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

Votre avenir . . .

Une mer de données . . . et plein d'emplois intéressants pour ceux qui peuvent la naviguer . . .

Meilleurs emplois 2019

BEST JOBS OF 2019

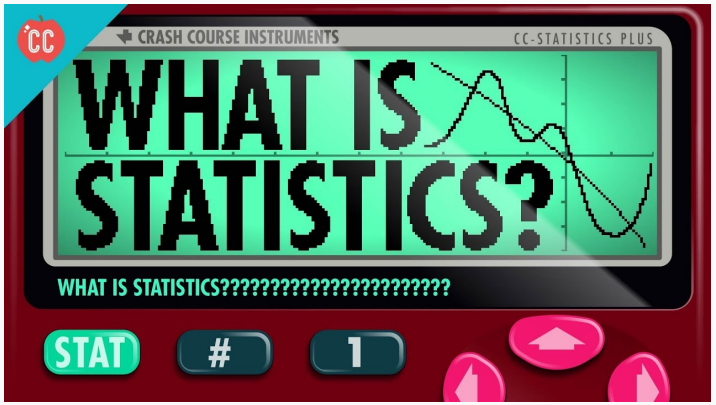
RANKING	PROFESSION	ANNUAL MEDIAN SALARY	GROWTH OUTLOOK (TO 2026)
1	Data scientist	"$118,370 "	19%
2	Statistician	"$88,190 "	33%
3	University professor	"$78,470 "	15%
4	Occupational therapist	"$84,270 "	24%
5	Genetic counselor	"$80,370 "	29%
6	Medical services manager	"$99,730 "	20%
7	Information security analyst	"$98,350 "	28%
8	Mathematician	"$88,190 "	33%
9	Operations research analyst	"$83,390 "	27%
10	Actuary	"$102,880 "	22%

Pires emplois 2019

AT THE BOTTOM OF THE LIST

RANKING	PROFESSION	ANNUAL MEDIAN SALARY	GROWTH OUTLOOK (TO 2026)
211	Broadcaster	"$66,880 "	0%
212	Advertising salesperson	"$51,740 "	-4%
213	Nuclear decontamination technician	"$42,030 "	17%
214	Disc jockey	"$31,990 "	-9%
215	Correctional officer	"$44,400 "	-7%
216	Enlisted military personnel	"$26,802 "	n/a
217	Retail salesperson	"$24,340 "	2%
218	Newspaper reporter	"$43,490 "	-9%
219	Logging worker	"$40,650 "	-13%
220	Taxi driver	"$25,980 "	5%

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.

Mauvaise utilisation de la statistique

Pass the Easter Egg! New study reveals that eating chocolate doesn't affect your Body Mass Index ... and can even help you LOSE weight!

- New research from Roy Morgan reveals there's no proof that chocolate consumption affects BMI
- Currently two thirds of Australians eat chocolate at least once a month
- A study from German researchers has also found there's a connection between cocoa diets and increased weight loss
- Chocolate also found to benefit brain, heart and stress levels

By SAM BAILEY FOR DAILY MAIL AUSTRALIA

PUBLISHED: 01:22 EST, 31 March 2015 | **UPDATED:** 16:14 EST, 31 March 2015

Share 1k shares 55 View comments

From the endless chocolate blocks passed around the office, to the glaring supermarket aisles and the family relatives who miraculously appear with baskets of eggs, Easter can be a minefield to navigate if you're trying to watch your waistline.

But according to new research, there's no need to go easy on the eggs this week, with a **Roy Morgan study** revealing there is no direct connection between chocolate consumption and an increasing Body Mass Index (BMI).

This should come as sweet relief for chocoholics when according to **Roy Morgan**, two thirds of Australians admit to munching on chocolate at least once a month.

Scroll down for video

Eggsellent news: A chocolate a day is found to not affect your Body Mass Index

I Fooled Millions Into Thinking Chocolate Helps Weight Loss. Here's How.

By John Bohannon | 5/27/15 4:23PM | Comments (476)

"Slim by Chocolate!" the headlines blared. A team of German researchers had found that people on a low-carb diet lost weight 10 percent faster if they ate a chocolate bar every day. It made the front page of *Bild*, Europe's largest daily newspaper, just beneath their update about the Germanwings crash. From there, it ricocheted around the internet and beyond, making news in more than 20

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.

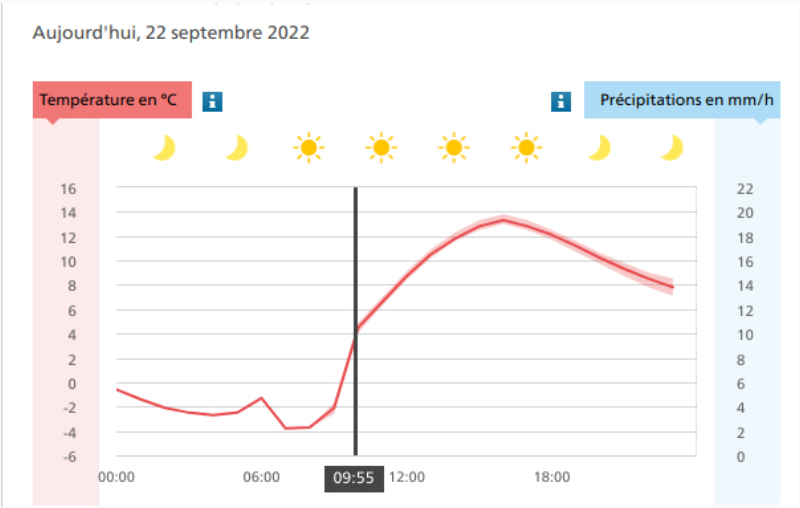
Etapes de la démarche statistique

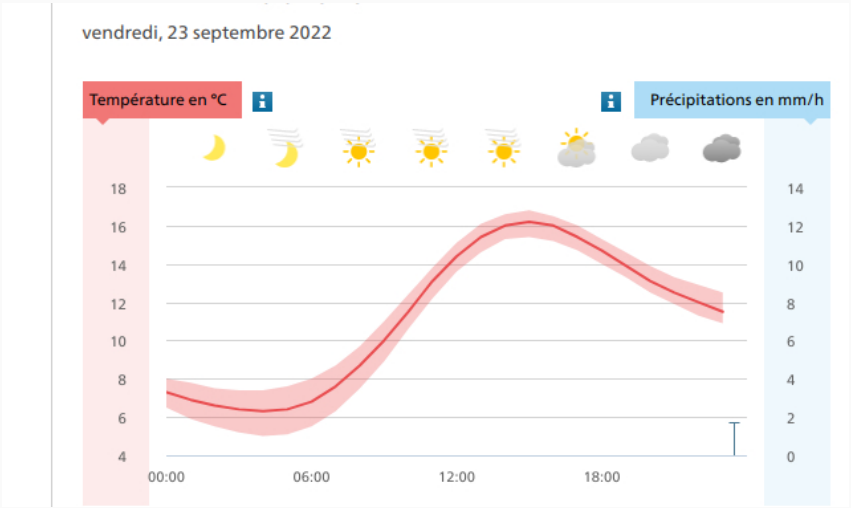
On peut identifier quatre étapes majeures dans la démarche statistique :

- 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





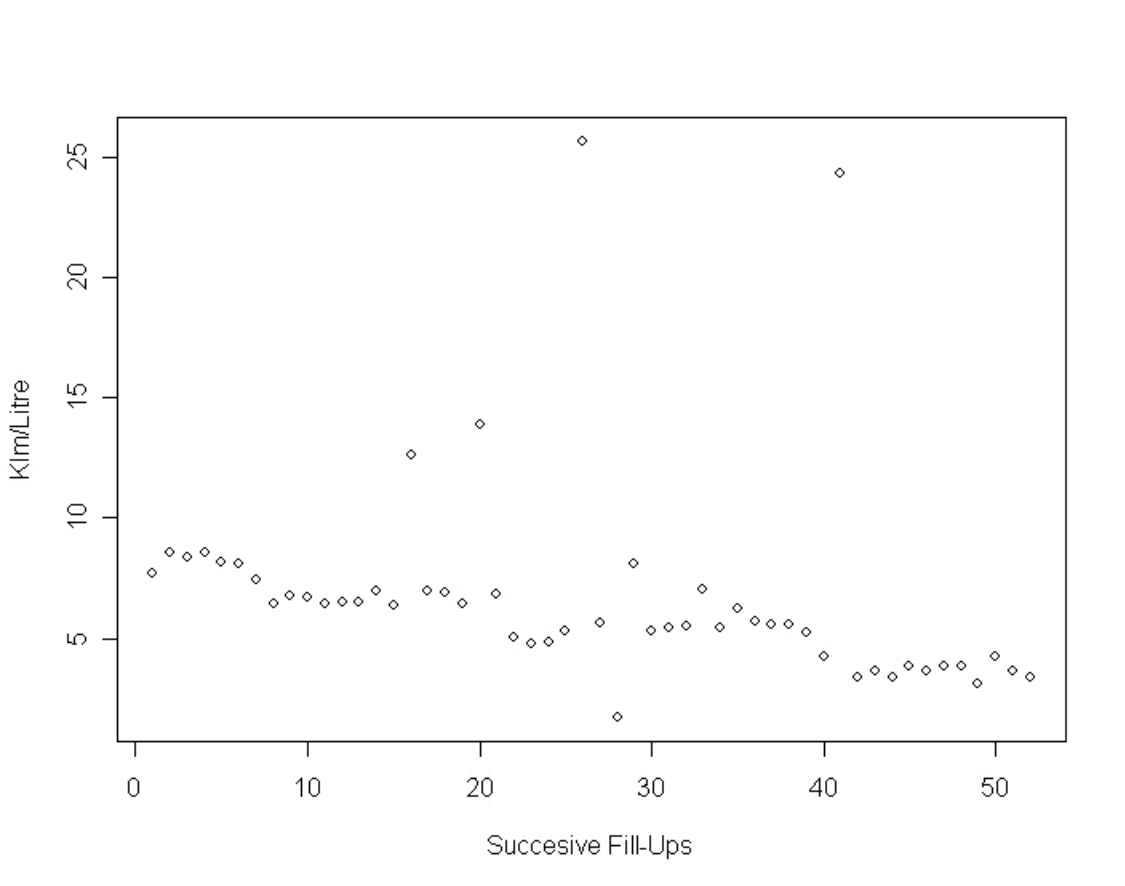
Analyse de données

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

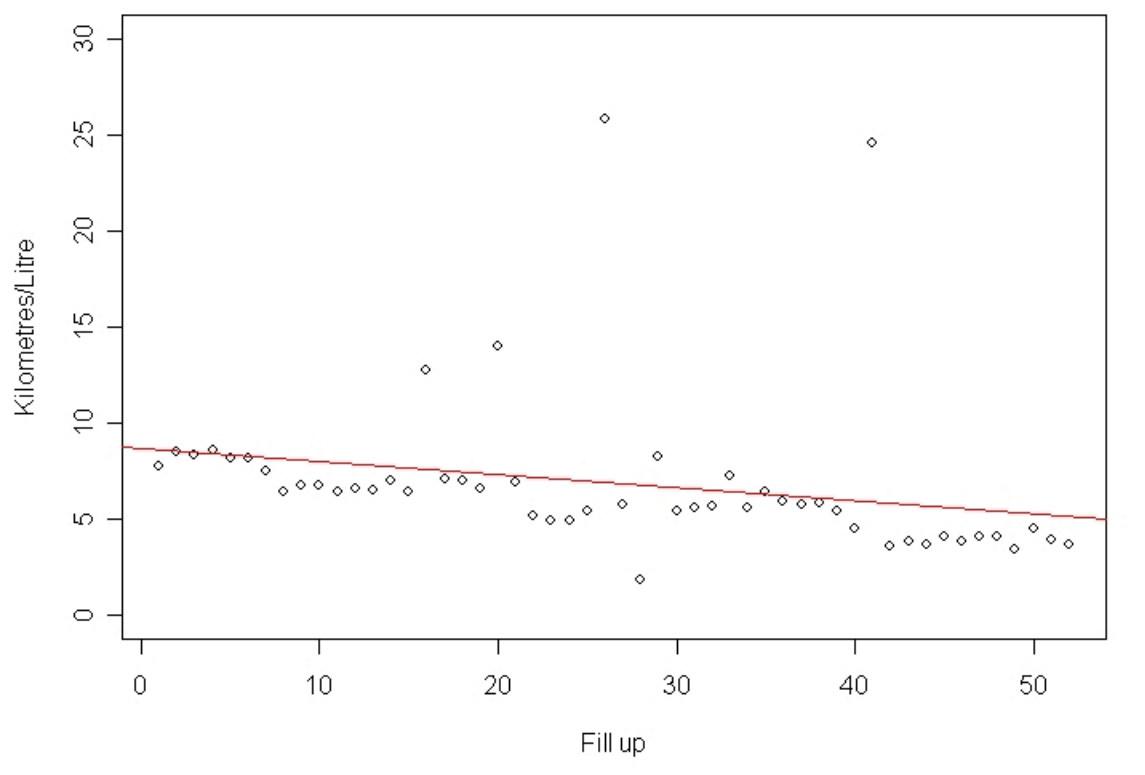
- **Phase 1 : l'analyse exploratoire** ("statistique descriptive") a recours principalement à des méthodes simples, flexibles, souvent graphiques. Elle permet d'étudier la structure des données et de détecter des structures spécifiques (tendances, formes, observations atypiques)
- Exemples :
  - dans 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** (2 semaines)—types de données, étude graphique des variables, synthèses numériques de distribution, le boxplot, la loi normale
2. **Calcul des probabilités** (6 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'EPFL).

Population : tout ensemble sur lequel porte une étude statistique

Echantillon : sous-ensemble de la population

Illustration:

- Population : ensemble des étudiants à l'EPFL
- Echantillon : ensemble des étudiants de 2me année à l'EPFL
- 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

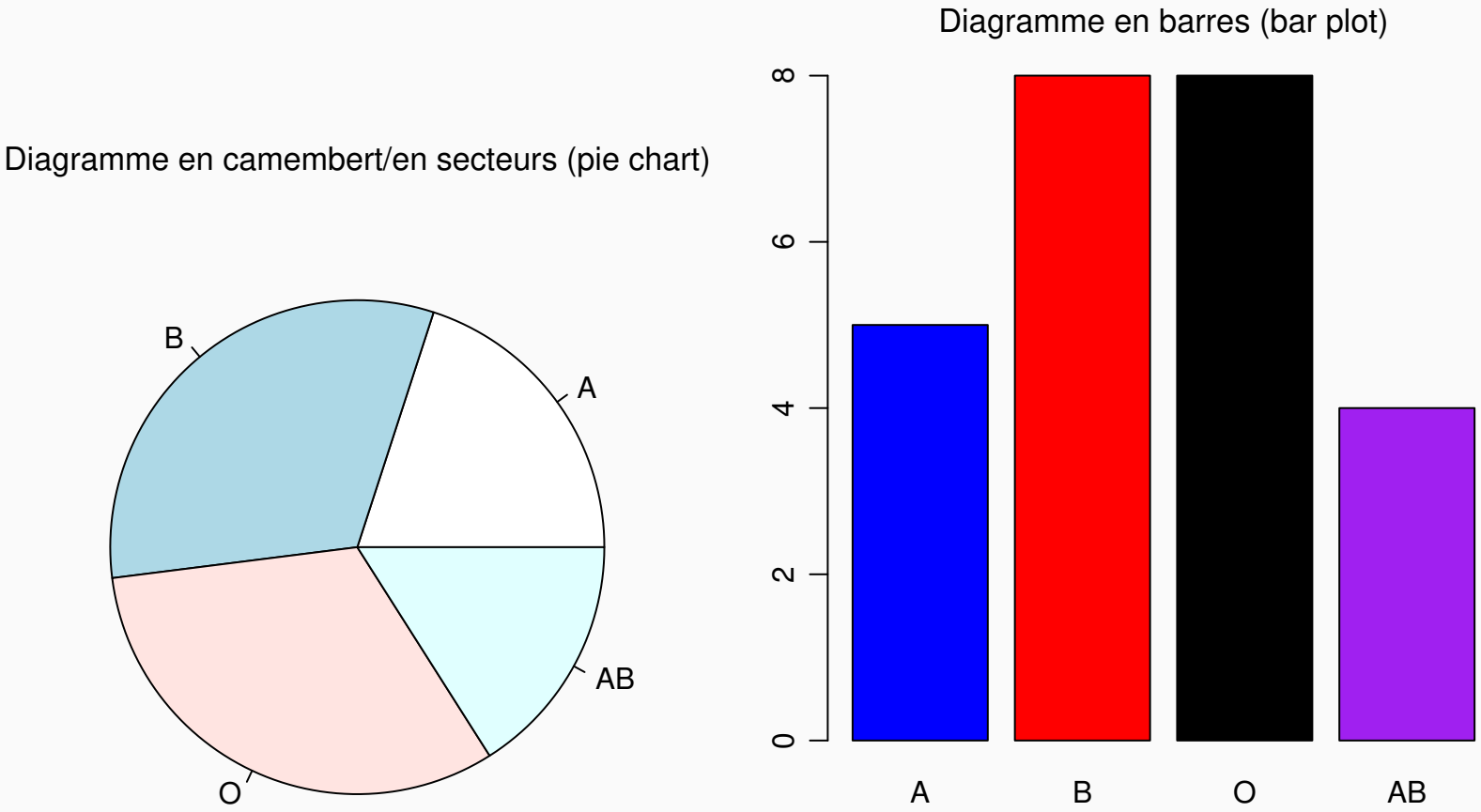
Le groupe sanguin de 25 donneurs a été relevé :

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

La table de fréquences est la suivante :

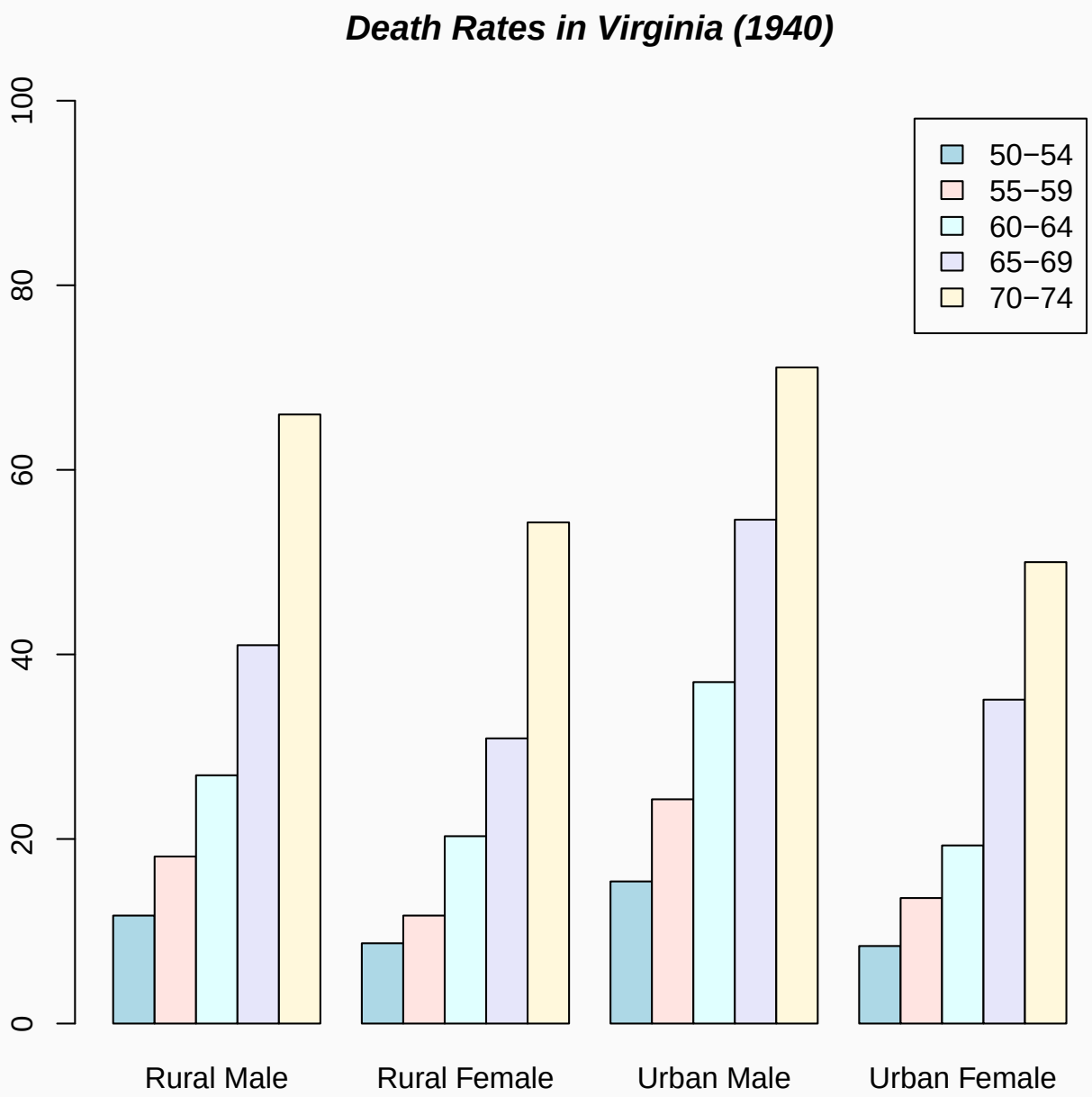
Classe	Fréquence absolue	Fréquence relative
A	5	$5/25 = 0.2$
B	8	$8/25 = 0.32$
O	8	$8/25 = 0.32$
AB	4	$4/25 = 0.16$
Total	25	25/25=1

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.

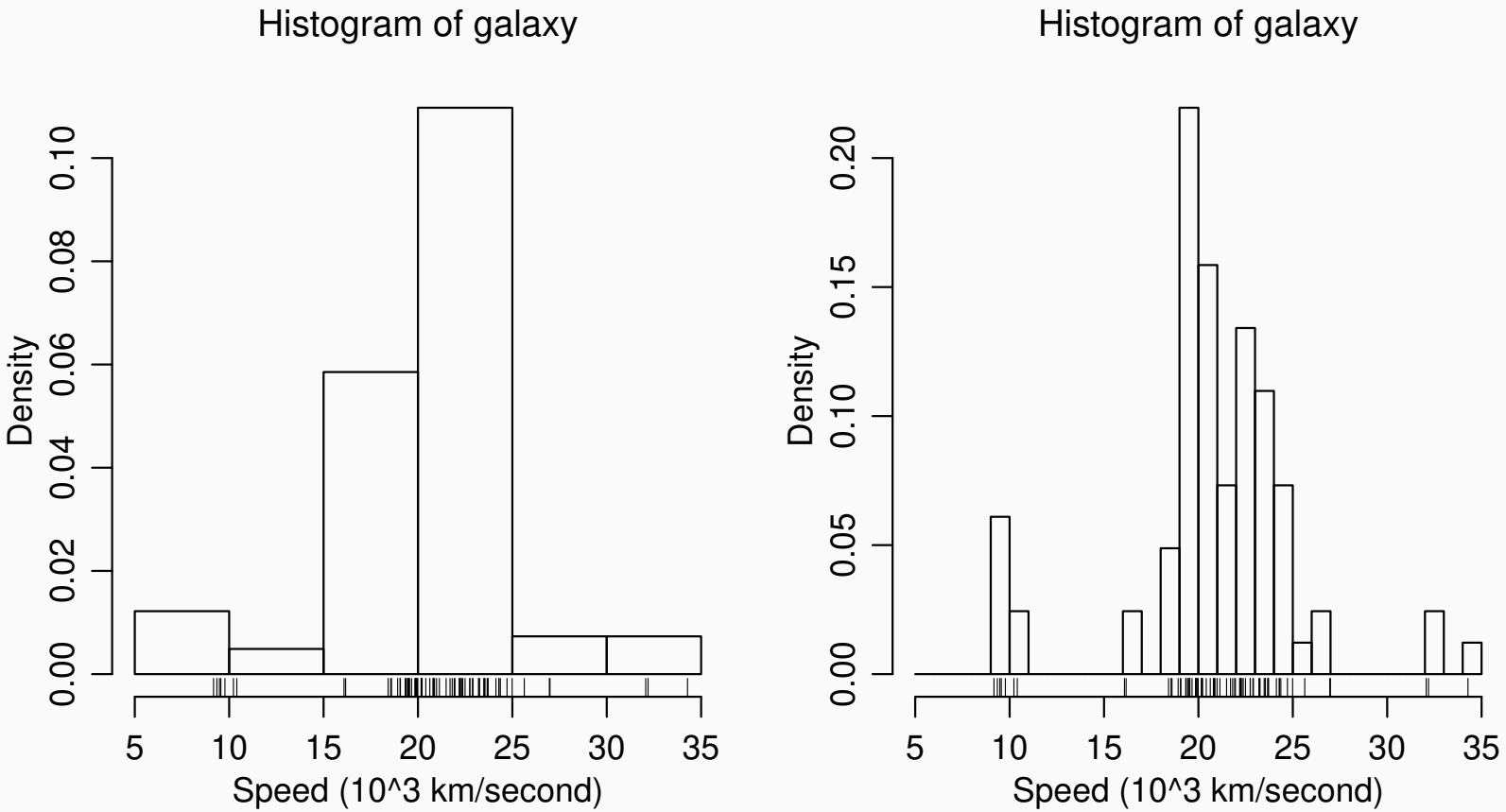
9.172	9.350	9.483	9.558	9.775	10.227	10.406	16.084	16.170	18.419
18.552	18.600	18.927	19.052	19.070	19.330	19.343	19.349	19.440	19.473
19.529	19.541	19.547	19.663	19.846	19.856	19.863	19.914	19.918	19.973
19.989	20.166	20.175	20.179	20.196	20.215	20.221	20.415	20.629	20.795
20.821	20.846	20.875	20.986	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

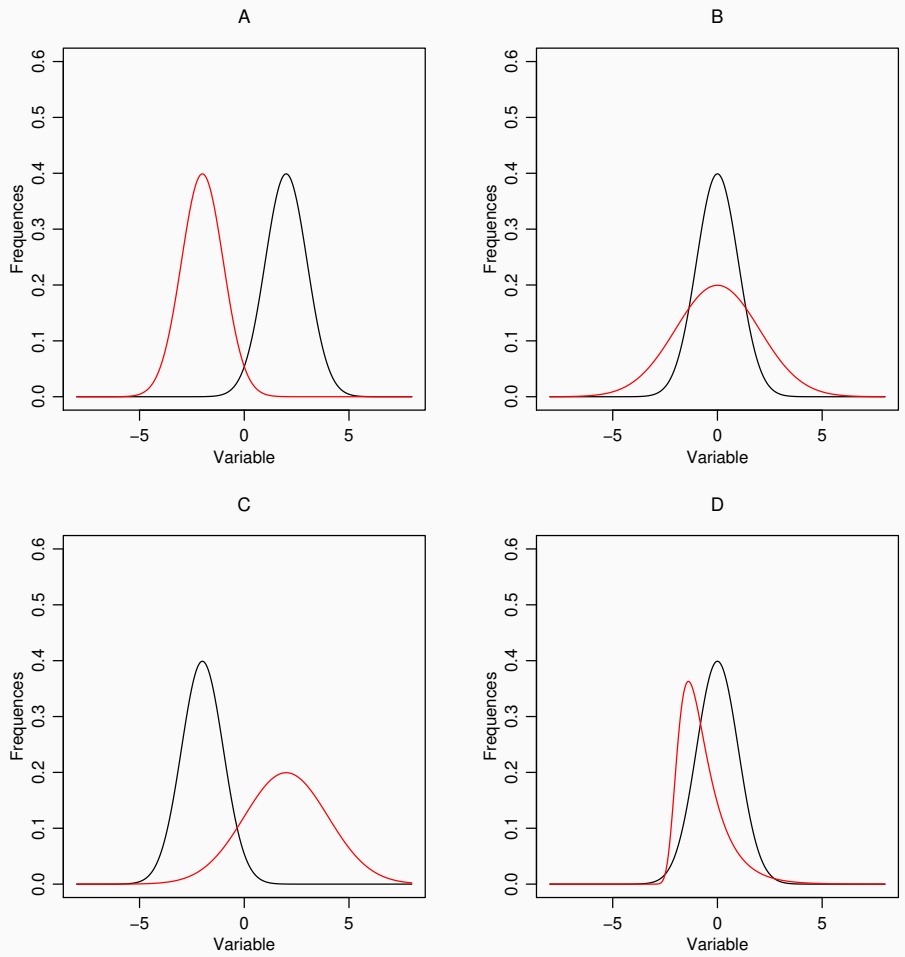
1.3 Synthèses numériques

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

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

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

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)**,

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

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									

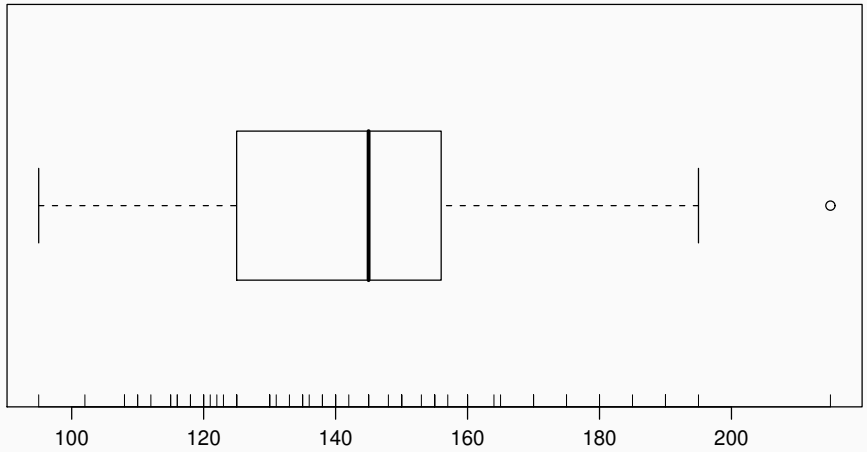
- Le **"five-number summary"** est la liste des cinq valeurs

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

  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)



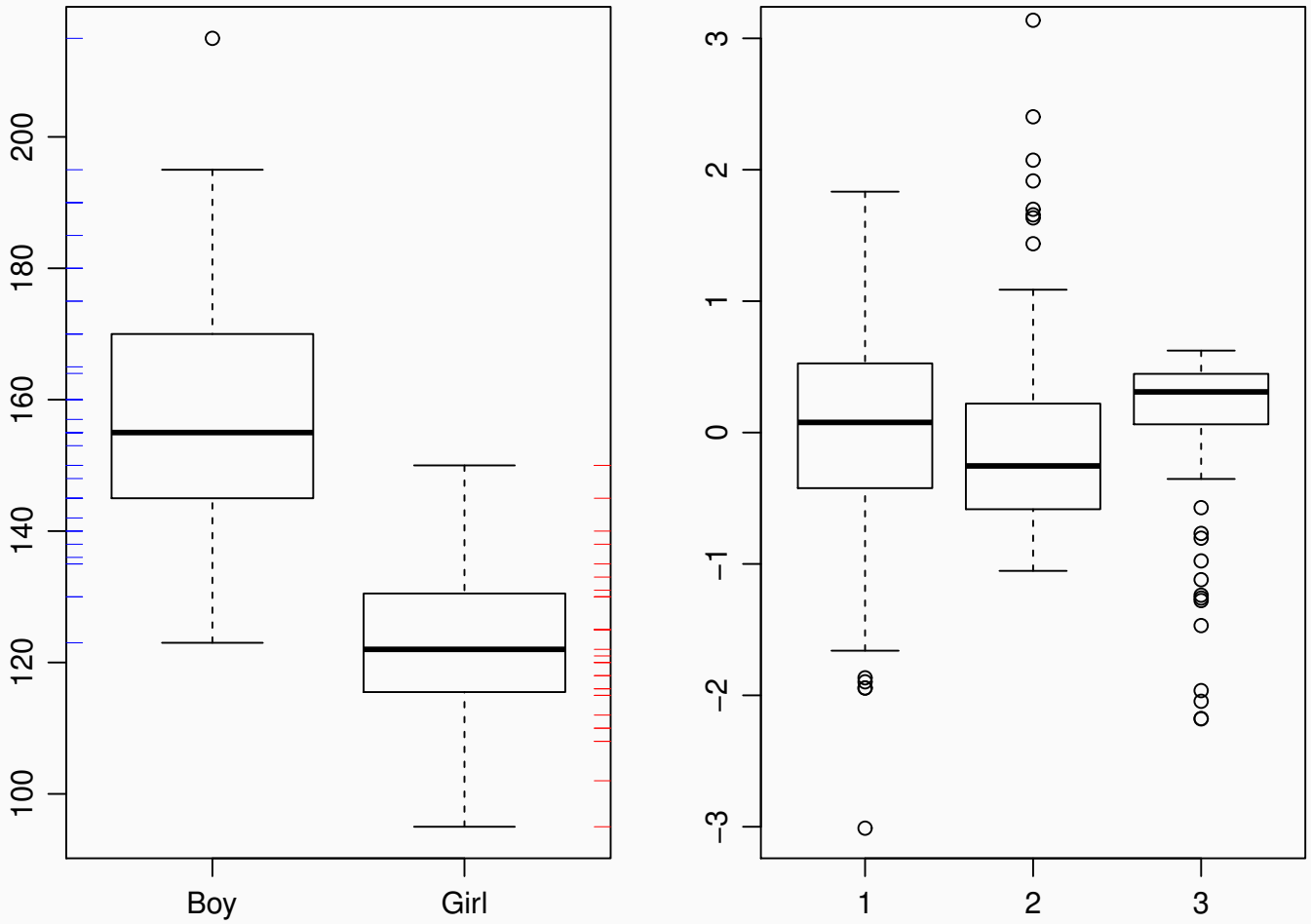
- Pour les poids, le "five-number summary" est $95, 125, 145, 156, 215$, et donc

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

- 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

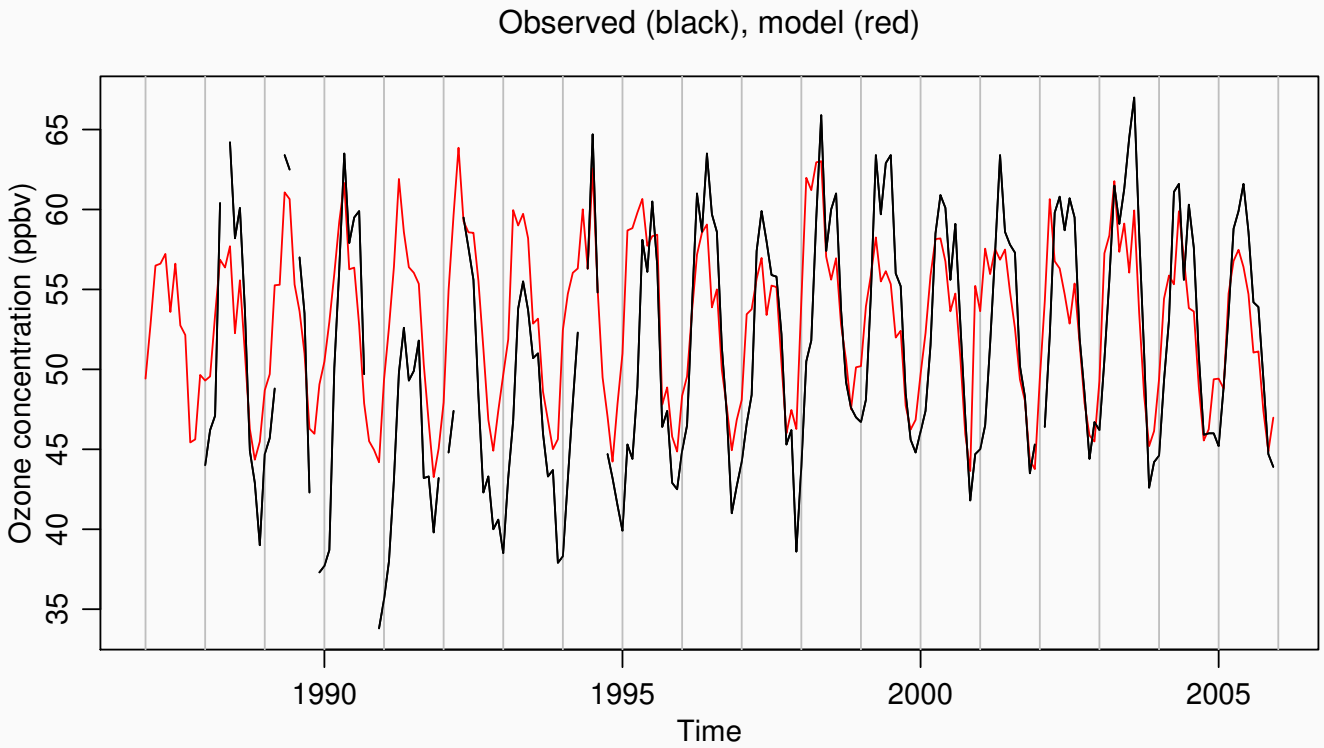
Boxplot (boîte à moustache)

- 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 :



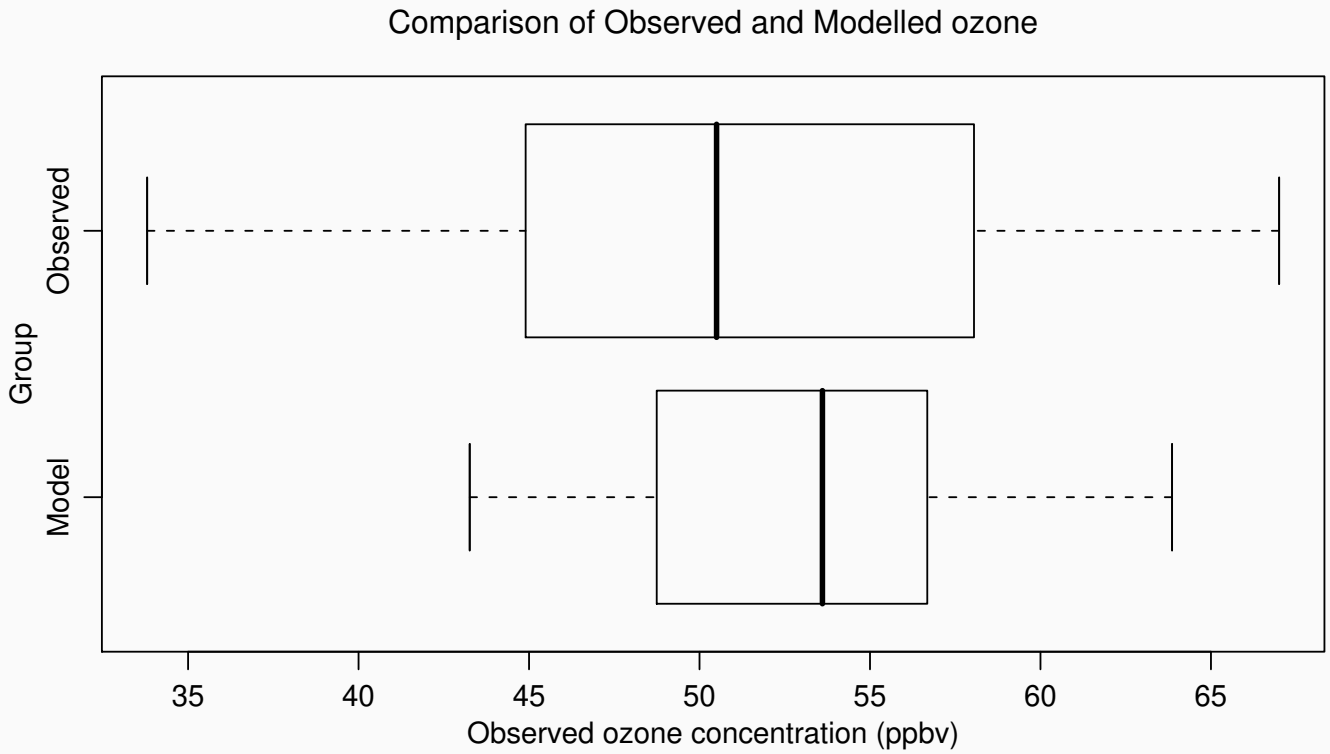
Ozone atmosphérique

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



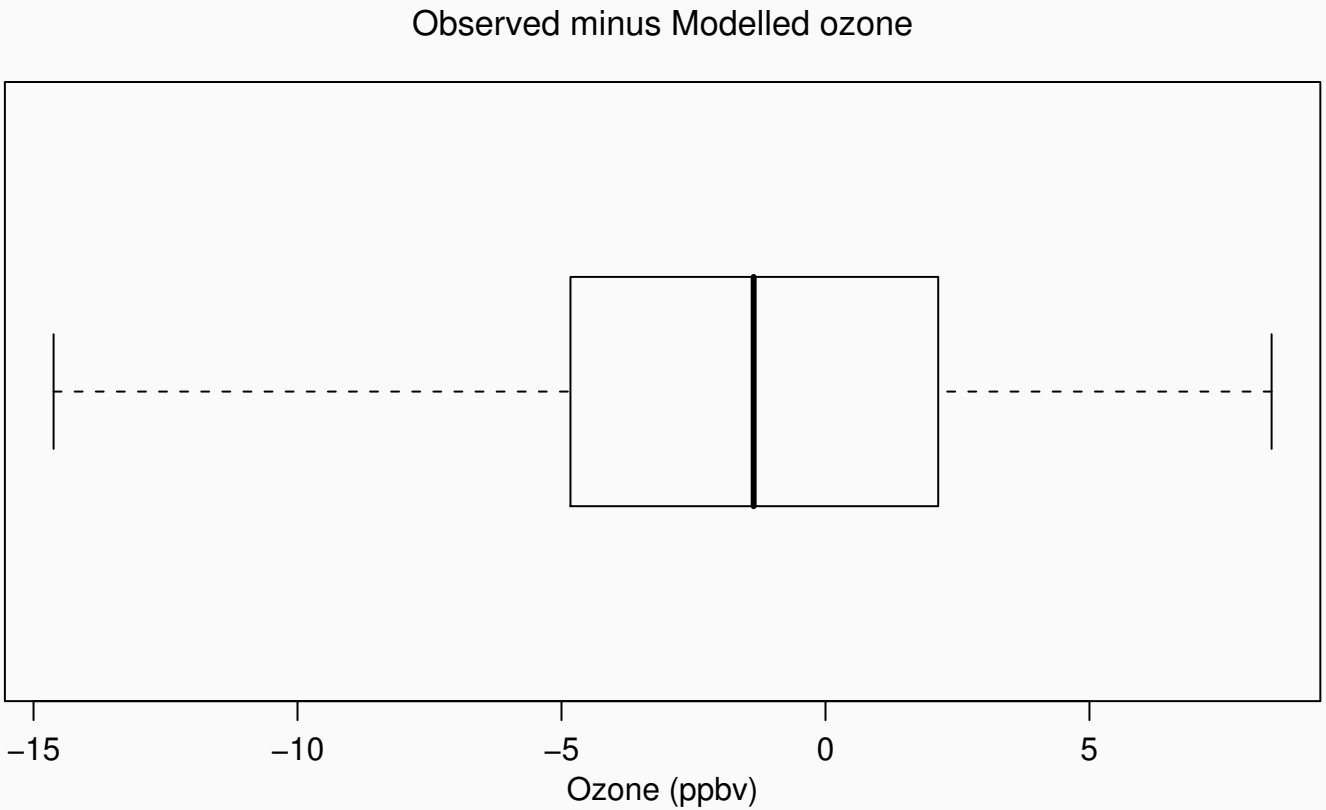
Est-ce que la modélisation est bonne ?

Ozone atmosphérique



Boxplot des données réeles et celles issues du modèle

Ozone atmosphérique



Différences des données réeles et celles issues du modèle

Commentaires

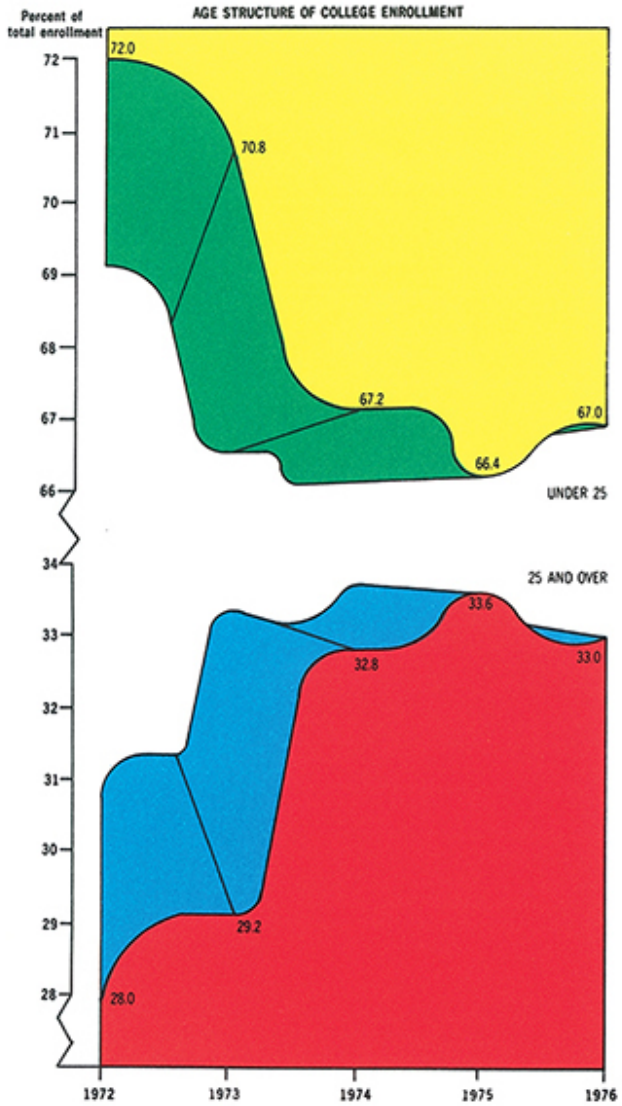
Il n'est pas toujours facile de créer de bons graphiques.

Quelques conseils :

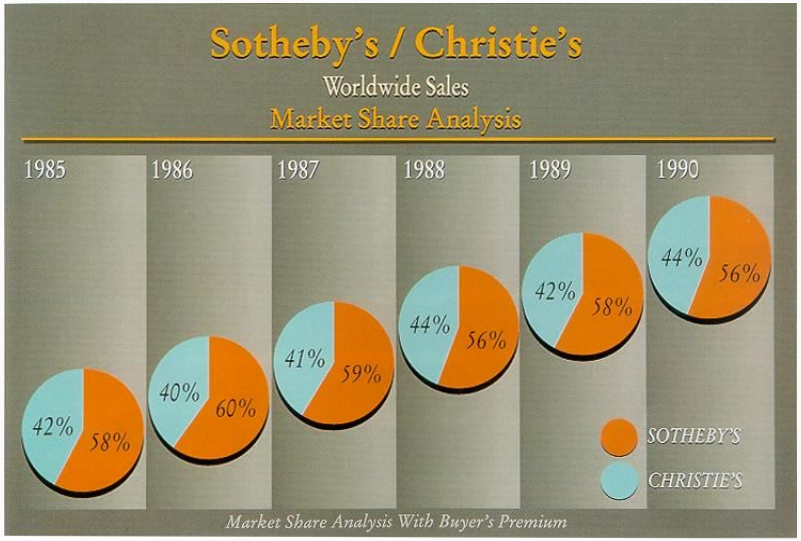
- essayer autant que possible de montrer les données elles-mêmes—pas de **fioritures/chart-junk** (couleurs/lignes/... inutiles etc.)
- mettre des unités et explications claires pour les axes et la légende
- pour comparer des quantités liées, utiliser les mêmes axes et mettre les graphiques en relation 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

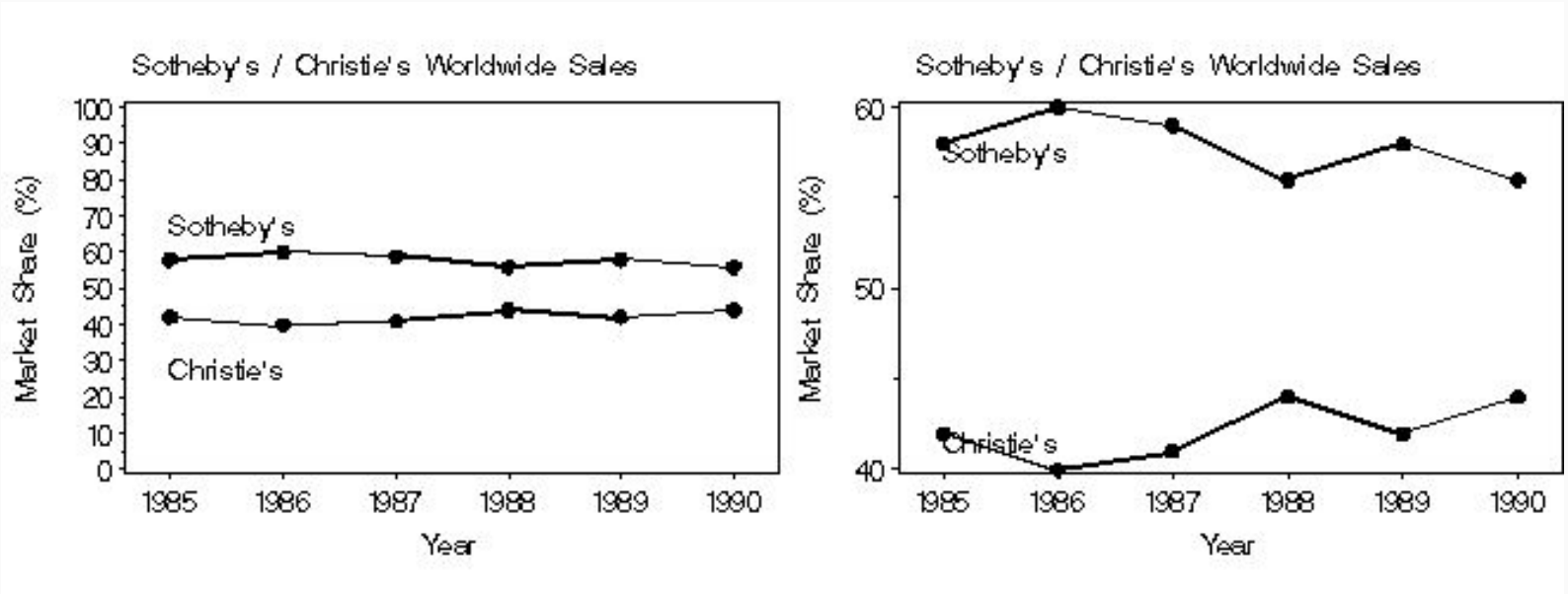
Chartjunk

Ce graphique montre 5 chiffres !



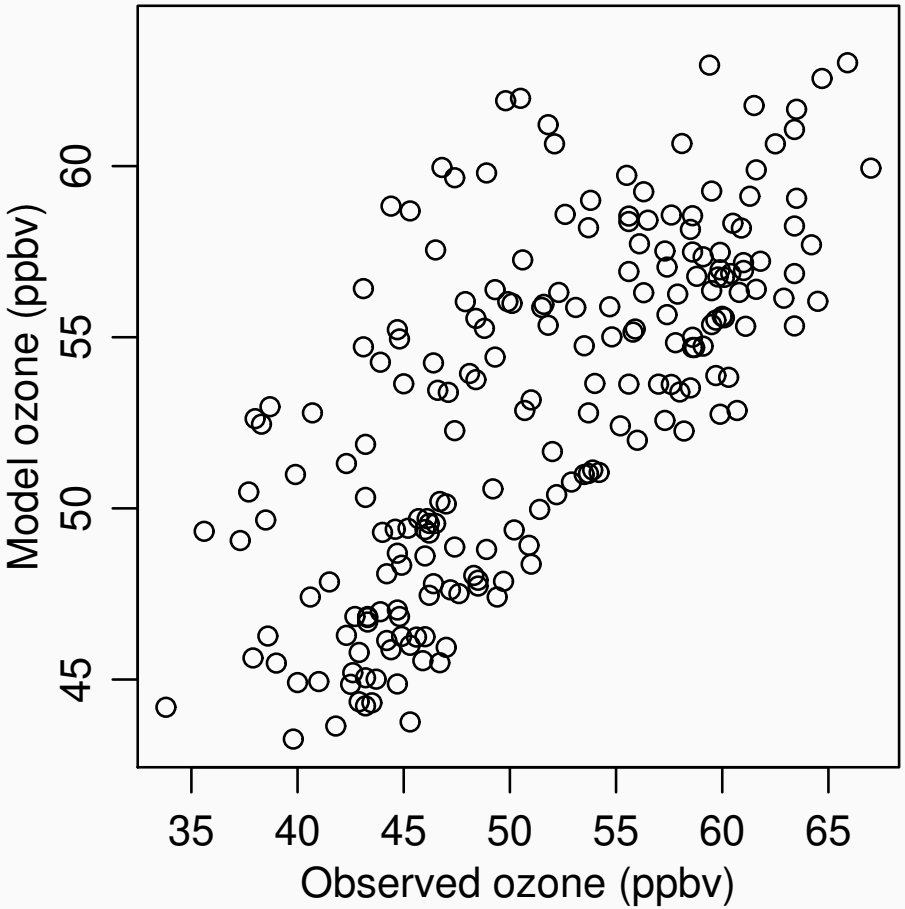
Chartjunk et échelle

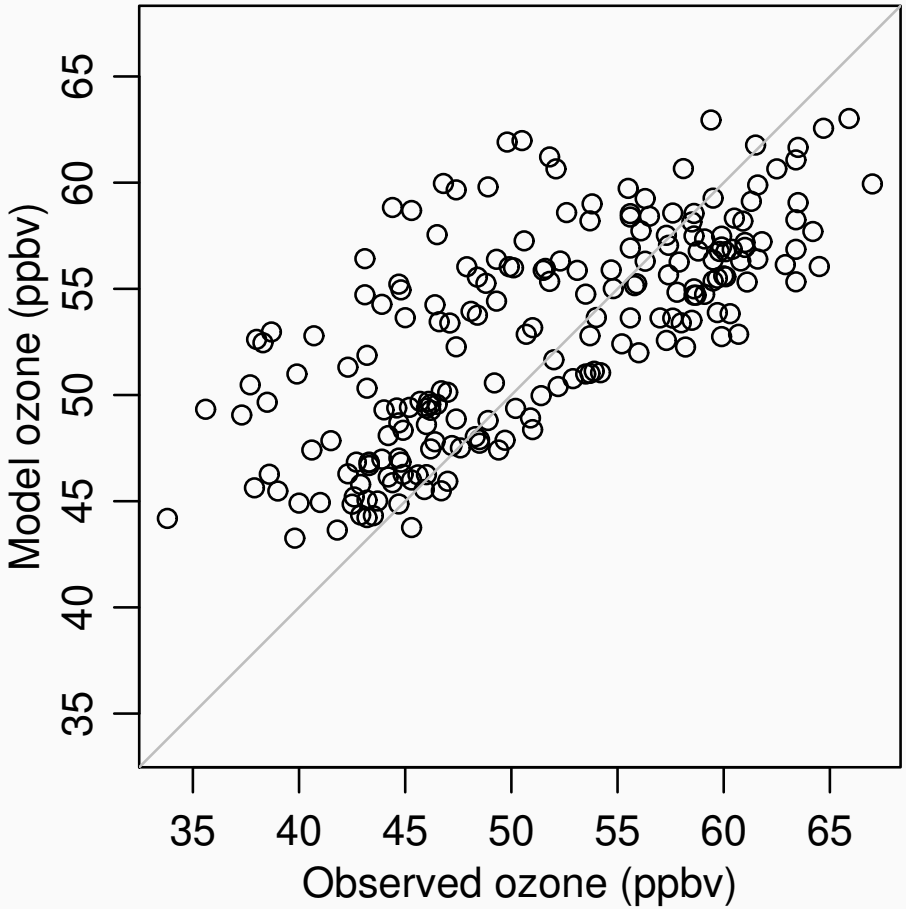




Choisir les bons axes

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

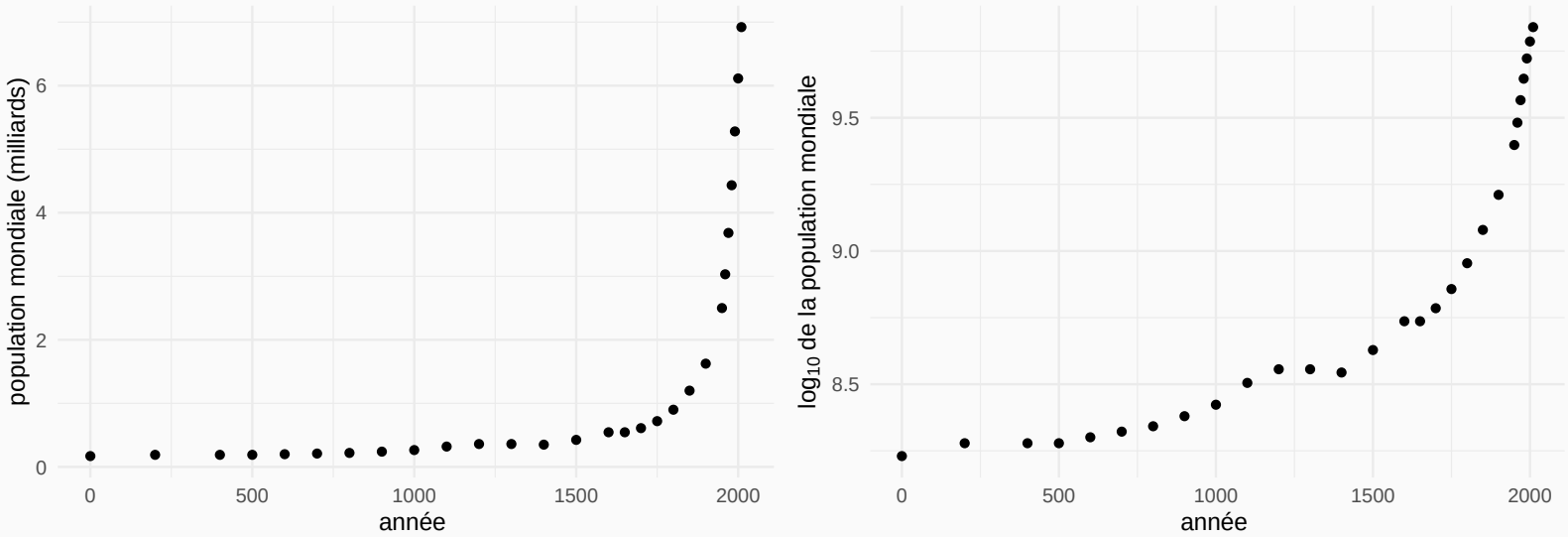




Changements d'échelles

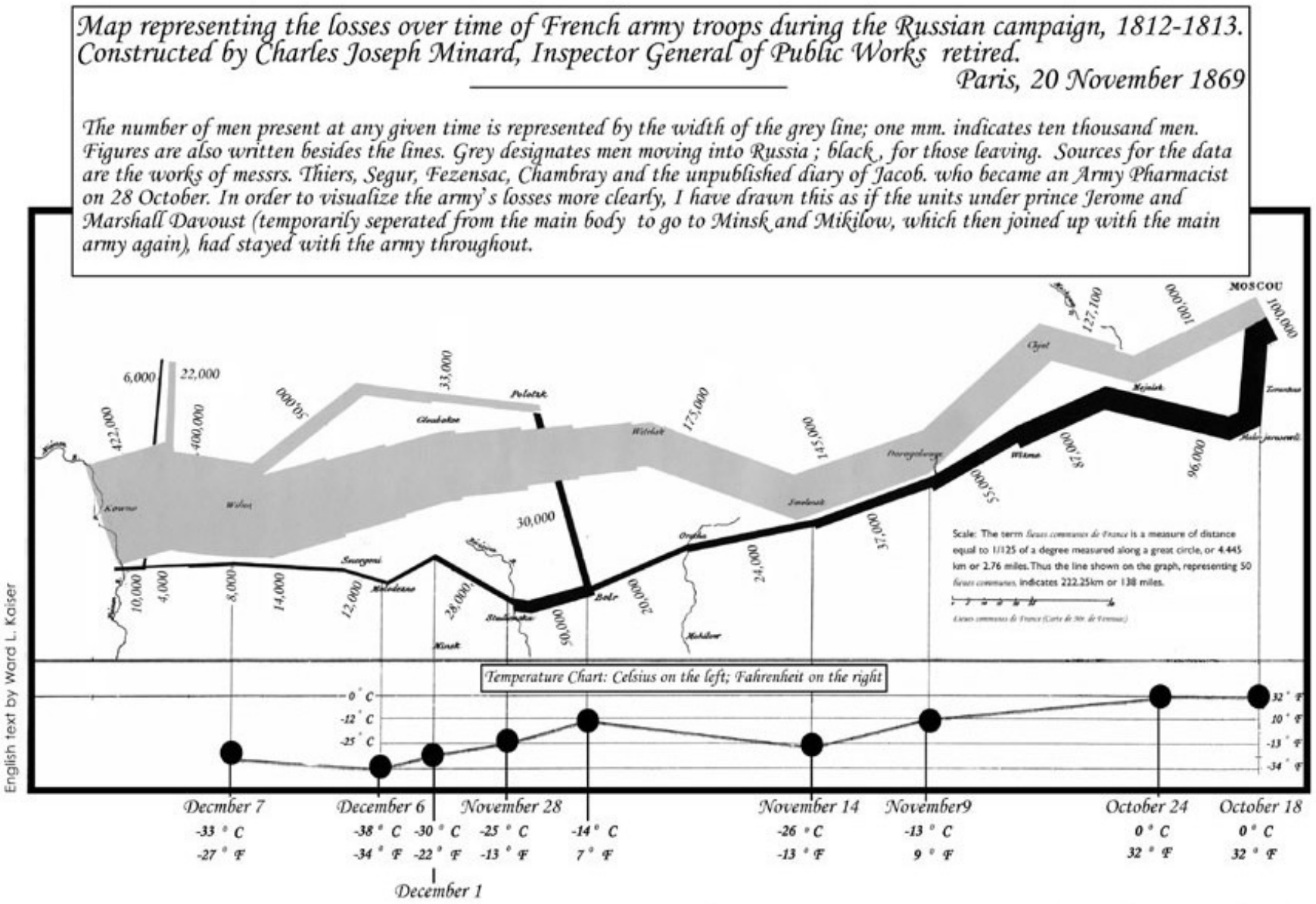
Pour certaines données, il est intéressant de les **transformer** avant de les représenter

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

Mesures de corrélation

- On veut souvent mesurer la dépendance de paires de données $(x_1, y_1), \ldots, (x_n, y_n)$ pour n individus (par exemple, $y =$ note lors d'un test, $x =$ quantité de bière consommée le soir avant)
- Souvent on utilise la **coefficient de la corrélation (empirique) (correlation coefficient)**,

$$r_{xy} = \frac{n^{-1}\sum_{j=1}^{n}(x_j - \bar{x})(y_j - \bar{y})}{\left\{n^{-1}\sum_{j=1}^{n}(x_j - \bar{x})^2 \times n^{-1}\sum_{j=1}^{n}(y_j - \bar{y})^2\right\}^{1/2}},$$

qui satisfait

(a) $-1 \leq r_{xy} \leq 1$;

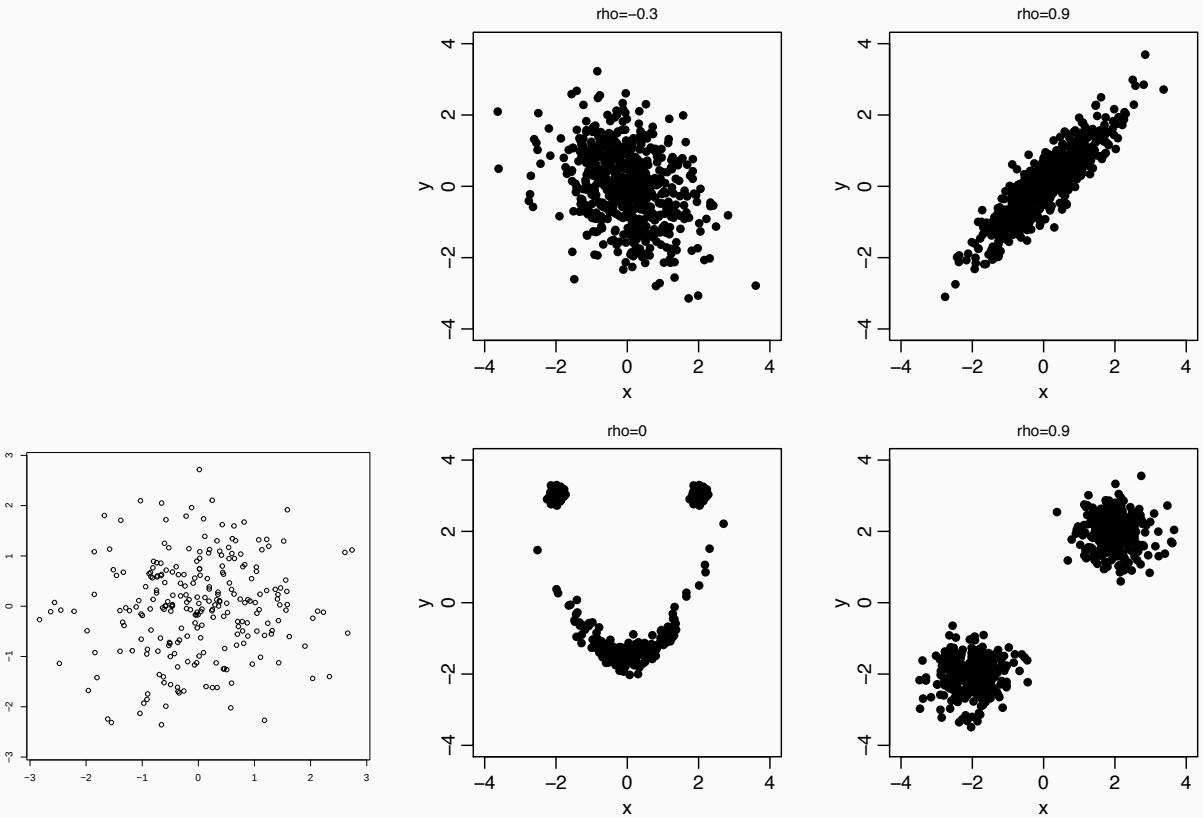
(b) si $r_{xy} = \pm 1$, alors les (x_j, y_j) sur une droite, de pente positive si $r_{xy} = 1$, et de pente négative si $r_{xy} = -1$

(c) si $r_{xy} = 0$ il n'y a **pas de dépendance LINÉAIRE** !

(d) si $(x_j, y_j) \mapsto (a + bx_j, c + dy_j)$ (avec $bd \neq 0$), alors $r_{xy} \mapsto \operatorname{sign}(bd) r_{xy}$

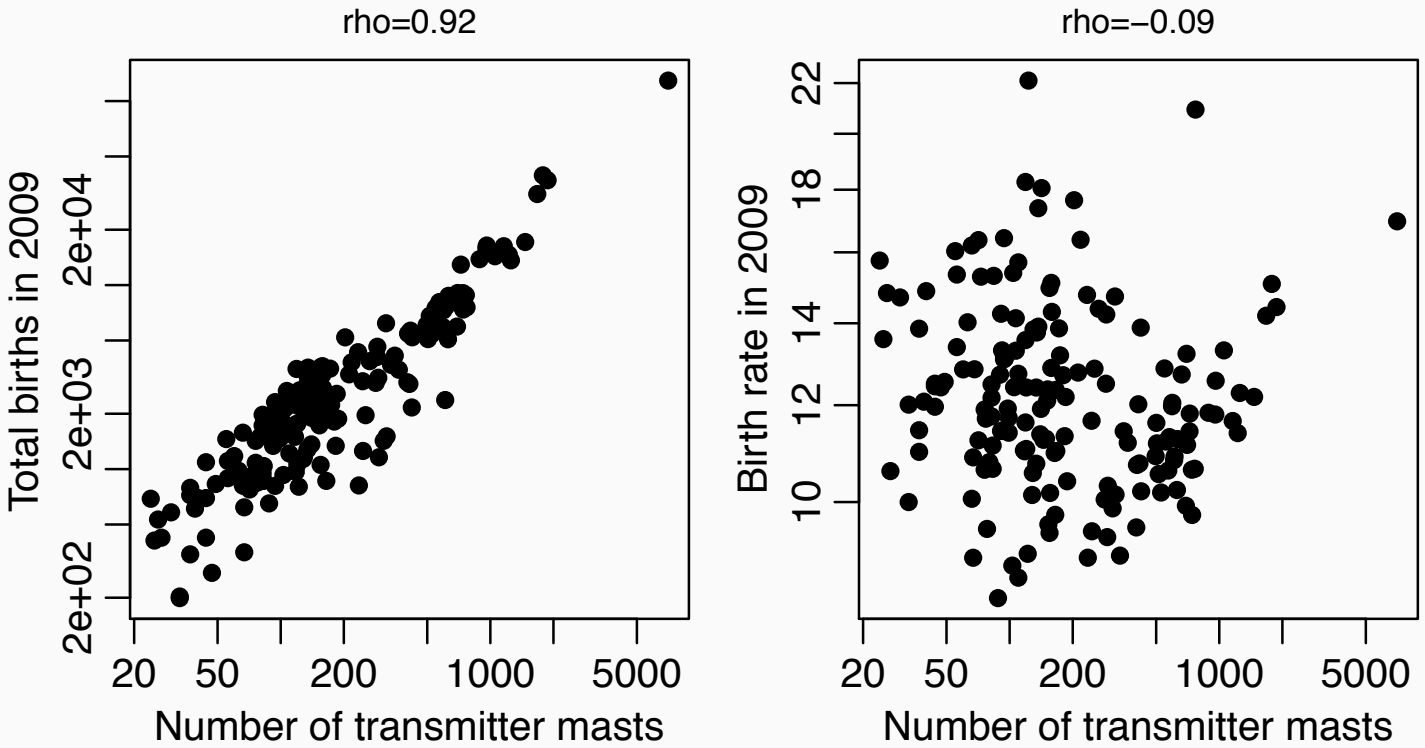
Limitations de la corrélation

- Une corrélation entre deux variables n'implique pas une causalité entre elles
- r_{xy} mesure la dépendance linéaire (panneaux supérieurs)
- On peut avoir $r_{xy} \approx 0$, mais dépendance forte mais non-linéaire (en bas au milieu)
- Une corrélation pourrait être forte mais **specieuse**, comme en bas à droite, ou deux sous-groupes, chacun sans corrélation, sont combinés



Corrélation $\neq$ causalité

Deux variables peuvent être très corrélées sans lien de causalité. Le graphique à gauche ici montre une corrélation forte entre le nombre de naissances et les mâts de communication dans les villes anglaises . . .



1.5 Stratégie

Analyse initiale des données

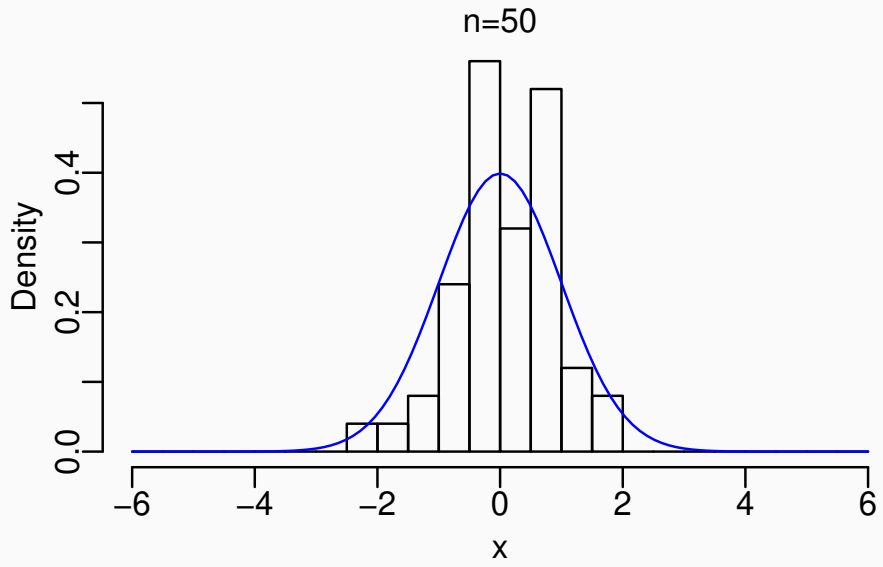
On a maintenant une **stratégie** pour explorer des données issues d'une variable quantitative :

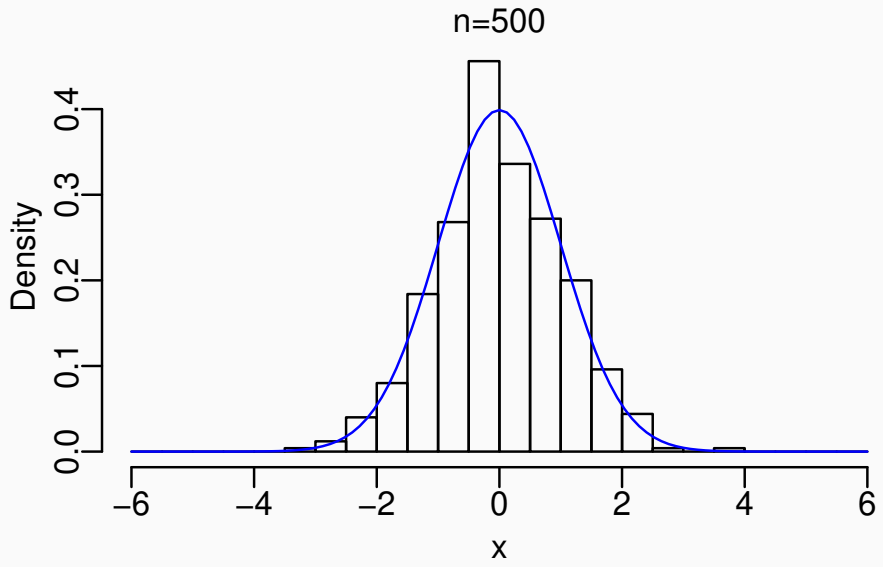
1) toujours faire des **représentations graphiques** d'abord
2) étudier la **structure globale** des données et identifier d'éventuelles valeurs atypiques / aberrantes ("outliers")—trouver pourquoi elles apparaissent
3) calculer des **synthèses numériques** pour décrire la tendance centrale (position / 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

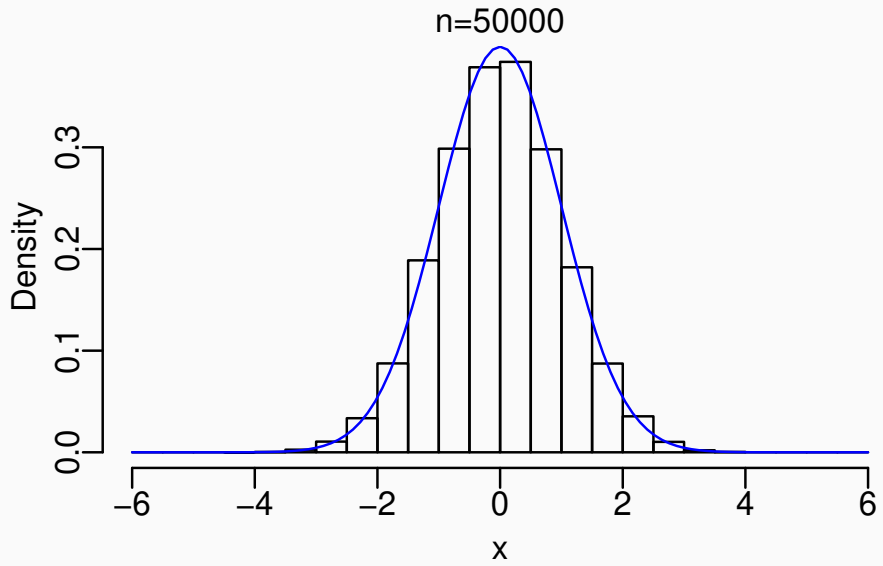
Modélisation des données

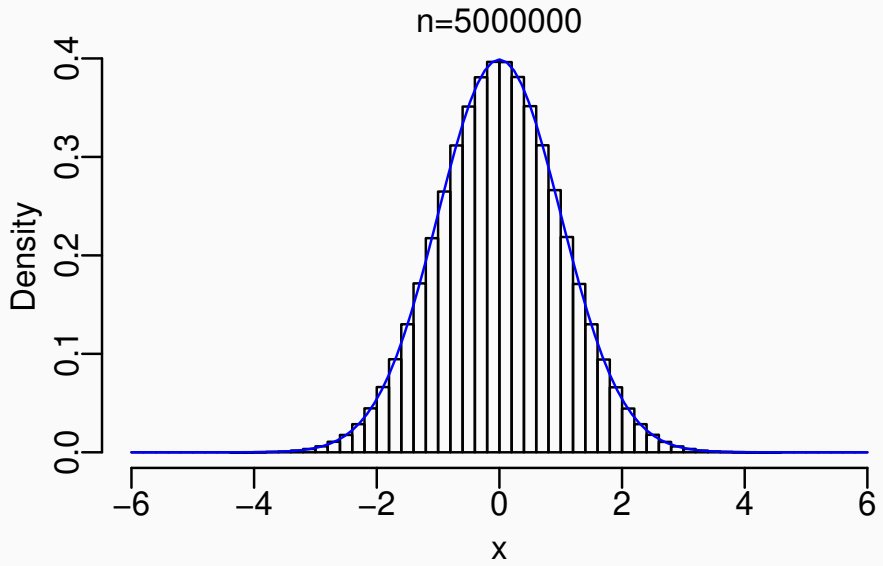
- Souvent on suppose que les données sont issues d'un échantillon aléatoire tiré d'une population d'intérêt
- Cette population est considérée comme très grande, d'une taille presque infinie
- En statistique ces modèles mathématiques sont souvent des **courbes de densité**, une fonction qui est toujours ≥ 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

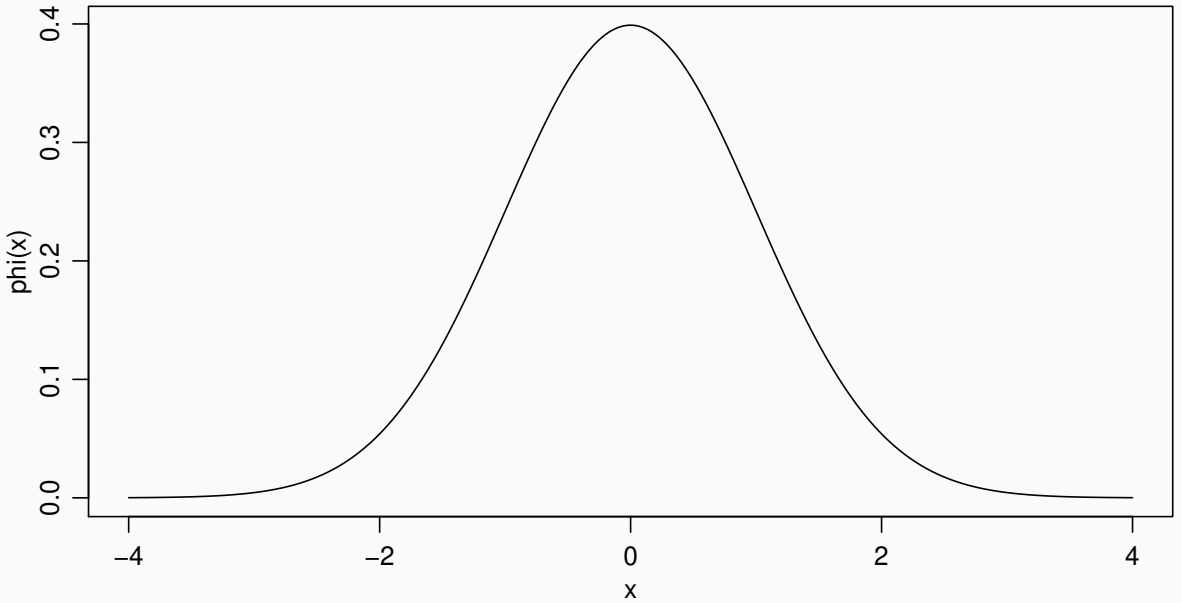
Distribution normale

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

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

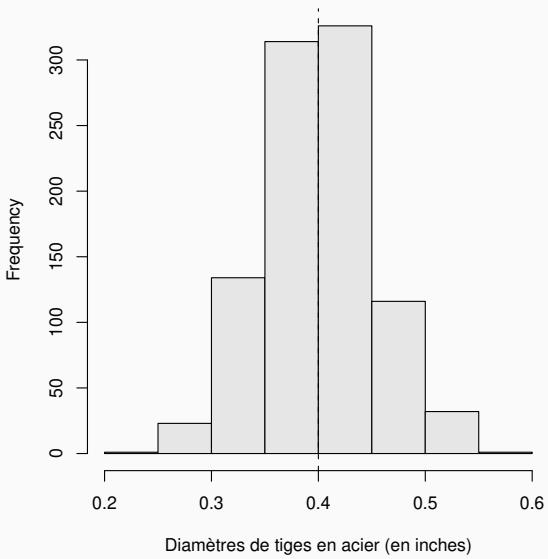
où μ est la **moyenne** et σ est l'**écart-type**

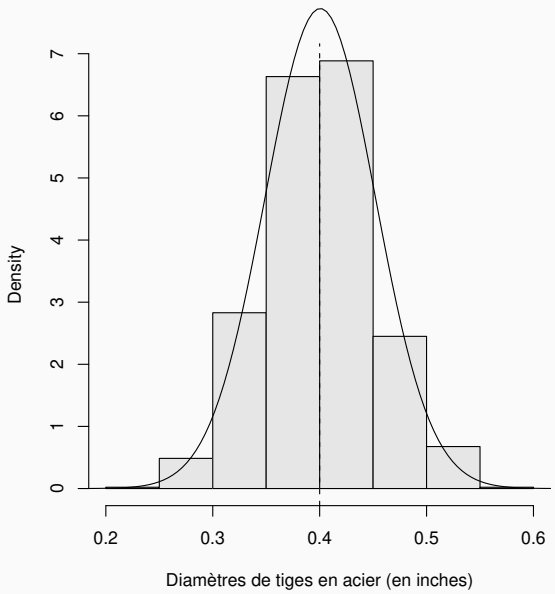
$f_{\mu,\sigma}(x)$ est la hauteur de la courbe au point x



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

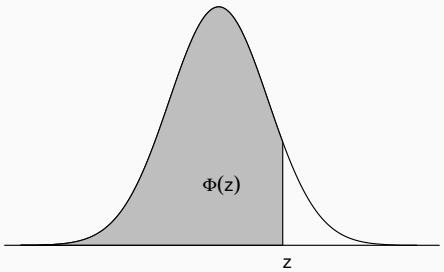
$$\mathcal{N}(\mu, \sigma^2) \mapsto \mathcal{N}(0,1)$$

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

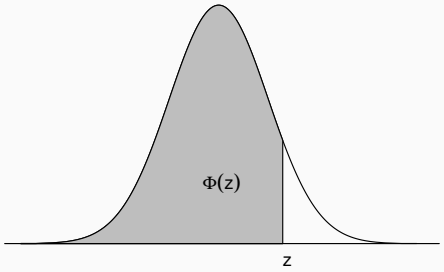
$$\Phi(z) = \int_{-\infty}^{z} \phi(x)\, dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-x^2/2}\, dx, \quad z \in \mathbb{R}$$



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

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

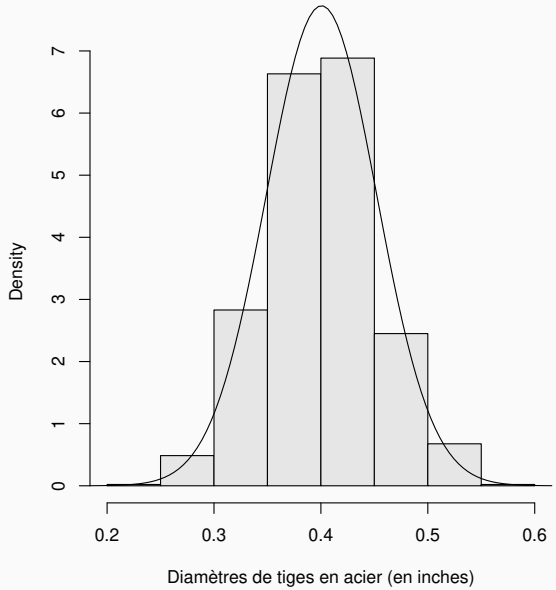
c'est à dire $\sim 95\%$ et $\sim 99.7\%$ de l'échantillon des tiges, respectivement.

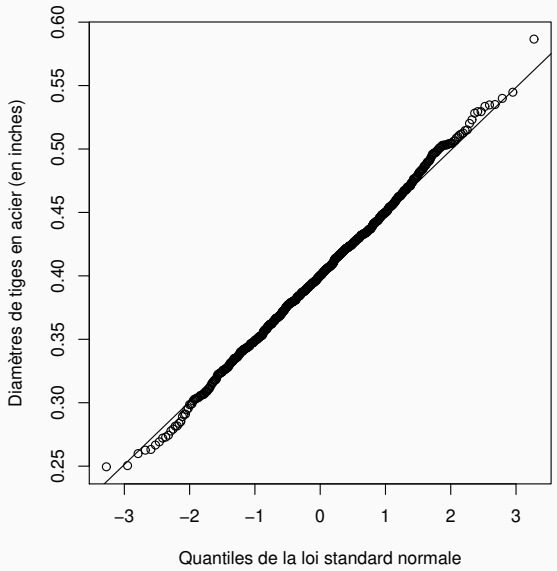
Q-Q plot normale

- L'histogramme ou le boxplot nous donnent des indices sur des propriétés d'une distribution normale, dont : pas de valeurs atypiques, symétrie, unimodalité
- Mais il faut le savoir plus précisément : **le meilleur outil pour "vérifier" la normalité graphiquement est le "Q-Q plot normal"**
- Si les points sur ce dernier sont proches d'une droite, cela signifie que les observations pourront être modélisées par un modèle normal
- Les valeurs aberrantes apparaissent comme des points isolés
- La pente et l'intercepte pour $x = 0$ donnent des estimations de σ et μ respectivement

Q-Q plot—tiges

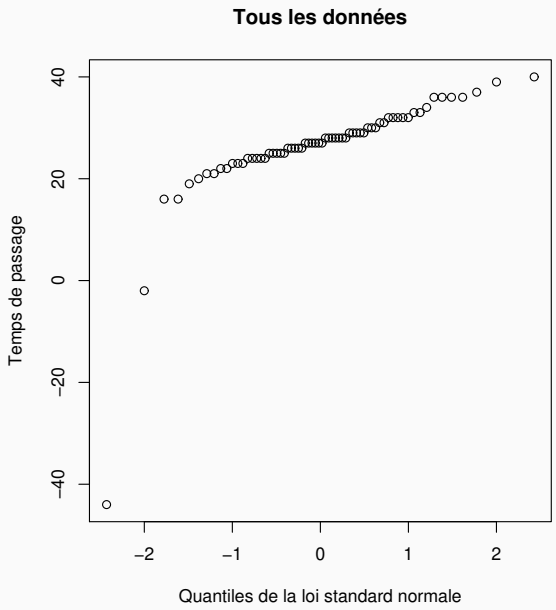
Histogramme et Q-Q plot normale du diamètre de 947 tiges en acier (en inches)

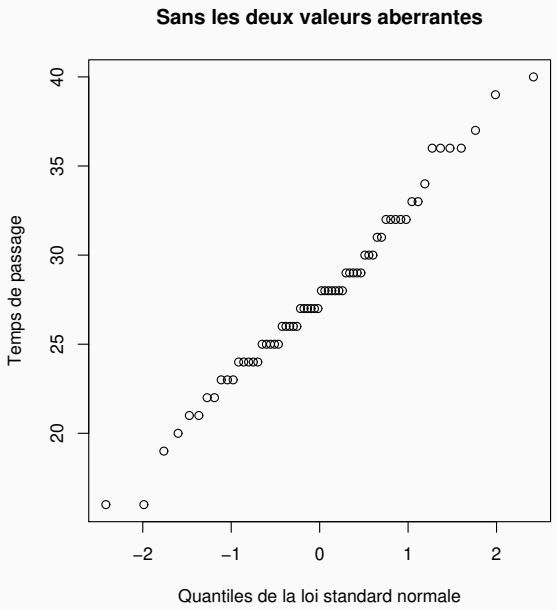




Q-Q plot—Newcomb

Q-Q plot normaux de 66 temps de passage de la lumière, mésurés par Simon Newcomb, pour traverser une distance connue :





2. Probabilité

Expériences aléatoires

La théorie des probabilités permet de décrire et modéliser les **phénomènes aléatoires**.

Les actions qui mènent à des résultats aléatoires sont appellées des **expériences aléatoires**. Plus précisément, une expérience est dite aléatoire s'il est impossible de prévoir son résultat. En principe, on admet qu'une expérience aléatoire peut être répétée (indéfiniment) dans des conditions identiques ; son résultat peut donc varier d'une réalisation à l'autre.

Exemples :

- lancer d'un dé ou d'une pièce de monnaie ;
- tirage d'une carte.

2.1. Probabilité d'événements

Modèles probabilistes d'une expérience aléatoire

- **Ensemble fondamental** Ω : tous les résultats possibles
- **Événement élémentaire** $\omega \in \Omega$: un résultat possible.
- **Événement** : un sous-ensemble (raisonnable) $A \subseteq \Omega$. Un événement peut réunir plusieurs événements élémentaires.
- On dit qu'un événement est **réalisé** si le résultat de l'expérience aléatoire (événement élémentaire) appartient à cet événement.

Exemple Lancer d'une pièce de monnaie :

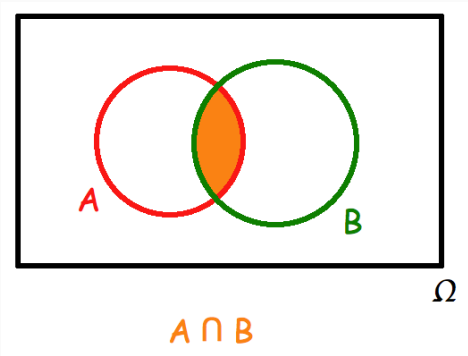
$$\begin{aligned} \Omega &= \{P, F\}. \\ A &= \{P\} = \text{“Pile” est un événement (élémentaire)} \end{aligned}$$

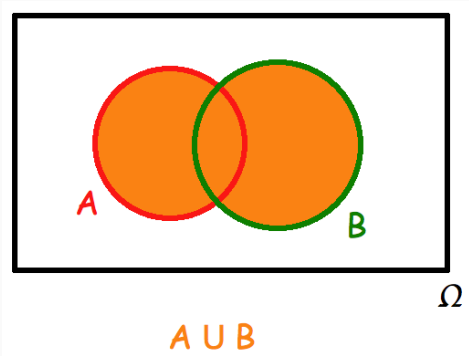
Exemple Lancer d'un dé :

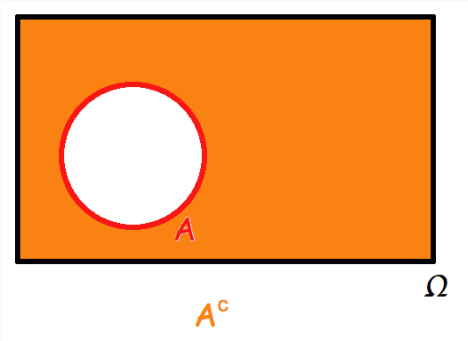
$$\begin{aligned} \Omega &= \{1, 2, 3, 4, 5, 6\}. \\ A &= \text{“obtenir 1”} = \{1\} \text{ est un événement (élémentaire).} \\ B &= \text{“obtenir un chiffre pair”} = \{2, 4, 6\} \text{ est un événement (composé).} \end{aligned}$$

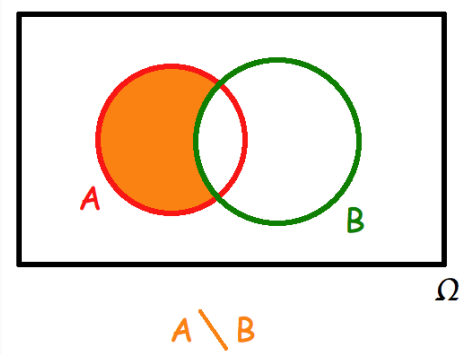
Diagramme de Venn et opérations entre événements

- $A \cup B = B \cup A$ union
- $A \cap B = B \cap A$ intersection
- A^c complémentaire
- $A \setminus B = A \cap B^c$ différence ; $A \setminus B \neq B \setminus A$
- $\emptyset$ ensemble vide
- $A = \{2, 4, 6\}$ (pair)
- $B = \{2, 3, 5\}$ (premier)









Fonction de probabilité

Définition: Les événements A et B sont **disjoints** si $A \cap B = \emptyset$.

Événements $A_1, A_2, \ldots, A_n$ sont disjoints si $A_i \cap A_j = \emptyset$ quand $i \neq j$.

Définition: Une fonction de probabilité, notée ici $\Pr$, est une fonction telle que

- $0 \leq \Pr(A) \leq 1$ pour tout événement A ;
- $\Pr(\Omega) = 1$, (événement certain) ;
- Si $A_1, \ldots, A_n$ est une collection disjointe d'événements, alors

$$\Pr\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} \Pr(A_i)$$

  De même pour une collection infinie dénombrable $A_1, A_2, \ldots$

Propriétés d'une fonction de probabilité

- $\Pr(\emptyset) = 0$, (événement impossible) ;
- $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$;
- $\Pr(A^c) = 1 - \Pr(A)$, (événement complémentaire de A) ;
- $A \subseteq B \Rightarrow \Pr(A) \leq \Pr(B)$.

Exemple Deux lancers d'une pièce de monnaie :

$$\Omega = \{PP, PF, FP, FF\}.$$

(a) Expliciter les événements $A =$"au moins un P", $B =$"au moins un F", $A \cap B$, et $A \cup B$.
(b) Trouver les probabilités correspondantes si

$$\Pr(\{PP\}) = \cdots = \Pr(\{FF\}) = 1/4.$$

Solution (diapositive 76)

Evénements élémentaires équiprobables

Sous l'hypothèse **d'équiprobabilité des événements élémentaires**, pour tout événement A de Ω,

$$\begin{aligned}\Pr(A) &= \frac{\text{nombre d'événements élémentaires dans } A}{\text{nombre total d'événements élémentaires dans } \Omega} \\ &= \frac{\text{nombre de cas favorables à } A}{\text{nombre total de cas possibles}}.\end{aligned}$$

Exemple Lancer d'un dé. **Supposons** que les six faces ont les mêmes chances d'apparaître (événements élémentaires équiprobables). Alors

$$\Pr(\{1\}) = \Pr(\{2\}) = \cdots = \Pr(\{6\}) = \frac{1}{6},$$

et

$$\begin{aligned}\Pr(\text{“obtenir un nombre pair”}) = \Pr(\{2,4,6\}) &= \Pr(\{2\}) + \Pr(\{4\}) + \Pr(\{6\}) \\ &= \frac{3}{6} = \frac{1}{2}.\end{aligned}$$

Exemple Lancers de deux dés. Trouver $\Pr(\text{“la somme des faces vaut 7”})$.

Solution (diapositive 78)

Probabilité conditionnelle et indépendance

La probabilité que l'événement A se réalise peut être influencée par la réalisation d'un autre événement B. Pour formaliser cette idée, on introduit les concepts de probabilité conditionnelle et d'indépendance :

Définition: La **probabilité conditionnelle** de A sachant que B s'est réalisé est définie par

$$\Pr(A \mid B) = \frac{\Pr(A \cap B)}{\Pr(B)}, \quad \text{si} \quad \Pr(B) > 0.$$

Définition: Deux événements A et B sont dits **indépendants** si $\Pr(A \cap B) = \Pr(A) \times \Pr(B)$.

Intuition : si $\Pr(B) > 0$, c'est équivalent à

$$\Pr(A \mid B) = \Pr(A).$$

Exemples

Exemple Deux lancers d'une pièce de monnaie. Trouver la probabilité d'obtenir pile au 2ème lancer sachant qu'on a obtenu pile au 1er lancer.

Exemple Lancer d'un dé Les événements $A = \{2, 4\}$ et $B = \{2, 4, 6\}$ sont-ils indépendants ?

Ne pas confondre indépendance et incompatibilité (A et B disjoints) !

Soient A, B disjoints tels que $\Pr(A), \Pr(B) > 0$. On a

$$\Pr(A \cap B) = \Pr(\emptyset) = 0, \quad \text{mais} \quad \Pr(A) \times \Pr(B) \neq 0,$$

donc A et B sont dépendants. Donc

$A \cap B = \emptyset \Rightarrow A$ et B dépendants, et ainsi, A et B indépendants $\Rightarrow A \cap B \neq \emptyset$.

Par ailleurs

$$A \cap B \neq \emptyset \not\Rightarrow A \text{ et } B \text{ indépendants.}$$

Solution diapositive 81

Solution diapositive 81

Indépendance : généralisation

Définition: Les événements $A_1, \ldots, A_n$ sont **indépendants** si, pour tout sous-ensemble d'indices $\{i_1, \ldots, i_k\} \subseteq \{1, \ldots, n\}$, on a

$$\Pr\left(\bigcap_{j=1}^{k} A_{i_j}\right) = \prod_{j=1}^{k} \Pr(A_{i_j}).$$

Exemple Un système de n composants est appelé **système en parallèle** s'il fonctionne dès qu'au moins un de ses composants fonctionne. Un **système en série** fonctionne si et seulement si tous ses composants fonctionnent.

(a) Si le $i$ème composant fonctionne indépendamment de tous les autres et avec une probabilité p_i, $i = 1, \ldots, n$, quelle est la probabilité de fonctionnement d'un système en parallèle ?

(b) Même question pour un **système en série**.

(c) Même question pour un **système composé**.

Solution diapositive 84

Formule des probabilités totales

Définition: Soit A un événement quelconque de Ω, et $\{B_i\}_{i=1,\ldots,n}$ une **partition** de Ω, c'est-à-dire,

$$B_i \cap B_j = \emptyset, \quad i \neq j, \qquad \bigcup_{i=1}^{n} B_i = \Omega.$$

La **formule des probabilités totales**

$$\Pr(A) \quad = \quad \sum_{i=1}^{n} \Pr(A \cap B_i) = \sum_{i=1}^{n} \Pr(A \mid B_i)\ \Pr(B_i).$$

Elle est également valide pour une partition infinie dénombrable.

Exemple Trois machines M_1, M_2 et M_3 fabriquent des pièces dans les proportions respectives 25%, 35% et 40%. On sait que respectivement 5%, 4% et 2% des pièces produites par M_1, M_2 et M_3 sont défectueuses. On choisit une pièce aléatoirement. Calculer

$$\Pr(\text{“la pièce est défectueuse”}).$$

Formule des probabilités totales : diagramme de Venn

Solution diapositive 86

Définissons les événements : $D =$ “la pièce est défectueuse” et pour $i = 1, 2, 3$, $A_i =$ “la pièce a été fabriquée par M_i”.

Théorème de Bayes

Théorème de Bayes Soient $A \subseteq \Omega$ et $\{B_i\}_{i=1,\ldots,n}$ une partition (éventuellement infinie dénombrable) de Ω. Si $\Pr(A) > 0$ alors on a, pour tout $i = 1, \ldots, n$,

$$\Pr(B_i \mid A) = \frac{\Pr(B_i \cap A)}{\Pr(A)} = \frac{\Pr(A \mid B_i)\Pr(B_i)}{\sum_{j=1}^{n} \Pr(A \mid B_j)\Pr(B_j)}.$$

- La formule de Bayes est très simple mais très utile, car elle permet une 'inversion du point de vue' dont on a souvent besoin en pratique.

Exemple Pour dépister une maladie, on applique un test. Si la maladie est présente, le test le découvre avec probabilité 0.99. Si la personne est saine, le test le trouve malade avec probabilité 0.02. Sachant qu'en moyenne un patient sur 1000 est atteint de la maladie, calculer la probabilité qu'un patient soit atteint sachant que son test a été positif. Comment améliorer ce resultat ?

Solution exemple Bayes

Soit M l'événement "le patient est atteint de la maladie", M^c l'événement complémentaire, et A l'événement "le résultat du test est positif".

Types d'indépendance

Les événements $A_1, \ldots, A_n$ sont **indépendants** si pour tout ensemble fini d'indices $F \subseteq \{1, \ldots, n\}$ qui est non-vide, on a

$$\Pr\left(\bigcap_{i \in F} A_i\right) = \prod_{i \in F} \Pr(A_i).$$

Définition: Les événements $A_1, \ldots, A_n$ sont **conditionnellement indépendants sachant** B si pour tout ensemble fini d'indices $F \subseteq \{1, \ldots, n\}$ qui est non-vide, on a

$$\Pr\left(\bigcap_{i \in F} A_i \mid B\right) = \prod_{i \in F} \Pr(A_i \mid B).$$

Exemples : indépendance conditionnelle

Exemple Une année donnée, la probabilité qu'un conducteur fasse une déclaration de sinistre à son assurance est μ, indépendamment des autres années. La probabilité pour une conductrice est de $\lambda < \mu$. Un assureur a le même nombre de conducteurs que de conductrices, et sélectionne une personne au hasard.

(a) Donner la probabilité que la personne déclare un sinistre cette année

(b) Donner la probabilité que la personne déclare des sinistres durant 2 années consécutives

(c) Si la compagnie sélectionne au hasard une personne ayant fait une déclaration, quelle est la probabilité que cette personne fasse une déclaration l'année suivante ?

(d) Montrer que la connaissance qu'une déclaration de sinistre ait été faite une année augmente la probabilité de déclarer un autre l'année suivante

Solution 92

2.2 Variables aléatoires

Définition

Exemple : lancer de deux dés. On s'intéresse à la somme obtenue plutôt qu'au fait de savoir si c'est le couple $\{1,6\}$, $\{2,5\}$, $\{3,4\}$, $\{5,2\}$ ou plutôt $\{6,1\}$ qui est apparu.

Après avoir effectué une expérience aléatoire, on s'intéresse davantage à une **fonction du résultat** qu'au résultat lui-même—c'est une variable aléatoire.

Définition: Soit Ω un ensemble fondamental. Une **variable aléatoire** définie sur Ω est une fonction de Ω dans $\mathbb{R}$ (ou dans un sous-ensemble $H \subseteq \mathbb{R}$) :

$$\begin{aligned} X : \quad & \Omega \longrightarrow \mathbb{R} \\ & \omega \longrightarrow X(\omega), \end{aligned}$$

où ω est un événement élémentaire.

L'ensemble H des valeurs prises par la variable aléatoire X peut être **discret** ou **continu**. Par exemple :

- Nombre de piles obtenus en n lancers d'une pièce : $H = \{0, 1, \ldots, n\}$.
- Nombre d'appels téléphoniques pendant une journée : $H = \{0, 1, \ldots\}$.
- Temps d'attente au M1 : $H = [0, T_{\max}]$.
- Quantité de pluie demain : $H = \mathbb{R}_+$.

Variables aléatoires discrètes

Définition: Une variable aléatoire X est dite **discrète** si elle prend un nombre fini ou dénombrable de valeurs. Dénotons $x_i, i = 1, 2, \ldots,$ les valeurs possibles de X. Alors la fonction

$$f_X(x_i) = \Pr(X = x_i)$$

est appelée **fonction de masse** (ou fonction des fréquences). Le comportement d'une variable aléatoire discrète X est complètement décrit par

- les valeurs $x_1, \ldots, x_k$ (k pas nécessairement fini) que X peut prendre ;
- les probabilités correspondantes

$$f_X(x_1) = \Pr(X = x_1), \ldots, f_X(x_k) = \Pr(X = x_k).$$

Fonction de masse

La **fonction de masse** f_X satisfait :

- $0 \leq f_X(x_i) \leq 1$, pour $i = 1, 2, \ldots$
- $f_X(x) = 0$, pour toutes les autres valeurs de x.
- $\sum_{i=1}^{k} f_X(x_i) = 1$.

Exemple On lance deux dés équilibrés. Trouver :
(a) la fonction de masse de la somme ; (b) la fonction de masse du maximum.

Solution 97 (a)

Solution 97 (b)

Fonction de répartition (cas discret ou continu)

Définition: La **fonction de répartition** F_X de la variable aléatoire (générale) X est

$$F_X(x) = \Pr(X \leq x), \quad x \in \mathbb{R}.$$

Elle a les propriétés suivantes :

- F_X prend des valeurs dans $[0, 1]$
- F_X est continue à droite et monotone non décroissante, avec

$$\lim_{x \to -\infty} F_X(x) = 0, \quad \lim_{x \to \infty} F_X(x) = 1$$

- $\Pr(a < X \leq b) = F_X(b) - F_X(a)$
- $\Pr(X > x) = 1 - F_X(x)$
- si X est discrète, alors

$$F_X(x) = \sum_{\{i:\ x_i \leq x\}} \Pr(X = x_i), x \in \mathbb{R}.$$

  et (sauf certains cas pathologiques) F_X est une fonction en escalier avec des sauts de taille $f_X(x_i)$ en x_i

Exemple Donner la fonction de répartition pour le maximum des résultats de deux dés.

Solution 100

Quelques notations (cas discret ou continu)

Par la suite, nous utilisons les notations suivantes :

- Les variables aléatoires sont notées en majuscules ($X, Y, Z, W, T, \ldots$).
- Les valeurs possibles des variables aléatoires sont notées en minuscules ($x, y, z, w, t, \ldots \in \mathbb{R}$).
- La fonction de répartition d'une variable aléatoire X est notée F_X.
- La fonction de masse (ou de densité dans le cas continu, cf plus loin) d'une variable aléatoire X est notée f_X.
- Ces dernières sont notées F ou f s'il n'y pas de risque de confusion.
- $X \sim F$ signifie "la variable aléatoire X suit la loi F, i.e., admet F pour fonction de répartition".
- $X \overset{\text{app}}{\sim} F$ signifie "la variable aléatoire X suit approximativement la loi F".

Loi de Bernoulli

Définition: Une **variable aléatoire de Bernoulli** satisfait

$$X = \begin{cases} x_1 = 0 & \text{si échec} \quad \text{probabilité } 1-p, \\ x_2 = 1 & \text{si succès} \quad \text{probabilité } p; \end{cases}$$

on écrit $X \sim \mathcal{B}(p)$. Sa loi de probabilité est donc

x_i	0	1	Total
$f_X(x_i) = \Pr(X = x_i)$	$1-p$	p	1

où p est la probabilité de succès.
Exemple du lancer d'une pièce de monnaie avec probabilité p fixée d'obtenir "Pile".

Loi binomiale

Définition: On effectue m fois indépendamment une expérience qui mène soit à un succès (avec probabilité p) soit à un échec (avec probabilité $1-p$). Soit X le nombre de succès obtenus. Alors on écrit $X \sim \mathcal{B}(m,p)$, et

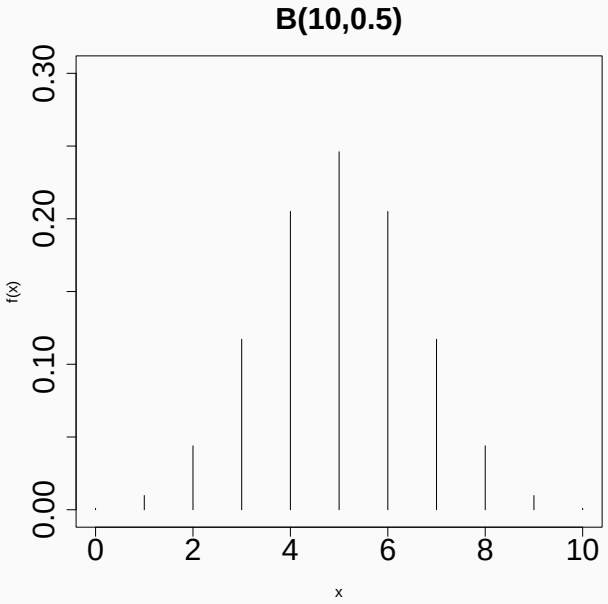
$$f_X(x) = \binom{m}{x} p^x (1-p)^{m-x}, \qquad x = 0, \ldots, m.$$

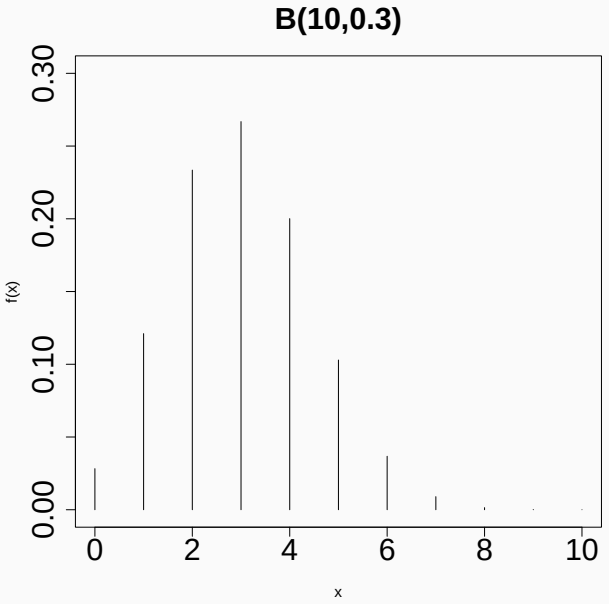
Ceci est la **loi binomiale** avec nombre d'essais m et probabilité p. Dans le cas $m = 1$, X est une variable de Bernoulli. m s'appelle **dénominateur** et p **probabilité de succès**.

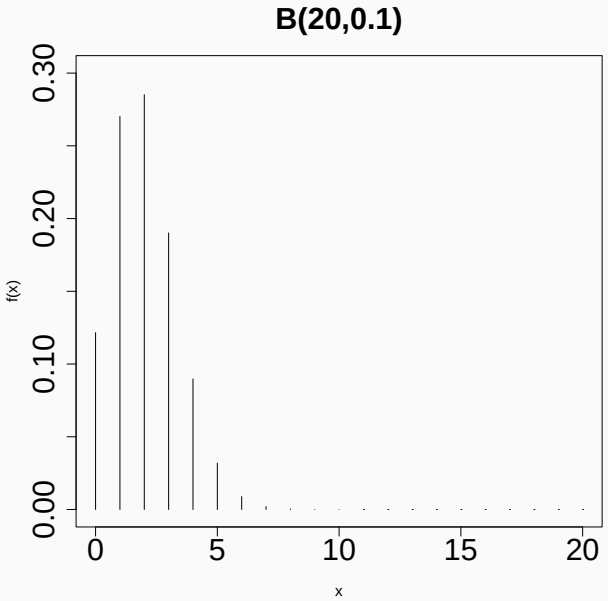
Exemple : m lancers indépendants d'une pièce de monnaie avec $\Pr(\text{“Pile”}) = p$ fixée.

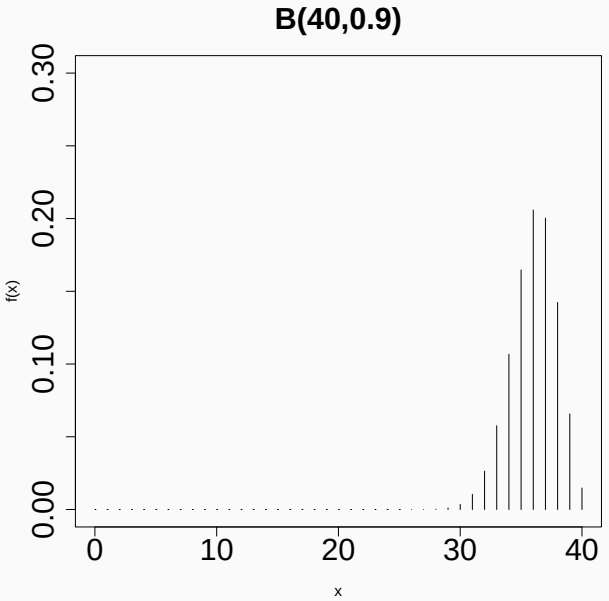
Exemple Trouver la loi du nombre X de personnes présentes à ce cours ayant leur anniversaire ce mois-ci.

Fonctions de masse binomiale









Solution Exemple 104

Variable aléatoire de Poisson

Définition: Une variable aléatoire X pouvant prendre pour valeurs $0, 1, 2, \ldots$ est dite de **Poisson** avec paramètre $\lambda > 0$ si

$$f_X(x) = \frac{\lambda^x}{x!} e^{-\lambda}, \quad x \in \{0, 1, 2, \ldots\}, \quad \lambda > 0.$$

On écrit $X \sim \text{Poiss}(\lambda)$.

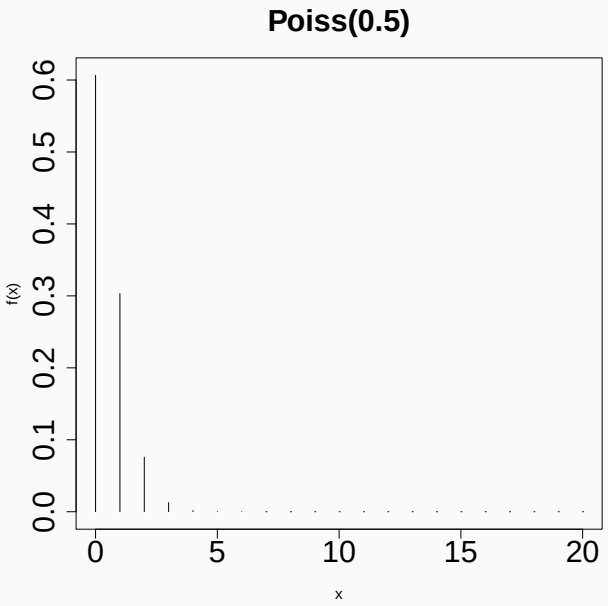
Applications :

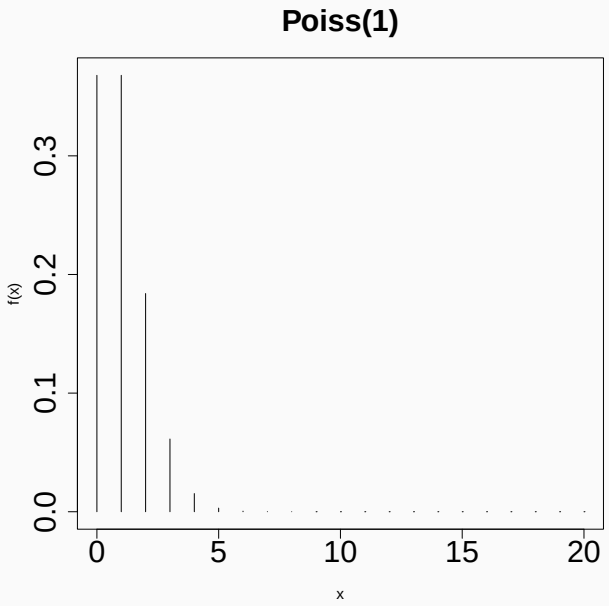
- nombre d'appels téléphoniques par minute dans une centrale téléphonique
- nombre de fautes de frappe dans les notes de cours
- nombre d'avalanches mortelles en Suisse cet hiver

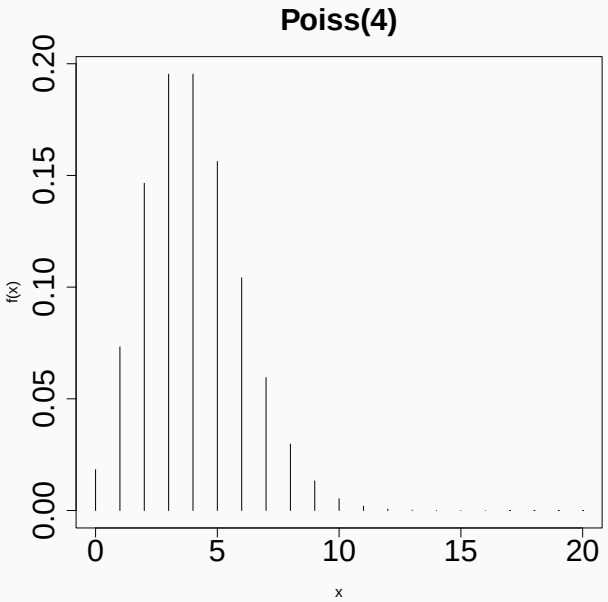
Exemple : E. coli Le niveau residuel des bactéries E. coli dans l'eau traitée est de 2/100 ml, en moyenne. (a) Trouver la probabilité qu'il y ait $k = 0, 1, 2, 3$ présent dans un échantillon de 200 ml d'eau.

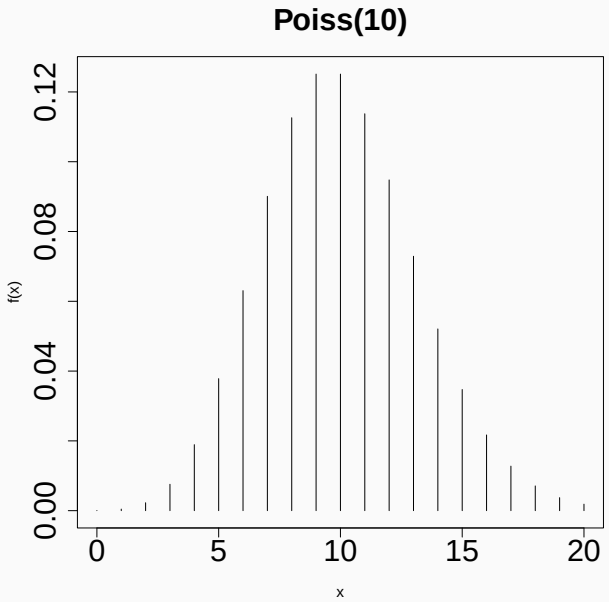
(b) Si on en trouve 10 dans un tel échantillon, l'eau est-elle bonne ?

Fonctions de masse Poisson









Solution Exemple 107

Approximation poissonienne de la loi binomiale

Soit $X \sim \mathcal{B}(m, p)$ avec m grand et p petit. Alors

$$X \overset{\text{app}}{\sim} \text{Poiss}(\lambda = mp).$$

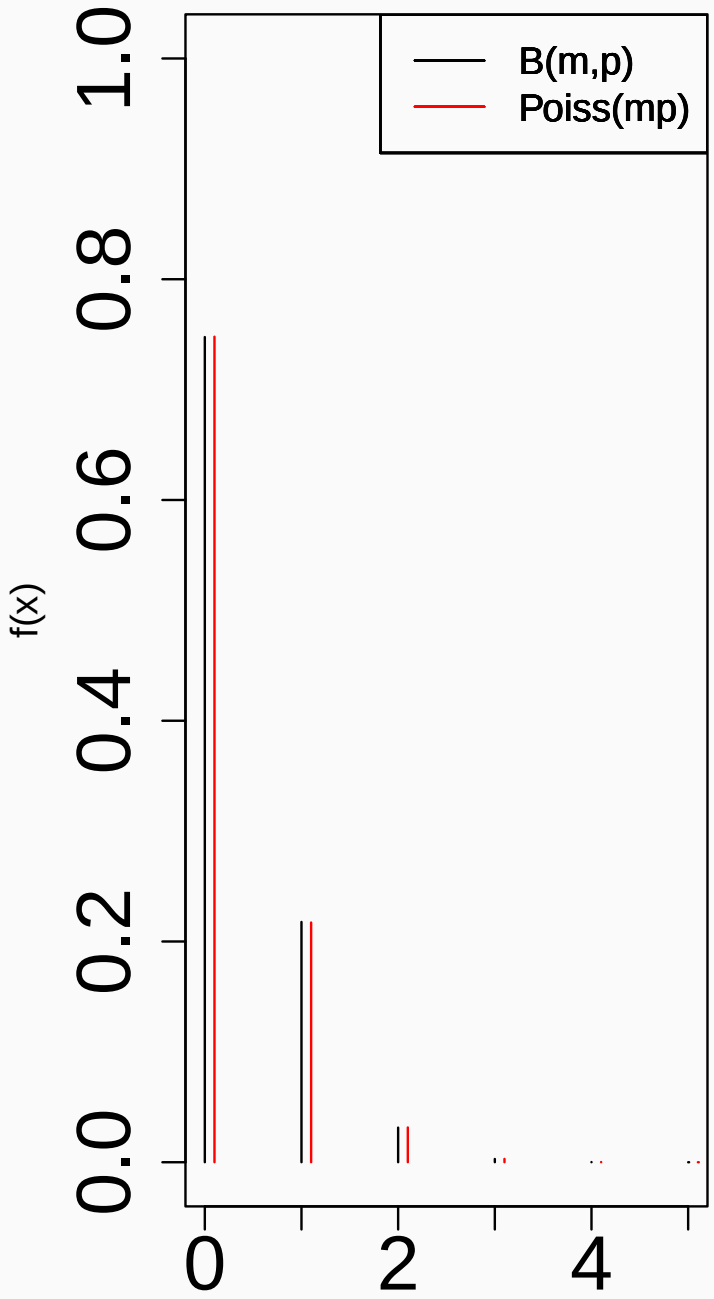
Ceci s'appelle parfois la **loi des petits nombres**.

Exemple D'après IS-Academia, vous êtes m étudiant(e)s.

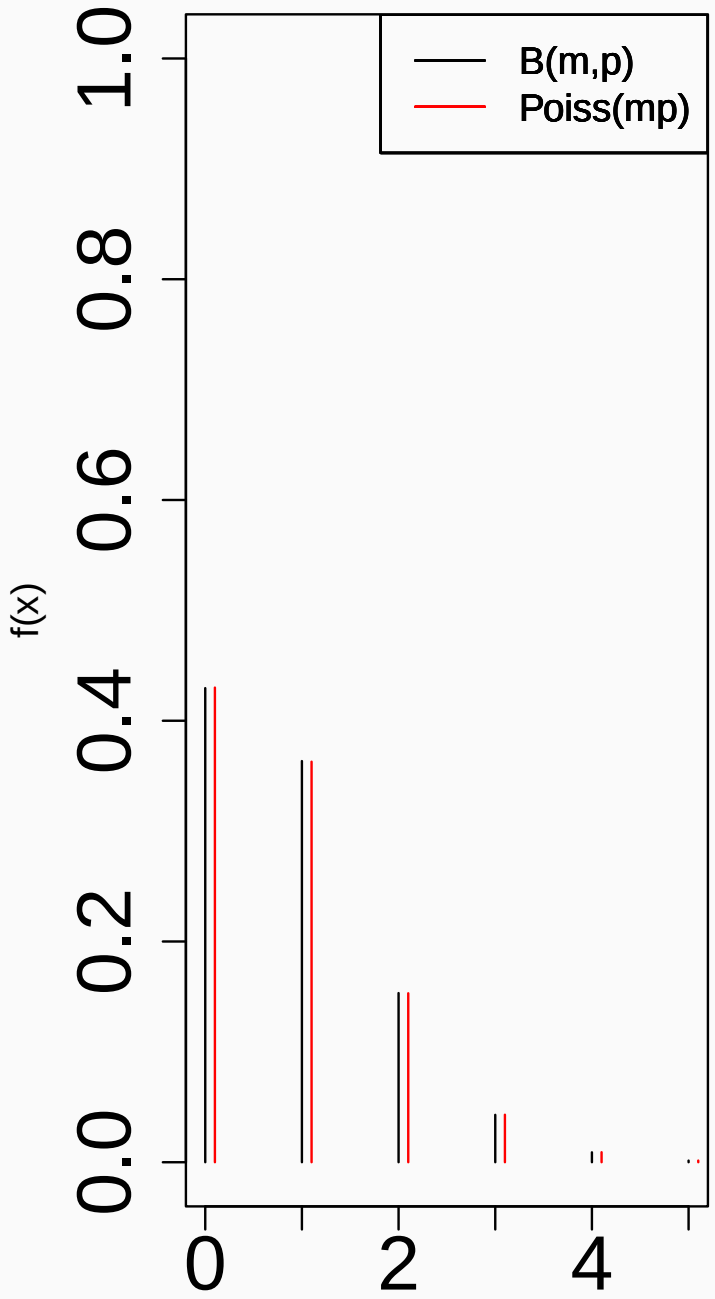
Soit X le nombre de personnes parmi vous dont l'anniversaire a lieu aujourd'hui.

Calculer les probabilités que $X = 0$, $X = 1$, et $X > 1$, sous la loi binomiale et son approximation poissonienne.

$m = 106, p = 1/365$



$m = 308, p = 1/365$



Variables aléatoires continues

Définition: On dit qu'une variable aléatoire X est **continue** s'il existe une fonction $f_X : \mathbb{R} \to [0, \infty)$ appelée **fonction de densité** telle que

$$\Pr(X \in A) = \int_A f_X(u)\mathrm{d}u,$$

où $A \subseteq \mathbb{R}$ est un ensemble 'raisonnable'. Par exemple, pour $A = (a, b]$,

$$\Pr(X \in A) = \Pr(a < X \leq b) = \int_a^b f_X(x)\mathrm{d}x.$$

f_X **n'est pas** une probabilité, mais une limite

$$f_X(x) = \lim_{h \to 0} \frac{1}{2h} \Pr(x - h \leq X \leq x + h)$$

Une variable continue peut prendre une infinité des valeurs, souvent dans un intervalle (borné, demi-droite, ou tout $\mathbb{R}$).

Fonctions de densité et de répartition : propriétés

- Propriétés de la **fonction de densité** :
  - $f_X(x) \geq 0$ pour tout $x \in \mathbb{R}$;
  - $\int_{-\infty}^{\infty} f_X(x)\mathrm{d}x = 1$.
- Si l'on pose $a = b$, on a

$$\Pr(X = a) = \int_a^a f_X(x)\mathrm{d}x = 0.$$

- La **fonction de répartition**, F_X, vérifie

$$F_X(a) = \Pr(X \leq a) = \Pr(X < a) = \int_{-\infty}^{a} f_X(x)\mathrm{d}x, \quad a \in \mathbb{R}.$$

- On a, pour tout $a, b \in \mathbb{R}$ tels que $a < b$,

$$\Pr(a < X \leq b) = F_X(b) - F_X(a) = \Pr(a < X < b).$$

- On a

$$f_X(x) = \frac{\mathrm{d}}{\mathrm{d}x} F_X(x) = F_X'(x), \quad x \in \mathbb{R}.$$

Quelques lois continues

- **Loi uniforme** : $X \sim U(a, b)$, pour $a < b$, de densité

$$f_X(x) = \begin{cases} 1/(b-a) & \text{si } a \leq x \leq b, \\ 0 & \text{sinon.} \end{cases}$$

- **Loi exponentielle** : $X \sim \exp(\lambda)$, pour $\lambda > 0$, de densité

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & \text{si } x \geq 0, \\ 0 & \text{sinon.} \end{cases}$$

- **Loi normale** : $X \sim \mathcal{N}(\mu, \sigma^2)$, pour $\mu \in \mathbb{R}, \sigma > 0$, de densité

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)}, \quad x \in \mathbb{R}.$$

Si $X \sim \mathcal{N}(\mu, \sigma^2)$, alors $Z = (X - \mu)/\sigma \sim \mathcal{N}(0, 1)$ ("standardisation").
Notations : $f_Z(z) = \phi(z)$ et $F_Z(z) = \Phi(z)$.

Quelques lois continues

Exemple

Exemple Le M1 passe toutes les 5.5 minutes. Si j'arrive à un moment choisi au hasard, quelle est la probabilité que je doive attendre (a) plus de 3 minutes ? (b) moins de 2 minutes ? (c) entre 1 et 4 minutes ?

Exemple

Exemple La probabilité qu'il pleuve pendant la journée est de 0.2. S'il pleut, la quantité de pluie journalière suit une loi exponentielle de parametre $\lambda = 0.05$ mm^{-1}. Trouver (a) la probabilité qu'il tombe au plus 5mm demain, (b) la probabilité qu'il tombe au moins 2mm demain.

Exemples

Exemple La quantité annuelle de pluie dans une certaine région est une variable aléatoire normale de moyenne $\mu = 140$ cm et de variance $\sigma^2 = 16$ cm^2. Quelle est la probabilité qu'il tombe entre 135 et 150 cm ?

2.2.3 Variables aléatoires conjointes

Variables aléatoires conjointes / simultanées

Soient X et Y deux variables aléatoires définies sur le même ensemble Ω. La **fonction de répartition conjointe (ou simultanée)** de X et Y est définie par

$$F_{X,Y}(x,y) = \Pr(X \leq x,\ Y \leq y), \qquad x,y \in \mathbb{R}.$$

- **Cas discret** (i.e., X et Y sont discrètes) : la loi de probabilité conjointe de X et Y est parfaitement déterminée si l'on connaît leur **fonction de masse conjointe**, i.e.,
$$f_{X,Y}(x_i,y_j) = \Pr(X = x_i, Y = y_j)$$
pour tous les couples (x_i, y_j) possibles.
- **Cas continu** (i.e., X et Y sont continues) : la loi de probabilité conjointe de X et Y est parfaitement déterminée si l'on connaît leur **fonction de densité conjointe**, définie (si elle existe) par
$$f_{X,Y}(x,y) = \frac{\partial^2 F_{X,Y}(x,y)}{\partial x \partial y}, \qquad x,y \in \mathbb{R}.$$

Cas discret : propriétés

- Propriétés de la fonction de masse conjointe :
  - $0 \leq f_{X,Y}(x_i, y_j) \leq 1, i, j = 1, 2, \ldots$
  - $f_{X,Y}(x, y) = 0$, pour toutes les autres valeurs de x et y.
  - $\sum_{i,j} f_{X,Y}(x_i, y_j) = 1$.

- La fonction de répartition conjointe vérifie

$$F_{X,Y}(x, y) = \sum_{\{(i,j):\ x_i \leq x, y_j \leq y\}} f_{X,Y}(x_i, y_j), \quad x, y \in \mathbb{R}.$$

Cas continu : propriétés

- Propriétés de la densité conjointe :
  - $f_{X,Y}(x,y) \geq 0, \quad x,y \in \mathbb{R}$.
  - $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(u,v)\mathrm{d}v\mathrm{d}u = 1$.
- La fonction de répartition conjointe vérifie

$$F_{X,Y}(x,y) = \Pr(X \leq x, Y \leq y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f_{X,Y}(u,v)\mathrm{d}v\mathrm{d}u, \quad x,y \in \mathbb{R}$$

- On a, pour tout $a_1, a_2, b_1, b_2 \in \mathbb{R}$ tels que $a_1 < b_1$ et $a_2 < b_2$,

$$\Pr(a_1 < X \leq b_1,\ a_2 < Y \leq b_2) = \int_{a_1}^{b_1}\int_{a_2}^{b_2} f_{X,Y}(u,v)\mathrm{d}v\mathrm{d}u.$$

Lois marginales

Définition: Soient X, Y deux variables aléatoires ayant pour densité (ou fonction de masse) conjointe $f_{X,Y}$. Les **densités marginales** du couple (X, Y) sont respectivement les densités de X et Y, i.e., f_X et f_Y. De même, les **fonctions de répartition marginales** du couple (X, Y) sont respectivement les fonctions de répartition de X et Y, i.e., F_X et F_Y.

Dans le cas des densités, on a

- **cas discret** : $f_X(x_i) = \sum_j f_{X,Y}(x_i, y_j), \quad f_Y(y_j) = \sum_i f_{X,Y}(x_i, y_j)$;
- **cas continu** : $f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\mathrm{d}y, \quad f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\mathrm{d}x.$

Concernant les fonctions de répartition, on a

- **cas discret** : $F_X(x) = \sum_{\{i:x_i \le x\}} f_X(x_i), \quad F_Y(y) = \sum_{\{j:y_j \le y\}} f_Y(y_j)$;
- **cas continu** : $F_X(x) = \int_{-\infty}^{x} f_X(u)\,\mathrm{d}u, \quad F_Y(y) = \int_{-\infty}^{y} f_Y(v)\,\mathrm{d}v.$

Exemple X, Y prennent les valeurs $(1, 2), (1, 4), (2, 3), (3, 2), (3, 4)$ avec probabilités égales. Trouver les lois marginales de X et de Y.

Solution 123 et 125

Exemple X, Y prennent les valeurs $(1,2),(1,4),(2,3),(3,2),(3,4)$ avec probabilités égales. Trouver les lois marginales de X et de Y.

Indépendance

Définition: Deux variables aléatoires X et Y sont **indépendantes** si

$$\Pr(X \leq x,\ Y \leq y) = \Pr(X \leq x) \times \Pr(Y \leq y), \quad \forall x, y \in \mathbb{R}.$$

Dans ce cas on écrit $X \perp\!\!\!\perp Y$.

- Donc $X \perp\!\!\!\perp Y \iff \forall x, y \in \mathbb{R} : F_{X,Y}(x, y) = F_X(x)F_Y(y)$
- si $X \perp\!\!\!\perp Y$ et f_X, f_Y sont connues, on peut obtenir $f_{X,Y}$. Ceci est **faux** pour des variables dépendantes
- si $X \perp\!\!\!\perp Y$, alors $g(X) \perp\!\!\!\perp h(Y)$ pour toutes fonctions g, h 'raisonnables'
- Pour des variables aléatoires **discrètes**

  $$\forall x, y \in \mathbb{R} : f_{X,Y}(x, y) = f_X(x) \times f_Y(y) \iff \forall x, y \in \mathbb{R} : F_{X,Y}(x, y) = F_X(x) \times F_Y(y)$$

- Pour des variables aléatoires **continues** $\Longrightarrow$ est vrai et pour montrer une **dépendance** il suffit de trouver x, y auxquels $f_{X,Y}$, f_X et f_Y sont continues et $f_{X,Y}(x, y) \neq f_X(x) \times f_Y(y)$

Exemple Les variables aléatoires X, Y de l'exemple précédant sont-elles indépendantes ?

Cas continu

La fonction de répartition conjointe est

$$\Pr(X \leq x,\ Y \leq y) = F_{X,Y}(x,y) = \int_{-\infty}^{y}\int_{-\infty}^{x} f_{X,Y}(u,v)\,\mathrm{d}u\,\mathrm{d}v.$$

Propriétés :

- $f_{X,Y}(x,y) \geq 0$ pour tout $(x,y) \in \mathbb{R}^2$
- $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(u,v)\,\mathrm{d}u\,\mathrm{d}v = 1$
- $f_{X,Y}(x,y) = \frac{\partial^2 F_{X,Y}(x,y)}{\partial x\,\partial y}$
- $\Pr(a_1 < X \leq b_1,\ a_2 < Y \leq b_2) = \int_{a_2}^{b_2}\int_{a_1}^{b_1} f_{X,Y}(u,v)\,\mathrm{d}u\,\mathrm{d}v$
- Plus généralement, pour $A \subseteq \mathbb{R}^2$ 'raisonnable'

$$\Pr((X,Y) \in A) = \int_A f_{X,Y}(u,v)\mathrm{d}u\mathrm{d}v$$

Exemple Soient $X \sim U[0,1]$ et $Y \sim U[0,2]$ indépendantes. Trouver $\Pr(X > Y)$.

Noter : $Y' = 2X \sim U[0,2]$ mais $\Pr(X > Y') = 0$; X et Y' sont dépendantes !

Solution 126

Densité conditionelle

Définition: La **densité conditionnelle** de X sachant $Y = y$ (tel que $f_Y(y) > 0$) est définie par

$$f_{X|Y}(x \mid y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}, \qquad x \in \mathbb{R}.$$

Si X et Y sont indépendantes, on a

$$f_{X|Y}(x \mid y) = f_X(x), \quad f_{Y|X}(y \mid x) = f_Y(y), \quad \text{pour tout } x \text{ et } y \in \mathbb{R}.$$

(mathématiquement, c'est pour 'presque' tout x, y)

Exemple Soient X et Y de densité conjointe

$$f_{X,Y}(x,y) = \begin{cases} x+y & \text{si} \quad 0 < x < 1, 0 < y < 1, \\ 0 & \text{sinon.} \end{cases}$$

Trouver les densités marginales de X et Y, et la densité conditionnelle $f_{X|Y}$. Les deux variables sont-elles indépendantes ?

Solution Exemple 128

2.3 Valeurs caractéristiques

Mesure de tendance centrale

Définition: L'**espérance** d'une variable aléatoire X est

$$\mathbb{E}(X) = \begin{cases} \sum_i x_i \, f_X(x_i), & X \text{ discrète,} \\ \int_{-\infty}^{\infty} x f_X(x) \, \mathrm{d}x, & X \text{ continue,} \end{cases}$$

si la somme/intégrale converge

Propriétés :

- Interprétation 1 : espérance $\equiv$ centre de gravité d'un ensemble de masses
- Interprétation 2 : espérance $\equiv$ moyenne pondérée par des masses
- si $X_1, \ldots, X_n$ sont des variables aléatoires et $a, b_1, \ldots, b_n$ des constantes, alors

$$\mathbb{E}\left(a + \sum_{i=1}^{n} b_i X_i\right) = a + \sum_{i=1}^{n} b_i \mathbb{E}(X_i)$$

- pour g fonction 'raisonnable', $\mathbb{E}\{g(X)\} = \begin{cases} \sum_i g(x_i) f_X(x_i), & X \text{ discrète} \\ \int_{-\infty}^{\infty} g(x) f_X(x) \mathrm{d}x, & X \text{ continue} \end{cases}$
- si X, Y sont indépendantes et g, h des fonctions 'raisonnables', alors

$$\mathbb{E}\{g(X)h(Y)\} = \mathbb{E}\{g(X)\}\mathbb{E}\{h(Y)\}$$

Exemples

Exemple Pour $X \sim \mathcal{B}(m, p)$, trouver $\mathbb{E}(X)$.

Exemple Pour $X \sim \mathrm{Poiss}(\lambda)$, trouver $\mathbb{E}(X)$ et $\mathbb{E}\{X(X-1)\}$.

Exemples

Exemple Soit $X \sim \mathcal{N}(\mu, \sigma^2)$, trouver $\mathbb{E}(X)$.

Mesure de dispersion

Définition: La **variance** d'une variable aléatoire X est définie comme

$$\text{var}(X) = \mathbb{E}[\{X - \mathbb{E}(X)\}^2] = \cdots = \mathbb{E}(X^2) - \mathbb{E}(X)^2$$

Propriétés :

- Interprétation physique : variance $\equiv$ moment d'inertie relatif au centre de masse
- $\text{var}(X) \geq 0$, et $\text{var}(X) = 0$ implique que X est constante
- la **déviation standard** de X est définie comme $\text{sd}(X) = \sqrt{\text{var}(X)} \geq 0$
- si a, b sont des constantes, alors $\text{var}(a + bX) = b^2\text{var}(X)$
- si $X_1, \ldots, X_n$ sont indépendantes et $a, b_1, \ldots, b_n$ des constantes, alors

$$\text{var}\left(a + \sum_{i=1}^{n} b_i X_i\right) = \sum_{i=1}^{n} b_i^2 \text{var}(X_i)$$

Exemple Si $X \sim \text{Poiss}(\lambda)$, montrer que $\text{var}(X) = \lambda$.

Exemple Si $X \sim \mathcal{B}(m, p)$, montrer que $\text{var}(X) = m\,p(1-p)$.

Exemple Si $X \sim \mathcal{N}(\mu, \sigma^2)$, montrer que $\text{var}(X) = \sigma^2$.

Exemples : variance

Exemple Si $X \sim \mathrm{Poiss}(\lambda)$, montrer que $\mathrm{var}(X) = \lambda$.

Exemple Si $X \sim \mathcal{B}(m, p)$, montrer que $\mathrm{var}(X) = m\,p(1-p)$.

Exemple Si $X \sim \mathcal{N}(\mu, \sigma^2)$, montrer que $\mathrm{var}(X) = \sigma^2$.

Covariance

Définition: La **covariance** des variables aléatoires X, Y est

$$\text{cov}(X, Y) = \mathbb{E}\left[\{X - \mathbb{E}(X)\}\{Y - \mathbb{E}(Y)\}\right] = \cdots = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y).$$

Interprétation : C'est une mesure de dépendance **linéaire** entre X et Y

Propriétés :

- la covariance dépend des unités dont on mesure X, Y
- $\text{cov}(X, Y) = \text{cov}(Y, X)$
- $\text{cov}(X, X) = \text{var}(X)$
- $\text{cov}(X+Y, Z+W) = \text{cov}(X, Z)+\text{cov}(Y, Z)+\text{cov}(X, W)+\text{cov}(Y, W)$
- si a, b, c, d sont des constantes, alors
  $\text{cov}(a\,X + b,\, c\,Y + d) = ac\,\text{cov}(X, Y)$
- $\text{var}(X \pm Y) = \text{var}(X) + \text{var}(Y) \pm 2\text{cov}(X, Y)$
- si X et Y sont indépendantes, alors $\text{cov}(X, Y) = 0$. Mais attention, l'inverse **n'est pas vraie** en général !

Exemple

Exemple (voir diapositive 128) Soient X et Y de densité conjointe

$$f_{X,Y}(x,y) = \begin{cases} x+y & \text{si} \quad 0<x<1,\ 0<y<1, \\ 0 & \text{sinon.} \end{cases}$$

Trouver $\mathrm{Var}(X)$, $\mathrm{Var}(Y)$, et $\mathrm{Cov}(X,Y)$.

Corrélation

Définition: La **corrélation** de X et Y est

$$\rho_{X,Y} = \rho(X, Y) = \mathrm{corr}(X, Y) = \frac{\mathrm{cov}(X, Y)}{\sqrt{\mathrm{var}(X)\mathrm{var}(Y)}}$$

(zéro si une des variances est zéro).

Propriétés :

- $\rho_{X,Y}$ mesure la dépendance **linéaire** (et seulement linéaire !) entre X et Y
- $\rho(a + bX, c + dY) = \mathrm{sign}(bd)\rho(X, Y)$
- $\mathrm{corr}(X, Y) = \mathrm{corr}(Y, X)$
- $\mathrm{corr}(X, X) = 1$ (si X n'est pas constante)
- $\mathrm{corr}(X, -X) = -1$ (si X n'est pas constante)
- $-1 \leq \mathrm{corr}(X, Y) \leq 1$ (inegalité de Cauchy–Schwarz)
- si X et Y sont indépendantes, alors $\mathrm{corr}(X, Y) = 0$, mais la réciproque est faux !
- corrélation $\neq$ causalité !

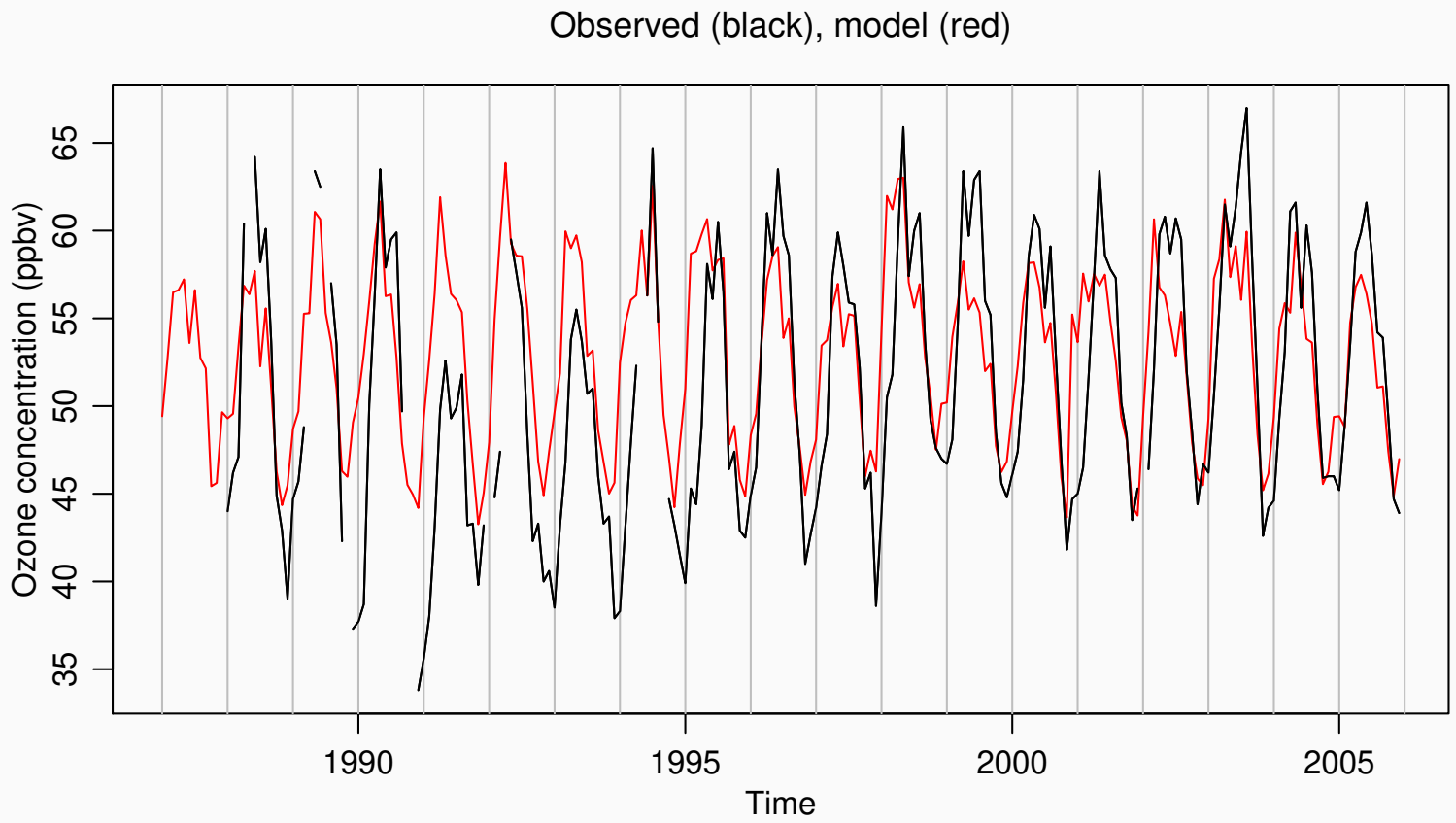
Corrélation empirique

Version empirique (si $\Pr((X = x_i, Y = y_i) = 1/n$ pour $i = 1, \ldots, n)$

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

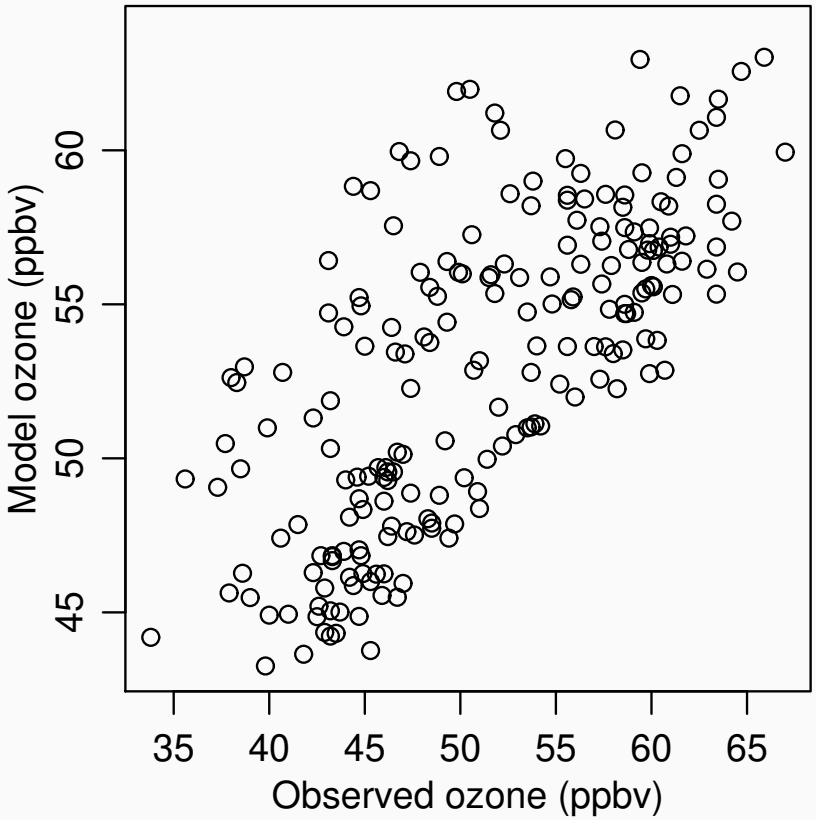
Exemple : ozone atmosphérique

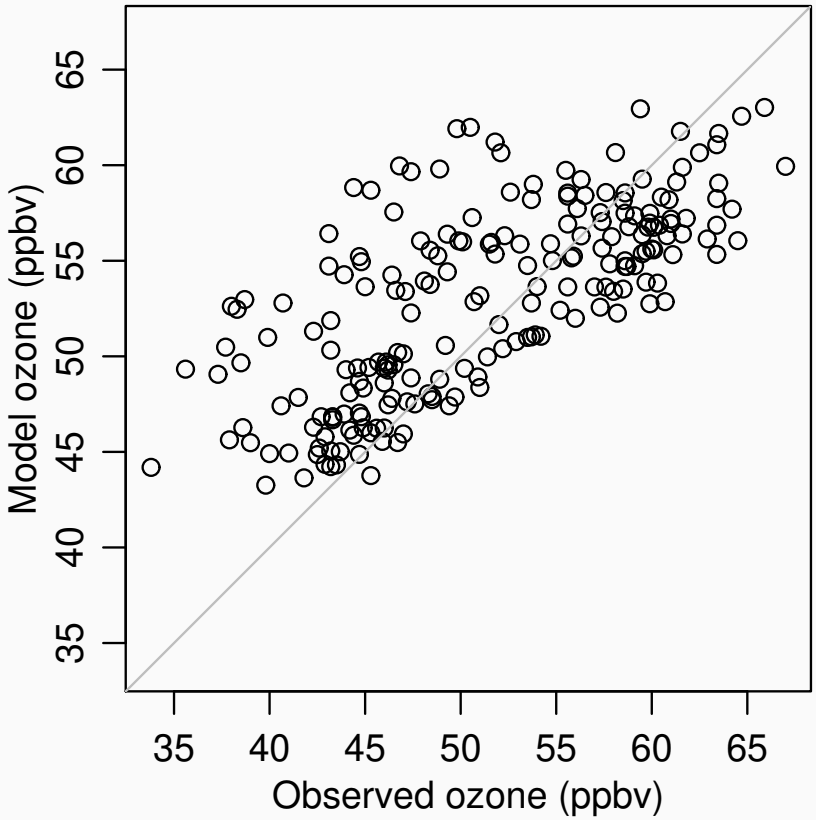
Prof. Isabelle Bey (SIE) : observations de la concentration d'ozone au Jungfraujoch de janvier 1987 à décembre 2005 (quelques valeurs manquantes), et résultats d'une modélisation.



La modélisation vous paraît-elle bonne ?

Exemple : ozone atmosphérique

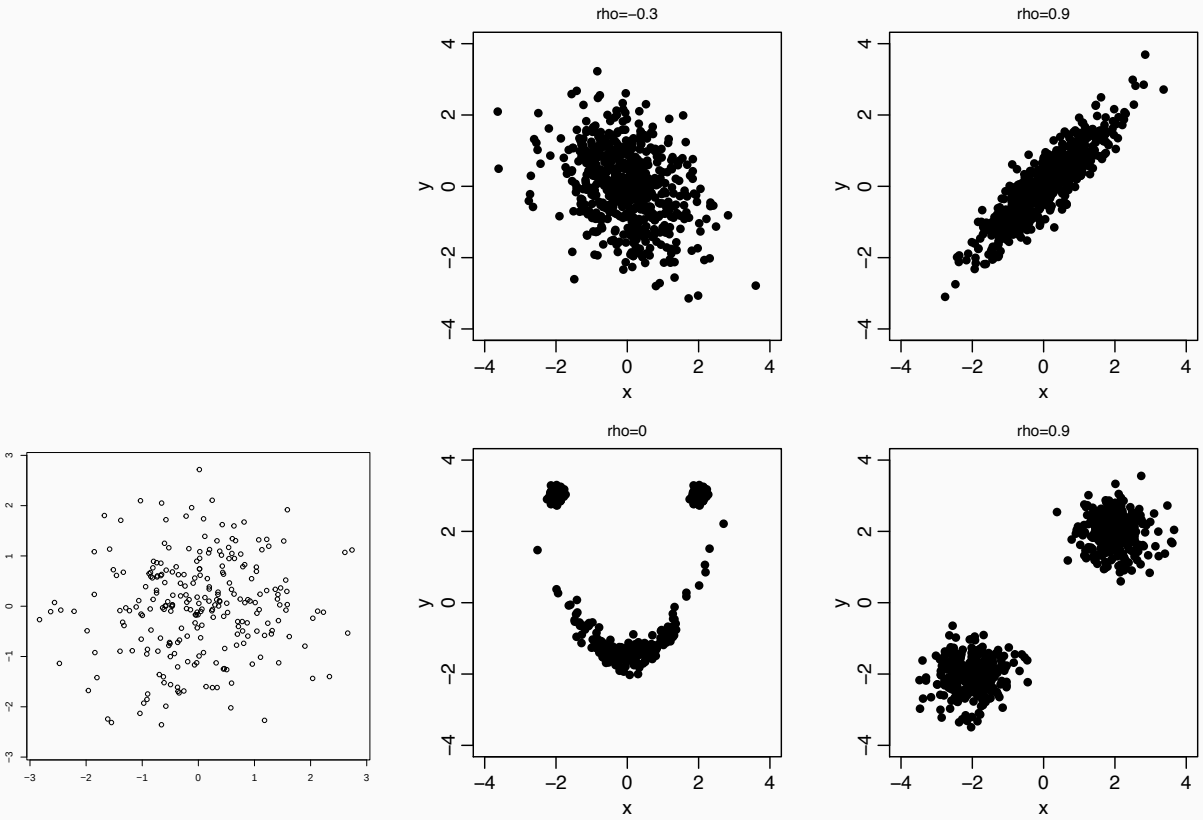




La corrélation empirique est $\rho = 0.707$.

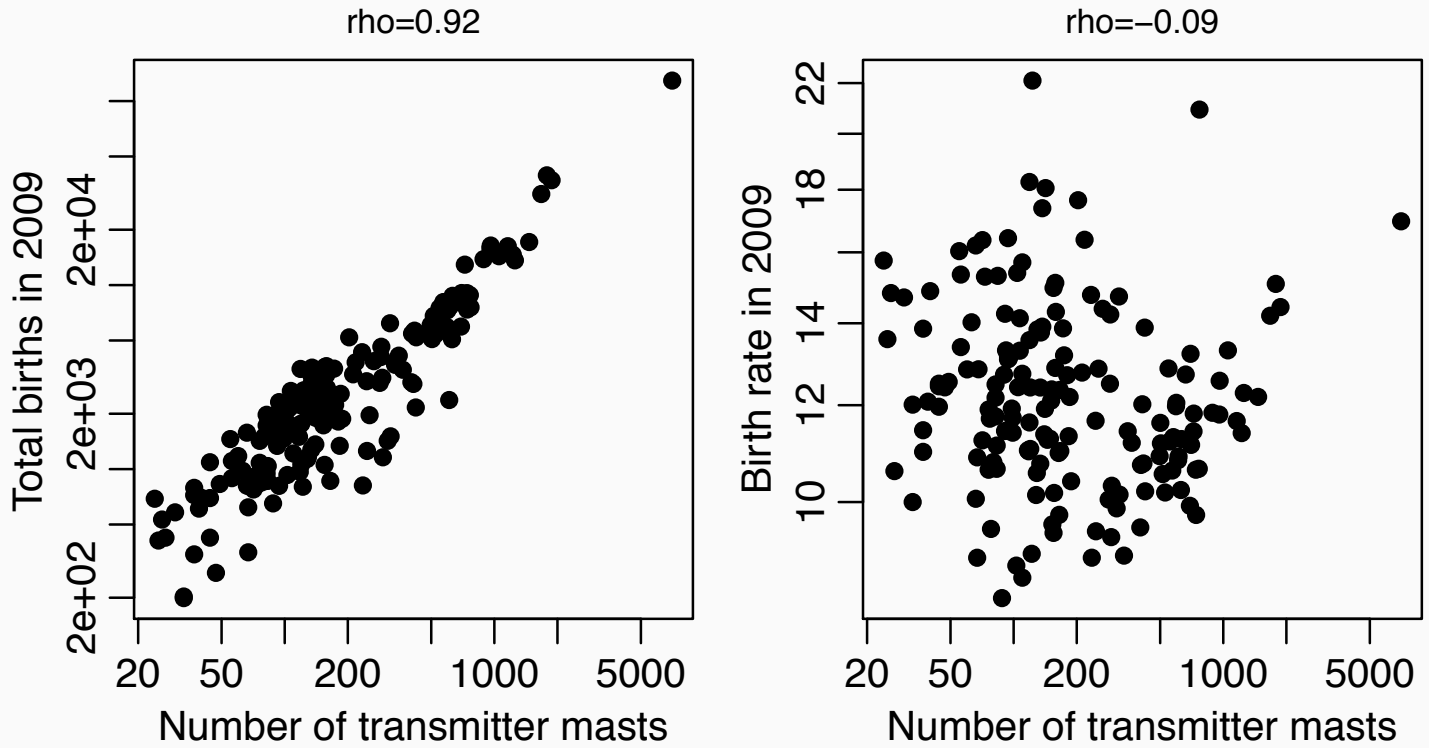
Limitations de la corrélation

- ρ mesure la dépendance linéaire (panneaux supérieurs)
- On peut avoir $\rho \approx 0$, mais dépendance forte mais non-linéaire (en bas au milieu)
- Une corrélation pourrait être forte mais **specieuse**, comme en bas à droite, ou deux sous-groupes, chacun sans corrélation, sont combinés
- Une corrélation entre deux variables n'implique pas une causalité entre elles



Corrélation $\neq$ causalité

Deux variables peuvent être très corrélées sans lien de causalité. Le graphique à gauche ici montre une corrélation forte entre le nombre de naissances et les mâts de communication dans les villes anglaises . . .



Danger !

- Les espérances/variances/covariances/corrélations ne sont pas définies si les intégrales/sommes ne convergent pas
- Ceci est notamment le cas lorsque la distribution de X a des **queues lourdes** : la densité de X décroît trop lentement vers zéro, et X a une probabilité élevée de prendre des valeurs énormes.

Exemple

- Considérons la fonction de densité $f(x) = \alpha x^{-1-\alpha}$ sur $[1, \infty)$ et $f(x) = 0$ pour $x < 1$ (loi **Pareto**). Pour $r \in \mathbb{R}$ on a

$$\mathbb{E}(X^r) = \alpha \int_1^\infty x^{r-1-\alpha} dx = \begin{cases} \frac{\alpha}{\alpha - r} & r < \alpha \\ \infty & r \geq \alpha \end{cases}$$

- En particulier, $\mathbb{E}(X) < \infty$ si et seulement si $\alpha > 1$, et $\text{var}(X) < \infty$ si et seulement si $\alpha > 2$
- Pour α petit la densité tend lentement vers zéro

Espérance d'une variable aléatoire mixte

Théorème de l'espérance totale Pour une partition $A_1, A_2, \ldots$

$$\mathbb{E}(X) = \sum_i \mathbb{E}(X|A_i)\Pr(A_i)$$

Exemple : pluie (diapositive 117) La probabilité qu'il pleuve pendant la journée est 0.2. S'il pleut, la quantité de pluie qui tombe suit une loi exponentielle de parametre $\lambda = 0.05\text{mm}^{-1}$. Trouver l'espérance de la quantité de pluie journalière.

Quantiles

Définition: Soit $0 < p < 1$. On définit le $p$ième **quantile** d'une fonction de répartition F par

$$x_p = \inf\{x : F(x) \geq p\}$$

- Pour des variables aléatoires continues, $F(x_p) = p$, donc x_p est tel que $\Pr(X \leq x_p) = p$
- Pour la plupart des variables aléatoires continues, ceci implique que $x_p = F^{-1}(p)$, où F^{-1} est la fonction inverse de F
- "La plupart" : celles ayant une fonction de densité strictement positive (sur $\{x : 0 < F(x) < 1\}$)
- Pour des variables aléatoires discrètes la situation est plus complexe
- Les quantiles empiriques (diapositive 36) sont des estimations (cf les prochains cours) des quantiles à partir des données à disposition.

En particulier, on appelle le 0.5ème quantile la **médiane** de F

Exemple quantiles

Exemple Calculer les quantiles des lois (a) $U(a, b)$, (b) Pareto (diapositive 144)

2.4 Théorèmes fondamentaux de probabilité

Approche expérimentale

- Considérons l'expérience de jeter une pièce de monnaie 10'000 fois et observons le nombre de "face" obtenues
- Soient $X_1, \ldots, X_n$ les variables aléatoires indépendantes

$$X_i = \begin{cases} 1, & \text{si le } i\text{ème jet donne "face"}, \\ 0, & \text{si le } i\text{ème jet donne "pile"} \end{cases} \qquad \sim B(1, p)$$

- Donc $S_n = X_1 + \cdots + X_n$ représente le nombre de "face" sur n essais et

$$S_n \sim \mathcal{B}(n, p)$$

- La proportion de "Face" sur n jets est $\overline{X}_n := S_n/n$ et

$$\begin{aligned} \mathbb{E}(\overline{X}_n) &= n^{-1}\mathbb{E}(S_n) = n^{-1}\, np = p, \\ \text{var}(\overline{X}_n) &= n^{-2}\text{var}(S_n) = n^{-2} np(1-p) = p(1-p)/n \to 0 \end{aligned}$$

  quand $n \to \infty$
- Donc $\overline{X}_n$ se concentre de plus en plus autour de p

Des inégalités de concentration

Conaissance de seulement l'espérance et de la variance d'une variable aléatoire X donne beaucoup d'information sur X

Théorème (inégalité de Markov) Soit X une variable aléatoire nonnegative et d'espérance finie. Alors pour tout $\epsilon > 0$,

$$\Pr(X \geq \epsilon) \leq \frac{\mathbb{E}(X)}{\epsilon}$$

Preuve. $X \geq \epsilon I(X \geq \epsilon)$ et donc $\mathbb{E}(X) \geq \mathbb{E}(\epsilon I(X \geq \epsilon)) = \epsilon \Pr(X \geq \epsilon)$

Théorème (inégalité de Chebyshev) Soit Y une variable aléatoire d'espérance et variance finies. Alors pour tout $\epsilon > 0$,

$$\Pr(|Y - \mathbb{E}(Y)| \geq \epsilon) \leq \frac{\text{var}(Y)}{\epsilon^2}$$

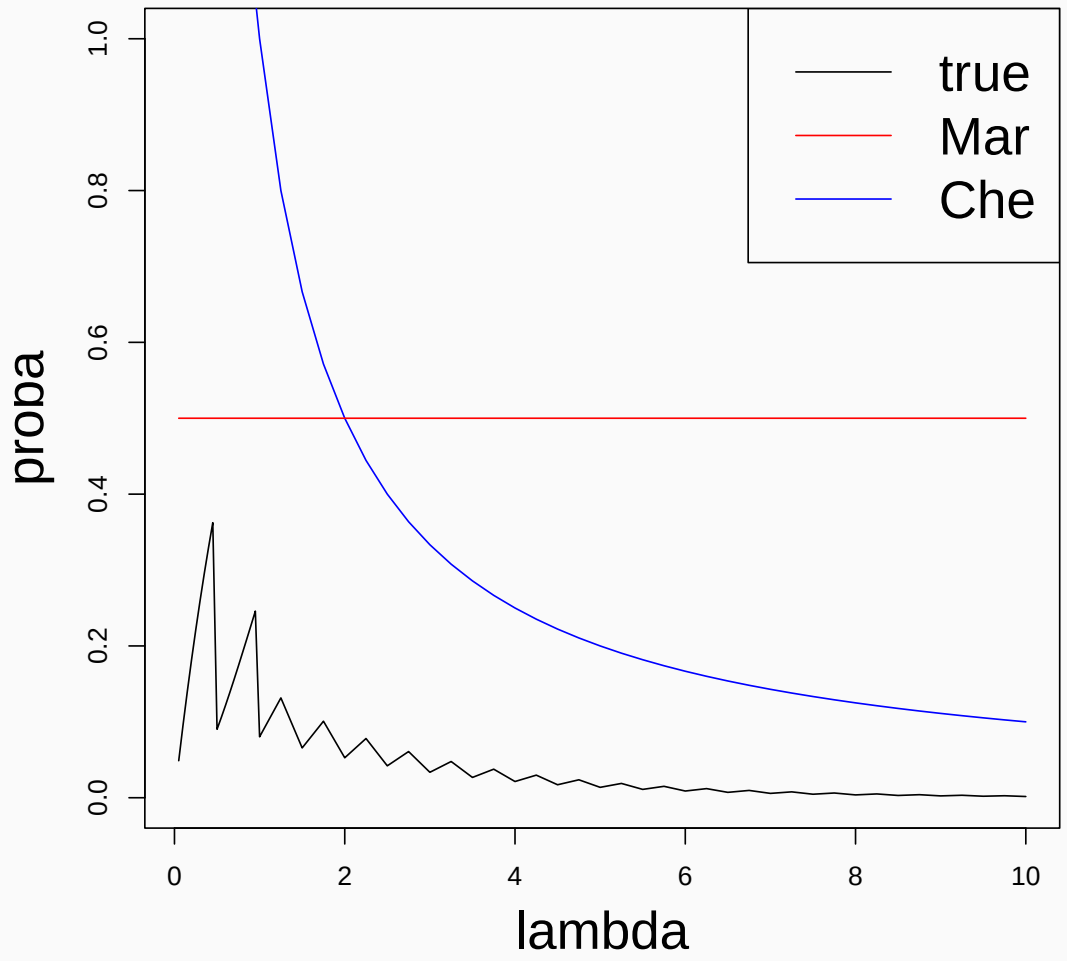
Preuve. Appliquer l'inégalité de Markov à $X = |Y - \mathbb{E}(Y)|^2$

Les **énoncés** sont examinables, mais les preuves ne sont pas examinables

Exemple Trouver une borne supérieure pour $\Pr(X \geq 2\lambda)$ où $X \sim$Poiss(λ)

Exemple des inégalités de concentration

Exemple Trouver une borne supérieure pour $\Pr(X \geq 2\lambda)$ où $X \sim$Poiss(λ)



Lois des grands nombres

Théorème (loi faible des grands nombres) Soient $X_1, X_2, \ldots$ des variables aléatoires indépendantes et identiquement distribuées d'espérance $\mu = \mathbb{E}(X_1)$ et variance $\sigma^2 = \text{var}(X_1)$ finies. Alors pour tout $\epsilon > 0$

$$\Pr(|\overline{X}_n - \mu| \geq \epsilon) \to 0, \qquad n \to \infty. \tag{1}$$

Théorème (loi forte des grands nombres) Soient $X_1, X_2, \ldots$ des variables aléatoires indépendantes et identiquement distribuées d'espérance $\mu = \mathbb{E}(X_1)$ finie. Alors

$$\Pr\left(\lim_{n\to\infty} \overline{X}_n = \mu\right) = 1 \tag{2}$$

Il est donc certain que $\overline{X}_n$ soit proche de μ pour n grand

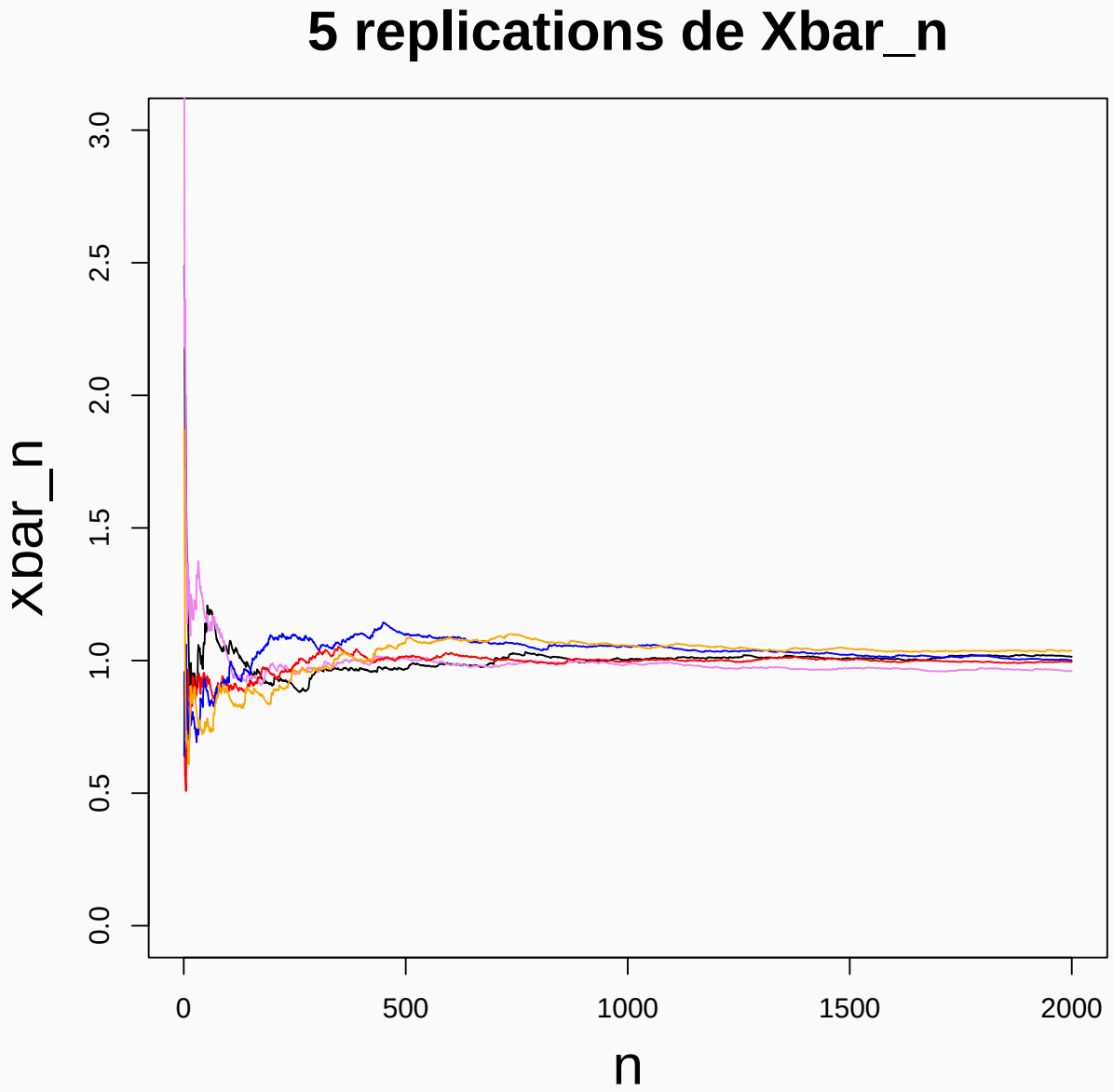
- La loi forte est plus forte parce que (2) implique (1) et la variance peut être infinie
- La loi faible utilise seulement $\text{cov}(X_i, X_j) = 0$ pour $i \neq j$

Preuve : loi faible

Théorème (loi faible des grands nombres) Soient $X_1, \ldots$ des variables aléatoires indépendantes et identiquement distribuées d'espérance $\mu = \mathbb{E}(X_1)$ et variance $\sigma^2 = \mathrm{var}(X_1)$ finies. Alors pour tout $\epsilon > 0$

$$\Pr(|\overline{X}_n - \mu| \geq \epsilon) \to 0, \qquad n \to \infty.$$

Illustration de la loi des grands nombres : exp(1)



Vitesse de convergence : Théorème central limite

- $\overline{X}_n \to \mu$ quand $n \to \infty$, mais à quelle vitesse ?
- Comme $\mathbb{E}(\overline{X}_n) = \mu$ et $\text{var}(\overline{X}_n) = \sigma^2/n \in (0, \infty)$, pour tout n

$$Z_n := \frac{\overline{X}_n - \mu}{\sqrt{\sigma^2/n}} = \sqrt{n}\frac{\overline{X}_n - \mu}{\sigma}$$

a espérance 0 et variance 1, suggérant que la vitesse est $\sqrt{n}$

Théorème central limite Soient $X_1, X_2, \ldots$ des variables aléatoires indépendantes et identiquement distribuées d'espérance μ et variance $\sigma^2 \in (0, \infty)$. Alors $Z_n := \sqrt{n}(\overline{X}_n - \mu)/\sigma$ satisfait

$$\Pr(Z_n \leq x) \to \Phi(x), \qquad x \in \mathbb{R}$$

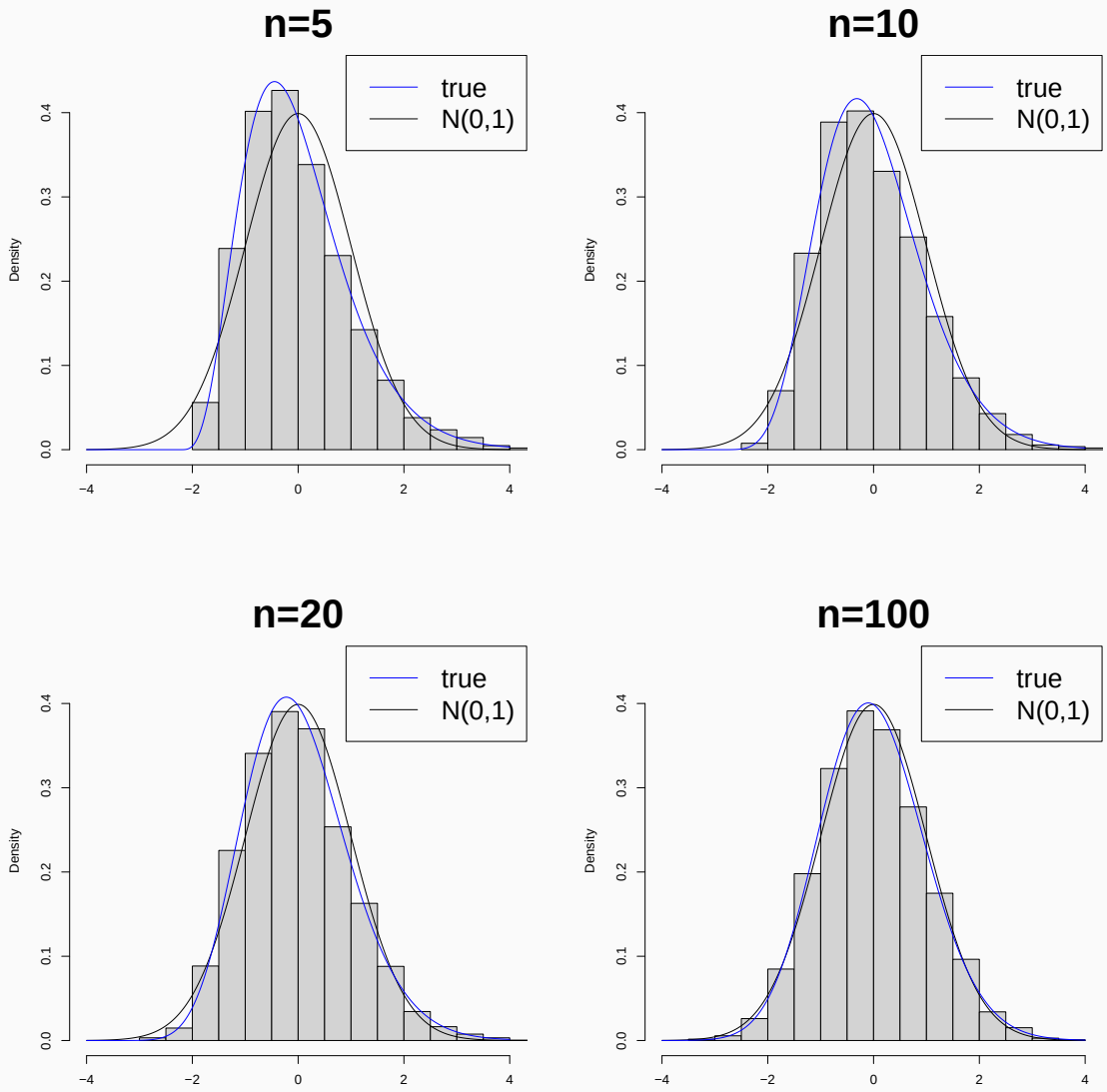
La convergence étant uniforme en x, on déduit

$$\Pr(\overline{X}_n \leq x) = \Pr(Z_n \leq \sqrt{n}(x - \mu)/\sigma) \approx \Phi(\sqrt{n}(x - \mu)/\sigma)$$

donc $\overline{X}_n$ suit **approximativement** une loi $\mathcal{N}(\mu, \sigma^2/n)$

Illustration avec des variables exp(1)

On calcule $\sqrt{n}(\overline{X}_n - \mathbb{E}(X_1))/\sqrt{\text{var}(X_1)}$, $R = 5000$ fois



Exemple

Exemple Soit $X \sim \mathcal{B}(m, p)$. Donner une approximation de $\Pr(X \leq r)$, pour $r \in \mathbb{R}$.

Solution Exemple 158 :
On a $X = \sum_{i=1}^{m} Y_i$, où $Y_1, \ldots, Y_m \overset{\text{iid}}{\sim} \mathcal{B}(p)$. De plus, $\mathbb{E}(Y_1) = p$ et $\mathrm{Var}(Y_1) = p(1-p)$. Le TCL nous donne donc que $X \overset{\text{app}}{\sim} \mathcal{N}(mp, mp(1-p))$ pour m grand. Ainsi, si Z désigne une variable aléatoire de loi $\mathcal{N}(0,1)$, on a, pour m grand,

$$\begin{aligned}\Pr(X \leq r) &= \Pr\left(\frac{X - mp}{\sqrt{mp(1-p)}} \leq \frac{r - mp}{\sqrt{mp(1-p)}}\right) \\ &\approx \Pr\left(Z \leq \frac{r - mp}{\sqrt{mp(1-p)}}\right) = \Phi\left(\frac{r - mp}{\sqrt{mp(1-p)}}\right).\end{aligned}$$

Utilisation du théorème central limite

- Le théorème central limite est utilisé pour approximer des probabilités impliquant des sommes de variables aléatoires indépendantes
- Sous les conditions précédentes, on a

$$\mathbb{E}\left(\sum_{j=1}^{n} X_j\right) = n\mu, \quad \text{var}\left(\sum_{j=1}^{n} X_j\right) = n\sigma^2 \in (0, \infty)$$

- On **standardise** la somme

$$\frac{\sum_{j=1}^{n} X_j - n\mu}{\sqrt{n\sigma^2}} = \frac{n(\bar{X}_n - \mu)}{\sqrt{n\sigma^2}} = \frac{n^{1/2}(\bar{X}_n - \mu)}{\sigma} = Z_n$$

- Par le théorème central limite Z_n est approximativement $\mathcal{N}(0, 1)$ et donc

$$\Pr\left(\sum_{j=1}^{n} X_j \leq r\right) = \Pr\left\{\frac{\sum_{j=1}^{n} X_j - n\mu}{\sqrt{n\sigma^2}} \leq \frac{r - n\mu}{(n\sigma^2)^{1/2}}\right\} \approx \Phi\left\{\frac{r - n\mu}{(n\sigma^2)^{1/2}}\right\}.$$

Exemple Un livre de 640 pages a un nombre aléatoire d'erreurs sur chaque page. Si le nombre d'erreurs par page suit une loi de Poisson d'espérance $\lambda = 0.1$, et est indépendant des autres pages, quelle est la probabilité que le livre contienne moins de 50 erreurs ?

Exemple : théorème central limite

Exemple Un livre de 640 pages a un nombre d'erreurs aléatoires à chaque page. Si le nombre d'erreurs par page suit une loi de Poisson d'espérance $\lambda = 0.1$, et est indépendant des autres pages, quelle est la probabilité que le livre contienne moins de 50 erreurs ?

Extensions et remarques

- Le théorème centrel limite est **remarquable**, car la distribution des X_i **n'a pas d'importance** : seulement l'espérance et la variance apparaissent. $\sum X_i$ a approximativement la même distribution si $X_i \sim Exp(1)$ ou $X_i \sim$Poiss(1)
- **Méthode delta** Si g est une fonction telle que $g'(\mu)$ existe, alors

$$\sqrt{n}\frac{g(\overline{X}_n) - g(\mu)}{\sigma} = g'(\mu)Z_n + o(Z_n)$$

  suit approximativement $\mathcal{N}(0, [g'(\mu)]^2)$ et donc $g(\overline{X}_n) \overset{\text{app}}{\sim} \mathcal{N}\left\{g(\mu), g'(\mu)^2\sigma^2/n\right\}$
- Version **générale** de la méthode delta : si $\sqrt{n}(Y_n - \theta) \overset{\text{app}}{\sim} Y$ pour une constante $\theta \in \mathbb{R}$, alors $\sqrt{n}(g(Y_n) - g(\theta)) \overset{\text{app}}{\sim} g'(\mu)Y$
- **Notation :** $\overset{\text{app}}{\sim}$ indique une distribution **approximative**
- Le théorème centrel limite dépend d'un effect de moyennement, et échoue quand tout dépend d'une fraction minuscule des variables. Il n'est donc pas valable pour les maxima, les minima, l'étendue, . . ., pour lequels on a d'autres théorèmes limites

3. Idées fondamentales de la statistique

Modèles statistiques

On étudie une **population** (ensemble d'individus ou d'éléments) à partir d'un **échantillon** (sous-ensemble).

- **modèle statistique** : X =la quantité étudiée (variable aléatoire) ; la loi F de X est supposée connue sauf un nombre fini de paramètres θ
- **échantillon** (doit être représentatif de la population) : "données" $x_1, \ldots, x_n$, souvent supposées comme étant une réalisation de $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F$ (**indépendantes et identiquement distribuées**) ou $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f$
- **statistique** : une fonction $T_n = h(X_1, \ldots, X_n)$ des variables aléatoires $X_1, \ldots, X_n$
- **estimateur** : une statistique utilisée pour estimer certains paramètres de F
- **Notations** :

$T_n = h(X_1, \ldots, X_n)$	est la statistique (variable aléatoire)
$t_n = h(x_1, \ldots, x_n)$	est la **réalisation** de T_n au moyen des x_i
$\widehat{\theta}$ ou $\widehat{\theta}_n$	est un **estimateur** d'un paramètre θ

Commentaires

Exemple Soient $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$ et $x_1, \ldots, x_n$ une réalisation correspondante. Alors

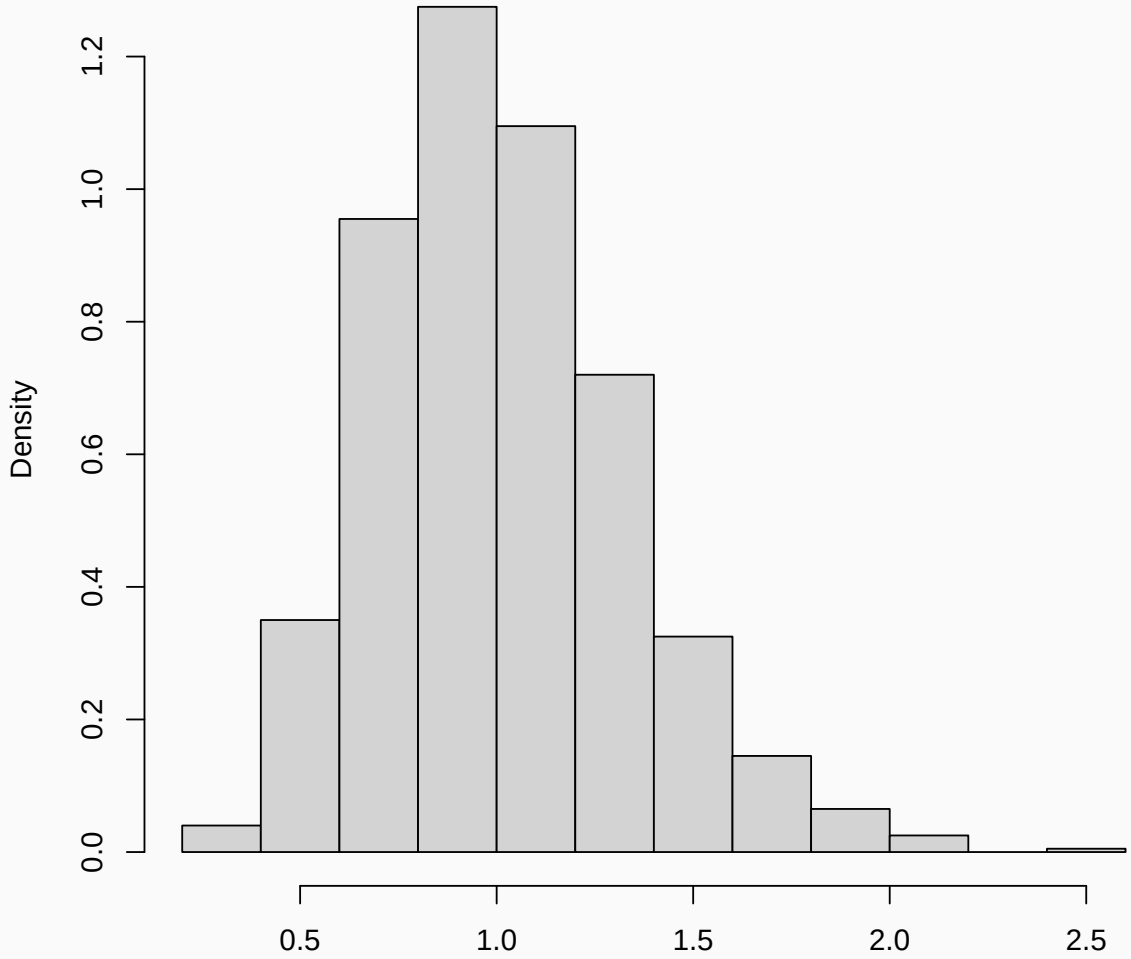
- $\widehat{\mu}_n = \overline{X}_n$ est une estimateur de μ, dont la valeur observée est $\overline{x}_n$
- $\widehat{\sigma}^2 = n^{-1} \sum_{i=1}^n (X_i - \overline{X}_n)^2$, est un estimateur de σ^2, dont la valeur observée est $n^{-1} \sum_{i=1}^n (x_i - \overline{x}_n)^2$

Remarques :

- une statistique T_n étant fonction des variables aléatoires $X_1, \ldots, X_n$, c'est elle-même une variable aléatoire !
- La loi de T_n dépend de la loi des X_i, et est appelée **distribution d'échantillonnage de T_n**
- Si on ne peut pas déduire la loi de T_n de celle des X_i, on doit se contenter parfois de connaître $\mathbb{E}(T_n)$ et $\text{var}(T_n)$, ou, si T_n est liée à $\overline{X}_n$, l'approximer à l'aide de $\mathbb{E}(X)$ et $\text{var}(X)$ et le théorème central limite

Loi d'échantillonnage

Histogramme de 1000 réalisations de $\overline{X}_n = \frac{1}{10}(X_1 + \ldots + X_{10})$ où les $X_i \overset{\text{iid}}{\sim} \exp(1)$



Problèmes attaqués par la statistique

On suppose un **modèle** (c'est à dire une famille de distributions $f(x;\theta)$) et on souhaite

- **estimer** les paramètres θ de ce modèle
- poser des questions au sujet de la valeur de ces paramètres, par exemple **tester** si $\theta = 0$
- **prédire** les valeurs des observations futures

Il existe plusieurs méthodes pour estimer les paramètres d'un modèle. On va décrire les suivantes :

- **méthode des moments** (simple)
- **méthode du maximum de vraisemblance** (souvent utilisée car optimale dans beaucoup de situations)

3.1 Estimation de paramètres

Méthode des moments

- Supposons que l'échantillon tiré soit représentatif de la population
- Pour obtenir des estimateurs pour les paramètres inconnus de la population, on égalise les "moments" de l'échantillon ("empirique") à ceux de la population ("théorique")
- $k$ème moment
  - Population ("théorique") : $m_k = \mathbb{E}(X^k)$. Comme la loi de X dépend de θ, $m_k = m_k(\theta)$
  - Echantillon ("empirique") : $\widehat{m}_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$
  - L'estimateur des moments s'obtient on égalisant m_k et $\widehat{m}_k$, ce qui donne équation(s) pour $\theta \in \mathbb{R}^p$
- On a donc besoin d'autant de moments (supposés finies !) que de paramètres inconnus

Exemple Soient $X_1, \ldots, X_n \overset{\text{iid}}{\sim} U(0, \theta)$. Estimer θ.

Exemple Soient $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$. Estimer μ et σ^2.

Méthode des moments

Exemple Soient $X_1, \ldots, X_n \overset{\text{iid}}{\sim} U(0, \theta)$. Estimer θ.

Méthode des moments

Exemple Soient $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$. Estimer μ et σ^2.

Méthode du maximum de vraisemblance

Définition: Soient $x_1, \ldots, x_n$ des données supposées être une réalisation d'un échantillon aléatoire $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x;\theta)$. La **vraisemblance** pour θ est la fonction

$$L(\theta) = f(x_1;\theta) \times f(x_2;\theta) \times \cdots \times f(x_n;\theta) = \prod_{i=1}^{n} f(x_i;\theta).$$

f est la fonction de densité ou de masse (on suppose qu'elle existe)

Définition: L'**estimateur du maximum de vraisemblance** (maximum likelihood) $\widehat{\theta}_{\text{ML}}$ d'un paramètre θ est celui qui maximise la fonction de vraisemblance parmi tous les θ possibles :

$$L(\widehat{\theta}_{\text{ML}}) \geq L(\theta) \quad \text{pour tout } \theta$$

Il est plus facile de maximiser $\ell(\theta) := \log L(\theta)$, souvent en résolvant $d\ell(\theta)/d\theta = 0$, et vérifiant qu'il s'agit bien d'un maximum (par exemple si la deuxième dérivée est négative $d^2\ell(\theta)/d\theta^2 < 0$)

Exemple $x_1, \ldots, x_n$ réalisations d'une loi $\exp(\lambda)$ avec $\lambda > 0$. Estimer λ

Exemple : maximum de vraisemblance

Exemple Supposons que $x_1, \ldots, x_n$ soient des réalisations i.i.d. d'une loi exponentielle,

$$f(x; \lambda) = \lambda e^{-\lambda x}, \ x \geq 0, \quad \lambda > 0.$$

Trouver $\widehat{\lambda}_{\mathrm{ML}}$.

Erreur quadratique moyenne

Définition: L'**erreur quadratique moyenne** de l'estimateur $\widehat{\theta}$ de θ est

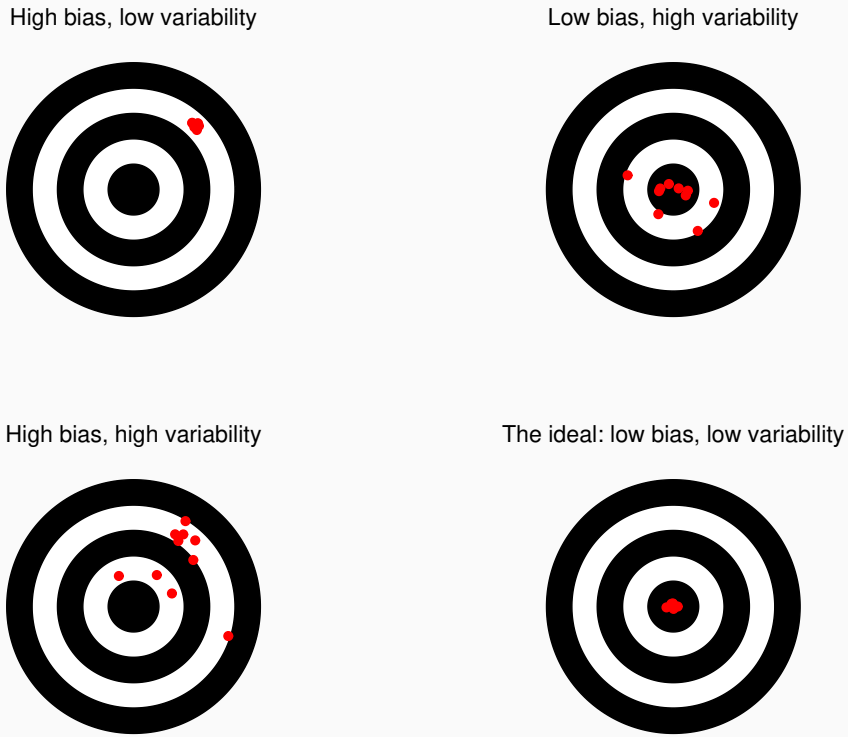
$$\mathrm{EQM}_\theta(\widehat{\theta}) = \mathbb{E}_\theta\{(\widehat{\theta} - \theta)^2\} = \cdots = \mathrm{Var}_\theta(\widehat{\theta}) + [b_\theta(\widehat{\theta})]^2,$$

où $b_\theta(\widehat{\theta}) = \mathbb{E}_\theta(\widehat{\theta}) - \theta$ est le **biais** de $\widehat{\theta}$

- La distribution de $\widehat{\theta}$ dépend de celle des X_i et donc de θ
- Si $\widehat{\theta}$ et $\widehat{\theta}'$ sont deux estimateurs du même paramètre θ et $\mathrm{EQM}_\theta(\widehat{\theta}) \leq \mathrm{EQM}_\theta(\widehat{\theta}')$, on préfère $\widehat{\theta}$
- si $b_\theta(\theta) < 0$, alors $\widehat{\theta}$ sous-estime θ
- si $b_\theta(\theta) > 0$, alors $\widehat{\theta}$ sur-estime θ
- si $b_\theta(\theta) \equiv 0$, alors $\widehat{\theta}$ est **non biaisé**, et $\mathrm{EQM}_\theta(\widehat{\theta}) = \mathrm{var}(\widehat{\theta})$

Exemple Soient $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$. La médiane M_n et la moyenne $\overline{X}_n$ sont (à peu près) non-biaisés pour μ mais $\mathrm{var}(M_n) > \mathrm{var}(\overline{X}_n)$. Lequel des estimateurs $\overline{X}_n$ et M_n de μ est préférable ? Et si des valeurs aberrantes peuvent apparaître ?

Biais et variance



- $\theta =$ "bulle centrale", supposée être la vraie valeur
- fléchettes rouges $=$ réalisations de $\widehat{\theta}$ qui estime θ

Exemple Soient $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, $\widehat{\mu} = \overline{X}_n$, et $\widehat{\sigma}_n^2 = n^{-1} \sum_{i=1}^n (X_i - \bar{X})^2$. On admet que $\text{var}_{\mu,\sigma^2}(\sum_{i=1}^n (X_i - \bar{X})^2) = 2\sigma^4(n-1)$. Trouver le biais et la variance de $\widehat{\mu}$. Trouver les valeurs de a qui minimisent le biais, la variance, et la EQM, pour $\widehat{\sigma}^a := a\widehat{\sigma}_n^2$.

Biais et variance : exemple

Exemple Soient $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, $\widehat{\mu} = \overline{X}_n$, et $\widehat{\sigma}_n^2 = n^{-1} \sum_{i=1}^n (X_i - \bar{X})^2$. On admet que $\text{var}_{\mu,\sigma^2}(\sum_{i=1}^n (X_i - \bar{X})^2) = 2\sigma^4(n-1)$. Trouver le biais et la variance de $\widehat{\mu}$. Trouver les valeurs de a qui minimisent le biais, la variance, et la EQM, pour $\widehat{\sigma}^a := a\widehat{\sigma}_n^2$.

Retour sur le maximum de vraisemblance

- Soient $x_1, \ldots, x_n$ des données supposées être une réalisation d'un échantillon aléatoire $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x; \theta)$ (densité / masse). La **vraisemblance** pour θ est la fonction

$$L(\theta) = f(x_1; \theta) \times f(x_2; \theta) \times \cdots \times f(x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta).$$

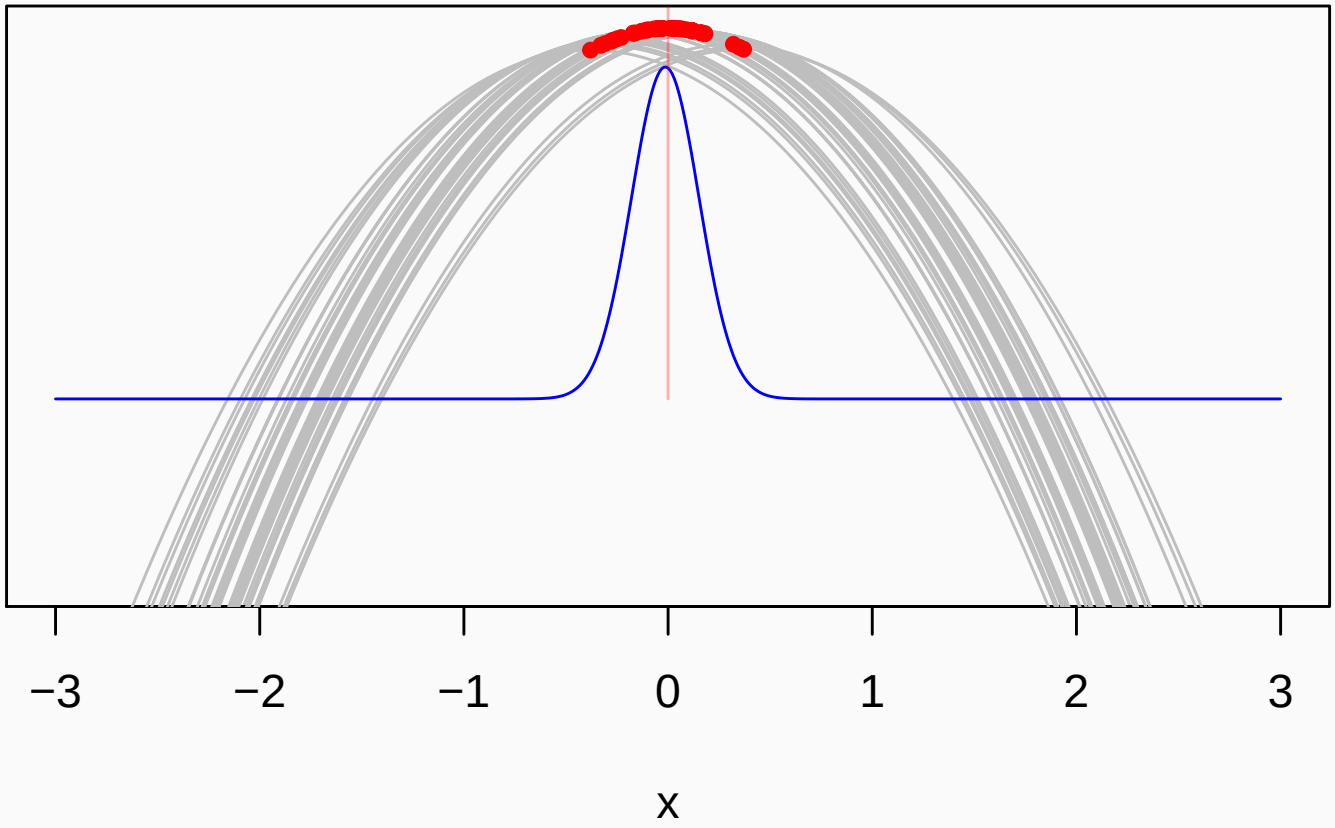
- $\widehat{\theta}_{\mathrm{ML}}$ satisfait

$$L(\widehat{\theta}_{\mathrm{ML}}) \geq L(\theta) \quad \text{pour tout } \theta$$

- **Interprétation** : dans le cas discret, on maximise $\Pr_\theta(X_1 = x_1, \ldots, X_n = x_n)$
- Il est plus facile de maximiser $\ell(\theta) := \log L(\theta)$, souvent en résolvant $d\ell(\theta)/d\theta = 0$, et vérifiant qu'il s'agit bien d'un maximum

Exemple $x_1, \ldots, x_n$ réalisations d'une loi Poiss(λ) avec $\lambda > 0$

La vraisemblance est une fonction aléatoire



Fonctions de vraisemblances correspondantes à 50 échantillons de taille $n = 40$. La vraie valeur est $\theta = 0$ et les 50 maxima sont en rouge.

Distribution asymptotique de $\widehat{\theta}_{\mathrm{ML}}$

- Dans des modèles “réguliers" (normal, exponentiel, Poisson, ...) on a $\ell'(\widehat{\theta}_{\mathrm{ML}}) = 0$
- Comme $\ell(\theta)$ est une somme de variables aléatoires iid, le théorème central limite s'applique
- Grâce à la méthode delta $\widehat{\theta}_{\mathrm{ML}} \overset{\mathrm{app}}{\sim} \mathcal{N}(\theta, 1/I_n(\theta))$ avec **l'information de Fisher**

$$I_n(\theta) = -\mathbb{E}_\theta(\ell''(\theta)) = \mathbb{E}_\theta(J_n(\theta)), \qquad J_n(\theta) = -\ell''(\theta)$$

  $1/I_n(\theta)$ est la **variance asymptotique** de $\widehat{\theta}_{\mathrm{ML}}$, $I_n(\theta)$ est la courbure
- **L'information observée** est $J_n(\widehat{\theta}_{\mathrm{ML}})$
- **Attention !** Certains modèles, tel que $U[0, \theta]$, ne sont pas réguliers

Exemple $x_1, \ldots, x_n$ réalisations d'une loi $\exp(\lambda)$ avec densité $\lambda e^{-\lambda x} I(x \geq 0)$, $\lambda > 0$

Exemple

Exemple $x_1, \ldots, x_n$ réalisations d'une loi $\exp(\lambda)$ avec densité $\lambda e^{-\lambda x} I(x \geq 0)$, $\lambda > 0$

3.2 Estimation par intervalle

Les intervalles de confiance

Un élément clé de la statistique est de donner une idée de l'incertitude d'un constat

Soit θ un paramètre inconnu, et soit $\tilde{\theta}_n = 1$ une estimation de θ basée sur $y_1, \ldots, y_n$:

- si $n = 10^5$ on est beaucoup plus sûr que $\theta \approx \tilde{\theta}_n$ que si $n = 10$
- pour exprimer ceci on aimerait donner un intervalle qui serait plus large quand $n = 10$ que quand $n = 10^5$, pour expliciter l'incertitude liée à $\tilde{\theta}_n$

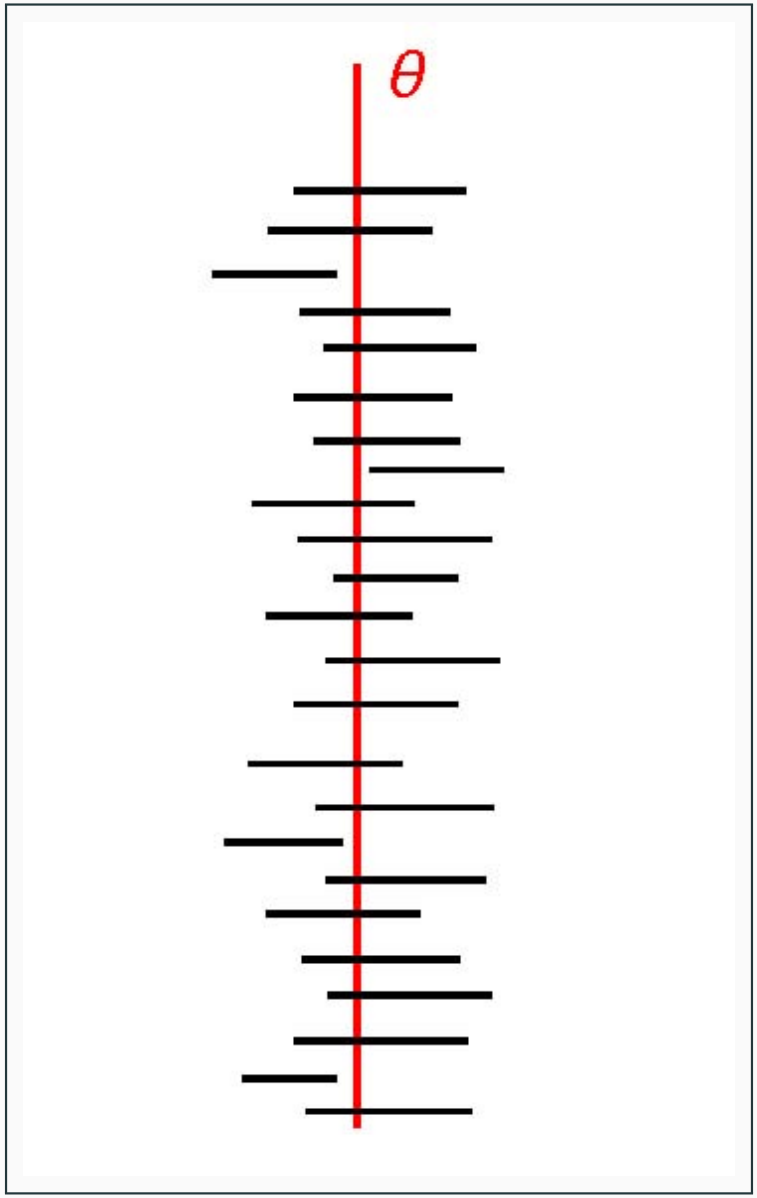
Définition: Soient $Y \equiv Y_1, \ldots, Y_n$ des données issues d'une loi F de paramètre $\theta \in \mathbb{R}$. Un **intervalle de confiance** (IC, 'confidence interval' en anglais) (L_n, U_n) pour θ est une statistique sous forme d'intervalle qui contient θ avec une probabilité spécifiée $1 - \alpha$

$$\Pr_\theta(L_n \leq \theta \leq U_n) = 1 - \alpha \qquad \forall \theta$$

- $1 - \alpha \in (0, 1)$ est le **niveau**, souvent $\alpha \in \{0.05, 0.01, 0.1\}$
- Les bornes $L_n = L_n(Y)$, $U_n = U_n(Y)$ sont des statistiques et non pas des inconnus
- Un IC **bilatéral**, de la forme (L_n, U_n), est le plus souvent utilisé
- Un IC **unilatéral à droite** est $(-\infty, U_n]$ tel que $\Pr_\theta(U_n \geq \theta) = 1 - \alpha$
- Un IC **unilatéral à gauche** est $[L_n, \infty)$ tel que $\Pr_\theta(L_n \leq \theta) = 1 - \alpha$

Interprétation d'un intervalle de confiance

- (L_n, U_n) est un intervalle aléatoire qui contient θ avec probabilité $(1 - \alpha)$
- Si on répète l'expérience avec d'autres données, on aura un autre intervalle de confiance
- Nous ne savons pas si notre IC contient θ, mais cette procédure nous fournit une garantie statistique que cet événement a une probabilité $(1 - \alpha)$



Constuire un IC dans le cas normal

- Soient $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, 1)$, et

$$\overline{Y}_n = \frac{1}{n}\sum_{j=1}^{n} Y_j$$

la moyenne de l'échantillon. On peut montrer que $\overline{Y}_n \sim \mathcal{N}(\mu, 1/n)$. Donc

$$Z_n = n^{1/2}(\overline{Y}_n - \mu) \sim \mathcal{N}(0, 1)$$

a une distribution qui ne dépend pas de μ. On obtient

$$\Pr_\mu\left(\overline{Y}_n - \frac{z_{1-\alpha/2}}{\sqrt{n}} \leq \mu \leq \overline{Y}_n + \frac{z_{1-\alpha/2}}{\sqrt{n}}\right) = 1 - \alpha$$

avec $z_\beta = \Phi^{-1}(\beta)$ les quantiles de la loi $\mathcal{N}(0, 1)$

- Si $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, et **on connaît** σ^2, alors

$$\frac{Z_n}{\sigma} = \frac{n^{1/2}(\overline{Y}_n - \mu)}{\sigma} \sim \mathcal{N}(0, 1)$$

et l'intervalle de confiance pour μ est

$$\left[\overline{Y}_n - \frac{z_{1-\alpha/2}}{\sqrt{n}}\sigma, \overline{Y}_n + \frac{z_{1-\alpha/2}}{\sqrt{n}}\sigma\right]$$

- La largeur $\downarrow$ avec n, $\uparrow$ avec σ et avec le niveau $(1-\alpha)$, car $\Pr_\theta(L_n \leq \theta \leq U_n) \uparrow$

Exemple

Exemple On suppose que la résistance Y d'un certain type d'équipements électriques est distribuée selon une loi normale avec $\sigma = 0.12$ ohm.

Un échantillon de taille $n = 64$ a donné comme moyenne la valeur $\bar{y}_n = 5.34$ ohm.

Trouver un intervalle de confiance pour μ au niveau 95%.

Intervalle de Student

- Soient $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$ avec σ^2 **inconnue**
- $\left[\overline{Y}_n - \frac{z_{1-\alpha/2}}{\sqrt{n}}\sigma, \overline{Y}_n + \frac{z_{1-\alpha/2}}{\sqrt{n}}\sigma\right]$ n'est plus un intervalle de confiance
- On suppose $n > 1$ et estime σ^2 par

$$S_n^2 = \frac{1}{n-1}\sum_{j=1}^{n}(Y_j - \overline{Y}_n)^2$$

- On remplace σ^2 par S_n : soit

$$T_n = \frac{Z_n}{S_n} = n^{1/2}\frac{\overline{Y}_n - \mu}{S_n} = \frac{Z_n/\sigma}{\sqrt{S_n^2/\sigma^2}}$$

- Nominateur et dénominateur indép, leurs lois ne dépendent pas de (μ, σ^2)
- T_n suit une **loi de Student avec $n-1$ degrés de liberté**, $T_n \sim t_{n-1}$, qui ne dépend de $\theta = (\mu, \sigma^2)$. C'est une loi symétrique, on dénote les quantiles par $t_{n-1,\beta}$
- L'intervalle de confiance qui en résulte est

$$\left[\overline{Y}_n - \frac{t_{n-1,1-\alpha/2}}{\sqrt{n}}S_n, \overline{Y}_n + \frac{t_{n-1,1-\alpha/2}}{\sqrt{n}}S_n\right]$$

- La largeur $\downarrow$ avec n, $\uparrow$ avec S_n et avec le niveau $(1-\alpha)$

Exemple

Exemple Pour déterminer le point de fusion μ d'un certain alliage, on a procédé à $n = 9$ observations qui ont donné une moyenne $\overline{y}_n = 1040^\circ$ avec $s_n = 16^\circ$.

Trouver un intervalle de confiance pour μ à niveau 95%.

Intervalles de confiance pour μ dans le cas normal : résumé

Soient $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$ avec μ inconnu

- Pour σ **connue** $n^{1/2}\frac{\overline{Y}_n - \mu}{\sigma} \sim \mathcal{N}(0,1)$ et on a les IC de niveau $1-\alpha$

$$\left[\overline{Y}_n - \frac{z_{1-\alpha/2}}{\sqrt{n}}\sigma, \overline{Y}_n + \frac{z_{1-\alpha/2}}{\sqrt{n}}\sigma\right]$$

$$\left[\overline{Y}_n - \frac{z_{1-\alpha}}{\sqrt{n}}\sigma, \infty\right] \qquad \left[-\infty, \overline{Y}_n + \frac{z_{1-\alpha}}{\sqrt{n}}\sigma\right]$$

  avec $z_\beta = \Phi^{-1}(\beta)$ les quantiles de la loi $\mathcal{N}(0,1)$

- Pour σ **inconnue** (et $n > 1$) $n^{1/2}\frac{\overline{Y}_n - \mu}{S_n} \sim t_{n-1}$ et on a les IC de niveau $1-\alpha$

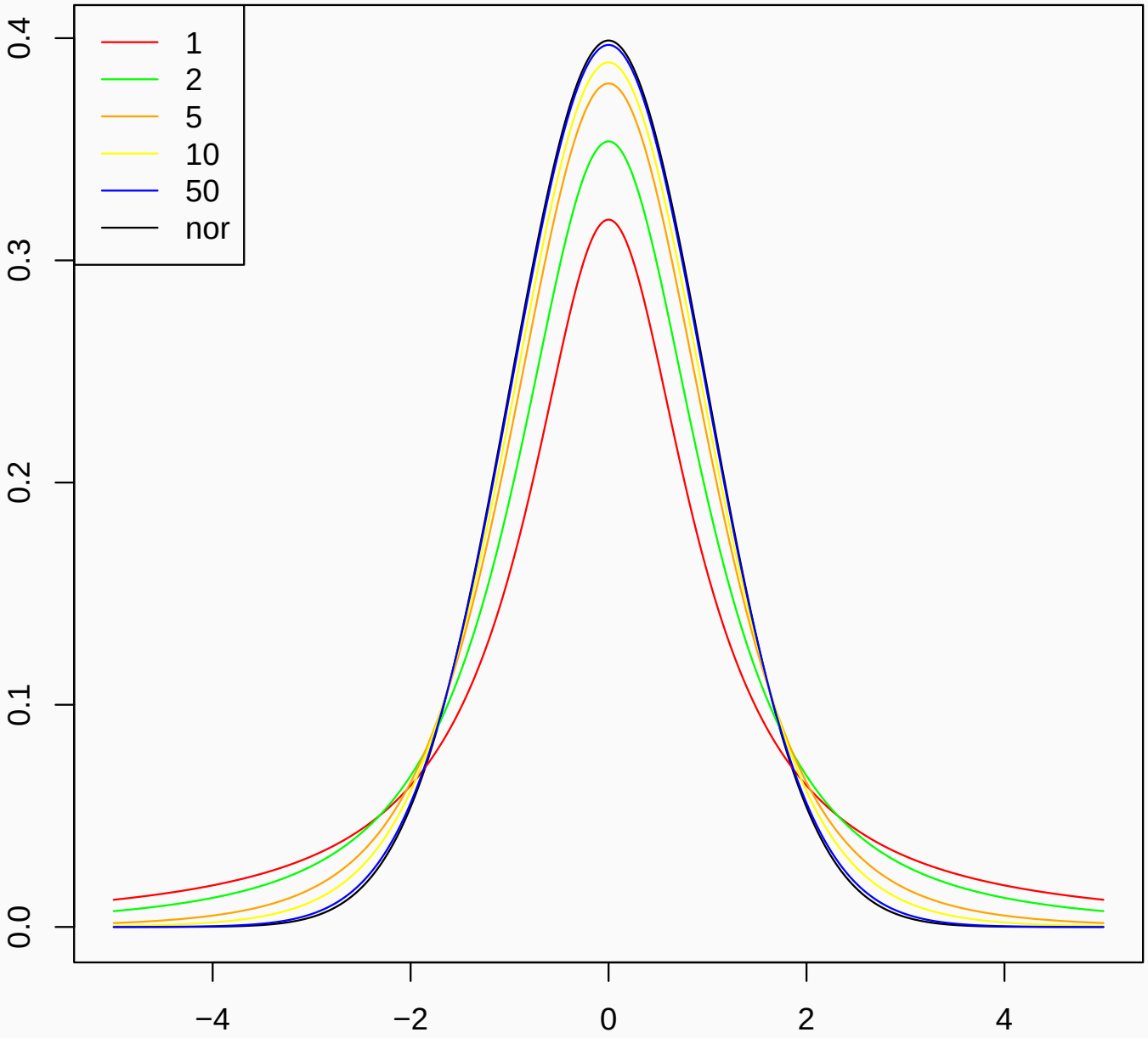
$$\left[\overline{Y}_n - \frac{t_{n-1,1-\alpha/2}}{\sqrt{n}}S_n, \overline{Y}_n + \frac{t_{n-1,1-\alpha/2}}{\sqrt{n}}S_n\right]$$

$$\left[\overline{Y}_n - \frac{t_{n-1,1-\alpha}}{\sqrt{n}}S_n, \infty\right] \quad \text{ou} \quad \left[-\infty, \overline{Y}_n + \frac{t_{n-1,1-\alpha}}{\sqrt{n}}S_n\right]$$

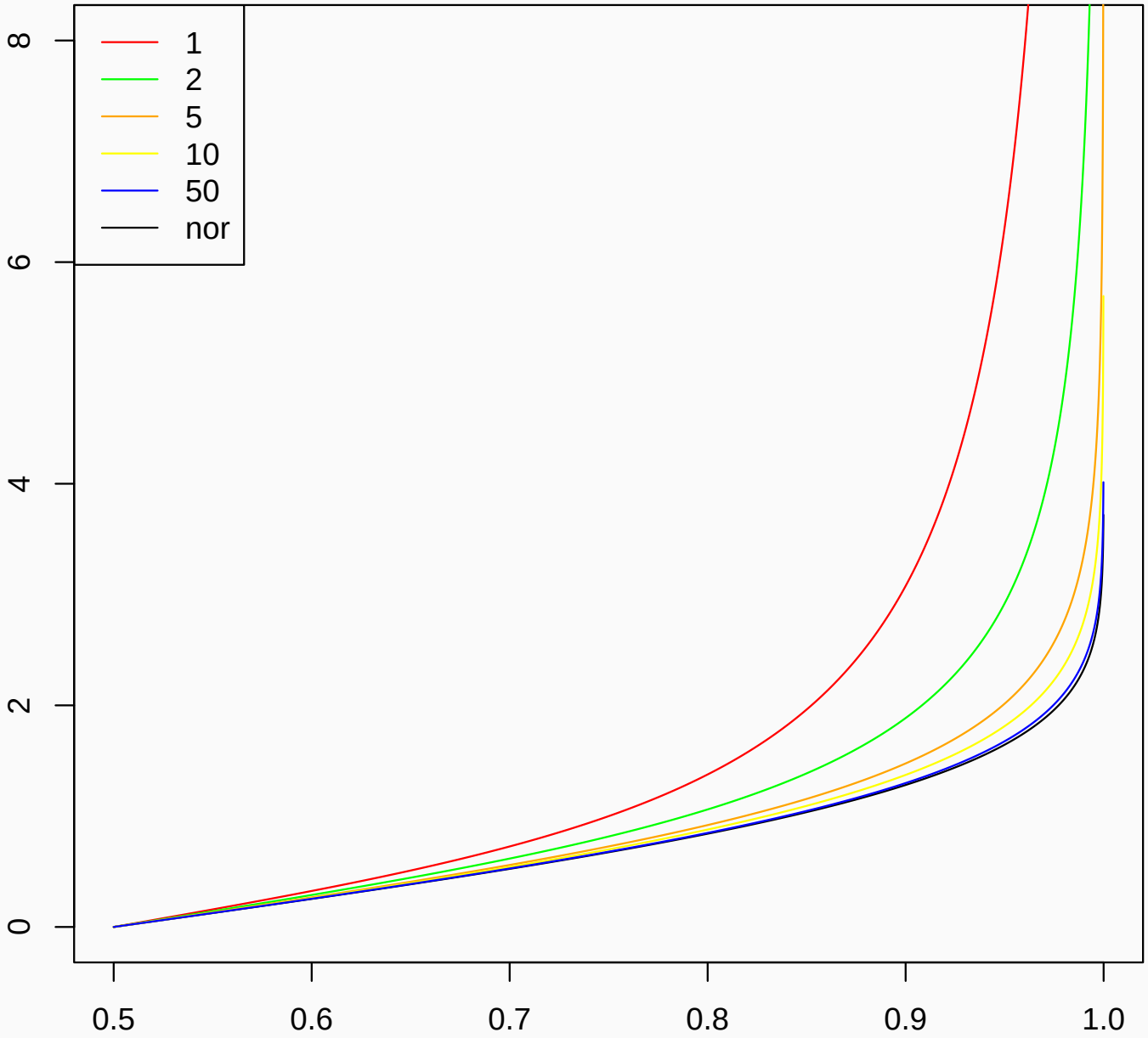
  avec t_{n-1} la **loi de Student avec $n-1$ degrés de liberté**, $t_{n-1,\beta}$ sont les quantiles de cette distribution, S_n définie à la diapositive 186.

t_{n-1} symétrique comme $\mathcal{N}(0,1)$, mais quantiles plus grands : $|t_{n-1,\beta}| > |z_\beta|$, $\beta \neq \frac{1}{2}$

Densités de t_k et $\mathcal{N}(0,1)$



Quantiles de t_k et $\mathcal{N}(0,1)$



Pivots

- Une fonction $T(Y_1, \ldots, Y_n, \theta)$ dont la loi est connue et ne dépend pas de θ s'appelle un **pivot**
- **Attention !** Ce n'est pas une statistique car c'est une fonction de θ
- Si $a \leq b$,

$$\alpha_1 = \Pr(T < a), \quad \alpha_2 = \Pr(T > b), \quad \alpha_1 + \alpha_2 = \alpha \in [0, 1]$$

(souvent $\alpha_1 = \alpha_2 = \alpha/2$) alors

$$\Pr_\theta(a \leq T \leq b) = \Pr_\theta(T \leq b) - \Pr_\theta(T < a) = (1 - \alpha_2) - \alpha_1 = 1 - \alpha$$

- Si on peut isoler θ, on peut trouver des variables aléatoires L et U telles que

$$\Pr_\theta(L \leq \theta \leq U) = 1 - \alpha$$

Exemples T_n (diapositive 186), Z_n (diapositive 183).

Pivot : cas uniforme

Exemple Soient $Y_1, \ldots, Y_n \sim U(0, \theta)$, $M_n = \max(Y_i)$. Alors $T(Y_1, \ldots, Y_n, \theta) = M_n/\theta$ est un pivot.

Intervalles de confiance approximatifs

- En pratique la plupart des intervalles de confiance se basent sur des approximations fournies par le théorème central limite, étant de la forme
$$(L_n, U_n) = (\widehat{\theta}_n - \sqrt{V_n} z_{1-\alpha/2}, \widehat{\theta}_n + \sqrt{V_n} z_{1-\alpha/2}),$$
où V_n est une estimateur de $\mathrm{var}_\theta(\widehat{\theta}_n)$ dont la racine s'appelle **erreur type (standard error)** de $\widehat{\theta}_n$. Sa réalisation $v_n^{1/2}$ est aussi appelée erreur type
- L'intervalle de confiance est approximatif dans le sens que
$$\Pr_\theta(L_n \leq \theta \leq U_n) \to 1 - \alpha, \qquad n \to \infty$$
- Dans des modèles réguliers, la variance asymptotique de $\widehat{\theta}_{\mathrm{ML}}$ (voir diapositive 178) est $1/I_n(\theta)$, estimée par $1/J_n(\widehat{\theta}_{\mathrm{ML}})$ et donc
$$(L_n, U_n) = \left(\widehat{\theta}_{\mathrm{ML}} - z_{1-\alpha/2}/\sqrt{J_n(\widehat{\theta}_{\mathrm{ML}})}, \widehat{\theta}_{\mathrm{ML}} + z_{1-\alpha/2}/\sqrt{J_n(\widehat{\theta}_{\mathrm{ML}})}\right)$$

Exemples

Exemple Soient $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \text{Poiss}(\lambda)$, $\lambda > 0$ inconnu. Trouver une erreur type pour $\widehat{\lambda}_{\text{ML}}$, et ainsi donner un intervalle de confiance approximatif de niveau 90% pour λ.

Exemples

Exemple Soient $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} U(0, \theta)$ avec θ inconnu et $\overline{Y}_n = n^{-1} \sum_{j=1}^{n} Y_j$. Utiliser le théorème central limite pour trouver un intervalle de confiance approximatif de niveau 95% pour θ.

3.3 Tests statistiques

Démarche scientifique

Toute **démarche scientifique** s'effectue selon le même schéma. Afin d'analyser la plausibilité d'une théorie, on itère les étapes suivantes :

- Enoncé d'une hypothèse (théorie) pouvant être contredite par des données.
- Récolte de données
- Comparaison des données avec les prédictions/implications de l'hypothèse.
- Non-rejet, rejet ou modification éventuelle de l'hypothèse.

Dans un cadre statistique, en supposant que l'on dispose d'un modèle pour le phénomène étudié, on itère les étapes suivantes :

- Enoncé d'une hypothèse (typiquement sur les paramètres du **modèle statistique**). Cette hypothèse peut être contredite par des données (via une statistique, appelée **statistique de test**).
- Récolte de données
- **Rejet (ou non) de l'hypothèse** à partir de la comparaison entre les données et les implications de l'hypothèse. En cas d'écart, à partir de quel seuil juge-t-on cet écart **significatif**, i.e., suffisamment important pour justifier le rejet de l'hypothèse ?

Exemple

Exemple Question : L'alcool ralentit-il les réflexes ?

Afin d'étudier l'effet de l'alcool sur les réflexes, on fait passer à 14 sujets un test de dextérité avant et après qu'ils aient consommé 100 ml de vin. Leurs temps de réaction (en ms) avant et après sont donnés dans le tableau suivant :

Sujet	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Avant	57	54	62	64	71	65	70	75	68	70	77	74	80	83
Après	55	60	68	69	70	73	74	74	75	76	76	78	81	90

Cadre statistique : [1] Hypothèse nulle et alternative

Etant donné un modèle statistique (de densité $f(x;\theta)$), nous voulons choisir entre deux théories concurrentes à propos du paramètre θ. Ces dernières forment une paire d'hypothèses :

$$H_0 : \text{l'hypothèse } \underline{\text{nulle}} \qquad \text{vs} \qquad H_1 : \text{l'hypothèse } \underline{\text{alternative}}.$$

Exemple. Dans une population décrite par la loi $\mathcal{N}(\mu, \sigma^2)$, nous pouvons former des hypothèses sur μ comme suit :

$$\underbrace{\left\{ \begin{array}{l} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{array} \right\}}_{\text{paire bilatérale}} \quad \text{ou} \quad \underbrace{\left\{ \begin{array}{l} H_0 : \mu = \mu_0 \\ H_1 : \mu > \mu_0 \end{array} \right\} \quad \text{ou} \quad \left\{ \begin{array}{l} H_0 : \mu = \mu_0 \\ H_1 : \mu < \mu_0 \end{array} \right\}}_{\text{paires unilatérales}}.$$

Cadre statistique : [2] Statistique de test

Comment choisir entre les deux hypothèses ?

- Nous tirons un échantillon $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x;\theta)$ tiré de la population. Comment l'utiliser pour prendre notre décision ?
- Nous choisissons une statistique $T = T_n = g(X_1, ..., X_n)$ qui a tendance à prendre des valeurs "typiques" sous l'hypothèse nulle H_0 (i.e., si H_0 est vraie) et "extrêmes" (dans la direction de l'hypothèse alternative H_1) sous H_1
- Ainsi, si on observe une valeur plutôt "extrême" ("extrême" dans la direction de l'hypothèse alternative H_1) de T, nous avons de l'évidence contre H_0.

Notre règle de décision est donc :

- Rejeter H_0 si la valeur observée de T est assez extrême (au-delà d'une valeur critique à déterminer).
- Ne pas rejeter H_0 si la valeur observée de T n'est pas assez extrême.

Exemple : paire bilatérale

Soient $X_1, ..., X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, où σ^2 est inconnue, et considérons la paire d'hypothèses :

$$\left\{ \begin{array}{l} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{array} \right\}.$$

On parle de paire bilatérale car $\mu \neq \mu_0$ est équivalent à $\mu < \mu_0$ ou $\mu > \mu_0$.

Considérons la statistique de test $T = \dfrac{\bar{X} - \mu_0}{S/\sqrt{n}}$.

- Si H_0 est vraie, alors $T \sim t_{n-1}$ (donc si H_0 est vraie, T prend typiquement des valeurs proches de 0).
- Compte tenu de H_1, nous considérons donc les valeurs de T comme "extrêmes" si elles sont "éloignées" de 0. Notons qu'ici, la notion d'"extrême" dans la direction de l'hypothèse alternative H_1 signifie une valeur "extrême" de la valeur absolue de T.
- Nous allons rejeter H_0 si $|T|$ est suffisamment élevée, i.e., $|T| > v^*$, où $v^* > 0$ est une valeur critique à déterminer.

Exemple : paire unilatérale

Soient $X_1, ..., X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, où σ^2 est inconnu, et considérons la paire d'hypothèses :

$$\left\{ \begin{array}{l} H_0 : \mu = \mu_0 \\ H_1 : \mu < \mu_0 \end{array} \right\}.$$

Considérons la statistique de test $T_n = \dfrac{\bar{X} - \mu_0}{S/\sqrt{n}}$.

- Si H_0 est vraie, alors $T \sim t_{n-1}$.
- Compte tenu de H_1, nous considérons donc les valeurs de T comme "extrêmes" si elles sont fortement négatives. Donc ici, la notion d'"extrême" dans la direction de l'hypothèse alternative H_1 signifie une valeur "extrême" de $|\min(T, 0)|$ et non de $|T|$.
- Nous allons donc rejeter H_0 si T est suffisamment négative, i.e., $T < v_*$, où $v_* < 0$ est la valeur critique à déterminer.

Cadre statistique : [3] Significativité statistique

Choix de la valeur critique (par exemple v^* et v_*) : Comment définir suffisamment élevée ou suffisamment négative. En d'autres termes, quelle ampleur est considérée comme significative ?

Pour répondre à cette question, il faut considérer les deux types d'erreurs que l'on peut commettre lorsque l'on se décide en faveur de l'une des hypothèses :

Décision \ Vérité	H_0 vraie	H_0 fausse
Non-rejet de H_0		Erreur de type II
Rejet de H_0	Erreur de type I	

Erreur de type I (**faux positif**) considérée plus grave que l'erreur de type II (faux négatif) — filtre de spam

Cadre statistique : [3] Significativité statistique

- On ne peut pas contrôler les deux erreurs à la fois :

$$\Pr(\text{erreur type I}) \text{ petite}$$
$$\Updownarrow$$
$$\text{rejet uniquement pour des valeurs très extrêmes}$$
$$\Updownarrow$$
$$\text{difficile de rejetter}$$
$$\Updownarrow$$
$$\Pr(\text{erreur type II}) \text{ grande}$$

- L'asymétrie entre la gravité des erreurs nous aide à choisir H_0 et H_1
- On va contrôler l'erreur de type I, qui est plus grave
- Parfois on choisit H_0 par convenience mathématique (de sorte que mathématiquement il serait plus facile de contrôler l'erreur de type I)

Cadre statistique : [3] Significativité statistique

- Nous choisissons la valeur maximale que l'on tolère pour la probabilité d'erreur de type I (éventuellement en tenant compte de l'avis d'un spécialiste). Cette quantité est notée α et appelée **niveau/seuil de significativité du test** ; $\alpha \in (0, 1)$. On choisit généralement une valeur faible pour α. Typiquement, $\alpha = 0.1, 0.05, 0.01, 0.001$; le plus souvent, $\alpha = 0.05$.

- La valeur critique est déterminée de manière à ce que

$$\Pr_{H_0}[\text{Rejet de } H_0] = \alpha.$$

- Ainsi, la **valeur critique** est telle que

$$\Pr_{H_0}[|T| > \text{ valeur critique}] = \alpha \quad \text{(cas bilatéral)},$$

$$\Pr_{H_0}[T < \text{ valeur critique}] = \alpha \quad \text{ou}$$

$$\Pr_{H_0}[T > \text{ valeur critique}] = \alpha \quad \text{(cas unilatéral)}.$$

- Les probabilités sont sous l'hypothèse que H_0 **est vraie** !

Cadre statistique : [3] Significativité statistique

Exemple, paire bilatérale : Soient $X_1, ..., X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, où σ^2 est inconnu, et considérons la paire $H_0 : \mu = \mu_0$ contre $H_1 : \mu \neq \mu_0$.

Nous allons rejeter H_0 si $|T| = \left| \dfrac{\bar{X} - \mu_0}{S/\sqrt{n}} \right|$ est assez large, c'est à dire $|T| > v^*$.

Soit α le niveau de significativité. La valeur critique v^* satisfait

$$\Pr_{H_0}[|T| > v^*] = \alpha,$$

i.e.,

$$\Pr_{H_0}[T < -v^* \text{ ou } T > v^*] = \alpha.$$

Quand H_0 est vraie $T \sim t_{n-1}$. On doit donc choisir

$$v^* = t_{n-1,1-\alpha/2},$$

où $t_{n-1,1-\alpha/2}$ est le $(1 - \alpha/2)$ quantile de la loi de Student t_{n-1}.

Cadre statistique : [4] La p-valeur

Au lieu d'utiliser des valeurs critiques pour choisir entre H_0 et H_1, nous pouvons utiliser une autre approche, basée sur la notion de **p-valeur**.

- La p-valeur (notée p_{obs}) est la probabilité d'obtenir une valeur de la statistique de test au moins aussi élevée (élevée dans la direction de H_1) que celle que nous avons observée si H_0 était vraie.
- Supposons que la réalisation de la statistique de test sur nos données est $T = t_{\text{obs}}$. Alors :
  - Cas bilatéral : $p_{\text{obs}} = \Pr_{H_0}[|T| > |t_{\text{obs}}|]$,
  - Cas unilatéral à gauche : $p_{\text{obs}} = \Pr_{H_0}[T < t_{\text{obs}}]$,
  - Cas unilatéral à droite : $p_{\text{obs}} = \Pr_{H_0}[T > t_{\text{obs}}]$.
- Petite valeurs de p_{obs} s'opposent à H_0 car elles démontrent que la realité observée serait très improbable si l'hypothèse nulle H_0 était vraie.
- Cas bilatéral (207) : $p_{obs} = 2(1 - F_{t_{n-1}}(|t_{obs}|))$, où $F_{t_{n-1}}$ est la fonction de répartition de la loi de Student t_{n-1}.

On peut utiliser la p-valuer pour faire un test d'hypothèse :

$$\boxed{\text{rejetter } H_0 \qquad \Longleftrightarrow \qquad p_{obs} < \alpha}$$

Exemple bilatérale (207) $p_{obs} = 2(1 - F_{t_{n-1}}(t_{obs}))$ donc

$$p_{obs} < \alpha \iff F_{t_{n-1}}(t_{obs}) > 1 - \alpha/2 \iff t_{obs} > t_{n-1,1-\alpha/2}$$

De manière générale, l'approche de la p-valeur est équivalente à l'approche des valeurs critiques. Cependant, la p-valeur p_{obs} fournit une information plus facilement interprétable que la valeur t_{obs}. Il s'agit d'une mesure de l'évidence contre H_0 contenue dans les données.

Attention : la p-valeur **n'est pas** la probabilité que H_0 soit vraie.

Résumé : les éléments d'un test

A Une hypothèse nulle H_0 à tester contre une hypothèse alternative H_1.

B Une statistique de test T, choisie de telle sorte que des valeurs "extrêmes" de T (en direction de H_1) suggèrent que H_0 est fausse. La valeur observée de T est t_{obs}.

C Un niveau de significativité α, qui la probabilité d'erreur de type I (rejet de H_0 quand H_0 est vraie) maximale que nous allons tolérer.

D1 Des valeurs critiques, telles que quand T tombe au-delà de ces valeurs, nous rejetons H_0 en faveur de H_1. Les valeurs critiques sont choisies pour respecter le niveau de significativité α.

Au lieu de D1, nous pouvons utiliser l'approche équivalente D2 :

D2 Une valeur p_{obs} donnant la probabilité d'observer une valeur de T aussi élevée que t_{obs} sous H_0. On rejette alors H_0 en faveur de H_1 quand $p_{\text{obs}} < \alpha$.

Exemple

Exemple On a contrôlé 10 compteurs d'électricité nouvellement fabriqués et obtenu les valeurs suivantes (en MW) :
983 1002 998 996 1002 983 994 991 1005 986.
On suppose qu'il s'agit de réalisation d'un échantillon iid d'une loi normale. On aimerait savoir s'il y a un écart entre la moyenne attendue de 1000 MW et la moyenne réelle des compteurs qui sortent de la fabrication. Nous avons obtenu $\bar{x} = 994 < 1000$. S'agit-il d'un hasard ou une faute de production ?

On va prendre $\alpha = 5\%$.

Solution Exemple 212

Supposons que nos observations $x_1, \ldots, x_n$ soient des réalisations de variables aléatoires $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, avec σ^2 inconnu. On veut tester : $H_0 : \mu = \mu_0$ contre $H_1 : \mu \neq \mu_0$, où $\mu_0 = 1000$. On prend comme statistique de test

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t_{n-1} \text{ sous } H_0 : \mu = \mu_0.$$

Dans notre cas $n = 10$, $\mu_0 = 1000$, $\bar{x} = 994$, et

$$s^2 = \frac{1}{9}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{9}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right) = 64.88,$$

donc $t_{\text{obs}} = -2.35$.

On rejette H_0 si et seulement si $|t_{\text{obs}}| > t_{n-1,1-\alpha/2}$ (cas bilatéral). , $t_{n-1,1-\alpha/2} = 2.262$ (voir les tables), et comme $t_{\text{obs}} = -2.35 < -2.262$, on rejette l'hypothèse H_0.

Intervalles de confiance et tests

- Soient $Y_1, \ldots, Y_n \sim \mathcal{N}(\mu, \sigma^2)$ avec μ, σ inconnus
- Considérons $H_0 : \mu = 0$ et $H_1 : \mu \neq 0$
- On rejette H_0 au niveau α si et seulement si

$$|T_n| = \left| \sqrt{n} \frac{\overline{Y}_n}{S_n} \right| > t_{n-1,1-\alpha/2}$$

- Intervalle de confiance (IC) de niveau $1 - \alpha$

$$\left[\overline{Y}_n - \frac{t_{n-1,1-\alpha/2}}{\sqrt{n}} S_n, \overline{Y}_n + \frac{t_{n-1,1-\alpha/2}}{\sqrt{n}} S_n \right]$$

- Cet intervalle contient zéro si et seulement si $|\overline{Y}_n| \leq t_{n-1,1-\alpha/2} S_n / \sqrt{n}$ et donc

$$\text{on rejette } H_0 \iff 0 \text{ n'est pas dans l'intervalle de confiance}$$

- De manière générale, rejet de $H_0^{\theta_0} : \theta = \theta_0$ est équivalent à $\{$l'IC ne contient pas $\theta_0\}$ si on se base sur la même statistique de test
- α petit $\iff$ difficile à rejetter $\iff$ IC de niveau $1 - \alpha$ large

Intervalles de confiance (IC) et tests

- Rejet $\iff$ $p_{obs} \leq \alpha$ $\iff$ $\theta_0 \notin$IC de niveau $1-\alpha$
- **IC** : α fixé, pour quels $\theta_0 \in \mathbb{R}$ $H_0^{\theta_0}$ n'est pas rejetée ?
- ***p*-valeur** : θ_0 fixé, pour quels $\alpha \in (0,1)$ $H_0^{\theta_0}$ est rejetée ?

3.4 Tests khi-deux

Le test khi-deux

- On se pose la question de l'adéquation d'une distribution théorique à des données
- Supposons que dans une expérience on observe n résultats différents avec des
  - **fréquences observées** dans k classes disjointes $o_1, \ldots, o_k$, alors que les
  - **fréquences théoriques** correspondantes sont $e_1, \ldots, e_k$,
  - où $\sum_{i=1}^{k} o_i = \sum_{i=1}^{k} e_i = n$
- On a H_0 : "les observations proviennent de la loi théorique spécifiée"
- Une mesure de l'écart entre les o_j et les e_j est donnée par la **statistique khi-deux**

$$T_n = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$$

- Si n est grand et les e_i ne sont pas trop petites (règle de pouce : $e_i \geq 5$ pour la plupart), alors $T_n \overset{\text{app}}{\sim} \chi^2_\nu$ sous H_0, où
  - χ^2_ν est la loi **khi-carré**, la loi de $\sum_{i=1}^{\nu} Z_i^2$ où $Z_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$
  - $\nu = k - 1$ si les e_i peuvent être calculés sans devoir estimer des paramètres inconnus
  - $\nu = k - 1 - c$ si les e_i sont calculés après avoir estimé c paramètres

Exemple

$$\text{rejet de } H_0 \text{ au niveau } \alpha \quad \Longleftrightarrow \quad t_{\text{obs}} = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i} > \chi^2_{\nu,1-\alpha}$$

Exemple $n = 60$ jets d'un dé ont donné la répartition suivante :

Valeur x_i	1	2	3	4	5	6	
Valeur o_i	8	10	9	16	13	4	60

Ici $k = 6$ et H_0 : "équilibre du dé" est équivalente au modèle

$$\Pr(X = x) = 1/6, \quad x \in \{1, \ldots, 6\}.$$

Exemple

Exemple L'intelligence (QI) X de $n = 1000$ personnes est testée :

Score X	$[0, 70)$	$[70, 85)$	$[85, 100)$	$[100, 115)$	$[115, 130)$	$[130, \infty)$
Nombre o_i	34	114	360	344	120	28

Est-il plausible que $X \sim \mathcal{N}(100, 15^2)$?

On a

$$e_i = n\Pr_{\mathcal{N}(100,15^2)}(a_i \leq X \leq b_i) = n\Pr\left(\frac{a_i - 100}{15} \leq \frac{X_i - 100}{15} \leq \frac{b_i - 100}{15}\right)$$

$$= n\Phi\left(\frac{b_i - 100}{15}\right) - n\Phi\left(\frac{a_i - 100}{15}\right)$$

Tableaux de contingence

Un **tableau de contingence** est une classification de n objets ou individus selon plusieurs critères

- Une question fondamentale concerne **l'indépendance** des critères
- Supposons qu'on observe deux caractères A (h classes) et B (k classes) sur chacun de n individus, donnant le **tableau de contingence** suivant :

	B						
A	1	2	$\cdots$	j	$\ldots$	k	Σ
1	n_{11}	n_{12}	$\cdots$	n_{1j}	$\cdots$	n_{1k}	$n_{1.}$
2	n_{21}	n_{22}	$\cdots$	n_{2j}	$\cdots$	n_{2k}	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	n_{i1}	n_{i2}	$\cdots$	n_{ij}	$\cdots$	n_{ik}	$n_{i.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
h	n_{h1}	n_{h2}	$\cdots$	n_{hj}	$\cdots$	n_{hk}	$n_{h.}$
Σ	$n_{.1}$	$n_{.2}$	$\cdots$	$n_{.j}$	$\cdots$	$n_{.k}$	$n_{..} = n$

Indépendance

- Soit n_{ij} le nombre de personnes tombant dans la classe i du caractère A et dans la classe j du caractère B, et soit

$$n_{i\cdot} = \sum_{j=1}^{k} n_{ij}, \quad n_{\cdot j} = \sum_{i=1}^{h} n_{ij}, \quad i \in \{1, \ldots, h\}, j \in \{1, \ldots, k\}$$

- On désire tester H_0 : A **et** B **sont indépendants**. Dans ce cas

$$\Pr(A = i, B = j) = \Pr(A = i) \times \Pr(B = j), \quad i \in \{1, \ldots, h\}, j \in \{1, \ldots, k\},$$

et les probabilités empiriques sont

$$\widehat{\Pr(A = i)} = \frac{n_{i\cdot}}{n}, \quad \widehat{\Pr(B = j)} = \frac{n_{\cdot j}}{n}$$

Ainsi, sous H_0, l'(i, j)ième élément du tableau de contingence est

$$E_{ij} = n \times \Pr(A = i, B = j) = n \times \Pr(A = i) \times \Pr(B = j)$$

qu'on va estimer par

$$e_{ij} = n \frac{n_{i\cdot}}{n} \frac{n_{\cdot j}}{n} = \frac{n_{i\cdot} n_{\cdot j}}{n}$$

Calcul sous H_0

- Sous H_0 et pour n grand, la statistique T_n suit une distribution χ^2_ν avec $\nu = (h-1)(k-1)$, car on a dû estimer $c = (h-1)+(k-1)$ probabilités, et

$$hk - 1 - c = kh - 1 - (h-1) - (k-1) = (h-1)(k-1)$$

- Pour tester à un niveau de significativité α fixé, on rejette H_0 si $t_{\text{obs}} > \chi^2_{(h-1)(k-1),1-\alpha}$, sinon on ne rejette pas H_0

Exemple On a relevé parmi 95 personnes la couleur de leurs yeux (caractère A) et celle de leurs cheveux (caractère B) et on a obtenu les résultats suivants :

	B		
A	Cheveux clairs	Cheveux sombres	
Yeux bleus	$n_{11} = 32$	$n_{12} = 12$	$n_{1.} = 44$
Yeux bruns	$n_{21} = 14$	$n_{22} = 22$	$n_{2.} = 36$
Autres	$n_{31} = 6$	$n_{32} = 9$	$n_{3.} = 15$
Σ	$n_{.1} = 52$	$n_{.2} = 43$	$n = 95$

On désire tester si la couleur des cheveux est indépendante de celle des yeux

Donc on a H_0 : indépendance entre couleur des cheveux et couleur des yeux

Régularité (non-examinable)

Les conditions de régularité sont compliquées. Elles sont fausses le plus souvent quand

- un des paramètres est discret
- le support de $f(y;\theta)$ dépend de θ
- le vrai θ se trouve sur une borne des valeurs possibles

Elles sont satisfaites dans la grande majorité des cas rencontrés en pratique

Exemple Soient $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} U(0,\theta)$, trouver la vraisemblance $L(\theta)$ et le $\widehat{\theta}_{\text{ML}}$. Montrer que la loi limite de $n(\theta - \widehat{\theta}_{\text{ML}})/\theta$ quand $n \to \infty$ est exp(1). Discuter.

Preuve (non-examinable)

- Les fonctions de densité et de répartition de y_j sont

$$f(y;\theta) = \theta^{-1} I(0 \le y \le \theta), \quad F(y) = y/\theta, \quad 0 \le y \le \theta.$$

- L'indépendance donne

$$L(\theta) = \prod \theta^{-1} I(Y_j < \theta) = \theta^{-n} I(\max Y_j \le \theta), \quad \theta > 0,$$

qui est maximisée au point $\widehat{\theta}_{\mathrm{ML}} = M_n = \max Y_j$

- On a

$$\Pr(M_n \le x) = \prod_{j=1}^{n} \Pr(Y_j \le x) = (x/\theta)^n, \quad 0 \le x \le \theta$$

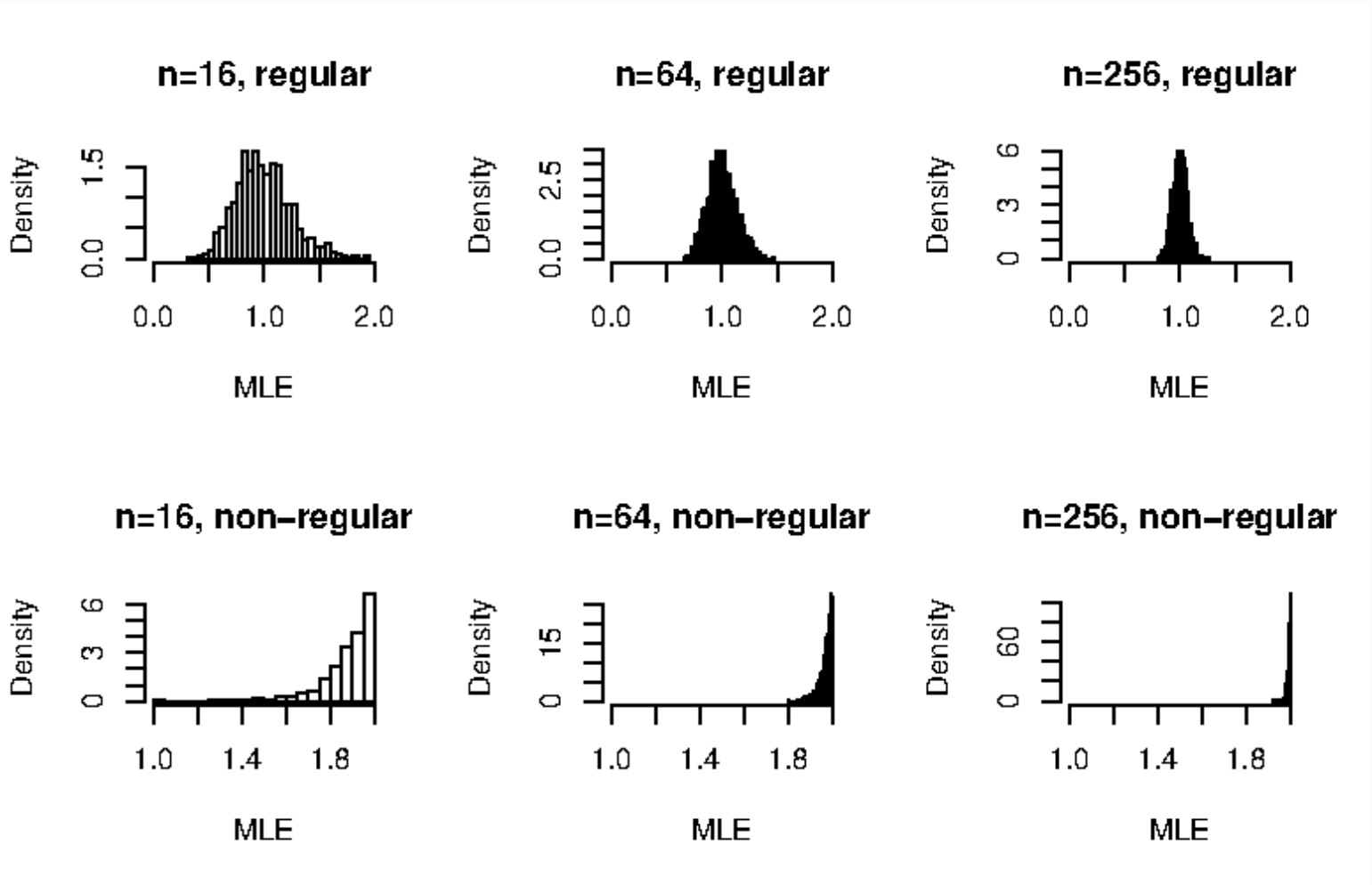
- Donc pour $x \ge 1$ et $n \ge x$,

$$\Pr\left\{n(\theta - \widehat{\theta}_{\mathrm{ML}})/\theta \le x\right\} = \Pr(\widehat{\theta}_{\mathrm{ML}} \ge \theta - x\theta/n) = 1 - \{(\theta - x\theta/n)/\theta\}^n$$
$$\to 1 - \exp(-x),$$

comme requis

Exemple (non examinable)

Comparaison des lois de $\widehat{\theta}$ dans un cas régulier (en haut, avec écart-type $\propto n^{-1/2}$ et loi limite normale) et dans un cas non-régulier (en bas, avec écart-type $\propto n^{-1}$ et loi limite non-normale)



4. Régression

4.1 Introduction

Motivation

La **régression** concerne la relation entre une variable d'intérêt et d'autres variables. On note

- la variable d'intérêt, la **variable de réponse**, y, et on la considère comme variable aléatoire
- les autres variables, les **covariables** (variables explanatoires) sont notées $x^{(1)}, \ldots, x^{(p)}$, on les considère comme fixées

On peut s'interesser à

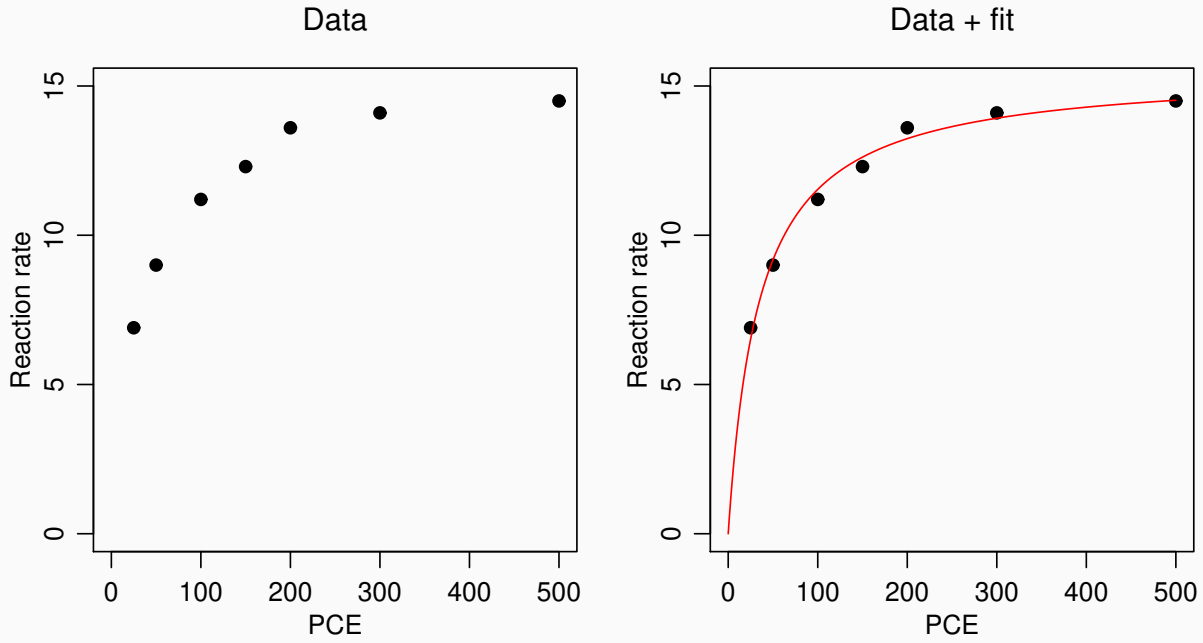
- l'**estimation** d'une relation eventuelle entre y et les $x^{(j)}$, ou
- la **prévision** des valeurs futures/manquantes de y sur la base des $x^{(j)}$ correspondantes

Réaction chimique

Professeur Christophe Holliger (SIE) : on essaye de déterminer les paramètres kinétiques d'une 'reductive dehalogenase dechlorinating tetrachloroethene (PCE)'. Ceci dépend de la concentration du substrat, et la vitesse de la réaction peut être exprimé par l'équation de Michaelis–Menten

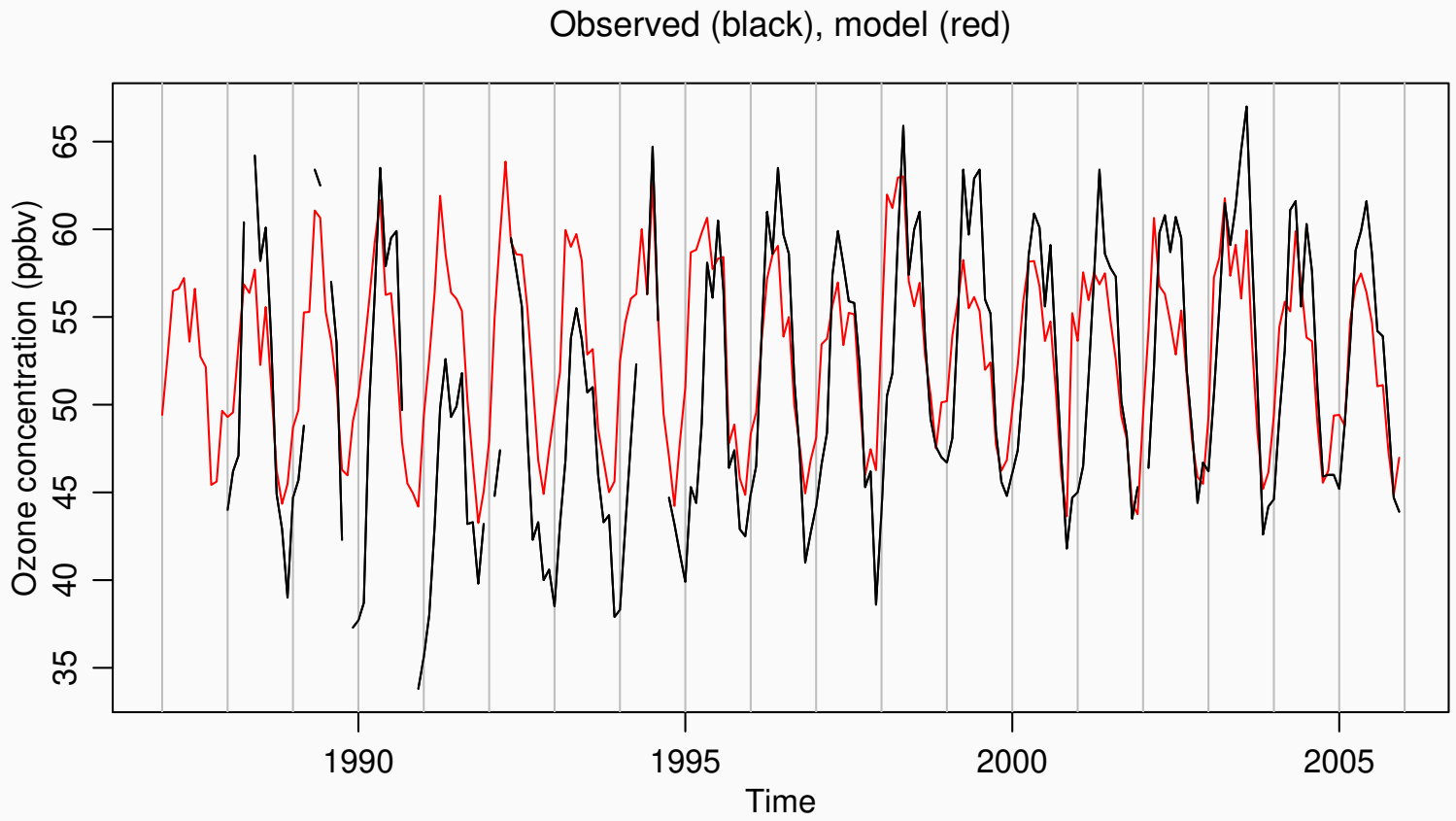
$$y = \frac{\gamma_0 x}{\gamma_1 + x},$$

où x est la concentration de PCE, γ_0 est la vitesse maximale, et γ_1 est la concentration quand $y = \gamma_0/2$. Comment estimer γ_0 et γ_1 ? Quelles sont leurs incertitudes ?



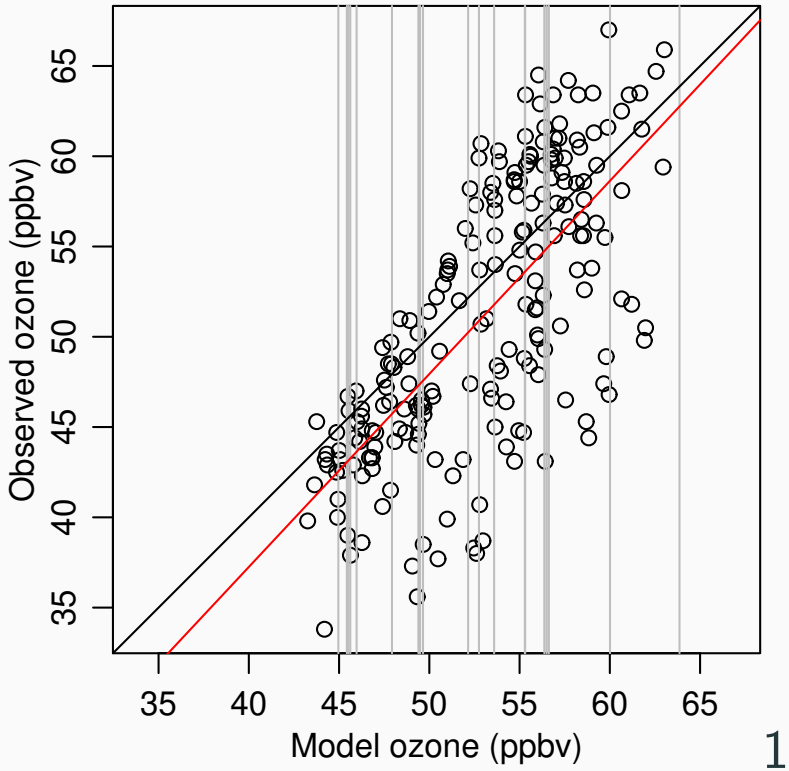
Ozone atmosphérique

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



Soient y les données réeles et x les resultats du modèle

Relation linéaire ?



120

Les lignes verticales grises montrent des x dont les y sont manquants. La ligne noire montre la relation $y = x$, et la ligne rouge montre la meilleure estimation d'une relation linéaire entre y et x.

Comment utiliser la relation entre les resultats du modèle x et les données observées y pour estimer les y manquants ?

Problème d'ajustement

- On considère une variable de réponse y que l'on cherche à expliquer par une covariable x
- On dispose d'un ensemble de points

$$\begin{pmatrix} x_1 \\ y_1 \end{pmatrix}, \ldots, \begin{pmatrix} x_n \\ y_n \end{pmatrix}$$

  qu'on peut représenter par un nuage de points (scatterplot) comme ceux d'auparavant
- D'une manière générale, le **problème d'ajustement** consiste à trouver une courbe $y = \mu(x)$ qui résume "le mieux possible" le nuage de points. La fonction $\mu(x)$ dépend de paramètres qu'il faut estimer
- S'il y a une **relation linéaire**, on peut utiliser la corrélation pour mesurer la dépendance linéaire entre les y et x. La régression linéaire permet de résumer cette dépendance par une droite

Moindres carrés

- Les écarts verticaux entre les données y_j et la courbe $\mu(x_j)$ sont

$$y_j - \mu(x_j), \quad j = 1, \ldots, n$$

- On cherche les paramètres de la fonction $\mu(x)$ pour minimiser la **somme des carrés** des écarts verticaux

$$\sum_{j=1}^{n} \{y_j - \mu(x_j)\}^2$$

- L'ajustement est dit **linéaire** si $\mu(x) = a + \beta x$. Dans ce cas, il faut minimiser

$$S(a, \beta) = \sum_{j=1}^{n} \{y_j - \mu(x_j)\}^2 = \sum_{j=1}^{n} \{y_j - (a + \beta\, x_j)\}^2$$

Estimateurs de moindres carrés

Théorème Soient $(x_1, y_1), \ldots, (x_n, y_n)$ issues d'un relation $y = a + \beta x$ et telles que pas tous les x_j sont égaux. Alors les **estimateurs de moindres carrés** de a et β sont

$$\widehat{a}_n = \overline{y}_n - \widehat{\beta}_n \overline{x}_n, \quad \widehat{\beta}_n = \frac{\sum_{j=1}^n (x_j - \overline{x}_n) y_j}{\sum_{j=1}^n (x_j - \overline{x}_n)^2}$$

Définition: La droite

$$\widehat{a}_n + \widehat{\beta}_n x$$

s'appelle la **droite des moindres carrés**, la **valeur ajustée** qui correspond à (x_j, y_j) est

$$\widehat{y}_j = \widehat{a}_n + \widehat{\beta}_n x_j,$$

et la différence

$$r_j = y_j - \widehat{y}_j = y_j - (\widehat{a}_n + \widehat{\beta}_n x_j)$$

s'appelle un **résidu**

Preuve (non-examinable)

Il faut minimiser

$$S(a,\beta) = \sum_{i=1}^{n}(y_i - a - \beta x_i)^2$$

en a et β. On calcule

$$\frac{dS}{da}(a,\beta) = -2\sum_{i=1}^{n}(y_i - a - \beta x_i) = 2na + 2n\beta\overline{x}_n - 2n\overline{y}_i$$

$$\frac{dS}{d\beta}(a,\beta) = -2\sum_{i=1}^{n} x_i(y_i - a - \beta x_i) = 2na\overline{x}_i + 2\beta\sum_{i=1}^{n} x_i^2 - 2\sum_{i=1}^{n} x_i y_i$$

$$\frac{d^2S}{da^2}(a,\beta) = 2n > 0 \qquad \frac{d^2S}{d\beta^2}(a,\beta) = 2\sum_{i=1}^{n} x_i^2 \qquad \frac{d^2S}{d\beta da}(a,\beta) = 2n\overline{x}_n$$

La matrice hessienne est donc

$$H = \begin{pmatrix} 2n & 2n\overline{x}_n \\ 2n\overline{x}_n & 2\sum_{i=1}^{n} x_i^2 \end{pmatrix} \qquad \text{définie positive}$$

car $2n > 0$ et $\det(H) = 4n[(\sum_{i=1}^{n} x_i^2) - n\overline{x}_n] = 4n\sum_{i=1}^{n}(x_i - \overline{x}_n)^2 > 0$

Propriétés

- La droite de moindres carrés passe par $(\overline{x}_n, \overline{y}_n)$
- $\sum_{j=1}^{n} r_j = 0$
- $\sum_{j=1}^{n} x_j\, r_j = \sum_{j=1}^{n} x_j(y_j - \widehat{y}_j) = 0$
- $\sum_{j=1}^{n} \widehat{y}_j\, r_j = 0$

(voir série d'exercices). Donc

$$\sum_{j=1}^{n}(y_j - \bar{y}_n)^2 = \sum_{j=1}^{n}\Big(\underbrace{y_j - \widehat{y}_j}_{r_j} + \widehat{y}_j - \bar{y}_n\Big)^2 = \cdots = \sum_{j=1}^{n}(\widehat{y}_j - \bar{y}_n)^2 + \sum_{j=1}^{n} r_j^2,$$

nous donnant la **décomposition de la somme des carrés** total

$$\mathrm{SC}_{\mathrm{Total}} = \mathrm{SC}_{\mathrm{R}} + \mathrm{SC}_{\mathrm{E}}$$

en une partie due à la régression (variation expliquée par le modèle) et une partie due à l'erreur (variation non-expliquée par le modèle)

Ozone atmosphérique

- Il y a $n = 207$ paires (Observée, Modèle) $= (y_j, x_j)$, et en plus 21 valeurs de x sans valeur observée
- Avec les n paires complètes on trouve comme droite des moindres carrés

$$\widehat{y} = \widehat{a}_n + \widehat{\beta}_n x = -5.511 + 1.069x$$

avec décomposition de la somme des carrés

$$\mathrm{SC}_{\mathrm{Total}} = \mathrm{SC}_{\mathrm{R}} + \mathrm{SC}_{\mathrm{E}} = 5813 + 5832.$$

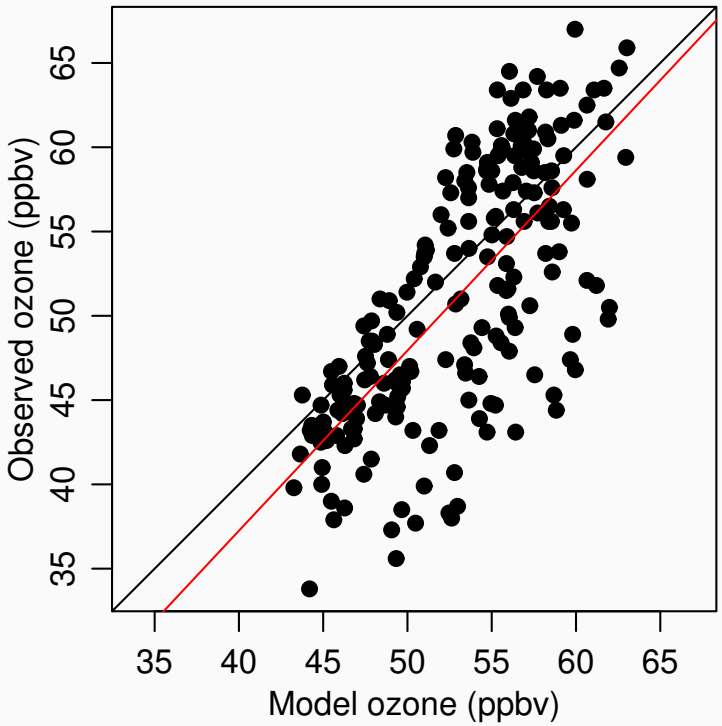
La régression explique donc une moitié de la somme des carrés totale

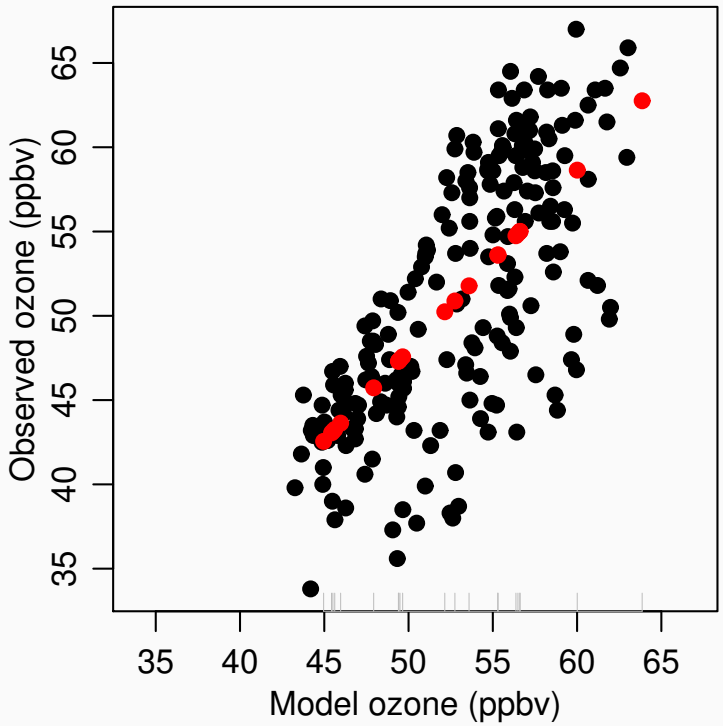
- Pour un paire (Observée, Modèle) $= (\,?,\, x_+)$ dont la valeur observée manque, on peut la remplacer par la valeur ajustée correspondante,

$$\widehat{y}_+ = \widehat{a}_n + \widehat{\beta}_n x_+.$$

On parle d'**imputation** de donnée

Modèle ajusté

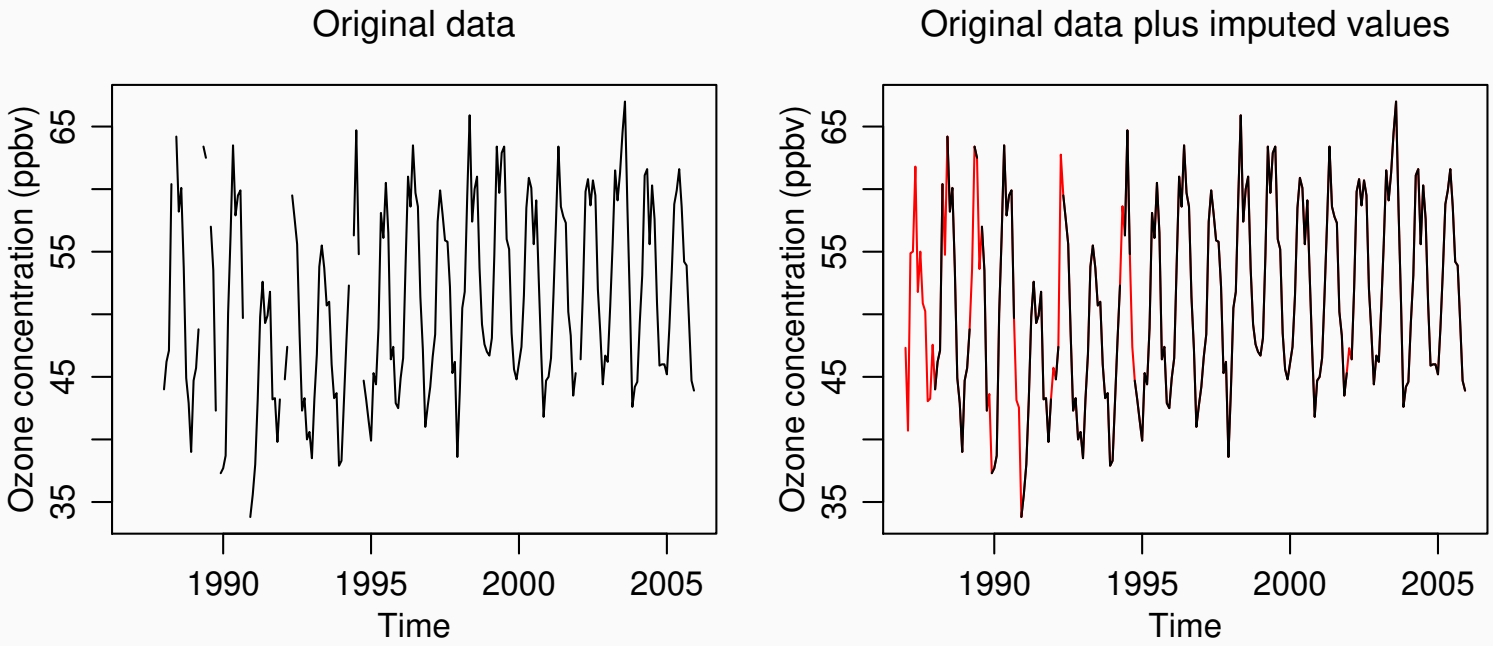




Gauche : droite $y = x$ et droite ajustée $\widehat{y} = \widehat{a}_n + \widehat{\beta}_n x = -5.511 + 1.069x$

Droite : valeurs ajustées pour des valeurs manquantes de x

Données imputées



Gauche : données originales

Droite : données originales (noir) avec valeurs imputées (rouge). Comparer avec la diapositive 230

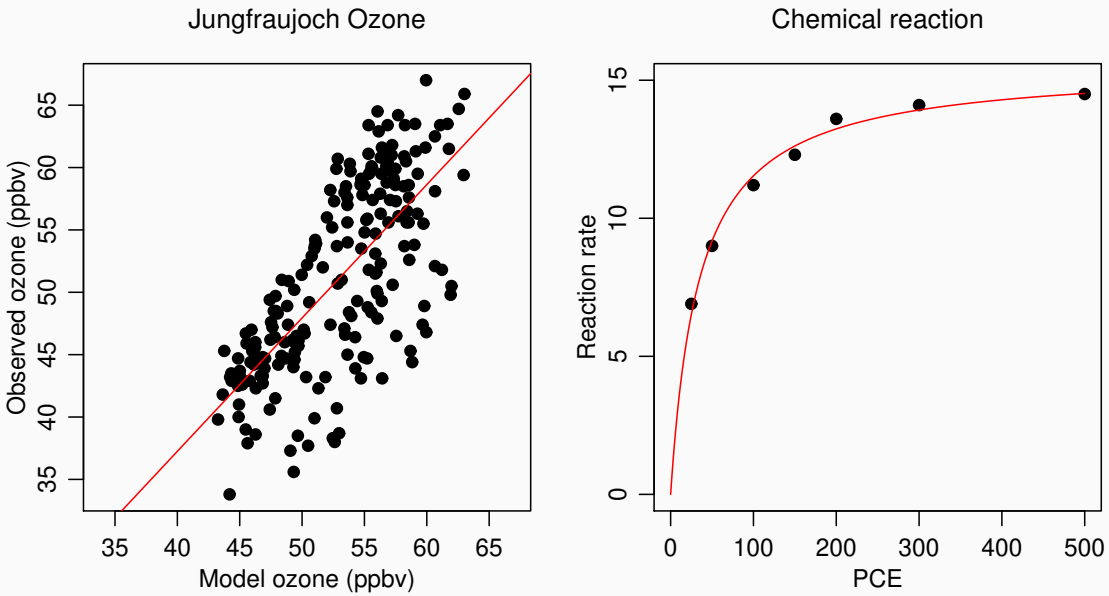
4.2 Modèle statistique

Modèle normale

- On observe une version perturbée d'une relation $y = \mu(x)$
- Pour modéliser ceci, on peut souvent supposer que

$$y_j \overset{\text{ind}}{\sim} \mathcal{N}\left\{\mu(x_j), \sigma^2\right\} \quad \text{ou bien} \quad y_j = \mu(x_j) + \epsilon_j, \quad \epsilon_j \overset{\text{ind}}{\sim} \mathcal{N}(0, \sigma^2)$$

- Ainsi la dépendance entre la réponse y et la variable explicative x est donnée par $\mathbb{E}(y) = \mu(x)$, alors que le bruit dépend de σ^2
- À gauche : $\mu(x)$ linéaire, σ^2 grand, donc beaucoup de bruit
- À droite : $\mu(x)$ non-linéaire, σ^2 petite, donc peu de bruit



Linéarité

- La linéarité du modèle concerne les paramètres :

$$y = a + \beta x + \epsilon,$$

où $\epsilon \sim \mathcal{N}(0, \sigma^2)$ est la différence entre y et la droite $a + \beta x$

- Le modèle

$$y = a + \beta x + \gamma x^2 + \delta x^3 + \epsilon$$

est linéaire en $(a, \beta, \gamma, \delta)$.

- Le modèle

$$y = \gamma_0 x^{\gamma_1} \eta, \quad \eta \sim \exp(1),$$

devient linéaire après transformation logarithmique :

$$\log y = \log \gamma_0 + \gamma_1 \log x + \log \eta = a + \beta x' + \log \eta$$

- Le modèle

$$y = \frac{\gamma_0 x}{\gamma_1 + x} + \epsilon$$

n'est pas linéaire en les paramètres γ_0, γ_1

Estimation des paramètres

- Dans le cas $\mu(x) = a + \beta x$ il y a trois paramètres inconnus : (intercepte, pente, bruit), $\theta = (a, \beta, \sigma^2) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}_+$
- Nous utilisons la méthode de maximum de vraisemblance pour les estimer
- La log vraisemblance est

$$\ell(a, \beta, \sigma^2) = -\frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \overbrace{\sum_{j=1}^{n} \{y_j - (a + \beta x_j)\}^2}^{S(a,\beta)} - \frac{n}{2} \log(2\pi),$$

et en maximisant celle-ci par rapport à θ nous trouvons

$$\widehat{a}_n = \overline{y}_n - \widehat{\beta}_n \overline{x}_n, \quad \widehat{\beta}_n = \frac{\sum_{j=1}^{n} (x_j - \bar{x}_n) y_j}{\sum_{j=1}^{n} (x_j - \bar{x}_n)^2}, \quad \widehat{\sigma}_n^2 = n^{-1} \sum_{j=1}^{n} r_j^2$$

avec $r_j = y_j - \widehat{y}_j$ les **résidus** et $\widehat{y}_j = \widehat{a}_n + \widehat{\beta}_n x_j$ les **valeurs ajustées**

- Les estimateurs $\widehat{a}_n$ et $\widehat{\beta}_n$ sont les estimateurs de moindres carrés et sont sans biais, mais $\mathbb{E}(\widehat{\sigma}_n^2) < \sigma^2$, et on utilise souvent l'estimateur non-biaisé (comparer avec 174)

$$S_n^2 = \frac{1}{n-2} \sum_{j=1}^{n} r_j^2 = \frac{1}{n-2} \sum_{j=1}^{n} (y_j - \widehat{y}_j)^2$$

Inférence pour les paramètres du modèle linéaire simple

- Le coefficient β (pente) est plus intéressant que a (ordonnée à l'origine). On se concentre donc ici sur le premier
- On peut montrer que

$$\mathrm{Var}(\widehat{\beta}_n) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \overline{x}_n)^2}$$

- On estime σ^2 par S^2 pour estimer cette variance. En prenant la racine carrée on obtient **l'erreur type** (standard error)

$$\widehat{\mathrm{sd}}(\widehat{\beta}_n) = \frac{S_n}{\sqrt{\sum_{i=1}^n (x_i - \overline{x}_n)^2}}$$

- On peut montrer que

$$\frac{\widehat{\beta}_n - \beta}{\widehat{\mathrm{sd}}(\widehat{\beta}_n)} \sim t_{n-2}$$

On a donc un pivot. On peut construire des intervalles de confiance et tester des hypothèses

Intervalles de confiance pour β

On en déduit des intervalles de confiance pour β au niveau de confiance $1-\alpha$, pour $\alpha \in (0,1)$:

- Intervalle de confiance bilatéral symétrique :
$$\left[\widehat{\beta}_n - t_{n-2,1-\alpha/2}\frac{S_n}{\sqrt{\sum_{i=1}^n (x_i-\bar{x})^2}}, \widehat{\beta}_n + t_{n-2,1-\alpha/2}\frac{S_n}{\sqrt{\sum_{i=1}^n (x_i-\bar{x})^2}}\right].$$
- Intervalle de confiance unilatéral à gauche :
$$\left(-\infty, \widehat{\beta}_n + t_{n-2,1-\alpha}\frac{S_n}{\sqrt{\sum_{i=1}^n (x_i-\bar{x})^2}}\right].$$
- Intervalle de confiance unilatéral à droite :
$$\left[\widehat{\beta}_n - t_{n-2,1-\alpha}\frac{S_n}{\sqrt{\sum_{i=1}^n (x_i-\bar{x})^2}}, \infty\right).$$

Comparer avec diapositive 186 : $[\widehat{\theta} \pm t_{k,1-\alpha/2}\widehat{sd}(\widehat{\theta})]$: en 186, $k = n-1, \widehat{\theta} = \overline{Y}_n$, ici $k = n-2, \widehat{\theta} = \widehat{\beta}_n$

Tests pour β

On peut effectuer les tests statistiques classiques au niveau de significativité α, pour $\alpha \in (0, 1)$:

- Test bilatéral $H_0 : \beta = \beta_0$ contre $H_1 : \beta \neq \beta_0$. On rejette H_0 si et seulement si $|t_{\text{obs}}| > t_{n-2,1-\alpha/2}$.
- Test unilatéral à gauche $H_0 : \beta = \beta_0$ contre $H_1 : \beta < \beta_0$. On rejette H_0 si et seulement si $t_{\text{obs}} < t_{n-2,1-\alpha}$.
- Test unilatéral à droite $H_0 : \beta = \beta_0$ contre $H_1 : \beta > \beta_0$. On rejette H_0 si et seulement si $t_{\text{obs}} > t_{n-2,1-\alpha}$.

La statistique de test est

$$T = \frac{\widehat{\beta_n} - \beta_0}{\widehat{\text{sd}}(\widehat{\beta_n})} = \frac{\widehat{\beta_n} - \beta_0}{S_n / \sqrt{\sum_{i=1}^n (x_i - \overline{x}_n)^2}}$$

qui suit la loi t_{n-2} quand H_0 est vraie

Nos données

```
> JungOzone
   Observed Model
1        NA 49.42
2      40.7 52.79
3        NA 56.49
4        NA 56.61
5      61.8 57.22
6        NA 53.59
7        NA 56.61
8        NA 52.75
9        NA 52.15
10       NA 45.43
...
> MM <- data.frame(
+   x=c(25, 50, 100, 150, 200, 300, 500),
+   y=c(6.9, 9.0, 11.2, 12.3, 13.6, 14.1, 14.5))
> MM
    x    y
1  25  6.9
2  50  9.0
3 100 11.2
4 150 12.3
5 200 13.6
6 300 14.1
7 500 14.5
```

Inférence

Voici le résultat de l'ajustement du modèle linéaire aux données d'ozone :

```
> fit <- lm(Observed~Model,data=JungOzone)
> summary(fit)
...
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.51072    3.98014  -1.385    0.168
Model        1.06903    0.07479  14.294   <2e-16 ***
---
Signif. codes:  0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1

Residual standard error: 5.334 on 205 degrees of freedom
  (21 observations deleted due to missingness)
Multiple R-Squared: 0.4992,Adjusted R-squared: 0.4967
F-statistic: 204.3 on 1 and 205 DF,  p-value: < 2.2e-16
```

Exemple : données d'ozone (inférence)

- On sait d'après les slides précédentes que l'intervalle de confiance bilatéral symétrique pour β au niveau de confiance $1-\alpha$ est
$$\left[\hat{\beta}_n - t_{n-2,1-\alpha/2}\hat{\mathrm{sd}}(\hat{\beta}_n), \hat{\beta}_n + t_{n-2,1-\alpha/2}\hat{\mathrm{sd}}(\hat{\beta}_n)\right].$$
- Ainsi, en lisant les sorties du logiciel, on obtient qu'une réalisation de l'IC précédent pour β au niveau de confiance 95% est donnée par
$$1.06903 \pm t_{205,0.975} \times 0.07479 \approx 1.07 \pm 1.97 \times 0.07 = [0.93, 1.21].$$
- Souvent, on veut tester si le terme impliquant la covariable est significatif. Cela revient à tester $H_0 : \beta = 0$.
- Ici, le scatter plot semble clairement indiquer que β est différent de 0 et on effectue donc plutôt le test $H_0 : \beta = 1$. On choisit comme niveau de significativité $\alpha = 0.05$. On rejette H_0 si et seulement si la valeur absolue de la réalisation t_{obs} de
$$T = \frac{\hat{\beta}_n - 1}{\hat{\mathrm{sd}}(\hat{\beta}_n)}$$
est strictement supérieure à $t_{n-2,1-\alpha/2} = t_{205,0.975} \approx 1.97$. On a $t_{\mathrm{obs}} \approx 0.92$ et on ne rejette donc pas H_0.

Modèle nonlinéaire (non-examinable)

- Les mêmes idées s'appliquent aux modèles nonlinéaires, mais comme approximations
- Il faut donner des valeurs initiales pour γ_0 et γ_1, en principe il faut en essayer plusieurs, car il est possible que la vraisemblance ait des maximas locaux
- Pour ajuster le modèle $\mu(x) = \gamma_0 x/(\gamma_1 + x)$ aux données chimiques :

```
> fit <- nls(y~g0*x/(g1+x),data=MM, start=c(g0=1,g1=1))
> summary(fit)

Formula: y ~ g0 * x/(g1 + x)

   Estimate Std. Error t value Pr(>|t|)
g0  15.5269     0.2876   53.99 4.12e-08 ***
g1  34.5990     2.8777   12.02 7.02e-05 ***
---
Signif. codes:  0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1

Residual standard error: 0.3341 on 5 degrees of freedom
```

Coefficient de détermination

- Nous avons dêjà vu la **décomposition de la somme des carrés** total

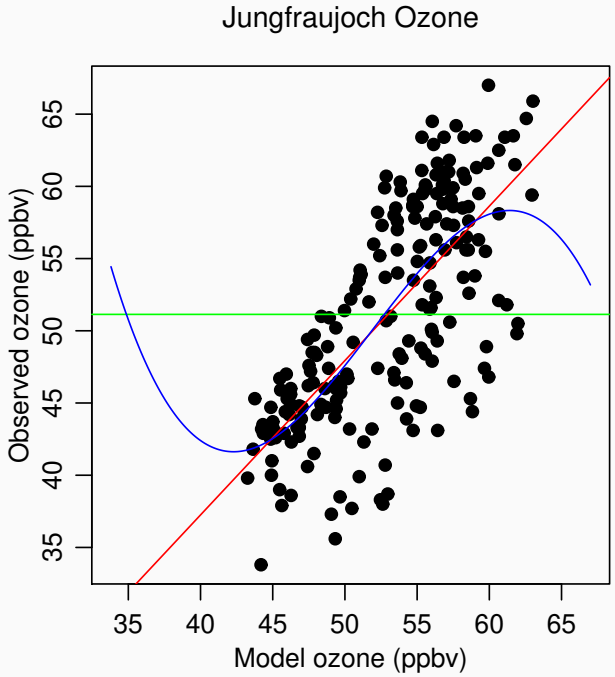
$$\sum_{j=1}^{n}(y_j-\bar{y})^2=\sum_{j=1}^{n}(\widehat{y}_j-\bar{y})^2+\sum_{j=1}^{n}r_j^2,\quad \text{soit}\quad \mathrm{SC}_{\mathrm{Total}}=\mathrm{SC}_{\mathrm{R}}+\mathrm{SC}_{\mathrm{E}},$$

  en une partie SC_{R} due à la régression et une partie SC_{E} due à l'erreur
- Le proportion (ou pourcentage) de la variation totale expliquée par le modèle

$$R^2=\frac{\mathrm{SC}_{\mathrm{R}}}{\mathrm{SC}_{\mathrm{Total}}}=\frac{\mathrm{SC}_{\mathrm{Total}}-\mathrm{SC}_{\mathrm{E}}}{\mathrm{SC}_{\mathrm{Total}}}$$

  est appelé **coefficient de détermination** ; $0\leq R^2\leq 1$
- Si $R^2\approx 1$, alors $y_j\approx\widehat{y}_j$ pour tout j et donc tous les $r_j\approx 0$, et donc le modèle explique les données presque parfaitement
- Si $R^2\approx 0$, alors l'inclusion de x n'explique presque rien de la variation totale
- Pour les données d'ozone, $R^2=0.5$, donc la moitié de la variance est expliquée par le modèle
- Pour les données chimiques, $R^2=0.99$, donc le modèle explique presque toute la variation

Comparaison des modèles



- Voici trois modèles :

$$\begin{aligned} \text{constant (vert)} : \quad y &= a + \epsilon, \\ \text{linéaire (rouge)} : \quad y &= a + \beta x + \epsilon, \\ \text{cubique (bleu)} : \quad y &= a + \beta x + \gamma x^2 + \delta x^3 + \epsilon? \end{aligned}$$

- Le rouge semble être bien meilleur que le vert, mais que le rouge et le bleu semblent être semblables. Comment tester ces constats ?

Décomposition de la variance

- Comparons le modèle constante $y = a + \epsilon$ et le modèle linéaire $y = a + \beta x + \epsilon$
- Pour tester s'il vaut la peine d'ajouter βx, on calcule

$$F = \frac{\mathrm{SC_R}/1}{\mathrm{SC_E}/(n-2)} \quad \sim \quad F_{1,n-2}$$

  si l'hypothèse nulle $H_0 : \beta = 0$ que le modèle est constant est vraie
- F_{d_1,d_2} est la **loi de Fisher(–Snedecor)** avec d_1 et d_2 degrés de liberté
- Pour un niveau de significativité $\alpha \in (0,1)$ donné, il faut comparer la valeur observée de F avec le $1-\alpha$ quantile $F_{1,n-2,1-\alpha}$ (rejet pour grandes valeurs de F)
- Pour les données d'ozone, on trouve $f_{obs} = 204.32$, à comparer avec $F_{1,205,0.95} = 3.887$
- Ce test est équivalent au t-test pour $H_0 : \beta = 0$ vu précédemmant car : $T \sim t_\nu \Longrightarrow T^2 \sim F_{1,\nu}$

Statistique F

- Pour tester $H_0 : \beta_{q+1} = \cdots = \beta_p = 0$ dans le modèle linéaire

$$y_i = \beta_0 + \beta_1 x_i^{(1)} + \cdots + \beta_q x_i^{(q)} + \beta_{q+1} x_i^{(q+1)} + \cdots \beta_p x_i^{(p)} + \epsilon,$$

on a deux sommes des carrés, l'un $\mathrm{SC}_{E,p}$ qui correspond au modèle avec $x^{(1)}, \ldots, x^{(p)}$ et l'autre $\mathrm{SC}_{E,q}$ qui correspond au modèle réduit avec $x^{(1)}, \ldots, x^{(q)}$, $q < p$. On a $\mathrm{SC}_{E,p} \leq \mathrm{SC}_{E,q}$, et pour tester H_0 on calcule

$$F = \frac{(\mathrm{SC}_{E,q} - \mathrm{SC}_{E,p})/(p-q)}{\mathrm{SC}_{E,p}/(n-p-1)} \sim F_{p-q,n-p-1}$$

si $H_0 : \beta = 0$ est vraie

- On rejette H_0 au niveau α si $f_{obs} > F_{p-q,n-p-1,1-\alpha}$
- Pour les données d'ozone, pour tester $\gamma = \delta = 0$ dans le modèle cubique

$$y = a + \beta x + \gamma x^2 + \delta x^3 + \epsilon,$$

on a $n = 207$, $p = 3$, $q = 1$, et

$$F = \frac{(5831.9 - 5712.2)/(3-1)}{5712/(207-3-1)} = 2.13 \sim F_{3-1,207-3-1} = F_{2,203},$$

dont le 0.95 quantile est $F_{2,203,0.95} = 3.04$.

Validation du modèle de régression linéaire (non-examinable)

- Le modèle normale $y \sim \mathcal{N}\left\{\mu(x), \sigma^2\right\}$ implique que

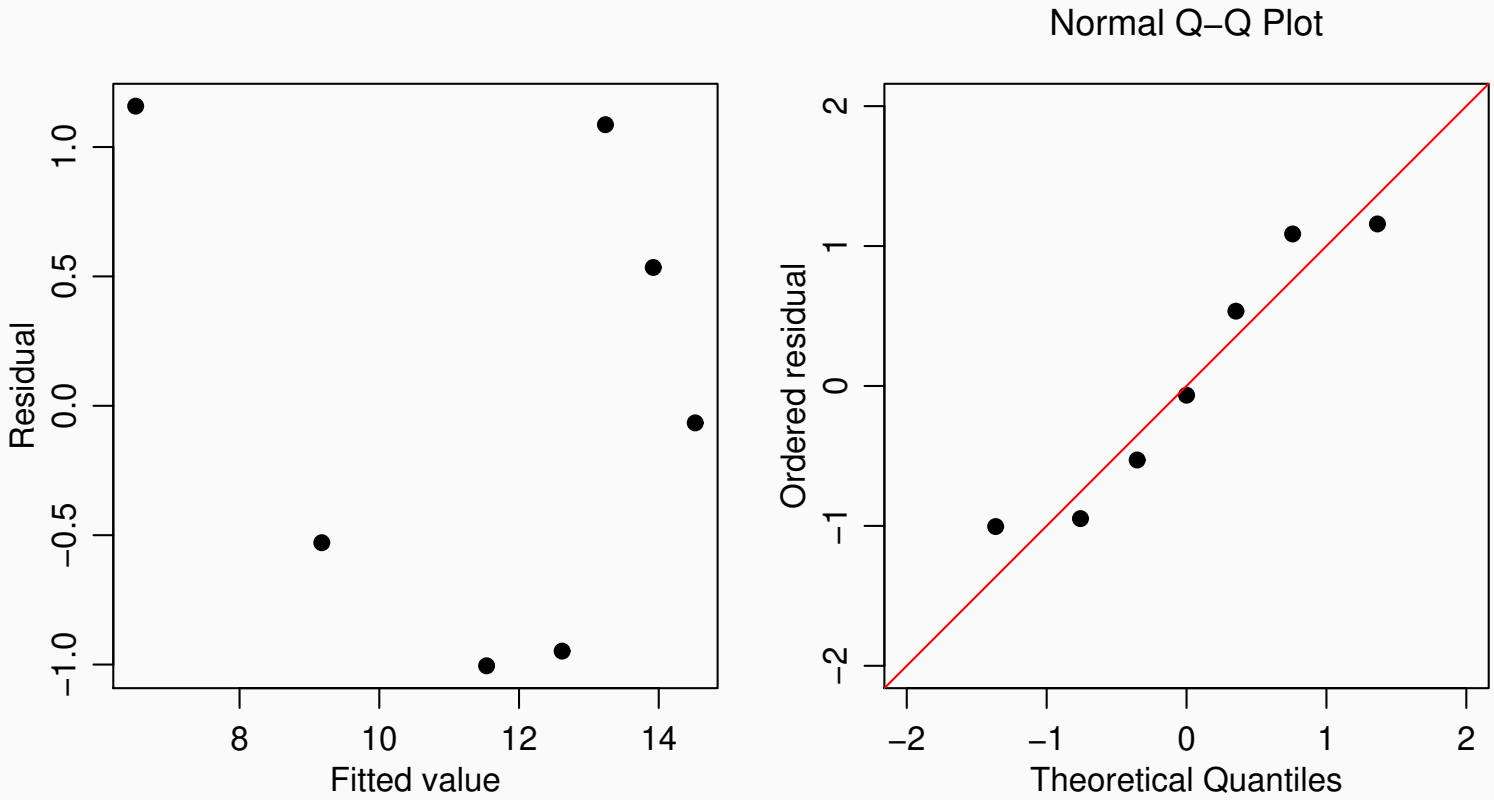
$$\frac{y - \mu(x)}{\sigma} \sim \mathcal{N}(0, 1),$$

et donc que le **résidu standardisé**

$$r_j^S = \frac{r_j}{s_n} = \frac{y_j - \widehat{y}_j}{s_n} = \frac{r_j}{s_n} = \frac{y_j - (\widehat{a}_n + \widehat{\beta}_n x_j)}{s_n} \overset{\text{app}}{\sim} \mathcal{N}(0, 1)$$

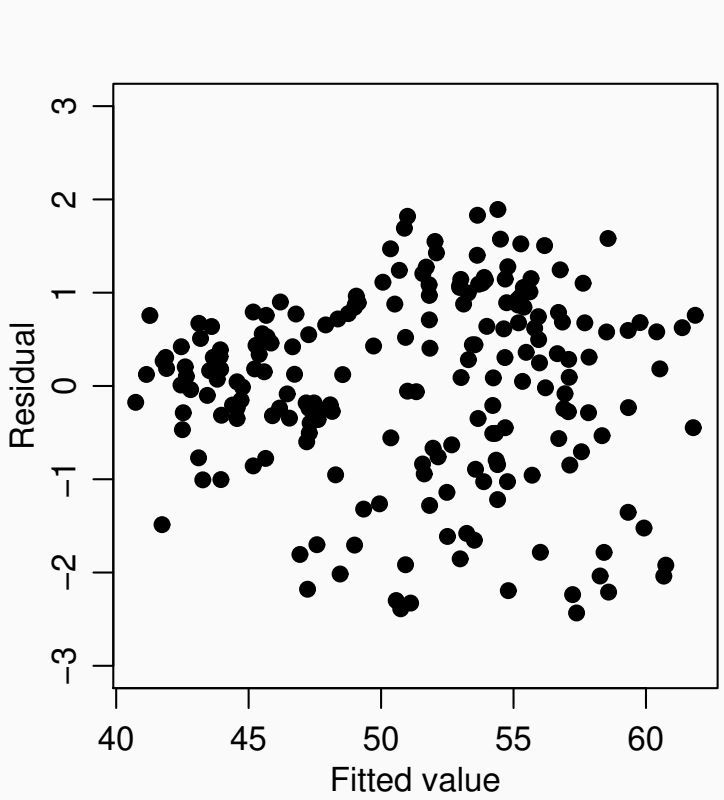
- On teste cela graphiquement avec un quantile-quantile plot (Q-Q plot) normal. C'est un graphique des quantiles empiriques des données (ici les résidus standardisés) contre les quantiles théoriques d'une loi $\mathcal{N}(0, 1)$. Si les r_i^S suivent effectivement la loi $\mathcal{N}(0, 1)$, alors les points du Q-Q plot doivent se trouver (plus ou moins) sur la diagonale $y = x$. Des écarts trop importants par rapport à la diagonale indiquent une violation de l'hypothèse de normalité des erreurs.
- Par ailleurs, il faut qu'il n'y ait pas de relation entre les r_j^S et les valeurs ajustées $\widehat{y}_j$

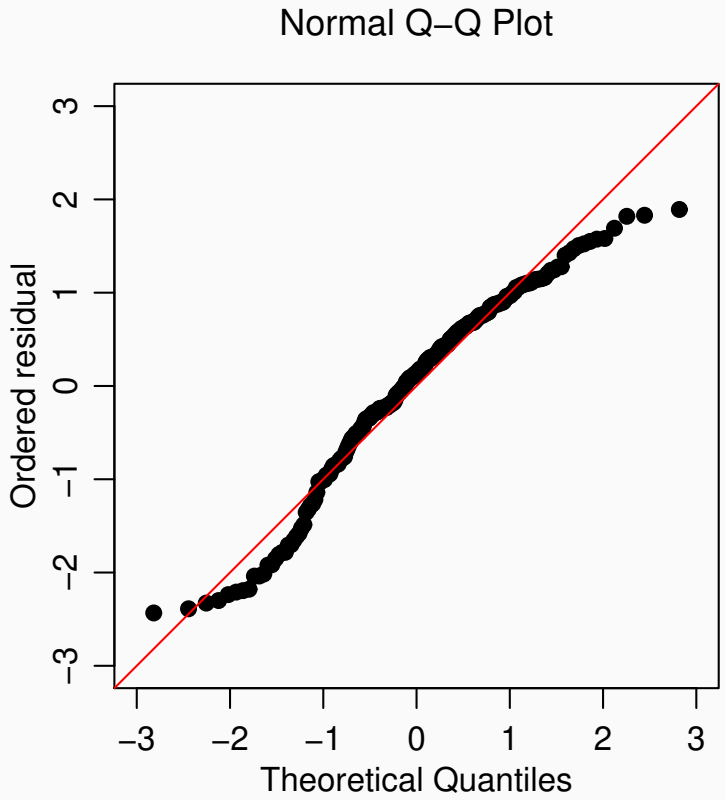
Données chimiques (non-examinable)



- À gauche : r_j^S contre $\widehat{y}_j$
- À droite : QQplot des r_j^S
- Avec $n = 7$, il est presque impossible de contradire le modèle

Données d'ozone (non-examinable)





- À gauche : r_j^S contre $\widehat{y}_j$
- À droite : QQplot des r_j^S
- La loi des erreurs n'est pas normale, mais asymétrique, et la variance semble changer avec $\mathbb{E}(y)$