Méthodes quantitatives

Paul Hobeika

2022-02-02

Contents

À propos de ce document **5**

1 Données et vocabulaire de la statistique **7**
1.1 Les sources statistiques en sociologie 7
1.2 Le vocabulaire de la statistique 9
1.3 Variables 12
1.4 Mesures de tendance centrale 14

2 Statistique descriptive univariée **15**
2.1 Variables qualitatives 15
2.2 Variables quantitatives 17
2.3 La loi normale : une distribution importante 27

3 Analyse bivariée et corrélation I **31**
3.1 Les tableaux croisés 31
3.2 Statistiques descriptives et statistiques inférentielles 34
3.3 Le test du χ^2 35

Références **41**

À propos de ce document

Cette page accueille les notes de cours de méthodes quantitatives du M1 de Science Po Strasbourg pour l'année 2021-2022. Il s'agit d'une introduction aux statistiques destinée à des étudiant · es de master de sociologie politique. Elle ne requiert pas de bagage préalable en statistique ou en mathématique. Il a été généré par l'extension bookdown de Yihui Xie, et le code source est disponible sur GitHub.

Chapter 1

Données et vocabulaire de la statistique

1.1 Les sources statistiques en sociologie

Nous avons évoqué la semaine dernière l'importance de la connaissance des sources statistiques pour la production de savoirs quantitatifs en sciences sociale. Il en existe différents types, qu'il est important de savoir identifier.

1.1.1 Les enquêtes par questionnaire produites par les chercheur-es

C'est par exemple le cas des données exploitées dans *La distinction* [Bourdieu, 1979] dont nous avons parlé au premier semestre. A partir d'une problématique de départ parfois abstraite (dans le sens pas directement quantifiable), l'élaboration d'un questionnaire a souvent pour objectif de trouver des éléments empiriques concrets qui permettent de rendre opérationnelles certaines notions ou concepts. Par exemple, dans *La distinction*, le questionnaire porte sur les pratiques culturelles et permet d'opérationaliser empiriquement la notion de *capital culturel*.

> Remarque : si vous souhaitez produire vous-même des données dans le cadre de votre TER et de la validation du cours c'est tout à fait possible, mais nous n'aborderons pas la méthodologie du questionnaire dans ce cours. De bons manuels sont toutefois disponibles, je vous recommande par exemple celui de Bugeja-Bloch and Couto [2021], chapitres 3 et 4.

1.1.2 Les autres source de "première main"

En réalité, les chercheur-es peuvent effectuer des traitement quantitatifs sur d'autres types de sources que les données issues d'un questionnaire. Pour cette raison, Fanny Bugeja-Bloch et Marie-Paule Couto font une distinction entre les **techniques d'enquête** et les **techniques d'analyse** des données.

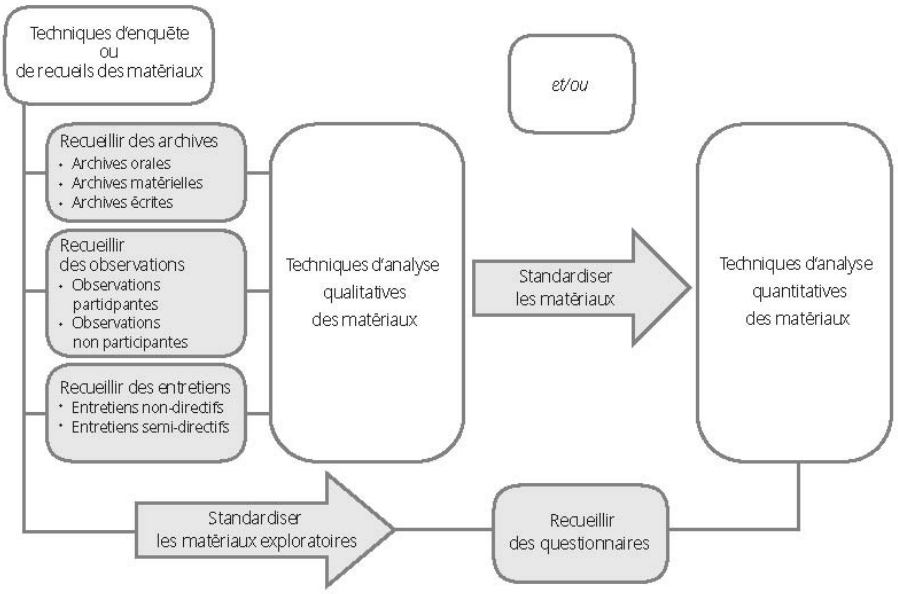


Figure 1.1: Techniques d'enquête et techniques d'analyse [Bugeja-Bloch and Couto, 2021]

Les **techniques d'enquête** désignent les différents modes de recueil des données : données d'archives, d'entretien ou encore issues d'observation. Les matériaux ainsi produits peuvent ensuite se prêter à différentes formes d'analyse. C'est seulement à ce niveau que l'on peut distinguer méthodes qualitatives et quantitatives. Une analyse qui se fondera sur le commentaire d'un ou quelques extrait d'entretien par exemple sera qualifiée de *qualitative*. Mais ces mêmes matériaux, lorsqu'ils sont *standardisés* et *mis en série* peuvent également être l'objet de techniques d'analyse quantitative. On peut produire des statistiques à partir d'archives [Lemercier and Zalc, 2008], à partir d'entretiens (le questionnaire en est un cas particulier) ou encore à partir d'observations [1].

[1] Un exemple tiré de la sociologie du travail est celui de l'enquête de Jean Peneff sur les urgences. Effectuant une enquête par observation participante en tant que brancardier dans un service d'urgence, il fait un certain nombre de comptages dans l'objectif d'objectiver certaines dimensions du travail aux urgences [Peneff, 1992].

1.1.3 L'analyse secondaire des données

Dans de nombreux cas, ce ne sont pas les sociologues ou politistes qui produisent les données qu'ils ou elles exploitent. On parle alors d'**analyse secondaire des données**. C'est le cas lorsqu'on travaille sur des données de l'Insee ou n'importe quelle base de donnée produite par une administration.

Quelques liens pour accéder aux données de la statistique publique française :

- le site de l'Adisp (Archives de données issues de la statistique publique) , qui rassemble les données de l'Insee et des directions statistiques ministérielles (santé, travail, culture, etc.)
- les données de l'Ined (Institut national de la recherche démographique)

1.1.4 Données d'enquête et données de gestion

Parmi l'ensemble des données accessibles produites par la statistique publique, on distingue en général deux grandes catégories [Desrosières, 2005] . D'un côté les bases de données produites via une **enquête par questionnaire** comme évoqué plus haut : elles sont réalisées à partir d'un échantillonage au sein d'une population plus large (voir plus loin pour des définitions de ces termes), et comportent un grand nombre de variables, qui correspondent en général à des questions qui sont posées directement par des enquêteurs ou enquêtrices. De l'autre côté, certaines bases de données sont le **résultat du travail de gestion de certaines administrations** : par exemple, les employeurs effectuent chaque année ce qu'on appelle une "déclaration annuelle de données sociales", dans laquelle ils renseignent une série d'informations sur leurs différents salarié · es (parmi lesquelles leur salaire et leur profession). Ces "DADS" constituent un exemple de base de données administrative. Ils sont largement utilisés pour étudier les salaires. Ces bases de données sont intéressantes mais en général moins riches que les données d'enquête, car elles ne sont pas réalisées dans le but de produire de la connaissance. Je vous conseille d'éviter de choisir une telle base de données pour votre rendu du semestre, car il est souvent plus difficile d'en tirer des résultats intéressants à moins de savoir exactement ce qu'on cherche.

1.2 Le vocabulaire de la statistique

1.2.1 Bases de données

Il est temps d'expliquer plus précisément ce qu'on entend par "base de données". En voilà un premier exemple, issu du package R `titanic` :

Une **base de données** se présente sous la forme d'un tableau. Les lignes décrivent les **individus** : ici ce sont des passagers, mais gardez en tête que la nature des individus peut être à peu près n'importe quoi (ça peut être des ménages, des villes, des bactéries, n'importe quoi). Chaque colonne apporte des

Table 1.1: Extrait de la base de données des passagers du Titanic

PassengerId	Survived	Age	Name
1	0	22	Braund, Mr. Owen Harris
2	1	38	Cumings, Mrs. John Bradley (Florence Briggs Thayer)
3	1	26	Heikkinen, Miss. Laina
4	1	35	Futrelle, Mrs. Jacques Heath (Lily May Peel)
5	0	35	Allen, Mr. William Henry
6	0	NA	Moran, Mr. James

Table 1.2: Extrait de la base USArrests

	Murder	Assault	UrbanPop	Rape
Alabama	13.2	236	58	21.2
Alaska	10.0	263	48	44.5
Arizona	8.1	294	80	31.0
Arkansas	8.8	190	50	19.5
California	9.0	276	91	40.6
Colorado	7.9	204	78	38.7

éléments permettant de caractériser les individus (leur nom, leur âge, etc.). On appelle ces caractéristiques des **variables**.

1.2.2 Un autre exemple

Dans cet exemple, les individus sont des États des États-Unis. Les variables correspondent à des taux d'arrestation par la police pour meurtre, agression et viol pour 10000 habitants en 1973, ainsi que le pourcentage de la population urbaine.

1.2.3 Données "tidy"

Dans R, on qualifie certaines base de données de "tidy" [Wickham, 2014]. C'est la structure qu'on souhaite avoir en général. Ces bases de données ont un individu par ligne, une variable par colonne. Dans chaque case, on trouve la modalité d'une variable correspondant à l'individu décrit dans la ligne L'exemple présenté sur la figure 1.2 en bas à droite n'est pas "tidy", car il existe une variable (dont on ne connaît pas le nom) dont les modalités sont réparties dans deux colonnes différentes, qui représenteny les années 1999 et 2000, c'est-à-dire les modalités d'une autre variable qui indique l'année. En replaçant l'année et la variable observée chacune dans une colonne, on obtient un tableau 'tidy' (en bas à gauche).

country	year	cases
Afghanistan	1999	745
Afghanistan	2000	2666
Brazil	1999	37737
Brazil	2000	80488
China	1999	212258
China	2000	213766

country	1999
Afghanistan	745
Brazil	37737
China	212258

table4

Figure 1.2: Tidy data sets

Table 1.3: Extrait de la base de données beaver1

day	time	temp	activ
346	840	36.33	0
346	850	36.34	0
346	900	36.35	0
346	910	36.42	0
346	920	36.55	0
346	930	36.69	0

1.2.4 Séries temporelles

On parle parfois de **série temporelle** lorsque une base de donnée concerne un même individu statistique à différents instants. On parle dans ce cas là plutôt d'**observations** que d'individus. En voilà un exemple : la base de données `beaver1` accessible dans R présentent la température corporelle d'un castor en fonction du temps.

1.2.5 Autres types de bases de données

- Des données qui décrivent différents individus à un moment donné sont parfois qualifiéed de **données en coupe** (ou *cross-sectional dataset*). Les données du titanic ou de USArrest en sont des exemples.
- Certaines bases de données décrivent un même groupe d'individus statistique de manière répétée dans le temps. On nomme ce genre de données des **données de panel** (exemple : enquête Emploi en continu).

1.3 Variables

1.3.1 Définition

Les variables sont les éléments qui permettent de décrire les individus présents dans le base de données. Lorsque les données sont issues d'un questionnaire, chaque question correspond en général à une variable.

Exemple :

- le sexe
- la catégorie socioprofessionnelle
- le niveau de diplôme
- le revenu

On appelle **modalités** les différentes valeurs que peuvent prendre une variable.

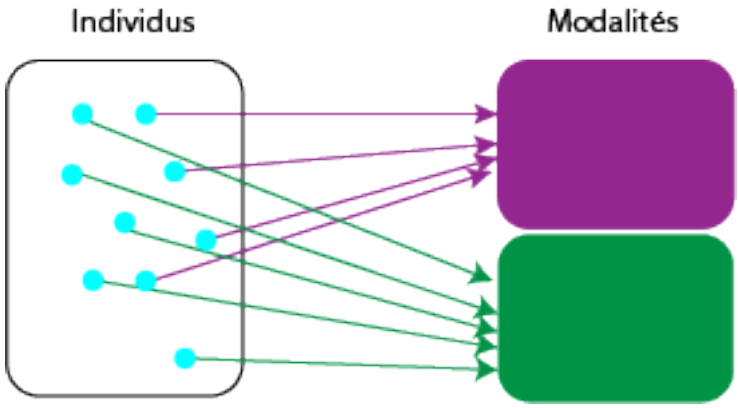


Figure 1.3: Les variables associent les individus à leur modalités

1.3.2 Variables qualitatives et variables quantitatives

On distingue les variables en fonction des opérations statistiques qu'on peut effectuer à partir de leurs modalités. Les deux grandes catégories de variables sont les **variables quantitatives**, dont les modalités sont des nombres, et les **variables qualitatives**, qui sont les autres.

1.3.3 Variables qualitatives

Parmi les variables qualitatives, on distingue encore :

- Les variables **qualitatives ordinales**, qui sont celles pour lesquelles ont peut ordonner les modalités (exemple : niveau de diplôme)
- Les variables **qualitatives nominales** dont les modalités ne sont pas hiérarchisables (exemple : sexe, catégorie socioprofesionnelle)

1.3.4 Variables quantitatives

Les modalités des variables quantitatives (ou numériques) sont des nombres qui ont une signification (par exemple, le code postal n'est pas une variable quantitative). Parmi elles, on distingue :

- Les variables **continues**, qui peuvent prendre toutes les valeurs réelles dans un intervalle donné
- Les variables **discrètes**, qui ne peuvent prendre qu'un certain nombre de valeurs

Pourquoi toutes ces catégories ? À ces différents types de variables, on associe différentes méthodes statistiques. Il est donc important de comprendre et mémoriser ces définitions, car lorsque vous souhaiterez étudier une variable, la première chose à faire sera d'identifier son type pour ensuite utiliser les méthodes statistiques appropriées.

1.4 Mesures de tendance centrale

Ce sont des manières de résumer l'information contenue dans une variable. En fonction du type de variable, il existe plusieurs indicateurs.

- La **moyenne** : lorsqu'on parle de moyenne, on fait généralement référence à la moyenne arithmétique d'un ensemble de valeurs numériques (par opposition à la moyenne géométrique ou harmonique). C'est une mesure très utilisée car elle fournit un premier résumé de la distribution statistique. Elle existe uniquement pour les **variables quantitatives**.

$$\bar{X} = \frac{1}{N} \sum_{i}^{N} x_i$$

- La **médiane** est la modalité d'une variable qui permet de séparer la population en deux parts égales. C'est-à-dire que 50% des individus auront une modalité supérieure ou égale à la médiane, et 50% une modalité inférieure ou égale à la médiane.
- Les **quantiles** sont une généralisation de la médiane : si vous voulez diviser votre population en groupe de 10%, vous pouvez utiliser les **déciles**. On appelle la médiane le **quantile d'ordre 2**, tandis que les déciles sont les **quantiles d'ordre 10**. Les quantiles les plus utilisés sont la médiane, les quartiles, les déciles et les centiles.
- Les quantiles n'existent que pour les variables dont les modalités peuvent être hiérarchisées : toutes les variables quantitatives et les variables qualitatives ordinales.
- Le **mode** indique la modalité la plus fréquente d'une variable. Par exemple, la plupart des passagers du Titanic sont décédés dans le naufrage, donc le mode de la variable "Survived" est 0. Le mode existe pour tous les types de variables.

Chapter 2

Statistique descriptive univariée

Le cours précédent était consacré à vous présenter différents types de variables. Celui de cette semaine présente les premiers éléments de **statistique descriptive univariée**, les outils permettant la description d'une unique variable. Ces outils dépendent de la nature de la variable étudiée.

2.1 Variables qualitatives

2.1.1 Tris à plat

Pour décrire ce genre de variable, le principal traitement statistique est de compter le nombre d'individus correspondant à chaque modalité de la variable. C'est ce qu'on appelle un **tri à plat** (par opposition aux tris croisés qui font intervenir plusieurs variables). Un exemple issu des données du titanic (voir section 1.1).

	n	%
1	216	24.2
2	184	20.7
3	491	55.1
Total	891	100.0

À partir d'un tableau comportant une ligne par passager, on produit donc un tableau qui comporte seulement une ligne par classe de passagers. La colonne d'effectif montre le nombre de passagers par classe, tandis que la colonne de pourcentage indique le pourcentage de passagers des différentes classes parmi l'ensemble de passagers du Titanic. On peut lire le tableau de cette manière : parmi les 891 passagers du Titanic, 216 voyageaient en première classe. On cal-

cule le pourcentage de chaque catégorie en divisant l'effectif de chaque catégorie par l'effectif total, puis en multipliant par 100 ($\frac{216}{891} * 100 = 24,2\%$).

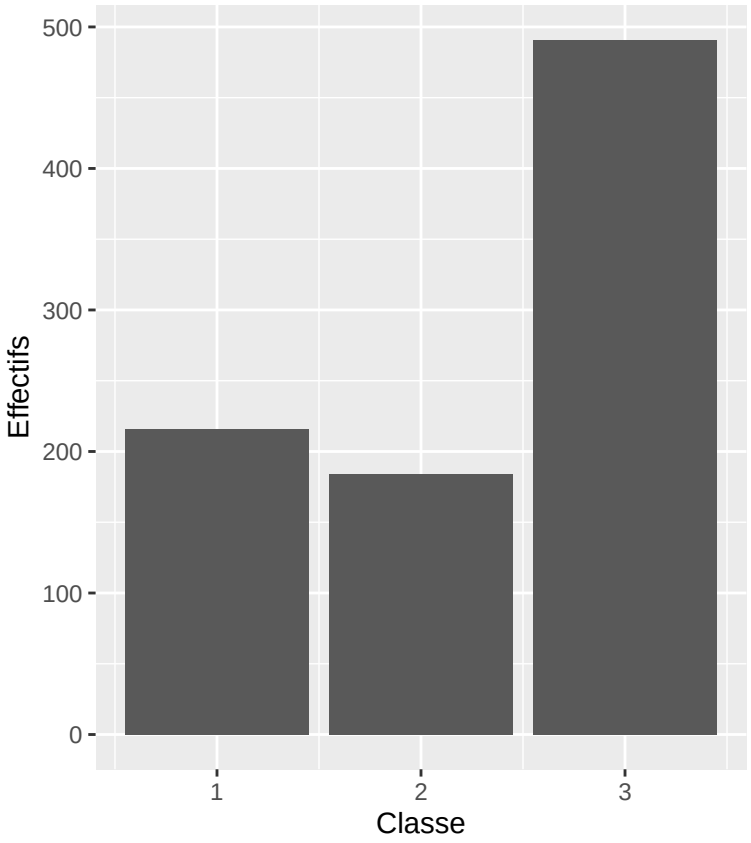
Dans le cas où la variable est qualitative ordinale (c'est-à-dire qu'on peut ordonner ses modalités de manière hiérarchique, comme c'est le cas pour la variable de classe), on peut présenter dans ce tableau les **pourcentages cumulés**.

	n	%	%cum
1	216	24.2	24.2
2	184	20.7	44.9
3	491	55.1	100.0
Total	891	100.0	100.0

Ici, le chiffre 44,9% représente le pourcentage des passagers qui voyageaient **au moins en seconde classe**. Il s'agit simplement de la somme des pourcentages des passagers des première et seconde classe. La ligne suivante indique le pourcentage de passagers qui voyageaient au moins en troisième classe, ce chiffre est donc logiquement égal à 100%.

2.1.2 Diagrammes en barre

La représentation graphique associée à ce décompte est ce qu'on appelle généralement un **diagramme en barre** (vous pouvez aussi trouver "diagramme en bâtons ou diagramme en"tuyaux d'orgue" qui désignent la même chose). Sur ce diagramme, on trace des barres verticales dont les hauteurs sont proportionnelles aux effectifs du tri à plat. Seule la hauteur des barres à une signification, la largeur est totalement arbitraire.



2.2 Variables quantitatives

Si la statistique univariée est très simple pour une variable qualitative, elle peut faire l'objet d'analyses plus approfondies lorsqu'on dispose de variables quantitatives.

2.2.1 Mesures de dispersion

La semaine dernière, je vous ai présenté quelques **mesures de tendance centrale**. Elles donnent des renseignements importants pour décrire une variable, mais n'en résument qu'une dimension. Deux séries statistiques peuvent avoir la même moyenne tout en étant très différentes.

Comparez par exemple ces deux séries de chiffres, qui représentent des profits (fictifs) en dollars de deux agriculteurs de deux régions A et B :

- A: 14, 16, 18, 20, et 22
- B: 2, 8, 18, 29, et 33

La somme de ces deux série est la même, 90 dollars, mais il apparaît rapidement que l'une des séries est beaucoup plus **dispersée** que l'autre, c'est-à-dire que les écarts par rapport à la moyenne sont en général beaucoup plus grands (la série B). Notre vision des risques et des profits liés à l'agriculture est informée par cette différence, et nous devrions en inclure des indices dans toute description statistique de cette variable.

Pour faire cela, nous avons besoin de mesures permettant de décrire la disperion des modalités de la variable autour de sa moyenne.

2.2.1.1 L'étendue

C'est la différence entre la plus grande et la plus petite valeur de la série :

$$R = X_{max} - X_{min}$$

C'est une mesure de dispersion assez basique. Son défaut est assez évident : elle dépend uniquement des valeurs extrêmes, et aucunement du reste de la distribution.

2.2.1.2 L'écart interquartile

Pour prendre en compte plus que les deux valeurs extrêmes, on peut calculer la différence entre deux quantiles, des quartiles par exemple (voir définition dans le premier cours)

$$Q_d = Q_3 - Q_1$$

C'est un mesure un peu meilleure que l'étendue, parce que le maximum et le minimum sont des valeurs qui donnent généralement peu d'information sur la distribution en général. Cet écart représente l'étendue de la moitié de la distribution, moitié obtenue après avoir enlevé les 25% des valeurs les plus faibles et 25% des valeurs les plus hautes. L'écart interquartile est moins sensible aux valeurs extrêmes que l'étendue (puisqu'on les a supprimées), mais résume tout de même l'ensemble de données sans prendre en compte la variabilité des données entre le premier et le 3ème quartile. Les mesures suivantes n'ont pas ces défauts.

2.2.1.3 La variance

Ce qu'on voudrait, c'est l'équivalent de la moyenne, mais pour mesurer la dispersion. On pourrait donc se dire qu'il suffirait de faire la **moyenne des écarts à la moyenne** de cette manière [1]:

[1] Les deux côtés de l'équation représentent la même chose, il s'agit juste d'une différence de notation. À gauche, on utilise … pour indiquer la série de termes supplémentaires qu'on va inclure dans l'addition mais qu'on n'écrira pas. À droite, l'opérateur $\sum_{i=1}^{n}$ est la somme pour i allant de 1 à n de l'expression qui est à droite. Cela signifie qu'il faut remplacer i par

$$\frac{1}{N}((X_1 - X_m) + (X_2 - X_m) + ... + (X_n - X_m) = \frac{1}{n}\sum_{i=1}^{n}(X_i - X_m)$$

Le problème, en faisant ça, c'est que compte tenu de la définition de la moyenne, les écarts à la moyenne vont se compenser terme à terme. On peut le voir facilement si l'on sépare la somme en deux :

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - X_m) = \frac{1}{n}\sum_{i=1}^{n}X_i - \frac{1}{n}\sum_{i=1}^{n}X_m$$

Du côté droit de l'équation, le terme de gauche ($\frac{1}{n}\sum_{i=1}^{n}X_i$) est la définition de la moyenne (c'est la somme des termes $X_1 + X_2 + ... + X_n$ divisé par l'effectif total n), tandis qu'à droite ($\frac{1}{n}\sum_{i=1}^{n}X_m$) on ajoute n fois la moyenne X_m puis on la divise par n, donc on obtient encore la moyenne. Au final, cette somme est toujours égale à 0.

Pour éviter ce problème, on définit la variance comme la somme des écarts à la moyenne **au carré**.

$$Var = \frac{1}{n}\sum_{i=1}^{n}(X_i - X_m)^2$$

Cette définition a l'avantage de donner un résultat non nul, excepté dans le cas où la variable est une constante (qui est alors toujours égale à sa moyenne). Surtout, les écarts à la moyenne s'ajoutent, qu'ils soient générés par des valeurs supérieures ou inférieures à la moyenne.

2.2.1.4 L'écart type

La variance a beaucoup de propriétés intéressantes et on l'utilise très largement en statistique. Malgré tout, elle a un dernier inconvénient, c'est de s'exprimer comme un carré de l'unité dans laquelle est mesurée la variable. Par exemple, si l'on mesure la taille des étudiant-es de la classe puis qu'on calcule la variance, on aura un résultat en centimètres ou en mètres au carré. Dans notre exemple, on obtient une mesure en "dollars au carré", dont le sens est difficile à interpréter.

On résoud ce problème de manière simple en calculant la racine carrée de la variance. Cette opération nous permet d'obtenir notre dernière mesure de dispersion, **l'écart-type**.

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - X_m)^2}$$

1, puis par 2, 3, etc. jusqu'à n et faire la somme de tous les éléments ainsi obtenu. Ce qui revient exactement à ce qui est écrit de manière plus longue de l'autre côté de l'équation.

L'écart-type est la mesure de dispersion la plus utile et la plus fréquente. Elle est meilleure que la variance car elle se mesure dans la même unité que la variable en question. Par exemple, on peut dire que dans la région A, la moyenne des revenus agricoles est de 18 dollars, avec un écart-type de 2,8 dollars.

2.2.2 Représentations graphiques

Il existe plusieurs manières de représenter graphiquement la distribution d'une variable quantitative.

2.2.2.1 Histogramme

Les histogrammes sont l'équivalent des diagrammes en barres pour les variables quantitatives. Chaque barre (ou rectangle) qui compose l'histogramme a une aire qui est proportionnelle au nombre d'observation dont les valeurs sont dans l'intervalle sur lequel s'étend le rectangle.

Comme il s'agit de variables quantitatives, on peut choisir le nombre de rectangles comme on le souhaite (contrairement aux variables qualitative dont les modalités sont définies une fois pour toutes). Ici, on représente la même variable avec moins de rectangles.

Ici avec un plus grand nombre de rectangle (largeur = 1 an)

2.2.2.2 Densité

On représente parfois les variables quantitatives par une courbe que l'on appelle une 'densité'. C'est la courbe qu'on pourrait obtenir si on avait un très grand nombre de passagers et qu'on représentait l'histogramme avec des rectangles très fins.

On peut également produire une estimation de cette courbe à partir d'une transformation effectuée sur les histogrammes. Si l'on trace une ligne qui passe au milieu de chacun des segments supérieurs des rectangles qui composent l'histogramme, on obtient alors un graph qu'on nomme un **polygone de fréquence**.

La forme de ce polygone peut être être "lissée" à l'aide de techniques mathématiques, pour donner la courbe de densité recherchée. Elle donne une idée de la forme du polygone de fréquence si l'on avait un très grand nombre d'individu dans notre échantillon.

Ces représentations graphiques sont un bon moyen de visualiser la **forme** de la distribution d'une variable quantitative continue, et spécifiquement de la manière dont les données sont réparties autour de leur valeur "centrale".

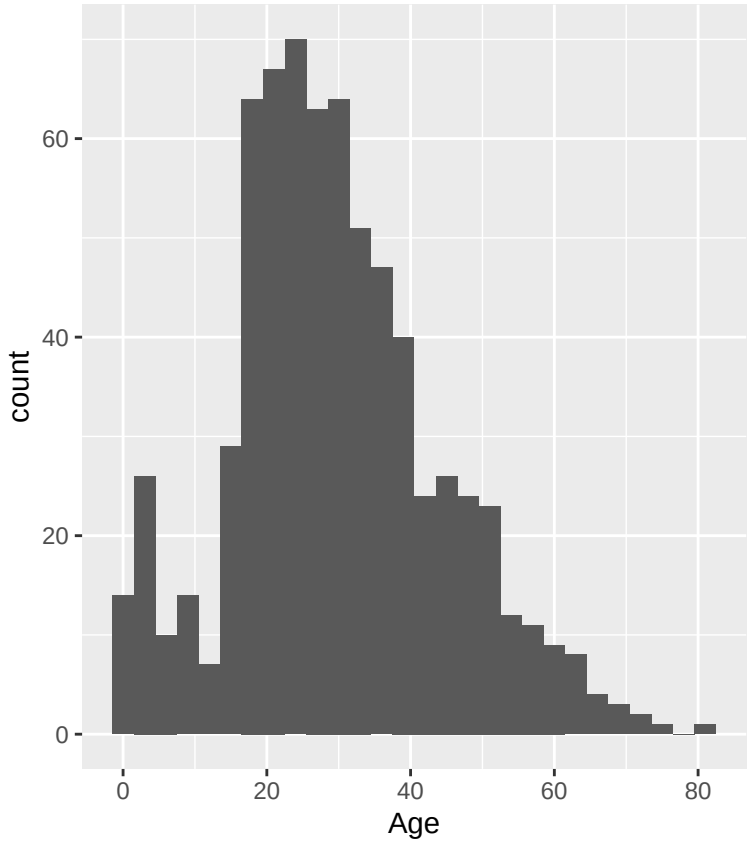


Figure 2.1: Distribution de l'âge des passagers du Titanic. Chaque rectangle a une largeur de 3 ans

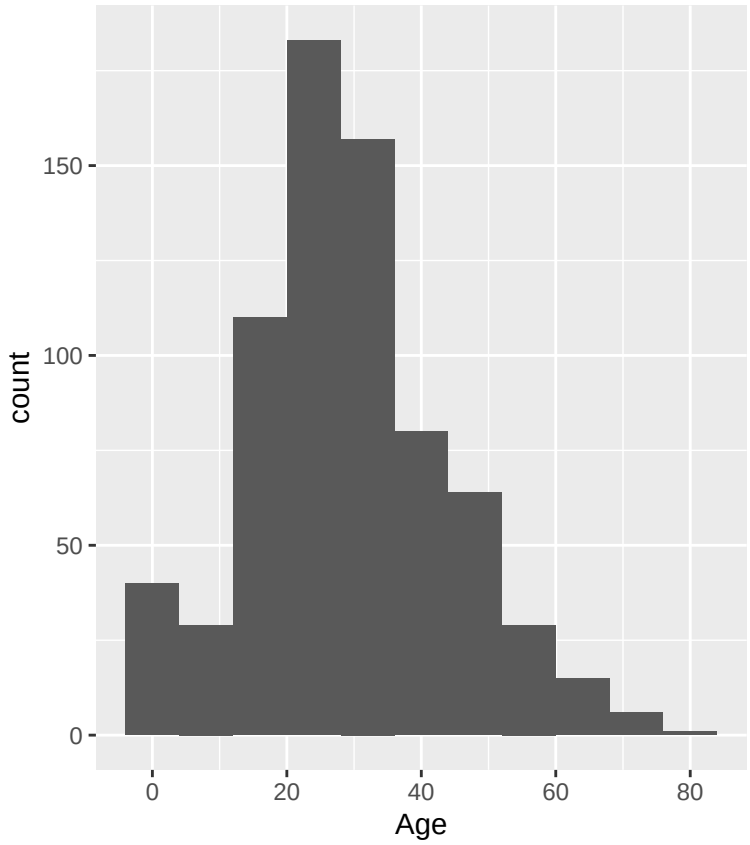


Figure 2.2: Même figure avec une largeur de 8 ans

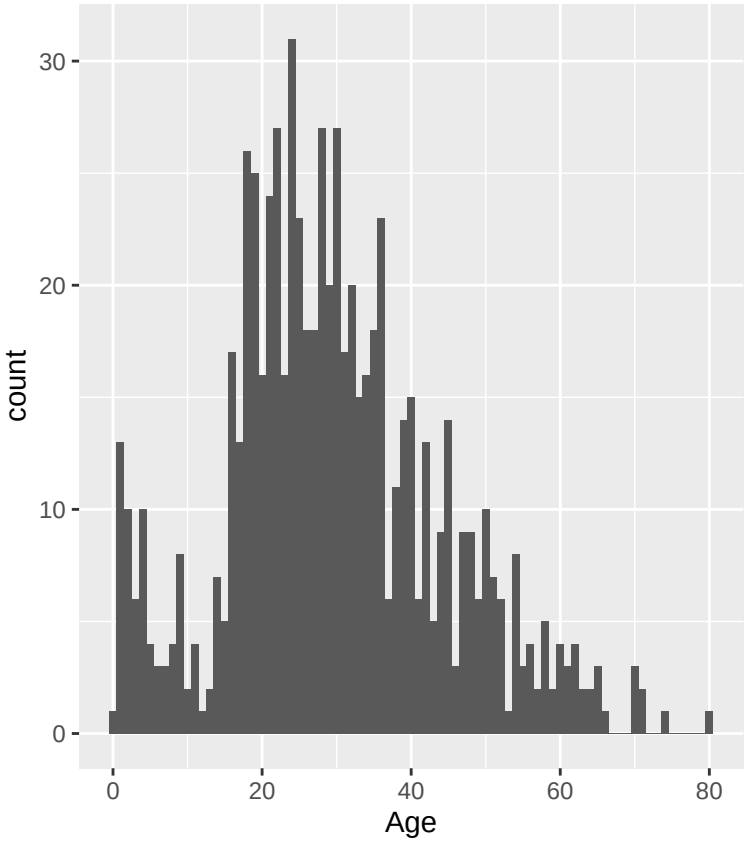


Figure 2.3: Même figure avec une largeur d'un an

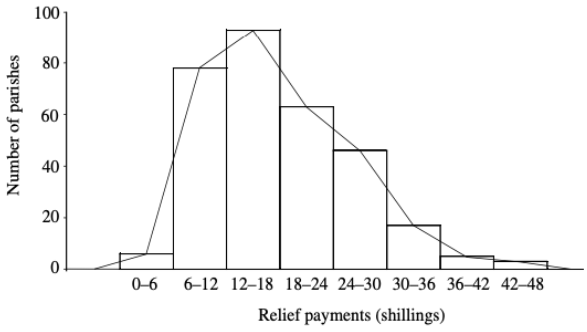


Figure 2.4: *Per capita* relief payments in 311 parishes in 1831 (Fenstein & Thomas, p. 41)

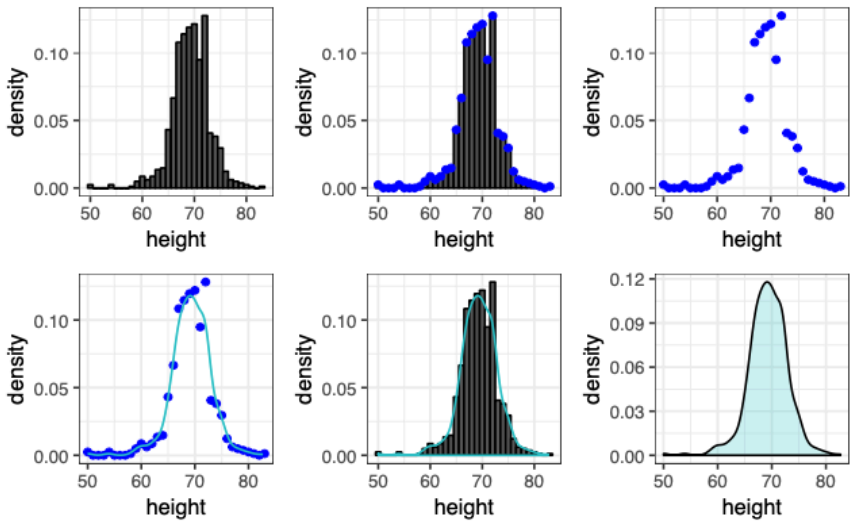


Figure 2.5: D'un histogramme à une densité de probabilité

2.2.2.3 Asymétrie (*skewness*)

Une manière de caractériser les distribution est leur symétrie par rapport à la moyenne. Les valeurs peuvent en effet être réparties de manière symétrique de part et d'autre de la moyenne, ou bien de manière asymétrique (*skewed*).

Différentes mesures permettent de quantifier l'asymétrie d'une distribution.

- Elles doivent être indépendantes de l'unité de mesure
- Et elle doivent être nulles lorsque la distribution est symétrique.

Un exemple de coefficient d'asymétrie est le suivant, mais il en existe d'autres :

$$Skewness = \frac{3 * (Mean - Median)}{\sigma}$$

Un exemple de variable dont la distribution est asymétrique est la distribution des revenus dans la population française. L'histogramme suivant représente la distribution du niveau de vie (c'est le revenu des ménages divisé par le nombre d'unités de consommation). Le niveau de vie médian (compris dans la portion verte du graphique) est inférieur au niveau de vie moyen, car les ménages au niveaux de vie très élevés (en bleu) tirent la moyenne vers le haut, mais n'ont pas d'effet sur la médiane.

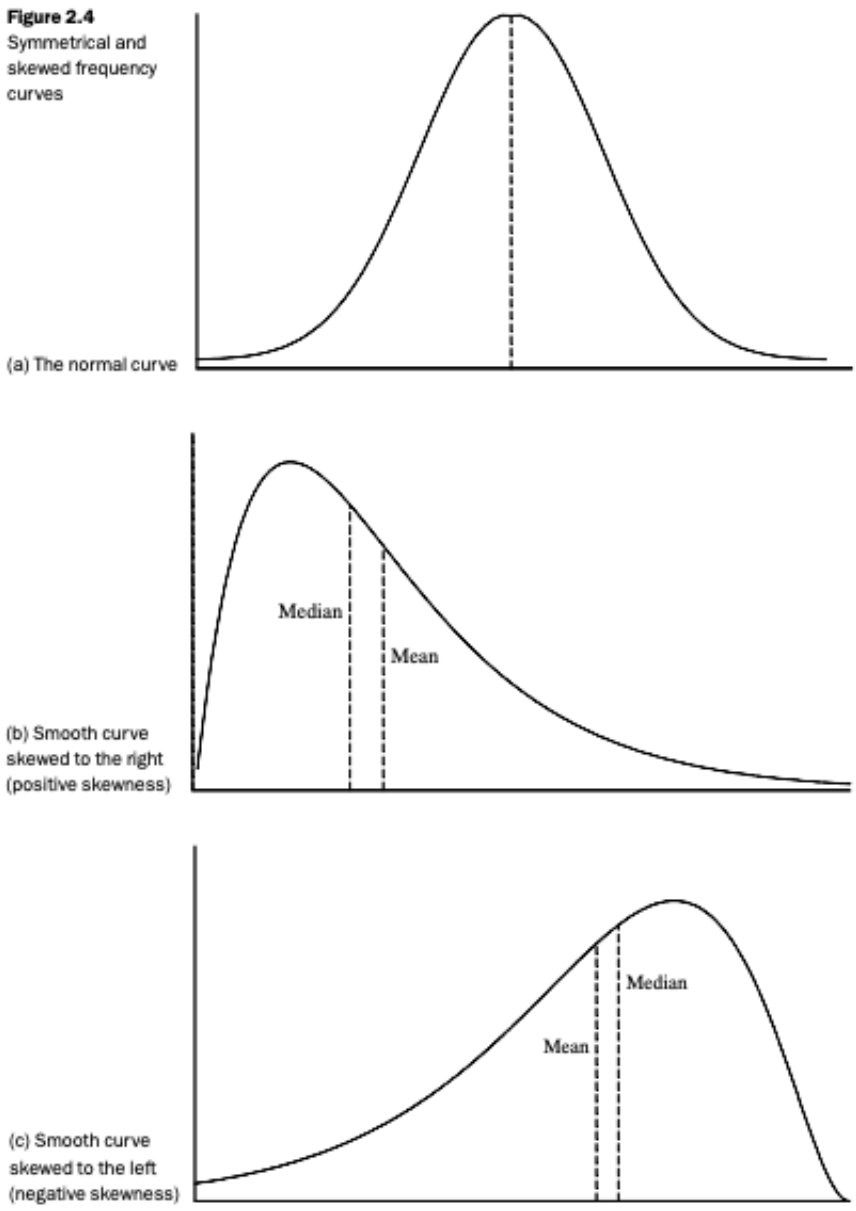


Figure 2.6: Distributions symétriques et asymétriques [Feinstein and Thomas, 2002, p.54]

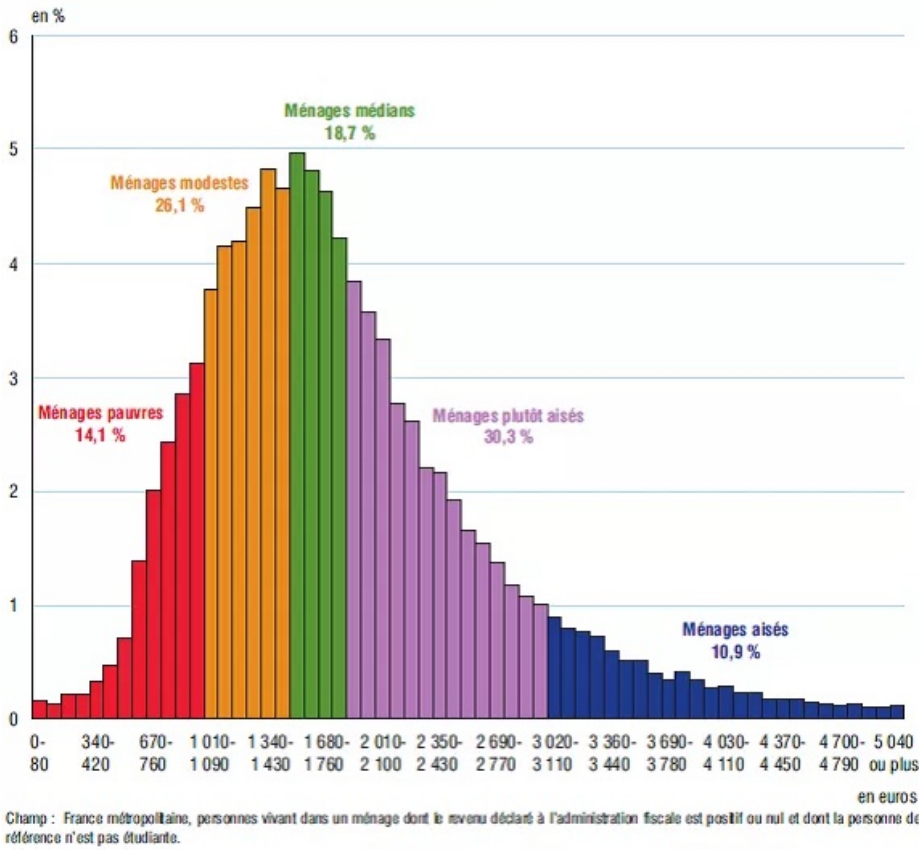


Figure 2.7: Distribution des niveaux de vie mensuels en 2014 en France (Source : Insee, Portrait social 2014)

2.3 La loi normale : une distribution importante

La loi normale est une distribution théorique, définie à partir de son expression mathématique. Mais bien que théorique, c'est une distribution très importante, car elle est souvent utilisée comme approximation de distributions réelles. Je vous la présente ici rapidement, on la retrouvera dans des prochaines séances.

Pour définir une loi normale, il faut connaître deux constantes : sa moyenne X_m et l'écart type σ. L'équation donne la valeur de Y (la hauteur de la courbe, qui apparaît sur l'axe des ordonnées) pour tout valeur de X (mesuré sur l'axe des abscisses) :

$$Y(X) = \frac{1}{\sigma\sqrt{2\pi}} exp(-\frac{(X - X_m)^2}{2 * \sigma^2})$$

La fonction $exp()$ qui apparaît dans la formule est la **fonction exponentielle**. Si vous ne connaissez pas cette fonction, sachez qu'elle est définie par le fait qu'il s'agit de l'unique fonction $f(x)$ qui est toujours égale à sa dérivée (la fonction dérivée est celle qui mesure la pente de la courbe en chaque point, on la note $f'(x)$: elle est positive lorsque f est croissante, et négative lorsqu'elle est décroissante) et qui est égale à 1 lorsque $x = 0$. Comme elle est toujours égale à sa dérivée, plus x est élevé, plus la fonction exponentielle doit avoir une dérivée élevée, donc plus elle doit croître rapidement.

```
curve(exp(x), from=-5, to=5, , xlab="x", ylab="y")
```

Dans la distribution de la loi normale, la fonction exponentielle contient une expression qui est toujours inférieure ou égale à zéro. Son maximum est donc atteint lorsque X est égal à sa moyenne X_m, auquel cas $Y(X_m) = \frac{1}{\sigma\sqrt{2\pi}}$. Plus X va s'éloigner de sa moyenne, plus $Y(X)$ sera faible, on dit que la distribution *tend vers 0 lorsque X tend vers "moins l'infini" ou "plus l'infini". Le graphe de la loi normale ressemble donc à un dos d'âne, ce qui explique qu'on l'appelle aussi "la courbe en cloche".

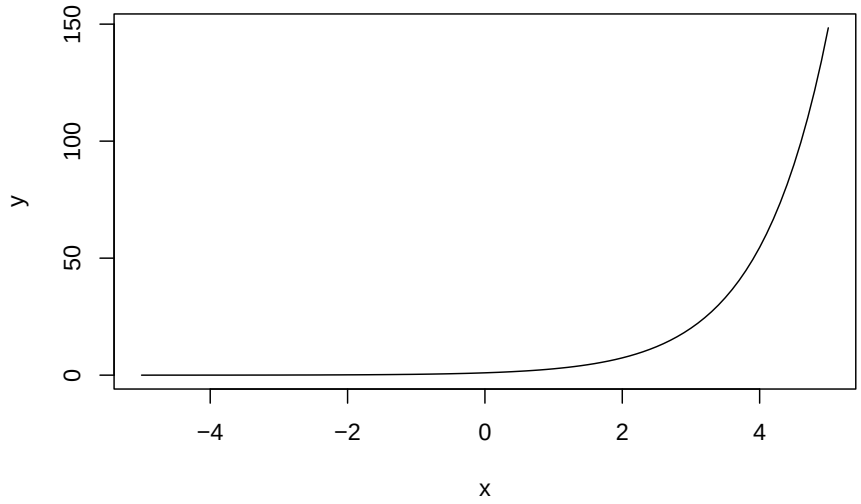
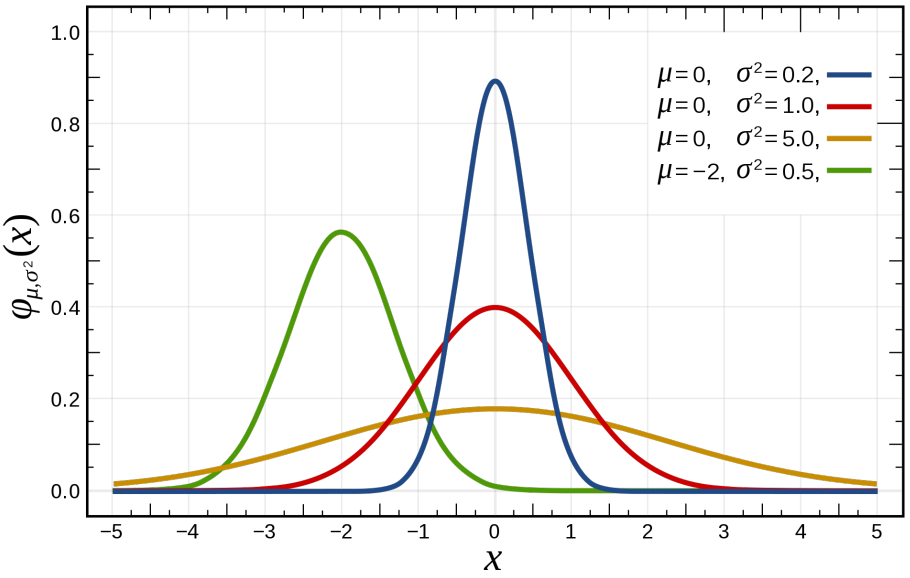


Figure 2.8: Graphe de la fonction exponentielle entre -5 et 5



Ce dernier graphe permet de constater que, si les lois normales ont toutes la même allure, leur forme dépend de la moyenne et de l'écart-type de la

distribution. Comme déjà évoqué, la moyenne indique le maximum de la courbe. L'écart-type détermine lui la "largeur" de la bosse, c'est-à-dire à quel point les données s'étalent autour de la moyenne.

Une propriété importante de la loi normale est que, quelque soit sa moyenne et son écart-type, il y a toujours une même proportion d'observations qui seront distribués à une certaine distance de la moyenne (que l'on peut mesurer en calculant l'aire sous la courbe), mesurée en nombre d'écarts-type.

Par exemple :

- **90% des observations** sont situés à moins de **1,645 écarts-type** autour de la moyenne, laissant 5% de chaque côté.
- **95% des observations** sont situés à moins de **1,96 écarts-type** autour de la moyenne, laissant 2,5% de chaque côté.
- **99% des observations** sont situés à moins de **2,58 écarts-type** autour de la moyenne, laissant 0,5% de chaque côté.

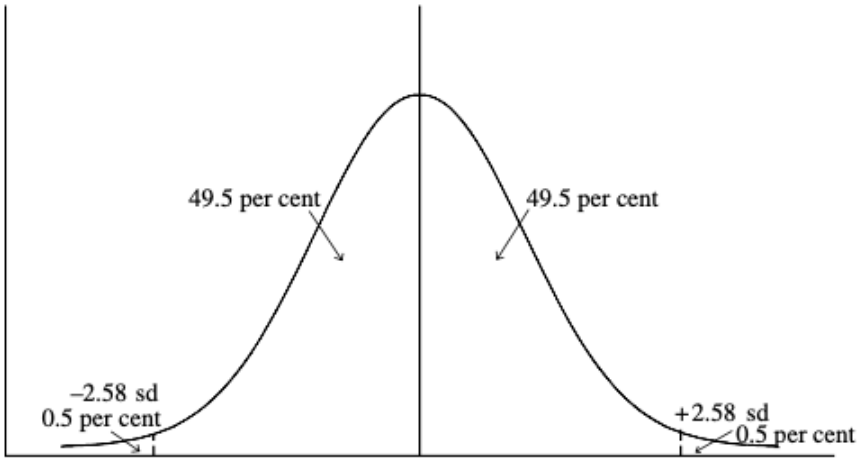


Figure 2.9: Aire sous la courbe

Chapter 3

Analyse bivariée et corrélation I

Jusque là, on a vu différents outils qui nous permettent de décrire plusieurs types de variables :

- tris à plat (variables qualitatives)
- indices de tendance centrale
- indices de dispersion
- représentations graphiques (diagrammes en barres, histogrammes)

C'est-à-dire que ce qu'on sait faire, c'est prendre une variable (par exemple la catégorie socioprofessionnelle d'une personne ou le revenu d'un ménage), et, en fonction du type de variable, en proposer une sorte de résumé, dont la forme dépend de la nature de la variable À partir de maintenant, on va commencer à voir comment étudier les relations entre plusieurs variables. C'est ce qui nous intéresse en général. Le cours de cette semaine ne traite que le cas de **deux variables qualitatives**.

3.1 Les tableaux croisés

Un **tableau de fréquence** ou **tableau croisé** est l'outil statistique le plus fréquemment utilisé pour étudier le lien entre deux variables qualitatives. C'est un tableau qui indique la distribution des effectifs d'une population en fonction des modalités de deux variables qualitatives. En voici un premier exemple :

Un tel tableau qui répartit N individus dans 4 cases constitue un système de classification **exclusif** (chaque individu est dans une seul case) et **exhaustif** (tous les individus sont dans une case)

Table 3.1: Répartition des passagers du Titanic par classe et par sexe

	1	2	3	Total
female	94	76	144	314
male	122	108	347	577
Total	216	184	491	891

Table 3.2: Répartition des passagers du Titanic par classe et par sexe

	1	2	3	Total
female	94	76	144	**314**
male	122	108	347	**577**
Total	**216**	**184**	**491**	**891**

3.1.1 Distributions marginales

La dernière colonne du tableau et la dernière ligne indiquent les totaux pour chaque ligne et colonne. On les appelle les **distributions marginales**. Un tableau croisé doit toujours comporter ces distributions marginales. Elles correspondent aux tris à plat des deux variable séparées, ici les variables *Sexe* et *Fréquentation du théâtre*.

3.1.2 Distributions conditionnelles

On appelle par contraste les effectifs présents à l'intérieur du tableau les **distribution conditionnelles**. Chaque case correspond au nombre d'individus concernés simultanément par les deux modalités des deux variables (en ligne et en colonne). Ce tableau est la *distribution de la fréquentation du théâtre en fonction du sexe*.

Le problème avec ce tableau est qu'il est difficile à commenter, car le nombre total de femmes et d'hommes parmi les passagers est différent, et de même le nombre de passagers dans chaque classe est différent. On ne peut donc pas facilement comparer les distributions conditionnelles (le nombre de femmes et d'hommes dans chaque classe). Pour cela, il faut transformer le tableau en utilisant des pourcentages.

Table 3.3: Répartition des passagers du Titanic par classe et par sexe

	1	2	3	Total
female	**94**	**76**	**144**	314
male	**122**	**108**	**347**	577
Total	216	184	491	891

Table 3.4: Pourcentages d'hommes et de femmes parmi chaque classe du Titanic

	1	2	3	Ensemble
female	43.5	41.3	29.3	35.2
male	56.5	58.7	70.7	64.8
Total	100.0	100.0	100.0	100.0

Il y a différentes façons de former une table en pourcentages à partir du tableau de fréquence. La première idée serait de calculer les pourcentage de chaque case **par rapport à l'effectif total**.

	1	2	3	Total
female	10.5	8.5	16.2	35.2
male	13.7	12.1	38.9	64.8
Total	24.2	20.7	55.1	100.0

La transformation effectuée ici consiste à diviser chaque chiffre du tableau par l'effectif total. On obtient ainsi le pourcentage de chaque catégorie parmi l'ensemble des passagers. On peut alors lire les distributions conditionnelles de cette manière : les femmes de 1ère classe représentent 10,5% de l'ensemble des passagers du Titanic. Les distributions marginales donnent à nouveau les tri à plat des deux variables, exprimés en pourcentages : 35,2% des passagers sont des femmes, 24,2% des passagers sont en première classe.

On observe donc que, malgré cette transformation, il est toujours impossible de comparer directement les pourcentages des distributions conditionnelles, car ils sont encore dépendants des distributions marginales. Pour comparer les chiffres présents dans les différentes lignes ou colonnes, il faut se ramener à une situation où le nombre d'hommes ou de femmes serait égal, ou bien le nombre de passagers de chaque classe serait égal.

Pour cela, on calcule des pourcentage **en ligne** ou des pourcentage **en colonne**.

3.1.3 Pourcentages en ligne et en colonne

Ce tableau est un exemple de pourcentages en colonne Pour l'obtenir, plutôt que calculer des pourcentages par rapport à l'effectif total, on calcule le pourcentage d'hommes et de femmes pour chaque classe, c'est-à-dire qu'on divise les distributions conditionnelles par les effectifs totaux de chaque classe (par exemple, $94/216 * 100 = 43,5\%$ de femmes parmi les passagers de première classe). On se ramène donc à une situation fictive, dans laquelle chaque classe du Titanic aurait 100 passagers et passagères au total, **mais dont la proportion d'hommes et de femmes parmi chaque classe serait la même que la proportion réelle**.

Se ramener à 100 passagers par classe permet alors de comparer les pourcentages de femmes et d'hommes pour ces différentes classe. Pour commenter le tableau,

Table 3.5: Distribution par classe des femmes et des hommes passagers du Titanic

	1	2	3	Total
female	29.9	24.2	45.9	100
male	21.1	18.7	60.1	100
Ensemble	24.2	20.7	55.1	100

il faut d'abord lire les pourcentages marginaux : on voit qu'au total il y a 35% de femmes et 65% d'hommes parmi les passagers. Ainsi, même si les femmes sont minoritaires en première classe (43,5%), elles sont surreprésentées par rapport à leur pourcentage parmi l'ensemble des passagers. Elles le sont également en seconde classe, où elles représentent 41,3% des passagers. Elles sont à l'inverse sous-représentées en troisième classe, où elles sont seulement 29,3%.

Il est également possible de calculer des pourcentages en ligne.

Ici, on se ramène à une sitaution où il y aurait 100 hommes et 100 femmes sur le bateau, et on compare leur distribution par classe. Ce tableau permet de comparer directement les proportions d'hommes et de femmes parmi les différentes classes. On observe par exemple que la proportion de femmes en première classe (29,9%) est plus élevée que celle des hommes (21,1%), et inversement en 3ème classe. On peut encore comparer avec les distributions marginales, mais ici comme il n'y a que deux modalités c'est moins important.

3.2 Statistiques descriptives et statistiques inférentielles

Une des caractéristiques des données des passagers du Titanic est qu'elles sont **exhaustives**, c'est-à-dire que l'on détient des informations pour l'ensemble des passagers. Produire un tableau croisé permet ainsi de pouvoir établir sans ambiguïté les liens entre deux variables qualitatives (par exemple, ici, on peut affirmer que les femmes sont surreprésentées dans les deux premières classes). On parle dans ce cas de **statistiques descriptives**.

Produire cette affirmation est plus complexe dans le cas où l'on dispose de données produites sur **un échantillon** d'individus, et que l'on souhaite en déduire des résultats sur une population plus large. C'est cette question qui est au cœur de la **statistique inférentielle**. Prenons donc un autre exemple, issu des données de l'enquête "Histoires de vie" réalisée en 2003 par l'Insee, et dont le package `questionr` fournit un extrait. On va s'intéresser à la pratique du bricolage.

On peut de la même manière produire un tableau avec des pourcentages en colonne

Table 3.6: Pratique du bricolage par sexe

	Non	Oui	Total
Homme	384	515	899
Femme	763	338	1101
Total	1147	853	2000

Table 3.7: Pratique du bricolage par sexe

	Non	Oui	Ensemble
Homme	33.5	60.4	45
Femme	66.5	39.6	55
Total	100.0	100.0	100

On lit ainsi que, parmi les individus qui font partie de l'échantillon, les hommes sont 60,4% à déclarer pratiquer le bricolage, contre 39,4% des femmes. La question est alors de savoir si l'on peut généraliser ce résultat à l'ensemble de la population, c'est-à-dire affirmer que, parmi la population française de plus de 15 ans, les hommes pratiquent plus le bricolage que les femmes.

On conçoit que la réponse à cette question dépend de la manière avec laquelle les individus qui composent l'échantillon ont été sélectionnés. Si les hommes ont été interrogés à la sortie d'un magasin de bricolage, tandis que les femmes ont été sélectionnée d'une autre manière, il est évident que ces résultats ne seront pas généralisable, car l'échantillon ne sera pas **représentatif** de la population qu'on cherche à décrire.

Pour répondre à cette question, il est donc nécessaire d'avoir un "bon" échantillon. On va donc faire comme si l'échantillonnage avait été réalisé de manière **aléatoire** (ce qui n'est pas forcément vrai pour ce jeu de données qui est seulement un extrait de la base de données 'histoire de vie'). Si l'échantillonnage est aléatoire, on peut alors préciser notre question, qui devient : quelle probabilité y a-t-il que les différences observées entre les déclarations des hommes et des femmes interrogées au sujet du bricolage soient l'effet du hasard ? Autrement dit, est-il possible d'expliquer que les hommes soient majoritaires à se déclarer bricoleurs dans notre échantillon par le fait qu'on aurait *par hasard* interrogé des hommes particulièrement bricoleurs, ou des femmes particulièrement peu bricoleuses ?

3.3 Le test du χ^2

Pour répondre à cette question, on effectue ce qu'on appelle un **test d'hypothèse**. Il existe beaucoup de tests différents, mais le test qu'on va utiliser s'appelle le test du χ^2 (prononcé ki-deux). Il s'agit d'une procédure

Table 3.8: Pratique du bricolage par sexe. Effectifs observés.

	Non	Oui	Total
Homme	384	515	899
Femme	763	338	1101
Total	1147	853	2000

permettant d'évaluer le **niveau de significativité d'une relation statistique entre deux variables qualitatives**, ici la relation entre la variable "Bricolage" et la variable "Sexe".

Les tests d'hypothèse suivent tous la même logique : on commence par faire une hypothèse de départ **sur la population générale**, et on va ensuite **tester le caractère plus ou moins plausible de cette hypothèse à partir des données de notre échantillon** (répétons qu'il doit s'agir d'un échantillon aléatoire). Ici, l'hypothèse est que les deux variables "Sexe" et "Bricolage" ne sont pas corrélées ; on l'appelle **l'hypothèse nulle**.

3.3.1 Principe du test d'hypothèse

Prenons un exemple plus simple pour bien se représenter le principe. Imaginons qu'on lance une pièce de monnaie pour savoir si elle est équilibrée ou non (c'est-à-dire qu'elle a la même probabilité de tomber sur pile ou face). D'un côté, on fait l'hypothèse qu'elle est équilibrée. De l'autre, on la lance 100 fois, et on observe qu'elle tombe 55 fois sur face et 45 fois sur pile. On chercher à accepter ou à rejeter notre hypothèse de départ à partir de ces chiffres.

On ne peut pas répondre à cette question de manière certaine, mais seulement estimer le **risque de se tromper**. Accepter ou rejeter l'hypothèse ont chacun leur risque associé :

- le risque d'accepter l'hypothèse alors qu'elle est fausse (par exemple, ici, dire que la pièce est équilibrée)
- le risque de rejeter l'hypothèse alors qu'elle est vraie (ici, dire que la pièce n'est pas équilibrée alors qu'elle l'est)

Pour calculer ces deux risques, le principe est toujours de comparer la distribution statistique (ici 45/55) et la distribution théorique à laquelle on s'attendrait si l'hypothèse de départ était vérifiée (ici, 50/50). On appelle ces distributions les **effectifs observés** et les **effectifs théoriques**

3.3.2 Effectifs observés et effectifs théoriques

Si l'on revient à notre exemple, les effectifs observés sont simples à obtenir, il s'agit de notre tableau croisé contenant les effectifs des différentes catégories. Ce sont donc les données déjà présentées :

Table 3.9: Pratique du bricolage par sexe. Effectifs théorique.

	Homme	Femme	Total
Non	515	631	1147
Oui	383	469	853
Total	899	1101	2000

Pour savoir quels sont les effectifs théoriques, il faut calculer l'équivalent du "50/50" pour la pièce de monnaie, mais dans le cas de notre tableau statistique. La question est donc de savoir, dans le cas où les deux variables ne sont pas corrélées, quels seraient les effectifs d'hommes et de femmes déclarant ou non pratiquer le bricolage.

Répondre à cette question est plus simple qu'il n'y paraît, car si les variables ne sont pas corrélées, la probabilité de pratiquer le bricolage doit être la même pour les hommes et pour les femmes, c'est donc le nombre d'individus déclarant bricoler (853), divisé par l'effectif total (2000). Pour obtenir l'effectif théorique d'hommes bricoleurs, il faut donc multiplier le nombre d'hommes dans notre échantillon (899) par la probabilité d'être bricoleur (853/2000). On obtient de la même façon toute la distribution conditionnelle théorique (remarquez bien qu'on ne change rien aux distributions marginales, elles sont identiques pour les effectifs observés et les effectifs théoriques).

$$n_{ij}^{th} = \frac{n_i * n_j}{n}$$

où n_{ij}^{th} est l'effectif théorique de la ligne i et de la colonne j (par exemple les hommes bricoleurs), n_i l'effectif de la colonne i (les bricoleurs et bricoleuses), n_j l'effectif de la colonne j (les hommes), et n l'effectif total.

3.3.3 Calcul du chi-2

Le résultat du test va dépendre de la différence entre les valeurs observées et les valeurs théoriques. Pour estimer ces différences, on calcule ce qu'on appelle le χ^2 de cette manière :

1. on prend la différence pour chaque case du tableau
2. on met ces différences au carré
3. on divise le résultat par les fréquences observées
4. on fait la somme de ces valeurs

$$\chi_2 = \sum_{ij} \frac{(n_{ij}^{th} - n_{ij}^{obs})^2}{n_{ij}^{th}}$$

Dans notre exemple, le chi-2 serait égal à :

$$\chi_2 = \frac{(805-881)^2}{805} + \frac{(1024-948)^2}{1024} + \frac{(539-462)^2}{539} + \frac{(685-762)^2}{685}$$

Mais on aura jamais à le faire à la main, les logiciels le font automatiquement. Une fois cette valeur obtenue, on peut presque répondre à la question. Il nous manque juste un élément : le lien entre cette valeur qui donne une idée de la différence entre les effectifs observés et les effectifs théoriques, et le risque d'erreur que l'on cherchait au début. Là il s'agit d'une question de mathématique qui est au-delà du niveau du cours et qu'il n'est pas nécessaire de maîtriser pour comprendre le principe du test. Admettons donc que nous sommes en mesure de déduire de cette valeur du χ^2 le risque d'erreur recherché.

Voici ce que nous indique R lorsqu'on lui demande de calculer le χ^2 pour le tableau présentant le bricolage en fonction du sexe :

```
table(hdv2003$sexe, hdv2003$bricol) %>% chisq.test()
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  .
## X-squared = 141.93, df = 1, p-value < 2.2e-16
```

R nous affiche d'abord le χ^2 que nous avons calculé plus haut à la main. `df` indique le nombre de **degrés de liberté** du tableau, qui correspond au nombre de colonnes moins 1 multiplié par le nombre de lignes moins 1 (ici $1*1=1$). Ce chiffre est nécessaire pour estimer le risque d'erreur (ce qu'on cherche) à partir de la valeur du χ^2, mais pour nous il n'est pas très important. Il affiche enfin la valeur recherchée, nommée *p-value* : il s'agit du risque de se tromper dans le cas où l'on rejette l'hypothèse nulle, c'est-à-dire l'hypothèse selon laquelle il n'y a pas de lien de corrélation entre les variables. Cette valeur est comprise entre 0 et 1. Pour un même nombre de degrés de liberté, plus la valeur du χ^2 est élevée, plus ce risque est faible. Ici il est égal à $2.2e-16$, ce qui signifie $0,000000000000022\%$.

Si l'échantillonnage est bien aléatoire, on peut donc affirmer avec confiance qu'au delà de l'échantillon observé, la pratique du jardinage est correlée au sexe des individus. En général, on se fixe un seuil de significativité *a priori* (par exemple 1%, ou $\alpha = 0.01$). Lorsque la *p-value* est inférieure à cette valeur, on va dire qu'on rejette l'hypothèse nulle avec un risque de 1%, ou encore que **la corrélation observée est significative au seuil de 1%**. Attention : si la *p-value* est supérieure à ce seuil, on ne peut pas conclure à la non significativé de la corrélation étudiée. On ne peut simplement pas affirmer avec le seuil de certitude choisi que la corrélation statistique est significative.

Cette méthode permet ainsi d'évaluer le risque d'erreur lorsqu'on cherche à généraliser une corrélation observée sur un échantillon à l'ensemble d'une population. On doit réaliser ce test à chaque fois qu'on veut commenter une relation

de corrélation entre deux variables qualitatives, car sinon on risque toujours de commenter en réalité des écarts qui sont liés au hasard de l'échantillonnage. Il faut enfin prendre garde à ne pas faire dire au χ^2 plus que ce qu'il ne permet d'affirmer. En particulier, le test ne dit rien sur l'intensité de la relation de corrélation, et cela quelle que soit la valeur de la *p-value* obtenue.

Références

Bibliography

Pierre Bourdieu. *La distinction: critique sociale du jugement*. Les Editions de minuit, Paris, France, 1979. ISSN: 0768-049X.

Fanny Bugeja-Bloch and Marie-Paule Couto. *Les méthodes quantitatives*. Que sais-je ? PUF, 2021. OCLC: 1285669386.

Alain Desrosières. Décrire l'État ou explorer la société : les deux sources de la statistique publique. *Geneses*, no 58(1):4–27, 2005. URL https://www.cairn.info/journal-geneses-2005-1-page-4.htm. Bibliographie_available: 0 Cairn-domain: www.cairn.info Cite Par_available: 1 Publisher: Belin.

Charles H. Feinstein and Mark Thomas. *Making History Count: A Primer in Quantitative Methods for Historians*. Cambridge University Press, Cambridge ; New York, 08 2002.

Claire Lemercier and Claire Zalc. *Méthodes quantitatives pour l'historien*. Number 507 in Repères. la Découverte, Paris, 2008.

Jean Peneff. *L'hôpital en urgence: étude par observation participante*. Métailié : Diffusion, Seuil, Paris, 1992.

Hadley Wickham. Tidy data. *Journal of Statistical Software*, 59:1–23, 09 2014. doi: 10.18637/jss.v059.i10. URL https://doi.org/10.18637/jss.v059.i10.