. Dans ce cas, on introduit artificiellement un maximum très grand ou un minimum très petit (« très négatif »). Par exemple, pour calculer  $\mathbb{P}[X \geq 19,5]$  on pourra à la place calculer  $\mathbb{P}[19,5 \leq X \leq 999999999999999]$ . De même, pour déterminer  $\mathbb{P}[X \leq 26,6]$  on pourra à la place calculer  $\mathbb{P}[-999999999999999 \leq X \leq 26,6]$ .

## 4.3.2 Lecture inverse de la table

Il arrive de se poser la question inverse, trouver des seuils donnant une probabilité fixée.

Quelques exemples sont donnés ci-dessous, où on prendra bien garde que  $x$  et  $z$  désignent des nombres tandis que  $X$  et  $Z$  désignent des variables aléatoires.

- Trouvez  $z$  tel que  $F(z) = 0,67$ .
  - On cherche dans la table page 3 du formulaire s'il y a une des valeurs de  $F(z)$  qui vaut 0,67. On constate que c'est le cas dans la ligne 0,4 à la colonne 4, c'est à dire que  $F(0,44) \simeq 0,67$ .
  - Donc la réponse est  $z \simeq 0,44$ .
- Trouvez  $z$  tel que  $F(z) = 0,05$ .
  - En cherchant dans la table du formulaire, on ne risque pas de trouver la valeur 0,05, car les valeurs qui apparaissent dans la table sont toutes supérieures ou égales à 0,5. Il en est ainsi parce que cette table indique uniquement les valeurs de  $F(z)$  lorsque  $z \geq 0$  (et dans ce cas là,  $F(z) \geq 0,5$ ). Pour obtenir  $F(z) = 0,05$ , il va donc falloir considérer que  $z < 0$ .
  - Maintenant qu'on sait que  $z < 0$ , on sait que  $F(z) = 1 - F(|z|)$ . On cherche donc à avoir  $1 - F(|z|) = 0,05$ , c'est à dire  $F(|z|) = 0,95$ .
  - On cherche donc dans la table page 3 du formulaire s'il y a une des valeurs qui vaut 0,95. On n'en trouve et ce qu'on trouve de plus proche est  $F(1,64) = 0,9495$  et  $F(1,65) = 0,9505$ . Donc  $|z|$  se situe entre 1,64 et 1,65.  
**Remarque** : La table en bas de la page 10 donne de manière plus précise la valeur  $|z| \simeq 1,645$  lorsque  $F(|z|) = 0,95$ .
  - En se souvenant que  $z$  était négatif, on conclue que  $z \simeq -1,645$ .  
La réponse est donc :  $z \simeq -1,645$ .
- Soit  $Z \sim \mathcal{N}(0; 1)$ . Trouvez  $z$  tel que  $\mathbb{P}[Z \geq z] = 95\%$ .
  - Comme  $Z \sim \mathcal{N}(0; 1)$ , on sait que  $\mathbb{P}[Z \geq z] = 1 - F(z)$ . On cherche donc à avoir  $1 - F(z) = 0,95$ , donc  $F(z) = 0,05$ . On vient de trouver (exemple précédent) qu'on a alors  $z = -1,645$ .

- La réponse est donc à nouveau :  $z \simeq -1,645$ .
- Soit  $X \sim \mathcal{N}(20; 5)$ . Trouvez  $x$  tel que  $\mathbb{P}[X \geq x] = 95\%$ .
- La condition  $\mathbb{P}[X \geq x] = 95\%$  revient à demander que  $\mathbb{P}\left[\underbrace{\frac{X-20}{5}}_Z \geq \frac{x-20}{5}\right] = 95\%$ .

Or on vient de calculer que pour avoir  $\mathbb{P}[Z \geq z] = 95\%$  il faut que  $z \simeq -1,645$ . Donc ici, il faut avoir  $\frac{x-20}{5} \simeq -1,645$ . On cherche donc à résoudre cette équation :

$\frac{x-20}{5} \simeq -1,645$  revient au même que  $\frac{x-20}{5} \times 5 \simeq -1,645 \times 5$  c'est à dire  $x - 20 \simeq -1,645 \times 5$

Mais  $x - 20 \simeq -1,645 \times 5$  revient au même que  $x - 20 + 20 \simeq -1,645 \times 5 + 20$  c'est à dire  $x \simeq -1,645 \times 5 + 20$

On obtient ainsi que  $x \simeq 11,78$

**Remarque** : pour la fin du calcul, vous avez probablement déjà vu des méthodes plus rapides pour la résolution de l'équation, où on « passe un terme de l'autre côté de l'équation ». Cela revient au même et les étapes données ci-dessus servent simplement à justifier que quand on « passe le 5 de l'autre côté de l'équation », la division par 5 devient une multiplication par 5, et que quand on « passe le 20 de l'autre côté de l'équation », la soustraction de 20 devient une addition de 20.

**Remarque** La calculatrice sait aussi « trouver  $x$  tel que  $\mathbb{P}[X \geq x] = 95\%$  » (y compris avec d'autres valeurs que 95%). Il faut utiliser pour cela la fonction `InvN` de la casio (dans le même menu que `Ncd`) ou la fonction `invNorm` (ou `NormalInverse` selon le modèle) de la TI (dans le même menu que `normalFRep` ou `normalcdf`). Ces fonctions varient un peu d'un modèle de calculatrice à l'autre, n'hésitez pas à regarder le formulaire de ce cours et/ou la documentation de votre calculatrice.

## 4.4 Effectifs théoriques

**Effectifs théorique** Si on considère une population de  $n$  individus, on appelle effectif théorique d'un événement sa probabilité multipliée par  $n$ .

Dans le cas d'une loi normale, il faut toutefois traiter différemment la première et la dernière classe, comme dans l'exemple suivant qui correspond à la loi de l'exercice 37 : on suppose que  $X \sim \mathcal{N}(10,42; 2,18)$ , et si par exemple  $n = 1000$  on obtient le tableau ci-dessous :

Désirabilité	[5; 7[	[7; 9[	[9; 11[	[11; 13[	[13; 15[	[15; 17[
Effectif théorique	58,3	199,1	347,5	276,8	100,4	17,9

où l'on a calculé les effectifs théoriques des différents intervalles. Dans cette situation on *adapte* le premier et le dernier intervalles pour s'assurer que la somme des effectifs soient bien égale à  $n$  :

- Pour la classe  $[5; 7[$ , on calcule en fait  $\mathbb{P}[X \leq 7]$ . Les événements où  $X < 5$  sont ainsi ajoutés artificiellement à cette colonne afin qu'ils apparaissent dans une des colonnes pour que la somme des probabilités fasse bien 100%. On obtient  $\mathbb{P}[X \leq 7] \simeq 0,1$ , d'où l'effectif théorique  $0,1 \times 1000 = 100$ .
- Pour la dernière classe on fait de même en ajoutant les événements où  $X \geq 17$  qui se seraient sinon dans aucune colonne. On calcule donc  $\mathbb{P}[X \geq 15] \simeq 0,0$ , d'où l'effectif théorique  $0,0 \times 1000 = 0$ .
- Pour les autres classes on garde l'intervalle écrit dans la première ligne du tableau. Par exemple, pour la 2ème classe, on calcule donc  $\mathbb{P}[X \geq 7] \simeq 0,2$ , d'où l'effectif théorique  $0,2 \times 1000 = 200$ .

## 4.5 Approximation de loi binomiale

Au début du chapitre, on a introduit la loi normale comme une approximation de la loi binomiale. Pour être plus précis l'affirmation est la suivante :

Si  $n \geq 30$ ,  $np > 5$  et  $n(1-p) > 5$ , alors  $\mathcal{B}(n; p) \approx \mathcal{N}(np; \sqrt{np(1-p)})$

**Remarque** Pour améliorer la précision de cette approximation, on pourra faire une « *correction de continuité* », c'est à dire que l'on prend en compte le passage d'un diagramme en bâtons (figure 4.1) à un histogramme (figure 4.2a où les bâtons deviennent des rectangles, avec une largeur).

**Exemple** : Calculer  $\mathbb{P}[8 \leq X \leq 12]$  lorsque  $X \sim \mathcal{B}(50; 0,2)$ .

On remarque tout d'abord que  $n = 50 \geq 30$ , que  $np = 10 > 5$ , et que  $n(1 - p) = 40 > 5$ , donc on peut approcher  $X$  par la loi  $\mathcal{N}(10; 2,83)$  (car  $2,83 \simeq \sqrt{50 \times 0,2 \times (1 - 0,2)}$ ).

Pour calculer  $\mathbb{P}[8 \leq X \leq 12]$ , avec une loi normale, on va en fait calculer  $\mathbb{P}[7,5 \leq X \leq 12,5]$  (ce qui prend en compte la largeur des rectangles pour approcher la loi normale). Une des façons de le comprendre est de dire que les lois normales sont des variables continues (leur valeurs ont plein de chiffres derrière la virgule), alors que les lois binomiales prennent uniquement des valeurs entières. Donc ça serait plutôt l'arrondi de la loi normale (pour en faire un nombre entier) qui approxime la loi binomiale. Et si l'arrondi d'un nombre est entre 8 et 12, c'est que le nombre de départ était entre 7,5 et 12,5 (d'où le calcul de  $\mathbb{P}[7,5 \leq X < 12,5]$  avec la loi normale).

$$\begin{aligned} \text{Ainsi, on a } \mathbb{P}[8 \leq X \leq 12] &\simeq \mathbb{P}[7,5 \leq X < 12,5] \simeq \mathbb{P}\left[\frac{7,5-10}{2,83} \leq Z < \frac{12,5-10}{2,83}\right] \\ &\simeq \mathbb{P}[-0,88 \leq Z < 0,88] \simeq F(0,88) - F(-0,88) \\ &\simeq F(0,88) - 1 + F(0,88) \simeq 0,8106 - 1 + 0,8106 \simeq 0,62 \end{aligned}$$

**Remarque** l'approximation par une loi normale étant assez imprécise, il est inutile de garder trop de chiffres après la virgule dans le résultat final (le plus souvent on en gardera 2 ou 3).





# Chapitre 5

## Estimation

### 5.1 « Prédiction » de la valeur d'une loi normale

Ce chapitre utilisera beaucoup la loi normale et avant de commencer le chapitre, il est utile de revenir le « pic » formé par sa courbe en cloche.

Remarquons tout d'abord que pour une loi normale centrée réduite  $\mathcal{N}(0; 1)$ , on peut calculer  $\mathbb{P}[-1,96 \leq Z \leq 1,96] = F(1,96) - F(-1,96) \simeq 0,9750 - (1 - 0,975) \simeq 0,975 - 0,025 \simeq 0,95$ . Comme  $\mathbb{P}[-1,96 \leq Z \leq 1,96] \simeq 0,95$ , on considérera que « avec la confiance 95% » la loi normale centrée réduite est entre -1,96 et 1,96. La table en bas de la page 10 du formulaire donne d'autres valeurs pour d'autres niveau de confiance que 0,95 : par exemple, avec la confiance 0,99, la loi normale centrée réduite est entre -2,576 et 2,576 (c'est à dire que  $\mathbb{P}[-2,576 \leq Z \leq 2,576] \simeq 0,99$ ). La figure 5.1a représente cela graphiquement.

Lorsque l'on considère une loi normale avec une moyenne et un écart type, on peut procéder de même, en choisissant un intervalle qui s'étend autant de chaque côté de la moyenne. Pour la confiance 95%, par exemple, on considérerait l'intervalle  $[\mu - 1,96\sigma ; \mu + 1,96\sigma]$  :

$$\mathbb{P}[\mu - 1,96\sigma \leq X \leq \mu + 1,96\sigma] = \mathbb{P}\left[\underbrace{\frac{\mu - 1,96\sigma - \mu}{\sigma}}_{-1,96} \leq \underbrace{\frac{X - \mu}{\sigma}}_Z \leq \underbrace{\frac{\mu + 1,96\sigma - \mu}{\sigma}}_{1,96}\right] = \mathbb{P}[-1,96 \leq Z \leq 1,96] \simeq 0,95$$

De même, pour la confiance 99% on considérerait l'intervalle  $[\mu - 2,576\sigma ; \mu + 2,576\sigma]$  et de même avec les autres niveaux de confiance que l'on trouve en page 10 du formulaire. La figure 5.1b le représente pour la loi  $\mathcal{N}(50; 10)$ .

Dans la suite de ce chapitre, pour expliquer comment estimer des proportions, on pourra donc utiliser ce qu'on vient de voir. à savoir que pour un niveau de confiance fixé, on construit dont on affirme (avec ce niveau de confiance) que la loi normale doit se trouver dans l'intervalle.

Le plus fréquent est de considérer une confiance de 95%, et c'est que l'on fera dans la plupart des exemples. Toutefois il est utile de savoir procéder pour n'importe quel niveau de confiance, donc lorsque les procédures générales seront énoncées, on décrira comment procéder pour une confiance arbitraire.

### 5.2 Estimation d'une proportion

À titre d'introduction, cherchons à déterminer quelle proportion des étudiant·e·s du campus de Dijon souffrent de problèmes de stress. Pour cela on interrogera 75 étudiants·e·s au hasard et on désignera par  $X$  le nombre d'étudiants stressés au sein de l'échantillon.  $X$  est une variable aléatoire, puisque sa valeur dépend du groupe de 75 étudiant·e·s (choisi·e·s au hasard) que l'on interroge.

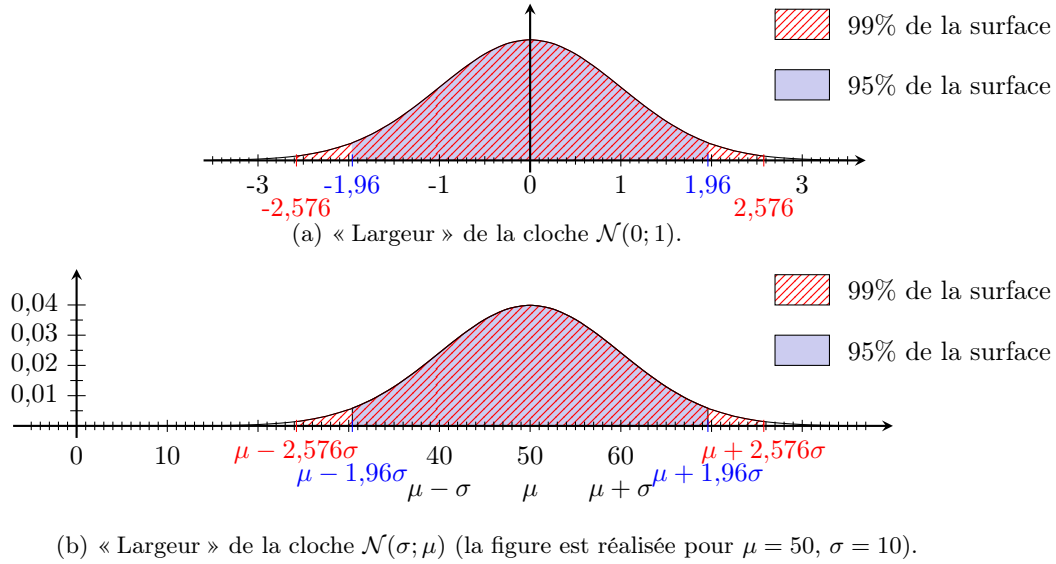


FIGURE 5.1: « Largeur du pic » de la loi normale.

**Exemple de calcul** que l'on pourrait faire si l'on connaissait la proportion d'étudiants stressés au sein du campus de Dijon :

Imaginons par exemple qu'il y ait 40% des étudiants qui souffrent de stress (sur l'ensemble du campus). Dans ce cas  $X \approx \mathcal{B}(75; 0,4)$  (car en tout il y a environ 30 000, et 75 est beaucoup plus petit que 30 000). De plus, d'après la chapitre précédent  $\mathcal{B}(75; 0,4) \approx \mathcal{N}(\underbrace{30}_{75 \times 0,4}; \underbrace{4,24}_{\sqrt{75 \times 0,4 \times 0,6}})$  (car  $n = 75 > 5$ ,  $np = 75 \times 0,4 = 30 > 5$  et  $n(1 - p) = 75 \times 0,6 = 45 > 5$ ).

On en déduit avec la confiance 95% que  $X$  doit être entre  $30 - 1,96 \times 4,24 \simeq 21,69$  et  $30 + 1,96 \times 4,24 \simeq 38,31$ .

En conséquence, si on effectue l'expérience, et que le nombre d'étudiants stressés au sein de l'échantillon n'est pas entre 22 et 38, alors on sera en dehors de l'intervalle auquel on s'attend, et on trouvera cela incompatible avec les probabilités calculées. On conclura alors (avec la confiance 95%) que le calcul de probabilité est faux, c'est à dire que  $X$  ne suit pas la loi  $\mathcal{B}(75; 0,4)$ , et donc qu'il n'y a pas 40% d'étudiants stressés sur le campus dijonnais.

**Estimation de la proportion** On aimerait maintenant raisonner de la même manière pour exclure d'autres valeurs que 0,4, afin d'avoir un maximum d'informations sur la proportion d'étudiant·e-s dijonnais·es stressé·e-s :

Pour cela on va cette fois-ci raisonner de manière plus abstraite, en introduisant la proportion  $p$  d'étudiant·e-s stressé·e-s à Dijon. On va l'utiliser dans les calculs de probabilité alors qu'on ne connaît pas sa valeur a priori. On va quand même supposer que  $75 \times p > 5$  et que  $75 \times (1 - p) > 5$  pour pouvoir approximer la loi binomiale ( $X \sim \mathcal{B}(75, p)$ ) par la loi normale  $\mathcal{N}(75p; \sqrt{75 \times p \times (1 - p)})$ . Dans ces condition, on peut aussi considérer la proportion aléatoire  $P$  d'étudiant·e-s stressé·e-s :  $P = \frac{X}{75} \approx \mathcal{N}(p; \sqrt{\frac{p \times (1 - p)}{75}})$  (cette loi normale découle de celle de  $X$  en divisant la moyenne et l'écart type par 75, car  $P = \frac{X}{75}$ ). Pour simplifier les notation, la quantité  $\sqrt{\frac{p \times (1 - p)}{75}}$  sera parfois notée  $s$  dans la suite. D'après ce qu'on a vu dans la partie 5.1, on s'attend (avec la confiance 95%) à ce que  $P$  soit dans l'intervalle  $[p - 1,96s; p + 1,96s]$ , c'est à dire qu'avec la confiance 95% on affirme que  $|P - p| \leq 1,96s$ .

En pratique on ne connaît pas  $p$ , et on cherche à estimer sa valeur. En revanche on peut mesurer  $P$  en faisant l'expérience. On interroge donc 75 étudiant·e-s, et on en trouve 18 qui sont stressé·e donc une

proportion  $\frac{18}{75} = 0,24$ . Cette valeur que l'on a mesuré pour  $P$  est la « proportion expérimentale » notée  $p_e$ . Elle correspond à la valeur de  $P$  pour un échantillon précis d'étudiants, et ne doit pas être confondue avec la variable aléatoire  $P$  elle-même ( $P$  désigne une variable aléatoire, qui varie d'un échantillon d'étudiants à l'autre). Avec la confiance 95% on considère que  $|p_e - p| \leq 1,96s$ , c'est à dire que  $p \in [p_e - 1,96s; p_e + 1,96s]$ . Comme on ne connaît pas très précisément la valeur de  $p$ , on est contraint de remplacer  $p$  par  $p_e$  dans le calcul de  $s$ , afin de trouver une valeur approximative de  $s$  :  $s \simeq \sqrt{\frac{0,24(1-0,24)}{75}} \simeq 0,049$ . Dès lors on déduit que  $1,96s \simeq 1,96 \times 0,049 \simeq 0,096$ , et que  $p_e - 1,96s \simeq 0,24 - 0,096 = 0,144$  et  $p_e + 1,96s \simeq 0,24 + 0,096 = 0,336$ . Ainsi comme on pensait que  $p \in [p_e - 1,96s; p_e + 1,96s]$ , on affirme qu'avec la confiance 95%,  $p$  est dans l'intervalle  $[0,144; 0,336]$ , c'est à dire que la proportion d'étudiant·e·s stressé·e·s au sein du campus dijonnais se trouve entre 14% et 34%.

L'intervalle  $[0,144; 0,336]$  s'appelle « intervalle de confiance » : avec la confiance 95%, on pense que la proportion  $p$  appartient à cet intervalle.

### Procédure générale (page 5 du formulaire)

On suppose que l'on dispose d'un échantillon de taille  $n$  choisi au sein d'une population qui contient beaucoup d'individus (la population entière contient  $N$  individus et  $N > 10n$ ). On désigne par  $p$  la proportion d'individus ayant une certaine caractéristique au sein de la population totale, et par  $p_e$  la proportion au sein de l'échantillon. De plus, on se donne un niveau de confiance  $c$  ou un risque d'erreur  $\alpha = 1 - c$  (par exemple la confiance  $c = 95\%$  correspond à risque d'erreur  $\alpha = 5\%$ ).

- Pour pouvoir effectuer une estimation il faut tout d'abord vérifier que  $n \geq 30$ , que  $np_e \geq 5$  et que  $n(1 - p_e) \geq 5$ . Si ces hypothèses sont satisfaites alors on peut suivre les étapes suivantes et déterminer un intervalle de confiance (car ces hypothèses autorisent à utiliser une loi normale) :
- La table du formulaire indique un nombre  $z_\alpha$  associée à ce niveau de confiance (en l'occurrence il satisfait  $F(z_\alpha) = 1 - \frac{\alpha}{2} = \frac{c+1}{2}$ ). Par exemple, pour la confiance 95%,  $z_\alpha = 1,96$ .
- On calcule  $a_\alpha = z_\alpha \sqrt{\frac{p_e(1-p_e)}{n}}$ , puis on calcule les nombres  $p_e - a_\alpha$  et  $p_e + a_\alpha$ .
- Avec la confiance  $c = 1 - \alpha$ , on peut affirmer que la proportion  $p$  (au sein de la grande population) se trouve dans l'intervalle  $[p_e - a_\alpha, p_e + a_\alpha]$ .

**Exemple** Si on fixe une confiance de 95% et on applique cette procédure pour un échantillon de 75 étudiant·e·s du campus Dijonnais, parmi lesquels 18 sont stressé·e·s, on obtient la même chose que ce qu'on a calculé précédemment :

- $p_e = \frac{18}{75} = 0,24$
- $n = 75 > 30$ ,  $n \times p_e = 75 \times 0,24 = 18 > 5$  et  $n \times (1 - p_e) = 75 (1 - 0,24) = 57 > 5$ , donc on peut utiliser cette procédure.
- Pour la confiance 95%, on a  $z_\alpha \simeq 1,96$ , et on obtient  $a_\alpha \simeq 1,96 \sqrt{\frac{0,24(1-0,24)}{75}} \simeq 0,097$ , puis  $p_e - a_\alpha \simeq 0,24 - 0,097 = 0,143$  et enfin  $p_e + a_\alpha \simeq 0,24 + 0,097 = 0,337$
- On conclue donc avec la confiance 95% que la proportion  $p$  d'étudiant·e·s stressé·e·s sur le campus dijonnais est dans l'intervalle  $[0,143; 0,337]$ .

**Taille de l'échantillon** Dès lors, une des questions que l'on peut se poser est « quelle taille d'échantillon dois-je prendre si je veux que mon intervalle de confiance soit suffisamment étroit ? »

Pour préciser la question, on va considérer que « l'étroitesse » de l'intervalle est donnée par le nombre  $a_\alpha$  qu'on a calculé ci-avant, et qui indique à quel point l'intervalle s'étend de part et d'autre de  $p_e$ .

Il s'avère que si on veut une précision  $h$  (c'est à dire si l'on veut que  $a_\alpha < h$ ), alors il convient de prendre un échantillon de taille  $n > z_\alpha^2 \frac{p_e(1-p_e)}{h^2}$  (cette formule est donnée en page 5 du formulaire).

Si c'est avant de faire l'expérience que l'on s'interroge sur la taille de l'échantillon à considérer, alors on ne dispose pas de la proportion expérimentale  $p_e$ , mais on peut prendre  $n > z_\alpha^2 \frac{1}{4h^2}$  (cela suffit pour être sûr que  $a_\alpha < h$ ).

## 5.3 Estimation d'une moyenne

L'estimation d'une moyenne correspond à une situation similaire : on dispose d'une variable statistique  $X$ , définie sur une grande population. On cherche à estimer sa moyenne  $\mu$  sur cette grande population, à partir des données récoltées sur un petit échantillon de  $n$  individus.

Le procédé est similaire à l'estimation de proportion, mais il y a deux cas de figures selon que la taille  $n$  de l'échantillon soit plus grande ou plus petite que 30.

- Si  $n \geq 30$ , alors une approximation par une loi normale donne la procédure indiquée dans le formulaire :
  - On fixe une confiance  $c$
  - La table indique le  $z_\alpha$  correspondant à ce niveau de confiance (c'est à dire tel que  $F(z_\alpha) = \frac{c+1}{2}$ )
  - On calcule  $a_\alpha = z_\alpha \frac{s_e}{\sqrt{n-1}}$  où  $s_e$  désigne l'écart type expérimental au sein de l'échantillon. On calcule ensuite  $m_e - a_\alpha$  et  $m_e + a_\alpha$  où  $m_e$  désigne la moyenne expérimentale au sein de l'échantillon.
  - Avec la confiance  $c$ , on peut affirmer que la moyenne sur l'ensemble de la population (c'est à dire  $\mu$ ) est dans l'intervalle  $[m_e - a_\alpha ; m_e + a_\alpha]$ .
- Si  $n < 30$ , alors la procédure ne fonctionne que si  $X$  suit une loi normale. Sous cette hypothèse,
  - On fixe une confiance  $c$
  - Le seuil  $z_\alpha$  (que l'on avait pour les estimations de proportions et pour les grands échantillons) est remplacé par un seuil  $t_\alpha$  que l'on lit dans la table de la loi de student à  $n - 1$  degrés de liberté. On trouve ce seuil en page 11 du formulaire, à la ligne  $n + 1$  (car il y a  $n + 1$  degrés de liberté) et à la colonne  $p = \alpha/2$  (dans l'explication de la procédure, page 5 du formulaire, les valeurs de  $p$  sont indiquées par une table, ce qui indique donc quelle colonne il faut regarder dans la page 11 du formulaire).
  - On calcule  $a_\alpha = t_\alpha \frac{s_e}{\sqrt{n-1}}$  où  $s_e$  désigne l'écart type expérimental au sein de l'échantillon. On calcule ensuite  $m_e - a_\alpha$  et  $m_e + a_\alpha$  où  $m_e$  désigne la moyenne expérimentale au sein de l'échantillon.
  - Avec la confiance  $c$ , on peut affirmer que la moyenne sur l'ensemble de la population (c'est à dire  $\mu$ ) est dans l'intervalle  $[m_e - a_\alpha ; m_e + a_\alpha]$ .

La seule différence pratique entre ces procédures est donc que  $z_\alpha$  est remplacée par  $t_\alpha$  si  $n \geq 30$ .

**Remarque** Cette année on ne dit pas comment déterminer précisément si le résultat d'une expérience est conforme à une loi normale. Cela fait partie des choses que vous verrez en deuxième année, et pour cette année on devra parfois supposer qu'on a des lois normales, sans avoir réellement de moyen de s'en assurer.

**Exemple 1** On utilise les données de l'exercice 8 pour estimer le nombre moyen de personnes par ménage au sein de la population Française, avec une confiance de 95%. On a trouvé en séance de TD que  $n = 6326$ , que  $m_e = 2,55$  et  $s_e = 1,3$ . On applique donc la procédure correspondant au cas où  $n \geq 30$  : On lit sur la table inverse que pour la confiance 0,95 on a  $z_\alpha \simeq 1,96$  d'où  $a_\alpha = z_\alpha \frac{s_e}{\sqrt{n-1}} \simeq 0,03204$ . On estime donc que  $\mu$  est dans l'intervalle  $[2,55 - 0,03204 ; 2,55 + 0,03204] \simeq [2,52 ; 2,58]$ , avec la confiance  $c = 0,95$ .

- Remarques**
- La procédure du formulaire s'appuie sur une approximation par une loi normale. Il est pourtant clair que la variable  $T$  (nombre de personnes d'un ménage) ne suit pas elle-même une loi normale (ne serait-ce que parce qu'elle prend uniquement les valeurs 1, 2, 3, 4, 5, et 6). En fait ce qui suit une loi normale est « la moyenne de  $T$  au sein d'échantillon de 6326 ménages », qui est bien une variable aléatoire (elle varie d'un échantillon à l'autre).
  - Il n'est pas très pertinent d'estimer le nombre d'habitants par ménage à partir de nombreux ménages issus de la même ville, car on s'attend à ce que le nombre d'habitants par ménage dépende du fait que l'on plûtôt en milieu rural ou urbain. L'intervalle que l'on a calculé s'appliquerait donc aux zones géographiques comparables à cette ville, plutôt qu'à la population française dans son ensemble.

**Exemple 2** On utilise les données de l'exercice 25 pour estimer le niveau d'anxiété moyen des personnes victimes d'agressions, avec la confiance 95%. On considère le niveau d'anxiété en l'absence de thérapie, donc la ligne « Avant » du tableau de données de l'exercice 25, qui est notée  $X$  dans le corrigé de l'exercice 25 mis en ligne.

Comme indiqué dans le corrigé de l'exercice 25, on trouve que  $n = 16$ , que  $m_e = 28,94$  et  $s_e = 7,09$ . On va supposer que  $X$  suit une loi normale et appliquer la procédure correspondant au cas où  $n < 30$  :

On cherche  $t_\alpha$  dans la table page 11 du formulaire, en ligne 15 (car  $16 - 1 = 15$ ) et colonne 0,025 (car  $c = 0,95$ ), On lit  $t_\alpha \simeq 2,1314$  d'où  $a_\alpha = t_\alpha \frac{s_e}{\sqrt{n-1}} \simeq 3,9018$ .

On estime donc que  $\mu$  est dans l'intervalle  $[28,94 - 3,9018; 28,94 + 3,9018] \simeq [25,04; 32,84]$  avec la confiance  $c = 0,95$

**Remarque** Bien qu'on suppose que  $X$  suit une loi normale, on a utilisé la table de la loi de student (table page 11 du formulaire). En effet, même si  $X$  suit une loi normale, la procédure du formulaire s'appuie sur « la moyenne de  $X$  au sein d'échantillon aléatoire de 16 victimes », qui suit dans ce cas une loi « de student ».

**Taille de l'échantillon** Comme pour la moyenne on peut se demander : « quelle taille d'échantillon dois-je prendre si je veux que mon intervalle de confiance soit suffisamment étroit ? »

Il s'avère que si on veut une précision  $h$  (c'est à dire si l'on veut que  $a_\alpha < h$ ), alors il convient de prendre un échantillon de taille  $n > z_\alpha^2 \frac{(s_e)^2}{h^2}$ .

- Remarques**
- Cette formule est donnée en page 5 du formulaire.
  - Dans cette formule,  $z_\alpha$  correspond à une loi normale, c'est le nombre qu'on aurait utilisé dans la procédure où  $n \geq 30$ .

## 5.4 Estimation d'un écart type

On considère encore une fois une variable statistique  $X$ , définie sur une grande population. On cherche à estimer son écart type  $\sigma$  sur cette grande population, à partir des données récoltées sur un petit échantillon de  $n$  individus.

Cette fois, la procédure donnée dans le formulaire ne fonctionne que si on suppose que  $X$  suit une loi normale. Cette procédure est la suivante :

- On lit dans la table du  $\chi^2$  à  $n - 1$  degrés de libertés deux nombres  $x_1$  et  $x_2$  :
  - On se concentre sur la ligne  $n - 1$  de la table en page 12 du formulaire (car ce qui apparaît cette fois-ci est une « du  $\chi^2$  à  $n - 1$  degrés de libertés »).
  - On trouve  $x_1$  à la colonne  $q = \frac{1-c}{2}$  et  $x_2$  à la colonne  $p = \frac{1-c}{2}$ . Par exemple si la confiance est 0,95, alors  $x_1$  est dans la colonne  $q = \frac{1-c}{2}$  et  $x_2$  dans la colonne  $p = \frac{1-c}{2}$ . Pour lire ces valeurs on

prendra bien garde que le titre des colonnes de la table contient deux lignes (une pour les valeurs de  $p$  et une pour les valeurs de  $q$ ).

- Avec la confiance  $c$ , on peut affirmer que l'écart type sur l'ensemble de la population (c'est à dire  $\sigma$ ) est dans l'intervalle  $\left[ s_e \sqrt{\frac{n}{x_2}}; s_e \sqrt{\frac{n}{x_1}} \right]$ .

**Exemple** On utilise à nouveau les données de l'exercice 25, mais pour estimer cette fois-ci l'écart type du niveau d'anxiété des personnes victimes d'agressions, avec la confiance 95%. Comme précédemment on a  $n = 16$ , et  $s_e = 7,09$ . On suppose que  $X$  suit une loi normale et on applique la procédure du formulaire : On lit dans la table du  $\chi^2$  à 15 degrés de liberté la valeur  $x_1 = 6,262$  correspondant à  $p = 0,025$  et la valeur  $x_2 = 27,49$  correspondant à  $q = 0,025$

On obtient alors  $s_e \sqrt{\frac{n}{x_2}} \simeq 7,09 \sqrt{\frac{16}{27,49}} \simeq 5,41$  et  $s_e \sqrt{\frac{n}{x_1}} \simeq 7,09 \sqrt{\frac{16}{6,262}} \simeq 11,3$

On estime donc que  $\sigma$  est dans l'intervalle  $[5,41; 11,3]$  avec la confiance  $c = 0,95$ .

## 5.5 Que conclure d'un intervalle de confiance ?

### Comparer un intervalle à une valeur

Il arrive qu'un intervalle de confiance serve à répondre à une question comme « Y a t'il plus de 30% des étudiants dijonnais qui sont stressés ? » où l'on doit comparer l'intervalle de confiance que l'on a calculé avec une valeur particulière (ici 30%). Il y a alors deux cas de figure :

- On obtient parfois que cette valeur précise est à l'intérieur de l'intervalle. Dans ce cas l'intervalle que l'on a obtenu contient à la fois des nombres supérieurs à cette valeur et des nombres inférieurs, donc l'intervalle qu'on a obtenu ne permet pas de tirer de conclusion qui réponde à la question.

C'est par exemple le cas quand on demande « Y a t'il plus de 30% des étudiants dijonnais qui sont stressés ? », sachant qu'on avait obtenu que la proportion d'étudiants stressés à dijon était dans l'intervalle  $[0,143; 0,337]$  : la valeur 30% est à l'intérieur de l'intervalle, cet intervalle contient à la fois des valeurs plus grandes que 30% et des valeurs plus petites, donc cet intervalle ne permet pas de conclure.

- Si l'on obtient au contraire que cette valeur précise est à l'extérieur de l'intervalle, alors on a soit que toutes les valeurs de l'intervalle sont plus grande que cette valeur soit qu'elle sont toutes plus petites que cette valeur précise. Dans les deux cas, l'intervalle permet de conclure et de répondre à la question avec la confiance  $c$ .

Par exemple, si on demande « Y a t'il plus de 40% des étudiants dijonnais qui sont stressés ? », alors toutes les valeurs de l'intervalle  $[0,143; 0,337]$  sont plus petites que 40% donc on pourra affirmer (avec la confiance 95%) qu'il y a moins de 40% des étudiants dijonnais qui sont stressés.

### Comparer deux intervalles entre eux

On est parfois amené à comparer des valeurs estimées dans différentes conditions. Par exemple, avec les données de l'exercice 25, on pourrait demander « le niveau moyen d'anxiété est il plus faible après la thérapie qu'avant ? ». Pour cela on a calculé qu'avant la thérapie le niveau moyen d'anxiété est dans l'intervalle  $[25,04; 32,84]$  (avec la confiance 95%). On pourrait de même estimer ce qu'il en est après la thérapie et on obtiendrait l'intervalle  $[16,88; 25,36]$ , et il conviendrait alors de comparer les deux intervalles.

Dans ce type de questions, deux cas de figure peuvent survenir :

- Si les intervalles se chevauchent (c'est à dire qu'ils ont des valeurs en commun), alors ils ne permettent pas de conclure.

C'est par exemple le cas ici : les deux intervalles  $[25,04; 32,84]$  et  $[16,88; 25,36]$  se chevauchent. Il serait possible que le niveau d'anxiété moyen soit plus élevé avant la thérapie (par exemple s'il est de 30,24 avant la thérapie et de 19,71 après la thérapie) ou bien qu'il soit plus élevé après la thérapie (par exemple s'il est de 25,15 avant la thérapie et de 25,25 après la thérapie).

- Au contraire, si les intervalles ne se chevauchent pas, on peut conclure et indiquer laquelle des deux valeurs est plus grande.

### Comparer les moyennes de deux variables appariées

Si on compare les moyennes de deux variables appariées, alors on peut parfois obtenir des conclusions plus précises en calculant leurs différences. Par exemple, avec les données de l'exercice 25, on obtient :

Avant	33	31	39	31	23	24	37	28	23	14	31	37	23	38	19	32
Après	21	14	29	28	13	16	28	21	25	8	24	19	11	38	16	27
Différence	12	17	10	3	10	8	9	7	-2	6	7	18	12	0	3	5

On a alors  $n = 16$ , et on peut calculer que la moyenne de la différence (au sein de l'échantillon) est  $m_e = 7,81$  tandis que son écart type est  $s_e = 5,31$ .

Cela permet de faire une estimation de la moyenne de la différence :

On lit  $t_\alpha \simeq 2,1314$  d'où  $a_\alpha = t_\alpha \frac{s_e}{\sqrt{n-1}} \simeq 2,9222$ .

On estime donc que  $\mu$  est dans l'intervalle  $[7,81 - 2,9222; 7,81 + 2,9222] \simeq [4,89; 10,73]$  avec la confiance  $c = 0,95$

Cette fois ci, on est en mesure de conclure que la moyenne de la différence est positive, c'est à dire que les niveaux d'anxiété sont en moyenne plus haut avant la thérapie (avec la confiance 95%).