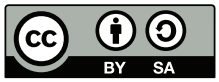
Statistiques en L1 de psychologie

(première moitié du cours)

Sébastien Leurent

Année 2019-2020



Introduction

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

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

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

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

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

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

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

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

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

Déroulement du semestre

Les présentes notes de cours[1] correspondent aux chapitres 1 à 3 uniquement. Elles sont disponibles à l'adresse `http://leurent.perso.math.cnrs.fr/stats_ps1/2019-2020/coursA.pdf`. L'adresse `http://leurent.perso.math.cnrs.fr/stats_ps1/` permet aussi d'accéder aux transparents de cours, feuilles d'exercice, formulaire, etc.

À chaque séance, il vous est demandé d'avoir le formulaire distribué en début de semestre, ainsi qu'une calculatrice scientifique (par exemple, les modèles «Graph 35+ USB» de Casio et «TI-83» sont parfaitement adaptés). 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
- Une note de contrôle continue (CC) est obtenue au cours du semestre, et constitue l'autre moitié de la note de l'UE. Elle n'est pas rattrapable en deuxième session et est constituée
  - pour moitié d'un contrôle commun à tous les groupes de TD, a le 10 mars 2020 à 17h15.
  - pour moitié d'une note attribuée au sein de chaque groupe de TD, à partir de note(s) de contrôle(s) mais qui peut aussi prendre en compte la participation de chaque étudiant.

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

Rappels et compléments de mathématiques

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

Rappels

Règles de calcul

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

$$-x + \frac{30}{3+2} + 3(-4+9) - 8^2$$

Notation

- Le symbole « x » désigne un nombre arbitraire, avec lequel on peut procéder à des calculs même si on ne connaît pas sa valeur.
- Le symbole « $+$ » désigne l'addition
- Le symbole « - » peut soit désigner une soustraction, soit « L'opposé » d'un nombre :
  - par exemple le nombre négatif -4 est l'opposé de 4
  - de même $-x$ désigne l'opposé de « x »
- La multiplication est notée par le symbole $\times$, ou par le symbole « $*$ » sur les calculettes. Parfois, on n'écrit pas du tout la multiplication et le lecteur doit « deviner » sa présence, afin de donner un sens à une formule mathématique.

  Par exemple l'expression « $3(-4+9)$ » n'aurait aucun sens si on n'ajoute pas une multiplication, de sorte que l'on comprends qu'elle signifie en fait « $3 \times (-4+9)$ ».
- Les divisions sont notés indifféremment par le symbole « $\div$ » (souvent utilisé par les calculatrices Casio), le symbole « / » (souvent utilisé par les calculatrices TI) ou un trait de fraction.

  Le trait de fraction présente l'avantage d'une plus grande lisibilité en évitant d'avoir à écrire explicitement certaines parenthèses. Ainsi, l'expression $\frac{30}{3+2}$ signifie $30 \div (3+2)$ (ou $30/(3+2)$ 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*.
- 8^2 désigne le carré de 8. Sur certaines calculatrices ce carré est noté 8^2.

Priorités de calcul

Pour calculer une expression comme $-x + \frac{30}{3+2} + 3(-4+9) - 8^2$, on effectue les étapes suivantes :

1. On calcule tout ce qui est dans des parenthèses, ou au numérateur ou dénominateur d'une fraction. Pour l'expression $-x + \frac{30}{3+2} + 3(-4+9) - 8^2$, on calcule donc $3+2=5$ et $-4+9=5$. On obtient donc
$$-x + \frac{30}{3+2} + 3(-4+9) - 8^2 = -x + \frac{30}{5} + 3 \times 5 - 8^2$$
2. On calcule ensuite les puissances (ici le carré). On calcule donc $8^2 = 64$ et on obtient
$$-x + \frac{30}{5} + 3 \times 5 - 8^2 = -x + \frac{30}{5} + 3 \times 5 - 8^2$$
3. On effectue ensuite les multiplications et les divisions : on calcule donc $\frac{30}{5} = 6$ et $3 \times (5) = 15$. On obtient donc
$$-x + \frac{30}{5} + 3 \times 5 - 8^2 = -x + 6 + 15 - 64$$
4. Enfin, on termine par l'addition :
$$-x + 6 + 15 - 64 = -x - 43$$

Pour cet exemple on obtient donc, à l'issue des calculs, que $-x + \frac{30}{3+2} + 3(-4+9) - 8^2 = -x - 43$.

Ces étapes s'effectuent toujours dans cet ordre (parenthèses, puissances, multiplications et enfin additions). On peut en pratique s'en remettre aux calculatrices ou ordinateurs, qui connaissent cet ordre, mais lorsqu'on veut manipuler soi même ce type d'expressions sans erreurs de parenthèses, il est utile de connaître ces règles de priorité, qui signifient par exemple que $5 \times 2 + 8$ est égal à $(5 \times 2) + 8$ et pas à $5 \times (2 + 8)$.

Autres règles de calcul

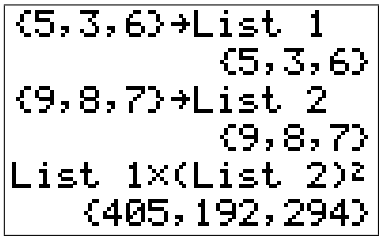
- Lorsqu'on multiplie quelque chose par une somme (c'est à dire une addition ou une soustraction) on peut « développer ». Par exemple cela signifie que
$$3 \times (5 - 8) = 3 \times 5 - 3 \times 8$$
- On peut changer l'ordre des termes dans une addition ou une soustraction (mais dans une soustraction, il faut faire attention aux signes). De même on peut changer l'ordre des facteurs dans une multiplication ou une division. Par exemple, cela signifie que $9 - 5 + 3 = 9 + 3 - 5$, et que $\frac{2}{3} \times 8 = \frac{8 \times 2}{3}$.
- Lorsqu'on multiplie une fraction par une expression, on multiplie simplement le numérateur par cette expression. Lorsqu'on divise une fraction par une expression, on multiplie simplement le dénominateur par cette expression.
- On peut ajouter et soustraire la même quantité sans modifier la valeur d'une expression. On peut de même, sans changer la valeur d'une expression, la multiplier et la diviser par la même quantité. Ainsi, on a par exemple $8 + x - 8 = x$, et $\frac{8 \times 6}{8} = 6$.

Calculatrice

Si vous n'êtes pas habitués à utiliser la calculatrice, vous êtes très vivement encouragés à faire les premiers exercice du chapitre de la feuille d'exercice. Un corrigé de ces ces exercices (et d'une partie des autres exercices) est disponible à l'adresse `http://leurent.perso.math.cnrs.fr/stats_ps1/2019-2020/Exercices_corriges.pdf`.

Ayez bien en tête que

- Sur certaines calculatrices (en particulier les calculatrices 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 :
  - `6.9E13` signifie $6{,}9 \times 10^{13}$ (c'est à dire $6{,}9 \times 10000000000000$). Quand on multiplie 6,9 par 10^{13}, on décale la virgule 13 fois vers la droite, aboutissant à 69000000000000.
  - `6.9E-11` signifie $6{,}9 \times 10^{-11}$ (c'est à dire $6{,}9 \times 0{,}00000000001$). Quand on multiplie 6,9 par 10^{-11}, on décale la virgule 11 fois vers la gauche, aboutissant à 0,000000000069.
- La calculette peut garder en mémoire le résultat de certains calculs, pour éviter des erreurs d'arrondis en recopiant des valeurs arrondies. D'une part, « Ans » désigne le résultat du dernier calcul. D'autre part, « → » (tapé avec [→] ou [Sto] selon les calculatrices) permet d'enregistrer le résultat d'un calcul. Les exercice 2 et 5 en donnent des exemples d'utilisation.
- La calculatrice peut aussi garder en mémoire plusieurs valeurs à la fois, sous la forme d'une liste, afin d'effectuer des opérations simultanées sur chaque valeur de la liste et/ou d'en calculer (entre autre la somme).
  - On peut créer des listes soit avec « → » (comme ci-contre), soit avec l'éditeur de liste (accessible depuis [MENU] sur Casio, ou depuis [STAT] sur TI).
  - Lorsqu'on fait des opérations sur les listes (comme les multiplier par des nombres ou même par d'autres listes), on réalise en fait l'opération sur chaque élément de la liste. Par exemple, la capture d'écran ci-contre, l'expression `List 1×(List 2)²` désigne la liste dont les éléments sont 5×9^2, 3×8^2 et 6×7^2 (c'est à dire que pour calculer `List 1×(List 2)²`, on a multiplié à chaque fois un élément de `List 1` par le carré d'un élément de `List 2`)



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. Par exemple, on peut interroger des étudiants (typiquement au moyen d'un questionnaire) et ainsi connaître leur nombre de frères et soeurs, leur taille, la couleur de leur 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"). Le tableau ci-dessous indique alors quels calculs font sens pour chacun de ces types de données :

	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 2019/2020 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

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 de 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 :

Guillaume
de bonne humeur 0 frère/soeur 1m82

Genevieve
de bonne humeur 1 frère/soeur 1m74

Alain
de mauvaise humeur 1 frère/soeur 1m71

Justine
de très bonne humeur 1 frère/soeur 1m69

Xavier
de bonne humeur 2 frères/soeurs 1m94

Serge
de relativement bonne humeur 0 frère/soeur 1m85

Gérard
de bonne humeur 0 frère/soeur 1m83

Marie
de mauvaise humeur 1 frère/soeur 1m72

Laurent
de bonne humeur 0 frère/soeur 1m86

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

1.2.1 Regroupement par modalités

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

Modalité	de mauvaise humeur	de relativement bonne humeur	de bonne humeur	de très bonne humeur
Effectif	2	1	5	1

(a) Humeur

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

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 (1.2) 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	[1,65 ; 1,70[	[1,70 ; 1,75[	[1,75 ; 1,80[	[1,80 ; 1,85[	[1,85 ; 1,90[	[1,90 ; 1,95[
Effectif	1	3	0	2	2	1

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

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

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

Notation

On désigne par r le nombre de colonnes.
On désigne par a_1 le minimum de la classe de la première colonne,
par a_2 le minimum de la classe de la deuxième colonne,
...
par a_r le minimum de la classe de la dernière colonne,
par a_{r+1} le maximum de la classe de la dernière colonne.
De même on désigne par n_1 l'effectif de la première colonne,
par n_2 l'effectif de la deuxième colonne,
...
par n_r l'effectif de la dernière colonne.

Exemple
Dans la table 1.2, on a $a_1 = 1{,}65$, $a_2 = 1{,}70$, $a_3 = 1{,}75$, $a_4 = 1{,}80$, $a_5 = 1{,}85$, $a_6 = 1{,}90$, $a_7 = 1{,}95$; $n_1 = 1$, $n_2 = 3$, $n_3 = 0$, $n_4 = 2$, $n_5 = 2$, et $n_6 = 1$.
Il y a 6 colonnes, d'où $r = 6$.
L'effectif total est $n = 1+3+2+2+1 = 9$

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 $f_3 = \frac{n_3}{n} = \frac{1}{9} \simeq 0{,}111$. Cela signifie que $\mathbb{P}_r[X = 2] \simeq 0{,}111$, c'est à dire que parmi cet échantillon, il y a environ 0,111% des étudiants qui ont exactement 2 frères et soeurs.

Remarque Dans ce cours, on utilisera simplement le terme « fréquence » pour désigner les fréquences relatives.

Humeur	mauvaise	relativement bonne	bonne	très bonne
Effectif	2	1	5	1
Fréquence	0,222	0,111	0,556	0,111
Fréquence cumulée	0,222	0,333	0,889	1

(a) Humeur

Nombre de frères/soeurs	0	1	2
Effectif	4	4	1
Fréquence	0,444	0,444	0,111
Fréquence cumulée	0,444	0,888	0,999

(b) Nombre de frères et soeurs

Taille	[1,65 ; 1,70[	[1,70 ; 1,75[	[1,75 ; 1,80[	[1,80 ; 1,85[	[1,85 ; 1,90[	[1,90 ; 1,95[
Effectif	1	3	0	2	2	1
Fréquence	0,111	0,333	0,0	0,222	0,222	0,111
Fréquence cumulée	0,111	0,444	0,444	0,666	0,888	0,999

(c) Taille

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

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 $f_1+f_2+f_3 =$

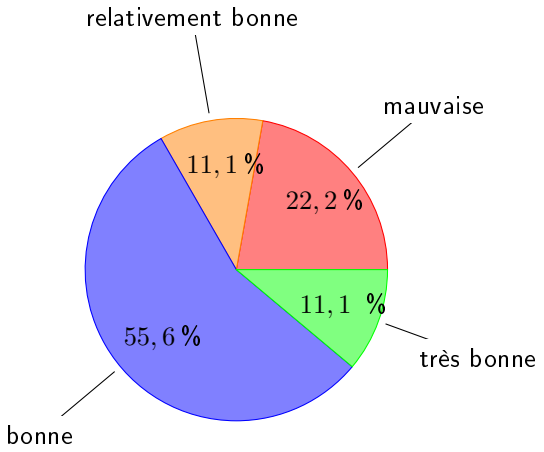


FIGURE 1.1: Humeur des étudiants : représentation en diagramme circulaire ("camembert")

$0{,}444 + 0{,}444 + 0{,}111 = 0{,}999$. Cela signifie que $\mathbb{P}_r[X \leqslant 2] \simeq 0{,}999$, c'est à dire que parmi cet échantillon, il y a environ 99,9 % des étudiants qui ont au maximum 2 frères et soeurs.

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

Par exemple pour la table 1.3c, la fréquence cumulée de la deuxième colonne est $f_1 + f_2 = 0{,}111 + 0{,}333 = 0{,}444$. Cela signifie que $\mathbb{P}_r[Y < 1{,}75] \simeq 0{,}444$, où l'on note bien que cette fréquence cumulée correspond à la taille "1m75".

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.

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

  Dans le cas présent, l'histogramme met en évidence que la taille de ces étudiants 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.

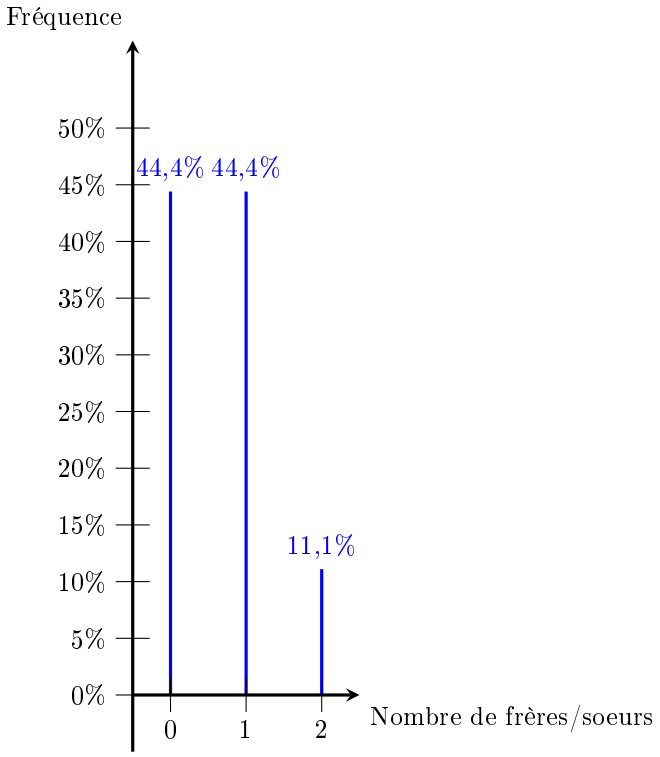


FIGURE 1.2: Nombre de frères et soeurs : représentation en bâtons

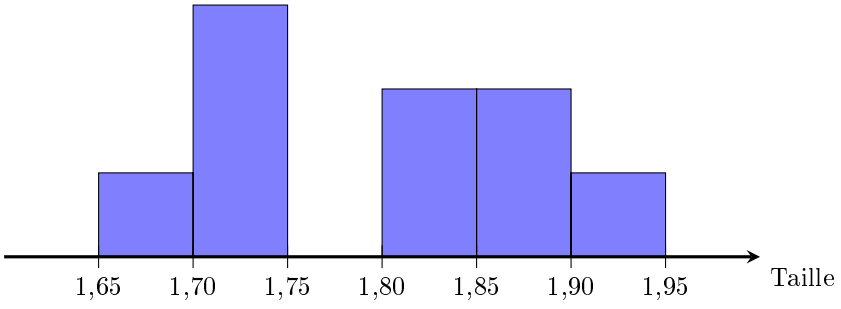


FIGURE 1.3: Histogramme des fréquences, pour la taille des étudiants de l'échantillon

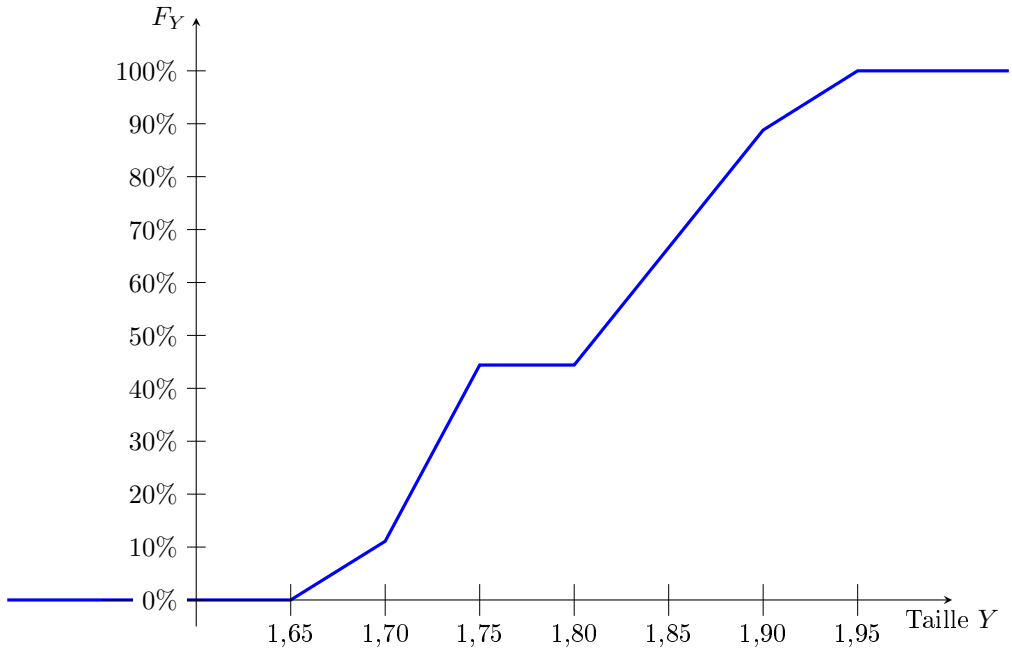


FIGURE 1.4: Polygone des fréquences cumulées, pour la taille des étudiants de l'échantillon

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 1.4 représente ce polygone des fréquences cumulées pour la taille des étudiants de l'échantillon.

Décrivons la façon dont il est construit :

- La taille des étudiants est notée Y, donc on considère la fonction de répartition F_Y définie par $F_Y(a) = \mathbb{P}_r[Y \leqslant a]$.
- Les fréquences cumulées calculées indiquent que $\mathbb{P}_r[Y \leqslant 1{,}7] \simeq 0{,}111$, $\mathbb{P}_r[Y \leqslant 1{,}75] \simeq 0{,}333$, etc,
  c'est à dire que $F_Y(1{,}7) \simeq 0{,}111$, $F_Y(1{,}75) \simeq 0{,}333$, etc,
  donc on place des points de coordonnées $(1{,}7; 0{,}111)$, $(1{,}75; 0{,}333)$, 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 $\mathbb{P}_r[Y \leqslant 1{,}65] = 0$, donc on ajoute un point de coordonnées $(1{,}65; 0)$
- Entre ces points on manque d'information pour approximer la fonction F_Y. On choisit de relier ces points par des segments de droites.
- Lorsque $t \leqslant 1{,}65$, on a $\mathbb{P}_r[Y \leqslant t] = 0$. En conséquence on trace une demi-droite horizontale à gauche du point de coordonnées $(1{,}65; 0)$.
- De même, lorsque $t > 1{,}95$, on a $\mathbb{P}_r[Y \leqslant t] = 1$. En conséquence on trace une demi-droite horizontale à droite du point de coordonnées $(1{,}95; 1)$.

1.4 Calcul d'indicateurs

1.4.1 Médiane

Définition La médiane est définie si X est une variable numérique ou ordinale. C'est une modalité notée 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 M R B B B B B 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 cumulées croise la droite d'ordonnée 50%. Comme indiqué en figure 1.5, on lit que la taille médiane est environ $1{,}813m$.
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 [1,8 ; 1,85 [. On note donc $a_i = 1{,}8$ et $a_{i+1} = 1{,}85$. On a donc $F_Y(a_i) = 0{,}444$ et $F_Y(a_{i+1}) = 0{,}666$, d'où $\text{Med} \simeq 1{,}8 + \dfrac{1{,}85 - 1{,}8}{0{,}666 - 0{,}444}\,(0{,}5 - 0{,}444) \simeq 1{,}813$.

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

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

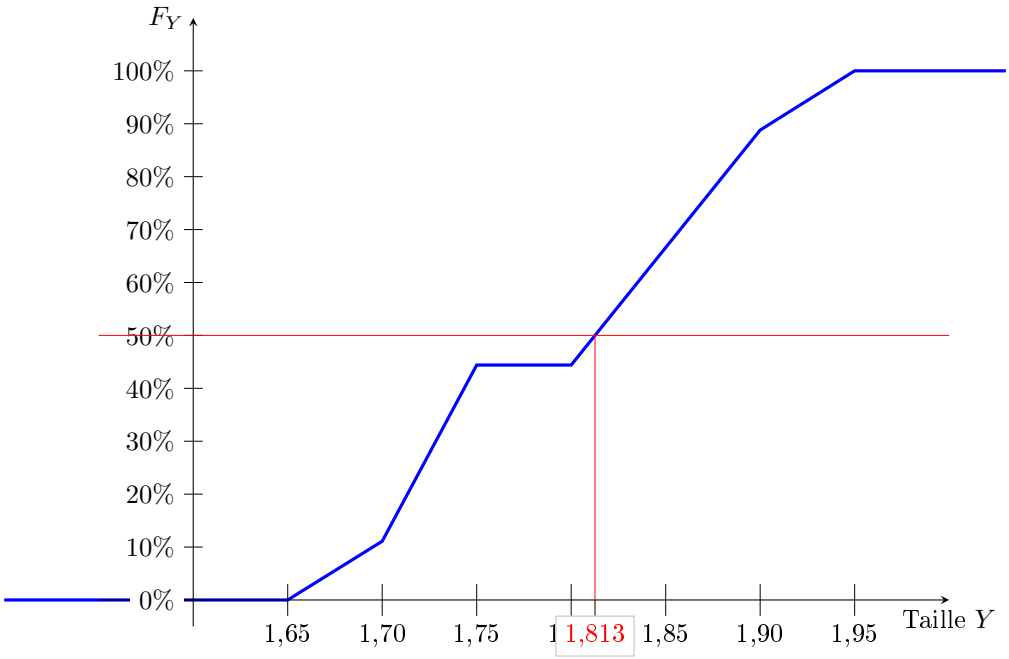


FIGURE 1.5: Lecture graphique de la taille médiane des étudiants de l'échantillon

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 $x_1 = 0$, $x_2 = 1$, $x_3 = 1$, $x_4 = 1$, $x_5 = 2$, $x_6 = 0$, $x_7 = 0$, $x_8 = 1$ et $x_9 = 0$. On a donc $m(X) = \frac{0+1+1+1+2+0+0+1+0}{9} \simeq 0{,}667$.

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

La formule s'écrit donc $m(X) = \frac{4\times 0+4\times 1+1\times 2}{9} \simeq 0{,}667$.

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 $c_1 = \frac{1{,}65+1{,}7}{2} = 1{,}675$, $c_2 = \frac{1{,}7+1{,}75}{2} = 1{,}725$, etc. En conséquence la formule devient $m(Y) \approx \frac{1{,}675+3\times1{,}725+0\times1{,}775+2\times1{,}825+2\times1{,}875+1{,}925}{9} \simeq 1{,}797$.

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 $m(Y) = \frac{1{,}82+1{,}74+1{,}71+1{,}69+1{,}94+1{,}85+1{,}83+1{,}72+1{,}86}{9} \simeq 1{,}796$.

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

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

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

Exemple Pour le nombre de frères et soeurs, on obtient
$m(X^2) = \frac{4\times(0^2)+4\times(1^2)+1\times(2^2)}{9} \simeq 0{,}889$.
D'où $\mathrm{Var}(X) \simeq 0{,}889 - (0{,}667)^2 \simeq 0{,}444$ et $s(X) \simeq \sqrt{0{,}444} \simeq 0{,}666$.

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 37 à 86 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	74	86	49	72	74	53	50	45	43	37	86	56	66	51	48
Y	34	21	27	21	23	39	37	48	59	47	22	51	34	25	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

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

Exemples

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

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

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

	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 manipulable 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 **correlation** 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 correlation

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.

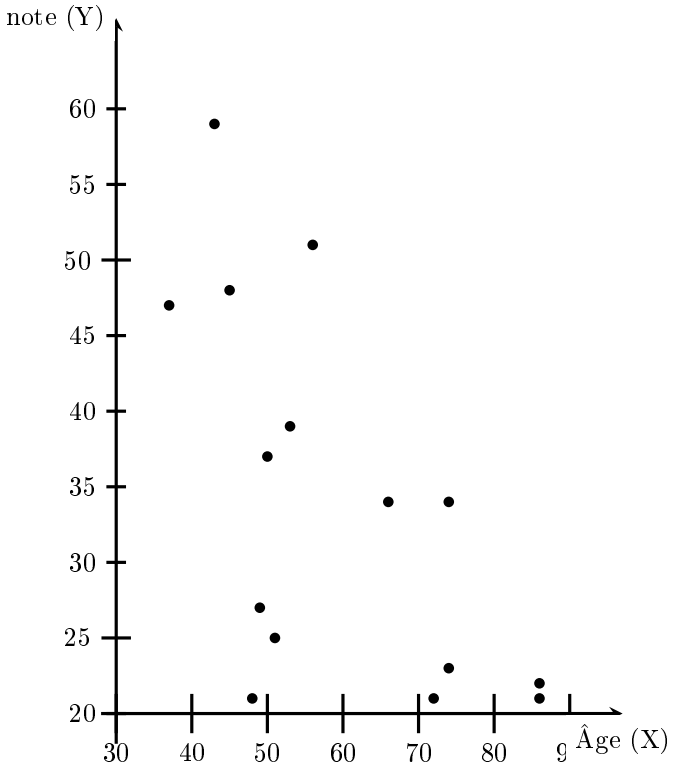


FIGURE 2.1: Nuage de points indiquant la note en fonction de l'âge

2.3.1 Coefficient de corrélation linéaire

Définition

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

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: $m(X) = \frac{\sum x_i}{n} = \frac{74+86+\cdots+48}{15} = \frac{890}{15} \simeq 59{,}33$
$m(X^2) = \frac{\sum x_i^2}{n} = \frac{74^2+86^2+\cdots+48^2}{15} = \frac{56278}{15}$
$Var(X) = m(X^2) - m(X)^2 = \frac{56278}{15} - \left(\frac{890}{15}\right)^2 \simeq 231{,}42$
Écart-type: $s(X) = \sqrt{Var(X)} \simeq 15{,}21$

moyenne: $m(Y) = \frac{\sum x_i}{n} = \frac{34+21+\cdots+21}{15} = \frac{509}{15} \simeq 33{,}93$
$m(Y^2) = \frac{\sum x_i^2}{n} = \frac{34^2+21^2+\cdots+21^2}{15} = \frac{19487}{15}$
$Var(Y) = m(Y^2) - m(Y)^2 = \frac{19487}{15} - \left(\frac{509}{15}\right)^2 \simeq 147{,}66$
Écart-type: $s(Y) = \sqrt{Var(Y)} \simeq 12{,}15$

$m(XY) = \frac{\sum x_i\, y_i}{n} = \frac{74\times34+86\times21+\ldots+48\times21}{15} = \frac{28487}{15} \simeq 1\,899{,}133$
$Cov(X,Y) = m(XY) - m(X)\,m(Y) = \frac{28487}{15} - \frac{890}{15}\frac{509}{15} \simeq -114{,}244$
$r(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} = \frac{-114{,}244}{\sqrt{231{,}422\times147{,}662}} \simeq -0{,}618$

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 $10^{\text{ème}}$ individu, X prend la valeur 37, qui est le plus petite valeur. On y associe le rang 1, et on pose $x'_{10} = 1$.
- Chez le $9^{\text{ème}}$ individu, X prend la valeur 43, qui est la deuxième plus petite valeur. On y associe le rang 2 (pour « deuxième plus petite valeur »), et on pose $x'_9 = 2$.
- De même la troisième plus petite valeur est 45, qui correspond au $8^{\text{ème}}$ individu, de sorte que l'on pose $x'_8 = 3$.
- On continue en posant $x'_{15} = 4$, $x'_3 = 5$, $x'_7 = 6$, $x'_{14} = 7$, $x'_6 = 8$, $x'_{12} = 9$, $x'_{13} = 10$ et $x'_4 = 11$.
- La valeur suivante est 74 qui correspond aux individus numérotés 1 et 5. On devrait attribuer le rang 12 à l'un et 13 à l'autre, mais pour éviter de choisir arbitrairement auquel des deux attribuer le rang 12, on décide de leur attribuer à tous les deux le rang 12,5.

  On pose donc $x'_1 = 12{,}5$ et $x'_5 = 12{,}5$.
- Se souvenant que les rangs que l'on vient d'attribuer correspondent à 12 et à 13, on reprend au rang 14, qui correspond à la valeur 86. Or deux individus ont cette valeur : le $2^{\text{ème}}$ et le $11^{\text{ème}}$ individu. On devrait leur attribuer les rangs 14 et 15, mais comme précédemment, on leur attribue le rang 14,5.

  On pose donc $x'_2 = 14{,}5$ et $x'_{11} = 14{,}5$.
- Ce procédé a ainsi attribué à chaque individu un rang x'_i qui ordonne les individus en fonction de la valeur de X. On a ainsi obtenu la table 2.1.

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

Individu	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Âge X	74	86	49	72	74	53	50	45	43	37	86	56	66	51	48
Rang x'_i	12,5	14,5	5	11	12,5	8	6	3	2	1	14,5	9	10	7	4

TABLE 2.1: Calcul des rangs pour la variable X

Âge X	74	86	49	72	74	53	50	45	43	37	86	56	66	51	48
Note Y	34	21	27	21	23	39	37	48	59	47	22	51	34	25	21
Rang x'_i	12,5	14,5	5	11	12,5	8	6	3	2	1	14,5	9	10	7	4
Rang y'_i	8,5	2	7	2	5	11	10	13	15	12	4	14	8,5	6	2

TABLE 2.2: Rangs des deux variables X et Y

Coefficient de correlation 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 $r_s(X;Y) \approx 1 - \left(6\frac{(12.5-8.5)^2+(14.5-2)^2+(5-7)^2+\cdots+(4-2)^2}{15(15^2-1)}\right) \simeq -0{,}555$.

Organisation du calcul

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

Âge (X)	74	86	49	72	74	53	50	45	43	37	86	56	66	51	48	
Note (Y)	34	21	27	21	23	39	37	48	59	47	22	51	34	25	21	
Rang x'_i	12,5	14,5	5	11	12,5	8	6	3	2	1	14,5	9	10	7	4	
Rang y'_i	8,5	2	7	2	5	11	10	13	15	12	4	14	8,5	6	2	*Total*
$(x'_i - y'_i)^2$	16	156,25	4	81	56,25	9	16	100	169	121	110,25	25	2,25	1	4	871

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

Interpretation

- Le coefficient de corrélation des rangs de Spearman est toujours entre -1 et $+1$. Plus il est proche de zéro moins cela traduit de lien entre les variables. Plus il est proche de 1 ou -1, plus cela indique un fort lien entre les deux variables.

- Si le coefficient de corrélation de Spearman est supérieur à 0,75 ou inférieur à $-0{,}75$, cela traduit un fort lien entre les deux variables, c'est à dire que le nuage de points est concentré le long d'une courbe, qui peut être une droite ou bien avoir une autre forme.
- S'il est entre 0,5 et 0,75, ou entre $-0{,}75$ et $-0{,}5$, cela traduit déjà un lien entre les deux variables, même s'il n'est pas très fort.

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

Remarque

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

2.4 Droites de régression

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

2.4.1 Existence de deux droites distinctes

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

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

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

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

2.4.2 Équation des deux droites

Détermination de Y à partir de X

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

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

Exemple
Pour estimer la note d'une personne de 80 ans, on calcule cette droite :
on pose $a = \frac{Cov(X,Y)}{Var(X)} \simeq \frac{-114,244}{231,42} \simeq -0{,}494$ et $b = m(Y) - a\,m(X) \simeq 33{,}933 - (-0{,}494) \times 59{,}333 \simeq 63{,}244$
D'où l'équation de la droite $D_{Y|X} : Y = -0{,}494\,X + 63{,}244$
Donc pour $x = 80$, on s'attend à $y = -0{,}494 \times 80 + 63{,}244 = 23{,}724$.

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 $a' = \frac{Cov(X,Y)}{Var(Y)} \simeq \frac{-114,244}{147,66} \simeq -0{,}774$ et $b' = m(X) - a'\,m(Y) \simeq 59{,}333 - (-0{,}774) \times 33{,}933 \simeq 85{,}597$
D'où l'équation de la droite $D_{X|Y} : X = -0{,}774\,Y + 85{,}597$
Donc pour $y = 20$, on s'attend à $x = -0{,}774 \times 20 + 85{,}597 \simeq 70{,}117$.

2.4.3 Remarque

Une telle régression linéaire est d'autant plus pertinente que le coefficient de correlation 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 chifres 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	$\frac{1}{16} \simeq 0{,}063$	$\frac{2}{16} = 0{,}125$	$\frac{3}{16} \simeq 0{,}188$	$\frac{4}{16} = 0{,}25$	$\frac{3}{16} \simeq 0{,}188$	$\frac{2}{16} = 0{,}125$	$\frac{1}{16} \simeq 0{,}063$

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

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

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

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

EFMP	EFPM	FEMP	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 $\frac{4}{24} \simeq 0{,}167$.
2. Une autre méthode consisterait à changer de point de vue sur les « cas » qu'il faut lister. On pourrait considérer qu'un « cas » est donnée par l'ensemble des deux premières personnes dessinées (sans se soucier de l'ordre). Si l'on procède ainsi, on obtient la liste ci-dessous :

   EF EM EP FM FP MP

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

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

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

Nombre de cas

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

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

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

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 15 personnes parmi lesquelles 4 fument quotidiennement. Si on choisit au hasard 3 personnes parmi ce groupe de 15, quelle est la probabilité de choisir un fumeur et deux non-fumeurs ?

- On commence par compter le nombre de cas possibles : il y a $\binom{15}{3}$ choix possible de 3 personnes au sein de l'échantillon. La calculette nous indique que $\binom{15}{3} = 455$ donc il y a 455 cas.
- On compte ensuite le nombre de cas où on a un fumeur et deux non-fumeurs, sachant que le groupe de 15 personne comptait 4 fumeurs et 11 non-fumeurs. Il y a donc $\binom{11}{2}$ choix possibles pour choisir deux non-fumeurs, et pour chacun de ces choix, il y a 4 choix d'un fumeur (pour compléter et obtenir trois personnes). Ainsi le nombre de cas qui nous intéresse est $4\times\binom{11}{2}$. Une calculette nous indique que cela vaut 220.
- On conclue que la probabilité recherchée vaut $\frac{220}{455} \simeq 0{,}484$.

3.3 Loi binomiale : répétition d'une expérience

3.3.1 Exemple

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

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

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

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

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

  On a de plus calculé la probabilité de chaque cas : par exemple pour être dans le cas « AGA » il faut à la fois n'avoir ressenti aucun gêne le premier jour (celà a 60% de chances d'arriver) puis en avoir ressenti le second jour (celà a 40% de chances d'arriver) et enfin ne pas en ressentir le troisième jour (celà a 60% de chances d'arriver). Ainsi l'évènement noté « AGA » a la probabilité $0{,}6\times 0{,}4\times 0{,}6$.
- On calcule $\mathbb{P}[X=0]$: comme l'évènement « $X=0$ » correspond précisément au cas « AAA », sa probabilité est $0{,}6\times 0{,}6\times 0{,}6 = 0{,}216$.

- On calcule $\mathbb{P}[X = 1]$: comme l'évènement « $X = 1$ » correspond aux trois cas « AAG », »AGA » et « GAA », sa probabilité est $0{,}6 \times 0{,}6 \times 0{,}4 + 0{,}6 \times 0{,}4 \times 0{,}6 + 0{,}4 \times 0{,}6 \times 0{,}6 \simeq 0{,}432$.

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

En conséquence on obtient que la loi de X est :

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

3.3.2 Formule générale et définition

Le calcul effectué ici sur un exemple peut 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

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

2ème méthode : utilisation de la calculatrice

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 :

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

Avec cette méthode il faut prendre garde que pour des entiers a et b arbitraires, on a $\mathbb{P}[a \leqslant X \leqslant b] = \mathbb{P}[X \leqslant b] - \mathbb{P}[X \leqslant a - 1]$. 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 27% 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 $X \sim \mathcal{B}(3; 0{,}27)$, 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 27% de chances de « succès ».

En conséquence on a les probabilités suivantes :

Évènement	$X = 0$	$X = 1$	$X = 2$	$X = 3$
Probabilité	0,389	0,432	0,16	0,02

On peut constater que $\mathbb{P}[X = 1] = 0{,}432$ diffère sensiblement de la probabilité ($\frac{220}{455} \simeq 0{,}484$) obtenue dans l'exemple page 30 à la fin de la section 3.2.2. Pourtant, on choisissait bien trois personnes au hasard parmi un échantillon où la proportion de fumeurs était $\frac{4}{15} \simeq 0{,}27$.

La raison pour laquelle les probabilités diffèrent est que dans l'exemple page 30 on choisissait parmi 15 individus, de sorte qu'une fois le premier individu choisi, il n'y avait plus que 14 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.