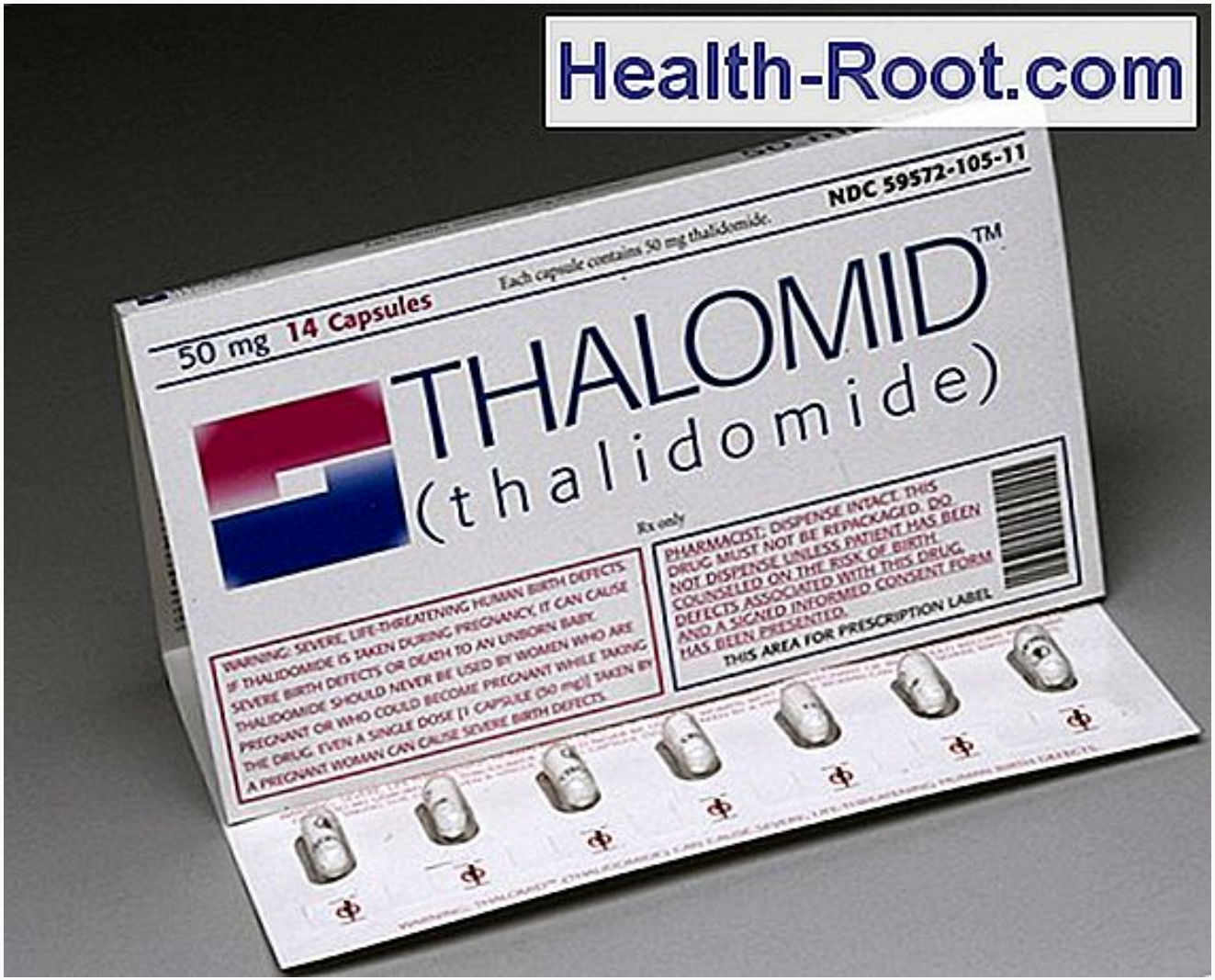
Probabilité et statistique

Yoav Zemel
Adapté des cours de D. Kuonen, A. C. Davison, V. M. Panaretos, G. Dehaene, E. Thibaud, E. Koch, et M. Wilhelm

Introduction

Chimie

- Accidents industriels $\longrightarrow$ besoin de contrôle statistique
- Scandale de Thalidomide (1960s) $\longrightarrow$ tests cliniques rigoureux



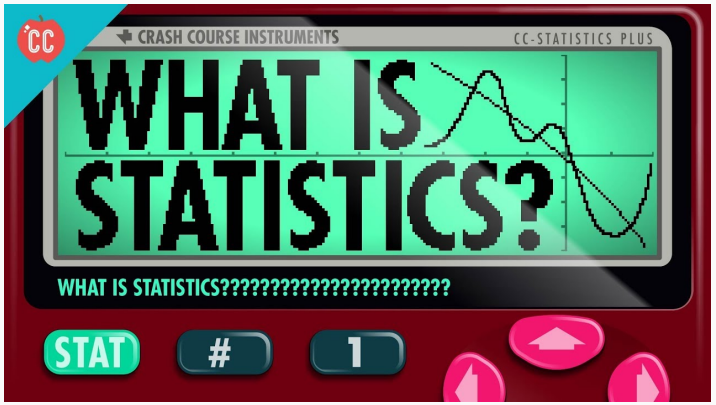
Sciences forensiques

- ADN (1980s) $\longrightarrow$ probabilités de concordance
- Erreurs judiciaires $\longrightarrow$ importance des taux d'erreur

Génie électrique

- Radar (2me guerre mondiale) $\longrightarrow$ détection de signal
- Shannon (1948) $\longrightarrow$ théorie de l'information

Statistique : définition ?



utiliser les mathématiques

pour

extraire des informations

à partir de

données

en présence d'

incertitude.

Statistique : objectifs

Entre autres :

- Description de données.
- Modélisation de données (ajustement d'un modèle statistique) pour, par exemple :
  - effectuer des prévisions (météorologiques, climatiques, économiques, politiques, ...) ;
  - analyser le risque associé à certains phénomènes (calcul de la probabilité d'événements extrêmes, ...).
- Evaluation de l'exactitude d'une théorie scientifique (en physique, chimie, médecine, pharmacologie, ...) en comparant les implications de la théorie et les données.

Et les probabilités ?

La théorie des probabilités nous aide pour la partie “incertitude”. Il s’agit de la discipline mathématique qui étudie les phénomènes **aléatoires** (ou stochastiques).

- Elle sert de base permettant de construire des modèles statistiques prenant en compte le caractère aléatoire du phénomène étudié de manière adéquate.
- Elle fournit également un cadre et de nombreux outils permettant de comprendre et quantifier l’effet de la présence d’aléas sur les informations (conclusions) que l’on extrait des données.

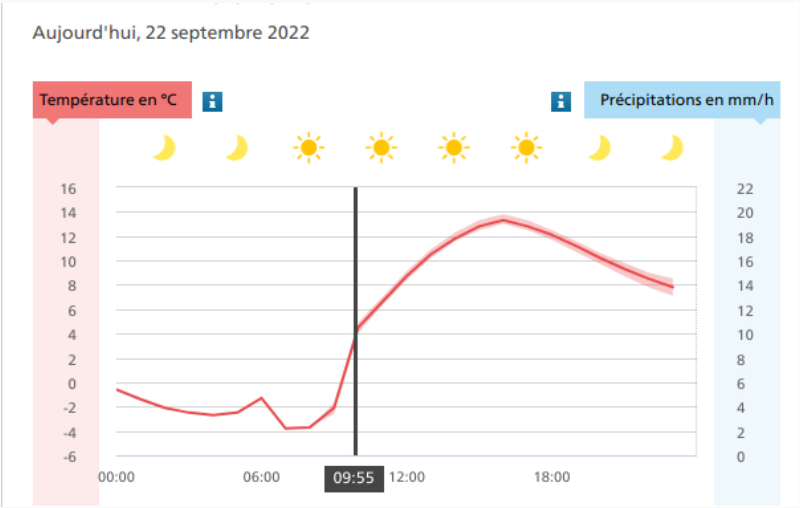
Etapes de la démarche statistique

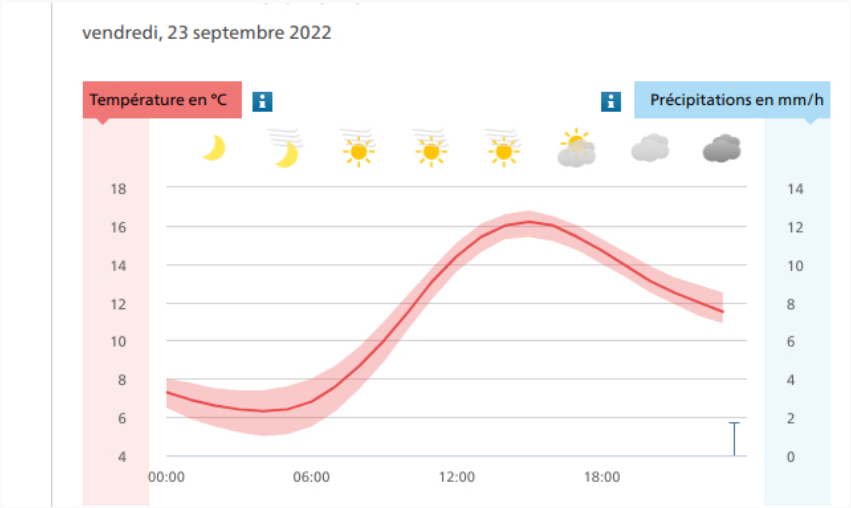
On peut identifier quatre étapes majeures dans la démarche statistique :

- Planification de l'expérience (description théorique du problème, élaboration du plan expérimental) ;
- Recueil des données ;
- **Analyse des données** ;
- Présentation et interprétation des résultats, suivies de conclusions pratiques et d'actions potentielles, toute en prenant en compte l'**incertitude**.

Dans ce cours on va se concentrer sur **l'analyse des données**.

Quantifier l'incertitude





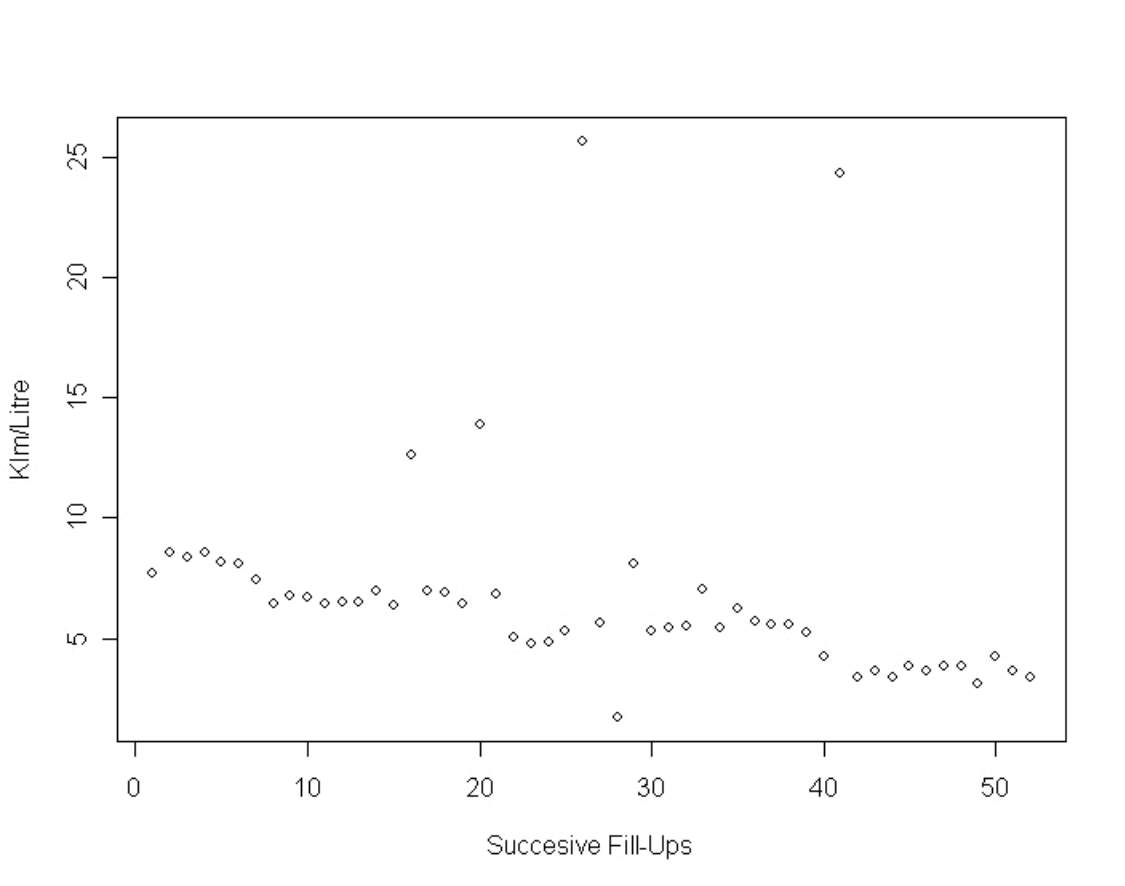
Analyse de données

L'analyse de données est souvent décrite comme comprenant deux phases :

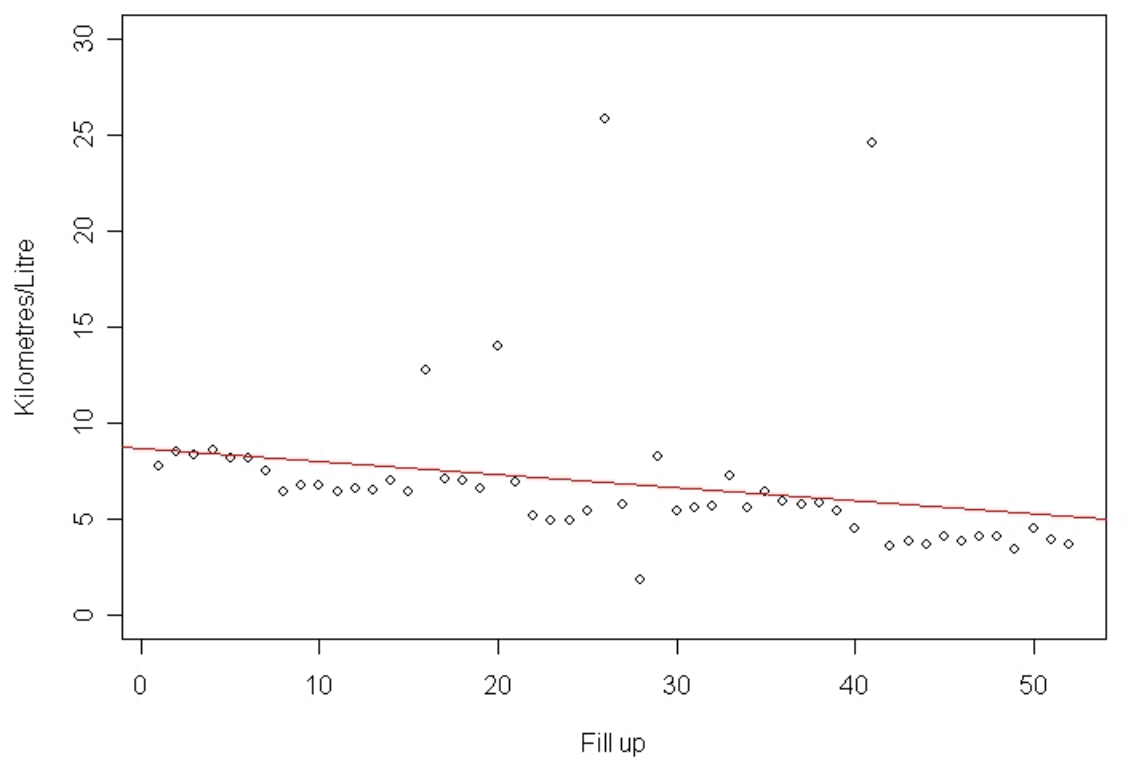
- **Phase 1 : l'analyse exploratoire** ("statistique descriptive") a recours principalement à des méthodes simples, flexibles, souvent graphiques. Elle permet d'étudier la structure des données et de détecter des structures spécifiques (tendances, formes, observations atypiques)
- Exemples :
  - dans quel intervalle la majorité de vos tailles se situe-t-elle ?
  - est-ce que vos tailles et vos poids sont associées ?
  - y-a-t il des personnes "extraordinaires" ?
- Cette phase n'utilise pas des idées probabilistes de façon explicite, elle suggère des hypothèses de travail et des modèles pouvant être formalisés et vérifiés dans la Phase 2 (en principe pas avec les mêmes données !)
- **Phase 2 : l'inférence statistique** conduit à des conclusions statistiques en utilisant des notions probabilistes — des méthodes de test, d'estimation et de prévision

Le camping car du professeur

Le camping car du professeur



Le camping car du professeur



Structure du cours

Ce cours sera divisé en quatre chapitres :

1. **Statistique exploratoire** (1.5 semaines)—types de données, étude graphique des variables, synthèses numériques de distribution, le boxplot, la loi normale
2. **Calcul des probabilités** (5.5 semaines)—probabilités d'événements, variables aléatoires, valeurs caractéristiques, théorèmes fondamentaux
3. **Idées fondamentales de la statistique** (4–5 semaines)—modèles statistiques et estimation des paramètres, estimation par intervalles, tests statistiques, tests khi-deux
4. **régression linéaire** (2–1 semaines)—introduction, principe des moindres carrées, régression linéaire simple, régression linéaire multiple

Matériel de cours

De bons livres de **probabilités** sont

- Ross, S. M. (2007) *Initiation aux probabilités*. PPUR : Lausanne
- Dalang, R. C. et Conus, D. (2018) *Introduction à la théorie des probabilités*, deuxième édition. PPUR : Lausanne
- mais il y a aussi beaucoup d'autres excellents livres de base : regarder au RLC

En **statistiques** : *Introduction à la statistique*, S. Morgenthaler, PPUR, 2014.

Notes de cours en ligne

1. Statistique exploratoire

1.1 Idées de base

Population, échantillon

Imaginons qu'une étude statistique s'intéresse à une caractéristique spécifique (une **variable statistique**, par exemple le poids) chez les individus d'un certain type (par exemple les étudiants de l'UNIL).

Population : tout ensemble sur lequel porte une étude statistique

Echantillon : sous-ensemble de la population

Illustration:

- Population : ensemble des étudiants à l'UNIL
- Echantillon : ensemble des étudiants de 2me année à l'UNIL
- Individu : Un(e) étudiant(e) de 2me année
- Donnée : le poids de l'individu

Types de variables

- Une variable peut être **quantitative** ou **qualitative**
- Une **variable quantitative** peut être **discrète** (souvent entière) ou **continue** :
  - variables quantitatives discrètes : nombre d'enfants dans une famille
  - variables quantitatives continues : poids en kilos
- Une **variable qualitative** (catégorielle) peut être **nominale** (non-ordonnée) ou **ordinale** (ordonnée)
  - variables qualitatives nominales : le groupe sanguin (A, B, AB, O)
  - variables qualitatives ordinales : le plat du jour (bon, passable, mauvais)

  Parfois on convertit des variables quantitatives en variables catégorielles : la taille en cm $\Rightarrow$ (S, M, L, ...)

1.2 Étude graphique de variables

Étude d'une variable qualitative

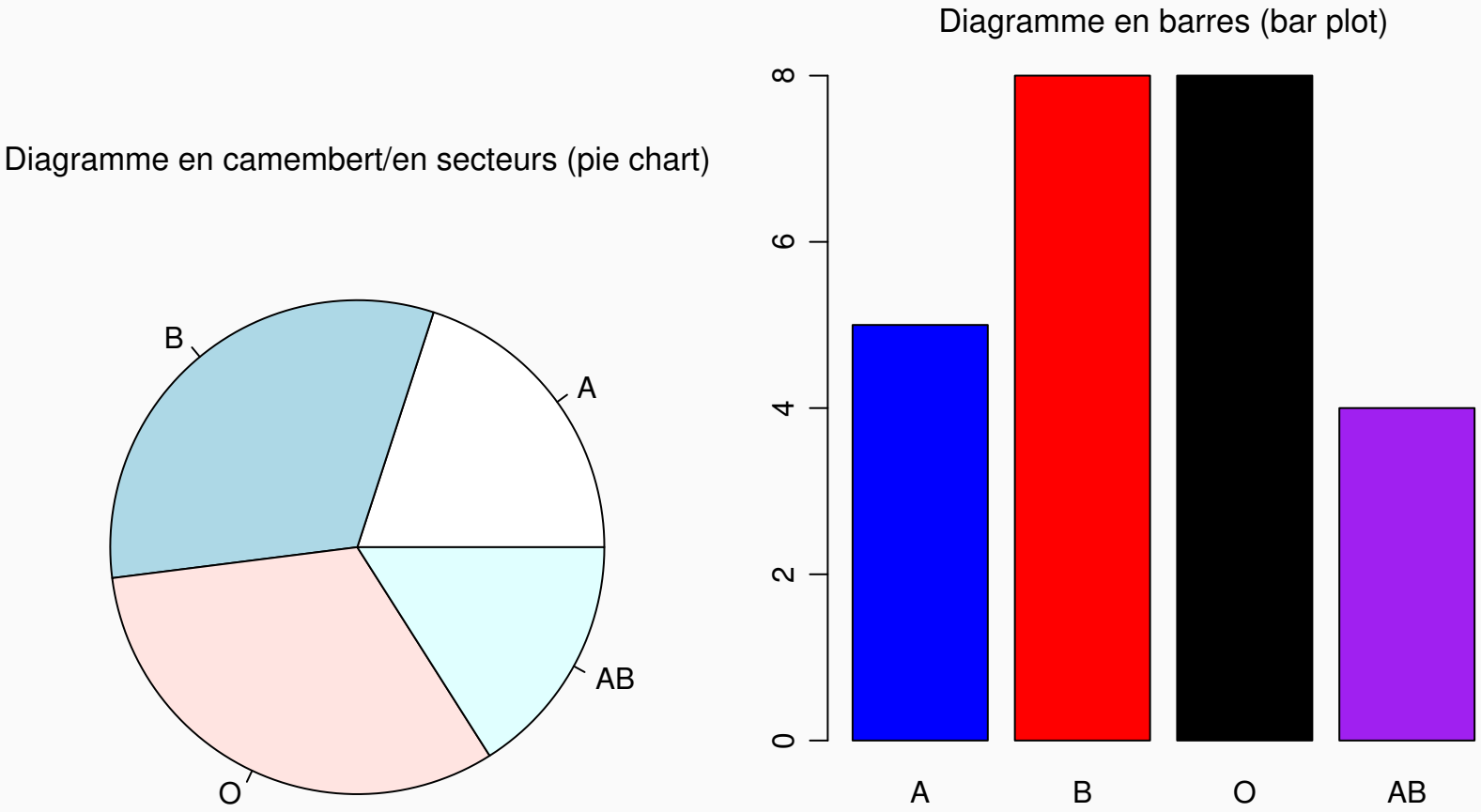
Le groupe sanguin de 25 donneurs a été relevé :

$$\begin{array}{ccccc} AB & B & A & O & B \\ O & B & O & A & O \\ B & O & B & B & B \\ A & O & AB & AB & O \\ A & B & AB & O & A \end{array}$$

La table de fréquences est la suivante :

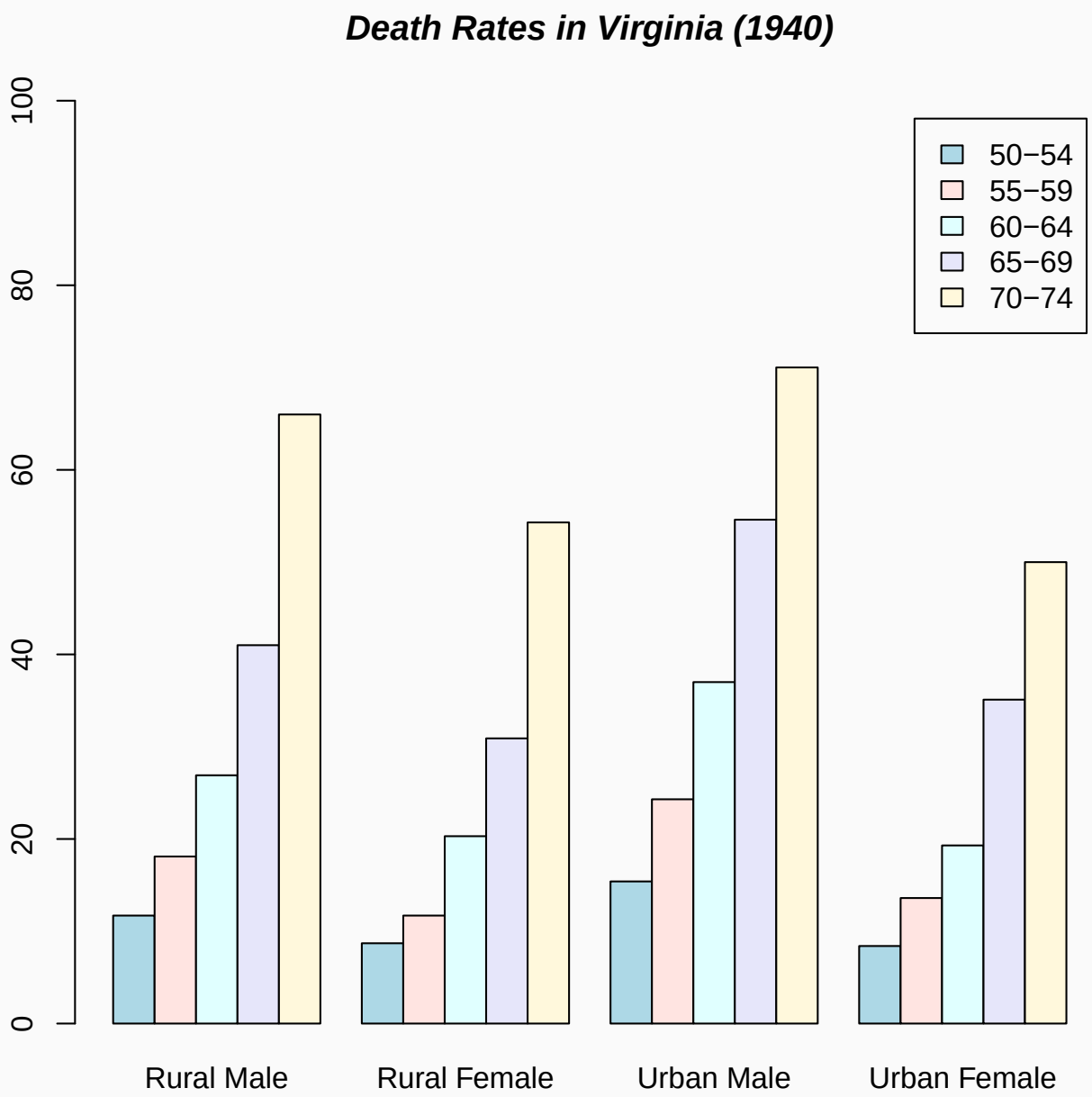
Classe	Fréquence absolue	Fréquence relative
A	5	$5/25 = 0.2$
B	8	$8/25 = 0.32$
O	8	$8/25 = 0.32$
AB	4	$4/25 = 0.16$
Total	25	25/25=1

Diagrammes en camembert et en barres



Nous jugeons mieux les distances que les angles, donc le diagramme en barres est meilleur (et aussi plus flexible)

Diagramme en barres



Histogramme

- Un histogramme montre le nombre d'observations dans des classes issues d'une division en intervalles de même longueur $h > 0$ avec un point de départ $a \in \mathbb{R}$.
- L'histogramme normalisé est l'histogramme divisé par nh.
- Pour construire un histogramme, il est utile de disposer d'une table de fréquences. Celle-ci peut être considérée comme un résumé des valeurs observées.

Histogramme : exemple

Les vitesses (en 1000km/s) avec lesquelles $n = 82$ galaxies de la région couronne boréale sont en train de diverger de notre galaxie.

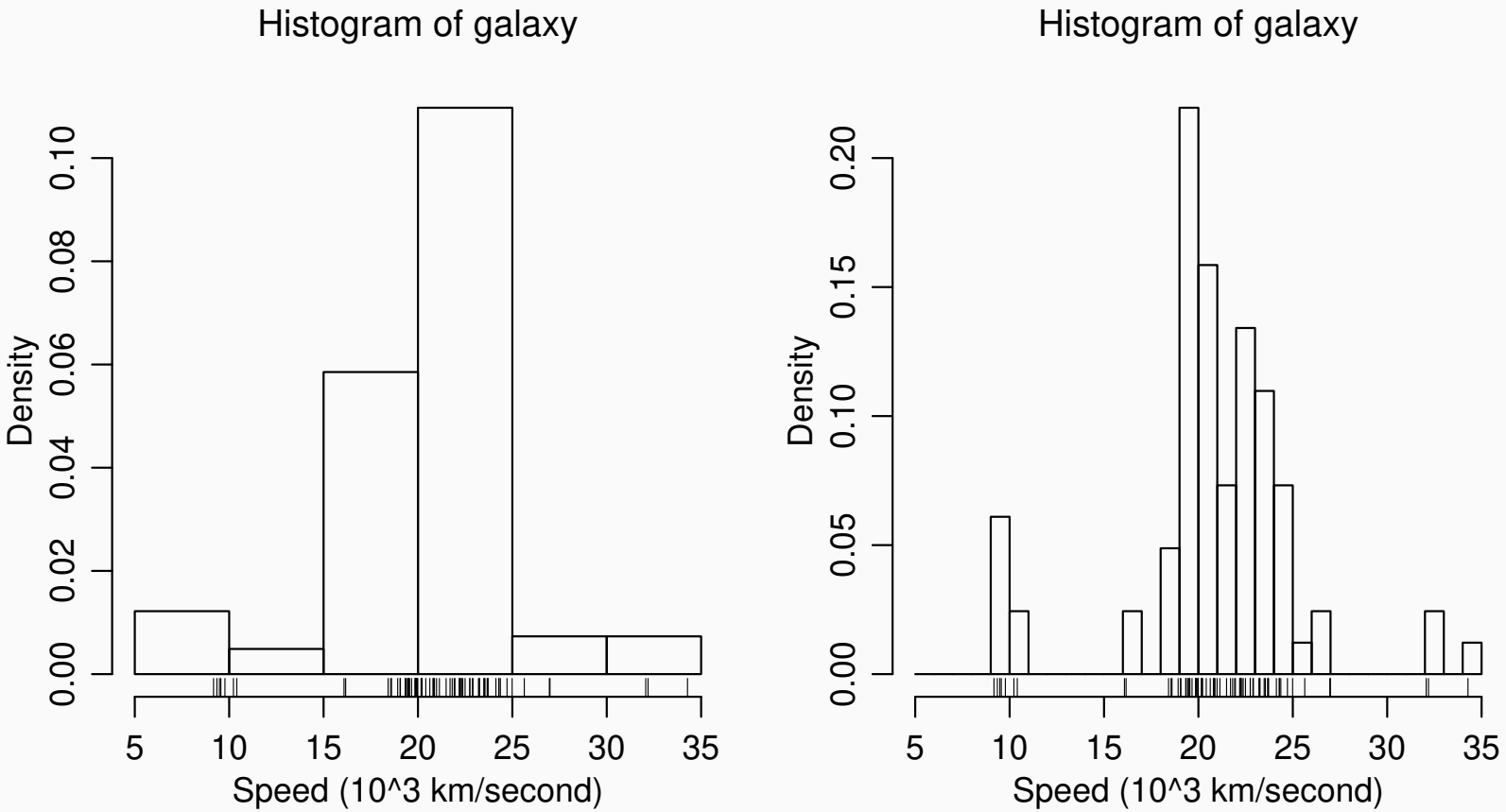
9.172	9.350	9.483	9.558	9.775	10.227	10.406	16.084	16.170	18.419
18.552	18.600	18.927	19.052	19.070	19.330	19.343	19.349	19.440	19.473
19.529	19.541	19.547	19.663	19.846	19.856	19.863	19.914	19.918	19.973
19.989	20.166	20.175	20.179	20.196	20.215	20.221	20.415	20.629	20.795
20.821	20.846	20.875	20.986	21.137	21.492	21.701	21.814	21.921	21.960
22.185	22.209	22.242	22.249	22.314	22.374	22.495	22.746	22.747	22.888
22.914	23.206	23.241	23.263	23.484	23.538	23.542	23.666	23.706	23.711
24.129	24.285	24.289	24.366	24.717	24.990	25.633	26.960	26.995	32.065
32.789	34.279								

Exemple de table de fréquences avec $a = 5$ et $h = 5$:

Classe	Fréquence absolue	Histogramme normalisé
$[5, 10)$	5	0.012
$[10, 15)$	2	0.005
$[15, 20)$	24	0.059
$[20, 25)$	45	0.109
$[25, 30)$	3	0.007
$[30, 35)$	3	0.007

Histogramme : exemple

Histogrammes pour les données des vitesses des galaxies, avec deux choix de h ; les données sont représentées à l'aide des 'tapis' en-dessous



Histogramme, remarques

- **Avantage** : l'histogramme peut être appliqué tout aussi bien à un grand nombre de données qu'à un petit nombre
- **Inconvénients** : les principaux inconvénients de l'histogramme sont la perte d'informations en raison de l'absence des valeurs des observations et le choix délicat de la largeur des boîtes. Il y a différentes possibilités d'interprétation !
- **Remarque** : Il existe des améliorations de l'histogramme, tel que l'estimateur de noyau

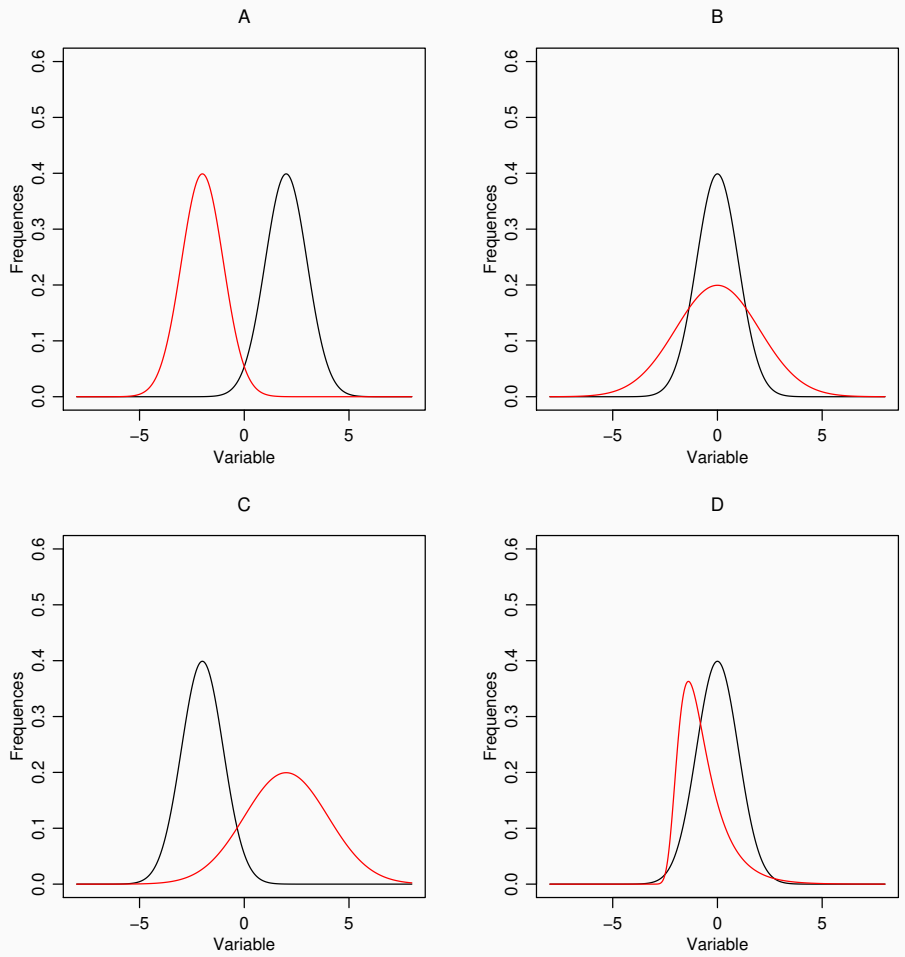
1.3 Synthèses numériques

Caractéristiques principales des données

Pour des **variables quantitatives**, on s'intéresse généralement aux caractéristiques suivantes :

1. la **tendance centrale** qui informe sur le "milieu" (la position/lieu, le centre), par exemple la moyenne et la médiane
2. la **dispersion** qui renseigne sur la variabilité des données autour du centre, par exemple l'étendue, l'écart-type et l'étendue interquartile
3. la **symétrie** ou **asymétrie** par rapport au centre
4. le nombre de **modes** ("bosses")
5. la présence éventuelle de **valeurs aberrantes (outliers)**, qui pourraient provenir d'erreurs de mesures (et donc sont à supprimer), mais pourraient aussi être les données les plus intéressantes, si elles sont correctes

Formes des densités



Centre / dispersion différents ; symétrie vs asymétrie

Tendance centrale

Indicateurs de tendance centrale (mesures de position) :

- La **moyenne** (arithmétique) est

$$\overline{y} = \frac{y_1 + \cdots + y_n}{n} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

  Exemple : la moyenne des vitesses des glaxies est de 20834 km/s.

- La **médiane** est la valeur qui partage l'ensemble des observations **ordonnées** en deux parties de même taille. Ainsi, 50% des données sont plus petites que la médiane et 50% sont plus grandes. Elle est notée $\mathrm{med}(y_1, \ldots, y_n)$ ou $\mathrm{med}(\mathrm{y})$ si $y \in \mathbb{R}^n$ est un vecteur de données.

Médiane

- Afin de définir la médiane, on ordonne les données

$$\min(y_1, \ldots, y_n) = y_{(1)} \leq y_{(2)} \leq \cdots \leq y_{(n)} = \max(y_1, \ldots, y_n).$$

- **Définition:** $\mathrm{med}(y) = y_{(\lceil n/2 \rceil)}$, où $\lceil y \rceil$ est le plus petit entier $\geq y$.
- **Exemple** avec $n = 7$: 1, 4, 7, 14, 10, 12, 9

Médiane

- Afin de définir la médiane, on ordonne les données

$$\min(y_1, \ldots, y_n) = y_{(1)} \leq y_{(2)} \leq \cdots \leq y_{(n)} = \max(y_1, \ldots, y_n).$$

- **Définition:** $\mathrm{med}(y) = y_{(\lceil n/2 \rceil)}$, où $\lceil y \rceil$ est le plus petit entier $\geq y$.
- **Exemple** avec $n = 7$: 1, 4, 7, 14, 10, 12, 9 $\mathrm{med}(y) = y_{(\lceil 7/2 \rceil)} = y_{(4)} = 9$

- **Exemple** avec $n = 8$: 1, 4, 7, 25, 10, 12, 14, 9

Médiane

- Afin de définir la médiane, on ordonne les données

$$\min(y_1, \ldots, y_n) = y_{(1)} \leq y_{(2)} \leq \cdots \leq y_{(n)} = \max(y_1, \ldots, y_n).$$

- **Définition:** $\mathrm{med}(y) = y_{(\lceil n/2 \rceil)}$, où $\lceil y \rceil$ est le plus petit entier $\geq y$.
- **Exemple** avec $n = 7$: 1, 4, 7, 14, 10, 12, 9 $\mathrm{med}(y) = y_{(\lceil 7/2 \rceil)} = y_{(4)} = 9$
- **Exemple** avec $n = 8$: 1, 4, 7, 25, 10, 12, 14, 9 $\mathrm{med}(y) = y_{(\lceil 8/2 \rceil)} = y_{(4)} = 9$
- Parfois on utilise une définition symétrique :

$$\begin{cases} y_{((n+1)/2)}, & n \text{ impaire}, \\ (y_{(n/2)} + y_{(n/2+1)})/2, & n \text{ paire}. \end{cases}$$

- **Exemple** calculer la version symétrique dans les deux exemples ci-dessus

Médiane

- Afin de définir la médiane, on ordonne les données

$$\min(y_1, \ldots, y_n) = y_{(1)} \leq y_{(2)} \leq \cdots \leq y_{(n)} = \max(y_1, \ldots, y_n).$$

- **Définition:** $\mathrm{med}(y) = y_{(\lceil n/2 \rceil)}$, où $\lceil y \rceil$ est le plus petit entier $\geq y$.
- **Exemple** avec $n = 7$: 1, 4, 7, 14, 10, 12, 9 $\mathrm{med}(y) = y_{(\lceil 7/2 \rceil)} = y_{(4)} = 9$
- **Exemple** avec $n = 8$: 1, 4, 7, 25, 10, 12, 14, 9 $\mathrm{med}(y) = y_{(\lceil 8/2 \rceil)} = y_{(4)} = 9$
- Parfois on utilise une définition symétrique :

$$\begin{cases} y_{((n+1)/2)}, & n \text{ impaire}, \\ (y_{(n/2)} + y_{(n/2+1)})/2, & n \text{ paire}. \end{cases}$$

- **Exemple** calculer la version symétrique dans les deux exemples ci-dessus
  Pour $n = 8$, $\mathrm{med}(x) = \frac{1}{2}(x_{(4)} + x_{(4+1)}) = \frac{1}{2}(9 + 10) = 9.5$

Moyenne et médiane

- Si la distribution est symétrique, alors la moyenne $\approx$ la médiane
- La moyenne est plus sensible aux données atypiques (aberrantes) que la médiane :

$$y_1 = 1, \quad y_2 = 2, \quad y_3 = 3 \quad \Rightarrow \quad \begin{cases} \bar{y} = 2, \\ \mathrm{med}(y) = 2, \end{cases}$$

$$y_1 = 1, \quad y_2 = 2, \quad y_3 = 30 \quad \Rightarrow \quad \begin{cases} \bar{y} = 11, \\ \mathrm{med}(y) = 2, \end{cases}$$

On dit que la médiane est **résistante** (robuste).

Quantiles

La médiane partage les données $y_1, \ldots, y_n$ en 50%–50%. Et si on voulait les partager en 25%–75% ou bien une autre fraction ?

Définition: Pour $p \in (0,1)$ le ***p*ème quantile** de $y_1, \ldots, y_n$ est $\widehat{q}(p) := y_{(\lceil np \rceil)}$.

Cas particuliers importants :

- La médiane est $y_{(\lceil n/2 \rceil)}$
- les **quartiles** sont $\widehat{q}(0.25) = y_{(\lceil n/4 \rceil)}$ (**inférieur**) et $\widehat{q}(0.75) = y_{(\lceil 3n/4 \rceil)}$ (**supérieur**)

Parfois on parle de **pourcentile (percentile)** : le p-quantile est le $100p$-pourcentile

Exemple : Calculer des 0.32, 0.01 et 0.95 quantiles des données 42, 27, 31, 45, 31, 31, 29, 36, 34, 39

Quantiles

La médiane partage les données $y_1, \ldots, y_n$ en 50%–50%. Et si on voulait les partager en 25%–75% ou bien une autre fraction ?

Définition: Pour $p \in (0,1)$ le ***p*ème quantile** de $y_1, \ldots, y_n$ est $\widehat{q}(p) := y_{(\lceil np \rceil)}$.

Cas particuliers importants :

- La médiane est $y_{(\lceil n/2 \rceil)}$
- les **quartiles** sont $\widehat{q}(0.25) = y_{(\lceil n/4 \rceil)}$ (**inférieur**) et $\widehat{q}(0.75) = y_{(\lceil 3n/4 \rceil)}$ (**supérieur**)

Parfois on parle de **pourcentile (percentile)** : le p-quantile est le $100p$-pourcentile

Exemple : Calculer des 0.32, 0.01 et 0.95 quantiles des données 42, 27, 31, 45, 31, 31, 29, 36, 34, 39 Ordonner 27, 29, 31, 31, 31, 34, 36, 39, 42, 45, $n = 10$
$p = 0.32 \Rightarrow \lceil np \rceil = \lceil 3.2 \rceil = 4 \Rightarrow y_{(4)} = 31$ donc $\widehat{q}(0.32) = 31$.

De même $\widehat{q}(0.01) = 27$, $\widehat{q}(0.95) = 45$

Les quantiles sont utiles car :

- ils sont faciles à calculer
- ils suggèrent la forme d'une loi sous-jacente
- ils résistent bien aux valeurs aberrantes

Mesures de dispersion

- **l'écart-type (standard deviation)**,

$$s = \left\{ \frac{1}{n-1} \sum_{j=1}^{n} (y_j - \bar{y})^2 \right\}^{1/2} = \left\{ \frac{1}{n-1} \left(\sum_{j=1}^{n} y_j^2 - n\,\bar{y}^2 \right) \right\}^{1/2},$$

  où s^2 est la **variance de l'échantillon** (on verra plus tard pourquoi on divise par $n-1$)
- **l'étendue (range)**, $y_{(n)} - y_{(1)} = \max(y_1, \ldots, y_n) - \min(y_1, \ldots, y_n)$
- **l'étendue/écart interquartile (interquartile range, IQR)**,

$$\mathrm{IQR}(y) = y_{(\lceil 3n/4 \rceil)} - y_{(\lceil n/4 \rceil)}$$

Mesures de dispersion

- **l'écart-type (standard deviation)**,

$$s = \left\{ \frac{1}{n-1} \sum_{j=1}^{n} (y_j - \bar{y})^2 \right\}^{1/2} = \left\{ \frac{1}{n-1} \left(\sum_{j=1}^{n} y_j^2 - n\,\bar{y}^2 \right) \right\}^{1/2},$$

où s^2 est la **variance de l'échantillon** (on verra plus tard pourquoi on divise par $n-1$)

- **l'étendue (range)**, $y_{(n)} - y_{(1)} = \max(y_1, \ldots, y_n) - \min(y_1, \ldots, y_n)$
- **l'étendue/écart interquartile (interquartile range, IQR)**,

$$\mathrm{IQR}(y) = y_{(\lceil 3n/4 \rceil)} - y_{(\lceil n/4 \rceil)}$$

Dans l'exemple précédant 42, 27, 31, 45, 31, 31, 29, 36, 34, 39 on ordonne 27, 29, 31, 31, 31, 34, 36, 39, 42, 45 et trouve $\mathrm{IQR}(y) = y_{(8)} - y_{(3)} = 39 - 31 = 8$

1.4 Le boxplot (boîte à moustache)

Boxplot (boîte à moustache)

Poids (en *pounds*) de 92 étudiants d'une école américaine

140	145	160	190	155	165	150	190	195	138	160
155	153	145	170	175	175	170	180	135	170	157
130	185	190	155	170	155	215	150	145	155	155
150	155	150	180	160	135	160	130	155	150	148
155	150	140	180	190	145	150	164	140	142	136
123	155									
140	120	130	138	121	125	116	145	150	112	125
130	120	130	131	120	118	125	135	125	118	122
115	102	115	150	110	116	108	95	125	133	110
150	108									

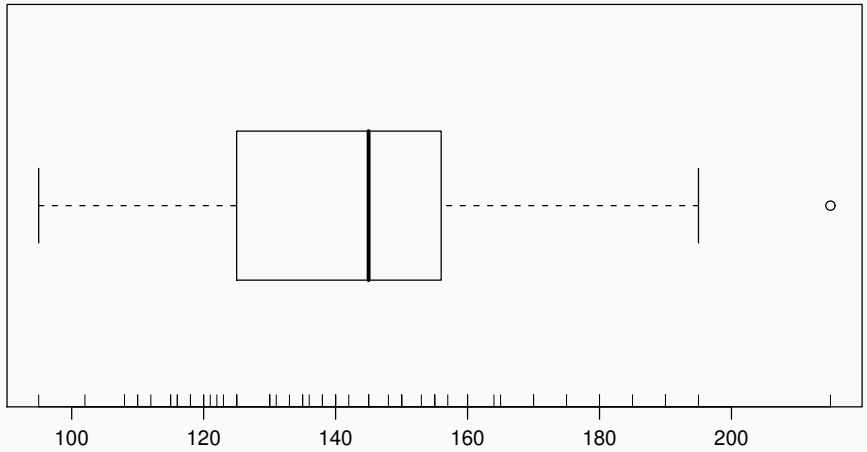
- Le **"five-number summary"** est la liste des cinq valeurs

$$y_{(1)}, \quad y_{(\lceil n/4 \rceil)}, \quad y_{(\lceil n/2 \rceil)}, \quad y_{(\lceil 3n/4 \rceil)}, \quad y_{(n)},$$

donnant un résumé numérique simple et pratique des données

- Cette liste est à la base de la **boîte à moustache (boxplot)**

Boxplot (boîte à moustache)



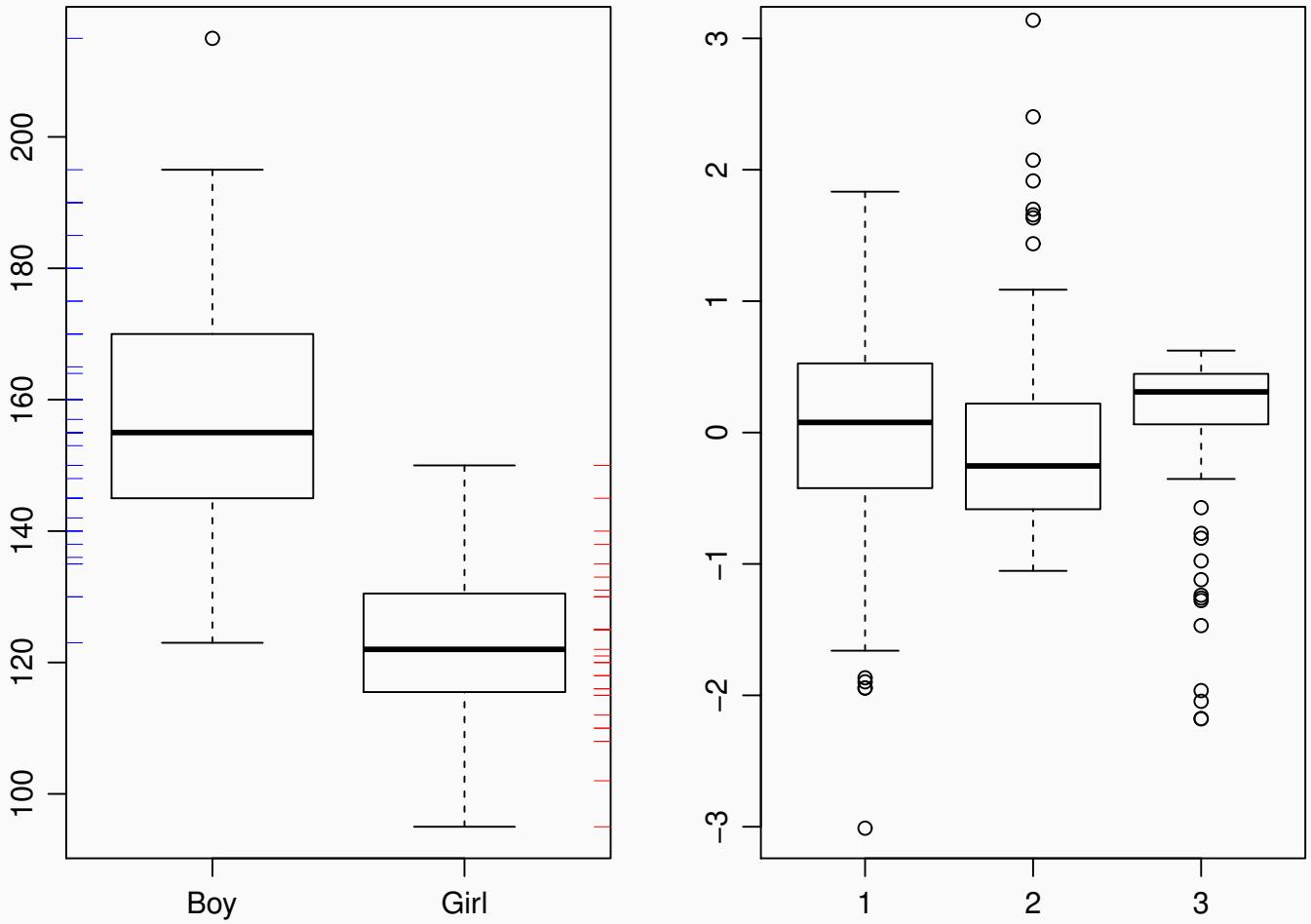
- Pour les poids, le "five-number summary" est $95, 125, 145, 156, 215$, et donc

$$\begin{aligned} \mathrm{IQR}(y) &= y_{(\lceil 3n/4 \rceil)} - y_{(\lceil n/4 \rceil)} = 156 - 125 = 31 \\ C &= 1.5 \times \mathrm{IQR}(y) = 1.5 \times 31 = 46.5 \\ y_{(\lceil n/4 \rceil)} - C &= 125 - 46.5 = 78.5 \\ y_{(\lceil 3n/4 \rceil)} + C &= 156 + 46.5 = 202.5 \end{aligned}$$

- Les limites de la moustache sont les y_i les plus extrêmes qui se trouvent à l'intérieur de l'intervalle $[y_{(\lceil n/4 \rceil)} - C, y_{(\lceil 3n/4 \rceil)} + C]$
- Les y_i à l'extérieur de la moustache sont montrés individuellement

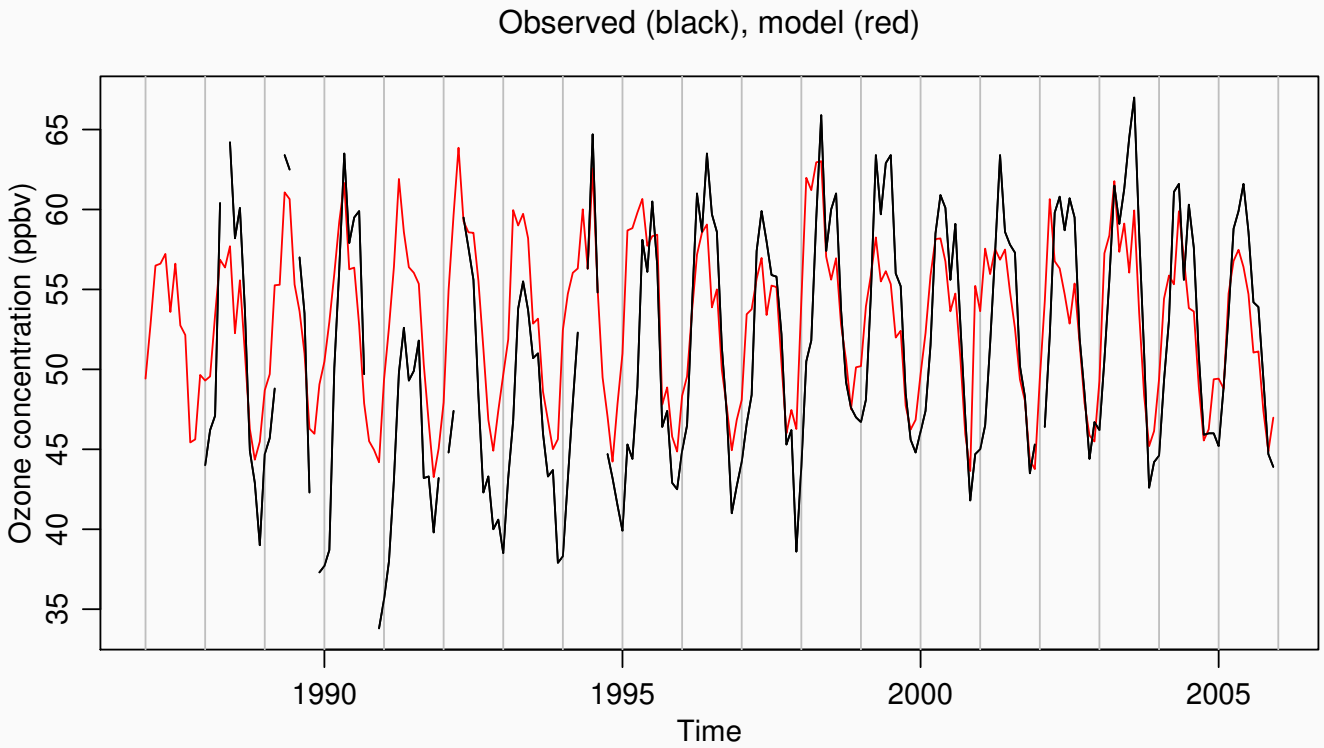
Boxplot (boîte à moustache)

- Le boxplot est utile pour la comparaison de groupes d'observations
- Boxplots du poids des étudiants selon le sexe, et de trois groupes d'observations simulées :



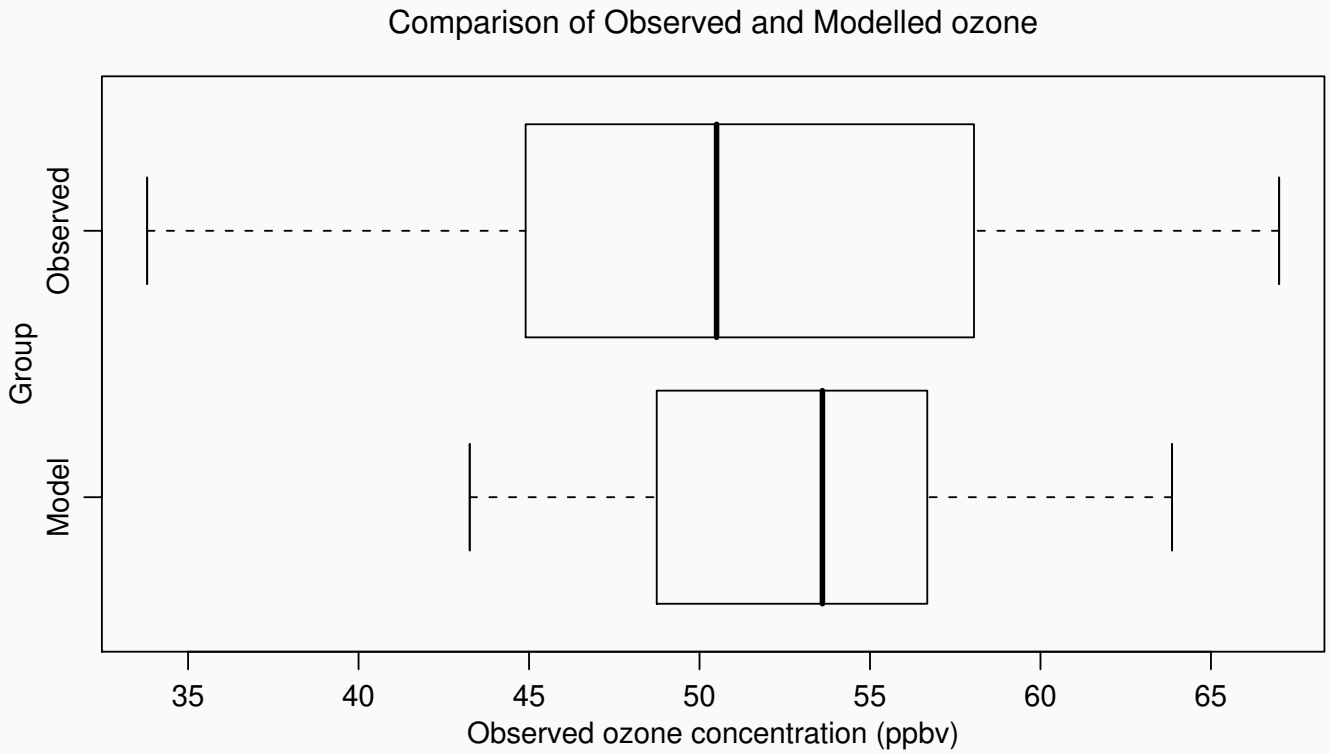
Ozone atmosphérique

Observations de la concentration de l'ozone au Jungfraujoch, de janvier 1987 à décembre 2005 (quelques valeurs manquantes), et résultats d'une modélisation



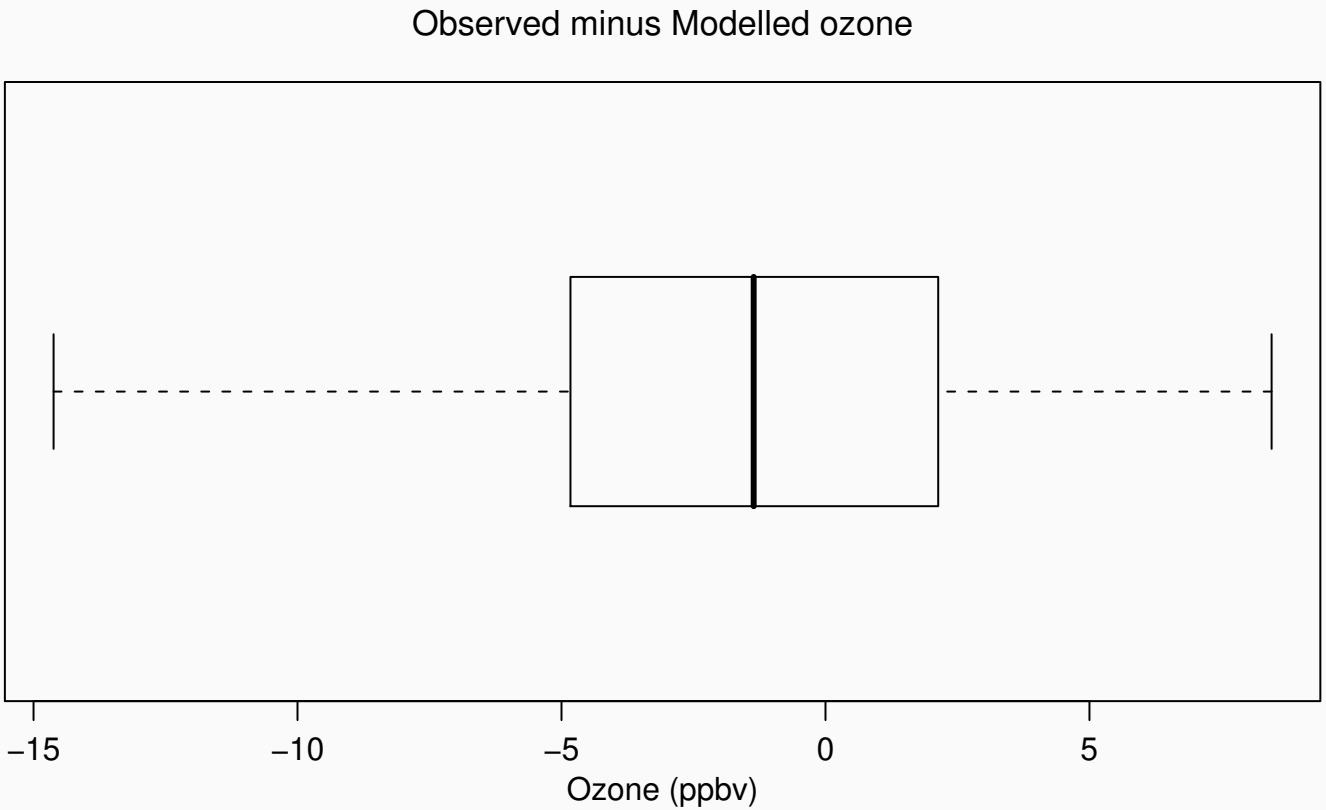
Est-ce que la modélisation est bonne ?

Ozone atmosphérique



Boxplot des données réeles et celles issues du modèle

Ozone atmosphérique



Différences des données réeles et celles issues du modèle

Commentaires

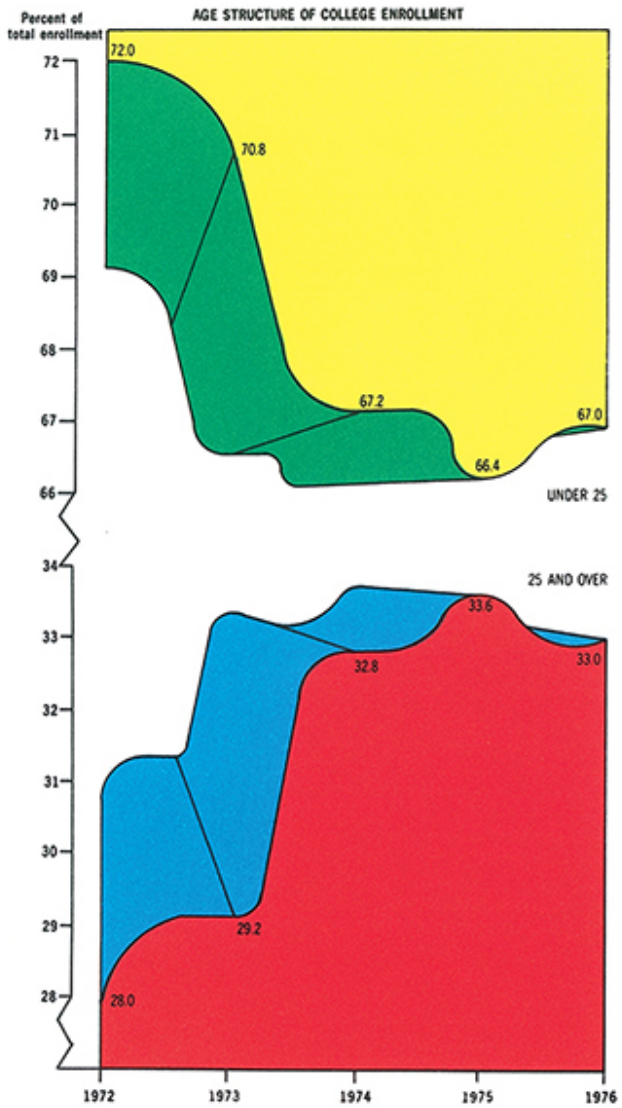
Il n'est pas toujours facile de créer de bons graphiques.

Quelques conseils :

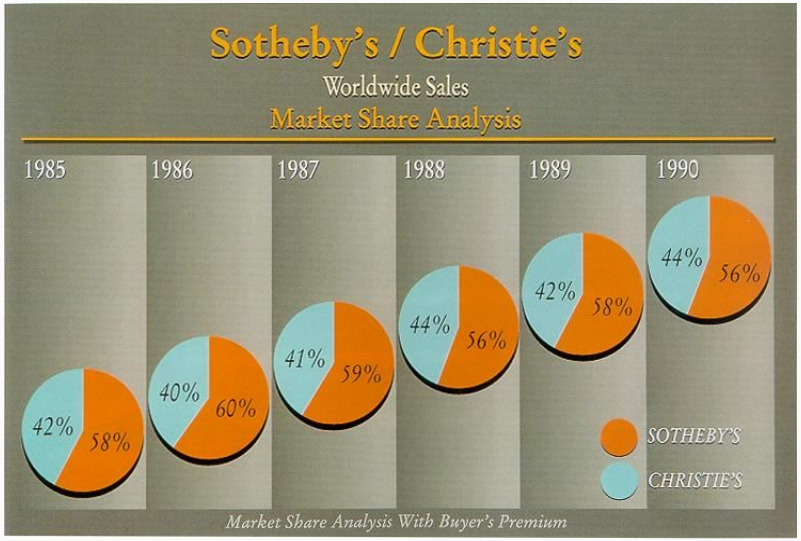
- essayer autant que possible de montrer les données elles-mêmes—pas de **fioritures/chart-junk** (couleurs/lignes/... inutiles etc.)
- mettre des unités et explications claires pour les axes et la légende
- pour comparer des quantités liées, utiliser les mêmes axes et mettre les graphiques en relation proche
- choisir les echelles telles que les relations systématiques apparaissent à un angle de $\sim 45^\circ$ des axes
- **transformer** les données peut aider à la visualisation
- dessiner le graphique de sorte que les départs du 'standard' apparaissent comme départs de la linéarité ou d'un nuage aléatoire de points

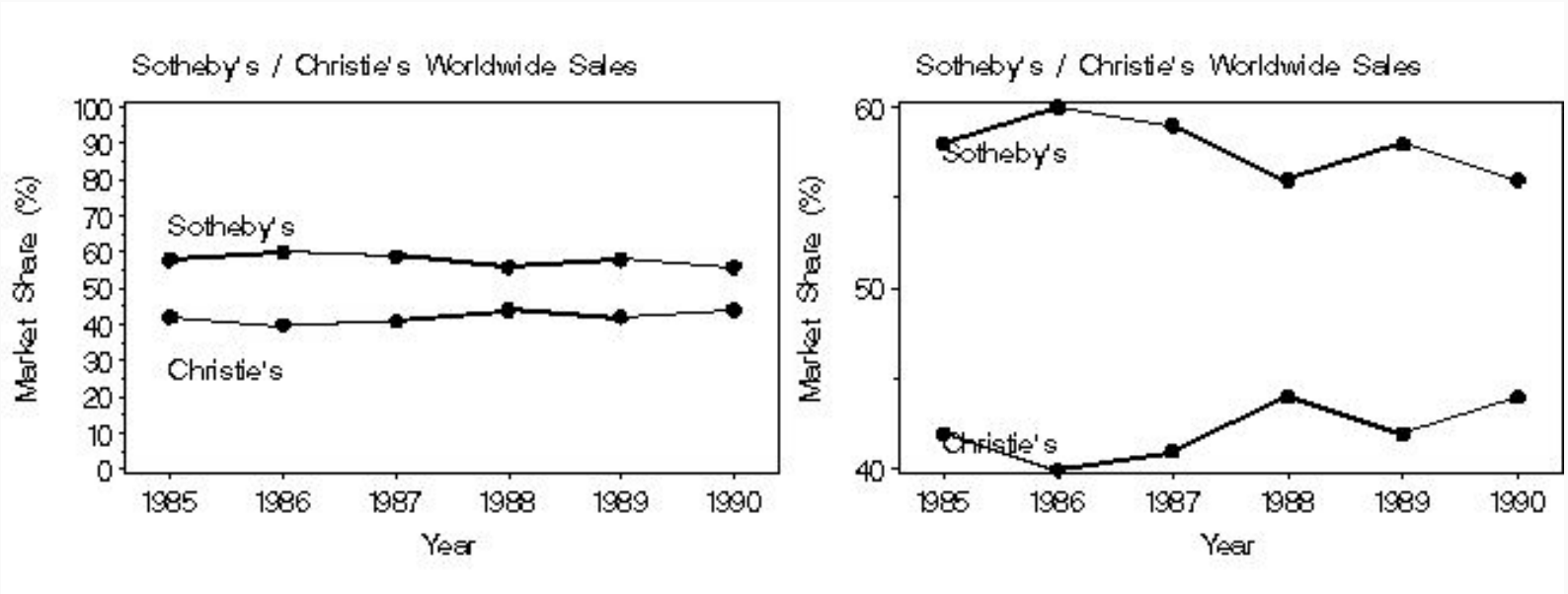
Chartjunk

Ce graphique montre 5 chiffres !



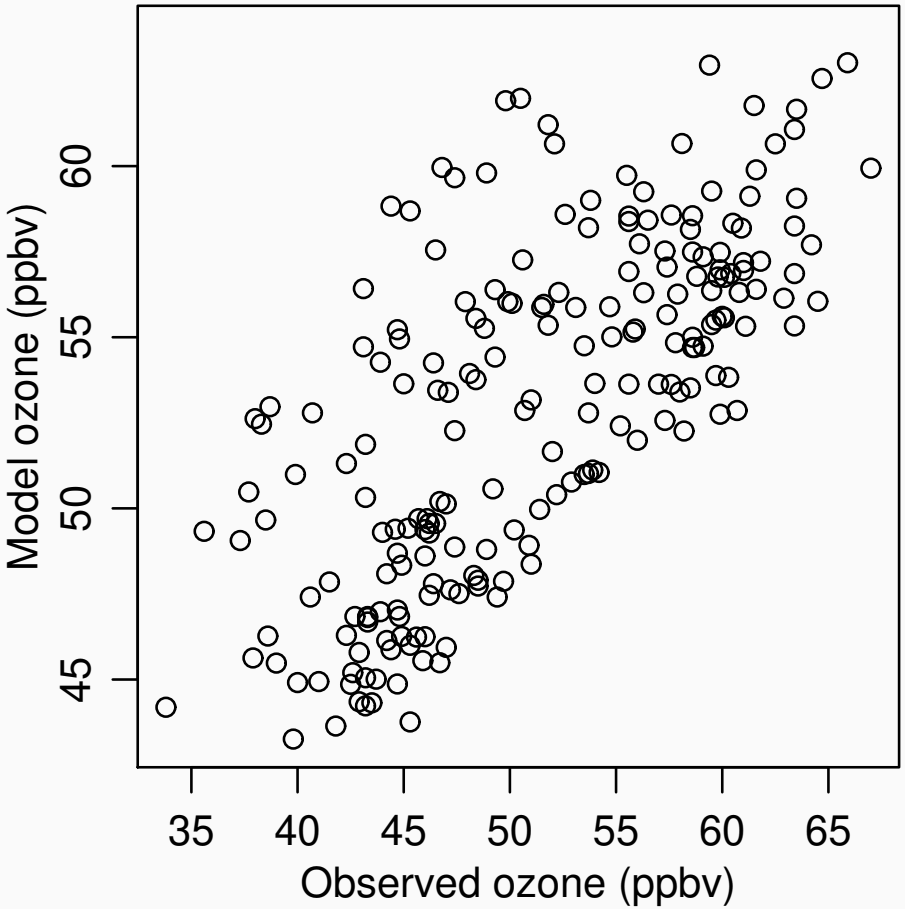
Chartjunk et échelle

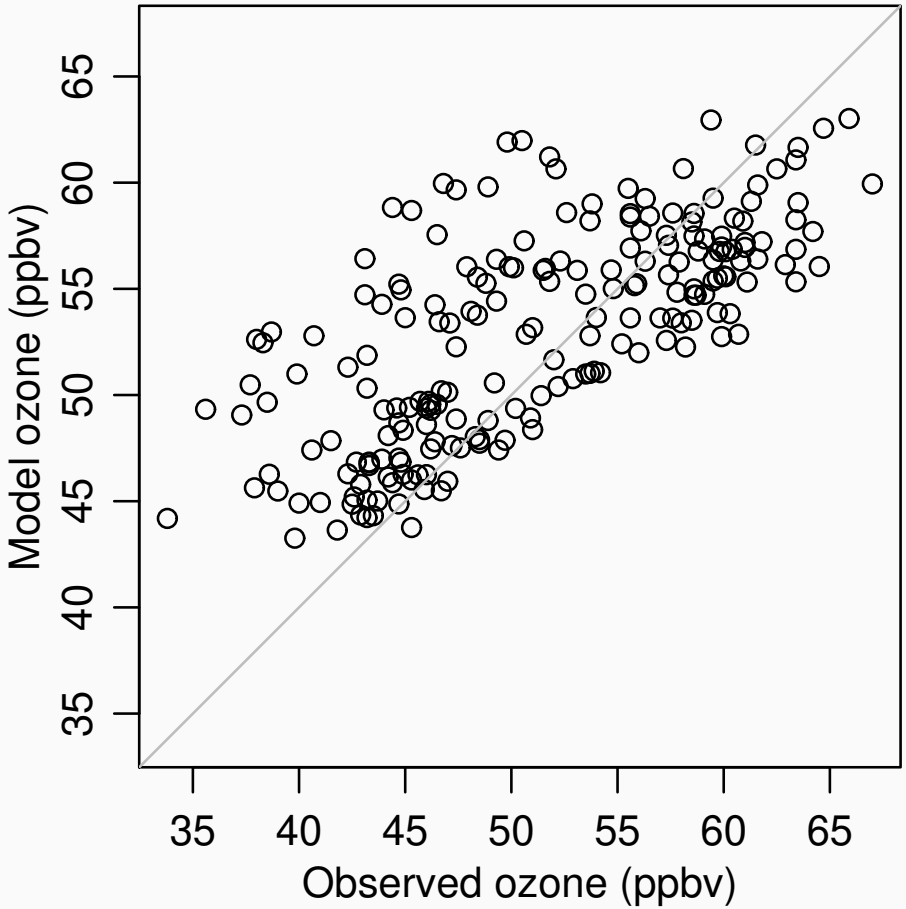




Choisir les bons axes

Effet du choix des axes sur la perception d'une relation :

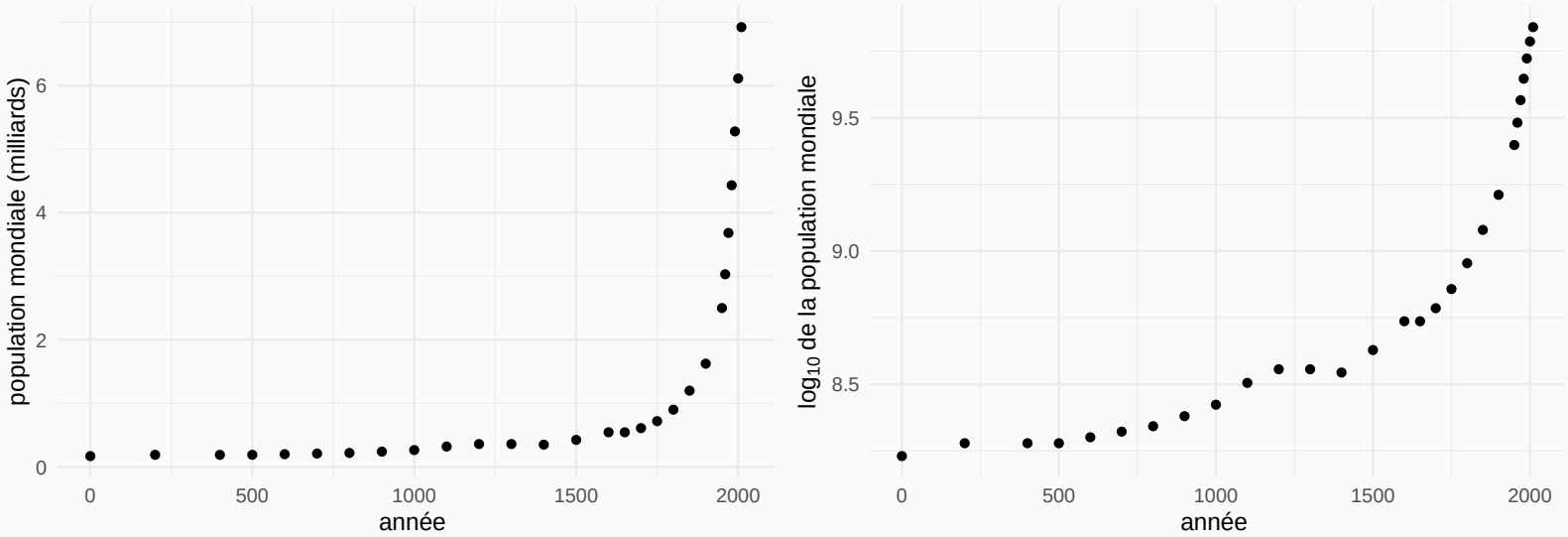




Changements d'échelles

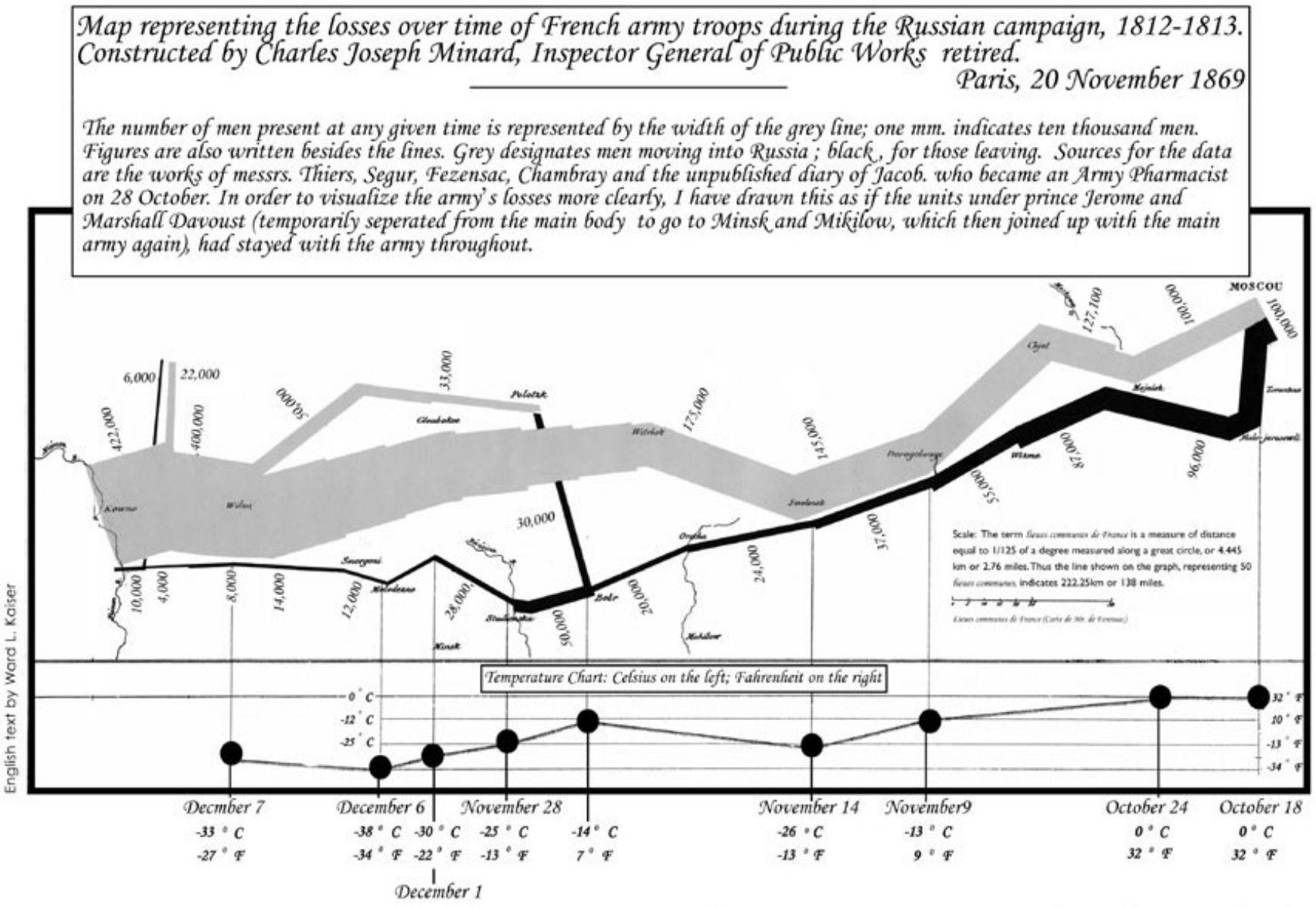
Pour certaines données, il est intéressant de les **transformer** avant de les représenter

Exemple : Population mondiale entre l'an 0 et 2000. L'échelle logarithmique permet de visualiser clairement le taux de croissance



La population en 1200 était de 360 millions, et en 1600 de 545 millions

La campagne russe de 1812



Editor s note: dates & temperatures are only referenced for the retreat from Moscow

1.5 Stratégie

Analyse initiale des données

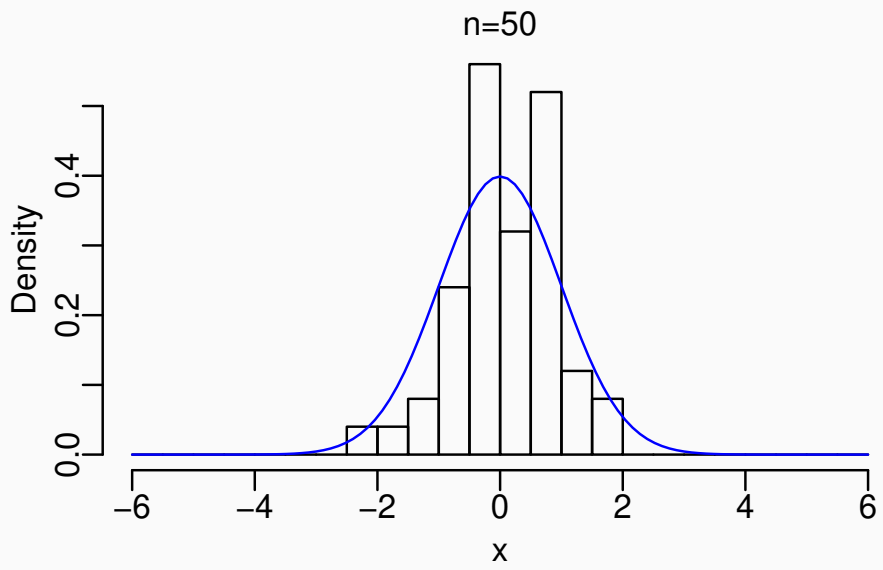
On a maintenant une **stratégie** pour explorer des données issues d'une variable quantitative :

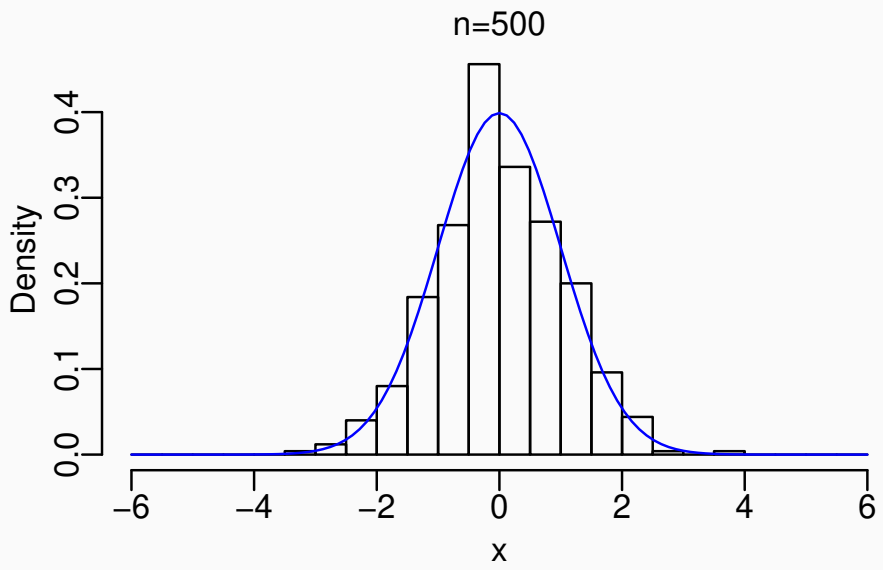
1) toujours faire des **représentations graphiques** d'abord
2) étudier la **structure globale** des données et identifier d'éventuelles valeurs atypiques / aberrantes ("outliers")—trouver pourquoi elles apparaissent
3) calculer des **synthèses numériques** pour décrire la tendance centrale (position / centre / lieu) et la dispersion (échelle)
4) souvent, la structure globale est si régulière qu'on aimerait la décrire par une courbe lisse. Cette courbe est une description mathématique pour la distribution des données

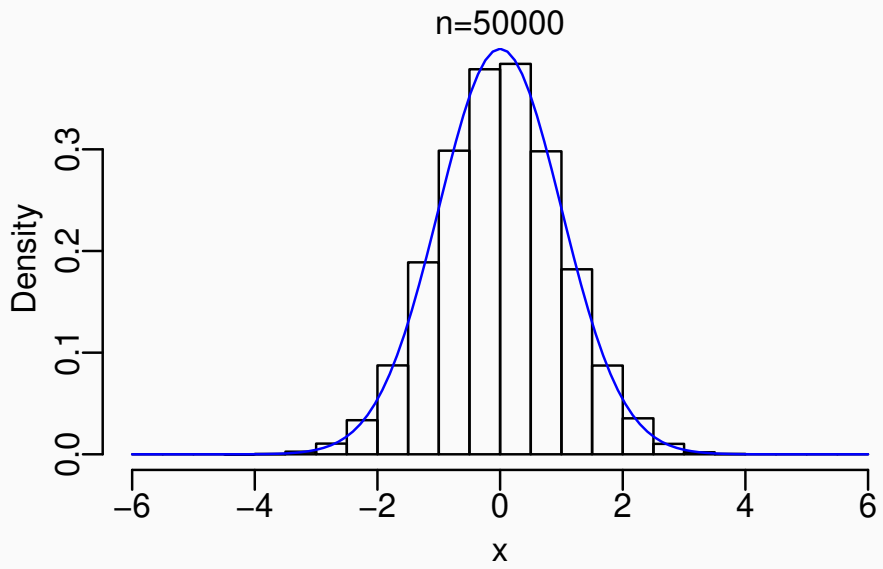
Modélisation des données

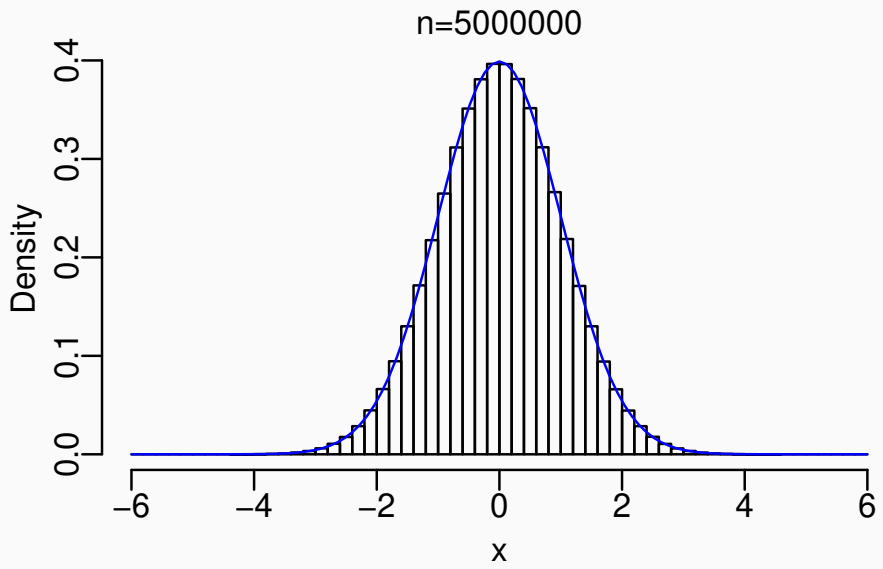
- Souvent on suppose que les données sont issues d'un échantillon aléatoire tiré d'une population d'intérêt
- Cette population est considérée comme très grande, d'une taille presque infinie
- En statistique ces modèles mathématiques sont souvent des **courbes de densité**, une fonction qui est toujours ≥ 0 et qui s'intègre à 1 ; l'aire sous cette courbe est la fréquence relative
- On peut comprendre la courbe de densité comme la limite d'un histogramme normalisé décrivant la structure d'un population de taille n, quand $n \to \infty$ et $h \to 0$

Modélisation des données, courbe de densité









1.5 La loi normale

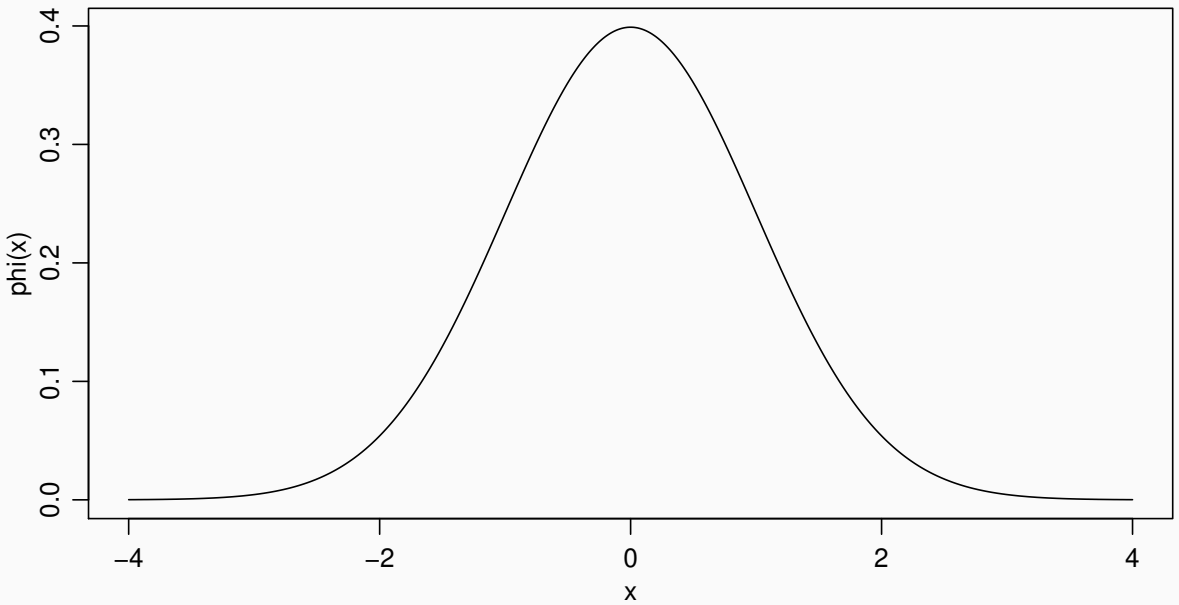
Distribution normale

Une classe particulière et importante de densités est la **densité normale (densité gaussienne)**, $\mathcal{N}(\mu, \sigma^2)$

$$f_{\mu,\sigma}(x) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, \quad -\infty < x, \mu < \infty, \sigma > 0,$$

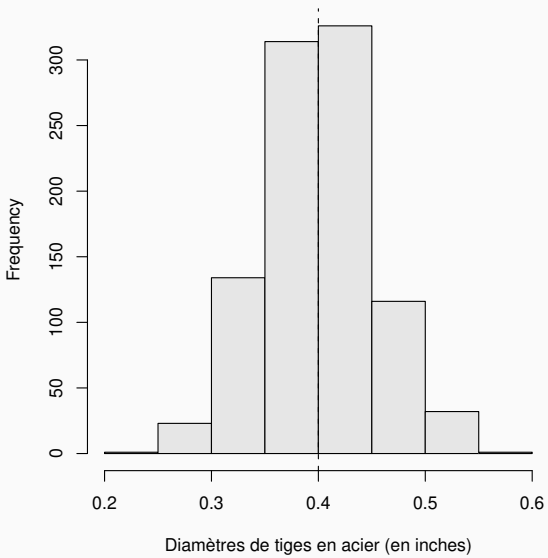
où μ est la **moyenne** et σ est l'**écart-type**

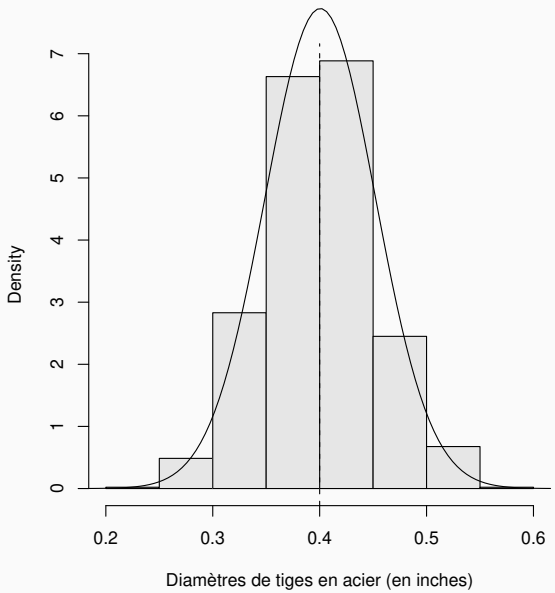
$f_{\mu,\sigma}(x)$ est la hauteur de la courbe au point x



Tiges en acier

Diamètres de 947 tiges en acier en pouces (inches)





Tiges en acier

- Pour obtenir les paramètres, on calcule la moyenne $\bar{x} = 0.4$ et l'écart-type $s = 0.051$
- Courbe précédente : $\mathcal{N}(\mu = 0.40,\ \sigma^2 = 0.051^2)$
- 472 des 947 tiges ont un diamètre ≤ 0.4 inches. Donc leur fréquence relative est
$$\frac{472}{947} = 0.498$$
- L'aire correspondante de la surface sous la courbe précédente vaut 0.5 — proche de 0.498, donc donne une bonne approximation

Propriétés de $\mathcal{N}(\mu, \sigma^2)$

Il y a une infinité des densités normales selon le choix de μ et σ, mais toutes ont des propriétés communes. En voici quelques-unes :

- La majorité des observations d'une "population normale" est proche du centre μ
- La règle "68-95-99.7" :

$$\mathcal{N}(\mu, \sigma^2) \Rightarrow \begin{cases} 68\% \text{ des observations sont dans } [\mu \pm \sigma] \\ 95\% \text{ dans } [\mu \pm 2\sigma] \\ 99.7\% \text{ dans } [\mu \pm 3\sigma] \end{cases}$$

Exemple des tiges: Diamètres de 947 tiges d'acier :

$$\begin{array}{lll} 69.06\% & \text{dans} & [\bar{x} \pm s] \\ 92.05\% & \text{dans} & [\bar{x} \pm 2s] \\ 99.8\% & \text{dans} & [\bar{x} \pm 3s]. \end{array}$$

Le modèle normal semble-t-il être une bonne approximation ?

Si oui, comment calculer ces mêmes proportions à l'aide de ce modèle ?

Standardisation

Si x est une observation issue d'une densité de moyenne μ et d'écart-type σ, alors la **valeur standardisée** de x est

$$z = \frac{x - \mu}{\sigma}$$

z est une observation issue d'une densité de moyenne 0 et d'écart-type 1

Exemple de tiges: Ici, $n = 947$, $\bar{x} = 0.400$, $s = 0.051$, et alors si on met $\mu = \bar{x}$ et $\sigma = s$, on a

$$x_{(644)} = 0.4239 \Rightarrow z_{(644)} = \frac{0.4239 - 0.400}{0.051} = 0.452$$

et de même, la tranformée $x \mapsto z = (x - \mu)/\sigma$ donne

$$\bar{x} = 0.400 \quad \Rightarrow \quad \bar{z} = 0$$

$$s_x = 0.051 \quad \Rightarrow \quad s_z = 1$$

Distribution $\mathcal{N}(0,1)$

La transformée $x \mapsto z = (x-\mu)/\sigma$ donne

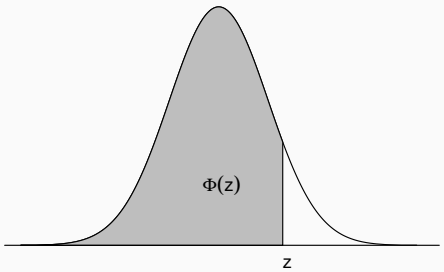
$$\mathcal{N}(\mu, \sigma^2) \mapsto \mathcal{N}(0,1)$$

Ici $\mathcal{N}(0,1)$ dénote la **distribution normale centrée réduite** (loi normale standard), dont la densité est

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \quad z \in \mathbb{R}$$

On définit aussi

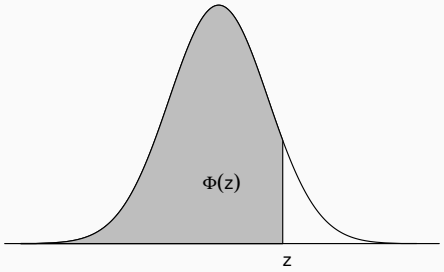
$$\Phi(z) = \int_{-\infty}^{z} \phi(x)\, dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-x^2/2}\, dx, \quad z \in \mathbb{R}$$



Par symétrie de $\phi(z)$ autour de $z = 0$, $\Phi(-z) = 1 - \Phi(z)$

La proportion d'observations dans $[z_1, z_2]$ est $\Phi(z_2) - \Phi(z_1)$

Tableau de $\mathcal{N}(0,1)$



z	0	1	2	3	4	5	6	7	8	9
0.0	.50000	.50399	.50798	.51197	.51595	.51994	.52392	.52790	.53188	.53586
0.1	.53983	.54380	.54776	.55172	.55567	.55962	.56356	.56750	.57142	.57535
0.2	.57926	.58317	.58706	.59095	.59483	.59871	.60257	.60642	.61026	.61409
0.3	.61791	.62172	.62552	.62930	.63307	.63683	.64058	.64431	.64803	.65173
0.4	.65542	.65910	.66276	.66640	.67003	.67364	.67724	.68082	.68439	.68793
0.5	.69146	.69497	.69847	.70194	.70540	.70884	.71226	.71566	.71904	.72240
0.6	.72575	.72907	.73237	.73565	.73891	.74215	.74537	.74857	.75175	.75490
0.7	.75804	.76115	.76424	.76730	.77035	.77337	.77637	.77935	.78230	.78524
0.8	.78814	.79103	.79389	.79673	.79955	.80234	.80511	.80785	.81057	.81327
0.9	.81594	.81859	.82121	.82381	.82639	.82894	.83147	.83398	.83646	.83891
1.0	.84134	.84375	.84614	.84850	.85083	.85314	.85543	.85769	.85993	.86214
1.1	.86433	.86650	.86864	.87076	.87286	.87493	.87698	.87900	.88100	.88298
1.2	.88493	.88686	.88877	.89065	.89251	.89435	.89617	.89796	.89973	.90147
1.3	.90320	.90490	.90658	.90824	.90988	.91149	.91309	.91466	.91621	.91774
1.4	.91924	.92073	.92220	.92364	.92507	.92647	.92786	.92922	.93056	.93189
1.5	.93319	.93448	.93574	.93699	.93822	.93943	.94062	.94179	.94295	.94408
1.6	.94520	.94630	.94738	.94845	.94950	.95053	.95154	.95254	.95352	.95449
1.7	.95543	.95637	.95728	.95818	.95907	.95994	.96080	.96164	.96246	.96327
1.8	.96407	.96485	.96562	.96638	.96712	.96784	.96856	.96926	.96995	.97062
1.9	.97128	.97193	.97257	.97320	.97381	.97441	.97500	.97558	.97615	.97670
2.0	.97725	.97778	.97831	.97882	.97932	.97982	.98030	.98077	.98124	.98169

Exemple

Exemple des tiges: Supposons le modèle normal avec $\mu = \bar{x}$ et $\sigma^2 = s^2$, alors la proportion de x's dans $[\bar{x} - s, \ , \bar{x} + s]$ est la même que celle de z's dans $[-1, 1]$, car

$$[\bar{x} - s, \bar{x} + s] \mapsto \frac{[\bar{x} - s, \bar{x} + s] - \bar{x}}{s} = [-1, 1].$$

Donc la proportion est

$$\Phi(1) - \Phi(-1) = \Phi(1) - \{1 - \Phi(1)\} = 2\Phi(1) - 1 = 0.6826.$$

De même on trouve 0.9544 et 0.9973 pour les proportions des tiges dans

$$[\bar{x} - 2s, \bar{x} + 2s] \mapsto [-2, 2], \qquad [\bar{x} - 3s, \bar{x} + 3s] \mapsto [-3, 3],$$

c'est à dire $\sim 95\%$ et $\sim 99.7\%$ de l'échantillon des tiges, respectivement.