Méthodes quantitatives

Paul Hobeika

2022-03-01

Table des matières

À propos de ce document

Cette page accueille les notes de cours de méthodes quantitatives du M1 de Sciences 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 requiert peu de bagages préalables en statistique et en mathématique. Ce document a été généré par l'extension bookdown de Yihui Xie, et le code source est disponible sur GitHub.

Chapitre 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.

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

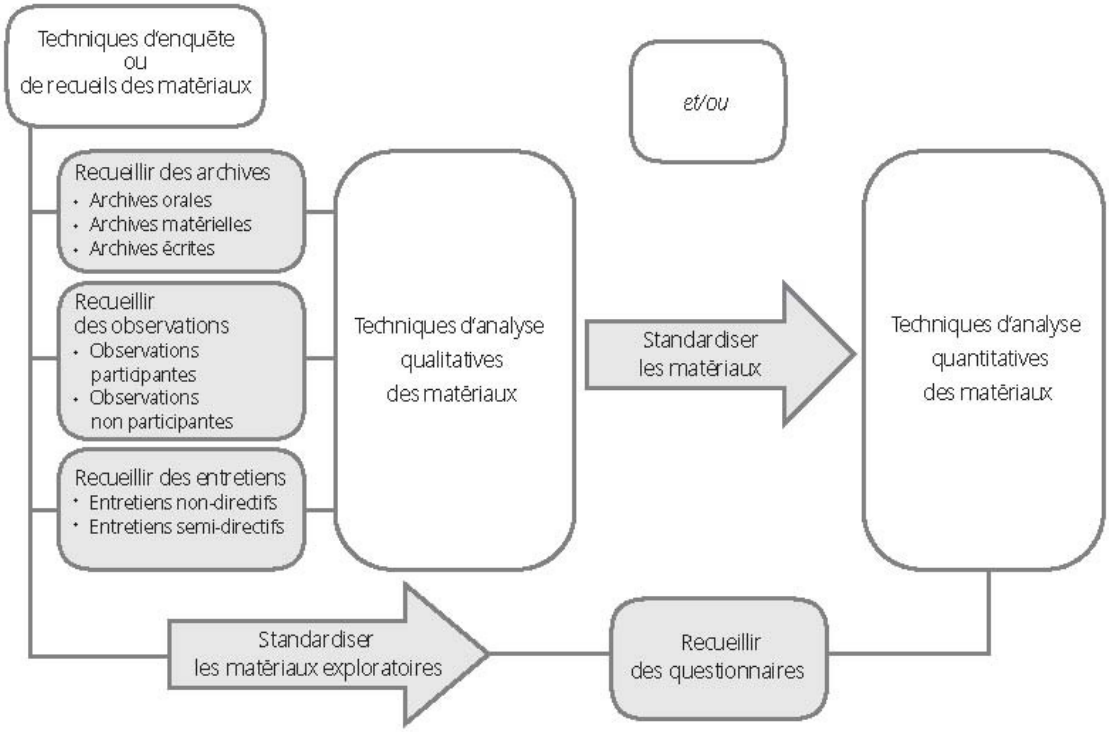


FIG. 1.1 – Techniques d'enquête et techniques d'analyse [Bugeja-Bloch and Couto, 2021]

[Lemercier and Zalc, 2008], à partir d'entretiens (le questionnaire en est un cas particulier) ou encore à partir d'observations [1].

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

[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].

TAB. 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

TAB. 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

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 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.

TAB. 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

1.2.3 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).

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.

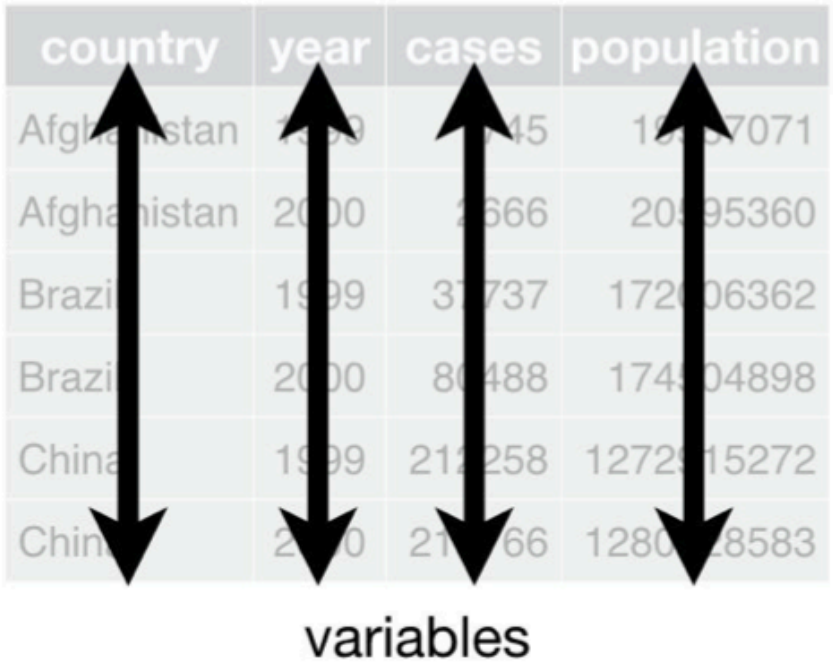


FIG. 1.2 – Tidy data sets

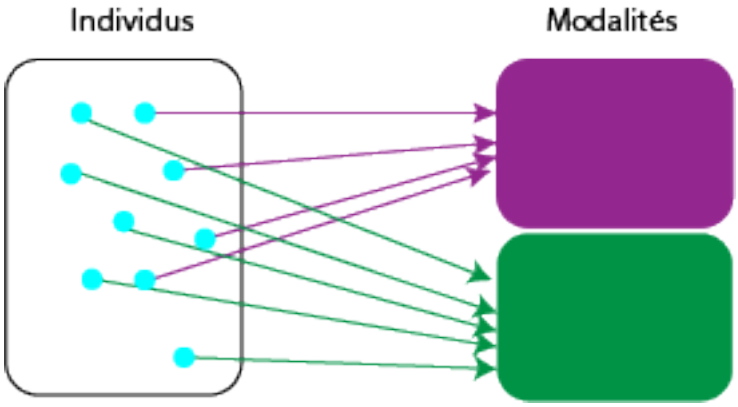


FIG. 1.3 – Les variables associent les individus à leur modalités

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.

Chapitre 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 calcule 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**.

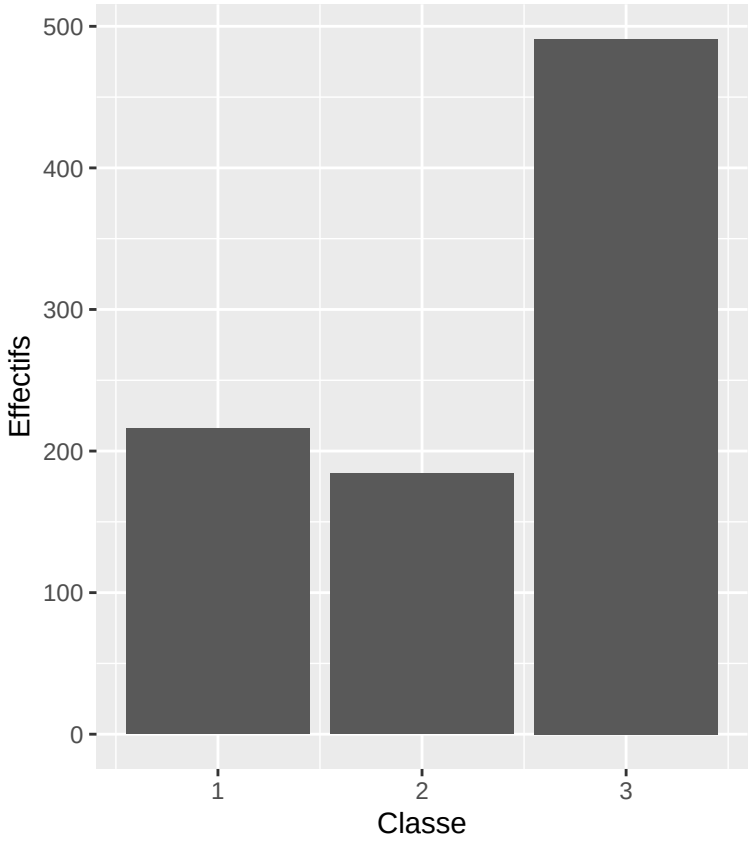
	n	%	%cum
1	216	24.2	24.2
2	184	20.7	44.9
3	491	55.1	100.0
Total	891	100.0	100.0

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] :

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

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 :

[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 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.

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

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 étudiantes 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**.

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

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.

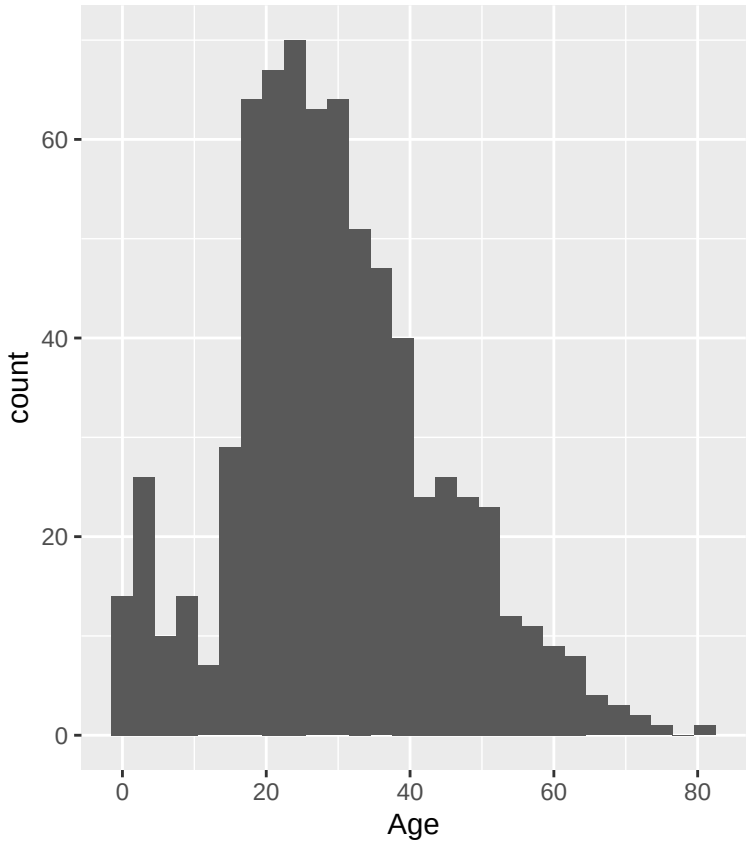


FIG. 2.1 – Distribution de l'âge des passagers du Titanic. Chaque rectangle a une largeur de 3 ans

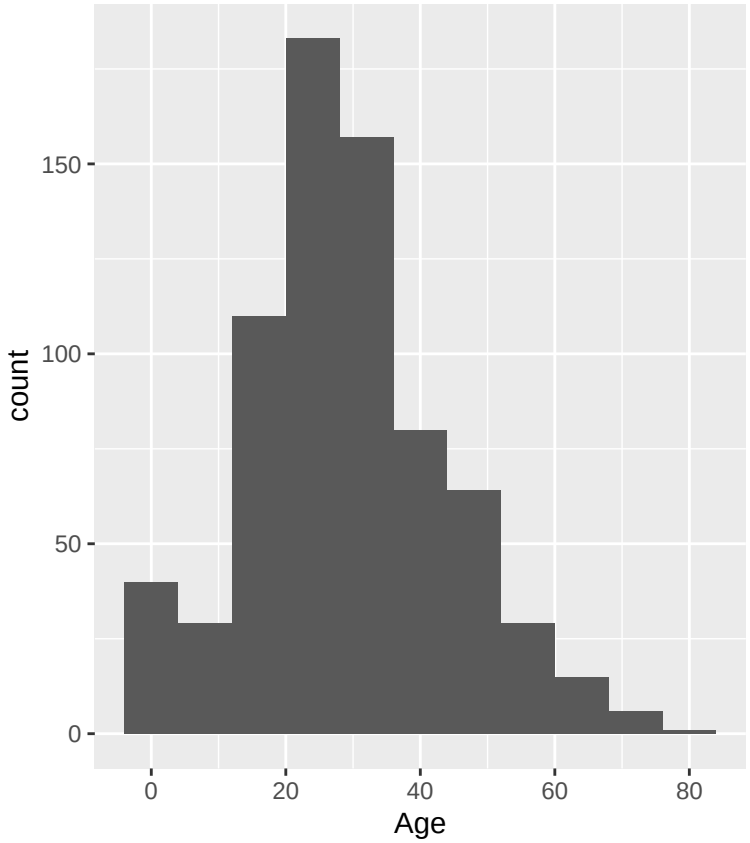


FIG. 2.2 – Même figure avec une largeur de 8 ans

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

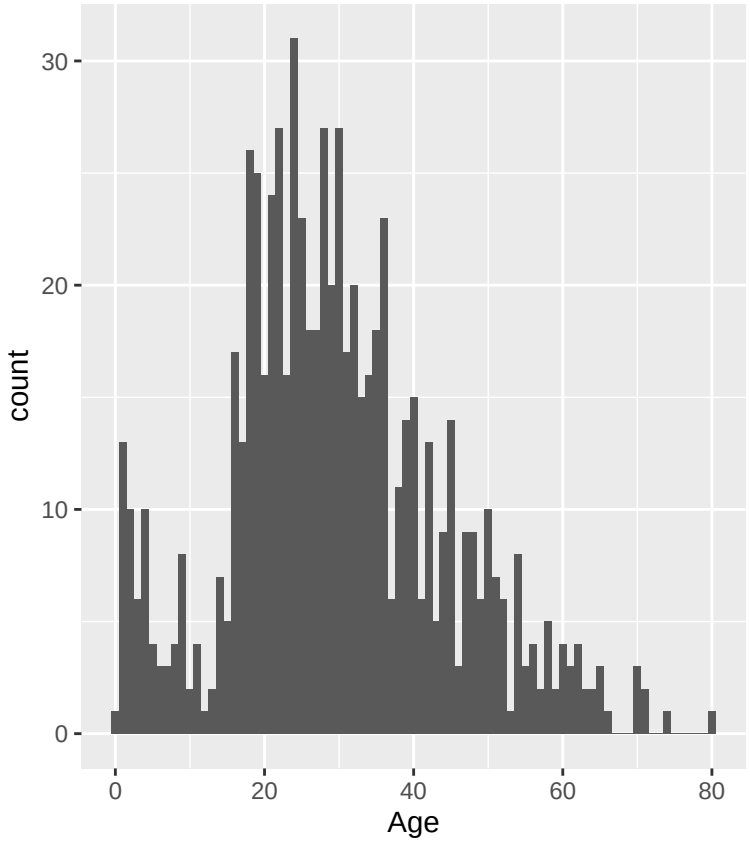


FIG. 2.3 – Même figure avec une largeur d'un an

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".

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*).

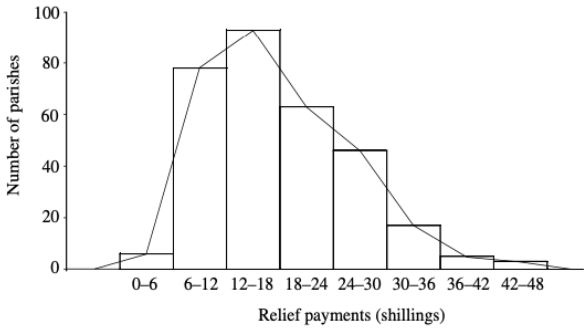


FIG. 2.4 – *Per capita* relief payments in 311 parishes in 1831 (Fenstein & Thomas, p. 41)

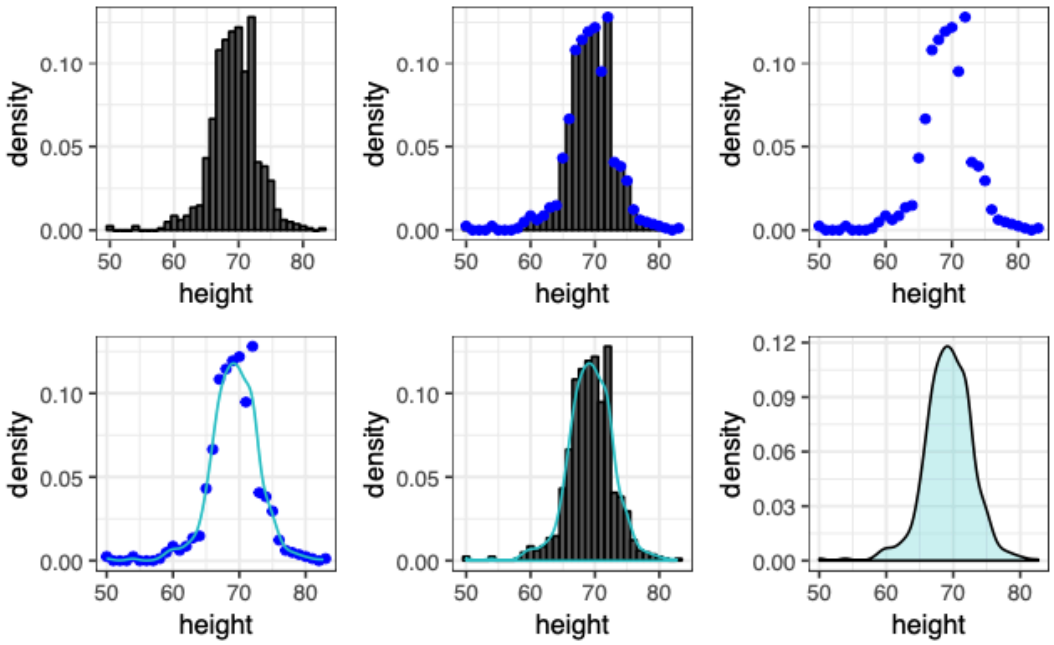


FIG. 2.5 – D'un histogramme à une densité de probabilité

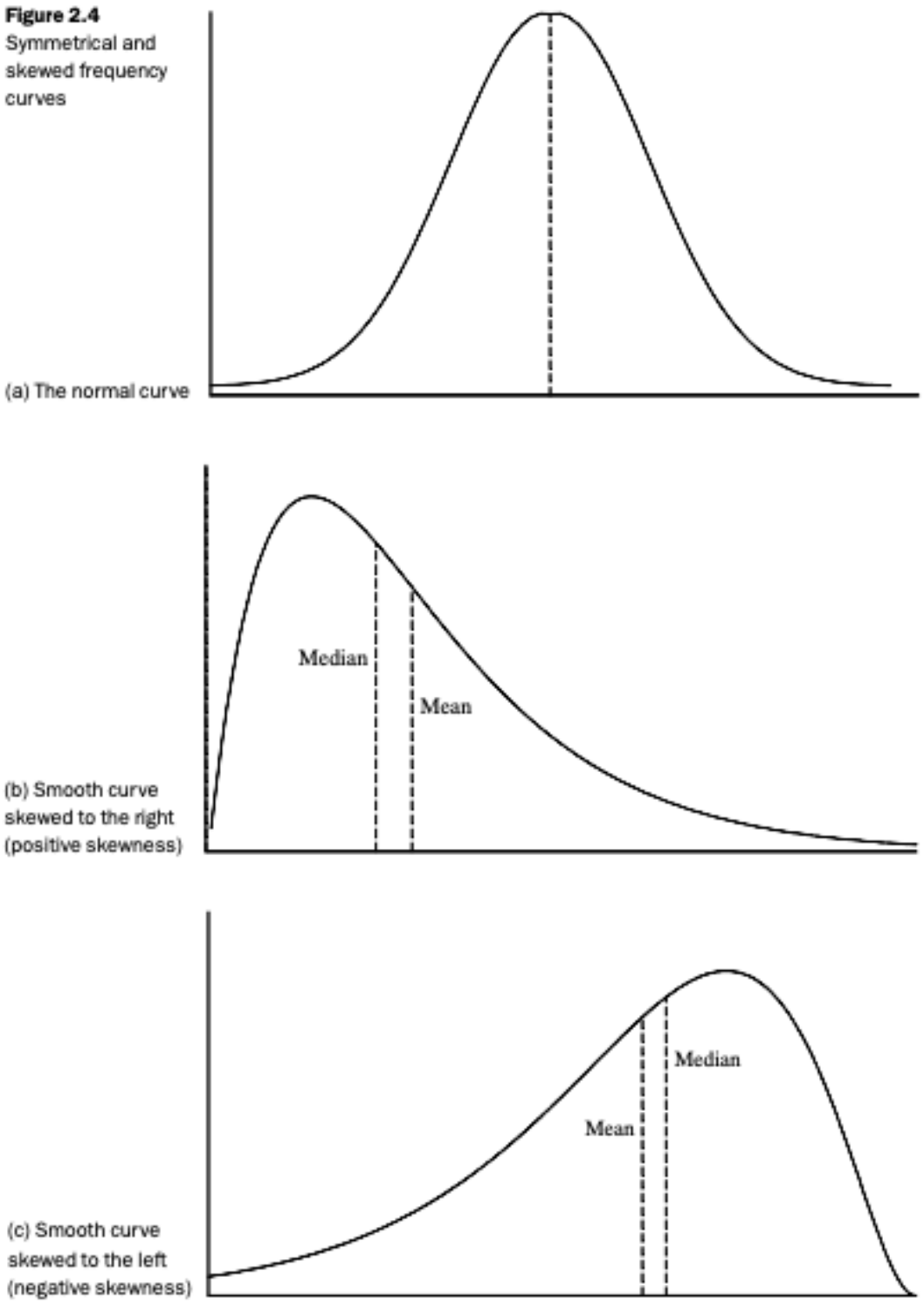


FIG. 2.6 – Distributions symétriques et asymétriques [Feinstein and Thomas, 2002, p.54]

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.

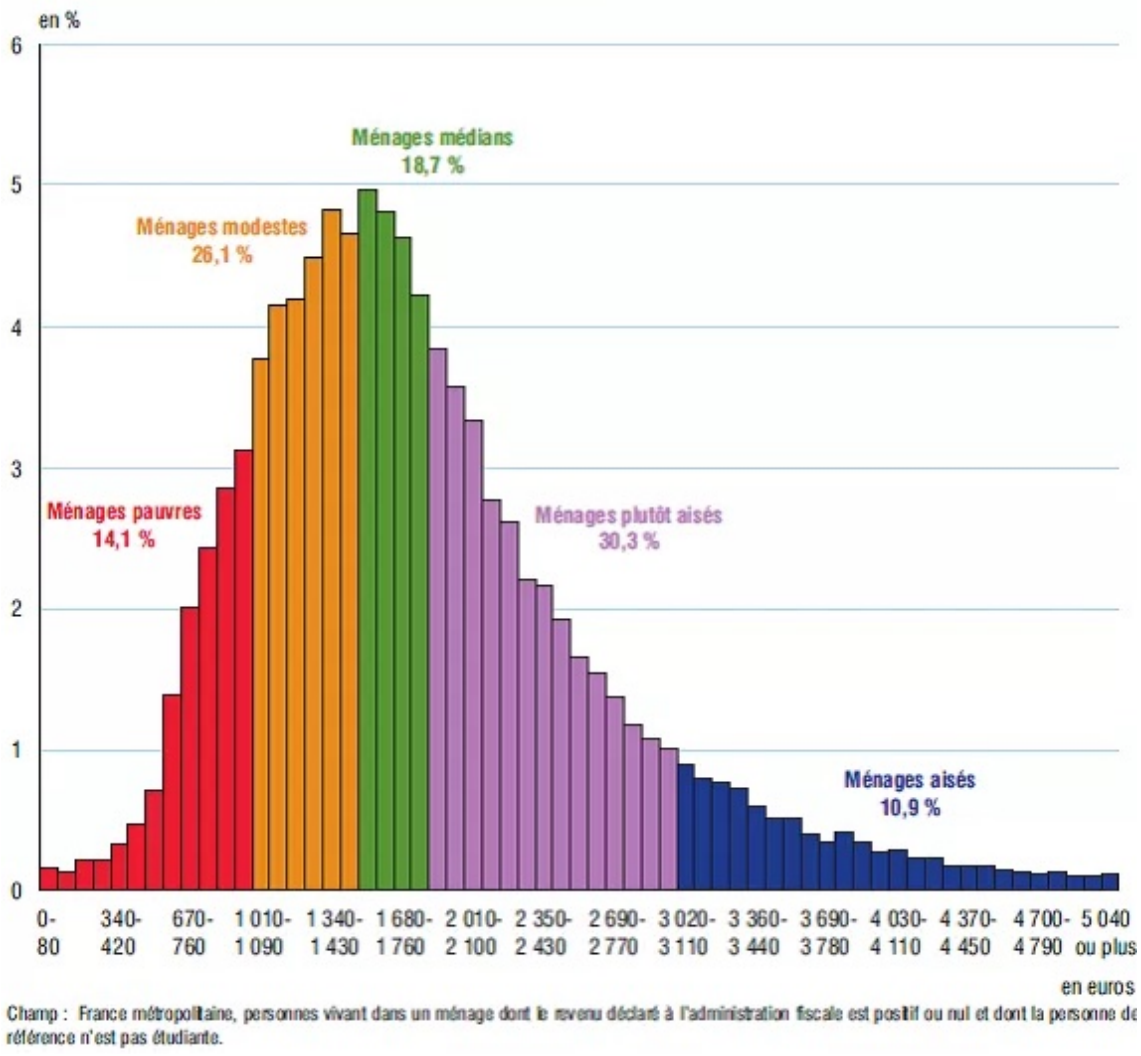


FIG. 2.7 – Distribution des niveaux de vie mensuels en 2014 en France (Source : Insee, Portrait social 2014)

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")
```

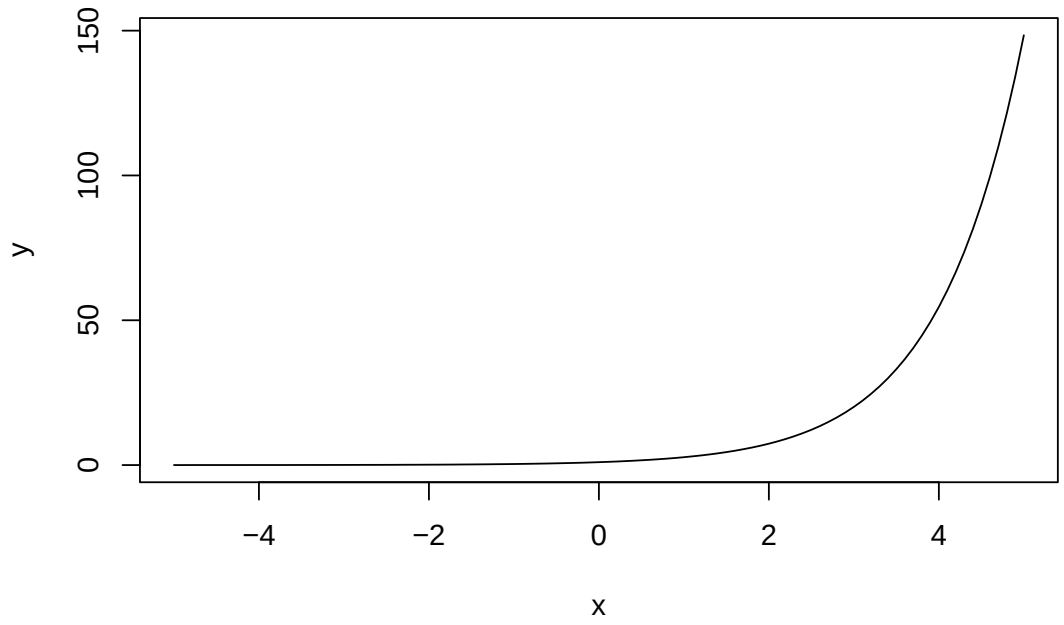
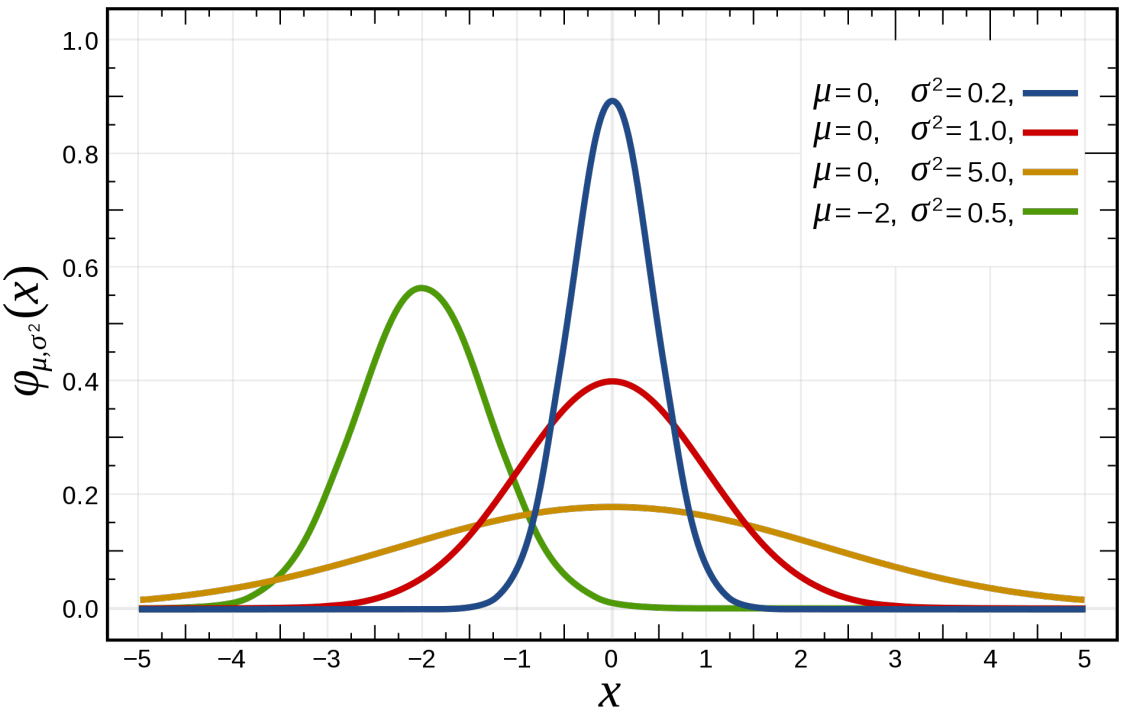


FIG. 2.8 – Graphe de la fonction exponentielle entre -5 et 5

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 une sorte de dos d'âne, ce qui explique qu'on l'appelle aussi "la courbe en cloche".



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.

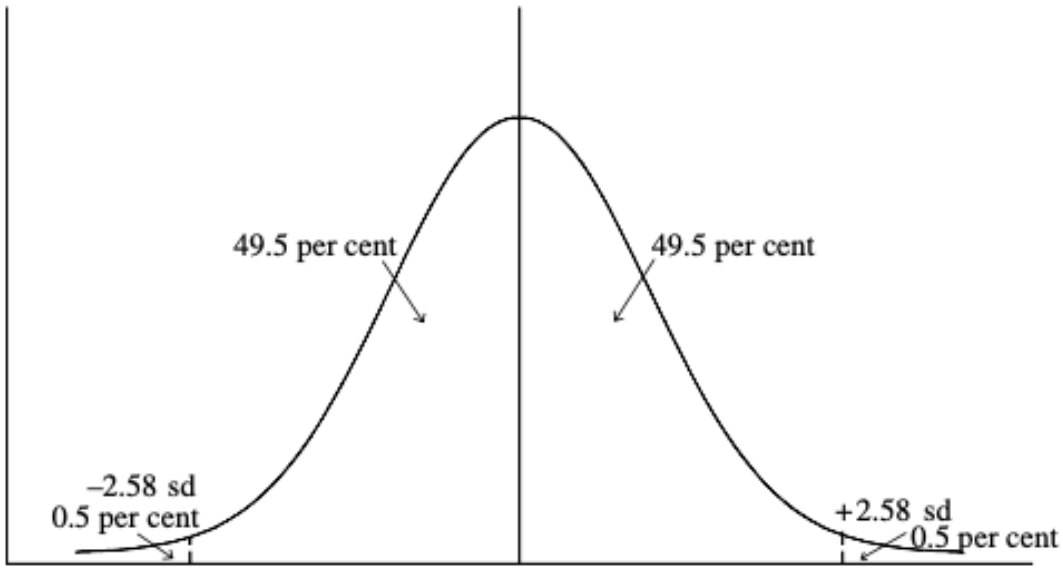


FIG. 2.9 – Aire sous la courbe

Chapitre 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 socio-professionnelle 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)

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

TAB. 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

TAB. 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**

TAB. 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

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.

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**.

TAB. 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

TAB. 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

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, 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

TAB. 3.6 – Pratique du bricolage par sexe

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

TAB. 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

bricolage.

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

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 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**.

TAB. 3.8 – Pratique du bricolage par sexe. Effectifs observés.

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

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 :

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.

TAB. 3.9 – Pratique du bricolage par sexe. Effectifs théorique.

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

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 :

$$\chi_2 = \frac{(515-384)^2}{515} + \frac{(631-515)^2}{631} + \frac{(383-763)^2}{383} + \frac{(469-338)^2}{469} = 141.93$$

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.

Chapitre 4

Inférence et variables quantitatives

Lors du dernier cours, on a abordé certaines notions de statistique inférentielle à partir de l'étude de la corrélation entre deux variables quantitatives. Le cours de cette semaine est consacré à la présentation de l'analyse inférentielle d'une variable quantitative. L'enjeu est de savoir, lorsqu'on réalise des statistiques sur un échantillon, dans quelle mesure il sera possible de généraliser les résultats tels que la moyenne observée d'une variable quantitative. Avant de traiter cette question, je vous présente plus en détail certaines notions de statistique inférentielle déjà abordées, telles que les méthodes d'échantillonnage.

4.1 Méthodes d'échantillonage

La statistique inférentielle repose sur la méthode dite des sondages, qui consiste à étudier une population à partir d'un échantillon d'individus sélectionné parmi l'ensemble de cette population.

Savoir si oui ou non il est possible de généraliser les résultats obtenus à partir de l'étude d'un échantillon dépend largement de la méthode d'échantillonnage. Il existe en effet des bons et des mauvais échantillons. Contrairement à une croyance longtemps répandue dans la pratique de la statistique, la taille de l'échantillon n'est pas le facteur déterminant de la qualité d'un échantillon. C'est ce qu'a montré George Gallup en 1936, qui a prédit la victoire de Roosevelt aux élections présidentielles états-uniennes à partir d'un échantillon comportant 5000 individus. Dans le même temps, les grands journaux états-uniens prédisaient la victoire de son concurrent à partir de la collecte de plus de deux millions d'intentions de vote. S'il faut bien sur un effectif minimal, il vaut donc mieux avoir un petit échantillon de bonne qualité plutôt qu'un gros échantillon de mauvaise qualité.

Quels sont alors les éléments qui font la qualité d'un échantillon ? Un échantillon est de bonne qualité lorsqu'il a la même structure que la population. On dit alors qu'il est **représentatif**. On peut dire par exemple que l'échantillon constitué par les lecteurs des quotidiens états-uniens est **non-représentatif** : les individus de cet échantillon auront des caractéristiques particulières par rapport au reste de la population (par exemple voter plus fréquemment pour le parti républicain). Leurs intentions de vote ne permettent donc pas de prédire les résultats aux élections présidentielles. De la même manière, l'enquête sur les classes sociales menée par l'équipe de sociologues réuni·es autour de Mike Savage au début des années 2010, la *Great british class survey*, a été réalisée par internet en collaboration avec la BBC qui en faisait la promotion à l'aide de spots télévisés. Malgré le succès de l'enquête mesurée en termes du nombre de personnes qui ont participé, l'un des premiers résultats est que les classes popu-

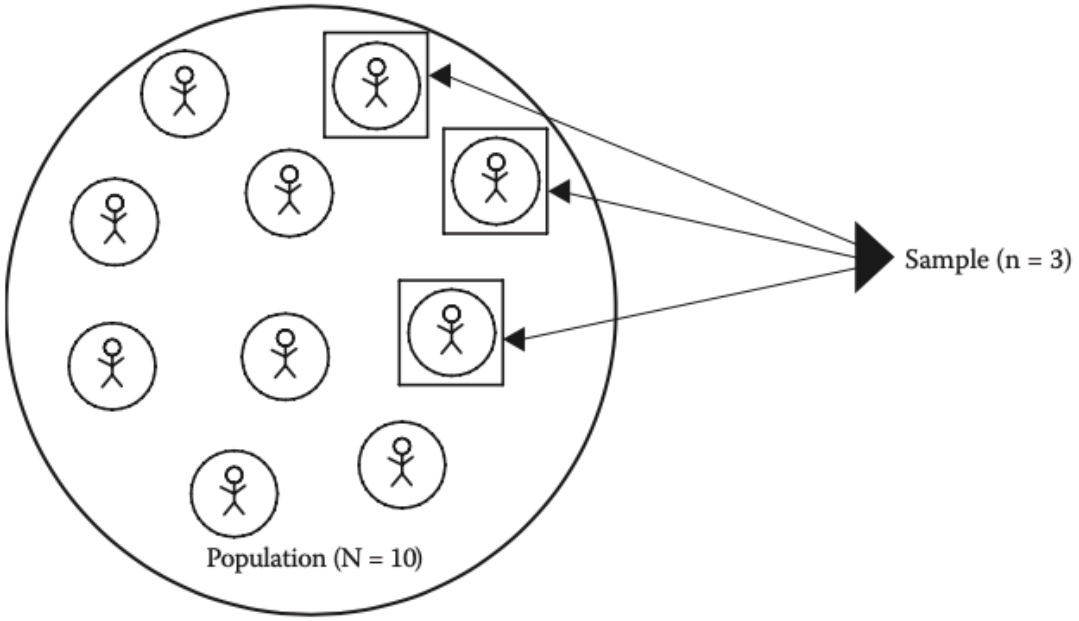


FIG. 4.1 – On utilise souvent des données qui concernent un échantillon pour décrire une population plus large

laires sont largement sous représentées dans l'échantillon, et en particulier les fractions les plus précaires [Savage et al., 2013].

4.1.1 Les échantillons aléatoires

Il existe en pratique plusieurs méthodes d'échantillonnage qui permettent de produire des échantillons de plus ou moins bonne qualité. La meilleure méthode d'un point de vue statistique est de **sélectionner l'échantillon de manière aléatoire**.

Une première manière de faire est de réaliser un tirage aléatoire lors duquel tous les individus de la population ont la même probabilité d'être choisis. On parle alors d'**échantillonnage aléatoire simple**. Cette méthode n'est pas toujours possible à mettre en œuvre car elle implique de disposer d'une liste exhaustive de la population à laquelle on s'intéresse. À partir de ce document qu'on appelle une *base de sondage*, on peut réaliser un tel tirage aléatoire. Dans le cas de la population française, la seule institution à disposer d'une telle liste est l'Insee, ce qui lui confère un forme de monopole sur la production d'échantillons aléatoires, et qui explique que de nombreuses enquêtes produites par l'Ined ou l'Inserm le sont en partenariat avec l'Insee [Bugeja-Bloch and Couto, 2021, p.66].

Dans certains cas, plutôt qu'attribuer une même probabilité de tirage à tous les individus, on souhaite que certaines catégories d'individus soient surreprésentées dans l'échantillon. On parle alors d'**échantillon aléatoire stratifié**. Un tel échantillon n'est pas représentatif de la population dans son ensemble, mais chaque strate (les différentes catégories d'individus auxquels on a attribué des probabilités différentes d'être sélectionnés) est représentative de la catégorie d'individus qu'elle représente. Par exemple, l'échantillon de l'enquête sur la sexualité des français conduite par l'Ined en 2008 surreprésente volontairement les jeunes, ce qui permet de produire des analyses détaillées de cette sous-population [Toulemon and Razafindratsima, 2008].

Enfin, une manière de produire un échantillonnage aléatoire sans disposer d'une base de

sondage portant sur les individus est de réaliser ce qu'on appelle un **échantillonnage par grappe** (ou échantillonnage aréolaire). Cela est possible lorsque les individus sont réunis naturellement en groupes relativement homogènes, et que l'on dispose d'une liste exhaustive de ces groupes, dont la nature peut être variée. Par exemple, l'enquête **Sans-Domicile 2001** de l'Insee est basée sur un échantillonnage des usagers des services de distributions de repas chaud en hiver. Cet échantillonnage est aréolaire car il repose sur un échantillonnage des villes de plus de 20000 habitants dans lesquels sont localisés ces services, et un autre échantillonnage des distributions de repas eux-mêmes [Brousse, 2005].

4.1.2 Les échantillons non aléatoires

Lorsque l'échantillon n'est pas réalisé selon une des méthodes décrite ci-dessus, il n'est pas aléatoire. Les individus n'ont alors pas tous la même probabilité de faire partie de l'échantillon. La méthode non aléatoire la plus utilisée est la méthode dite **des quotas**, mise en œuvre notamment par les instituts de sondage (IFOP, IPSOS, etc.). Dans ce cas, l'échantillon est constitué à partir d'une définition *a priori* des critères importants de représentativité de la population (sexe, age, catégories socioprofessionnelles par exemple). Il faut donc connaître certaines caractéristiques de la population de référence pour construire un échantillon par quota. L'échantillon ne sera toutefois représentatif que des critères précis sélectionnés en amont, contrairement à un échantillon aléatoire, qui est représentatif quel que soit le critère envisagé (et cela sans avoir à spécifier aucun de ces critères). Un défaut important de ce mode d'échantillonnage est que les individus qui ne souhaitent pas répondre disparaissent de l'échantillon, et il n'est donc pas possible d'en décrire les caractéristiques [Lehingue, 2007].

Vous pouvez garder en tête que les méthodes de la statistique inférentielle supposent en général que l'on dispose d'un échantillon aléatoire.

4.2 Vocabulaire de la statistique inférentielle

4.2.1 Paramètres et estimateurs

Revenons maintenant à la question qui nous intéresse. On suppose qu'on dispose d'un échantillon aléatoire et d'une variable quantitative permettant de décrire une caractéristique des individus (par exemple leur taille). On s'intéresse à la différence entre ce que l'on mesure sur cet échantillon (par exemple leur taille moyenne) et la valeur qu'on cherche à décrire pour l'ensemble de la population. Pour distinguer ces deux objets, on utilise des termes et des notations différentes :

— Lorsqu'on mesure une grandeur relative à l'échantillon, on parle de **statistique** ou d'**estimateur**. Nous revenons sur la différence entre ces deux termes plus bas.
— Mais lorsqu'on veut désigner la grandeur correspondante pour la population, on parle de **paramètre**.

Pour signifier visuellement la différence entre paramètre et statistiques, on utilise des lettres différentes : les paramètres sont généralement désignés par des lettres grecques, tandis que les estimateurs et statistiques ont des lettres romaines et parfois des traits ou des chapeaux.

Par exemple, pour la moyenne de la population, on utilise souvent la notation μ :

$$\mu = \frac{1}{N} \sum_{i=1}^{N} X_i$$

où la somme porte sur l'ensemble de la population, dont l'effectif est noté N. Et on note $\overline{X}$ la moyenne de l'échantillon :

TAB. 4.1 – Les tailles des 10 individus de notre population

x
168
189
179
154
192
183
167
183
179
173

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

La somme porte ici sur l'ensemble de l'échantillon, dont l'effectif est noté n. On utilisera σ pour désigner l'écart-type de la population et s pour désigner l'écart-type de l'échantillon. Gardez en tête que les notations peuvent varier selon les auteurs.

4.2.2 Distribution d'échantillonage

Si les paramètres (ici μ) ont une valeur unique, les statistiques ($\overline{X}$) dépendent de l'échantillon sélectionné (aléatoirement) parmi notre population. Supposons que l'on chercher à estimer la taille moyenne d'une population de taille N = 10, avec un échantillon de taille n = 3.

Il y a plusieurs manières de choisir trois individus (donc trois tailles) parmi cette liste de 10. Il existe d'ailleurs une formule qui permet de calculer le nombre de manières différentes de former un échantillon de 3 individus distincts parmi 10 individus :

$$\binom{10}{3} = \frac{10!}{(10-3)!3!} = 120$$

Où $\binom{10}{3}$ qui se lit "trois parmi dix" est une notation usuelle pour résumer la formule indiquée dans le deuxième terme de l'équation. La notation $n!$, lue "factorielle n", désigne quant à elle le produit de tous les entiers inférieurs ou égal à n, c'est-à-dire : $n! = n*(n-1)*(n-2)*...*2*1$, par exemple $3! = 3*2*1 = 6$. Vous pouvez bien sûr oublier ça, je l'évoque simplement pour indiquer qu'il existe 120 manières différentes de sélectionner 3 individus parmi 10. Bien sûr, toutes ne donneront pas la même moyenne. Ce qui nous intéresse, c'est alors de savoir comment se distribuent ces 120 moyennes différentes, car si l'on connaît cette distribution, cela donne une idée de la probabilité d'obtenir une moyenne proche de celle qu'on cherche à estimer, c'est-à-dire la moyenne μ des tailles des 10 individus qui constituent la population.

On peut lister toutes les manières de sélectionner 3 tailles parmi les 10. Les premières pourraient être :

Taille1	Taille2	Taille3
168	189	179
168	189	154
168	189	192
168	189	183
168	189	167
168	189	183

Chacun de ces 120 échantillons a sa propre moyenne. Si l'on calcule ces 120 moyennes, on obtient une liste de valeurs moyennes dont on peut représenter la distribution par un histogramme (4.2). On l'appelle **distribution d'échantillonnage de la moyenne**.

Taille1	Taille2	Taille3	Moyenne
168	189	179	178.6667
168	189	154	170.3333
168	189	192	183.0000
168	189	183	180.0000
168	189	167	174.6667
168	189	183	180.0000

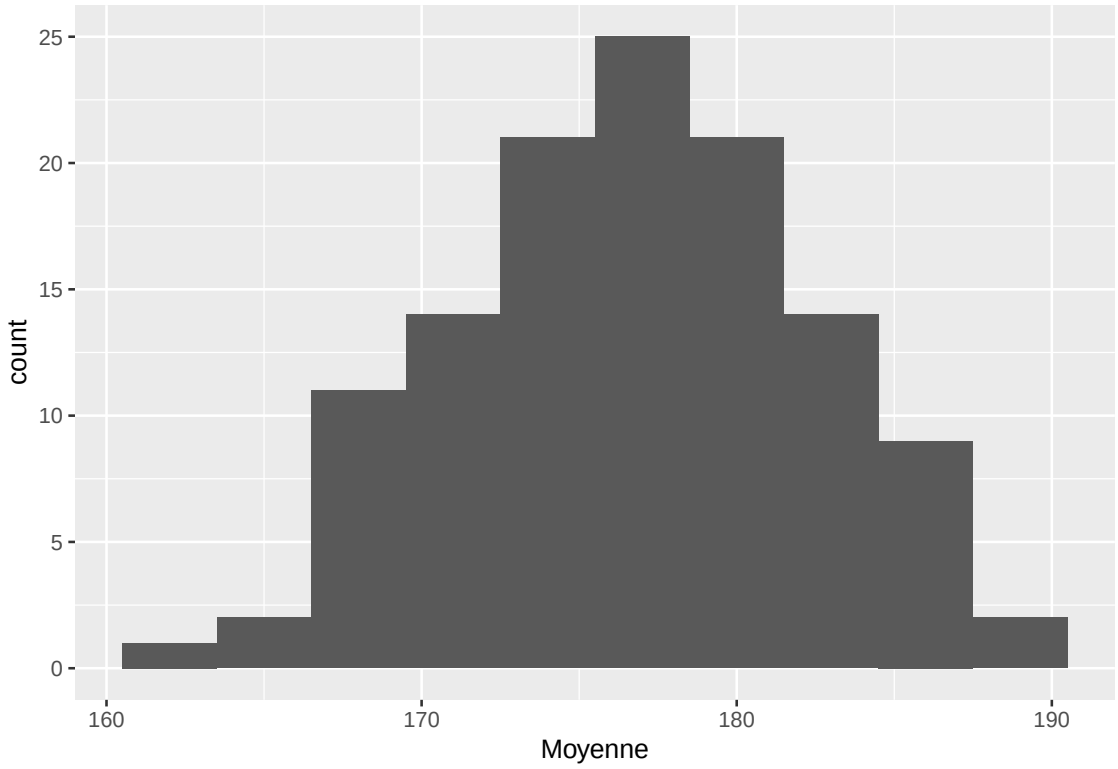


FIG. 4.2 – Distribution d'échantillonage de la moyenne

4.2.3 Erreur type

Comment cette distribution nous aide-t-elle à savoir si la moyenne estimée à partir d'un échantillon $\overline{X}$ est une bonne estimation de la moyenne de la population μ? Pour le savoir, on va d'abord admettre que la moyenne de cette distribution d'échantillonnage est la moyenne de la population. Ce qui permet alors de reformuler la question : on se demande alors quelle

sera en moyenne la distance entre notre estimation $\overline{X}$, qui est une valeur aléatoire de cette distribution, et la moyenne de la population μ.

Si vous vous souvenez du cours d'il y a deux semaines, il s'agit justement de la définition de l'**écart-type**, qui mesure un écart moyen à la moyenne. L'écart-type de la distribution d'échantillonnage de la moyenne (la distribution 4.2) nous donne donc l'erreur moyenne que l'on va réaliser en estimant la moyenne de la population μ à l'aide de la moyenne de l'échantillon $\overline{X}$. Il s'agit d'une grandeur importante, à laquelle on attribue le nom d'**erreur type**. Plus l'erreur type est importante, plus l'estimation réalisée à partir d'un échantillon sera en moyenne éloignée du paramètre que l'on cherche à mesurer.

Le problème, c'est qu'en général on ne va pas avoir accès à la distribution d'échantillonnage de la moyenne pour calculer l'erreur type, car on cherche justement à produire une estimation à partir de la seule connaissance d'un échantillon. Ici on doit utiliser un résultat de la théorie des probabilité, qui nous indique que si l'on fait l'hypothèse de l'indépendance statistique entre les différents individus de notre échantillon (ici que la taille d'un individu ne dépendra pas de la taille d'un autre), on peut estimer l'erreur type (notée SE, pour *standard error*) comme **l'écart type de l'échantillon divisé par la racine carré de son effectif**.

$$SE \approx \frac{s}{\sqrt{n}}$$

où :

— s est l'écart type de l'échantillon
— n l'effectif de l'échantillon

4.2.4 Théorème central limite

Ce résultat est une conséquence du **théorème central limite**, qui est un résultat fondamental de la théorie des probabilités, sans lequel il n'existerait pas de statistique inférentielle. Ce théorème indique que, pour toute variable quantitative de notre population (quelle que soit sa distribution), la distribution d'échantillonnage de la moyenne de cette variable va *tendre vers* une loi normale de moyenne μ et d'écart-type $\sigma/\sqrt{n}$ lorsque la taille de l'échantillon augmente [1]. En particulier, ce résultat permet de savoir que l'erreur que l'on va commettre, mesurée par l'écart-type de notre distribution d'échantillonnage, va diminuer lorsque n augmente. C'est un résultat relativement intuitif : plus la taille de l'échantillon augmente, plus l'estimateur (la moyenne de l'échantillon) sera proche du paramètre estimé (la moyenne de la population).

Formulé en termes abstraits, ce résultat peut être difficile à comprendre. Mais certaines représentations graphiques peuvent en donner une bonne intuition. Ci-dessous, on représente des distributions de probabilité pour un tirage aléatoire d'une pièce à pile ou face. On imagine qu'on réalise une série de tirages, en comptant 1 lorsqu'on obtient face et 0 sinon, puis on additionne tous les résultats. Les différentes courbes représentent les probabilités d'obtenir un total égal au chiffre indiqué en abscisses pour différents nombre de tirages. Lorsqu'on réalise un seul tirage, on a 50% de chance d'obtenir 0 et 50% de chance d'obtenir 1 (d'où le segment horizontal). Puis si on en fait un second, on a 25% de chance d'obtenir deux fois pile (0), 25% d'obtenir deux fois face (2), et 50% d'obtenir une fois pile et une fois face (1). Et ainsi de suite.

Remarquez qu'ici l'écart type de la distribution (la "largeur" de la cloche) augmente avec le nombre de tirages, car la distribution représentée correspond à la somme des valeurs obtenues

[1] 'Tendre vers' renvoie à la notion de limite en mathématique. Si l'on considère une suite de nombres réels x_n, on dit que la suite des valeurs $(x_1, x_2, ..., x_n)$ tend vers un nombre l si pour tout nombre ϵ (aussi petit soit-il), il existe un nombre entier k telle que pour tout $n > k$ on a $x_n - l < \epsilon$. On dit alors que l est la limite de la suite x_n quand n tend vers l'infini, et on note $\lim_{n\to\infty} x_n = l$. Par exemple, $x_n = 1/n$ tend vers 0 : on le montre facilement à partir de la définition, car si on se donne un nombre $\epsilon > 0$ (par exemple, 0,1) on sait que dès que $n > 1/\epsilon$ (10 dans notre exemple), on aura $x_n = 1/n < \epsilon$.

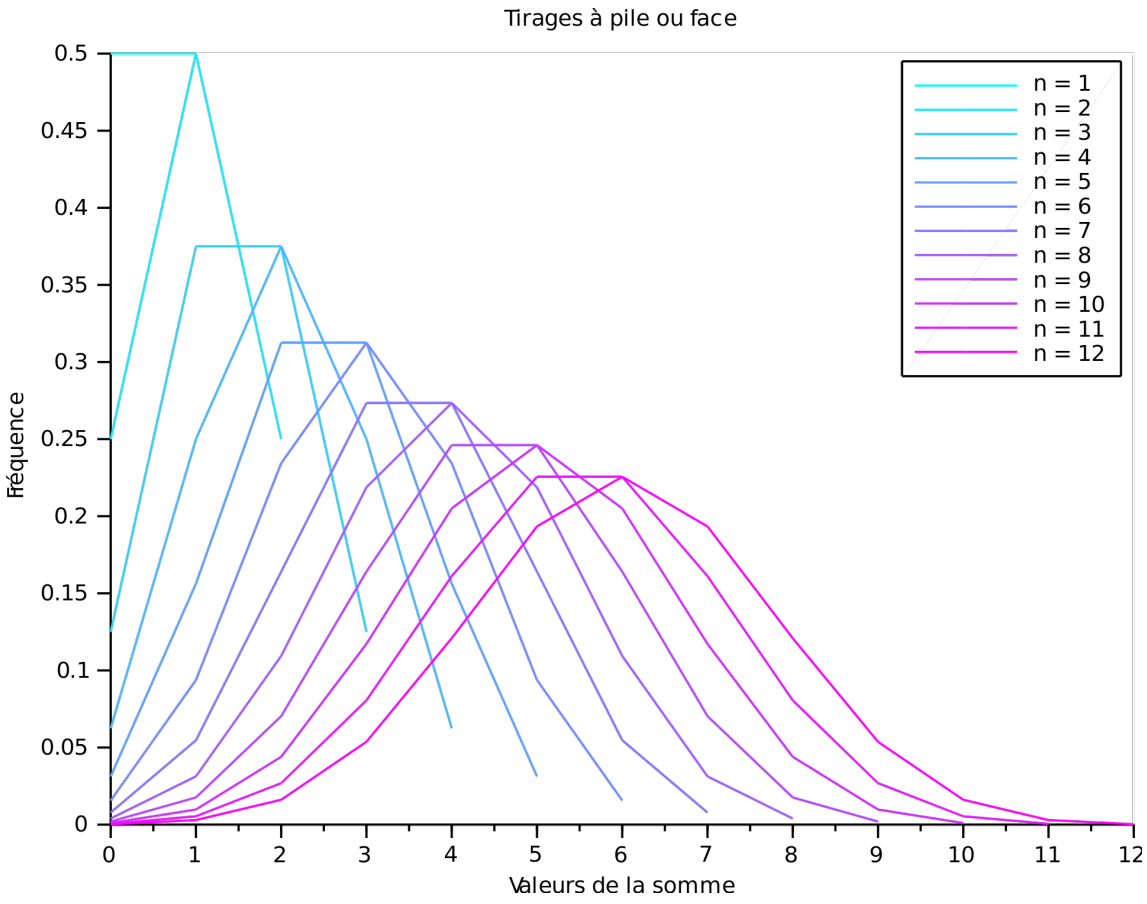


FIG. 4.3 – Fréquence d'apparition de la somme des valeurs d'un tirage alétaoire de 0 et 1

et non à leur moyenne. Pour obtenir la distribution de la moyenne obtenue, il faudrait diviser la somme obtenu par le nombre de tirages : on obtiendrait alors des courbes entre 0 et 1, centrées sur 0,5, et dont l'écart-type diminue avec le nombre de tirages comme évoqué plus haut.

Cela permet d'avoir une bonne idée du résultat auquel on aboutirait si on cherchait à représenter les distributions d'échantillonnage d'une variable quantitative qui ne prend que deux valeurs, pour différentes tailles d'échantillon. Alors que la distribution de cette variable n'a rien de normale au sein de la population, la distribution d'échantillonnage deviendrait de plus en plus proche d'une loi normale à mesure que l'on augmenterait l'effectif de l'échantillon.

4.3 Intervalles de confiance

4.3.1 *z*-distribution

Le théorème central limite permet donc d'estimer l'erreur type d'une variable quantitative relative à des données produites sur un échantillon, et donc de quantifier l'incertitude des grandeurs mesurées lorsqu'il s'agit de les généraliser à l'ensemble de la population. On souhaiterait maintenant avoir un moyen de présenter cette incertitude de manière la plus explicite possible. Pour cela, on va utiliser les propriétés de la loi normale.

Comme nous l'avons évoqué dans l'avant dernier cours (section 2.3), lorsque l'on sait qu'une variable est distribuée de manière normale, on peut connaître la probabilité qu'une valeur sélectionnée de manière aléatoire se trouve à moins d'une certaine distance de la moyenne. Par exemple, 99% des valeurs se situent à moins de 2,58 écart-type de la moyenne. Ces chiffres permettent de préciser des **intervalles de confiance** relatifs à une estimation. On va dire qu'on est la moyenne μ se trouve dans l'intervalle $[\overline{X} - 2,58 * SE \ ; \ \overline{X} + 2,58 * SE]$ avec une certitude de 99%.

Disons maintenant que l'on cherche l'intervalle de confiance à 95% plutôt qu'à 99%. Comment l'obtenir ? Pour cela on définit parfois à partir de notre distribution d'échantillonnage une nouvelle distribution que l'on nomme z-distribution de cette manière :

$$z = \frac{\overline{X} - \mu}{SE} = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

Si $\overline{X}$ est distribuée selon une loi normale de moyenne μ et d'écart-type $SE = \sigma/\sqrt{n}$, alors z sera distribuée selon une loi normale de moyenne 0 et d'écart-type égal à 1, ce qu'on appelle la **loi normale centrée réduite**. C'est une manière de toujours se ramener à la même distribution. On parle de **standardisation**. En se ramenant à une loi normale, on peut donc obtenir à partir d'une table de valeurs l'équivalent du chiffre 2,58 pour n'importe quelle pourcentage de certitude attendu.

Dans R, la fonction `qnorm()` permet par exemple d'obtenir ces valeurs :

```
qnorm(p = 0.975)
```

```
## [1] 1.959964
```

Cette fonction permet de déterminer n'importe quel quantile de la loi normale centrée réduite. C'est-à-dire qu'elle donne ici la valeur qui sépare les 97,5% de valeurs les plus faibles des 2,5% les plus élevées. On utilise 97,5% plutôt que 95% on souhaite également délimiter les 2,5% des valeurs les plus faibles. Comme la courbe est symétrique par rapport à 0, la valeur équivalente sera toujours l'opposé de celle cherchée ici. On peut le vérifier :

```
qnorm(p = 0.025)
```

```
## [1] -1.959964
```

Graphiquement, les deux valeurs obtenues permettent de délimiter 95% de l'aire comprise sous le graphe de la loi normale centrée réduite. Sur le graphique suivant, on représente en rouge les 2,5% de l'aire sous la courbe qui représentent les valeurs extrêmes de la distribution.

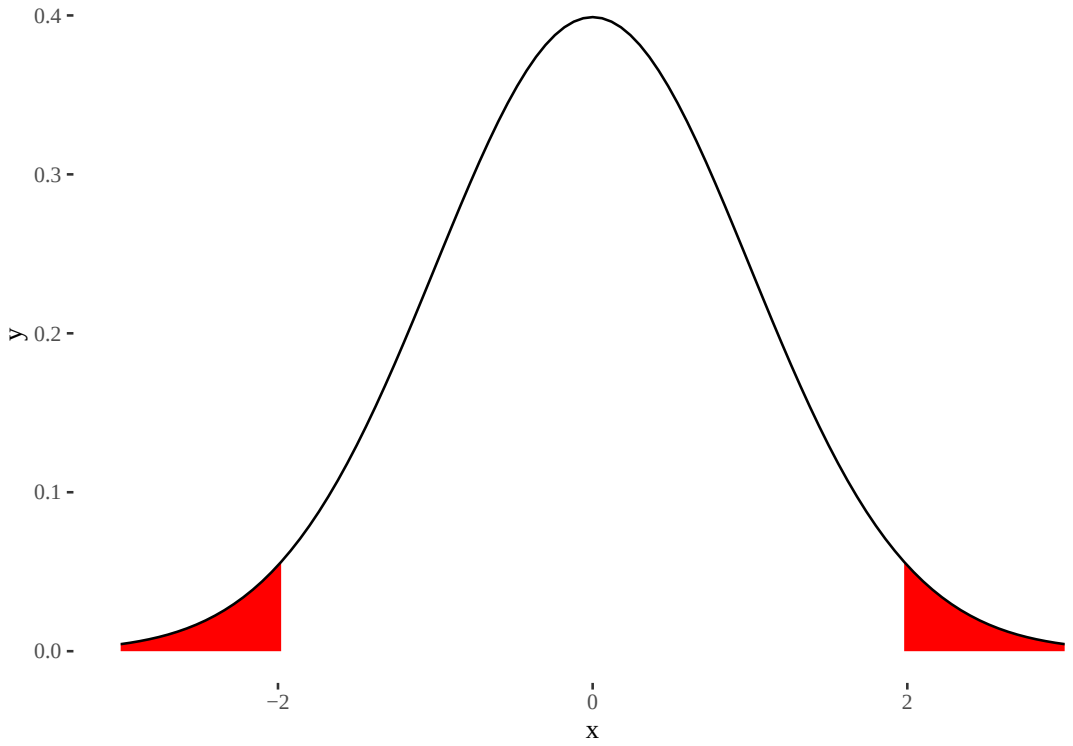


FIG. 4.4 – Représentation graphique de la loi normale centrée réduite, et en rouge des 5% de l'aire sous la courbe situées aux deux extrêmes de la distribution

4.3.2 *t*-distribution

En réalité, cette manière d'estimer un intervalle de confiance n'est pas la plus précise. Nous avons vu plus haut que $SE \approx s/\sqrt{n}$, où s est l'écart-type de l'échantillon et n son effectif. Les statisticiens ont montré que cette approximation n'est pas la meilleure estimation de l'erreur-type. D'après le théorème central limite, on sait que l'erreur type est égale à $SE = \sigma/\sqrt{n}$ où σ est l'écart-type de la population. L'approximation qui est discutable est donc celle qui consiste à estimer l'erreur type σ par la valeur $s/\sqrt{n}$.

C'est ici qu'on peut préciser la différence entre **statistique** et **estimateur**. Une statistique est la valeur mesurée sur notre échantillon, par exemple l'écart-type de l'échantillon s. Lorsqu'on parle d'estimateur, on sous-entend que l'on cherche la meilleure estimation possible d'un paramètre à partir des données de l'échantillon. Il existe des bons et des mauvais estimateurs. Un critère souvent utilisé est celui de la moyenne de l'estimateur lorsqu'on réalise plusieurs échantillons. Si il prend en moyenne la valeur que l'on cherche à estimer, on parle alors d'un estimateur **non biaisé**. Un exemple d'un tel estimateur est la moyenne de l'échantillon $\overline{X}$. Dans ce cas, $\overline{X}$ désigne à la fois la moyenne statistique de notre échantillon, et notre estimation de la moyenne de la population.

Contrairement à la moyenne, l'écart-type observé s n'est pas un bon estimateur de l'écart-type

σ. Pour une raison que l'on ne développera pas, il faut remplacer n par n-1 dans la définition de l'écart-type pour obtenir un estimateur non biaisé.

$$\hat{\sigma} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})}$$

Remarquez qu'ici, on doit utiliser deux notations différentes pour désigner l'écart type de notre échantillon s et l'estimateur de l'écart-type $\hat{\sigma}$ (les estimateurs sont souvent écrits avec un chapeau). On appelle alors $n-1$ le nombre de **degrés de libertés** de l'estimation. Lorsqu'on ne connaît pas l'écart-type μ de la population, il faut donc utiliser plutôt que la distribution z ce qu'on appelle la t-distribution (parfois aussi nommée "t de Student") :

$$t = \frac{\overline{X} - \mu}{\hat{\sigma}/\sqrt{n}}$$

La distribution t est très proche de la loi normale, mais dépend du nombre de degré de libertés de la distribution (égal à $n-1$). Comme pour la loi normale, on peut lire les valeurs que l'on cherche à partir d'un table, ou bien en utilisant la fonction `qt()` dans R. La table en lien permet de voir comment varie t avec le nombre de degrés de liberté (en lisant les données d'une même colonne). Vous pouvez voir que, pour chaque degré de certitude, par exemple 90%, la valeur t_{90} diminue avec le nombre de degrés de libertés. En bas du tableau est représenté les valeurs de z associées. On peut remarquer que t et z sont très différentes pour les faibles effectifs, mais dès que le nombre de degrés de liberté augmente (par exemple df = 100), leurs valeurs sont très proches.

4.3.3 Un exemple

Prenons la base de données `hdv2003` du package `questionr`. On va prendre l'ensemble des individus de cette base comme population de référence, et échantillonner aléatoirement 20 individus de cette base de données. On souhaite estimer le temps moyen quotidien passé devant la télévision avec un intervalle de confiance à 95%.

Voici les valeurs de la variable `heures.tv` pour les 20 individus et leur moyenne :

```
d$heures.tv
```

```
##  [1] 3.0 2.0 7.0 0.0 2.0 0.0 2.0 4.0 4.0 1.0 0.3 3.0 3.0 5.0 3.0 1.0 0.0 2.0 2.1
## [20] 2.0
```

```
mean(d$heures.tv)
```

```
## [1] 2.32
```

Il faut aussi calculer l'estimation de l'erreur type :

```
sd(d$heures.tv)*sqrt(1/19)
```

```
## [1] 0.4105668
```

On ne connaît pas l'écart-type de la population, on doit donc utiliser la t-distribution. Comme $n = 20$, on a $df = 20 - 1 = 19$. On peut utiliser cette table et lire les valeurs associées à $t_{.975}$, en sélectionnant la ligne $df = 19$ et "two tails" (car on cherche un intervalle symétrique). On lit directement la valeur $t_{.975} = 2.093$. Autrement, on peut utiliser la fonction `qt()` :

```
qt(p = 0.975, df = 19)
```

```
## [1] 2.093024
```

On obtient (heureusement) la même valeur. Cela nous permet d'établir notre intervalle de confiance à 95% :

$$I_{95} = [2,32 - 0,41 * 2,093 \ ; \ 2,32 + 0,41 * 2,093] = [1,46 \ ; \ 3,18]$$

Finalement, notre intervalle de confiance de la moyenne va de 1,46 à 3,18 heures passées à regarder la télévision. Si l'on souhaite une estimation plus précise, on peut prendre un échantillon plus grand. Le code suivant réalise les mêmes calculs avec un échantillon de 200 individus :

```
# On sélectionne alétaoirement 1/10 de la population, c'est-à-dire 200 individus
d2 <- hdv2003 %>% sample_frac(0.1)
# On calcule directement la borne inférieures de notre intervalle de confiance
mean(d2$heures.tv, na.rm = T) - qt(p= 0.975, df = 199) * sd(d2$heures.tv, na.rm = T)*sqrt(1/199)
```

```
## [1] 2.054105
```

```
# Et la borne supérieure
mean(d2$heures.tv, na.rm = T) + qt(p= 0.975, df = 199) * sd(d2$heures.tv, na.rm = T)*sqrt(1/199)
```

```
## [1] 2.605192
```

On peut observer que l'intervalle de confiance à 95% est plus réduit, il s'agit donc d'une meilleure estimation. Finalement, on peut calculer la moyenne recherchée :

```
mean(hdv2003$heures.tv, na.rm = TRUE)
```

```
## [1] 2.246566
```

Remarquez qu'on doit utiliser l'argument `na.rm = TRUE` dans la fonction `mean()` en raison de l'existence de valeurs manquantes dans la variable `heures.tv`.

4.3.4 Interpréter un intervalle de confiance

Une fois que l'on a calculé l'intervalle de confiance, le dernier problème est de savoir exactement quelle signification lui attribuer. On pourrait être tenté de dire "la moyenne que l'on essaie d'estimer a 95% de chance d'être dans l'intervalle". Mais ça n'est pas vraiment une bonne formulation, car la moyenne à estimer n'est pas aléatoire : soit elle est dans l'intervalle soit elle ne l'est pas, donc la probabilité qu'elle soit dans entre les deux bornes de l'intervalle est 0 (elle n'y est pas) ou 1 (elle y est) mais pas 0,95. Ce qui est aléatoire ici, c'est l'intervalle que l'on a estimé et non la moyenne μ. Il faudrait donc plutôt dire que "l'intervalle a 95% de chance de capturer la moyenne entre ses bornes". Autrement dit, si l'on calcule 100 intervalles de confiance à 95% (quels qu'ils soient, pas nécessairement sur la même distribution), on aura en moyenne 5 intervalles qui ne captureront pas la moyenne μ.

Chapitre 5

Analyse bivariée et corrélation II

Jusqu'à maintenant, on a passé en revue :

— l'étude univariée d'une variable qualitative ou d'une variable quantitative (cours 2 et 4).
— l'étude de la corrélation entre deux variables qualitatives (cours 3)

Le cours de cette semaine est destiné à présenter les notions essentielles impliquées dans l'étude de la corrélation entre plusieurs variables quantitatives. La première partie du cours présente la représentation graphique associée à une telle étude, la notion de covariance et le modèle de la régression linéaire, tandis que la seconde partie aborde les questions d'inférence statistique avec la présentation d'un nouveau test d'hypothèse, le t-test.

5.1 Deux variables quantitatives

Étudier la corrélation entre deux variables quantitatives permet en général d'utiliser un plus grand nombre d'outils que l'étude des variables qualitatives, car il est possible de faire des calculs à partir des modalités des variables considérées. Certaines représentations graphiques sont aussi plus adaptées à l'étude des variables quantitatives.

5.1.1 Représenter deux variables quantitatives

Lorsqu'on souhaite observer les liens entre deux variables quantitatives, on représente en général un **nuage de points**. Il s'agit d'un graphique dans lequel chacune des variables est représentée selon un axe (l'une en abscisses, l'autre en ordonnées), ce qui permet de positionner chaque individu statistique à partir des valeurs associées à chacune des variables, qui seront alors ses coordonnées dans le plan.

```
ggplot(USArrests) + geom_point(aes(x = Murder, y = Rape))
```

Sur cet exemple, chaque point représente un État des États-Unis. Plus le point est situé à droite du graphe, plus le taux d'arrestation pour meurtre correspondant est important. De même, plus il est situé en hauteur sur le graphe, plus le taux d'arrestation pour viol est important (les données datent de 1974).

Pour avoir une idée d'où se situent les différents État sur le graphe, on peut indiquer sur le graphe à côté de chaque point l'État auquel il correspond.

```
## Warning: ggrepel: 1 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```

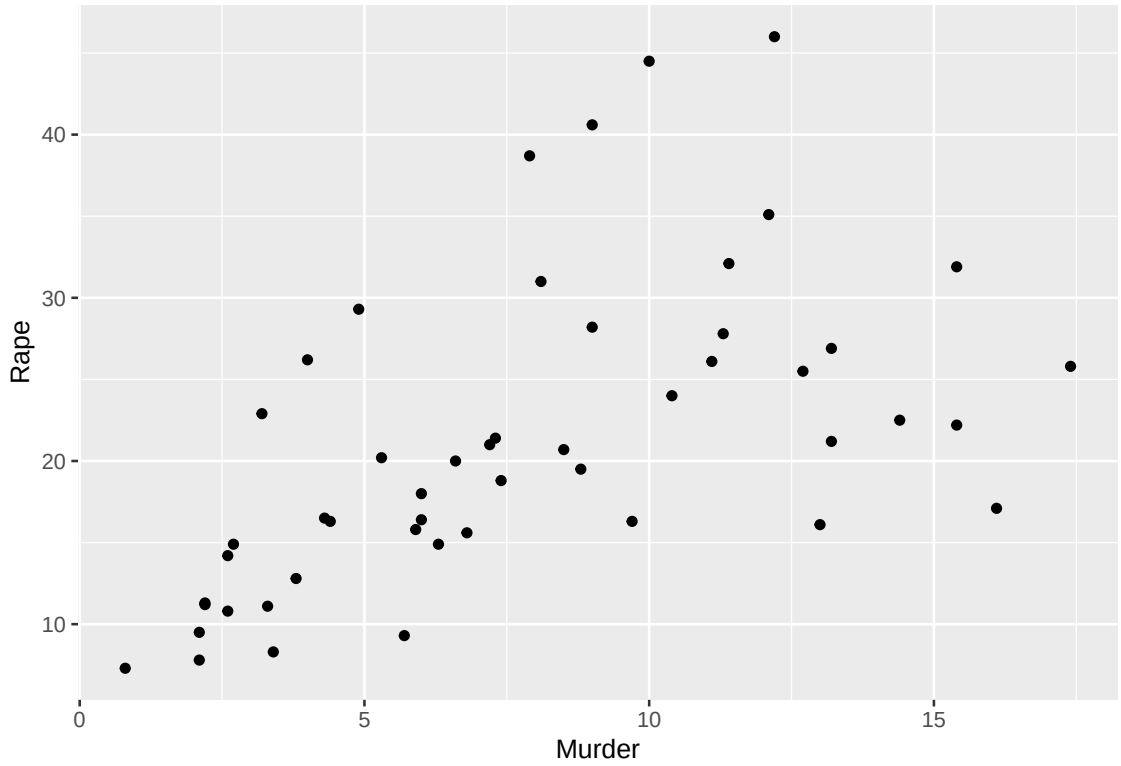
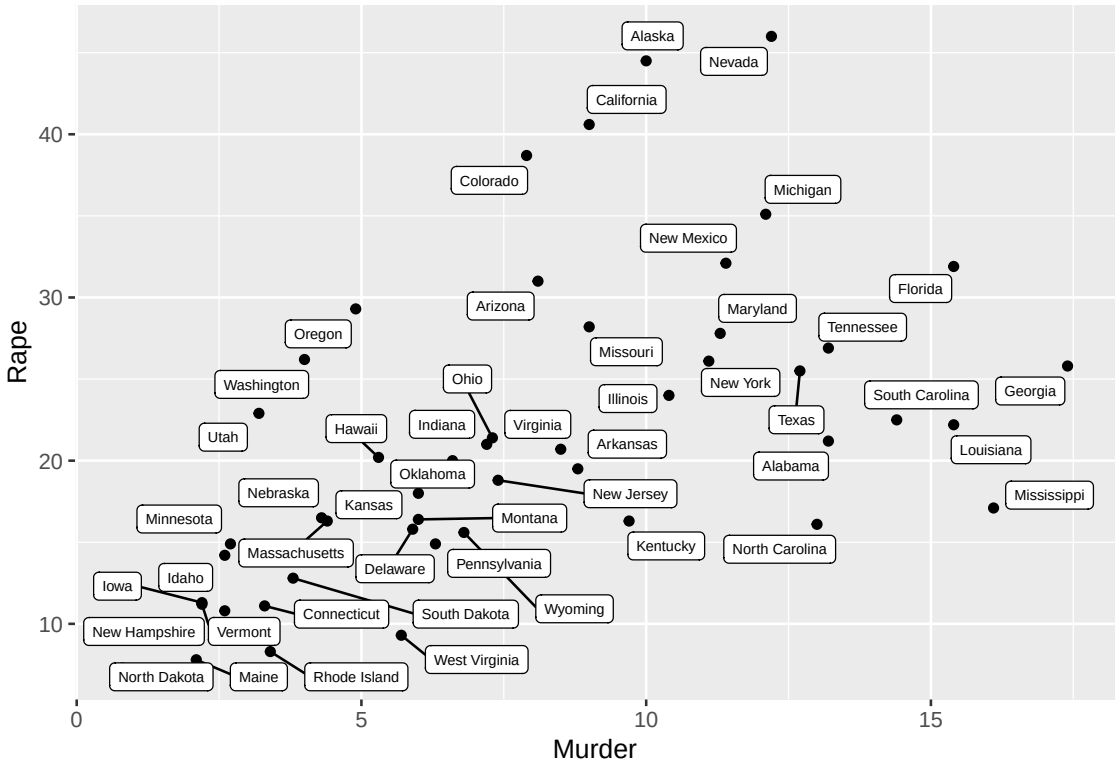


FIG. 5.1 – Un exemple à partir des données USArrests : chaque point représente un état des États-Unis, dont la position dépend de son taux d'arrestation pour meurtres (pour 100 000 habitants, selon x) et pour viol (pour 100 000 habitants, selon y) en 1974



La forme du nuage de point permet caractériser le lien entre les variables : on observe ici qu'il y a peu d'État dans en haut à gauche ou en bas à droite du graphique. C'est-à-dire que lorsque le taux d'arrestation pour meurtre est faible dans un État, le taux d'arrestation pour viol l'est aussi. Il semble donc exister un lien de corrélation entre ces deux variables : lorsqu'une augmente, on observe en général que l'autre augmente également.

5.1.2 La covariance

La covariance est une grandeur qui permet de mesurer la corrélation entre deux variables quantitatives. C'est une généralisation de la variance dans le cas de deux variables. Elle mesure la moyenne du produit des écarts à la moyenne de deux variables X et Y :

$$Cov(X,Y) = \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})$$

On peut remarquer immédiatement que $Cov(X,X) = Var(X)$, donc il s'agit bien d'une généralisation de la variance.

Pourquoi la covariance mesure-t-elle une corrélation entre X et Y? On peut le comprendre à partir d'un exemple. Disons que X mesure la taille d'un individu i, et Y son poids. Lorsque X est supérieur à sa moyenne, on aura par définition $(X_i - \overline{X}) > 0$. Mais les personnes les plus grandes seront aussi en moyenne plus lourdes que la moyenne, donc on aura le plus souvent $(Y_i - \overline{Y}) > 0$. À l'inverse, les personnes qui sont plus petites que la moyenne seront aussi en moyenne plus légères, donc lorsque $(X_i - \overline{X}) < 0$ on aura la plupart du temps $(Y_i - \overline{Y}) < 0$. Les écarts à la moyenne de X et Y auront donc le plus souvent le même signe, leur produit sera donc positif. Autrement dit, une corrélation positive entre les deux variables (c'est-à-dire le fait qu'une augmentation de X est généralement associée à une augmentation de Y) a pour conséquence une covariance positive. On pourrait montrer de la même manière qu'une corrélation négative induit une covariance inférieure à 0. Enfin, lorsque la covariance de X

et Y est nulle, on dit que les deux variables sont **indépendantes** : la valeur de l'une n'a en moyenne pas de lien avec la valeur de l'autre.

La covariance permet donc de retranscrire numériquement l'idée de corrélation. On peut par exemple la calculer pour les deux variables `Murder` et `Rape` représentées plus haut (figure 5.1).

```
cov(USArrests$Murder, USArrests$Rape)
```

```
## [1] 22.99141
```

On obtient bien un coefficient positif, comme le suggérait l'allure du nuage de points représenté. Le problème avec la covariance, c'est qu'au delà de son signe il est difficile de lui attribuer une signification. Cela est du fait que sa valeur dépend des unités de X et de Y. On préfèrerait un indice compris entre -1 et 1.

Pour l'obtenir, on calcule ce qu'on appelle **le coefficient de corrélation de Pearson**. Il règle le problème de l'unité de la covariance en la divisant par les écarts types de X et de Y :

$$r_{XY} = \frac{Cov(X,Y)}{\sigma_{\bar{X}}\sigma_{\bar{Y}}}$$

En normalisant la covariance, on conserve ses propriétés intéressantes, mais on obtient un indicateur plus facile à interpréter. Lorsque le coefficient de corrélation est égal à 1, les deux variables sont parfaitement corrélées. Si l'on représente leur nuage de points, on doit voir une droite de pente positive. À l'inverse, un coefficient de corrélation égal à -1 correspond à une droite de pente négative. Le coefficient mesure donc l'intensité de la corrélation : s'il est proche de 0, les variables sont faiblement corrélées, s'il est proche de 1 ou de -1, elle le sont fortement. Voici le résultat qu'on obtient pour notre exemple :

```
cov(USArrests$Murder, USArrests$Rape)/(sd(USArrests$Murder)*sd(USArrests$Rape))
```

```
## [1] 0.5635788
```

5.2 La regression linéaire

La méthode de la régression linéaire consiste à modéliser la relation entre deux variable quantitative par une droite, et cela même lorsque le coefficient de corrélation entre ces deux variables n'est pas égal à 1 ou -1. Graphiquement, il s'agit de trouver la droite qui résume le mieux le lien entre les deux variables.

Dans cette partie, on prendra comme exemple le jeu de données fictives `parenthood` proposé par Dan Navarro [Navarro, 2015]. Il contient plusieurs variables. L'une d'entre elles est le "niveau de mauvaise humeur de Dan" : on imagine qu'il le note chaque jour, en lui attribuant un score entre 0 et 100 (0 lorsqu'il est parfaitement de bonne humeur, 100 lorsqu'il est parfaitement de mauvaise humeur). La seconde variable que l'on va utiliser est son temps de sommeil.

Vous pouvez remarquer que les deux variables n'ont pas vraiment le même statut. La question que l'on va se poser est **l'effet du temps de sommeil sur la mauvaise humeur de Dan**. On dit que la variable "mauvaise humeur" est la **variable à expliquer**, tandis que la variable "temps de sommeil" est la **variable explicative**.

5.2.1 Principe de la régression

Comme il s'agit de modéliser la relation entre ces variables par une ligne droite, commençons par écrire l'équation d'une ligne droite :

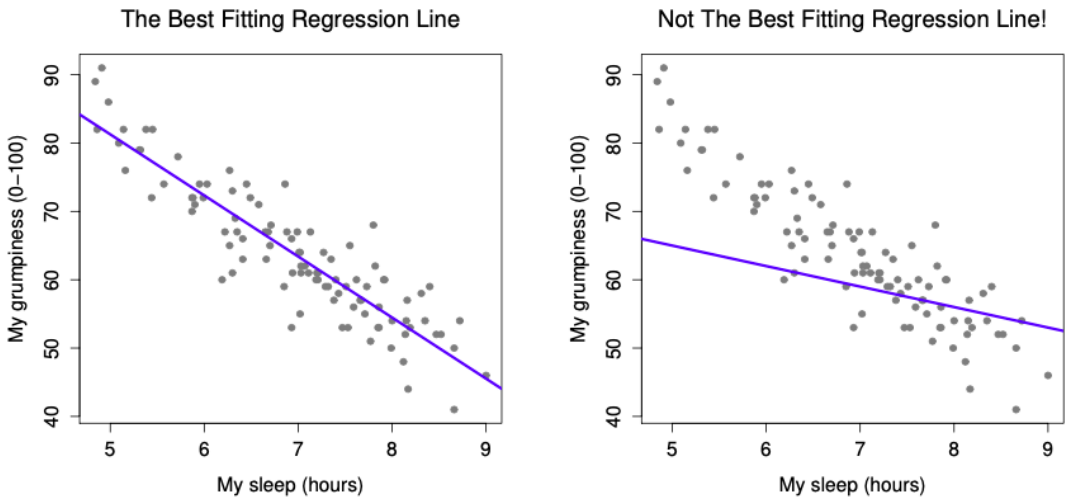


FIG. 5.2 – L'objectif est de trouver la droite qui résume le mieux le nuage de points. À gauche, la meilleure regression, à droite, une autre beaucoup moins bien

$$y = ax + b$$

y et x sont deux variables, et a et b deux coefficients : ce sont les nombres que l'on cherche à déterminer. On peut remarquer que l'asymétrie entre les deux variables se retrouve dans l'équation : y est la variable à expliquer, et x la variable explicative.

— a est la **pente** de la droite (c'est-à-dire l'augmentation de y lorsque x augmente d'une unité)
— b est **l'ordonnée à l'origine** (c'est-à-dire la valeur de y quand $x = 0$)

La formule qu'on utilise pour décrire une droite de régression est très similaire :

$$\hat{Y}_i = b_1 X_i + b_0$$

— on note X_i et Y_i plutôt que X et Y pour bien indiquer que cette relation est valable pour chacune des observations
— on note $\hat{Y}_i$ plutôt que Y_i, pour faire la différence entre les valeurs observées Y_i, et les valeurs estimées $\hat{Y}_i$, ce sont les valeurs qui sont prédites à partir du modèle.
— on note les coefficients b_1 et b_0 plutôt que a et b (ce qui permettra plus tard d'ajouter b_2, b_3, etc.).

Puisque notre estimation n'est pas parfaite, c'est-à-dire qu'elle diffère des données observées, il existe une différence entre les données observées et les données estimées qu'on appelle **résidus** :

$$\epsilon_i = Y_i - \hat{Y}_i$$

L'expression des résidus permet de réécrire l'équation du modèle de cette manière :

$$Y_i = b_1 X_i + b_0 + \epsilon_i$$

La question est alors de savoir quel critère retenir lorsqu'on cherche "la meilleure droite de régression". La figure suivante représente la grandeur des résidus pour deux droites de régression. Sur la gauche, on a l'intuition que la droite est une meilleure estimation du nuage de points : les résidus sont aussi en moyenne plus faibles.

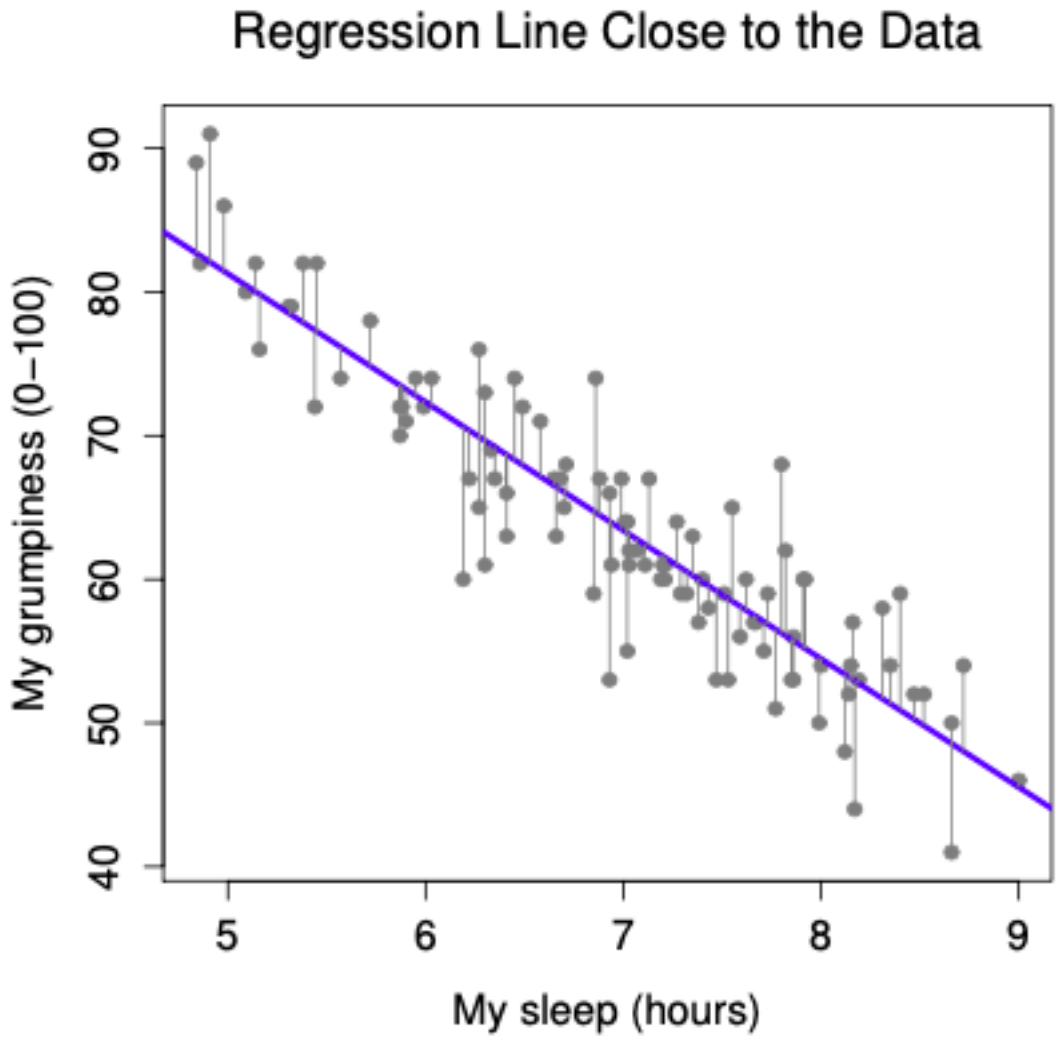


FIG. 5.3 – Une représentation graphique des résidus associés aux deux droites de régression. On remarque que les résidus sont moins grands en moyenne dans le premier cas que dans le second

5.2.2 Estimer une droite de regression

C'est ce critère que l'on va retenir pour déterminer quelle est la meilleure droite de régression. Les coefficients estimés, b_0 et b_1 sont ainsi ceux qui minimisent la somme des carrés des résidus, qu'on peut écrire :

$$\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2$$

ou encore :

$$\sum_{i=1}^{N}\epsilon_i^2$$

On appelle cette méthode d'estimation des coefficients la **méthode des moindres carrés ordinaires**. Trouver les coefficients qui minimisent la somme des carrés des résidus est un problème purement mathématique. L'idée est que vous n'avez pas besoin de savoir comment faire, mais que R peut le faire pour vous à l'aide de la fonction `lm()`.

5.2.3 Utiliser la fonction `lm()` dans R

La fonction `lm()` (pour *linear model*) est assez compliquée : si vous tapez `?lm` dans la console, vous verrez qu'il existe un grand nombre de manières d'utiliser cette fonction. Mais à ce stade, parmi les arguments qu'on peut passer à la fonction, il y en a seulement deux qui nous intéressent :

— `formula`. Une `formula` permet de préciser le modèle proposé pour la régression. Pour la régression linéaire simple, la formule est simplement `y ~ x`, où `y` est la variable à estimer, et `x` la variable explicative.
— `data`. La base de données où sont présentes les variables.

Le produit de la fonction `lm` est un amalgame d'information qu'on ne peut lire qu'à l'aide d'autres fonctions. En général, on stocke le résultat de la fonction `lm` dans un objet dédié :

```
regression.1 <- lm(formula = dan.grump ~ dan.sleep, data = parenthood)
print(regression.1)
```

```
##
## Call:
## lm(formula = dan.grump ~ dan.sleep, data = parenthood)
##
## Coefficients:
## (Intercept)    dan.sleep
##     125.956       -8.937
```

R nous donne les deux résultats principaux de cette régression linéiare : les coefficients b_0 et b_1. En d'autres termes, la meilleure droite de régression que l'on a représentée plus haut est la droite d'équation :

$$\hat{Y}_i = -8.94X_i + 125.96$$

Qu'est-ce-que cela signifie ? Selon notre modèle linéaire, le temps de sommeil de Dan Navarro est corrélé négativement à sa mauvaise humeur. C'est ce qu'indique le coefficient $b_1 = -8.94$: lorsque Dan dort une heure de plus, sa mauvaise humeur diminue en moyenne de 8.94 points. L'autre coefficient $b_0 = 125.96$ indique l'ordonnée à l'origine, c'est-à-dire le score de mauvaise humeur attendu dans le cas où Dan ne dort pas du tout. Il est ici incohérent avec la définition de ce score (qui est entre 0 et 100), ce qui permet de constater qu'une estimation linéaire ne produit pas nécessairement des valeurs qui ont du sens.

Pourquoi appelle-t-on cette équation un **modèle** linéaire ? Pour le comprendre, il suffit de bien se représenter ce que signifient les deux termes de l'équation. À gauche, il s'agit de notre

estimation de la variable Y (d'où le chapeau sur le Y). À droite, X n'a pas de chapeau, c'est donc la valeur observée du temps de sommeil de Dan. L'équation a un caractère **prédictif** : elle permet de calculer le score de mauvaise humeur de Dan attendu pour différentes valeurs de son temps de sommeil.

On peut constater qu'avec ces seuls résultats, nous ne sommes pas en mesure d'évaluer la qualité de l'estimation produite par le modèle de régression linéaire. On sait qu'il s'agit de la meilleure droite, mais rien n'indique que le modèle linéaire (le fait d'estimer le nuage de point par une droite) est vraiment approprié. On verra dans le prochain cours comment évaleur la qualité de l'estimation.

5.3 Le *t*-test

On a vu dans le cours précédent comment la distribution t permettait de produire des intervalles de confiance lorsqu'on cherche à inférer à partir d'un échantillon la moyenne d'une variable quantitative. Cette distribution est largement utilisée pour produire des tests d'hypothèse. Je vous en présente ici différentes versions. La première est très proche de l'estimation d'un intervalle de confiance qui a été abordée au cours précédent, la seconde permet d'utiliser la même méthode pour étudier des corrélations entre variables qualitatives et variables quantitatives.

5.3.1 *t*-test pour un seul échantillon

C'est un test statistique qui a été introduit par William Sealy Gosset, qui travaillait comme chimiste à la brasserie Guinness et a publié son travail sous le pseudonyme 'A Student' (Student, 1908). On parle encore aujourd'hui du t de Student.Pour rappel, on définit la distribution t de cette manière :

$$t = \frac{\bar{X} - \mu}{\hat{\sigma}/\sqrt{N}}$$

où $\bar{X}$ est la moyenne de l'échantillon, μ est la moyenne de la population, $\hat{\sigma}$ l'estimation de l'écart-type.

On peut comprendre la question à laquelle son test cherchait à répondre : il prélevait un échantillon de Guinness et en mesurait différentes caractéristiques (par exemple le pourcentage d'alcool). Sa question était alors de savoir si cette caractéristique mesurée était suffisamment proche de celle attendue (ie. le degré d'alcool indiqué sur la bouteille de Guinness). On utilise ainsi le t test pour un échantillon dans le cas où on cherche à comparer la moyenne de cet échantillon à une valeur définie par ailleurs.

Le principe du test revient à faire l'hypothèse que la moyenne de la population est égal à la moyenne attendue (l'hypothèse nulle), puis d'estimer sous cette hypothèse la probabilité d'obtenir la valeur mesurée sur notre échantillon. Comme pour le test du χ^2, R va nous indiquer une *p-value*, qui indique la probabilité de se tromper si l'on accepte l'hypothèse nulle, c'est-à-dire la probabilité d'avoir mesuré **par hasard** une valeur proche de la valeur attendue.

Reprenons l'exemple de la semaine dernière avec notre échantillon de 20 individus de la base `hdv2003`. On se demande si 1.9 heures par jour est une valeur plausible pour la moyenne du temps passé à regarder la télévision de notre population.

```
t.test(d$heures.tv, mu = 1.9)
```

```
##
##  One Sample t-test
##
```

```
## data:  d$heures.tv
## t = 1.0496, df = 19, p-value = 0.3071
## alternative hypothesis: true mean is not equal to 1.9
## 95 percent confidence interval:
##  1.482432 3.157568
## sample estimates:
## mean of x
##      2.32
```

On observe donc qu'avec cet échantillon de 20 individus, il est difficile de rejeter l'hypothèse nulle. La *p-value* de 0.3 indique bien qu'il y a 30% de chances de se tromper si l'on rejette l'hypothèse nulle. On peut également remarquer que R nous indique l'intervalle de confiance à 95% de la moyenne, intervalle qui dans 95% des cas capture la moyenne de la population entre ses bornes.

Répétons le test avec un échantillon de 200 individus sélectionnés aléatoirement parmi les 2000 individus de `hdv2003` :

```
d <- sample_frac(hdv2003, size = 0.1)
t.test(d$heures.tv, mu = 1.9)
```

```
##
##  One Sample t-test
##
## data:  d$heures.tv
## t = 1.684, df = 198, p-value = 0.09377
## alternative hypothesis: true mean is not equal to 1.9
## 95 percent confidence interval:
##  1.865873 2.333122
## sample estimates:
## mean of x
##  2.099497
```

On observe cette fois ci que 1.9 n'est pas compris dans l'intervalle de confiance, et que la *p-value* est beaucoup plus faible. Ce qui nous permet de rejeter l'hypothèse nulle : la moyenne mesurée sur l'échantillon permet avec une certitude relative de conclure que la moyenne du temps passé à regarder la télévision dans la population n'est pas de 1.9 heures par jour.

5.3.2 *t*-tests pour des échantillons indépendants

Même si le *t*-test pour un échantillon peut être utile, on utilise plus souvent le *t*-test pour deux échantillons indépendants. L'idée est que l'on dispose de deux échantillons différents, et que l'on cherche à savoir si leurs deux moyennes sont égales (ou plus précisément, on cherche à savoir quelle est la probabilité que les moyennes des deux populations sont égales). Il est important de noter que les deux échantillons ne sont pas nécessairement deux échantillons aléatoires de la même population : ils peuvent avoir des caractéristiques différentes. Par exemple, on peut prendre un échantillon d'hommes et un échantillon de femmes et comparer les temps moyen de travail domestique correspondant. On trouvera normalement un temps plus élevé pour le groupe des femmes. Mais si l'on veut s'assurer qu'ils sont bien différents (c'est-à-dire que la différence ne s'explique pas par hasard lié à l'échantillonnage), on effectuera un *t*-test.

On définit alors la distribution t de cette manière :

$$t = \frac{\bar{X}_1 - \bar{X}_2}{SE(\bar{X}_1 - \bar{X}_2)}$$

où

$$SE(\bar{X}_1 - \bar{X}_2) = \hat{\sigma}\sqrt{\frac{1}{N_1} + \frac{1}{N_2}}$$

et l'on admettra que, si l'hypothèse nulle est vérifiée (les moyennes des deux échantillons sont égales), cette distribution suit une loi t à $(N_1 + N_2 - 2)$ degrés de libertés, où N_1 est l'effectif du premier échantillon et N_2 celui du second. Pour plus de détails, je vous renvoie à la lecture de Navarro [Navarro, 2015, chapitre 13].

Pour exemple, on peut prendre encore la variable `heures.tv` des données `hdv2003`. On va se demander si les actifs en emploi regardent autant en moyenne la télévision que les chômeurs et les inactifs (hypothèse nulle). J'ai créé une nouvelle variable `act` dans la table hdv2003 qui permet de distinguer les actifs en emploi (`act = 1`) des chômeurs et inactifs (`act = 0`). Pour faire le test dans R, on peut encore utiliser la fonction `t.test()` de cette manière :

```
t.test(heures.tv ~ act, data = hdv2003, var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  heures.tv by act
## t = 11.601, df = 1993, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  0.7429832 1.0453003
## sample estimates:
## mean in group 0 mean in group 1
##        2.715823        1.821681
```

Comment lire les résultats de ce test ? R nous indique à la toute fin les moyennes calculées pour les deux échantillons : 2.7 pour le premier (les chômeurs et inactifs), et 1.8 pour le second (les actifs en emploi). Le test permet de quantifier la probabilité que la différence entre ces deux moyennes est due au hasard, à partir de l'estimation de la valeurs de la distribution $t = 11.378$. La *p-value* associée montre que cette probabilité est très faible. R donne également l'intervalle de confiance à 95% de la différence entre ces deux moyennes : on remarque que 0 ne fait pas partie de cet intervalle. Pour commenter ce test, on pourrait écrire quelque chose comme ça :

> Le temps quotidien passé à regarder la télévision est en moyenne de 2.7 heures pour les chômeurs et inactifs et de 1.8 heures pour les actifs en emploi. Un test de Student pour deux échantilons indépendant montre que cette différence est significative ($t = 11.378$, $p < 0.05$), ce qui suggère l'existence d'un lien de corrélation entre l'activité et le temps passé à regarder la télévision.

Cette forme de t-test permet donc d'estimer l'existence d'une corrélation entre une variable qualitative (ici l'activité) et une variable quantitative (ici le temps passé à regarder la télévision).

Un élément important que je n'ai pas évoqué est l'ensemble des hypothèses que doivent vérifier les deux variables pour que cette méthode soit correcte. Il y a trois hypohtèses :

1. La *normalité*. On fait l'hypohtèse que la variable quantitative est distribuée selon une loi normale.
2. L'*indépendance*. Les observations doivent être indépendantes les unes des autres.
3. L'*homogénéité de la variance* (aussi nommé *homoscédasticité*). Cette troisième hypothèse implique que l'écart-type de la population est le même pour les deux groupes (ici les actifs occupés et les autres).

Cette troisième hypothèse est en général peu réaliste car si l'on compare des populations différentes, il n'y a pas de raison pour que l'écart-type soit le même. Pour s'affranchir de cette hypothèse, on peut utiliser encore un autre test, le t-test de Welch. Ce test suit exactement le même principe que le test de Student : on définit t de la même manière, les seules différences sont dans l'estimation de l'erreur-type :

$$SE(\overline{X_1} - \overline{X_2}) = \sqrt{\frac{\hat{\sigma_1}^2}{N_1} + \frac{\hat{\sigma_2}^2}{N_2}}$$

qui dépend ici, contrairement au test de Student, des deux écart-types différents. Le nombre de degrés de liberté change également :

$$df = \frac{(\hat{\sigma_1}^2/N_1 + \hat{\sigma_2}^2/N_2)^2}{(\hat{\sigma_1}^2/N_1)^2/(N_1 - 1) + (\hat{\sigma_2}^2/N_2)^2/(N_2 - 1)}$$

On ne va pas commenter cette formule, je vous l'indique simplement pour que vous ne soyez pas surpris que le nombre de degré de liberté ne sera pas égal à un entier, ça peut ici être n'importe quel nombre positif. Il est en général un peu plus faible que le nombre de degrés de liberté utilisé pour le test de Student.

Le test se fait dans R de la même manière que le test de Student, avec l'option `var.equal = TRUE` en moins. Il s'agit en fait de la sorite par défaut de la fonction `t.test()`. Ce qui s'affiche est tout à fait similaire à ce que l'on a déjà vu.

```
t.test(heures.tv ~ act, data = hdv2003)
```

```
##
##  Welch Two Sample t-test
##
## data:  heures.tv by act
## t = 11.378, df = 1616.4, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  0.7399966 1.0482870
## sample estimates:
## mean in group 0 mean in group 1
##        2.715823        1.821681
```

Chapitre 6

Regréssion multilinéaire et regréssion logistique

La semaine dernière, je vous ai présenté la méthode de la régression linéaire : le principe est de **modéliser** la relation entre deux variables quantitatives par une relation linéaire, c'est-à-dire dont la représentation graphique dans un plan serait une droite. Le modèle donne en résultat deux coefficients qui caractérisent la droite obtenue : la pente de la droite et son ordonnée à l'origine. Ce modèle est utile lorsque l'on étudie les relations entre **deux variables quantitatives**. Dans le cours de cette semaine, nous nous intéressons dans un premier temps à la généralisation de ce modèle de la régression linéaire lorsqu'on cherche à inclure plusieurs variables explicatives. Dans une deuxième partie, nous aborderons le modèle de la régression logistique, qui lui permet de prendre pour variable à expliquer une variable qualitative dichotomique (*i.e.* qui ne prend que deux valeurs).

6.1 Régression multilinéaire

6.1.1 Principe

Le modèle de la régression multilinéaire est une simple généralisation du modèle de la régression linéaire. On prend encore une variable que l'on cherche à expliquer, cette fois non plus à l'aide d'une variable mais de plusieurs variables quantitatives différentes. Dans notre exemple de la semaine dernière, on étudiait le lien entre le temps sommeil de Dan Navarro (variable explicative) et son niveau de mauvaise humeur (variable expliquée). Si l'on cherche à prendre en compte également le temps de sommeil de son fils dans le modèle (autre variable explicative), on réalisera une régression multilinéaire.

Peut-être vous demandez-vous pourquoi on ne réalise pas plutôt deux régressions linéaires, l'une pour chaque variable explicative. La réponse est que l'on n'estime pas la même chose dans les deux cas. Lorsqu'on réalise une régression multilinéaire, les coefficients que l'on obtient nous indiquent la relation entre la variable explicative correspondante (le temps de sommeil de Dan par exemple) et la variable expliquée, *toutes choses égales par ailleurs* ou plus précisément toutes les autres variables explicatives incluses dans le modèle étant supposées constantes. Dans notre exemple, le coefficient considéré donnera le lien entre le temps de sommeil de Dan et sa mauvaise humeur *en neutralisant l'effet du temps de sommeil de son fils* sur sa mauvaise humeur. La régression multilinéaire permet ainsi d'isoler les effets des différentes variables explicatives.

D'un point de vue formel, le modèle multilinéaire s'écrit presque de la même manière que le modèle linéaire :

$$Y_i = (\sum_{k=1}^{K} b_k X_{ik}) + b_0 + \epsilon_i$$

Vous pouvez remarquer que pour K = 1, on retrouve la régression linéaire de la semaine dernière. Si l'on inclue deux variables explicatives (X_1 et X_2), les résultats principaux du modèle seront trois coefficients :

— b_0 qui sera toujours l'ordonnée à l'origine (la valeur attendue de Y lorsque $X_1 = X_2 = 0$)
— b_1 est la pente de la droite lorsqu'on représente Y en fonction de X_1
— b_2 est la pente de la droite lorsqu'on représente Y en fonction de X_2

6.1.2 Un exemple

On va reprendre le même exemple que la semaine dernière.

dan.sleep	baby.sleep	dan.grump	day
7.59	10.18	56	1
7.91	11.66	60	2
5.14	7.92	82	3
7.71	9.61	55	4
6.68	9.75	67	5
5.99	5.04	72	6

Pour rappel, voici ce qu'on obtenait en réalisant une régression linéaire :

```
lm(dan.grump ~ dan.sleep, data = parenthood)
```

```
##
## Call:
## lm(formula = dan.grump ~ dan.sleep, data = parenthood)
##
## Coefficients:
## (Intercept)    dan.sleep
##     125.956       -8.937
```

Pour étudier le niveau de mauvaise humeur de Dan en fonction à la fois de son temps de sommeil et du temps de sommeil de son fils (donc faire une régression multilinéaire), on rajoute la variable `baby.sleep` dans la formule à l'intérieur de la fonction `lm` :

```
lm(dan.grump ~ dan.sleep + baby.sleep, data = parenthood)
```

```
##
## Call:
## lm(formula = dan.grump ~ dan.sleep + baby.sleep, data = parenthood)
##
## Coefficients:
## (Intercept)    dan.sleep   baby.sleep
##   125.96557     -8.95025      0.01052
```

On obtient maintenant trois coefficients différents. On peut remarquer que le coefficient associé à la variable `dan.sleep` a légèrement varié entre les deux régressions. Même si la variation est faible, cela illustre l'effet de l'ajout d'une nouvelle variable dans le modèle.

Les coefficients s'interprètent presque de la même manière que pour la régression linéaire : 125.9 est l'estimation du niveau de mauvaise humeur de Dan lorsque son temps de sommeil et celui de son fils sont égal à 0. Le coefficient associé à `dan.sleep` indique dans le modèle

la variation du score de mauvaise humeur de Dan lorsqu'il dort une heure de plus, le temps de sommeil de son fils étant constant (il baisse d'environ 9 points). De même, le coefficient `baby.sleep` indique la variation de la mauvaise humeur de Dan lorsque son fils dort une heure supplémentaire, son temps de sommeil personnel étant constant. Ce dernier coefficient est faible : la mauvaise humeur de Dan augmente dans notre modèle de 0.01 point.

Il semble donc qu'à temps de sommeil constant pour Dan, le temps de sommeil de son fils a très peu d'effet sur son humeur. Cela invite à faire deux remarques.

6.1.2.1 Remarque 1. Interprétation des résultats

Cela ne signifie pas que le temps de sommeil de son fils et son humeur ne sont pas corrélées : si l'on fait une régression linéaire de `dan.grump` par `baby.sleep` sans inclure la variable `dan.sleep`, on obtient en effet

```
lm(dan.grump ~ baby.sleep, data = parenthood)
```

```
##
## Call:
## lm(formula = dan.grump ~ baby.sleep, data = parenthood)
##
## Coefficients:
## (Intercept)   baby.sleep
##      85.782       -2.742
```

On obtient ici un coefficient de corrélation entre `dan.grump` et `baby.sleep` très différent de celui obtenu dans la régression multilinéaire, car ici une heure de sommeil de son fils est associée à une réduction de 2.7 points de sa mauvaise humeur. Mais cela est estimé *toutes choses inégales par ailleurs*, en particulier le temps de sommeil de Dan lui-même. On voit ici l'intérêt de la régression multilinéaire qui, en neutralisant l'effet du temps de sommeil de Dan, permet de montrer que le temps de sommeil de son fils n'a pas d'effet propre sur sa mauvaise humeur. Autrement dit, ce n'est pas parce que son fils dort moins que Dan est de mauvaise humeur, c'est parce que son fils l'a réveillé, et donc qu'il a moins dormi lui même. On peut d'ailleurs s'en assurer en faisant une régression linéaire entre le temps de sommeil de son fils et celui de Dan (essayez d'interpréter le coefficient obtenu)

```
lm(dan.sleep ~ baby.sleep, data = parenthood)
```

```
##
## Call:
## lm(formula = dan.sleep ~ baby.sleep, data = parenthood)
##
## Coefficients:
## (Intercept)   baby.sleep
##      4.4897       0.3075
```

6.1.2.2 Remarque 2 : significativité des coefficients

Une autre question que soulève le faible coefficient obtenu pour la variable `baby.sleep` dans notre régression multilinéaire est plus générale : quand peut-on considérer qu'il y a une corrélation significative entre les deux variables (ici `dan.grump` et `baby.sleep`) ? C'est une question de statistique inférentielle. Pour répondre à cette question, on peut réaliser un t-test sur les coefficients obtenus lors de la régression. Pour afficher les résultats de ces tests dans R, on passe le résultat de la fonction `lm` à la fontion `summary()` :

```
lm(dan.grump ~ dan.sleep + baby.sleep, data = parenthood) %>% summary()
```

```
##
```

```
## Call:
## lm(formula = dan.grump ~ dan.sleep + baby.sleep, data = parenthood)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.0345  -2.2198  -0.4016   2.6775  11.7496
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 125.96557    3.04095  41.423   <2e-16 ***
## dan.sleep    -8.95025    0.55346 -16.172   <2e-16 ***
## baby.sleep    0.01052    0.27106   0.039    0.969
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.354 on 97 degrees of freedom
## Multiple R-squared:  0.8161, Adjusted R-squared:  0.8123
## F-statistic: 215.2 on 2 and 97 DF,  p-value: < 2.2e-16
```

La *p-value* calculée dans la dernière colonne du tableau indique le degré de significativité des coefficients obtenus. Les étoiles représentées à droit de la *p-value* permettent de repérer rapidement les variables dont la valeur du coefficient est significatif (la légende est indiquée en dessous du tableau). On ne veut pas commenter la valeurs des coefficients qui n'ont pas d'étoile, car leur valeurs a de grandes probabilités d'être le résultat du hasard de l'échantillonnage.

6.1.3 La qualité des modèles

La régression nous permet de caractériser les relations entre les variables qui entrent dans le modèle, mais la valeur des coefficients ne dit rien de la qualité du modèle (est-ce-que le nuage de points ressemble bien à une droite ?). Si les résultats du modèle indiquent que l'humeur de Dan s'améliore avec son temps de sommeil, on n'a pas regardé si ce modèle était bon ou non.

Pour étudier ça, on utilise encore une fois la somme des résidus au carré :

$$SS_{res} = \sum_i (Y_i - \hat{Y}_i)^2 = \sum_i \epsilon_i$$

On espère que cette somme est relativement petite. Plus précisément, on voudrait la comparer à la variance totale de la variable expliquée :

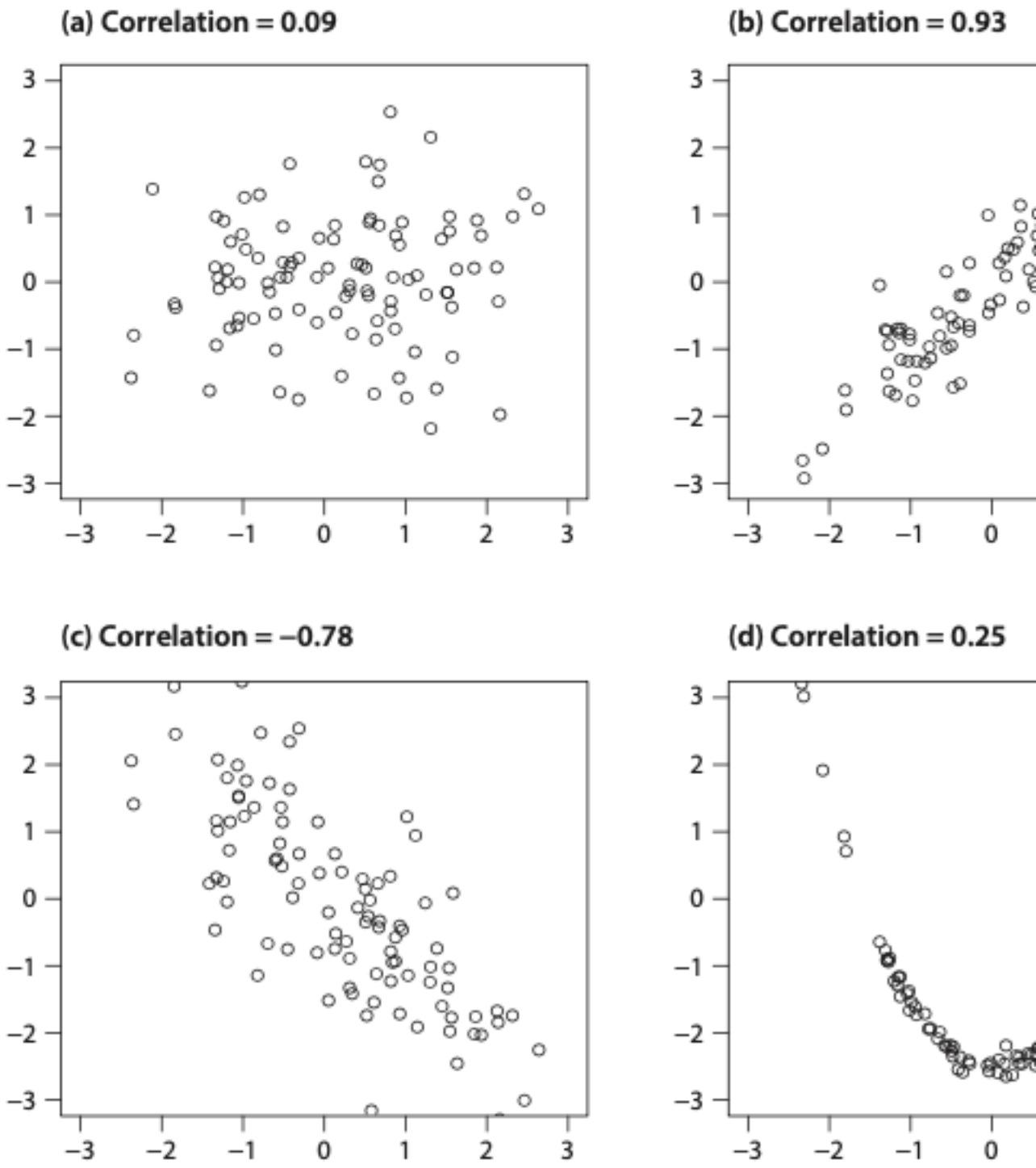
$$SS_{tot} = \sum_i (I_y - \bar{Y})^2$$

Ce qu'on voudrait, c'est calculer un nombre qui serait égal à 1 lorsque le modèle colle parfaitement aux données, donc lorsque la somme des résidus est strictement nulle. Inversement, si le modèle est totalement inutile, on voudrait avoir $R^2 = 0$. Ce qu'on entend par inutile, c'est que la somme des carrés des résidus serait égal à la variance totale, $SS_{res} = SS_{tot}$.

Au final, on peut proposer la valeur $R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$. En fait, on peut montrer que cette valeur est la même que le coefficient de corrélation de Pearsons (présenté au cours précédent). Dans le résultat de la fonction `summary()` affiché plus haut, cette grandeur est indiquée en bas, là où il est écrit `Multiple R squared`.

La figure suivante présente les valeurs du R^2 pour plusieurs régression linéaire, à côté du nuage de points associée. On peut remarquer que si une valeur de R^2 proche de 1 signifie

à la fois que les données sont corrélées et que le modèle linéaire est adapté (figures b et c), une valeur faible de R^2 est plus difficile à interpréter. Cela peut être du au fait que les deux variables sont peu corrélées (figure a), mais pas forcément. La figure d donne en effet un exemple où les variables sont très corrélées (les points se distribuent suivante une courbe, connaître la valeur de x permet quasiment de déterminer la valeur de y) mais le coefficient de corrélation est faible. Cela s'explique ici par le fait que le modèle linéaire n'est pas adapté : l'allure de la courbe indique qu'il faudrait utiliser un modèle non linéaire, par exemple $y = a * x^2 + b$. Si vous rencontrez cette situation, il vaut mieux essayer de changer de variable que de changer de modèle, c'est-à-dire prendre $Y = y$ et $X = x^2$, puis faire une régression linéaire entre X et Y.



6.1.4 Un exemple en sociologie

Un exemple parmi beaucoup d'autres, cette régression effectuée par Nicolas Herpin dans un article qui étude l'aspect social de la taille [Herpin, 2003]. Sa question est de savoir si la taille a un effet propre, toutes choses égales par ailleurs, sur la probabilité d'obtenir un emploi par exemple, ou d'être en couple. Avant d'étudier ces questions, il réalise une régression

sur les facteurs sociaux de la taille. C'est-à-dire que la taille des individus est la variable expliquée (variable quantitative) et que plusieurs variables sont proposées pour expliquer les différences de tailles observées entre les individus. Herpin réalise en fait 2 régressions séparées, l'une pour les hommes et l'autre pour les femmes, tou·tes ayant plus de 30 ans pour pouvoir considérer leur taille comme définitive.

Vous pouvez remarquer que les variables explicatives ne sont pas nécessairement des variables quantitatives, ici la plupart sont qualitatives (à l'exception de la corpulence et l'âge). Les modalités de ces variables sont traitées dans la régression comme autant de variable dichotomiques (égales à 0 ou 1), et chaque modalité a donc un coefficient différent, qu'il faut lire par rapport à la catégorie qui est définie comme référence. Par exemple, toute choses égales par ailleurs, les hommes agriculteurs font 2,26cm de plus que les hommes ouvriers.

Voici comment Nicolas Herpin commente cette régression, il lit les résultats et propose des interprétations basées sur le résultat observé que la taille dépend de manière inverse de l'intensité du travail physique et de sa précocité :

> La taille définitive est sensible à plusieurs facteurs dont les incidences respectives sont établies toutes choses égales (cf. tableau B). Hommes ou femmes sont d'autant plus grand(e)s qu'ils ou elles sont plus jeunes. Le Nord (départements du Nord et du Pas-de-Calais) et l'Est ont des habitants plus grands et ceux de l'Ouest (Bretagne, Poitou-Charentes et Val-de-Loire) plus petits que ceux résidant dans les autres régions (une régression de la taille sur les mêmes facteurs explicatifs mais effectuée sur la population de ceux qui résident dans leur région de naissance donne les mêmes résultats). Les ouvriers sont plus petits et se distinguent des hommes des autres catégories professionnelles (y compris des agriculteurs). L'origine sociale oppose les hommes issus des classes moyennes (fils d'employés et de professions intermédiaires), plus grands, aux autres origines sociales. Les fils de cadres ne sont pas différenciés par la taille des fils d'agriculteurs, des fils d'artisans ou de commerçants et des fils d'ouvriers. Enfin, le travail précoce – dont l'âge auquel la personne quitte l'école est l'indice – a des effets néfastes sur la taille définitive pour les hommes, toutes choses égales.
>
> La comparaison avec les femmes fait ressortir peu de différence. Cependant, la précocité au travail ne semble pas avoir des effets aussi forts ni aussi réguliers que chez les hommes. Les filles aident au travail domestique davantage dans les milieux populaires. Mais leur travail rémunéré ne nécessite pas autant de force que celui des hommes lorsqu'il commence avant la fin de l'adolescence. On peut alors comprendre pourquoi l'amélioration des conditions du travail manuel et le relatif déclin des emplois les moins qualifiés dans l'agriculture, l'industrie et la construction ont davantage profité au grandissement des hommes qu'à celui des femmes. Mais (cf. tableau B), il faut attribuer le plus rapide grandissement générationnel des hommes à leur entrée de plus en plus tardive sur le marché de l'emploi [Herpin, 2003, p. 74-75].

6.2 Régression logistique

Il n'est toutefois pas toujours possible de se ramener à un modèle linéaire. Dans certaines situations, on doit utiliser des modèles non linéaires, c'est-à-dire que la relation entre le deux variable ne sera pas modélisée graphiquement par une droite.

6.2.1 Principe de la régression logistique

Disons que l'on cherche à modéliser une **variable qualitative qui n'a que deux modalités**, par exemple la possession ou non d'une voiture. Si l'on cherche coûte que coûte à utiliser la

	Les hommes		
	Paramètre	Écart-type	Paramè
Constante	181,9***	0,74	168,0**
Corpulence	0,15	0,17	- 0,57
Âge de la personne	- 0,16***	0,01	- 0,09
Région habitée			
Région parisienne	- 1,10*	0,58	- 2,44
Bassin parisien	- 1,23***	0,48	- 1,85
Méditerranée	- 1,58***	0,54	- 1,78
Est	- 0,52	0,60	- 0,89
Ouest	- 2,21***	0,51	- 2,89
Sud-Ouest	- 1,74***	0,55	- 2,60
Centre-Est	- 1,65***	0,54	- 1,93
Nord	*Réf.*		*Réf.*
Profession de la personne			
Agriculteur	2,26***	0,54	0,79
Artisan, commerçant, entrepreneur	2,16***	0,45	1,60
Cadre, profession libérale, prof. intell. supérieure	2,67***	0,40	2,35
Profession intermédiaire	2,01***	0,33	1,38
Employé	1,72***	0,41	1,08
Ouvrier	*Réf.*		*Réf.*
Profession du père			
Agriculteur	- 0,17	0,37	0,55
Artisan, commerçant, entrepreneur	0,49	0,40	0,14
Cadre, profession libérale, prof. intell. supérieure	0,69	0,49	0,67
Profession intermédiaire	0,94**	0,43	0,95
Employé	1,10***	0,40	0,51
Ouvrier	*Réf.*		*Réf.*
Âge auquel la personne quitte l'école			
13 ans et moins	- 1,06***	0,39	- 0,55
14 ou 15 ans	- 1,04**	0,48	- 0,44
16 ou 17 ans	- 0,71	0,49	- 0,77
18 ou 19 ans	- 0,33	0,47	- 0,62
20, 21 ou 22 ans	- 0,35	0,48	- 0,81
23 ans et plus	*Réf.*		*Réf.*

Lecture : la taille de l'homme et celle de la femme sont régressées séparément sur le même ensemble de
*au seuil de 1 %, ** : significatif au seuil de 5 %, * : significatif au seuil de 10 %, Réf. : catégorie de référen*
Champ : 30 ans et plus, France métropolitaine.
Source : Panel européen, *vague 2001, Insee.*

FIG. 6.1 – Régression multilinéaire sur les facteurs socioéconomiques de la taille

méthode de la régression linéaire, on peut transformer cette variable en une variable quantitative qui est égale à 1 lorsqu'un ménage possède une voiture et 0 sinon. On voudrait modéliser **la probabilité qu'un ménage disposant d'un certain revenu détienne une voiture**.

Cette modélisation ne sera toutefois pas très satisfaisante si on utilise un modèle linéaire (voir la droite sur le graphique suivant).

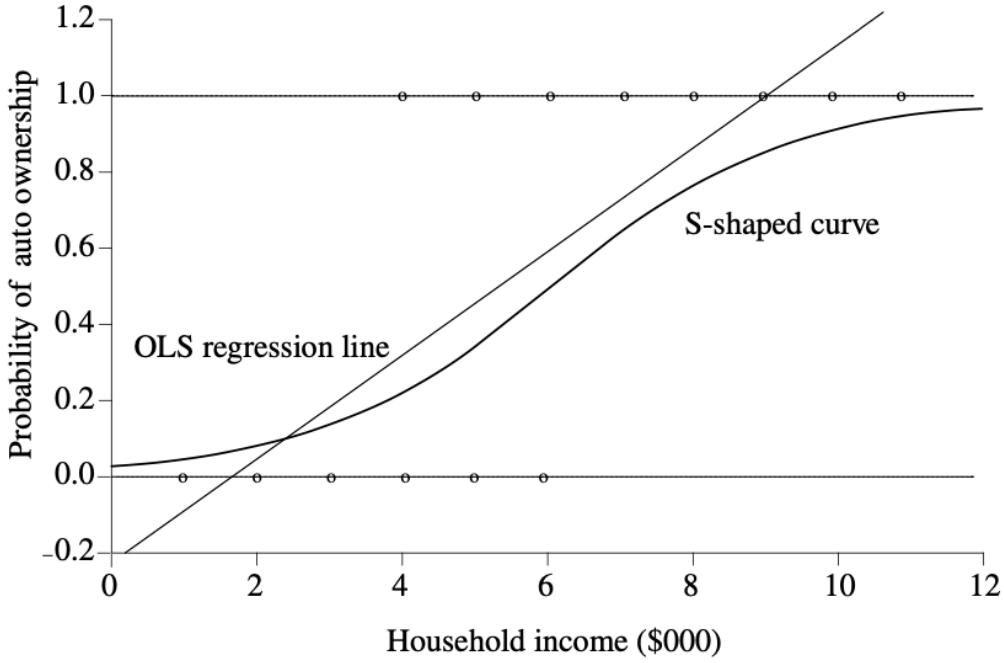


FIG. 6.2 – Possession d'automobile, Detroit, 1919

Quels sont les problèmes ?

1. **Les résultats bizarres**. Pour les ménages dont le revenu est inférieur à $1,540\$$, le modèle prédit une probabilité de posséder une voiture négative. Pour les ménages dont les revenus sont supérieurs à $8,815\$$, le modèle prédit une probabilité de posséder une voiture supérieure à 1. Ça n'a pas vraiment de sens.
2. On voit que la ligne droite n'est pas une forme appropriée pour modéliser la relation entre ces deux variables. Ce qu'on voudrait, c'est une fonction dont la courbe ressemble à la courbe en S sur le graphique : elle doit être croissante et varier de 0 à 1. Cela permet de modéliser une situation où les ménages à faibles revenus auront une faible probabilité de détenir une voiture, tandis que ceux aux revenus élevés auront au contraire une forte probabilité d'en avoir une.

Donc plutôt que de reprendre le modèle linéaire qui n'est pas adapté, on va s'adapter en modifiant quelques éléments :

1. **On laisse tomber l'hypothèse de linéarité**, et on utilise à la place une forme fonctionnelle de régression (c'est-à-dire que l'on va arrêter d'essayer de modéliser Y, mais plutôt $f(Y)$, où f est une fonction que l'on devra choisir.
2. On utilise ce qu'on appelle une **fonction de répartition**, qui donne la probabilité d'un évènement en fonction d'une variable réelle (par exemple, la probabilité de posséder une voiture si l'on a un salaire inférieur à $1,500\$$)

3. Ensuite, toute la question est de savoir quelle fonction utiliser dans notre modèle parmi l'ensemble des fonctions croissantes de 0 à 1. Les régression qui utilisent des fonction logarithmiques sont appelées **logit**, tandis que celles qui utilisent des lois normales sont appelées **probit**. Je vous présente seulement le modèle logit.

6.2.2 Le modèle de la regression logistique

Dans ce modèle, on considère que la probabilité de posséder une voiture se distribue selon une fonction logistique qui peut s'écrire de cette manière :

$$P(Y_i = 1) = \frac{exp(aX_i + b)}{1 + exp(aX_i + b)}$$

où $exp(X)$ est encore la fonction exponentielle.

La probabilité pour le $i$ème ménage de posséder une voiture est de même $P(Y_i = 0)$, qui vaut également $1 - P(Y_i = 1)$:

$$P(Y_i = 0) = 1 - \frac{exp(w_i)}{1 + exp(w_i)} = \frac{1}{1 + exp(w_i)}$$

où $w_i = aX_i + b$.

Une fois que l'on a décidé que l'on choisissait cette fonction pour modéliser la relation entre nos deux variables, la question est, comme pour la régression linéaire, de savoir quels sont les coefficients a et b qui ajustent le mieux la fonction P à la relation observée entre les deux variables. Pour la régression linéaire, on a vu que le principe pour estimer le meilleur modèle était de trouver la droite pour laquelle la somme des résidus était la plus faible. Ici c'est un peu différent, on parle de **maximum de vraissemblance**. Le princpe est un peu plus complexe : l'idée est que l'on a une série de données, et on cherche la distribution de probabilité (parmi une famille de fonction) qui permet le mieux d'expliquer les données. C'est-à-dire la distribution de probabilité qui rend les données obtenues **les plus vraissemblables**.

6.2.3 Dans R

On utilise la fonction `glm()`, pour "Generalized linear models". C'est une famille de modèles de régressions non-linéaires. Pour cette raison, il faut indiquer à la fonction la famille en question. Les modèles logit correspondent à l'argument `family = binomial` :

```
glm_fit <- titanic_train %>%
  glm(as.numeric(Survived) ~ Fare,
      data=.,
      family = "binomial")
```

On a ici utilisé la fonction glm pour faire une régression logistique entre les variables `Survived` (variable dichotomique, que l'on cherche à expliquer) de la table `titanic_train` et la variable `Fare` (variable quantitative, que l'on utilise comme une variable explicative). La question associée à cette régression est celle du lien entre le prix du ticket et la probabilité de survivre au naufrage du Titanic. Voici les résultats de la régression :

```
print(glm_fit)
```

```
##
## Call:  glm(formula = as.numeric(Survived) ~ Fare, family = "binomial",
##     data = .)
##
## Coefficients:
```

```
## (Intercept)         Fare
##     -0.9413       0.0152
##
## Degrees of Freedom: 890 Total (i.e. Null);  889 Residual
## Null Deviance:       1187
## Residual Deviance: 1118  AIC: 1122
```

Les signes des coefficients peuvent être interprétés de la même façon que pour les regression linéaires. Par exemple, si un coefficient est négatif, cela indique que la variable expliquée varie en sens opposé de la variable explicative concernée. En revanche, l'interprétation des coefficients est plus délicate que pour la régression linéaire.

Disons que l'on cherche à savoir quelle est la probabilité de survivre au naufrage du Titanic si l'on a payé un ticket 10$. La valeur de la fonction logit pour cette valeur du ticket **ne peut pas être calculée comme on aurait fait pour une régression linéaire**. C'est-à-dire que le calcul :

$$-0.9413 + 0.0152 * 10 = -0.7893$$

n'a pas de sens.

La première étape pour réussir à interpréter les coefficients est de reprendre l'équation précédente. Si on la réarrange, on peut obtenir :

$$e^w = \frac{P(Y_i = 1)}{P(Y_i = 0)}$$

où e^w est la valeur attendue de la variable expliquée dans la régression. Le ratio des deux probabilités à la droite de l'équation est appelé **odds ratio**

Ici, l'*odds ratio* est la probabilité de survivre au naufrage divisée par la probabilité d'y périr. Lorsque l'*odds ratio* est égal à 1, cela signifie que les deux probabilités sont égales. S'il est égal à 3, cela signifie qu'il y a trois fois plus de chance de survivre, etc. Les valeurs des *odds ratio* sont toujours des nombres strictement positifs.

Si l'on prend le logarithme des deux côtés, puisque $ln(exp(x)) = x$, le côté gauche devient seulement w, tandis que le côté droit est égal au logarithme de l'odds ratio. Donc,

$$w = aX + b = ln(\frac{P(Y_i = 1)}{P(Y_i = 0)})$$

Le logarithme du odds ratio est donc égal à la valeur estimée de la fonction logit pour un ensemble donné de variables explicatives.

Dans nos données du Titanic, la valeur mesurée du logit de la fonction, w, pour $X_1 = 10\$$ était -0.7893. Pour transformer cette valeur et obtenir la probabilité estimée de survivre au naugrage, $P(Y_i = 1)$, que l'on peut écrire $\hat{P}_y$, il faut suivre ces différentes étapes :

1. Prendre l'exponentielle de -0.7893. C'est égal à 0.4541 (la fonction R correspondante est `exp()`)
2. Diviser 0.4541 par 1.4541 pour obtenir la probabilité plutôt que l'odds ratio.

On a donc $\hat{P}_Y$ pour un ticket égal à $\$10$ est $0.4541/1.4541 = 0.3122$ ou 31.2%. Donc d'après notre modèle, un passager qui aurait payé sont ticket $10\$$ aurait environ $31,2\%$ de chance de survivre au naufrage. On pourrait faire les mêmes calculs pour estimer $\hat{P}_Y$ pour n'importe quel valeur du ticket.

Il n'est pas utile de savoir faire ces calculs, je vous les présente surtout pour que vous vous souveniez que l'interprétation des coefficients obtenus lors de la régression logistique sont plus difficile à interpréter que ceux de la régression linéaire. L'important est surtout de comprendre l'intérêt de la régression logistique, et de savoir lire un tableau qui présente les résultats à partir d'*odds ratio*.

6.2.4 Un exemple

Je reprends ici l'exemple proposé par Marie-Paule Couto et Fanny Bugeja-Bloch [Bugeja-Bloch and Couto, 2021, p. 100-104] de la régression logistique tiré d'un article de Sibylle Gollac [Gollac, 2005], dans lequel la sociologue s'intéresse à la place de la fonction publique dans les trajectoires des personnes issues de classes populaires. La question à laquelle elle souhaite répondre est celle de l'influence des origines sociales sur la probabilité d'être fonctionnaire. On pourrait se dire qu'il suffit de réaliser un tableau croisé entre une variable permettant de repérer les fonctionnaires et une autre indiquant l'origine sociale pour répondre à cette question. Ce que fait d'ailleurs l'auteure : en prenant la catégorie sociale du père pour identifier l'origine sociale des individus, elle remarque que 20,7% des enfants d'ouvriers sont fonctionnaires contre 27,6% des enfants de cadres [Gollac, 2005, p. 47]. On pourrait donc dire que, *toutes choses inégales par ailleurs*, les enfants de cadres on plus de chances de devenir fonctionnaires que les enfants d'ouvriers. Mais si l'on cherche à identifier l'effet propre de l'origine sociale, on doit remarquer que les emplois dans la fonction publique requièrent en moyenne un plus haut niveau de diplôme que dans le privé, c'est-à-dire que si l'on compare fonctionnaires et salariés du secteur privé, les cadres et les profession intermédiaires seront largement surreprésentés dans le premier groupe, ce qui explique en partie pourquoi les enfants de cadres deviennent plus fréquemment fonctionnaires : c'est ce qu'on appelle un **effet de structure**.

La régression logistique permet de contrôler cet effet de structure, et de mesurer l'effet de l'origine sociale en contrôlant l'effet du niveau de diplôme. Le résultat de la régression est présenté dans le tableau suivant. Il se lit toujours par rapport à une situation de référence : il s'agit ici de celle d'un homme âgé de 50 à 59 ans, sans diplôme, dont le père était cadre salarié du secteur privé. Ce choix ne modifie pas vraiment les résultats de la régression, mais il en dicte le sens de lecture.

Vous pouvez remarquer qu'ici les variables explicatives sont toutes qualitatives. Le signe des coefficients permet de savoir, par rapport à la situation de référence, si la probabilité d'être fonctionnaire est plus ou moins élevée, *toutes les autres variables étant maintenues constantes*. On peut donc lire sur ce tableau que les enfants d'agriculteurs, d'ouvriers ou d'employés ont plus de chances de devenir cadres toutes choses égales par ailleurs. Si l'on veut exprimer cette différence, on peut facilement calculer l'odds-ratio en prennant l'exponentielle du coefficient : par exemple, les enfants d'ouvriers ont `exp(0.32) = 1.37` fois plus de chance d'être fonctionnaires que les cadres.

Marie-Paul Couto et Fanny Bugeja-Bloch concluent que

> l'effet net dont rend compte ici la régression logistique ne va pas dans le même sens que l'effet brut décrit précédemment dans le tableau croisé […] L'analyse de régressions logistiques, ici heuristique, ne doit cependant pas faire oublier deux aspects : d'abord, que l'expression « toutes choses égales par ailleurs » est en partie abusive puisque seules sont contrôlées les variables introduites dans le modèle ; ensuite, que si elles démêlent les effets des différents facteurs, dans la réalité sociale ils demeurent fortement liés [Bugeja-Bloch and Couto, 2021, p. 104].

Variables	Coefficient	Écart-type	Seu s
Sexe			
Homme	*Référence*		
Femme	0,64	0,02	
Age			
15-24 ans	-0,97	0,05	
25-39 ans	-0,54	0,03	
40-49 ans	-0,10	0,03	
50-59 ans	*Référence*		
60 ans et plus	-0,48	0,08	
Diplôme			
Supérieur	1,34	0,04	
Baccalauréat + 2 ans	0,89	0,04	
Baccalauréat ou équivalent	0,58	0,04	
Diplôme inférieur au baccalauréat	0,29	0,03	
Sans diplôme	*Référence*		
Groupe social du père			
Agriculteurs	0,26	0,07	
Artisans, commerçants, chefs d'entreprise	0,14	0,07	
Cadres	*Référence*		
Professions intermédiaires	0,32	0,04	
Employés	0,35	0,04	
Ouvriers	0,32	0,04	
Statut du père			
A son compte	-0,08	0,06	
Salarié de l'État ou des collectivités locales	0,54	0,03	
Autre salarié	*Référence*		

FIG. 6.3 – Régression logistique sur la probabilité de devenir agent de l'État ou des collectivités locales [@gollac2005]

Références

Bibliographie

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

Cécile Brousse. Définir et compter les sans-abris en europe : enjeux et controverses. *Genèses*, 58(1) :48–71, 2005. URL https://www.cairn.info/revue-geneses-2005-1-page-48.htm. Bibliographie_available : 0 Cairndomain : www.cairn.info Cite Par_available : 1 Publisher : Belin.

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 Cairndomain : 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.

Sibylle Gollac. La fonction publique : une voie de promotion sociale pour les enfants des classes populaires ? *Societes contemporaines*, 58(2) :41–64, 2005. URL https://www.cairn.info/revue-societes-contemporaines-2005-2-page-41.htm. Bibliographie_available : 1 Cairndomain : www.cairn.info Cite Par_available : 1 Publisher : Presses de Sciences Po.

Nicolas Herpin. La taille des hommes : son incidence sur la vie en couple et la carrière professionnelle. *Economie et Statistique*, 361(1) :71–90, 2003. doi : 10.3406/estat.2003.7355. URL https://www.persee.fr/doc/estat_0336-1454_2003_num_361_1_7355. Publisher : Persée - Portail des revues scientifiques en SHS.

Patrick Lehingue. *Subunda : coups de sonde dans l'océan des sondages*. Savoir-agir. Editions du Croquant, Bellecombe-en-Bauges, 2007.

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

Daniel Navarro. *Learning Statistics with R*. null, version 0.5 edition edition, Feb 2015. ISBN 978-1-326-18972-3.

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

Mike Savage, Fiona Devine, Niall Cunningham, Mark Taylor, Yaojun Li, Johs Hjellbrekke, Brigitte Le Roux, Sam Friedman, and Andrew Miles. A new model of social class ? findings from the bbc's great british class survey experiment. *Sociology*, 47(2) :219–250, 2013. doi : 10.1177/0038038513481128. URL http://journals.sagepub.com/doi/10.1177/0038038513481128.

Laurent Toulemon and Nicolas Razafindratsima. *Plan de sondage et pondérations de l'enquête*. La Découverte, 2008. URL https://www.cairn.info/enquete-sur-la-sexualite-en-france--

9782707154293-page-45.htm. Bibliographie_available : 1 Cairndomain : www.cairn.info Cite Par_available : 0 Pages : 45-60 Publication Title : Enquête sur la sexualité en France.

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.