Statistiques en L1 de psychologie

Sébastien Leurent

Année 2025-2026

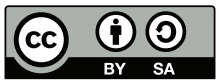


Table des matières

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 notamment à introduire quelques 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* dont on doit tenir compte lorsqu'on tire des conclusions 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 descriptive univariée
Chapitre 2 : Statistique descriptive bivariée } **Statistique descriptive** : Résumer des données expérimentales

Chapitre 3 : Probabilités } **Probabilité** : Outil de prédiction du résultat d'une expérience, sous certaines hypothèses

Chapitre 4 : Estimation } **Statistique inférentielle** : déductions à partir de mesures sur un échantillon

Déroulement du semestre et contrôle des connaissances

Les présentes notes de cours [1], ainsi que les 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`.

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 (réinitialisé ou en mode examen) et le formulaire (vierge de toute annotation) sont autorisés aux contrôles et examens.

Contrôle des connaissances

- Un contrôle terminal (CT) a lieu en fin de semestre et donne la moitié de la note de l'UE (la totalité en deuxième session).
- Une note de contrôle continue (CC) est obtenue au cours du semestre, et constitue l'autre moitié de la note de l'UE (en première session uniquement). Elle est constituée
  - pour deux tiers d'un contrôle commun à tous les groupes de TD, qui devrait avoir lieu le jeudi 12 mars (date à confirmer).
  - pour un tiers d'une note attribuée au sein de chaque groupe de TD, à partir d'au moins deux petits contrôles et éventuellement d'autres notes (exemple : participation).

La calculatrice (réinitialisée ou en « mode examen ») et le formulaire (vierge de toute annotation) sont autorisés aux différents contrôles et examens.

1. Ces notes de cours sont largement inspirées de notes de cours écrites par A. Jebrane, et ont été relues et modifiées par G. Massuyeau ; il convient de les remercier tous les deux 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 :

$$-x + \frac{102}{8+9} + 3(-3+7) - 5^2$$

Notations

- Le symbole « x » désigne un nombre arbitraire, avec lequel on peut faire des calculs « formels » 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 -3 est l'opposé de 3
  - de même $-x$ désigne l'opposé de « x »
- La multiplication est notée par le symbole « $\times$ », ou parfois « $\cdot$ », et par le symbole « $*$ » sur les calculatrices. 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 « $3(-3+7)$ » n'aurait aucun sens si on n'y insère pas un symbole de multiplication, de sorte qu'elle signifie en fait « $3 \times (-3+7)$ ».
- Les divisions sont notées 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 $\frac{102}{8+9}$ signifie $102 \div (8+9)$ (ou $102/(8+9)$ 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, c'est-à-dire 5×5. Sur certaines calculatrices, ce carré est noté 5^2.
- 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 $-x + \frac{102}{8+9} + 3(-3+7) - 5^2$, on effectue les étapes suivantes dans cet ordre :

1. On calcule tout d'abord ce qui est dans des parenthèses, ou au numérateur ou dénominateur d'une fraction. Pour l'expression $-x + \frac{102}{8+9} + 3(-3+7) - 5^2$, on calcule donc $8+9=17$ et $-3+7=4$. On obtient donc
$$-x + \frac{102}{8+9} + 3(-3+7) - 5^2 = -x + \frac{102}{17} + 3 \times 4 - 5^2$$
2. On calcule ensuite les puissances (ici le carré). On calcule donc $5^2 = 25$ et on obtient
$$-x + \frac{102}{17} + 3 \times 4 - 5^2 = -x + \frac{102}{17} + 3 \times 4 - 25$$
3. On effectue ensuite les multiplications et les divisions : on calcule donc $\frac{102}{17} = 6$ et $3 \times 4 = 12$. On obtient donc
$$-x + \frac{102}{17} + 3 \times 4 - 25 = -x + 6 + 12 - 25$$
4. Enfin, on termine par les additions/soustractions :
$$-x + 6 + 12 - 25 = -x - 7$$
Pour cet exemple on obtient donc, à l'issue des calculs, que $-x + \frac{102}{8+9} + 3(-3+7) - 5^2 = -x - 7$.

Ces étapes s'effectuent toujours dans cet ordre (parenthèses, puissances, multiplications/divisions et enfin additions/soustractions). 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 important de connaître ces règles de priorité. Par exemple, il faut savoir que $5 \times 8 + 4$ est égal à $(5 \times 8) + 4$ et non pas à $5 \times (8+4)$.

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
$$6 \times (2-4) = 6 \times 2 - 6 \times 4$$
- 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 $6 - 3 + 7 = 6 + 7 - 3$, et que $\frac{8}{9} \times 2 = \frac{2 \times 8}{9}$.
- 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 *simultanément* 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 *simultanément* par la même quantité. Ainsi, on a par exemple $5 + x - 5 = x$, et $\frac{5 \times 9}{5} = 9$.

Calculatrice

Si vous n'êtes pas habitués à utiliser la calculatrice, vous êtes vivement encouragés à faire les Exercices 1 à 4 de la liste d'exercices de TD. (Ces exercices font partie de ceux qui ne seront pas traités en séances, et dont un corrigé est disponible sur la page `plubel` du cours.) De plus, des vidéos sont mises en ligne sur plubel au fur et à mesure du semestre, pour montrer l'utilisation de plusieurs modèles de calculatrices pour effectuer les principaux calculs de ce cours.

Voici quelques points dont il faut avoir connaissance lorsqu'on utilise la calculatrice :

- Sur certaines calculatrices (en particulier les modèles de TI), il y a deux touches « - » : une touche [-] pour la soustraction, et une touche [(-)] 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 :
  - 9.9E16 signifie 9,9 × 10^{16} (c'est-à-dire 9,9 × 10000000000000000). Quand on multiplie 9,9 par 10^{16}, on décale la virgule 16 fois vers la droite, aboutissant à 99000000000000000.
  - 9.9E9 signifie 9,9 × 10^{9} (c'est-à-dire 9,9 × 0,1). Quand on multiplie 9,9 par 10^{9}, on décale la virgule -9 fois vers la gauche, aboutissant à 0,99.
- Toute calculatrice scientifique peut garder en mémoire le résultat de certains calculs. Lorsque les calculs s'enchaînent, cette fonctionnalité permet notamment d'éviter des erreurs d'arrondis. D'une part, la fonction « Ans » désigne le résultat du dernier calcul effectué. D'autre part, « → » (tapé avec [→] ou [STO] selon les calculatrices) permet d'enregistrer le résultat d'un calcul. Les exercices 2 et 5 en donnent des exemples d'utilisation.

Remarque : pour plusieurs modèles de calculatrice, des vidéos sont mises en ligne sur plubel pour décrire les principaux calculs utilisés dans ce cours.

Complément : Notation $\sum$

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

$$1^2 + 2^2 + 3^2 + 4^2 + 5^2 + 6^2 + 7^2 \tag{1}$$

Cette somme se décrit simplement par un 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 :

$$\sum_{i=1}^{7} i^2 \tag{2}$$

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 même 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 soeurs, 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 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 se 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 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 2025/2026 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 :

Brigitte
de mauvaise humeur 4 frères/soeurs 1m59

Jean-Pierre
de bonne humeur 0 frère/soeur 1m88

Chloe
de relativement bonne humeur 0 frère/soeur 1m64

Magali
de très bonne humeur 3 frères/soeurs 1m65

Myriam
de bonne humeur 3 frères/soeurs 1m69

David
de bonne humeur 2 frères/soeurs 1m80

Sylvie
de bonne humeur 3 frères/soeurs 1m59

Bernard
de bonne humeur 2 frères/soeurs 1m89

Karine
de très bonne humeur 2 frères/soeurs 1m59

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	1	1	5	2

(a) Humeur

Modalité	0	1	2	3	4
Effectif	2	0	3	3	1

(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 $x_1 = 0$, $x_2 = 1$, $x_3 = 2$, $x_4 = 3$, $x_5 = 4$; $n_1 = 2$, $n_2 = 0$, $n_3 = 3$, $n_4 = 3$, et $n_5 = 1$.
Il y a 5 colonnes, d'où $r = 5$.
L'effectif total est $n = 2 + 3 + 3 + 1 = 9$

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

$$n = n_1 + n_2 + \cdots + n_r \,. \tag{1.1}$$

On prendra l'habitude d'écrire la formule (1.1) de manière plus compacte, sous la forme

$$n = \sum_{i=1}^{r} n_i \,, \tag{1.2}$$

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	[1,55 ; 1,60[	[1,60 ; 1,65[	[1,65 ; 1,70[	[1,70 ; 1,75[	[1,75 ; 1,80[	[1,80 ; 1,85[	[1,85 ; 1,90[
Effectif	3	1	2	0	0	1	2

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 [1] 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 $a_1 = 1{,}55$, $a_2 = 1{,}60$, $a_3 = 1{,}65$, $a_4 = 1{,}70$, $a_5 = 1{,}75$, $a_6 = 1{,}80$, $a_7 = 1{,}85$, $a_8 = 1{,}90$; $n_1 = 3$, $n_2 = 1$, $n_3 = 2$, $n_4 = 0$, $n_5 = 0$, $n_6 = 1$, et $n_7 = 2$.
Il y a 7 colonnes, d'où $r = 7$.
L'effectif total est
$$n = 3 + 1 + 2 + 1 + 2 = 9.$$

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

$$f_i = \frac{n_i}{n} \tag{1.3}$$

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 $f_3 = \frac{n_3}{n} = \frac{3}{9} \simeq 0{,}333\,3$. Cela signifie que $\mathbb{P}_r[X = 2] \simeq 0{,}333\,3$, c'est à dire que parmi cet échantillon, il y a environ 33,33% 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 $f_1 + f_2 + f_3 \simeq 0{,}222\,2 + 0 + 0{,}333\,3 = 0{,}555\,5$. Cela signifie que $\mathbb{P}_r[X \leqslant 2] \simeq 0{,}556$, c'est à dire que parmi cet échantillon, il y a environ 55,6 % des étudiants qui ont au maximum 3 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.

Par exemple pour la table 1.3c, la fréquence cumulée de la deuxième colonne est $f_1 + f_2 \simeq 0{,}333\,3 + 0{,}111\,1 = 0{,}444\,4$. Cela signifie que $\mathbb{P}_r[Y < 1{,}65] \simeq 0{,}444$, où l'on note bien que cette fréquence cumulée correspond à la taille "1m65".

1. Rigoureusement, a_{r+1} s'appelle mathématiquement le "suprémum" de la dernière classe, et non pas son "maximum". Dans les notes et les résumés de ce cours, on passera sous silence cette distinction, et par abus de langage, on utilisera le terme "maximum" pour parler en fait de "supremum".

Humeur	mauvaise	relativement bonne	bonne	très bonne
Effectif	1	1	5	2
Fréquence	0,111	0,111	0,556	0,222
Fréquence cumulée	0,111	0,222	0,778	1,000

(a) Humeur

Nombre de frères/soeurs	0	1	2	3	4
Effectif	2	0	3	3	1
Fréquence	0,222	0,000	0,333	0,333	0,111
Fréquence cumulée	0,222	0,222	0,556	0,889	1,000

(b) Nombre de frères et soeurs

Taille	[1,55 ; 1,60[	[1,60 ; 1,65[	[1,65 ; 1,70[	[1,70 ; 1,75[	[1,75 ; 1,80[	[1,80 ; 1,85[	[1,85 ; 1,90[
Effectif	3	1	2	0	0	1	2
Fréquence	0,333	0,111	0,222	0,000	0,000	0,111	0,222
Fréquence cumulée	0,333	0,444	0,667	0,667	0,667	0,778	1,000

(c) Taille

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

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.

Les vidéos mises en ligne sur plubel décrivent une façon de calculer les fréquences et les fréquences cumulées sans erreurs d'arrondis.

Si vous avez quand même ces erreurs d'arrondis, il vaut mieux garder 0,999 ou 1,001 (et se rendre compte de l'imprécision du calcul) 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 situation :

- Pour des variables qualitatives on utilise fréquemment un diagramme en camembert comme celui de la figure 1.4. Les surfaces des différents quartiers sont proportionnels aux fréquences. Pour obtenir celà, il suffit de donner à chaque quartier l'angle $f \times 360°$, où f est la fréquence.
- Pour des variables numériques discrètes, on peut utiliser un diagramme en bâton, où la hauteur de

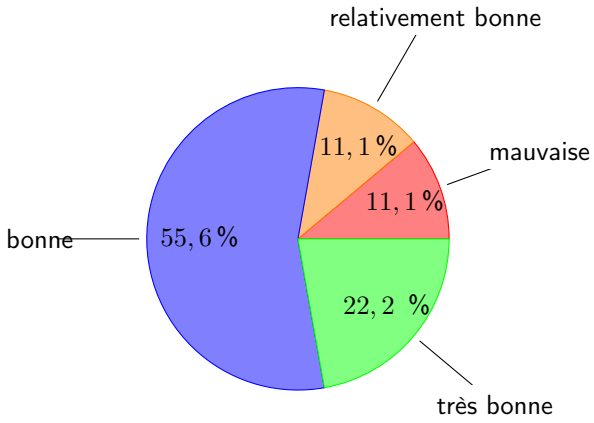


FIGURE 1.4 – Humeur des étudiants : représentation en diagramme circulaire ("camembert")

chaque bâton donne la fréquence d'une modalité. La figure 1.5 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 proportionelle à la fréquence. La figure 1.6 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 ce 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 $F_X(a) = \mathbb{P}_r[X \leqslant a]$.

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

Construction du polygone des fréquences cumulées La figure 1.7 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 $F_Y(a) = \mathbb{P}_r[Y \leqslant a]$.
- Les fréquences cumulées calculées indiquent que $\mathbb{P}_r[Y \leqslant 1{,}6] \simeq 0{,}333$, $\mathbb{P}_r[Y \leqslant 1{,}65] \simeq 0{,}444$, etc,
  c'est à dire que $F_Y(1{,}6) \simeq 0{,}333$, $F_Y(1{,}65) \simeq 0{,}444$, etc,
  donc on place des points de coordonnées $(1{,}6; 0{,}333)$, $(1{,}65; 0{,}444)$, etc.

  On a ainsi un point pour chaque colonne du tableau.

  Remarque : On peut choisir d'écrire ces fréquences en pourcentage sur le dessin, pour plus de lisibilité.
- De plus $\mathbb{P}_r[Y \leqslant 1{,}55] = 0$, donc on ajoute un point de coordonnées $(1{,}55; 0)$
- Entre ces points on manque d'information pour approximer la fonction F_Y. On choisit de relier ces points par des segments de droites.

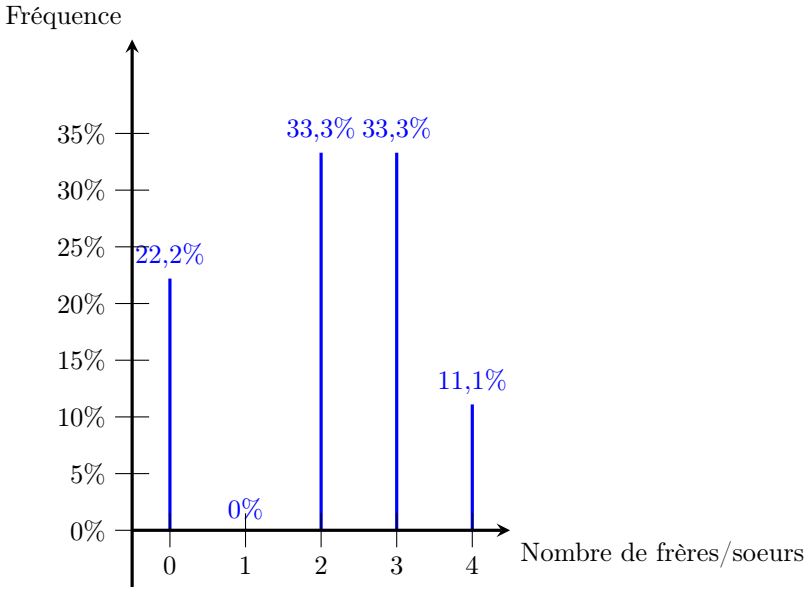


FIGURE 1.5 – Nombre de frères et soeurs : représentation en bâtons

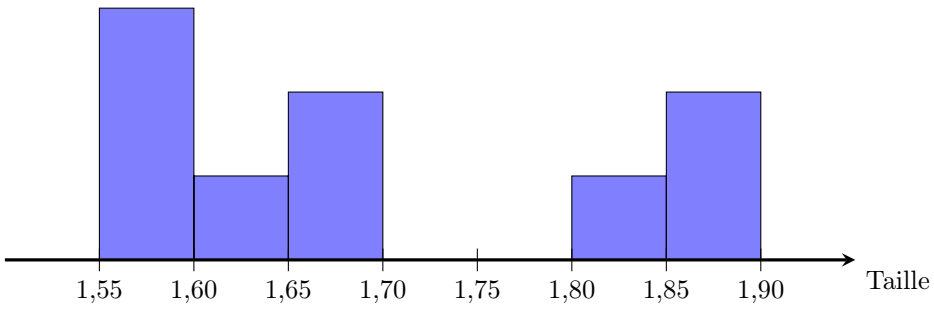


FIGURE 1.6 – Histogramme des fréquences, pour la taille des étudiants de l'échantillon

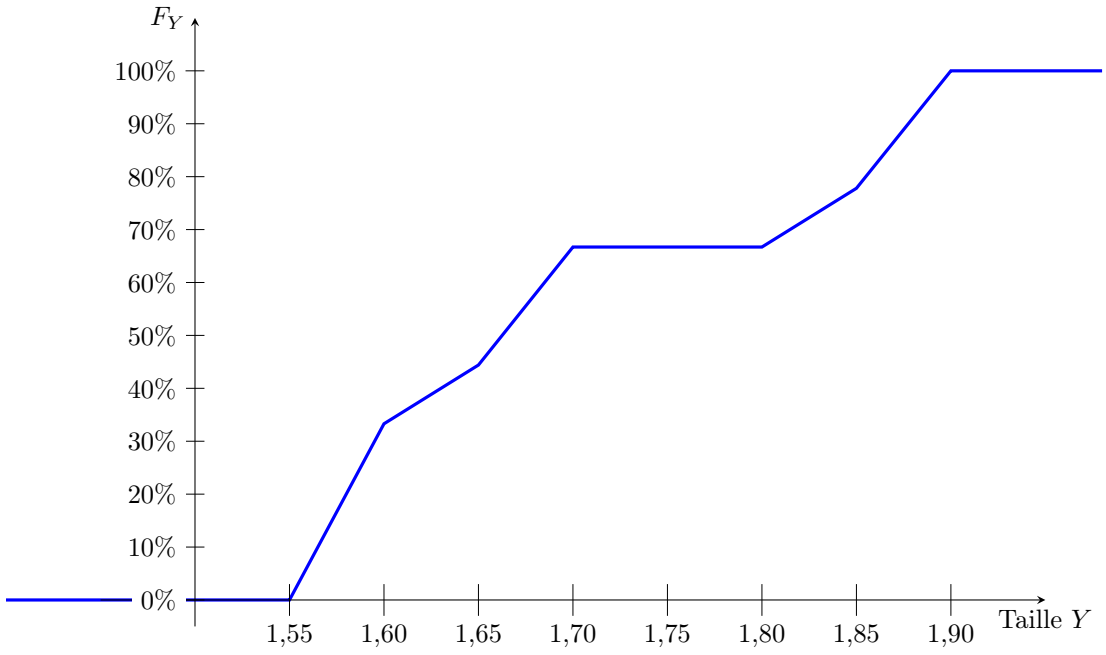


FIGURE 1.7 – Polygone des fréquences cumulées, pour la taille des étudiants de l'échantillon

- Lorsque $t \leqslant 1{,}55$, on a $\mathbb{P}_r[Y \leqslant t] = 0$. En conséquence on trace une demi-droite horizontale à gauche du point de coordonnées $(1{,}55; 0)$.
- De même, lorsque $t > 1{,}9$, on a $\mathbb{P}_r[Y \leqslant t] = 1$. En conséquence on trace une demi-droite horizontale à droite du point de coordonnées $(1{,}9; 1)$.

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 Med telle que $\mathbb{P}_r[X \geqslant \text{Med}] \geqslant 0{,}5$ et que $\mathbb{P}_r[X \leqslant \text{Med}] \geqslant 0{,}5$.

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 $n = 9$ donc $\frac{n+1}{2} = 5$. 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 R B B B B B T T, et la $5^{\text{ème}}$ valeur est "B". La médiane est donc *de bonne humeur*.

Cas de données regroupées en classes Pour des données regroupées en classe on résout de manière approchée l'équation $\mathbb{P}_r[x \leqslant \text{Med}] = 0{,}5$, 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

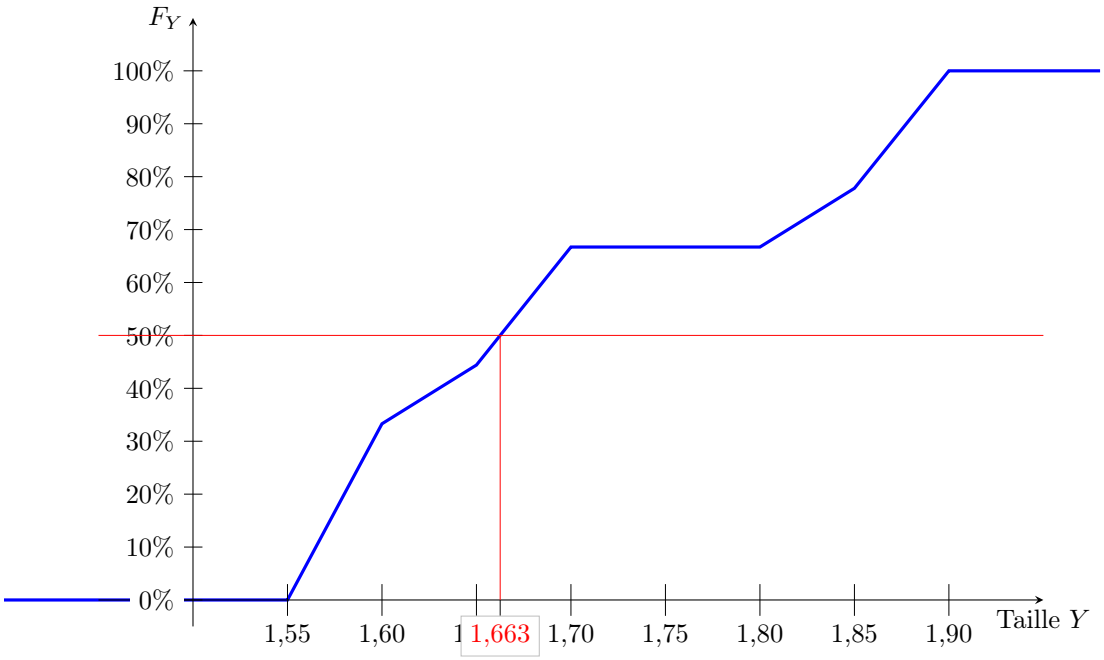


FIGURE 1.8 – Lecture graphique de la taille médiane des étudiants de l'échantillon

cumulées croise la droite d'ordonnée 50%. Comme indiqué en figure 1.8, on lit que la taille médiane est environ 1,663m.

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 $\text{Med} \simeq a_i + \dfrac{a_{i+1} - a_i}{F_X(a_{i+1}) - F_X(a_i)}\big(0{,}5 - F_X(a_i)\big)$.

   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 $[\,1{,}65\ ;\ 1{,}7\,[$. On note donc $a_i = 1{,}65$ et $a_{i+1} = 1{,}7$. On a donc $F_Y(a_i) = 0{,}444$ et $F_Y(a_{i+1}) = 0{,}667$, d'où $\text{Med} \simeq 1{,}65 + \dfrac{1{,}7 - 1{,}65}{0{,}667 - 0{,}444}\,(0{,}5 - 0{,}444) \simeq 1{,}663$.

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 $F_X(\text{Q}_1) = 25\%$ pour définir le premier quartile (Q_1) ou $F_X(\text{Q}_3) = 75\%$ 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, ..., 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'inidividus. 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 $x_1 = 4$, $x_2 = 0$, $x_3 = 0$, $x_4 = 3$, $x_5 = 3$, $x_6 = 2$, $x_7 = 3$, $x_8 = 2$ et $x_9 = 2$. On a donc $m(X) = \frac{4+0+0+3+3+2+3+2+2}{9} \simeq 2{,}111$.

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 $x_1 = 0$, $x_2 = 1$, $x_3 = 2$, $x_4 = 3$, $x_5 = 4$, $x_6 = 5$, $x_7 = 6$; $n_1 = 2$, $n_2 = 0$, $n_3 = 3$, $n_4 = 3$, $n_5 = 1$, $n_6 = 0$, et $n_7 = 0$.

La formule s'écrit donc $m(X) = \frac{2\times0+0\times1+3\times2+3\times3+1\times4}{9} \simeq 2{,}111$.

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 $c_1 = \frac{1{,}55+1{,}6}{2} = 1{,}575$, $c_2 = \frac{1{,}6+1{,}65}{2} = 1{,}625$, etc. En conséquence la formule devient $m(Y) \approx \frac{3\times1{,}575+1{,}625+2\times1{,}675+0\times1{,}725+0\times1{,}775+1{,}825+2\times1{,}875}{9} \simeq 1{,}697$.

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 $m(Y) = \frac{1{,}59+1{,}88+1{,}64+1{,}65+1{,}69+1{,}8+1{,}59+1{,}89+1{,}59}{9} \simeq 1{,}702$.

1.4.3 Écart type

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

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

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

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

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

Exemple Pour le nombre de frères et soeurs, on obtient
$m(X^2) = \frac{2\times(0^2)+0\times(1^2)+3\times(2^2)+3\times(3^2)+1\times(4^2)}{9} \simeq 6{,}111$.
D'où $\mathrm{Var}(X) \simeq 6{,}111 - (2{,}111)^2 \simeq 1{,}655$ et $s(X) \simeq \sqrt{1{,}655} \simeq 1{,}286$.

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 15 personnes de 38 à 85 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	47	68	55	45	39	38	53	55	85	53	48	40	71	79	64
Y	42	33	40	28	77	63	20	50	23	30	59	55	44	32	21

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

Variables 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 "contrôlable" 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*.

	variable indépendante	variable dépendante
Exemple	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 contrôlable soit en fixant artificiellement sa valeur (dosage d'un traitement, alimentation de rats, etc), soit en choisissant le 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 même 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 **corrélation** l'intensité de ce lien, que l'on mesurera à l'aide de *coefficients de corrélation*.
- Peut-on prédire la variable d'un 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 positionnée horizontalement en la valeur x_i que prend la variables 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 corrélations 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.

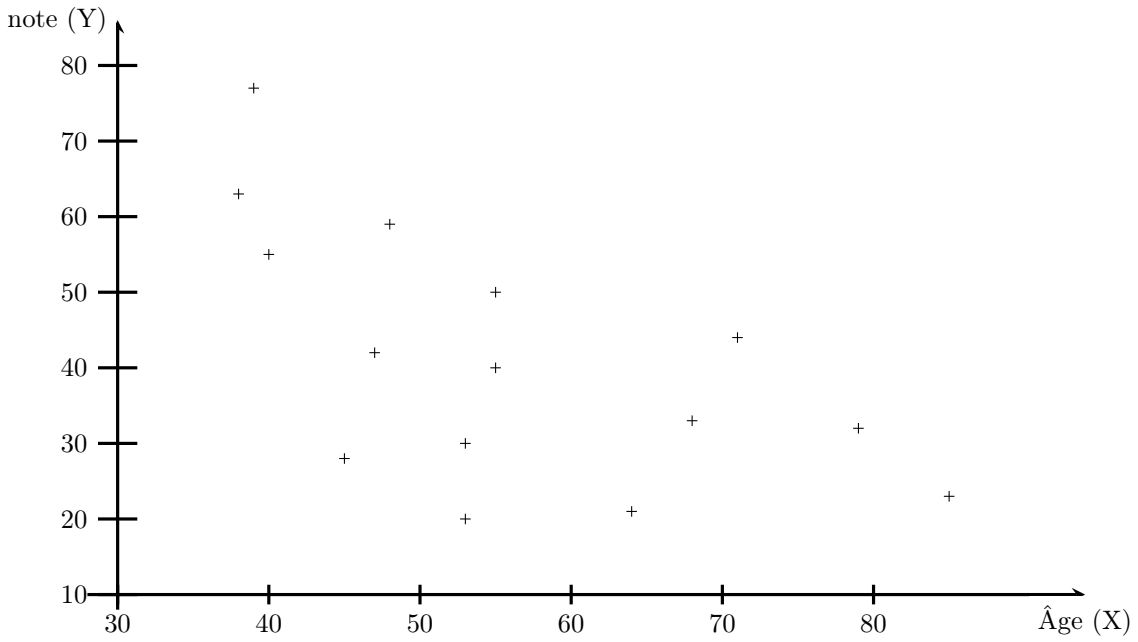


FIGURE 2.1 – Nuage de points indiquant la note en fonction de l'âge

2.3.1 Coefficient de corrélation linéaire

Définition

- On appelle covariance de deux variables quantitatives X et Y la quantité $\text{Cov}(X;Y) = m\left((X - m(X)) \times (Y - m(Y)\right)$ 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 : $r(X;Y) = \dfrac{\text{Cov}(X;Y)}{s(X) \cdot s(Y)}$

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 : $m(X\,Y) = \frac{\sum x_i y_i}{n}$.

Exemple Pour les données présentées en début de chapitre, on obtient :
moyenne: $m(X) = \frac{\sum x_i}{n} = \frac{47+68+55+\cdots+64}{15} = \frac{840}{15} = 56$
$m(X^2) = \frac{\sum x_i^2}{n} = \frac{47^2+68^2+55^2+\cdots+64^2}{15} = \frac{49\,998}{15}$
$Var(X) = m(X^2) - m(X)^2 = \frac{49\,998}{15} - \left(\frac{840}{15}\right)^2 = 197{,}2$
Écart-type: $s(X) = \sqrt{Var(X)} \simeq 14{,}04$
moyenne: $m(Y) = \frac{\sum x_i}{n} = \frac{42+33+40+\cdots+21}{15} = \frac{617}{15} \simeq 41{,}13$
$m(Y^2) = \frac{\sum x_i^2}{n} = \frac{42^2+33^2+40^2+\cdots+21^2}{15} = \frac{29\,371}{15}$
$Var(Y) = m(Y^2) - m(Y)^2 = \frac{29\,371}{15} - \left(\frac{617}{15}\right)^2 \simeq 266{,}12$
Écart-type: $s(Y) = \sqrt{Var(Y)} \simeq 16{,}31$
$m(X\,Y) = \frac{\sum x_i\, y_i}{n} = \frac{47\times42+68\times33+\ldots+64\times21}{15} = \frac{32\,458}{15} \simeq 2\,163{,}867$
$Cov(X,Y) = m(X\,Y) - m(X)\,m(Y) = \frac{32\,458}{15} - \frac{840}{15}\frac{617}{15} = -139{,}6$
$r(X,Y) = \frac{Cov(X,Y)}{s(X)\,s(Y)} = \frac{-139{,}6}{14{,}04\times16{,}31} \simeq -0{,}61$
et le coefficient de corrélation linéaire vaut donc environ -0,61.

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 $6^{\text{ème}}$ individu, X prend la valeur 38, qui est le plus petite valeur. On y associe le rang 1, et on pose $x'_6 = 1$.
- Chez le $5^{\text{ème}}$ individu, X prend la valeur 39, qui est la deuxième plus petite valeur. On y associe le rang 2 (pour « deuxième plus petite valeur »), et on pose $x'_5 = 2$.
- De même la troisième plus petite valeur est 40, qui correspond au $12^{\text{ème}}$ individu, de sorte que l'on pose $x'_{12} = 3$.
- On continue en posant $x'_4 = 4$, $x'_1 = 5$ et $x'_{11} = 6$.
- La valeur suivante est 53 qui correspond aux individus numérotés 7 et 10. On devrait attribuer le rang 7 à l'un et 8 à l'autre, mais pour éviter de choisir arbitrairement auquel des deux attribuer le rang 7, on décide de leur attribuer à tous les deux le rang 7,5.

  On pose donc $x'_7 = 7{,}5$ et $x'_{10} = 7{,}5$.
- Se souvenant que les rangs que l'on vient d'attribuer correspondent à 7 et à 8, on reprend au rang 9, qui correspond à la valeur 55. Or deux individus ont cette valeur : le $3^{\text{ème}}$ et le $8^{\text{ème}}$ individu. On devrait leur attribuer les rangs 9 et 10, mais comme précédemment, on leur attribue le rang 9,5.

  On pose donc $x'_3 = 9{,}5$ et $x'_8 = 9{,}5$.
- 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 2.2.

Individu	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Âge X	47	68	55	45	39	38	53	55	85	53	48	40	71	79	64
Rang x'_i	5	12	9,5	4	2	1	7,5	9,5	15	7,5	6	3	13	14	11

TABLE 2.2 – 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 2.2 on obtient la table 2.3.

Âge X	47	68	55	45	39	38	53	55	85	53	48	40	71	79	64
Note Y	42	33	40	28	77	63	20	50	23	30	59	55	44	32	21
Rang x'_i	5	12	9,5	4	2	1	7,5	9,5	15	7,5	6	3	13	14	11
Rang y'_i	9	7	8	4	15	14	1	11	3	5	13	12	10	6	2

TABLE 2.3 – 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 de 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 $r_s(X;Y) \approx 1 - \left(6\frac{(5-9)^2+(12-7)^2+(9{,}5-8)^2+\cdots+(11-2)^2}{15(15^2-1)}\right) \simeq -0{,}536$.

Organisation du calcul

Pour simplifier le calcul, on peut rajouter à la table 2.3 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 2.4

Âge (X)	47	68	55	45	39	38	53	55	85	53	48	40	71	79	64	
Note (Y)	42	33	40	28	77	63	20	50	23	30	59	55	44	32	21	
Rang x'_i	5	12	9,5	4	2	1	7,5	9,5	15	7,5	6	3	13	14	11	
Rang y'_i	9	7	8	4	15	14	1	11	3	5	13	12	10	6	2	*Total*
$(x'_i - y'_i)^2$	16	25	2,25	0	169	169	42,25	2,25	144	6,25	49	81	9	64	81	860

TABLE 2.4 – Table résumant les calculs permettant de calculer efficacement $r_s(X;Y)$

Interprétation

- 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 23, 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 :

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

Exemple
Pour estimer la note d'une personne de 80 ans, on calcule cette droite :
on pose $a = \frac{Cov(X,Y)}{Var(X)} = \frac{-139{,}6}{197{,}2} \simeq -0{,}708$ et $b = m(Y) - a\,m(X) \simeq 41{,}133 - (-0{,}708) \times 56 \simeq 80{,}781$
D'où l'équation de la droite $D_{Y|X} : Y = -0{,}708\,X + 80{,}781$
Donc pour $x = 80$, on s'attend à $y = -0{,}708 \times 80 + 80{,}781 = 24{,}141$.

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 :

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

Exemple
Pour estimer l'âge d'une personne qui a la note 25, on calcule cette droite :
on pose $a' = \frac{Cov(X,Y)}{Var(Y)} \simeq \frac{-139,6}{266,12} \simeq -0{,}525$ et $b' = m(X) - a'\, m(Y) \simeq 56 - (-0{,}525) \times 41{,}133 \simeq 77{,}595$
D'où l'équation de la droite $D_{X|Y} : X = -0{,}525\, Y + 77{,}595$
Donc pour $y = 25$, on s'attend à $x = -0{,}525 \times 25 + 77{,}595 = 64{,}47$.

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 :

$$\begin{array}{cccccccc} 1+1 & ; & 1+2 & ; & 1+3 & ; & 1+4 \\ 2+1 & ; & 2+2 & ; & 2+3 & ; & 2+4 \\ 3+1 & ; & 3+2 & ; & 3+3 & ; & 3+4 \\ 4+1 & ; & 4+2 & ; & 4+3 & ; & 4+4 \end{array}$$

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	$\frac{1}{16} = 0{,}062\,5$	$\frac{2}{16} = 0{,}125$	$\frac{3}{16} = 0{,}187\,5$	$\frac{4}{16} = 0{,}25$	$\frac{3}{16} = 0{,}187\,5$	$\frac{2}{16} = 0{,}125$	$\frac{1}{16} = 0{,}062\,5$

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 "$X \leqslant 3$" (constitué de "$X = 2$" et "$X = 3$"), "$X \geqslant 5$", 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 :

moyenne: $m(X) = \frac{\sum_i x_i\, n_i}{n} = \frac{2\times1+3\times2+4\times3+\cdots+8\times1}{16} = \frac{80}{16} = 5$

$m(X^2) = \frac{\sum_i x_i^2\, n_i}{n} = \frac{2^2\times1+3^2\times2+4^2\times3+\cdots+8^2\times1}{16} = \frac{440}{16}$

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

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

Un·e psychologue étudie les dessins d'enfants, auxquels on demande de dessiner leur famille. Il/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	FEPM	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 $n \times (n-1) \times (n-2) \times \cdots \times 2 \times 1$ (que l'on peut aussi écrire $1 \times 2 \times 3 \times \cdots \times n$).

Définition

Étant donné un entier n, le nombre $n \times (n-1) \times (n-2) \times \cdots \times 2 \times 1$ 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 $\frac{4}{24} \simeq 0{,}166\,7$.
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 $\frac{1}{6} \simeq 0{,}166\,7$.

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 $0 \leqslant k \leqslant n$).

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

- $\binom{4}{2} = \frac{4!}{2!\times 2!} = \frac{4\times 3\times 2}{(2\times 1)\times(2\times 1)} = (4\times 3)\times\frac{1}{2\times 1}$
  $= \frac{2\times 2\times 3}{2} = 2\times 3 = 6$
  c'est pourquoi on a obtenu 6 cas avec la deuxième méthode dans l'exemple en début de section 3.2.2.
- $\binom{6}{3} = \frac{6!}{3!\times 3!} = \frac{6\times 5\times\ldots\times 1}{(3\times 2\times 1)\times(3\times 2\times 1)} = (6\times 5\times 4)\times\frac{1}{3\times 2\times 1}$
  $= \frac{3\times 2\times 5\times 4}{3\times 2} = 5\times 4 = 20$
  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 18 personnes parmi lesquelles 5 fument quotidiennement. Si on choisit au hasard 3 personnes parmi ce groupe de 18, quelle est la probabilité de choisir un fumeur et deux non-fumeurs ?

- On commence par compter le nombre de cas possibles : il y a $\binom{18}{3}$ choix possible de 3 personnes au sein de l'échantillon. La calculette nous indique que $\binom{18}{3} = 816$ donc il y a 816 cas.
- On compte ensuite le nombre de cas où on a un fumeur et deux non-fumeurs, sachant que le groupe de 18 personne comptait 5 fumeurs et 13 non-fumeurs. Il y a donc $\binom{13}{2}$ choix possibles pour choisir deux non-fumeurs, et pour chacun de ces choix, il y a 5 choix d'un fumeur (pour compléter et obtenir trois personnes). Ainsi le nombre de cas qui nous intéresse est $5\times\binom{13}{2}$. Une calculette nous indique que cela vaut 390.
- On conclue que la probabilité recherchée vaut $\frac{390}{816} \simeq 0{,}477\,9$.

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 (cela a 60% de chances d'arriver) puis en avoir ressenti le second jour (cela a 40% de chances d'arriver) et enfin ne pas en ressentir le troisième jour (cela 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 $0{,}6 \times 0{,}6 \times 0{,}4 + 0{,}6 \times 0{,}4 \times 0{,}6 + 0{,}4 \times 0{,}6 \times 0{,}6 = 0{,}432$.

  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 $0{,}4^2 \times 0{,}6 = 0{,}096$. 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 $\mathbb{P}[X = 3] = 0{,}4^3 = 0{,}064$.

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 de généraliser à toute situations où l'on répète plusieurs fois une expérience qui a à chaque fois la même probabilité de succès (indépendemment 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 $m(X) = n\,p$.
- sa variance est $\mathrm{Var}(X) = n\,p\,(1-p)$.
- son écart type est $s(X) = \sqrt{n\,p\,(1-p)}$.

Procédure de calcul

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

1ère méthode : énumération des cas

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

Dans le cas présent on obtient $\mathbb{P}[1 \leqslant S \leqslant 2]$
$$\begin{aligned} &= \mathbb{P}[S=1] + \mathbb{P}[S=2] \\ &= \binom{3}{1} \times 0{,}4\,(1-0{,}4)^2 + \binom{3}{2} \times 0{,}4^2\,(1-0{,}4) \\ &\simeq 0{,}432 + 0{,}288 \\ &= 0{,}72 \end{aligned}$$

2ème méthode : utilisation de la calculatrice (voir aussi les vidéos mises en ligne sur plubel)

La calculatrice sait calculer les probabilités de la forme $\mathbb{P}[X \leqslant \dots]$. On se ramène donc à une soustraction de probabilités de ce type :
$\mathbb{P}[1 \leqslant S \leqslant 2] = \mathbb{P}[S \leqslant 2] - \mathbb{P}[S \leqslant 0] = 0{,}936 - 0{,}216 = 0{,}72$

Avec cette méthode il faut prendre garde que pour des entiers a et b arbitraires, on a $\mathbb{P}[a \leqslant X \leqslant b] = \mathbb{P}[X \leqslant b] - \mathbb{P}[X \leqslant a-1]$. 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 $\mathbb{P}[X \leqslant a-1]$ (car les valeurs strictement plus petites que a sont toutes les valeurs jusqu'à $a-1$).

3.3.3 Échantillonage

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 28% 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 $X \sim \mathcal{B}(3; 0{,}28)$, 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 28% de chances de « succès ».

En conséquence on a les probabilités suivantes :

Évènement	$X=0$	$X=1$	$X=2$	$X=3$
Probabilité	0,373	0,435	0,169	0,022

On peut constater que $\mathbb{P}[X=1] = 0{,}435$ diffère sensiblement de la probabilité ($\frac{390}{816} \simeq 0{,}477\,9$) obtenue dans l'exemple page 32 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 $\frac{5}{18} \simeq 0{,}278$.

La raison pour laquelle les probabilités diffèrent est que dans l'exemple page 32 on choisissait parmi 18 individus, de sorte qu'une fois le premier individu choisi, il n'y avait plus que 17 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.

3.4 La loi normale

3.4.1 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 3.2 ci-après.

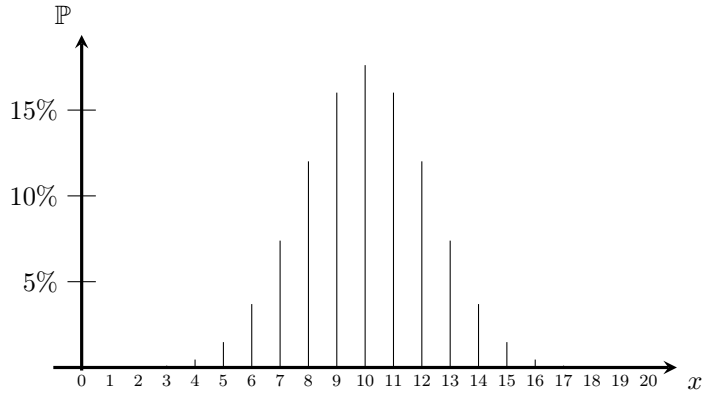


FIGURE 3.2 – Diagramme *en bâtons* pour la loi $\mathcal{B}(20; 0{,}5)$.

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 $\mathbb{P}[6 \leqslant X < 12] = \mathbb{P}[X = 6] + \mathbb{P}[X = 7] + \cdots + \mathbb{P}[X = 11]$, c'est-à-dire la somme des aires des rectangles grisés sur la figure 3.3a 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 3.3b ci-après.

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.

3.4.2 Loi normale

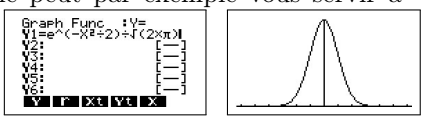
Loi normale centrée réduite Si la loi d'une variable aléatoire Z est donnée par

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

(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 3.3b, 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 $x = 11{,}5$.

  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.

- En pratique, les calculs avec des lois normales se feront en utilisant la calculatrice, car "mesurer l'aire sous une courbe" n'est pas envisageable à la main.

- Dans ce cours, les lettres X, Y, etc, désignent des variables aléatoires, ou des variables statistiques qui dépendent du contexte. On tachera de n'utiliser la lettre Z que pour des variables qui suivent la loi normale centrée réduite.

Loi normale Soit X est une variable aléatoire. S'il existe deux nombres μ et σ et une variable aléatoire Z qui suit la loi normale centrée réduite, tels que

$$X = \mu + Z \cdot \sigma$$

alors on dit que « X suit la loi normale de moyenne μ et d'écart type σ », et on note $X \sim \mathcal{N}(\mu; \sigma)$.

Remarques

- En particulier, la loi normale centrée réduite se note $\mathcal{N}(0; 1)$. En effet, $\mathcal{N}(0; 1)$ signifie qu'on a $X = 0 + Z \cdot 1$, c'est à dire $X = Z$, où Z suit la loi normale centrée réduite.

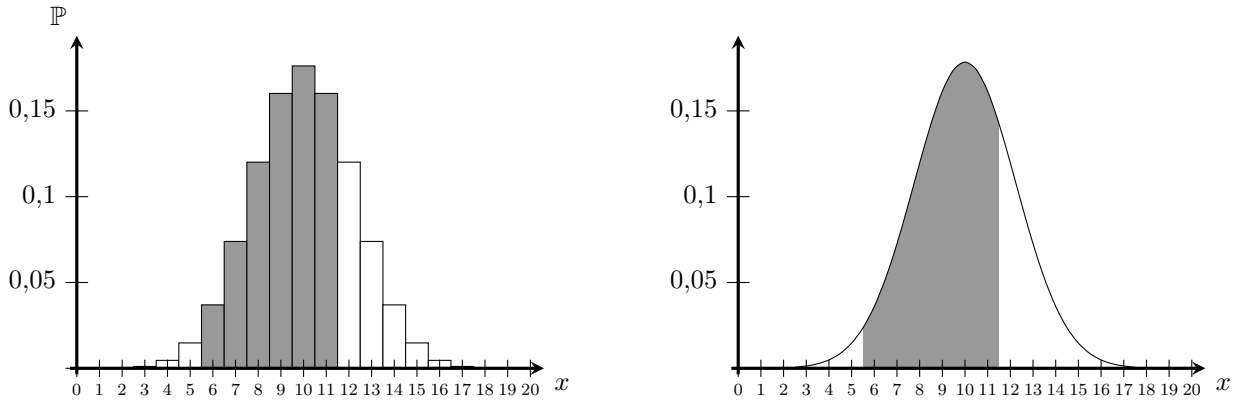
- Cette loi normale (avec une moyenne μ et un écart type σ) s'exprime aussi comme l'aire sous une courbe. Cette fois-ci la courbe (que l'on appelle « densité ») est donnée par la formule $\frac{\exp(-\frac{(x-\mu)^2}{2\sigma^2})}{\sqrt{2\pi\sigma^2}}$. La figure 3.4 montre la différence entre cette densité 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 σ).

- Pour la loi normale, les probabilités sont les mêmes pour des inégalités larges ou strictes, c'est à dire que les quatre probabilités $\mathbb{P}[a \leqslant Z \leqslant b]$, $\mathbb{P}[a \leqslant Z < b]$, $\mathbb{P}[a < Z \leqslant b]$, 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).

3.4.3 Probabilité d'un intervalle

On considère une variable X qui suit la loi $\mathcal{N}(30\ ;\ 10)$. On cherche à déterminer la probabilité $\mathbb{P}[23 \leqslant X \leqslant 35{,}94]$.

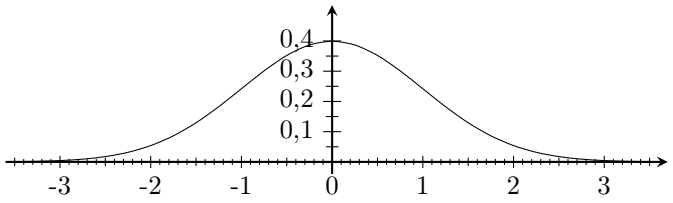
Les vidéos mises en ligne sur plubel indiquent comment procéder sur plusieurs modèles de calculatrices :



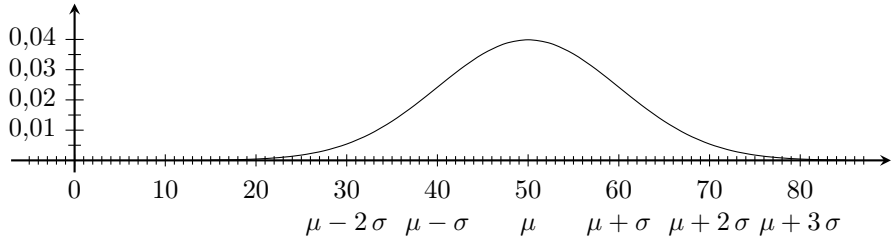
(a) $\mathbb{P}[6 \leqslant X < 12]$ pour $X \sim \mathcal{B}(20; 0{,}5)$.

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

FIGURE 3.3 – Approximation : loi binomiale et loi normale.



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

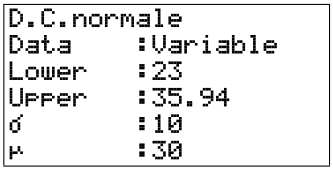


(b) Densité de la loi $\mathcal{N}(\sigma; \mu)$ (la figure est réalisée pour $\mu = 50$, $\sigma = 10$).

FIGURE 3.4 – Comparaison entre les densités de $\mathcal{N}(0; 1)$ et $\mathcal{N}(\mu; \sigma)$.

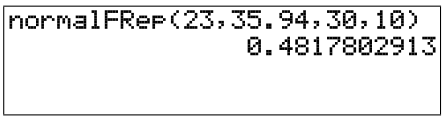
Calcul avec une calculatrice Casio

Dans [MENU], choisir STAT, puis dans DIST, choisir NORM puis [Ncd]. Dans le menu qui s'ouvre, entrer 23 dans la ligne Lower et 35.94 dans la ligne Upper (car on veut calculer $\mathbb{P}[23 \leqslant X \leqslant 35{,}94]$). Entrer enfin 10 dans la ligne σ et 30 dans la ligne μ (car on considère $X \sim \mathcal{N}(30\,;\,10)$). Entrer [CALC] (dans la ligne Execute) pour calculer et afficher la probabilité.



Calcul avec une calculatrice TI

On utilise la fonction normalFRep (ou normalcdf sur les calculettes anglophones). On trouve cette fonction dans le menu DISTR accessible par [2nd] [VARS]. On entre ensuite les valeurs 30, 10, 23, et 35.94 (séparées par des virgules), pour indiquer qu'on calcule $\mathbb{P}[23 \leqslant X \leqslant 35{,}94]$ et que $X \sim \mathcal{N}(30\,;\,10)$. On finit par [)] [ENTER]. La valeur qui s'affiche alors est la probabilité $\mathbb{P}[23 \leqslant X \leqslant 35{,}94]$.



Calcul avec une calculatrice Numworks

On choisit Probabilités dans le menu [⌂], et on y choisit la loi normale. On entre ensuite la valeur 30 pour la moyenne (notée μ par la calculatrice) et la valeur 10 pour l'écart type (noté σ par la calculatrice). On appuie sur suivant et la calculatrice nous demande de quel intervalle on cherche à calculer la probabilité. On lui indique alors P(23⩽X⩽ 35.94). Pour cela on déplace le curseur vers la gauche pour choisir l'icône qui indique que l'on spécifiera le minimum et le maximum de l'intervalle. Il suffit ensuite d'entrer les valeurs 23 et 35.94, et la calculette affiche alors P(23⩽X⩽ 35.94)=0.4817803.

Remarque La plupart des calculatrices calculent la probabilité d'un intervalle dont seraient indiqués le minimum et le maximum. Lorsque l'on doit calculer par exemple $\mathbb{P}[X \geqslant 23]$ ou $\mathbb{P}[X \leqslant 35{,}94]$, 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 $\mathbb{P}[X \geqslant 23]$ on pourra à la place calculer $\mathbb{P}[23 \leqslant X \leqslant 999999999999999]$. De même, pour déterminer $\mathbb{P}[X \leqslant 35{,}94]$ on pourra à la place calculer $\mathbb{P}[-99999999999999 \leqslant X \leqslant 35{,}94]$.

3.4.4 Intervalle de probabilité fixée

Il arrive de se poser la question inverse : trouver un intervalle dont la probabilité est fixée. En général il y a beaucoup d'intervalles qui ont la même probabilité (vous le verrez dans l'exercice 39) et il faut préciser la forme de l'intervalle (par exemple est-ce « $X \leqslant \cdots$ » ou « $X \geqslant \cdots$ » ?

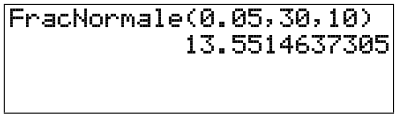
À nouveau, les vidéos mises en ligne sur plubel indiquent comment procéder sur plusieurs modèles de calculatrices.

On va traiter ici l'exemple suivant : « quand $X \sim \mathcal{N}(30\,;\,10)$, trouvez x tel que $\mathbb{P}[X \geqslant x] = 95\%$ ».

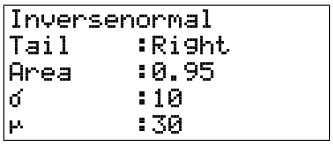
Détermination avec une calculatrice TI

La résolution peut dépendre de votre modèle de calculatrice, mais certaines calculatrices TI ne savent résoudre que les questions avec un $\mathbb{P}[X \leqslant \ldots]$ (par opposition au $\mathbb{P}[X \geqslant \ldots]$ que l'on considère présentement). Si c'est le cas de votre modèle de calculatrice, on cherchera à trouver x tel que $\mathbb{P}[X \leqslant x] = 5\%$ (ce qui revient au même que « $\mathbb{P}[X \geqslant x] = 95\%$ »).

On utilise la fonction `FracNormale` (ou `invNorm` sur les calculettes anglophones). On trouve cette fonction dans le menu `DISTR` accessible par [2nd] [VARS] .
Sur les modèles récents, on a alors un menu comme celui des calculatrices casio (voir ci-après), tandis que sur les modèles plus anciens, on entre les valeurs 0.05 (car on veut $\mathbb{P}[X \leqslant x] = 5\%$), puis 30 et 10 (car on considère $X \sim \mathcal{N}(30\,;\,10)$). On finit par [)] [ENTER] . La valeur qui s'affiche alors est le nombre x tel que $\mathbb{P}[X \leqslant x]$ soit environ égal à 5% (donc tel que $\mathbb{P}[X \geqslant x] = 95\%$).



Détermination avec une calculatrice Casio
Dans [MENU] , choisir `STAT`, puis dans `DIST`, choisir `NORM` puis [InvN].
Dans le menu qui s'ouvre, entrer `Right` dans la ligne `Tail` (pour indiquer qu'on cherche un intervalle de la forme « $X \geqslant \ldots$ »), 0.95 dans la ligne `Area` (car on veut que $\mathbb{P}[X \geqslant x]$ soit égale à 0,95). Entrer enfin 10 dans la ligne σ et 30 dans la ligne μ (car on considère $X \sim \mathcal{N}(30\,;\,10)$).
Entrer [CALC] (dans la ligne `Execute`) pour calculer et afficher la valeur de x.



Calcul avec une calculatrice Numworks

On choisit Probabilités dans le menu [⌂] , et on y choisit la loi normale. On entre ensuite la valeur 30 pour la moyenne (notée μ par la calculatrice) et la valeur 10 pour l'écart type (noté σ par la calculatrice). On appuise sur suivant et on arrive au même écran que précédemment (quand on avait calculé la probabilité d'un intervalle).

On déplace le curseur tout à gauche et on choisit ◺ car on considère une probabilité de la forme $\mathbb{P}[X \geqslant \ldots]$. On déplace ensuite le curseur tout à droite (là où se trouve la valeur d'une probabilité) et on entre 0.95 (pour demander que la probabilité soit 95%). La calculatrice se met à jour et affiche alors le bon intervalle.

3.4.5 Effectifs théoriques

Effectifs théorique Si on considère une population de n individus, on appelle effectif théorique d'un évènement 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 39 : on suppose que $X \sim \mathcal{N}(9{,}58\,;\,3{,}11)$, et si par exemple $n = 1000$ on obtient le tableau ci-dessous :

Désirabilité	$[1\,;\,4[$	$[4\,;\,7[$	$[7\,;\,10[$	$[10\,;\,13[$	$[13\,;\,16[$	$[16\,;\,19[$
Effectif théorique	36,4	167	350,3	310,6	116,2	19,5

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 $[\ 1\ ;\ 4\ [$, on calcule en fait $\mathbb{P}[X \leqslant 4]$. L'évènement « $X < 1$ » est ainsi ajouté artificiellement à cette colonne afin qu'il apparaisse quelque part et que la somme des probabilités fasse bien 100%. On obtient $\mathbb{P}[X \leqslant 4] \simeq 0{,}036$, d'où l'effectif théorique $0{,}036 \times 1\,000 = 36$.
- Pour la dernière classe, on procède de même en ajoutant l'évènement « $X \geqslant 19$ » qui ne serait sinon inclus dans aucune autre colonne. On calcule donc $\mathbb{P}[X \geqslant 16] \simeq 0{,}019$, d'où l'effectif théorique $0{,}019 \times 1\,000 = 19$.
- Pour les classes intermédiaires, on garde l'intervalle écrit dans la première ligne du tableau. Par exemple, pour la 2ème classe, on calcule donc $\mathbb{P}[4 \leqslant X \leqslant 7] \simeq 0{,}17$, d'où l'effectif théorique $0{,}17 \times 1\,000 = 170$.

3.4.6 Approximation de loi binomiale

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

Si $n\,p(1-p) \geqslant 10$, alors $\mathcal{B}(n;p) \approx \mathcal{N}\left(np; \sqrt{np(1-p)}\right)$, mais une correction de continuité est nécessaire sauf si $n\,p(1-p) \geqslant 1000$.

Remarque Cette « *correction de continuité* », signifie que l'on prend en compte le passage d'un diagramme en bâtons (figure 3.2) à un histogramme (figure 3.3a où les bâtons deviennent des rectangles, avec une largeur).

Exemple : Calculer $\mathbb{P}[8 \leqslant X \leqslant 12]$ lorsque $X \sim \mathcal{B}(50; 0{,}3)$.

On remarque tout d'abord que $n\,p\,(1-p) = 10{,}5 > 10$, donc on peut approcher X par la loi $\mathcal{N}(15\,;\,3{,}24)$ (car la calculatrice nous permet de calculer que $50 \times 0{,}3 = 15$ et que $\sqrt{50 \times 0{,}3 \times (1-0{,}3)} \simeq 3{,}24$).

Pour calculer $\mathbb{P}[8 \leqslant X \leqslant 12]$, avec une loi normale, on va en fait calculer $\mathbb{P}[7{,}5 \leqslant X < 12{,}5]$ (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 $\mathbb{P}[7{,}5 \leqslant X < 12{,}5]$ avec la loi normale).

Ainsi, on a $\mathbb{P}[8 \leqslant X \leqslant 12] \simeq \mathbb{P}[7{,}5 \leqslant X < 12{,}5] \simeq 0{,}21$

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).

Très grands échantillons : Lorsque $n\,p(1-p) \geqslant 1000$, il n'est pas nécessaire de faire de correction de continuité. Si on désigne par $P = \frac{X}{n}$ la proportion de « succès », on peut alors aussi approcher P par la loi $\mathcal{N}\left(p; \sqrt{\frac{p(1-p)}{n}}\right)$.

Remarque : L'approximation de la proportion $P = \frac{X}{n}$ par la loi $\mathcal{N}\left(p; \sqrt{\frac{p(1-p)}{n}}\right)$ est en fait pertinente dès que $n\,p\,(1-p) \geqslant 10$, mais quand on doit faire une correction de continuité elle est difficile à faire pour la proportion P. Dans le chapitre suivant, on passera sous silence cette correction de continuité, si bien que les résultats obtenus seront un peu moins précis que dans ce chapitre-ci.

Chapitre 4

Estimation

4.1 « Prédicition » 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 sur 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 $\mathbb{P}\left[-1{,}96 \leqslant Z \leqslant 1{,}96\right] \simeq 0{,}95$.
Comme $\mathbb{P}[-1{,}96 \leqslant Z \leqslant 1{,}96] \simeq 0{,}95$, on considérera que « avec la confiance 95% » la loi normale centré réduite est entre $-1{,}96$ et $1{,}96$.

En Table 4.1, cela se traduit par la troisième colonne, qui indique que pour une confiance $c = 0{,}95$, (c'est à dire un risque d'erreur $\alpha = 0{,}05$), on a un seuil $z_\alpha = 1{,}96$. De même la deuxième colonne de cette figure indique que « avec la confiance 90% », la loi normale centrée réduite est entre $-1{,}645$ et $1{,}645$ (c'est à dire que $\mathbb{P}[-1{,}645 \leqslant Z \leqslant 1{,}645] \simeq 0{,}90$).

confiance: c	0,9	0,95	0,96	0,98	0,99	0,995
risque d'erreur: α	0,1	0,05	0,04	0,02	0,01	0,005
seuil z_α	1,645	1,960	2,054	2,326	2,576	2,807

TABLE 4.1 – Seuils z_α pour différents niveaux de confiance

La figure 4.3a représente graphiquement les intervalles que l'on obtient pour la confiance 95% et pour la confiance 99%.

Obtention de ces seuils avec la calculatrice Pour trouver ces intervalles pour n'importe quel niveau de confiance, commencer par noter qu'on a désigné par α le risque d'erreur et que la confiance est $1-\alpha$ (c'est à dire 100% moins le risque d'erreur). Ce risque d'erreur est la probabilité d'être en dehors de l'intervalle recherché.

On choisit un intervalle où quand on est en dehors il y a autant de chances d'être à droite de l'intervale qu'à gauche, donc pour avoir une probabilité totale α d'être en dehors de l'intervalle, on doit avoir une probabilité $\alpha/2$ d'être en dehors à droite, et $\alpha/2$ d'être en dehors à gauche, comme sur la figure 4.2, qui est déssinée pour l'exemple d'une confiance $c = 0{,}90$.

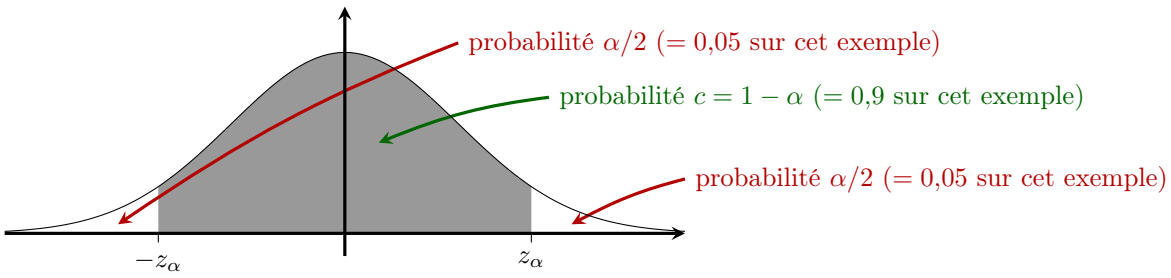


FIGURE 4.2 – Intervalle de probabilité $c = 1 - \alpha$ fixée (dessiné pour l'exemple de $c = 0{,}9$).

Dès lors, on voit sur le dessin que $\mathbb{P}[Z \geqslant z_\alpha] = \alpha/2$ et que $\mathbb{P}[Z \leqslant -z_\alpha] = \alpha/2$. Quand on demande à la calculatrice d'en déduire z_α (comme au pargraphe 3.4.4), on trouve bien les valeurs de la table 4.1. Par exemple, pour la confiance $c = 0{,}9$, si on demande $\mathbb{P}[Z \geqslant z_\alpha] = \alpha/2$ alors on obtient $z_\alpha \simeq 1{,}644854$ (que l'on a arrondi à 1,645 dans la table 4.1.

Avec moyenne et écart type Pour une moyenne μ et un écart type σ arbitraires, on peut faire de même pour une variable aléatoire X qui suive la loi $\mathcal{N}(\mu;\sigma)$. Dans ce cas, $X = \mu + Z\sigma$, où $Z \sim \mathcal{N}(0;1)$.

Avec la probabilitété 95%, on a $-1{,}96 \leqslant Z \leqslant 1{,}96$, c'est à dire $\mu - 1{,}96\sigma \leqslant \underbrace{\mu + Z\sigma}_{X} \leqslant \mu + 1{,}96\sigma$.

On considère donc que X est dans l'intervalle $[\mu - 1{,}96\sigma; \mu + 1{,}96\sigma]$ avec la confiance 95%. Pour d'autres niveaux de confiance, l'intervalle se calcule de même en remplaçant « 1,96 » par le seuil z_α qui correspond à un autre niveau de confiance, et que l'on peut soit lire dans la table 4.1 soit obtenir de la calculatrice comme montré au paragraphe précédent.

La figure 4.3b représente graphiquement ces intervalles que l'on obtient pour la loi $\mathcal{N}(50;10)$, pour la confiance 95% et pour la confiance 99%.

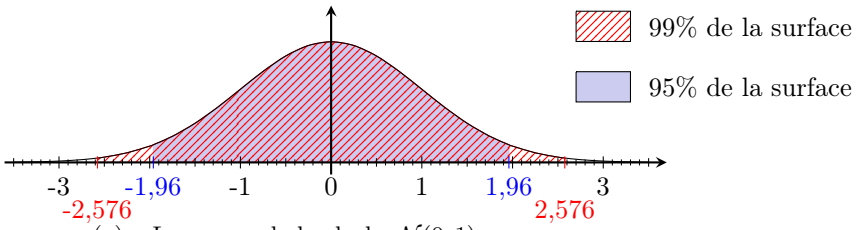
Calculatrices récentes Sur certaines calculatrices récentes, l'intervalle que l'on vient d'introduire s'appelle un « intervalle de fluctuation ». On n'utilisera pas dans ce cours cette notion d'*intervalle de fluctuation*, mais il peut être utile de savoir que sur beaucoup de calculatrices on obtient cet intervalle de la même façon qu'en paragraphe 3.4.4 : sur les calculatrices casio on met `Central` dans la ligne `Tail` (et le niveau de confiance dans la ligne `Area`). Sur les calculatrice numworks, on choisit (à gauche de l'écran qui donne les probabilités pour la loi normale qu'on considère), et on entre à droite de l'écran (à l'emplacement de la probabilité) la valeur de la confiance.

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

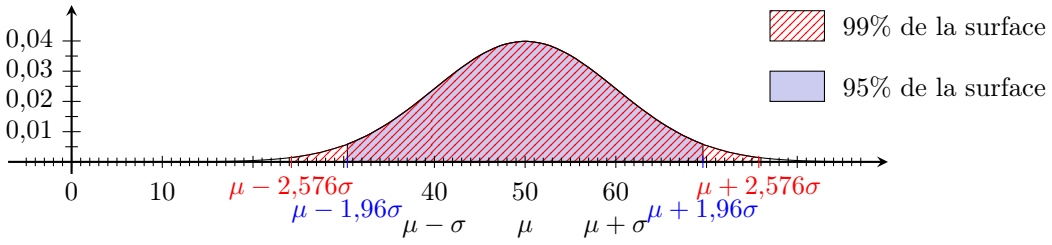
Le plus fréquent est de considérer une confiance de 95%, et c'est que l'on fera dans la plupart des exemples. Toutefois il est utile de savoir procéder pour n'importe quel niveau de confiance, donc lorsque les procédures générales seront énoncées, on décrira comment procéder pour une confiance arbitraire.

4.2 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.



(a) « Largeur » de la cloche $\mathcal{N}(0;1)$.



(b) « Largeur » de la cloche $\mathcal{N}(\sigma;\mu)$ (la figure est réalisée pour $\mu = 50$, $\sigma = 10$).

FIGURE 4.3 – « 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 étudiants sur le campus, 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 $n\,p\,(1-p) = 18 > 10$).

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 autant d'informations que possible 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 $75 \times p \times (1-p) \geqslant 10$ pour pouvoir approximer la loi binomiale ($X \sim \mathcal{B}(75,p)$) par la loi normale $\mathcal{N}(75\,p\,;\,\sqrt{75\times p\times(1-p)})$. 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 4.1, 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 $|P-p| \leqslant 1{,}96s$.

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 stress·e·é 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 $|p_e - p| \leqslant 1{,}96s$, 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 : $s \simeq \sqrt{\frac{0{,}24\,(1-0{,}24)}{75}} \simeq 0{,}049\,32$. Dès lors on déduit que $1{,}96s \simeq 1{,}96 \times 0{,}049 = 0{,}096\,04$, 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 dijonais 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 4 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 ce 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 $n\,p_e\,(1 - p_e) \geqslant 10$. Si cette hypothèse est satisfaite alors ont 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_α 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 :

- $p_e = \frac{18}{75} = 0{,}24$
- $n\,pe\,(1 - p_e) = 75 \times 0{,}24\ (1 - 0{,}24) = 13{,}68 > 10$, donc on peut utiliser cette procédure.
- Pour la confiance 95%, on a $z_\alpha \simeq 1{,}96$, et on obtient $a_\alpha \simeq 1{,}96 \sqrt{\frac{0{,}24\,(1-0{,}24)}{75}} \simeq 0{,}096\,66$, 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_α 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 4 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 $n > z_\alpha^2 \dfrac{1}{4h^2}$ (cela suffit pour être sûr que $a_\alpha < h$).

4.3 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 μ 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 $n \geqslant 30$, alors une approximation par une loi normale donne la procédure indiquée dans le formulaire :
  - On fixe une confiance c
  - La table indique le z_α 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 μ) 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_α (que l'on avait pour les estimations de proportions et pour les grands échantillons) est remplacé par un seuil t_α 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 μ) est dans l'intervalle $[m_e - a_\alpha\,;\, m_e + a_\alpha]$.

La seule différence pratique entre ces procédures est donc que z_α est remplacée par t_α si $n \geqslant 30$.

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 10 pour estimer le nombre moyen de personnes par ménage au sein de la population Française, avec une confiance de 95%. Comme vous l'avez vu lors de l'exercice 16 (et TD), on a $n = 5706$, $m_e = 2{,}91$ et $s_e = 1{,}33$. On applique donc la procédure correspondant au cas où $n \geqslant 30$:
On estime donc que μ est dans l'intervalle $[2{,}91 - 0{,}034\,51\,;\, 2{,}91 + 0{,}034\,51] \simeq [2{,}88\,;\, 2{,}94]$, 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 5706 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 pas ménage dépende du fait que l'on plutôt 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 28 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 28, qui est notée X dans le corrigé de l'exercice 28 mis en ligne.

Comme indiqué dans le corrigé de l'exercice 28, on trouve que $n = 16$, que $m_e = 29{,}81$ et $s_e = 7{,}32$. On va supposer que X suit une loi normale et appliquer la procédure correspondant au cas où $n < 30$:
On cherche t_α dans la table page 11 du formulaire, en ligne 15 (car 16 -1 = 15) et en colonne 0,025 (car $c = 0{,}95$),
On lit $t_\alpha \simeq 2{,}1314$ d'où $a_\alpha = t_\alpha \frac{s_e}{\sqrt{n-1}} \simeq 2{,}1314 \times \frac{7{,}32}{\sqrt{16-1}} \simeq 4{,}028\,4$.
On estime donc que μ est dans l'intervalle $[29{,}81 - 4{,}028\,4\,;\ 29{,}81 + 4{,}028\,4] \simeq [25{,}78\,;\ 33{,}84]$ 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_α correspond à une loi normale, c'est le nombre qu'on aurait utilisé dans la procédure où $n \geqslant 30$.

4.4 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 σ 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 χ^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 χ^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 σ) 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 28, 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 $s_e = 7{,}32$. On suppose que X suit une loi normale et on applique la procédure du formulaire :
On lit dans la table du χ^2 à 15 degrés de liberté la valeur $x_1 = 6{,}262$ correspondant à $q = 0{,}025$ et la valeur $x_2 = 27{,}49$ correspondant à $p = 0{,}025$
On obtient alors $s_e\sqrt{\frac{n}{x_2}} \simeq 7{,}32\sqrt{\frac{16}{27{,}49}} \simeq 5{,}58$ et $s_e\sqrt{\frac{n}{x_1}} \simeq 7{,}32\sqrt{\frac{16}{6{,}262}} \simeq 11{,}7$
On estime donc que σ est dans l'intervalle $[5{,}58; 11{,}7]$ avec la confiance $c = 0{,}95$.

4.5 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 28, 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 [25,78 ; 33,84] (avec la confiance 95%). On pourrait de même estimer ce qu'il en est après la thérapie et on obtiendrait l'intervalle [13,24 ; 26,64], 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 [25,78 ; 33,84] et [13,24 ; 26,64] se chevauchent. Il serait possible que le niveau d'anxiété moyen soit plus élevé avant la thérapie (par exemple s'il est de 31,15 avant la thérapie et de 17,71 après la thérapie) ou bien qu'il soit plus élevé après la thérapie (par exemple s'il est de 26,07 avant la thérapie et de 26,35 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 28, on obtient :

Avant	28	42	42	15	40	31	27	21	36	25	26	34	23	27	30	30
Après	7	41	25	4	44	32	16	15	31	18	9	31	6	17	10	13
Différence	21	1	17	11	−4	−1	11	6	5	7	17	3	17	10	20	17

On a alors $n = 16$, et on peut calculer que la moyenne de la différence (au sein de l'échantillon) est $m_e = 9{,}88$ tandis que son écart type est $s_e = 7{,}57$.

Cela permet de faire une estimation de la moyenne de la différence :
On lit $t_\alpha \simeq 2{,}1314$ d'où $a_\alpha = t_\alpha \frac{s_e}{\sqrt{n-1}} \simeq 2{,}1314 \times \frac{7{,}57}{\sqrt{16-1}} \simeq 4{,}166$.

On estime donc que μ est dans l'intervalle $[9{,}88 - 4{,}166\,;\, 9{,}88 + 4{,}166] \simeq [5{,}71\,;\, 14{,}05]$ 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%).