Probabilités et Statistique

Linda Mhalla
linda.mhalla@epfl.ch

Printemps 2025

Organisation

- ☐ Enseignant : Linda Mhalla, linda.mhalla@epfl.ch
- ☐ Assistant principal : Emil Bennewitz, emil.bennewitz@epfl.ch
- ☐ 2 heures de cours par semaine (les mardis de 08h15 à 10h00 en AAC 1 37).
- ☐ 2 heures d'exercices par semaine (les mercredis de 14h15 à 16h00 en INM 202).
- ☐ N'hésitez pas à poser des questions en cours, à la pause et après le cours !
- ☐ Les séances d'exercices vous aideront beaucoup, n'hésitez pas à solliciter vos assistants au maximum !
- ☐ Evaluation : un examen final (seuls un formulaire et une calculatrice non-programmable seront autorisés).

Organisation

- ☐ Matériel (disponible sur Moodle) :
  - Un polycopié contenant notamment tous les transparents utilisés en cours. Il s'agit d'une version largement remaniée de notes de cours des Profs. D. Kuonen, A. C. Davison, V. M. Panaretos, E. Thibaud et E. Koch.
  - Un examen blanc (et sa solution) similaire à l'examen final en termes de structure.
  - Le formulaire auquel vous aurez droit pour l'examen final.
  - Un document regroupant informations et conseils pour l'examen final.
  - Les exercices et leurs solutions (postées chaque mercredi à 18h00).
- ☐ Un ancien polycopié était (est) en vente à la bibliothèque : ne pas l'acheter.
- ☐ Une référence (pas besoin de l'acheter) : *Introduction à la statistique*, S. Morgenthaler, PPUR, 2014.

Statistique : définition

Commençons par les mathématiques :

Le terme "Mathématiques" vient du grec *máthēma* qui signifie "apprendre".

C'est une manière :

- ☐ d'exprimer une grande variété de notions complexes avec précision et cohérence ;
- ☐ de *"légitimer les conquêtes de notre intuition"* (selon Jacques Hadamard) — apprendre, comprendre et conclure correctement.

Statistique : définition

Et la statistique :

Science
utilisant les mathématiques
pour
extraire des informations
à partir de
données
en présence
d'aléatoire.

Statistique : objectifs

Entre autres :

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

Et les probabilités ?

La théorie des probabilités nous aide pour la partie "aléatoire". 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.

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

Analyse des données

L'analyse des données est formée de deux phases :

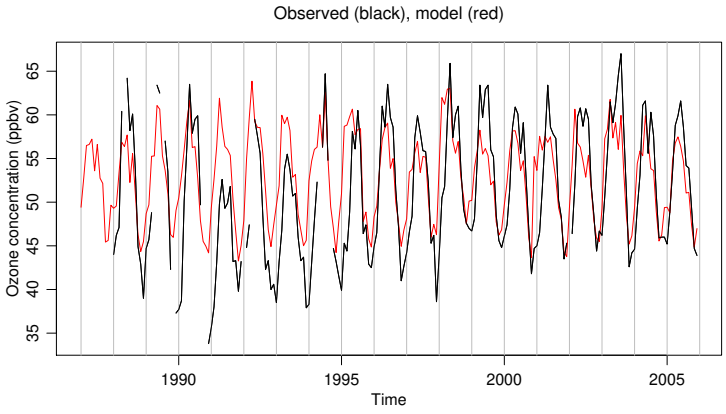
A. **L'analyse exploratoire des données** (statistiques exploratoires/descriptives) :
   - composée principalement de méthodes relativement simples, intuitives, flexibles et graphiques ;
   - permet d'étudier la "structure" des données et de détecter des caractéristiques spécifiques (tendances, formes, observations atypiques).

L'analyse exploratoire suggère des hypothèses de travail et des modèles pouvant être formalisés et vérifiés dans la seconde phase.

B. **L'inférence statistique** (analyse confirmatoire des données) :
   - conduit à des conclusions statistiques à partir des données en utilisant des notions de la théorie des probabilités ;
   - cette partie est plus formelle et concerne notamment la modélisation statistique ainsi que les méthodes de test, d'estimation, et de prédiction.

Exemple : ozone atmosphérique

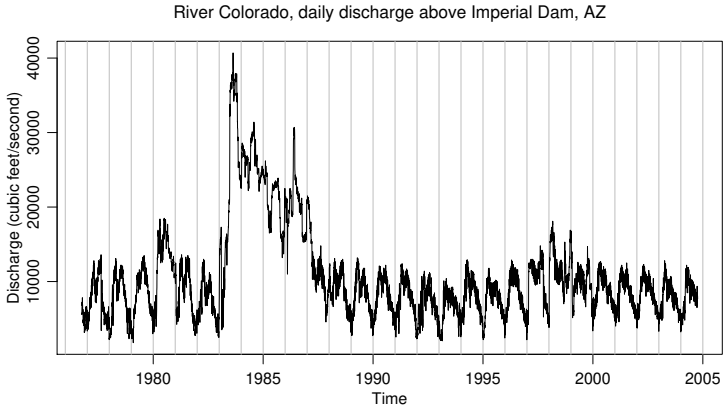
Prof. Isabelle Bey (SIE) : observations de la concentration d'ozone au Jungfraujoch de janvier 1987 à décembre 2005 (quelques valeurs manquantes), et résultats d'une modélisation.



La modélisation vous paraît-elle bonne ?

Exemple : le fleuve Colorado

Prof. Andrew Barry (SIE) : débits (en pieds cube par seconde) du fleuve Colorado au-dessus du barrage Imperial Dam, Arizona.



Y a-t-il des changements à long terme ?

Structure du cours

Le cours est divisé en quatre chapitres :

- **Statistique exploratoire** (2 semaines)—types de données, étude graphique des variables, synthèses numériques d'une distribution, boxplot, loi normale ;
- **Probabilités** (environ 5 semaines)—probabilités d'événements, variables aléatoires, valeurs caractéristiques, théorèmes fondamentaux ;
- **Notions fondamentales de la statistique** (environ 5 semaines)—modèles statistiques, estimation des paramètres, intervalles de confiance, tests statistiques, tests du khi-deux ;
- **Régression linéaire** (environ 2 semaines)—introduction, principe des moindres carrés, régression linéaire simple, régression linéaire multiple.

1. Statistique exploratoire slide 14

1.1 Types de données slide 15

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.

Exemple :

- Population : ensemble des étudiants de l'EPFL.
- Echantillon : ensemble des étudiants en 1ère année à l'EPFL.
- Individu : un(e) étudiant(e) en 1ère année à l'EPFL.
- Donnée : le poids de cet individu.

Types de variables

Une variable peut être **quantitative** ou **qualitative**.

Une **variable quantitative** peut être **discrète** (souvent entière) ou **continue** (c'est-à-dire qu'elle prend n'importe quelle valeur dans un intervalle).

- ☐ Variables quantitatives discrètes :
  - le nombre d'enfants dans une famille ;
  - le nombre d'étudiant(e)s dans cette salle.
- ☐ Variables quantitatives continues :
  - le poids en kg d'un individu ;
  - la taille en cm d'un individu.

Variables qualitatives

Une **variable qualitative** (catégorielle) peut être **nominale** (ses instances ne peuvent pas être ordonnées) ou **ordinale** (ses instances peuvent être ordonnées).

- ☐ Variables qualitatives nominales :
  - le sexe (masculin ou féminin) ;
  - les groupes sanguins (A, B, AB, O).
- ☐ Variables qualitatives ordinales :
  - la qualité du repas proposé au Vinci (bon, passable, mauvais) ;
  - l'intérêt pour les statistiques (très bas, bas, moyen, élevé, très élevé).

On convertit parfois des variables quantitatives en variables catégorielles pour des raisons descriptives ou autres. Par exemple : la taille en cm $\Rightarrow$ petit, moyen, grand.

1.2 Etude graphique des variables

Etude d'une variable qualitative

Exemple 1 *Le groupe sanguin de 25 donneurs a été relevé :*

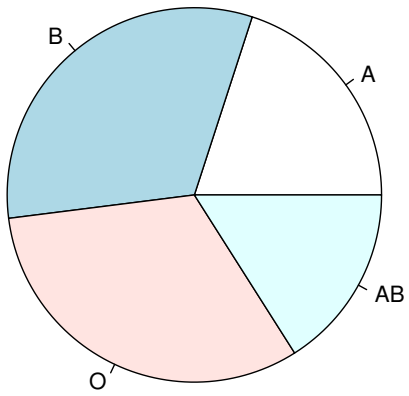
AB	B	A	O	B
O	B	O	A	O
B	O	B	B	B
A	O	AB	AB	O
A	B	AB	O	A

La table des 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*

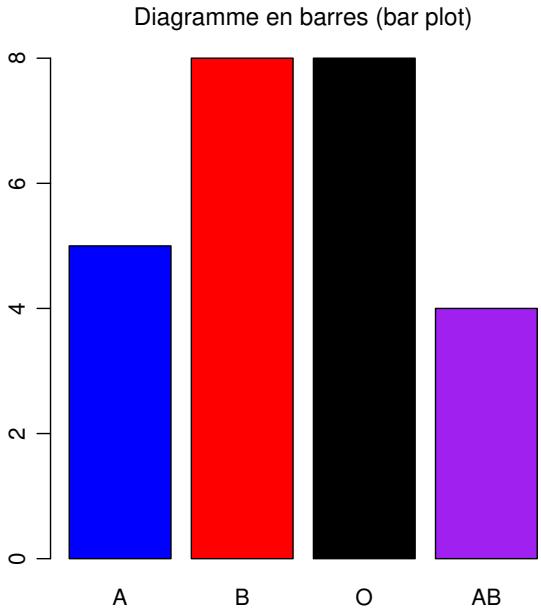
Diagramme en camembert

Diagramme en camembert/en secteurs (pie chart)



A éviter : difficile de comparer les fréquences.

Diagramme en barres



Etude d'une variable quantitative

Considérons une seule variable continue mesurée plusieurs (n) fois. On dispose ainsi de n observations

$$x_1, x_2, \ldots, x_n$$

de cette variable.

Ces valeurs peuvent être rangées dans l'ordre croissant. Les valeurs ainsi ordonnées seront notées

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}.$$

Le minimum est donc $x_{(1)}$ et le maximum $x_{(n)}$. Il existe d'autres notations : pour $i = 1, \ldots, n$, $x_{(i)}$ peut aussi être noté $x_{[i]}$ ou $x_{i/n}$ ou $x_{i:n}$ ou $x_{(i)|n}$.

Exemple

Exemple 2 *Le poids (plus rigoureusement la masse) de 92 étudiants d'une école américaine a été relevé, en livres anglaises (pounds) ; 1 lb ≈ 0.45 kg.*

Les données observées figurent dans le tableau suivant :

Garçons										
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*									
Filles										
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*									

Diagramme branches-et-feuilles (stem-and-leaf)

On sépare chaque poids entre le nombre de dizaines et le chiffre des unités. Par exemple, $95 \mapsto 9 \mid 5$, $102 \mapsto 10 \mid 2$, $108 \mapsto 10 \mid 8$. Puis, pour chaque nombre de dizaines, on reporte toutes les instances du chiffre des unités. On obtient le diagramme :

9	5
10	288
11	002556688
12	00012355555
13	0000013555688
14	00002555558
15	000000000355555555557
16	000045
17	000055
18	0005
19	00005
20	
21	5

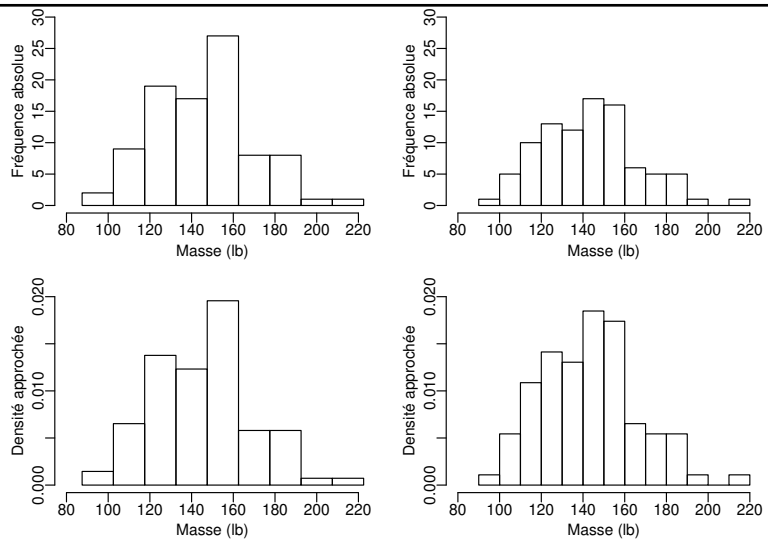
Histogramme

- Un histogramme montre le nombre d'observations (ou un équivalent, cf ci-après) dans des classes issues d'une division en intervalles de même longueur.
- 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.

Exemple de table de fréquences :

Classe	Centre	Fréquence absolue	Fréquence relative
$87.5 - 102.5^-$	95	2	0.022
$102.5 - 117.5^-$	110	9	0.098
$117.5 - 132.5^-$	125	19	0.206
$132.5 - 147.5^-$	140	17	0.185
$147.5 - 162.5^-$	155	27	0.293
$162.5 - 177.5^-$	170	8	0.087
$177.5 - 192.5^-$	185	8	0.087
$192.5 - 207.5^-$	200	1	0.011
$207.5 - 222.5^-$	215	1	0.011
Total		92	1

Histogramme



Histogrammes du poids des étudiants de l'école américaine, avec 9 classes (gauche) et 13 classes (droite). En haut, l'échelle est en fréquences absolues. En bas, l'échelle est en fréquences relatives renormalisées par la largeur des classes (*densité approchée*, qui correspond à la fréquence relative par livre).

Exemple

Exemple 3 *Concentration (en parties par million (ppm)) de métaux lourds à 259 lieux d'une région du Jura.*

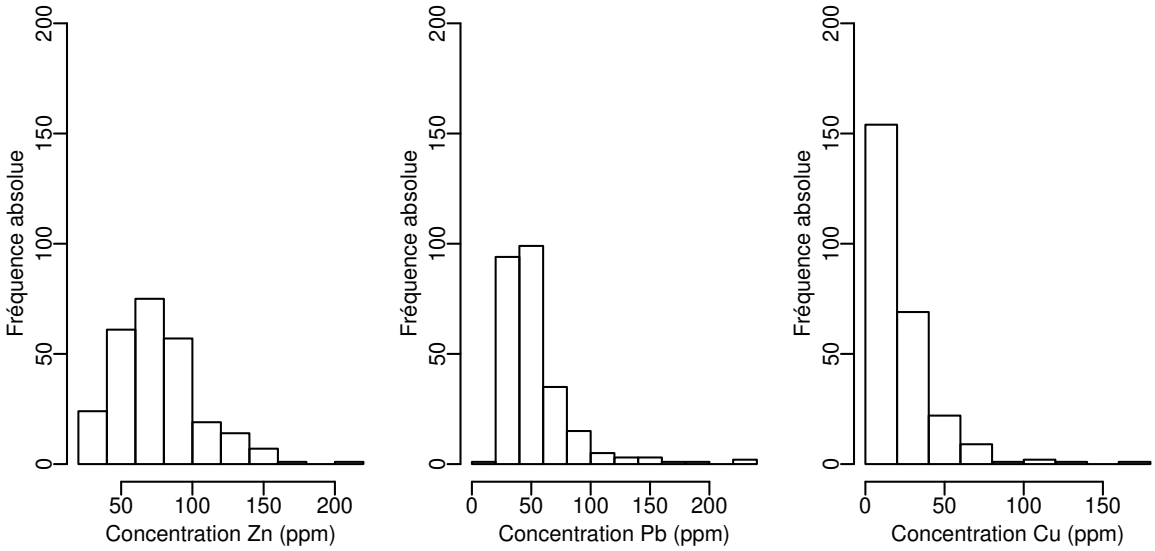
	Xloc	*Yloc*	*Cd*	*Co*	*Cr*	*Cu*	*Ni*	*Pb*	*Zn*
1	*2.39*	*3.08*	*1.74*	*9.32*	*38.32*	*25.72*	*21.32*	*77.36*	*92.56*
2	*2.54*	*1.97*	*1.33*	*10.00*	*40.20*	*24.76*	*29.72*	*77.88*	*73.56*
3	*2.81*	*3.35*	*1.61*	*10.60*	*47.00*	*8.88*	*21.40*	*30.80*	*64.80*
4	*4.31*	*1.93*	*2.15*	*11.92*	*43.52*	*22.70*	*29.72*	*56.40*	*90.00*
5	*4.38*	*1.08*	*1.56*	*16.32*	*38.52*	*34.32*	*26.20*	*66.40*	*88.40*
6	*3.24*	*4.52*	*1.15*	*3.51*	*40.40*	*31.28*	*22.04*	*72.40*	*75.20*
7	*3.92*	*3.79*	*0.89*	*15.08*	*30.52*	*27.44*	*21.76*	*60.00*	*72.40*
8	*2.12*	*3.50*	*0.53*	*4.20*	*25.40*	*66.12*	*9.72*	*141.00*	*72.08*
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Concentration de métaux : branches-et-feuilles pour Zn

2	57799
3	111223333577889
4	000012233444444556667777788899
5	00000112344445566677778889999
6	000001122222334445555555666666777789
7	011111122222344444445556666666778888888999
8	000011111122222333334444666666889
9	000000001111223455777789
10	002222244466788
11	00148
12	01334557
13	344667
14	023689
15	2
16	6
17	
18	
19	
20	
21	9

Concentration de métaux : histogrammes

Histogrammes de la concentration de Zinc (Zn), Plomb (Pb) et Cuivre (Cu).

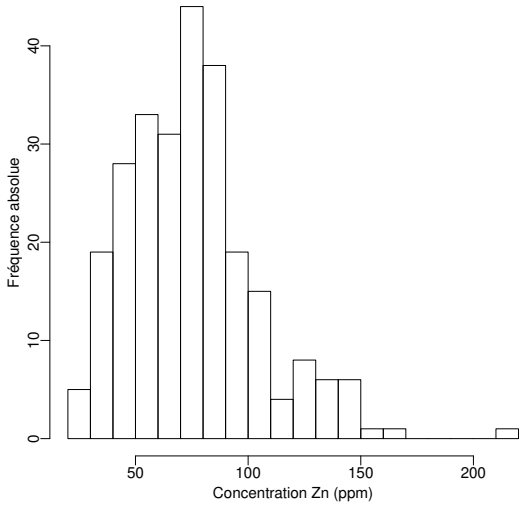


Histogramme

- **Avantage** : l'histogramme peut être utilisé tout aussi bien pour un grand nombre ou un petit nombre de données.
- **Inconvénients** :
  - Perte d'informations par rapport aux données initales en raison de l'absence des valeurs des observations.
  - Le choix de la largeur des classes est difficile. Cela mène à différentes possibilités d'interprétation !
- **Remarque** : Le diagramme branches-et-feuilles peut être vu comme un histogramme particulier obtenu par rotation. Il contient cependant davantage d'informations que ce dernier.
- **Remarque** : Il existe des versions améliorées de l'histogramme, par exemple l'estimateur à noyau de la densité.

Diagrammes branches-et-feuilles et histogrammes

2	57799
3	111223333577889
4	000012233444444455666777778889 9
5	00000112344445566677777888999 9
6	000001122222334445555555666666777 89
7	011111222223444444455566666667788888888999
8	000011111122222333334444666666688 9
9	000000001111223455777789
10	002222244466788
11	00148
12	01334557
13	344667
14	023689
15	2
16	6
17	
18	
19	
20	
21	9



Les différences entre les deux graphiques sont dues au fait que les données ont étés arrondies à l'entier le plus proche pour former le diagramme branches-et-feuilles.

Faire de bons graphiques

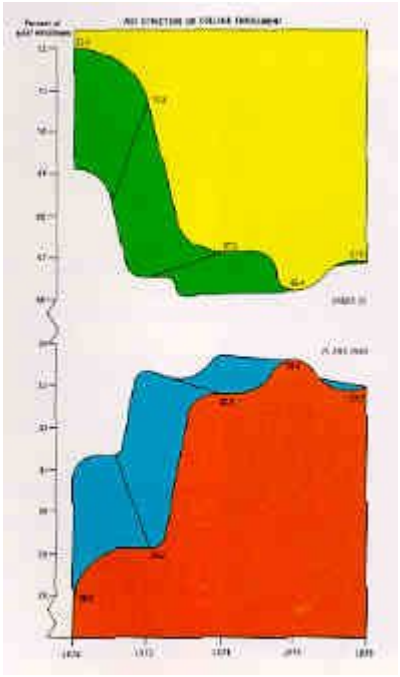
Il n'est pas facile de créer de bons graphiques. Souvent ceux générés par les logiciels standards (par exemple Excel) sont (très !) mauvais.

Quelques conseils :

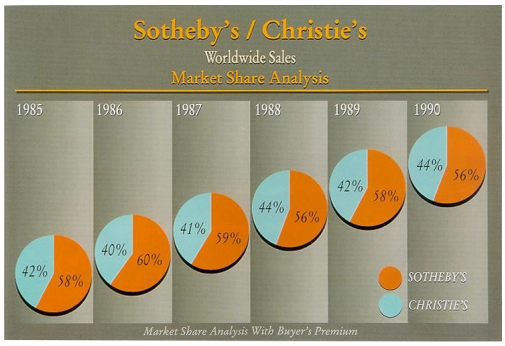
- ☐ Essayer autant que possible de montrer les données telles quelles—pas de **chartjunk** (couleurs/lignes/... inutiles).
- ☐ Indiquer variables et unités sur les axes et placer une légende claire.
- ☐ Choisir des **plages de valeurs (échelles) appropriées** pour les axes.
- ☐ Choisir les plages de valeurs sur les axes et l'aspect ratio pour que les relations systématiques apparaissent à un angle par rapport aux axes proche de $45°$.
- ☐ Faire varier l'aspect ratio peut révéler des choses intéressantes.
- ☐ Essayer de construire des graphiques de sorte que les écarts au "standard" apparaissent comme des écarts à la linéarité ou à un nuage aléatoire de points.

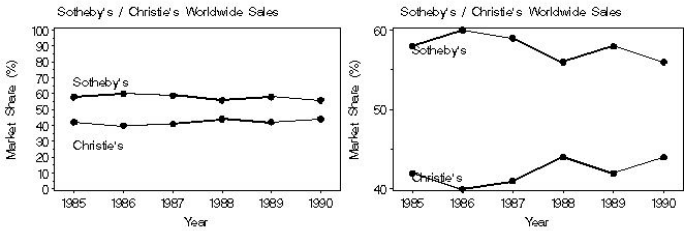
Chartjunk

Ce graphique montre 5 chiffres !



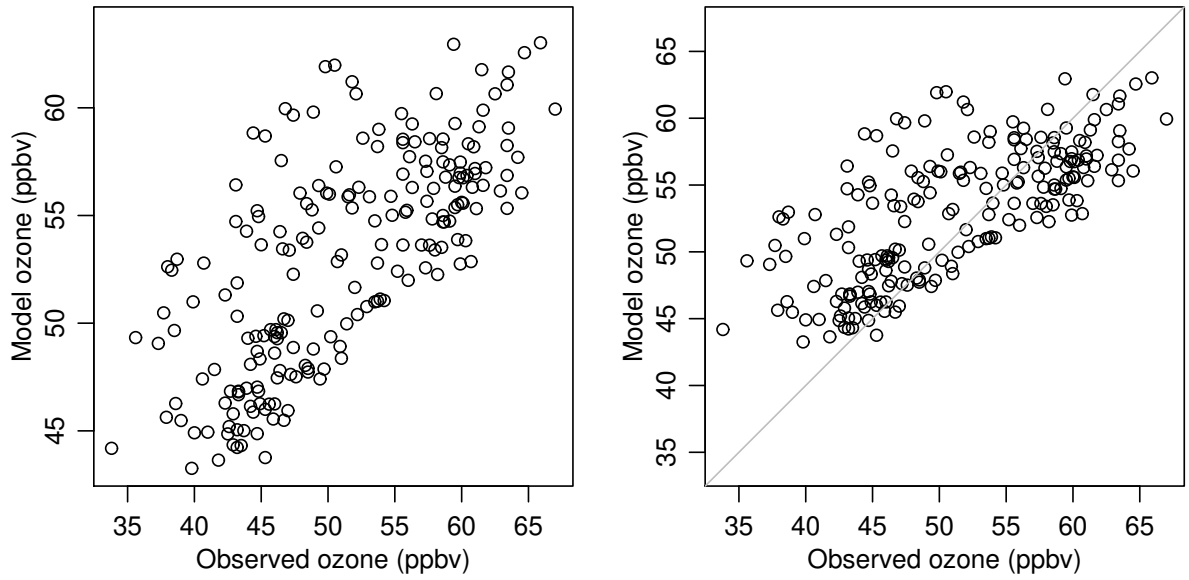
Chartjunk et plage de valeurs pour les axes





Choisir des plages de valeurs appropriées

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



La campagne russe de 1812

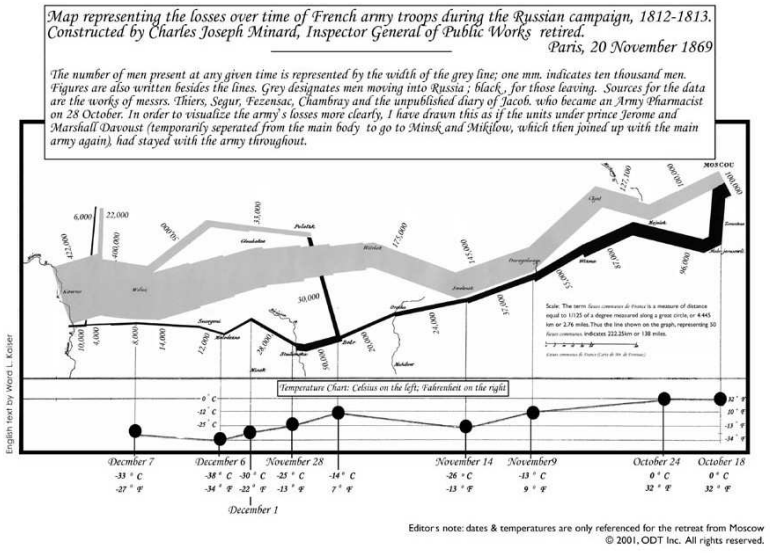


Figure 58. ***Minard s map of Napoleon s Russian campaign.***
This graphic has been translated from French to English and modified to most effectively display the temperature data.

1.3 Synthèses numériques des données

slide 38

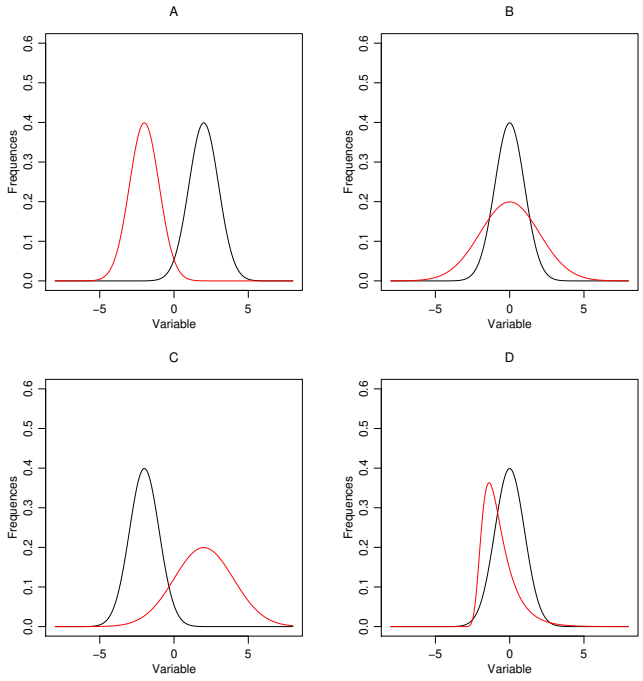
Caractéristiques principales des données

Pour les **variables quantitatives**, on s'intéresse le plus souvent aux caractéristiques suivantes :

- ☐ La **tendance centrale** qui informe sur le "milieu" (la position, le centre) des données. Des indicateurs souvent utilisés sont la moyenne et la médiane.
- ☐ La **dispersion** qui renseigne sur la variabilité des données autour de leur centre. Des indicateurs courants sont l'étendue, l'écart-type et l'étendue interquartile.
- ☐ La **symétrie** ou **asymétrie** par rapport au centre.
- ☐ Le nombre de **modes** ("bosses").

Pourquoi ?

Formes des distributions



A : Distributions semblables mais pas le même centre

B : Même centre, dispersions différentes

C : Dispersions et centres différents

D : Distribution rouge asymétrique

Tendance centrale

Indicateurs de tendance centrale (mesures de position) :

- La **moyenne** (arithmétique) est

$$\overline{x} = \frac{x_1 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

  Exemple 2 : la moyenne des poids des étudiants américains est de 145.15 lbs.

- La **médiane** : Il s'agit de 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 $\mathrm{med}(x)$.

Médiane

- Définition : $\mathrm{med}(x) = x_{(\lceil n/2 \rceil)}$, où $\lceil x \rceil$ est le plus petit entier $\geq x$.
- Données avec $n = 7$:

$$1,\ 4,\ 7,\ 9,\ 10,\ 12,\ 14 \ \Rightarrow\ \mathrm{med}(x) = x_{(\lceil 7/2 \rceil)} = x_{(4)} = 9.$$

  Données avec $n = 8$:

$$1,\ 4,\ 7,\ 9,\ 10,\ 12,\ 14,\ 25 \ \Rightarrow \mathrm{med}(x) = x_{(\lceil 8/2 \rceil)} = x_{(4)} = 9.$$

- Parfois on utilise une définition symétrique :

$$\mathrm{med}(x) = \begin{cases} x_{((n+1)/2)}, & n \text{ impaire}, \\ (x_{(n/2)} + x_{(n/2+1)})/2, & n \text{ paire}. \end{cases}$$

  Dans le cas ci-dessus avec $n = 8$, $\mathrm{med}(x) = \frac{1}{2}(x_{(4)} + x_{(4+1)}) = \frac{1}{2}(9 + 10) = 9.5$.

Moyenne et médiane

- Si la distribution est symétrique, alors la moyenne et la médiane sont proches.
- La moyenne est beaucoup plus sensible aux valeurs extrêmes (atypiques), appelées "outliers" que la médiane.
- Exemple :

$$x_1 = 1, \quad x_2 = 2, \quad x_3 = 3 \quad \Rightarrow \quad \begin{cases} \overline{x} = 2, \\ \mathrm{med}(x) = 2. \end{cases}$$

$$x_1 = 1, \quad x_2 = 2, \quad x_3 = 30 \quad \Rightarrow \quad \begin{cases} \overline{x} = 11, \\ \mathrm{med}(x) = 2. \end{cases}$$

Quantiles empiriques, quartiles

- Le concept de médiane (50%/50%) peut être généralisé en partageant les observations en quatre (ou davantage de) parties de même cardinal.
- Les bornes des classes ainsi obtenues sont appelées des **quantiles empiriques**, par exemple **quartiles** dans le cas de quatre parties.

Soit $\alpha \in (0, 1)$. Pour définir le **quantile empirique d'ordre** α, $\widehat{q}(\alpha)$, on ordonne les données

$$x_{(1)} \leq \cdots \leq x_{(n)},$$

et on calcule le nombre $n\alpha$. Si ce nombre n'est pas entier, on prend le plus petit nombre entier supérieur. On définit :

$$\widehat{q}(\alpha) = x_{(\lceil n\alpha \rceil)}.$$

Cas particulier : les **quartiles** ($\alpha = 0.25, 0.50, 0.75$, respectivement)

$\underbrace{\widehat{q}(25\%)}$ quartile inférieur (ou 1er quartile) $\quad$ $\underbrace{\widehat{q}(50\%)}$ médiane $\quad$ $\underbrace{\widehat{q}(75\%)}$ quartile supérieur (ou 3ème quartile)

Exemple

Exemple : Calcul du quantile empirique d'ordre $\alpha = 32\%$ des données suivantes ($n = 10$) :

$$27, 29, 31, 31, 31, 34, 36, 39, 42, 45.$$

On calcule

$$n\alpha = 10 \times \frac{32}{100} = 3.2 \Rightarrow \lceil 3.2 \rceil = 4 \Rightarrow \widehat{q}(32\%) = x_{(4)} = 31.$$

Indicateurs/mesures de dispersion

- **L'écart-type** :

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2} = \sqrt{\frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right)}.$$

  Il s'agit de l'indicateur le plus couramment utilisé. La quantité s^2 est la **variance empirique de l'échantillon**.

- **L'étendue** :

$$\max(x_1, \ldots, x_n) - \min(x_1, \ldots, x_n) = x_{(n)} - x_{(1)}.$$

  Ce n'est pas une mesure satisfaisante car très sensible aux valeurs extrêmes ou aberrantes (car on ne considère que les deux x_i les plus extrêmes).

- **L'écart ou étendue interquartile** :

$$\mathrm{IQR} = \widehat{q}(75\%) - \widehat{q}(25\%).$$

  Cette mesure est plus résistante aux valeurs extrêmes ou aberrantes.

1.4 Le boxplot

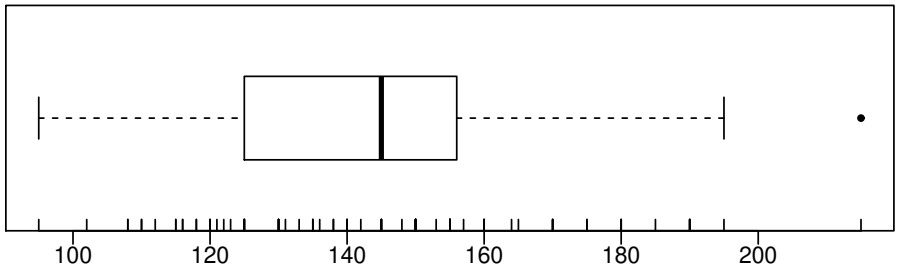
slide 47

"Five-number summary"

La liste des cinq valeurs

$$\min(x_1, \ldots, x_n) = x_{(1)},\ \widehat{q}(25\%),\ \text{médiane},\ \widehat{q}(75\%),\ \max(x_1, \ldots, x_n) = x_{(n)},$$

appelée **"five-number summary"**, donne un résumé numérique simple et pratique d'une distribution. Cette liste est à la base du "boxplot" (ou **boîte à moustache**).



Boxplot du poids des étudiants de l'école américaine.

La boîte centrale indique $\widehat{q}(25\%)$, la médiane et $\widehat{q}(75\%)$. Un point indique une valeur individuelle. Le calcul des limites de la moustache est décrit ci-dessous.

Boxplot : calcul des limites

- Poids des 92 étudiants américains. Le "five-number summary" est

$$95, \quad 125, \quad 145, \quad 156, \quad 215.$$

- On calcule

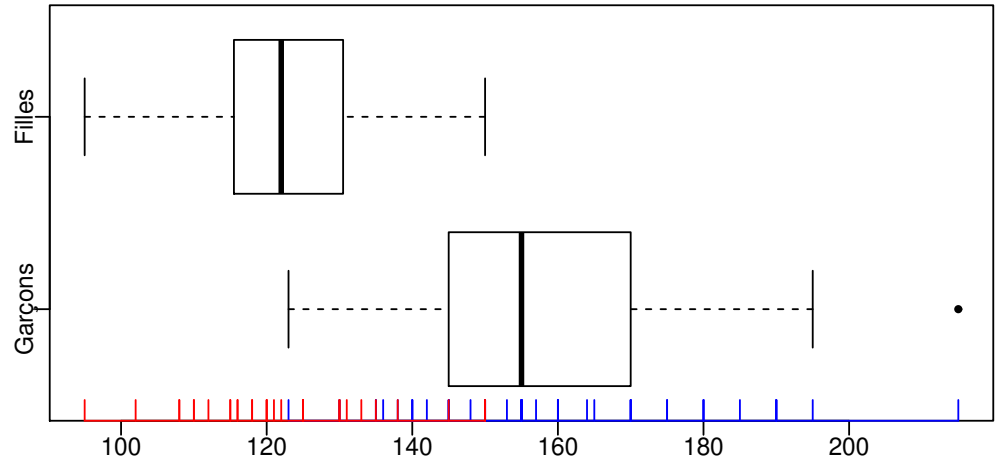
$$\begin{aligned} \text{IQR} &= \widehat{q}(75\%) - \widehat{q}(25\%) = 156 - 125 = 31, \\ C &= 1.5 \times \text{IQR} = 1.5 \times 31 = 46.5, \\ \widehat{q}(25\%) - C &= 125 - 46.5 = 78.5, \\ \widehat{q}(75\%) + C &= 156 + 46.5 = 202.5. \end{aligned}$$

- Les limites de la moustache sont respectivement le plus petit x_i supérieur à $\widehat{q}(25\%) - C$ et le plus grand x_i inférieur à $\widehat{q}(75\%) + C$.

- S'il y en a, les x_i à l'extérieur de la moustache sont indiqués individuellement.

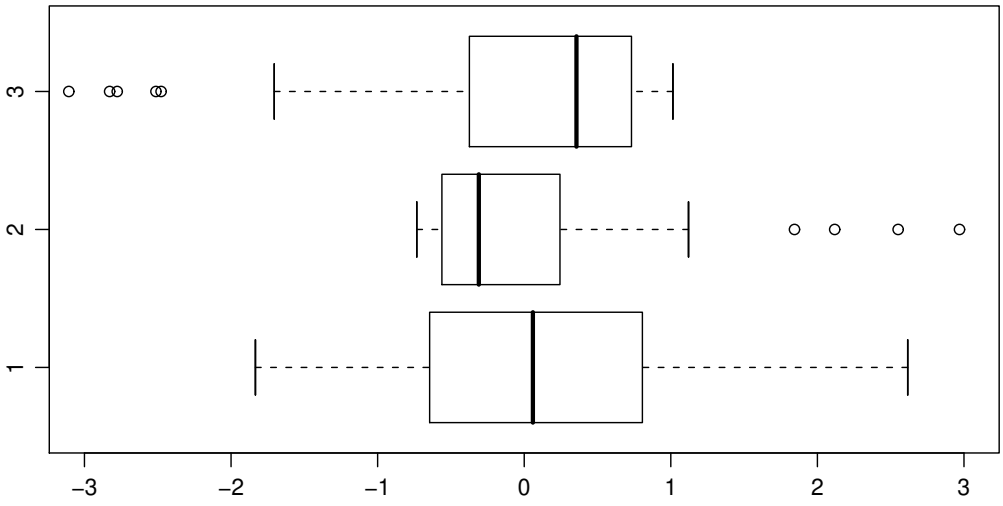
Boxplot : exemple 1

Le boxplot est très utile pour comparer plusieurs groupes d'observations :



Boxplot du poids des étudiants de l'école américaine, selon le sexe.

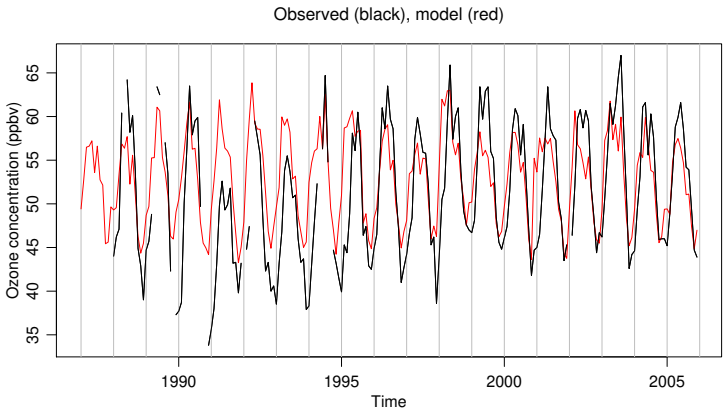
Boxplot : exemple 2



Boxplot de groupes d'observations symétriques et asymétriques.

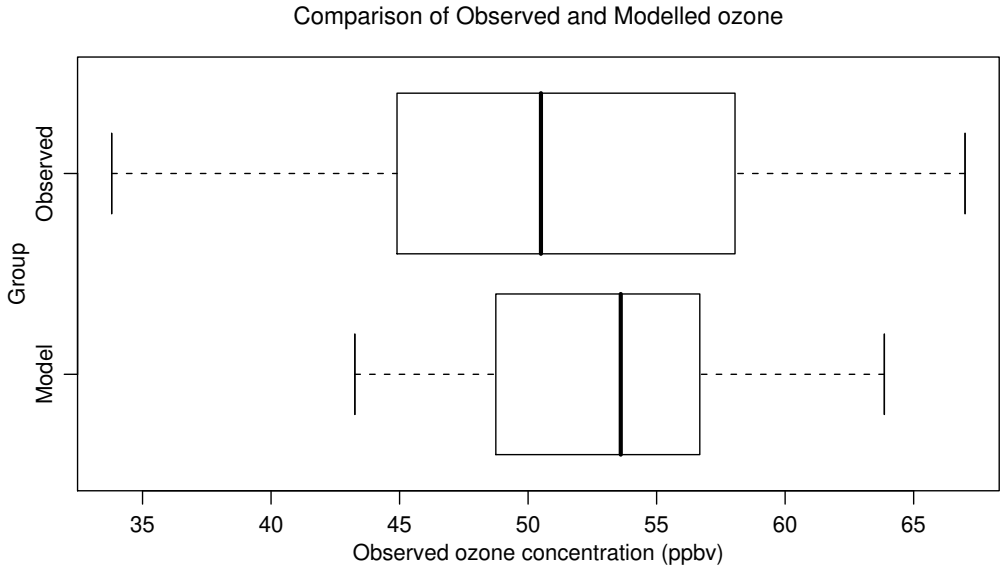
Ozone atmosphérique

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.



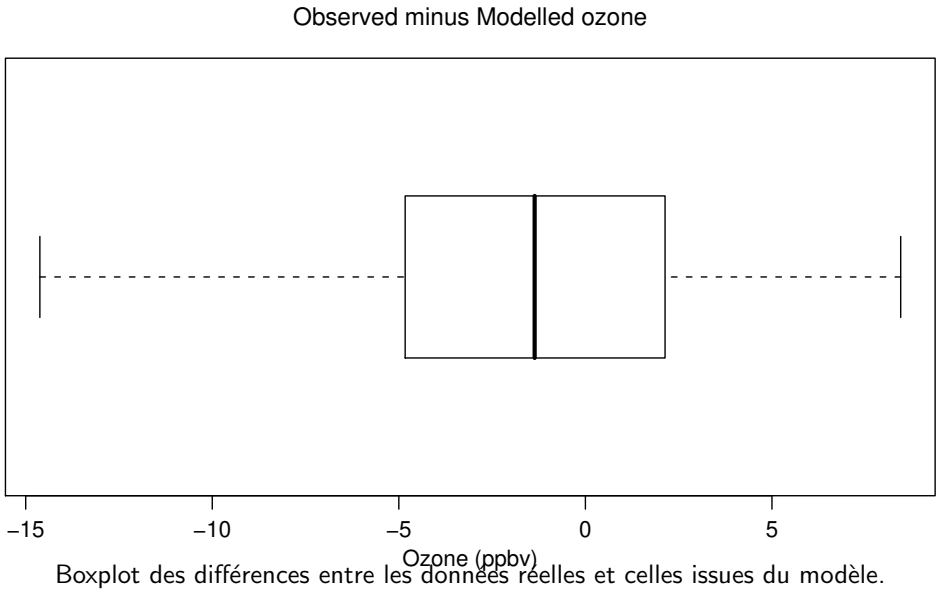
La modélisation vous paraît-elle satisfaisante ?

Ozone atmosphérique



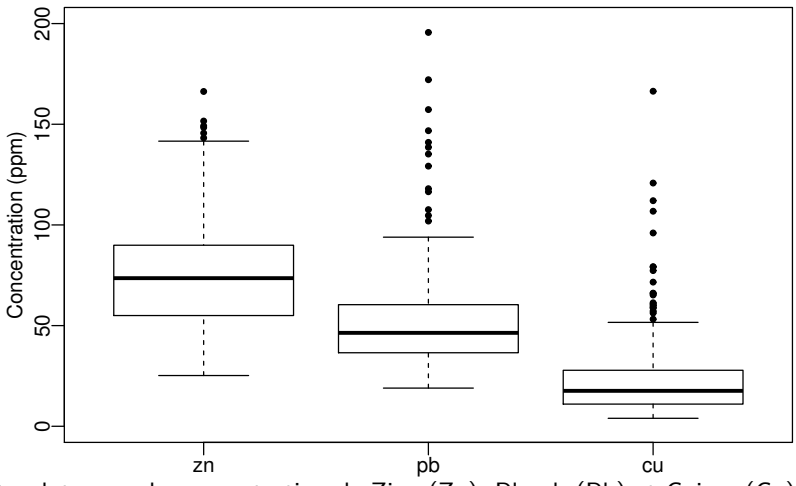
Boxplot des données réelles et de celles issues du modèle.

Ozone atmosphérique



Boxplot des différences entre les données réelles et celles issues du modèle.

Concentration de métaux



Boxplots pour la concentration de Zinc (Zn), Plomb (Pb) et Cuivre (Cu).

1.5 Stratégie

slide 56

Analyse initiale des données

Stratégie pour explorer des données issues d'une variable quantitative :

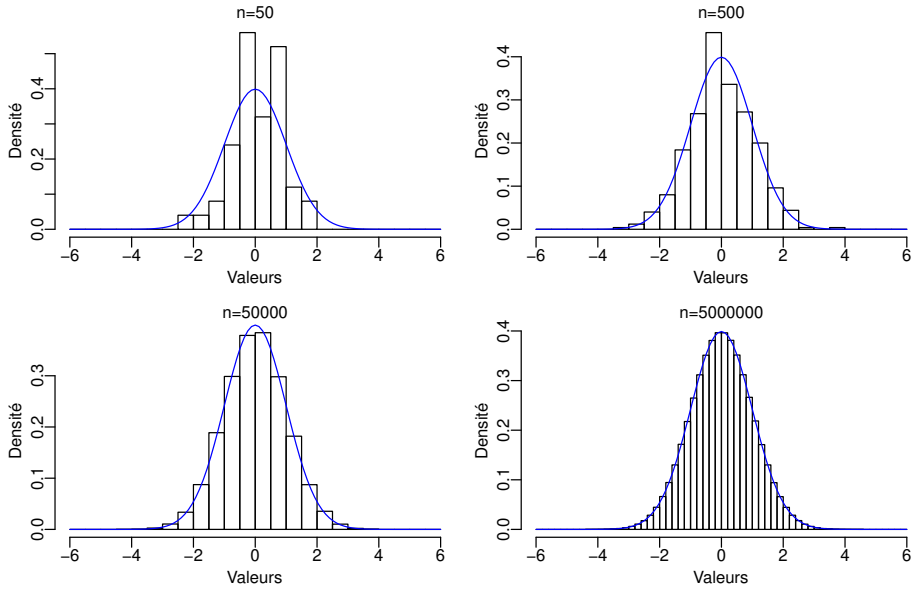
1. Toujours commencer par des **graphiques**.
2. Etudier la **structure globale** des données et identifier d'éventuelles valeurs atypiques/aberrantes ("outliers")—identifier s'il s'agit de vraies observations ou si elles résultent d'erreurs de mesure.
3. Calculer des **synthèses numériques** pour décrire la tendance centrale (position/centre/lieu) et la dispersion (échelle).

   Une étape supplémentaire très importante et utile :
4. Souvent, la structure globale est régulière et l'on peut la décrire par une courbe lisse. Il s'agit d'une **modélisation mathématique** de la distribution des données permettant de tirer des informations de ces dernières et de répondre à des questions d'intérêt.

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.
- ☐ Les modèles mathématiques pour ce type de population sont formalisés par des **courbes de densité**.
- ☐ On peut comprendre la courbe de densité comme la limite d'un histogramme décrivant la structure d'une population de taille n, quand $n \to \infty$ et quand le pas de l'histogramme tend vers 0.
- ☐ Les valeurs d'un histogramme indiquant les "densités approchées" sont ≥ 0 et l'aire d'un tel histogramme vaut 1. De même, la fonction de densité est ≥ 0 et s'intègre à 1.

Modélisation des données, courbe de densité



1.6 La loi normale

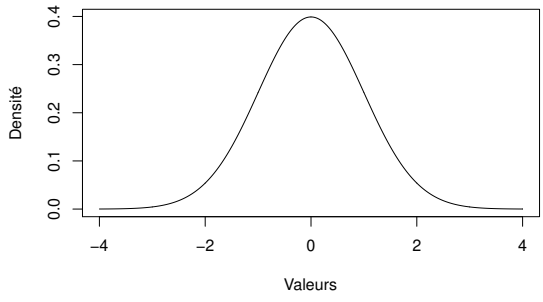
slide 60

Densité normale/gaussienne

Une densité particulièrement importante est la **densité normale/gaussienne**, associée à la distribution normale notée $\mathcal{N}(\mu, \sigma^2)$, où $\mu \in \mathbb{R}$ est la "**moyenne**" (plus rigoureusement l'espérance, cf plus tard) et $\sigma > 0$ est l'"**écart-type**" (plus rigoureusement la déviation standard, cf plus tard). Elle s'écrit

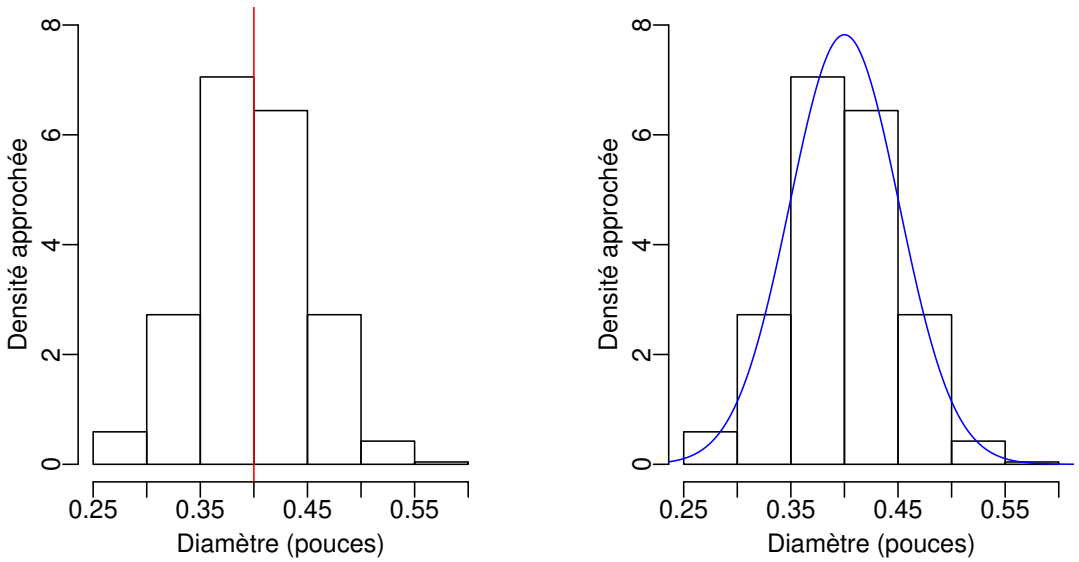
$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, \quad x \in \mathbb{R}.$$

Représentation dans le cas $\mu = 0$ et $\sigma = 1$:



Exemple : tiges en acier

Histogramme des diamètres (en pouces) de 947 tiges en acier.



Exemple : tiges en acier

- La densité précédente (en bleu) correspond à la distribution $\mathcal{N}(\mu = 0.40, \sigma^2 = 0.051^2)$.
- 472 des 947 tiges en acier ont un diamètre ≤ 0.4 pouces. Leur fréquence relative vaut donc

$$\frac{472}{947} = 0.498.$$

- L'aire correspondante sous la densité précédente (qui correspond à la probabilité donnée par le modèle) vaut 0.5. Ceci est proche de 0.498 et le modèle fournit donc une bonne approximation.

Propriétés de la distribution normale/gaussienne $\mathcal{N}(\mu, \sigma^2)$

Il y a une infinité de 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 :

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

Le modèle normal/gaussien vous 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 d'une variable aléatoire (caractérisée par sa densité) de "moyenne" μ et d'"écart-type" σ, la **valeur standardisée** de x est

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

  Alors z est une observation d'une variable aléatoire de "moyenne" 0 et d'"écart-type" 1 (expliqué dans la suite du cours), dite centrée réduite.

- ☐ Soient $x_1, \ldots, x_n$ les observations d'une certaine variable et notons $\overline{x}$ et s_x la moyenne et l'écart-type correspondants. Considérons leurs valeurs standardisées :

$$z_i = \frac{x_i - \overline{x}}{s_x}, \quad i = 1, \ldots, n.$$

  Il est facile de vérifier que leur moyenne et écart-type vérifient $\overline{z} = 0$ et $s_z = 1$.

Exemple des tiges : $n = 947$, $\overline{x} = 0.400$, $s = 0.051$, On a

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

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

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

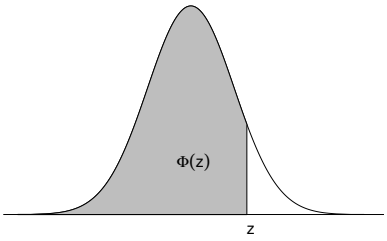
$$\mathcal{N}(\mu, \sigma^2) \mapsto \mathcal{N}(0,1).$$

La distribution $\mathcal{N}(0,1)$ est appelée **distribution normale centrée réduite** (ou encore loi normale standard). Sa densité est

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

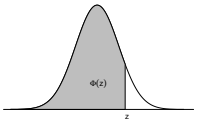
On définit aussi

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



Par symétrie de $\phi(z)$ autour de $z = 0$, $\Phi(-z) = 1 - \Phi(z)$.
De plus, la proportion d'observations dans $[z_1, z_2]$ est $\Phi(z_2) - \Phi(z_1)$.

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



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 que leur diamètre suit le modèle normal avec $\mu = \overline{x}$ et $\sigma^2 = s^2$. La proportion de x_i dans $[\overline{x} - s, \overline{x} + s]$ est la même que celle de z_i dans $[-1, 1]$ car

$$[\overline{x} - s, \overline{x} + s] \mapsto ([\overline{x} - s, \overline{x} + s] - \overline{x})/s = [-1, 1].$$

Donc la proportion recherchée est

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

De même on trouve 0.9544 pour la proportion de tiges dont le diamètre appartient à

$$[\overline{x} \pm 2s] \mapsto [-2, 2].$$

2. Probabilités slide 69

2.1 Probabilités d'événements slide 70

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.

Modèles probabilistes

- ☐ L'ensemble Ω de tous les résultats possibles d'une expérience aléatoire est appelé **ensemble fondamental**.
- ☐ Chaque élément de Ω (un résultat possible de l'expérience aléatoire) est un **événement élémentaire**.
- ☐ Tout sous-ensemble de Ω est appelé un **événement** de l'expérience aléatoire. 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 4 *Lancer d'une pièce de monnaie :*

$\Omega = \{P, F\}$.
$A = \{P\}$ = *"Pile" est un événement, et aussi un événement élémentaire.*

Exemple 5 *Lancer d'un dé :*

$\Omega = \{1, 2, 3, 4, 5, 6\}$.
$A =$ *"obtenir 1"* $= \{1\}$ *est un événement, et aussi un événement élémentaire.*
$B =$ *"obtenir un chiffre pair"* $= \{2, 4, 6\}$ *est un événement (composé).*

Operations sur les événements : intersection

- ☐ A **et** B, noté $A \cap B$ (intersection des événements A et B)
  - L'intersection de deux événements contient tous les événements élémentaires communs contenus dans les deux événements, et seulement ceux-là.
  - L'intersection est l'événement *vide* (ou *impossible*), noté $\emptyset$, si et seulement si il n'y a aucun événement élémentaire commun.
  - L'intersection d'événements est symétrique : $A \cap B = B \cap A$.

  Exemples pour le lancer d'un dé :
  - "obtenir un chiffre pair" et "obtenir un chiffre premier" :

$$\{2, 4, 6\} \cap \{2, 3, 5\} = \{2\}.$$

  - "obtenir un chiffre pair" et "obtenir 3" :

$$\{2, 4, 6\} \cap \{3\} = \emptyset.$$

Operations sur les événements : union

☐ A **ou** B, noté $A \cup B$ (union des événements A et B)

- L'union de deux événements contient tous les événements élémentaires contenus dans au moins un des deux événements.
- L'union de deux événements est l'événement *vide* (ou *impossible*) si et seulement si les deux événements sont vides.
- L'union d'événements est symétrique : $A \cup B = B \cup A$.

Exemple pour le lancer d'un dé :

- "obtenir un chiffre pair" ou "obtenir un chiffre premier" :

$$\{2,4,6\} \cup \{2,3,5\} = \{2,3,4,5,6\}.$$

Operations sur les événements : complémentaire

☐ **Pas** A, noté A^c (événement complémentaire de A)

- L'événement complémentaire de A, A^c, contient tous les événements élémentaires de Ω qui ne sont pas contenus dans A, et seulement ceux-là.
- L'événement complémentaire de A est *vide* (ou *impossible*) si et seulement si $A = \Omega$.
- Evidemment : $A \cup A^c = \Omega$, $A \cap A^c = \emptyset$.

Exemple pour le lancer d'un dé :

- Pas "obtenir un chiffre pair" :

$$\{2,4,6\}^c = \{1,3,5\}.$$

Operations sur les événements : différence

☐ A **mais pas** B, dénoté $A \setminus B = A \cap B^c$ (différence des événements A et B)

- La différence $A \setminus B$ contient tous les événements élémentaires contenus dans A sauf ceux qui sont aussi contenus dans B.
- Attention : la différence d'événements n'est en général pas symétrique !

$$A \setminus B = A \cap B^c \neq B \cap A^c = B \setminus A.$$

- $A \setminus B = \emptyset$ si et seulement si $A \subset B$.

Exemple pour le lancer d'un dé :

- "obtenir un chiffre pair" mais pas "obtenir un chiffre premier" :

$$\{2,4,6\} \setminus \{2,3,5\} = \{4,6\}.$$

Diagramme de Venn

Le diagramme de Venn est un outil simple pour visualiser les événements et les opérations entre événements.

☐ L'ensemble fondamental est représenté comme un rectangle.

☐ Les événements sont représentés comme des disques contenus dans ce rectangle.

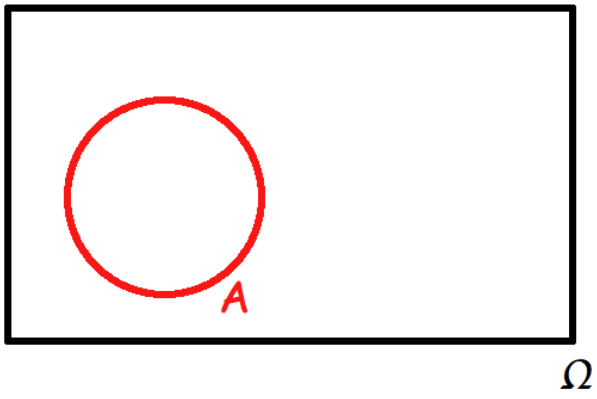
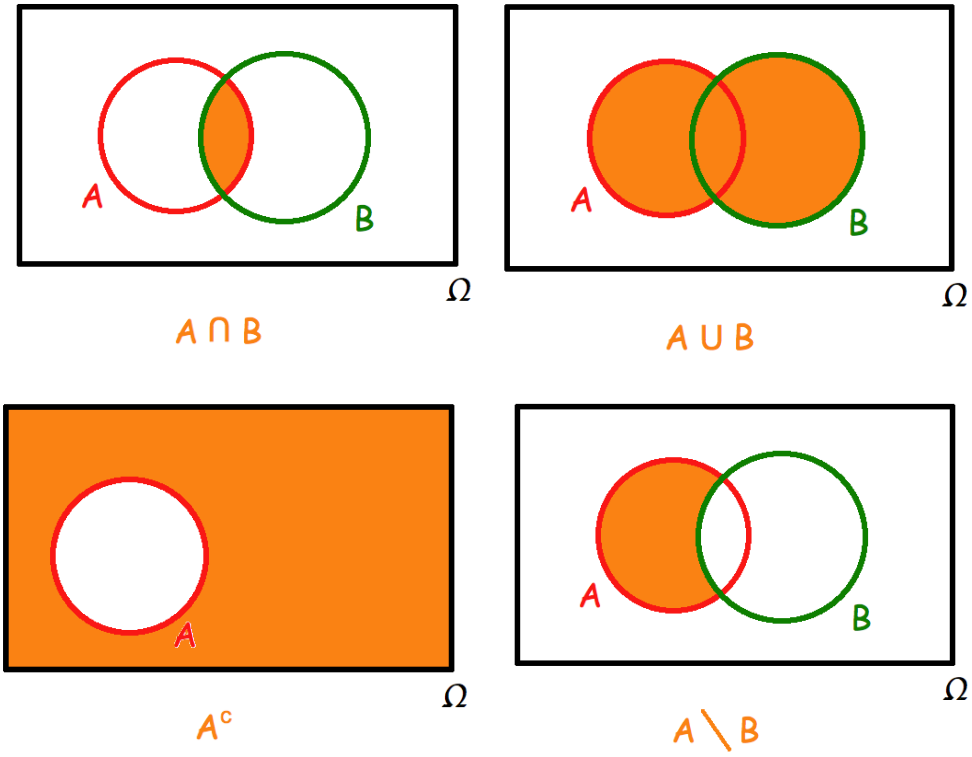


Diagramme de Venn et opérations entre événements



Propriétés d'une fonction de probabilité

Toute fonction de probabilité, notée ici $\Pr$, satisfait :

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

Exemple 6 *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 Exemple 6

On a

$$\begin{aligned} A &= \{PP, PF, FP\} \\ B &= \{FF, FP, PF\} \\ A \cap B &= \{PF, FP\} \\ A \cup B &= \{PP, PF, FP, FF\} = \Omega. \end{aligned}$$

Comme

$$A = \{PP, PF, FP\} = \{PP\} \cup \{PF\} \cup \{FP\},$$

nous obtenons

$$\Pr(A) = \Pr(\{PP\} \cup \{PF\} \cup \{FP\}) = \Pr(\{PP\}) + \Pr(\{PF\}) + \Pr(\{FP\}) = 3/4.$$

De même, on obtient $\Pr(B) = 3/4$, $\Pr(A \cap B) = 1/2$ et $\Pr(A \cup B) = 1$.

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

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

et

$$\Pr(\textit{“obtenir un nombre pair”}) = \Pr(\{2,4,6\}) = \Pr(\{2\}) + \Pr(\{4\}) + \Pr(\{6\}) = \tfrac{3}{6} = \tfrac{1}{2}.$$

Exemple 8 (Lancers de deux dés) *Trouver* $\Pr(\textit{“la somme des faces vaut 7”})$.

Solution Exemple 8

Soit A l'événement "la somme des faces vaut 7". L'ensemble Ω contient tous les 36 couples possibles, i.e.,

$$\Omega = \{(1,1),(1,2),\ldots,(6,6)\}.$$

La somme des deux faces est donnée par

D1/ D2	1	2	3	4	5	6
1	2	3	4	5	6	**7**
2	3	4	5	6	**7**	8
3	4	5	6	**7**	8	9
4	5	6	**7**	8	9	10
5	6	**7**	8	9	10	11
6	**7**	8	9	10	11	12

,

et on voit donc que $A = \{(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)\}$, qui donne, sous l'hypothèse d'équiprobabilité des événements élémentaires, $\Pr(A) = 6/36 = 1/6$.

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

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

Définition 2 *Deux événements A et B sont dits* **indépendants** *si et seulement si*

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

Une condition equivalente est : $\Pr(A \cap B) = \Pr(A) \times \Pr(B)$.

Exemples

Exemple 9 (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 10 (Lancer d'un dé) *Les événements $A = \{2, 4\}$ et $B = \{2, 4, 6\}$ sont-ils indépendants ?*

Ne pas confondre indépendance et incompatibilité (intersection vide) !
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

$$A \cap B = \emptyset \Rightarrow A \text{ et } B \text{ dépendants}, \quad \text{et ainsi,} \quad A \text{ et } B \text{ indépendants } \Rightarrow A \cap B \neq \emptyset.$$

Par ailleurs

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

Solution Exemple 9

On a

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

Soit A l'événement "obtenir pile au 1er lancer" et B l'événement "obtenir pile au 2ème lancer". On a donc $A = \{PP, PF\}$ et $B = \{PP, FP\}$, ce qui donne $A \cap B = \{PP\}$. Ainsi, sous l'hypothèse d'équiprobabilité des événements élémentaires,

$$\Pr(A) = 2/4 = 1/2, \quad \Pr(B) = 2/4 = 1/2, \quad \Pr(A \cap B) = 1/4,$$

et donc

$$\Pr(B \mid A) = \frac{\Pr(A \cap B)}{\Pr(A)} = \frac{1/4}{1/2} = \frac{1}{2} = \Pr(B).$$

Les événements A et B sont donc indépendants.

Solution Exemple 10

On a

$$\Omega = \{1, 2, 3, 4, 5, 6\}.$$

De plus, $A = \{2, 4\}$ et $B = \{2, 4, 6\}$, ce qui donne $A \cap B = \{2, 4\}$. Ainsi,

$$\Pr(A) = 1/3, \quad \Pr(B) = 1/2, \quad \Pr(A \cap B) = 1/3,$$

ce qui donne

$$\Pr(B \mid A) = \frac{\Pr(A \cap B)}{\Pr(A)} = \frac{1/3}{1/3} = 1 \neq \Pr(B).$$

Les événements A et B sont donc dépendants.
Avez-vous une idée pour voir cela plus directement ?

Indépendance : généralisation

Définition 3 *Les événements $A_1, \ldots, A_n$ sont* **indépendants** *si, pour tout sous-ensemble d'indices $\{i_1, \ldots, i_k\} \subset \{1, \ldots, n\}$, on a*

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

Exemple 11 *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 Exemple 11

Soit A_i l'événement "le composant i fonctionne", $i = 1, \ldots, n$. On a donc $\Pr(A_i) = p_i$ et $\Pr(A_i^c) = 1 - p_i$.
(a) On a, en utilisant l'indépendance des A_i,

$$\begin{aligned}
\Pr(\text{"le système fonctionne"}) &= 1 - \Pr(\text{"le système ne fonctionne pas"}) \\
&= 1 - \Pr(\text{"aucun composant ne fonctionne"}) \\
&= 1 - \Pr(A_1^c \cap A_2^c \cap \cdots \cap A_n^c) \\
&= 1 - \Pr(A_1^c)\Pr(A_2^c) \cdots \Pr(A_n^c) \\
&= 1 - \prod_{i=1}^{n} (1 - p_i).
\end{aligned}$$

(b) On a, en utilisant l'indépendance des A_i,

$$\begin{aligned}
\Pr(\text{"le système fonctionne"}) &= \Pr(\text{"tous les composants fonctionnent"}) \\
&= \Pr(A_1 \cap A_2 \cap \cdots \cap A_n) \\
&= \Pr(A_1)\Pr(A_2) \cdots \Pr(A_n) = \prod_{i=1}^{n} p_i.
\end{aligned}$$

Formule des probabilités totales

Définition 4 *Soit A un événement quelconque de Ω, et $\{B_i\}_{i=1,\ldots,n}$ une* **partition** *de Ω, c'est-à-dire,*

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

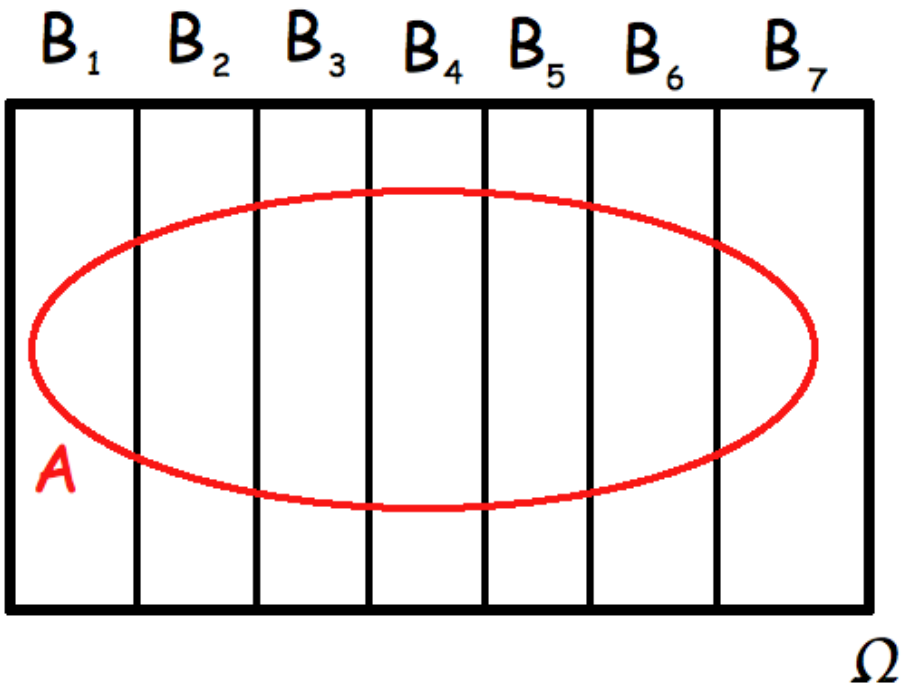
La **formule des probabilités totales** *donne*

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

Exemple 12 *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 Exemple 12

Définissons les événements : $D =$ "la pièce est défectueuse" et pour $i = 1, 2, 3$, $\tilde{M}_i =$ "la pièce a été fabriquée par M_i".
Les événements $\tilde{M}_1$, $\tilde{M}_2$ et $\tilde{M}_3$ forment une partition de l'ensemble fondamental, donc par la loi des probabilités totales,

$$\begin{aligned}
\Pr(D) &= \Pr(D \cap \tilde{M}_1) + \Pr(D \cap \tilde{M}_2) + \Pr(D \cap \tilde{M}_3) \\
&= \Pr(D \mid \tilde{M}_1)\Pr(\tilde{M}_1) + \Pr(D \mid \tilde{M}_2)\Pr(\tilde{M}_2) + \Pr(D \mid \tilde{M}_3)\Pr(\tilde{M}_3) \\
&= 5\% \times 25\% + 4\% \times 35\% + 2\% \times 40\% \\
&= 0.0345.
\end{aligned}$$

Théorème de Bayes

Théorème 1 (Bayes) *Soient $A \subset \Omega$ et $\{B_i\}_{i=1,\dots,n}$ une partition de Ω. On a, pour tout $i = 1, \dots, n$,*

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

Exemple 13 *On effectue dans une usine de production un test qui, avec probabilité 95%, détecte qu'une pièce défectueuse est défectueuse. On sait que le test donne un résultat faussement "positif" dans 1% des cas. Si 0.5% des pièces sont effectivement défectueuses, quelle est la probabilité qu'une pièce soit réellement défectueuse sachant que le test la déclare comme telle ?*

Solution Exemple 13

Soient les événements $T =$ "le test déclare la pièce défectueuse" et $D =$ "la pièce est défectueuse". On a $\Pr(T \mid D) = 0.95$ et $\Pr(T \mid D^c) = 0.01$. Par ailleurs, on sait que $\Pr(D) = 0.005$, ce qui donne $\Pr(D^c) = 1 - \Pr(D) = 0.995$. Le théorème de Bayes nous donne donc

$$\begin{aligned}
\Pr(D \mid T) &= \frac{\Pr(T \mid D)\Pr(D)}{\Pr(T)} \\
&= \frac{\Pr(T \mid D)\Pr(D)}{\Pr(T \mid D)\Pr(D) + \Pr(T \mid D^c)\Pr(D^c)} \\
&= \frac{0.95 \times 0.005}{0.95 \times 0.005 + 0.01 \times 0.995} \\
&\approx 0.323.
\end{aligned}$$

2.2 Variables aléatoires

Définition

Exemple 14 (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 5 *Soit* Ω *un ensemble fondamental. Une* **variable aléatoire** *définie sur* Ω *est une fonction de* Ω *dans* $\mathbb{R}$ *(ou dans un sous-ensemble* $H \subset \mathbb{R}$*) :*

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

où ω *est un événement élémentaire.*

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

- ☐ Somme des chiffres des faces supérieures lors du lancer de deux dés.
- ☐ Nombre de piles obtenus en n lancers d'une pièce : $H = \{0, 1, \ldots, n\}$.
- ☐ Nombre d'appels téléphoniques pendant une journée : $H = \{0, 1, \ldots\}$.
- ☐ Quantité de pluie demain : $H = \mathbb{R}_+$.

2.2.1 Variables aléatoires discrètes

Variables aléatoires discrètes

Définition 6 *Une variable aléatoire* X *est dite* **discrète** *si elle prend un nombre fini ou dénombrable de valeurs. Notons* $x_i, i = 1, 2, \ldots,$ *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 15 *On lance deux dés équilibrés et on note les chiffres des faces supérieures. Trouver : (a) la fonction de masse de la somme ; (b) la fonction de masse du maximum.*

Solution Exemple 15 (a)

L'ensemble Ω contient tous les 36 couples possibles, i.e.,

$$\Omega = \{(1,1), (1,2), \ldots, (6,6)\}.$$

La somme des deux faces est donnée par

D1/ D2	1	2	3	4	5	6
1	2	3	4	5	6	**7**
2	3	4	5	6	**7**	8
3	4	5	6	**7**	8	9
4	5	6	**7**	8	9	10
5	6	**7**	8	9	10	11
6	**7**	8	9	10	11	12

.

Soit X la variable aléatoire donnant la somme des deux nombres. La fonction de masse de X est donnée par

x_i	2	3	4	5	6	7	8	9	10	11	12	
$f_X(x_i) = \Pr(X = x_i)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$	$\Sigma = 1$

.

Solution Exemple 15 (b)

Le maximum des deux nombres est donné par

D1/ D2	1	2	3	4	5	6
1	1	2	3	**4**	5	6
2	2	2	3	**4**	5	6
3	3	3	3	**4**	5	6
4	**4**	**4**	**4**	**4**	5	6
5	5	5	5	5	5	6
6	6	6	6	6	6	6

.

Soit Y la variable aléatoire donnant le maximum des deux nombres. Sa fonction de masse est alors

y_i	1	2	3	4	5	6	
$f_Y(y_i) = \Pr(Y = y_i)$	$\frac{1}{36}$	$\frac{3}{36}$	$\frac{5}{36}$	$\frac{7}{36}$	$\frac{9}{36}$	$\frac{11}{36}$	$\Sigma = 1$

.

Fonction de répartition (cas discret ou continu)

Définition 7 *La* **fonction de répartition** F_X *d'une variable aléatoire* X *discrète ou continue est définie par*

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

Une telle fonction possède les propriétés suivantes :

- ☐ F_X prend ses valeurs dans $[0, 1]$.
- ☐ F_X est croissante.
- ☐ On a $\Pr(a < X \leq b) = F_X(b) - F_X(a)$.
- ☐ F_X est continue à droite en tout $x \in \mathbb{R}$ (voir plus loin dans le cas des variables aléatoires continues).
- ☐ Si X est une variable aléatoire discrète alors $F_X(x) = \sum_{\{i:\ x_i \leq x\}} \Pr(X = x_i), x \in \mathbb{R}$.
- ☐ Si X est une variable aléatoire discrète alors F_X est une fonction en escalier et est continue à droite en tout x_i, $i = 1, 2, \ldots$

Exemple 16 *Esquisser les fonctions de répartition correspondant à l'exemple 15 (b).*

Solution Exemple 16

Considérons la variable aléatoire Y qui donne les maximum des deux nombres. Par exemple, nous avons :

$$\begin{aligned} F_Y(4) &= \Pr(Y \leq 4) = \Pr(Y = 4) + \Pr(Y = 3) + \Pr(Y = 2) + \Pr(Y = 1) \\ &= \tfrac{7}{36} + \tfrac{5}{36} + \tfrac{3}{36} + \tfrac{1}{36} \\ &= \tfrac{16}{36}. \end{aligned}$$

De même

$$\begin{aligned} F_Y(1) &= \Pr(Y \leq 1) = \Pr(Y = 1) = \tfrac{1}{36} \\ F_Y(2) &= \tfrac{4}{36} \\ F_Y(3) &= \tfrac{9}{36} \\ F_Y(5) &= \tfrac{25}{36} \\ F_Y(6) &= 1. \end{aligned}$$

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".
- ☐ $X \dot{\sim} F$ signifie "la variable aléatoire X suit approximativement la loi F".

Loi de Bernoulli

Définition 8 *Une* **variable aléatoire de Bernoulli** *satisfait*

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

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

$X = x_i$	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 9 *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.*

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

Exemple 17 *Trouver la loi du nombre X de personnes présentes à ce cours ayant leur anniversaire ce mois-ci.*

Solution Exemple 17

Soir m le nombre de personnes présentes. On suppose que :

- ☐ les anniversaires arrivent aléatoirement durant l'année ;
- ☐ les personnes présentes sont indépendantes (pas de jumeaux, etc).

Dans ce cas, $X \sim \mathcal{B}(m, p)$, avec $p \approx 1/12$ (ou plus précisément $p = 31/365$).
Si par exemple $m = 60$ et si on prend $p = 1/12$, alors la fonction de masse de X est donnée par (calculs faits dans R avec "dbinom")

0	1	2	3	4	5	6	7
0.0054	0.0295	0.0790	0.1389	0.1800	0.1832	0.1527	0.1071

.

8	9	10	11	12	13	14	15
0.0645	0.0339	0.0157	0.0065	0.0024	0.0008	0.0002	0.0001

Loi de Poisson

Définition 10 *Une variable aléatoire X pouvant prendre pour valeurs $0, 1, 2, \ldots$ est dite de* **Poisson** *avec paramètre $\lambda > 0$ si*

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

Alors on écrit $X \sim \mathrm{Poiss}(\lambda)$.

Modélise un nombre d'événements (rares par exemple) :

- ☐ météorologie (nombre d'avalanches graves en Suisse cet hiver) ;
- ☐ télécommunications (nombre d'appels par minute dans une centrale téléphonique) ;
- ☐ finance.

Exemple 18 (E. coli) *Le niveau résiduel des bactéries E. coli dans l'eau traitée est de 2 dans 100 ml en moyenne.*
(a) Calculer la probabilité qu'il y ait un niveau résiduel de k (pour $k = 0, 1, 2, 3$) dans un échantillon de 200 ml d'eau traitée.
(b) Si on trouve $k = 10$ dans un échantillon de 200 ml d'eau quelconque, cette eau est-elle bonne ?

Solution Exemple 18

(a) Dans 200 ml la moyenne est de 4. Comme nous le verrons plus tard, la moyenne d'une variable de Poisson est égale à λ. On modélise donc le niveau résiduel à l'aide d'une loi de Poisson de paramètre $\lambda = 4$. On trouve les probabilités suivantes pour $k = 0, 1, 2, \ldots, 15$

k	0	1	2	3	4	5	6	7
p	0.0183	0.0733	0.1465	0.1954	0.1954	0.1563	0.1042	0.0595
k	8	9	10	11	12	13	14	15
p	0.0298	0.0132	0.0053	0.0019	0.0006	0.0002	0.0001	0.0000

(b) Dans de l'eau traitée, la probabilité d'observer $k = 10$ est d'environ 0.005. Plus intéressant, la probabilité d'observer $k \geq 10$ est d'environ 0.008. Ainsi il est peu vraisemblable que l'eau considérée ait été traitée.

Approximation poissonienne de la loi binomiale

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

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

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

Exemple 19 (Anniversaires) *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.

Solution Exemple 19

Nous effectuons les mêmes hypothèses que précédemment. On a

$$X \sim \mathcal{B}(m,p) \quad \text{avec} \quad m = 62 \text{ et } p = \tfrac{1}{365}.$$

Par exemple, la probabilité qu'exactement une personne parmi vous ait son anniversaire aujourd'hui est $\Pr(X = 1)$. On a

$$\Pr(X = 1) = \binom{m}{1} \tfrac{1}{365} \left(\tfrac{364}{365}\right)^{62} = 0.144.$$

L'approximation de Poisson donne

$$X \sim \text{Poiss}(\lambda = mp) \quad \text{avec} \quad \lambda = \tfrac{62}{365} = 0.1699, \quad \Pr(X = 1) = \lambda e^{-\lambda} = 0.143.$$

Pour les autres cas (j'ai utilisé R pour les calculs), pour la loi binomiale on a

$$\Pr(X = 0) = 0.84358 \quad \text{et} \quad \Pr(X > 1) = 0.01273,$$

et pour l'approximation de Poisson on trouve

$$\Pr(X = 0) = 0.84378 \quad \text{et} \quad \Pr(X > 1) = 0.01289.$$

2.2.2 Variables aléatoires continues

Variables aléatoires continues

Définition 11 *On appelle* **variable aléatoire continue** *une variable aléatoire qui peut prendre n'importe quelle valeur d'un intervalle (intervalle borné, demi-droite ou $\mathbb{R}$ tout entier).*
Le comportement d'une variable aléatoire continue X est décrit au moyen d'une fonction f_X appelée **fonction de densité** *ou simplement* **densité** *telle que*

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

où A est un ensemble de nombres réels.

Exemple 20 *Soit $A = (a, b]$ un intervalle, alors*

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

Fonctions de densité et de répartition : propriétés

☐ Propriétés essentielles 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

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

Exemple

Exemple 21 (Loi uniforme) *On choisit au hasard un nombre réel dans l'intervalle* $[0, 1]$. *Soit* X *le résultat de cette expérience.*
(a) Quelle est la distribution de X *?*
(b) Soient $0 < a < b < 1$. *Trouver* $\Pr(a < X \leq b)$.

Solution Exemple 21

(a) Par définition on a

$$F_X(x) = \Pr(X \leq x) = \begin{cases} x & \text{si} \quad 0 \leq x \leq 1 \\ 0 & \text{si} \quad x < 0 \\ 1 & \text{si} \quad x > 1. \end{cases}$$

Et donc

$$f_X(x) = F'_X(x) = \begin{cases} 1 & \text{si } \ 0 \leq x \leq 1 \\ 0 & \text{sinon.} \end{cases}$$

La quantité X est appelée variable aléatoire uniforme sur l'intervalle $[0, 1]$, ce que l'on note $X \sim U(0, 1)$.
(b) On a

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

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 $F_Z(z) = \Phi(z)$.

Exemples

Exemple 22 *Le M1 passe toutes les 12 minutes. Si j'arrive à un moment choisi au hasard, quelle est la probabilité que je doive attendre (a) plus de 8 minutes ? (b) moins de 2 minutes ? (c) entre 3 et 6 minutes ?*

Exemple 23 *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.*

Exemple 24 *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 ?*

Solution Exemple 22

On modélise le temps d'attente par une loi uniforme $T \sim U(0, 12)$. On a

$$\Pr(T > 8) = \int_8^\infty f_T(u)\mathrm{d}u = \int_8^{12} \frac{1}{12}\mathrm{d}u = 4/12 = 1/3.$$

Par ailleurs,

$$\Pr(T \leq 2) = \int_{-\infty}^2 f_T(u)\mathrm{d}u = \int_0^2 \frac{1}{12}\mathrm{d}u = 2/12 = 1/6.$$

$$\Pr(3 < T \leq 6) = \int_3^6 f_T(u)\mathrm{d}u = \int_3^6 \frac{1}{12}\mathrm{d}u = 3/12 = 1/4 = 0.25.$$

On peut également obtenir ces résultats à l'aide la fonction de répartition. Dans le cas de la loi uniforme sur $[a, b]$, $U(a, b)$, on a, pour $a \leq x \leq b$,

$$F_X(x) = \int_{-\infty}^x f_X(u)\mathrm{d}u = \int_a^x 1/(b-a)\mathrm{d}u = \frac{x-a}{b-a}.$$

Pour $x < a$, $F_X(x) = 0$ et pour $x > b$, $F_X(x) = 1$.

Solution Exemple 23

(a) Soient A et B les événements "il pleut demain" et "il pleut au plus 5mm demain". Tout d'abord, nous calculons la fonction de répartition de la loi exponentielle. Si $X \sim \exp(\lambda)$,

$$F_X(x) = \int_0^x \lambda e^{-\lambda t} dt = \left[-e^{-\lambda t}\right]_0^x = 1 - e^{-\lambda x}.$$

Maintenant, la loi des probabilités totales nous donne

$$\begin{aligned}\Pr(B) &= \Pr(B \mid A)\Pr(A) + \Pr(B \mid A^c)\Pr(A^c) \\ &= \{1 - \exp(-0.05 \times 5)\}0.2 + 1 \times 0.8 = 0.844.\end{aligned}$$

(b) Soit C l'événement "au moins 2mm tombent". Alors

$$\begin{aligned}\Pr(C) &= \Pr(C \mid A)\Pr(A) + \Pr(C \mid A^c)\Pr(A^c) \\ &= \exp(-0.05 \times 2) \times 0.2 + 0 \times 0.8 = 0.181.\end{aligned}$$

Solution Exemple 24

Soit $Z \sim N(0,1)$. On a

$$\begin{aligned}\Pr(135 < X \leq 150) &= \Pr\left(\tfrac{135-140}{4} < \tfrac{X-140}{4} \leq \tfrac{150-140}{4}\right) \\ &= \Pr(-1.25 \leq Z \leq 2.5) \\ &= \Phi(2.5) - \{1 - \Phi(1.25)\} \\ &= 0.9938 - (1 - 0.8944) = 0.8882\end{aligned}$$

en utilisant la table de la loi normale (ou alors plus simplement R).

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

- La fonction de répartition conjointe vérifie

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

- Propriétés essentielles 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.
  - $\sum_{i,j} f_{X,Y}(x_i,y_j) = 1$.

Cas continu : propriétés

☐ La fonction de répartition conjointe vérifie

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

☐ Propriétés essentielles de la densité conjointe :

– $$f_{X,Y}(x,y) \ge 0, \quad x,y \in \mathbb{R}.$$

– $$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(u,v)\mathrm{d}v\mathrm{d}u = 1.$$

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

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

Lois marginales

Définition 12 *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 25 *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 Exemple 25

On a

$$f_X(1) = \sum_j f_{X,Y}(1, y_j) = f_{X,Y}(1,2) + f_{X,Y}(1,4) = 2/5.$$

Le même raisonnement nous permet d'obtenir

$X = x_i$	1	2	3
$f_X(x_i)$	2/5	1/5	2/5

et

$Y = y_j$	2	3	4
$f_Y(y_j)$	2/5	1/5	2/5

Indépendance

Définition 13 *Deux variables aléatoires discrètes X et Y prenant des valeurs x_i et y_j sont dites* **indépendantes** *si et seulement si pour tout x_i et y_j,*

$$\Pr(X = x_i, Y = y_j) = \Pr(X = x_i) \times \Pr(Y = y_j).$$

Dans le cas continu, X et Y sont indépendantes si et seulement si

$$f_{X,Y}(x, y) = f_X(x) \times f_Y(y), \quad \textit{pour tout } x \textit{ et } y \in \mathbb{R},$$

ce qui est équivalent à

$$F_{X,Y}(x, y) = F_X(x) \times F_Y(y), \quad \textit{pour tout } x \textit{ et } y \in \mathbb{R}.$$

Donc, si X et Y sont indépendantes et l'on connaît f_X et f_Y, alors $f_{X,Y}$ est connue.

Exemple 26 *Les variables aléatoires X, Y de l'exemple 25 sont-elles indépendantes ?*

Définition 14 *On écrit $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f$ pour dire que $X_1, \ldots, X_n$ sont des variables aléatoires* **indépendantes et identiquement distribuées** *de densité f.*

Exemple 27 *Soient $X_1, X_2 \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$. Trouver leur densité conjointe. Si $\mu = 3$ et $\sigma^2 = 4$, trouver* $\Pr(X_1 \leq 1, -1 < X_2 \leq 5)$.

Solution Exemple 27

Par indépendance, la densité conjointe s'écrit

$$f_{X_1,X_2}(x_1,x_2) = f_{X_1}(x_1) \times f_{X_2}(x_2).$$

Ainsi

$$\begin{aligned}
&\Pr(X_1 \leq 1, -1 \leq X_2 \leq 5) \\
&= \int_{x_1=-\infty}^{1} \int_{x_2=-1}^{5} f_{X_1}(x_1) f_{X_2}(x_2) \mathrm{d}x_1 \mathrm{d}x_2 \\
&= \int_{x_1=-\infty}^{1} f_{X_1}(x_1) \mathrm{d}x_1 \times \int_{x_2=-1}^{5} f_{X_2}(x_2) \mathrm{d}x_2 \\
&= \Pr(X_1 \leq 1) \Pr(-1 < X_2 \leq 5) \\
&= \Pr\left(\frac{X_1-\mu}{\sigma} \leq \frac{1-\mu}{\sigma}\right) \Pr\left(\frac{-1-\mu}{\sigma} < \frac{X_2-\mu}{\sigma} \leq \frac{5-\mu}{\sigma}\right) \\
&= \Phi(-1) \times [\Phi(1) - \Phi(-2)] \\
&= \Phi(-1) \times [\Phi(1) - (1 - \Phi(2))] \\
&= 0.1299.
\end{aligned}$$

Densité conditionelle

Définition 15 *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

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

Exemple 28 *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. Les deux variables sont-elles indépendantes ?

Solution Exemple 28

Pour $x \in (0,1)$, on a

$$f_X(x) = \int_0^1 f_{X,Y}(x,y)\mathrm{d}y = \int_0^1 (x+y)\mathrm{d}y = \left[xy + \tfrac{y^2}{2}\right]_0^1 = x + \tfrac{1}{2}.$$

De même, pour $y \in (0,1)$,

$$f_Y(y) = \int_0^1 f_{X,Y}(x,y)\mathrm{d}x = \int_0^1 (x+y)\mathrm{d}x = \cdots = y + \tfrac{1}{2}.$$

Pour $x \notin (0,1)$, on a $f_X(x) = 0$ et pour $y \notin (0,1)$, $f_Y(y) = 0$. Enfin, pour $x, y \in (0,1)$,

$$f_{Y|X}(y \mid x) = \tfrac{f_{X,Y}(x,y)}{f_X(x)} = \tfrac{x+y}{x+1/2} \neq f_Y(y).$$

Donc X et Y ne sont pas indépendantes ! On peut aussi vérifier que $f_{X,Y}(x,y) \neq f_X(x)f_Y(y)$.

2.3 Quantités caractéristiques

slide 131

Mesure de tendance centrale : espérance

Définition 16 *L'***espérance** *d'une variable aléatoire X est définie par*

$$\mathrm{E}(X) = \begin{cases} \sum_i x_i f_X(x_i) & \textit{si } X \textit{ est discrète} \\ \int_{-\infty}^{\infty} x f_X(x)\mathrm{d}x & \textit{si } X \textit{ est continue.} \end{cases}$$

Interprétations :

- ☐ Interprétation 1 : somme des valeurs possibles multipliées par leurs probabilités théoriques.
- ☐ Interprétation 2 (physique) : centre de gravité d'un ensemble de masses (somme des positions des masses multipliées par leur masse normalisée).

Propriétés de l'espérance

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

Propriétés :

- ☐ Pour toute fonction g, on a (théorème de transfert)

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

- ☐ Pour toutes constantes $a, b \in \mathbb{R}$, on a $\mathrm{E}(aX + b) = a\mathrm{E}(X) + b$.
- ☐ Si X et Y sont deux variables aléatoires et $g : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, on définit $\mathrm{E}\{g(X, Y)\}$ comme ci-dessus à partir de la fonction de masse ou densité conjointe.
- ☐ Si X et Y sont deux variables aléatoires, alors $\mathrm{E}(X + Y) = \mathrm{E}(X) + \mathrm{E}(Y)$.
- ☐ Si $X_1, \ldots, X_n$ sont des variables aléatoires, alors $\mathrm{E}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \mathrm{E}(X_i)$.
- ☐ Si X, Y sont indépendantes et g, h des fonctions quelconques, alors

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

Exemples

Exemple 29 *Soit* $X \sim \mathcal{B}(m = 3, p = 0.1)$. *Calculer* $\mathrm{E}(X)$.

Exemple 30 *Soit* $X \sim \mathrm{Poiss}(\lambda)$. *Calculer* $\mathrm{E}(X)$ *et* $\mathrm{E}(X^2)$.

Exemple 31 *Soit* $X \sim \mathcal{N}(\mu, \sigma^2)$. *Calculer* $\mathrm{E}(X)$.

Exemple 23 (suite) Calculer l'espérance de la quantité de pluie de demain.

Solution Exemple 29

On a

$$f_X(x) = \Pr(X = x) = \binom{3}{x} 0.1^x (1 - 0.1)^{3-x}, \quad x = 0, 1, 2, 3,$$

x_i	0	1	2	3
$f_X(x_i)$	0.729	0.243	0.027	0.001

Donc

$$\mathrm{E}(X) = \sum_i x_i f_X(x_i) = 0 + 1 \times 0.243 + 2 \times 0.027 + 3 \times 0.001 = 0.3.$$

Dans le cas général, si $X \sim \mathcal{B}(m, p)$ alors on peut écrire $X = \sum_{i=1}^m Y_i$, où $Y_1, \ldots, Y_m \overset{\text{iid}}{\sim} \mathcal{B}(p)$. On en déduit donc que

$$\mathrm{E}(X) = \sum_{i=1}^m \mathrm{E}(Y_i) = m\mathrm{E}(Y_1) = m(p \times 1 + 0 \times (1 - p)) = mp.$$

Solution Exemple 30

Si $X \sim$Poiss(λ), alors

$$f_X(x) = \tfrac{\lambda^x}{x!} e^{-\lambda}, \qquad x = 0, 1, 2, 3, \ldots$$

Alors, en effectuant le changement de variable $u = x - 1$, on obtient

$$\mathrm{E}(X) = \sum_{x=0}^{\infty} x \tfrac{\lambda^x}{x!} e^{-\lambda} = 0 + \sum_{x=1}^{\infty} x \tfrac{\lambda^x}{x!} e^{-\lambda} = \lambda \sum_{x=1}^{\infty} \tfrac{\lambda^{x-1}}{(x-1)!} e^{-\lambda} = \lambda e^{-\lambda} \sum_{u=0}^{\infty} \tfrac{\lambda^u}{u!} = \lambda.$$

De la même façon,

$$\begin{aligned} \mathrm{E}(X^2) &= \sum_{x=0}^{\infty} x^2 \tfrac{\lambda^x}{x!} e^{-\lambda} = \sum_{x=1}^{\infty} x^2 \tfrac{\lambda^x}{x!} e^{-\lambda} = \lambda \sum_{x=1}^{\infty} x \tfrac{\lambda^{x-1}}{(x-1)!} e^{-\lambda} \\ &= \lambda \sum_{u=0}^{\infty} (u+1) \tfrac{\lambda^u}{u!} e^{-\lambda} = \lambda \mathrm{E}(X+1) = \lambda(\mathrm{E}(X) + 1) = \lambda(\lambda + 1). \end{aligned}$$

Solution Exemple 31

En effectuant le changement de variable $z = (x-\mu)/\sigma$ (qui donne $x = \mu + \sigma z$ et donc $\mathrm{d}x = \sigma \mathrm{d}z$), on a

$$\begin{aligned}
\mathrm{E}(X) &= \int_{-\infty}^{\infty} (x-\mu+\mu) f_X(x) \mathrm{d}x \\
&= \int_{-\infty}^{\infty} (x-\mu) f_X(x) \mathrm{d}x + \int_{-\infty}^{\infty} \mu f_X(x) \mathrm{d}x \\
&= \int_{-\infty}^{\infty} (x-\mu) \tfrac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \mathrm{d}x + \mu \times 1 \\
&= \int_{-\infty}^{\infty} \tfrac{z}{\sqrt{2\pi}} e^{-z^2/2} \sigma \mathrm{d}z + \mu \\
&= \mu,
\end{aligned}$$

car l'intégrande est une fonction impaire.

Solution Exemple 23 (suite)

Soit $X \sim \exp(\lambda)$. On a

$$\begin{aligned}
\mathrm{E}(X) &= \int_{-\infty}^{\infty} x f_X(x) \mathrm{d}x = \int_0^{\infty} x\lambda e^{-\lambda x} \mathrm{d}x = \lambda \int_0^{\infty} x e^{-\lambda x} \mathrm{d}x \\
&= \lambda \left(\left[-\frac{1}{\lambda} e^{-\lambda x} x \right]_0^{\infty} - \int_0^{\infty} -\frac{1}{\lambda} e^{-\lambda x} \mathrm{d}x \right) = \lambda \left[\frac{1}{\lambda} \int_0^{\infty} e^{-\lambda x} \mathrm{d}x \right] \\
&= \int_0^{\infty} e^{-\lambda x} \mathrm{d}x = \left[-\frac{1}{\lambda} e^{-\lambda x} \right]_0^{\infty} = \frac{1}{\lambda}.
\end{aligned}$$

Soit Y la quantité de précipitation demain et A l'événement "il pleut demain". On a

$$\mathrm{E}(Y) = \mathrm{E}(Y|A)\mathrm{Pr}(A) + \mathrm{E}(Y|A^c)\mathrm{Pr}(A^c) = \frac{1}{0.05} \times 0.2 = 4 \text{ mm}.$$

Mesure de dispersion : variance

Définition 17 *La* **variance** *d'une variable aléatoire X est définie par*

$$\mathrm{Var}(X) = \mathrm{E}[\{X - \mathrm{E}(X)\}^2] = \cdots = \mathrm{E}(X^2) - \mathrm{E}(X)^2.$$

Propriétés :

- ☐ $\mathrm{Var}(X) \geq 0$.
- ☐ $\mathrm{Var}(X) = 0$ implique que X est constante.
- ☐ La **déviation standard** de X est définie par $\mathrm{sd}(X) = \sqrt{\mathrm{Var}(X)} \geq 0$.
- ☐ Pour toutes constantes $a, b \in \mathbb{R}$, on a $\mathrm{Var}(aX + b) = a^2\mathrm{Var}(X)$.
- ☐ Si X et Y sont indépendantes, alors $\mathrm{Var}(X \pm Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$.

Exemple 32 *Si $X \sim \mathrm{Poiss}(\lambda)$, montrer que $\mathrm{Var}(X) = \lambda$.*

Exemple 33 *Si $X \sim \mathcal{B}(m, p)$, montrer que $\mathrm{Var}(X) = mp(1-p)$.*

Exemple 34 *Si $X \sim \mathcal{N}(\mu, \sigma^2)$, montrer que $\mathrm{Var}(X) = \sigma^2$.*

Solution Exemples 32 et 33

Soit $X \sim \mathrm{Poiss}(\lambda)$. On a vu que $\mathrm{E}(X) = \lambda$ et $\mathrm{E}(X^2) = \lambda(\lambda + 1)$. On a donc

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - [\mathrm{E}(X)]^2 = \lambda(\lambda + 1) - \lambda^2 = \lambda.$$

Soit $X \sim \mathcal{B}(m, p)$. On a $X = \sum_{i=1}^m Y_i$, où $Y_1, \ldots, Y_m \overset{\text{iid}}{\sim} \mathcal{B}(p)$. Si $Y \sim \mathcal{B}(p)$, on a $\mathrm{E}(Y^2) = 1 \times p + 0 \times (1-p) = p$ donc

$$\mathrm{Var}(Y) = \mathrm{E}(Y^2) - [\mathrm{E}(Y)]^2 = p - p^2 = p(1-p).$$

En utilisant l'indépendance des Y_i, on obtient

$$\mathrm{Var}(X) = \sum_{i=1}^m \mathrm{Var}(Y_i) = m\mathrm{Var}(Y_1) = mp(1-p).$$

Solution Exemple 34

Soit $X \sim \mathcal{N}(\mu, \sigma^2)$. On a vu que $\mathrm{E}(X) = \mu$. Ainsi, en utilisant le changement de variable $z = (x-\mu)/\sigma$ (qui donne $\mathrm{d}x = \sigma \mathrm{d}z$), on obtient

$$\begin{aligned}
\mathrm{Var}(X) &= \int_{-\infty}^{\infty} (x-\mu)^2 \times \tfrac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \mathrm{d}x \\
&= \int_{-\infty}^{\infty} \sigma^2 z^2 \times \tfrac{1}{\sqrt{2\pi}\sigma} e^{-\frac{z^2}{2}} \sigma \mathrm{d}z \\
&= \sigma^2 \int_{-\infty}^{\infty} z \times z \tfrac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \mathrm{d}z \\
&= \sigma^2 \left(\left[z \times \left(-\tfrac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} \right) \right]_{-\infty}^{\infty} - \int_{-\infty}^{\infty} -\tfrac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \mathrm{d}z \right) \\
&= \sigma^2.
\end{aligned}$$

Covariance

Définition 18 *La* **covariance** *entre les variables aléatoires X et Y est une mesure de dépendance entre elles définie par*

$$\mathrm{Cov}(X,Y) = \mathrm{E}\left[\{X - \mathrm{E}(X)\}\{Y - \mathrm{E}(Y)\}\right] = \cdots = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y).$$

Propriétés :

- ☐ $\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$, $\mathrm{Cov}(X,X) = \mathrm{Var}(X)$.
- ☐ $\mathrm{Cov}(X+Y,Z) = \mathrm{Cov}(X,Z) + \mathrm{Cov}(Y,Z)$.
- ☐ Pour $a,b,c,d \in \mathbb{R}$, $\mathrm{Cov}(aX+b, cY+d) = ac\mathrm{Cov}(X,Y)$.
- ☐ $\mathrm{Cov}(\cdot,\cdot)$ peut être considérée comme un produit scalaire.
- ☐ Du fait de la bilinéarité, la valeur de la covariance dépend des unités de mesure de X et Y.
- ☐ $\mathrm{Var}(X \pm Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) \pm 2\mathrm{Cov}(X,Y)$.
- ☐ Si X et Y sont indépendantes, alors $\mathrm{Cov}(X,Y) = 0$. Mais attention, l'inverse n'est pas vraie en général !

Exemple

Exemple 35 *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* $\mathrm{Cov}(X,Y)$.

Solution Exemple 35

En utilisant le résultat de l'exemple 28 pour la densité marginale de X, on obtient, pour $r \geq 1$,

$$\mathrm{E}(X^r) = \int_{-\infty}^{\infty} x^r f_X(x)\mathrm{d}x = \int_0^1 x^r(x+\tfrac{1}{2})\mathrm{d}x = \left[\tfrac{x^{r+2}}{r+2}\right]_0^1 + \tfrac{1}{2}\left[\tfrac{x^{r+1}}{r+1}\right]_0^1 = \tfrac{1}{r+2} + \tfrac{1}{2(r+1)}.$$

Ainsi, les lois marginales de X et Y étant identiques, on a $\mathrm{E}(X) = \mathrm{E}(Y) = 7/12$, $\mathrm{E}(X^2) = \mathrm{E}(Y^2) = 5/12$, et donc $\mathrm{Var}(X) = \mathrm{Var}(Y) = 60/144 - 49/144 = 11/144$.
Pour la covariance et la corrélation on calcule

$$\begin{aligned} \mathrm{E}(XY) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy f_{X,Y}(x,y)\mathrm{d}y\mathrm{d}x = \int_0^1 \left[\int_0^1 xy(x+y)\mathrm{d}y\right]\mathrm{d}x \\ &= \int_0^1 \left[x^2\tfrac{y^2}{2} + x\tfrac{y^3}{3}\right]_0^1 \mathrm{d}x = \int_0^1 (\tfrac{x^2}{2} + \tfrac{x}{3})\mathrm{d}x = \left[\tfrac{x^3}{6} + \tfrac{x^2}{6}\right]_0^1 = 1/3 \end{aligned}$$

et on en déduit $\mathrm{Cov}(X,Y) = 1/3 - 49/144 = -1/144$

Corrélation

Définition 19 *La* **corrélation** *entre X et Y est une mesure de dépendance entre X et Y définie par*

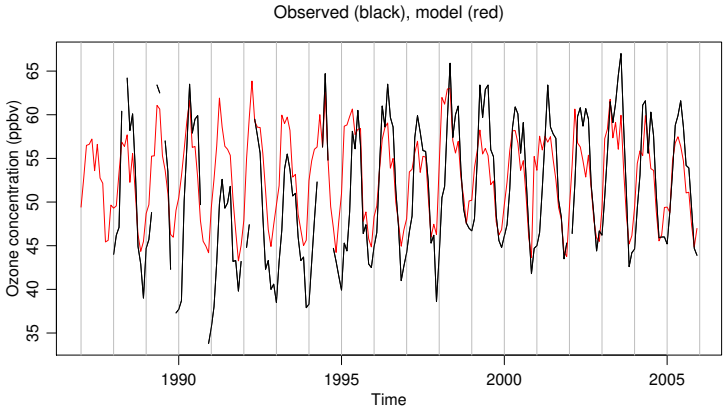
$$\rho_{X,Y} = \rho(X,Y) = \mathrm{Corr}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}.$$

Propriétés :

- ☐ $\rho_{X,Y}$ est une mesure de dépendance **linéaire** (seulement linéaire !) entre X et Y.
- ☐ $\mathrm{Corr}(X,Y) = \mathrm{Corr}(Y,X)$.
- ☐ $\mathrm{Corr}(X,X) = 1$.
- ☐ $\mathrm{Corr}(X,-X) = -1$.
- ☐ Pour $a,b,c,d \in \mathbb{R}$, $\mathrm{Corr}(aX+b, cY+d) = \mathrm{sgn}(ac)\mathrm{Corr}(X,Y)$, où sgn est la fonction signe.
- ☐ $-1 \leq \mathrm{Corr}(X,Y) \leq 1$ (conséquence de l'inégalité de Cauchy–Schwarz).
- ☐ Si X et Y sont indépendantes, alors $\mathrm{Corr}(X,Y) = 0$, mais la réciproque est fausse !

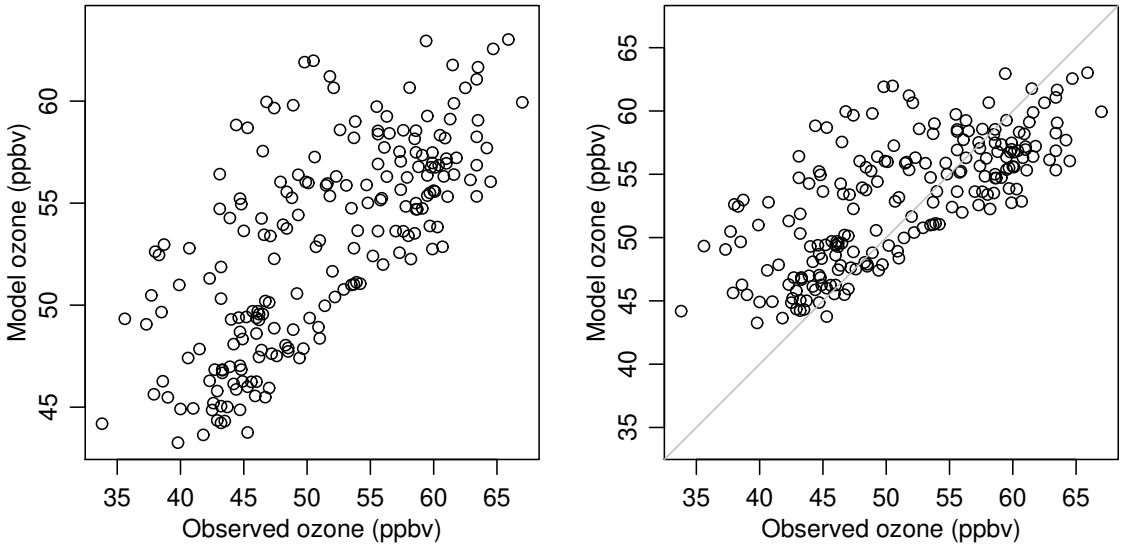
Exemple : ozone atmosphérique

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 ?

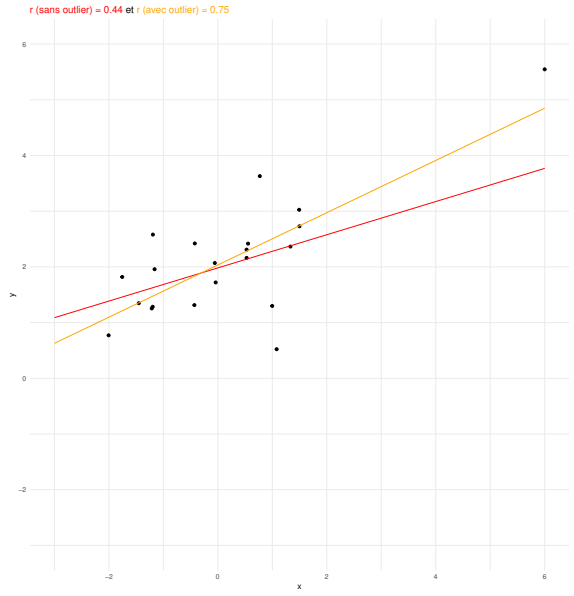
Exemple : ozone atmosphérique



La corrélation empirique est $\rho = 0.707$.

Erreurs fréquentes dans l'interpétation de la corrélation

☐ Valeurs aberrantes et anomalies : les anomalies peuvent fausser la corrélation et donc certaines conclusions



Erreurs fréquentes dans l'interpétation de la corrélation

☐ Taille de l'échantillon : La taille des données peut affecter de manière significative la fiabilité de la mesure de corrélation

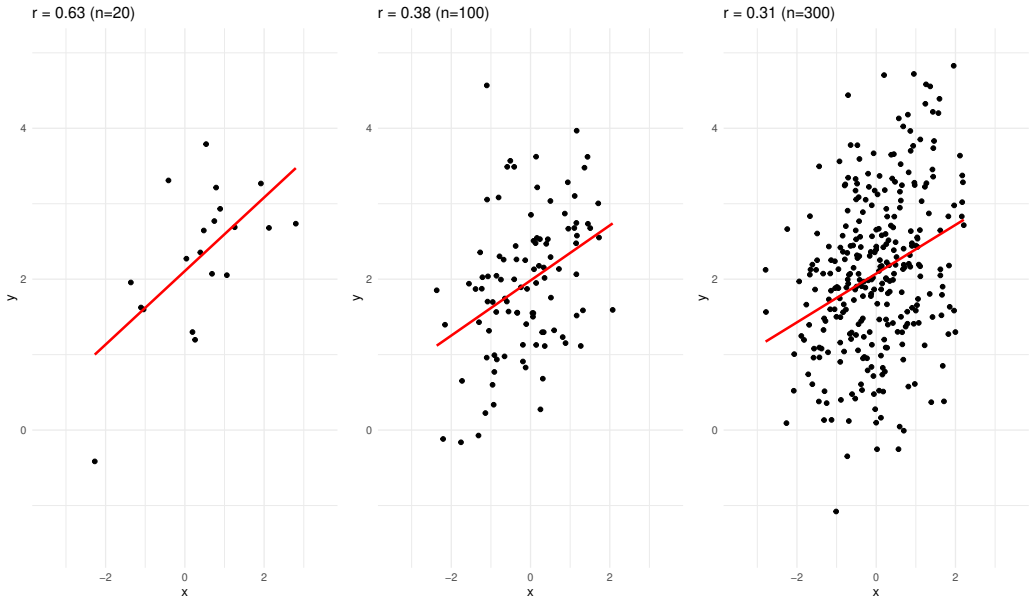
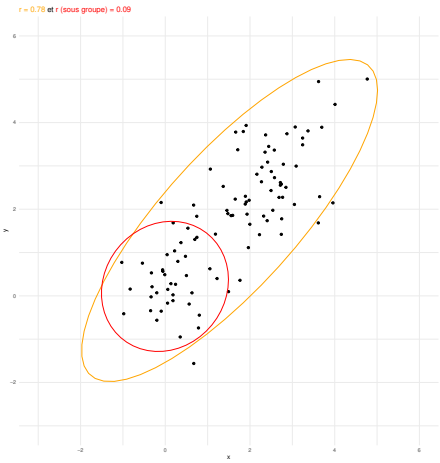


FIGURE 1 – Données simulées avec une vraie corrélation de 0.3

Erreurs fréquentes dans l'interprétation de la corrélation

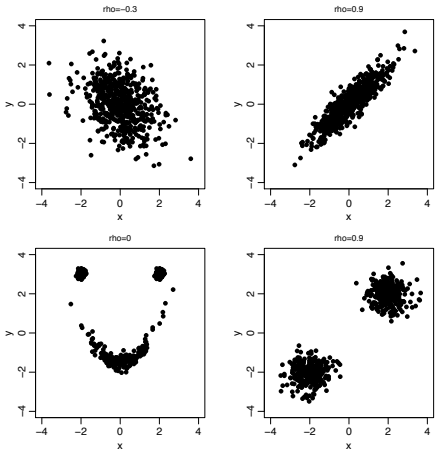
☐ Etendue des variables : les données n'incluent qu'une sous-catégorie des valeurs possibles d'une variable



⇒ Toujours inspecter le nuage de dispersion pour évaluer la présence d'une relation linéaire, de valeurs aberrantes ou encore de sous-groupes !

Limitations de la corrélation

- ☐ r_{xy} mesure la dépendance linéaire (panneaux supérieurs)
- ☐ On peut avoir $r_{xy} \approx 0$, mais dépendance forte mais non-linéaire (en bas à gauche)
- ☐ Une corrélation pourrait être forte mais spécieuse, comme en bas à droite, où deux sous-groupes, chacun sans corrélation, sont combinés
- ☐ Corrélation $\neq$ causalité !



Corrélation parasite

Des variables non liées peuvent être fortement corrélées

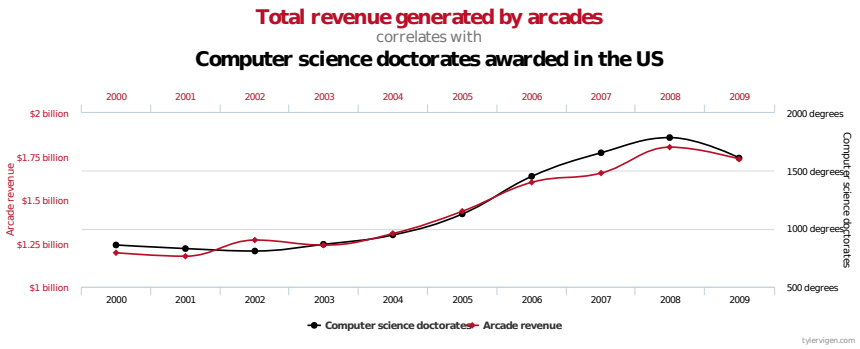


FIGURE 2 – Tirée de `https://www.tylervigen.com/spurious-correlations`

sans présence de lien causal...

⇒ important de prendre en compte le contexte global lors de l'interprétation des corrélations

Corrélation parasite

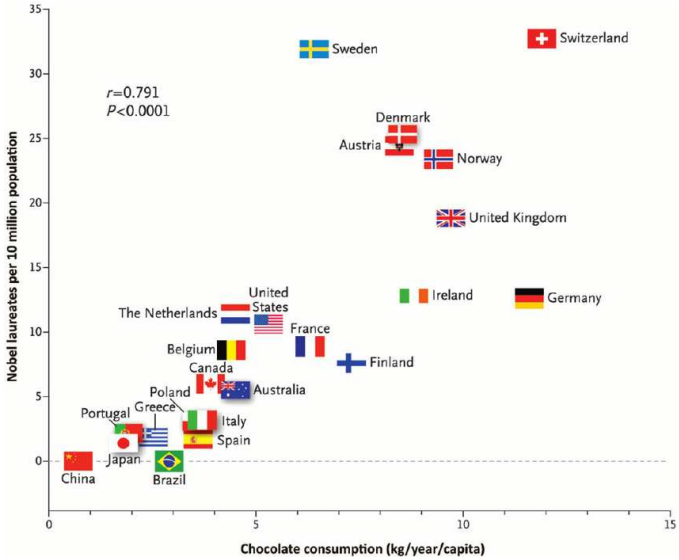
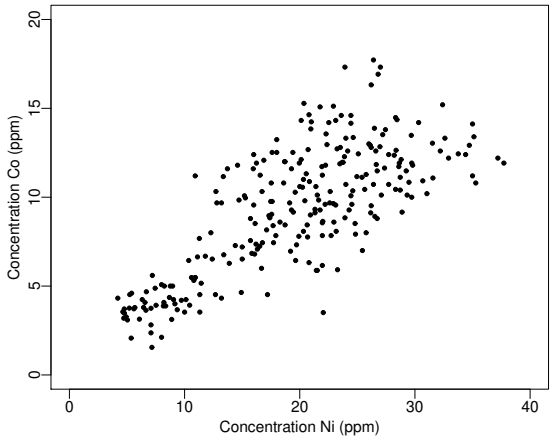


FIGURE 3 – Publiée dans Messerli (2012) Chocolate Consumption, Cognitive Function, and Nobel Laureates, New England Journal of Medicine.

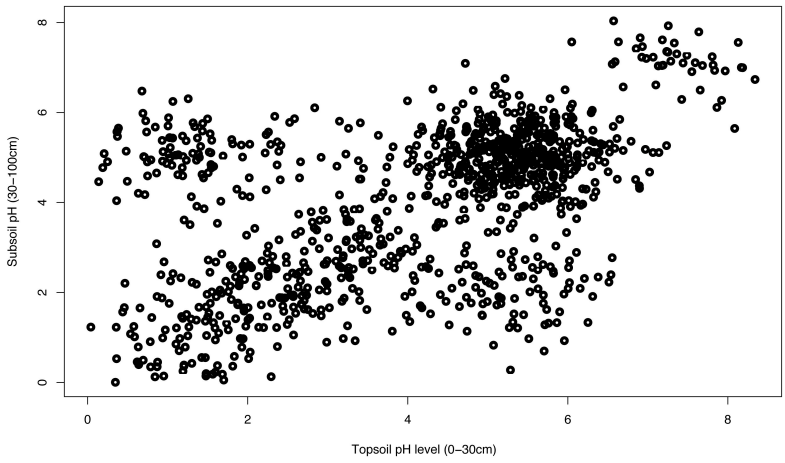
⇒ facteurs socio-économique, saisonniers, ou encore culturels peuvent influencer les données et donc l'interprétation de la corrélation

Exemple : concentration de métaux



Ici $\rho = 0.75$.

Exemple : acidité du sol



Ici $\rho = 0.5$.

Quantiles

Soit X une variable aléatoire et $\alpha \in (0, 1)$.

- Le quantile de X au niveau α, noté $q_X(\alpha)$, est défini par

$$q_X(\alpha) = \inf\{x \in \mathbb{R} : F_X(x) \geq \alpha\}.$$

- Si X est une variable aléatoire continue à support en un seul morceau, alors $q_X(\alpha)$ est l'unique solution de l'équation

$$F_X(x) = \alpha,$$

et donc

$$q_X(\alpha) = F_X^{-1}(\alpha).$$

- Les quantiles empiriques définis en Section 1.3 sont des estimations (cf les prochains cours) des quantiles à partir des données à disposition.

2.4 Théorèmes fondamentaux

slide 157

Approche expérimentale

Considérons l'expérience : on lance une pièce de monnaie 10'000 fois et on observe le nombre de "Face" obtenus.
Soient $X_1, \ldots, X_n$ des variables aléatoires indépendantes telles que

$$X_i = \begin{cases} 1 & \text{si le } i\text{-ème jet donne "Face"} \\ 0 & \text{si le } i\text{-ème jet donne "Pile"}, \end{cases}$$

et soit p est la probabilité d'obtenir "Face" (succès). Alors $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{B}(p)$. La quantité $X_1 + \cdots + X_n$ représente le nombre de "Face" obtenu en n lancers, et donc

$$X_1 + \cdots + X_n \sim \mathcal{B}(n, p).$$

La proportion de "Face" obtenue en n lancers est $\overline{X} = (X_1 + \cdots + X_n)/n$. Donc

$$\begin{aligned} \mathrm{E}(\overline{X}) &= n^{-1}\mathrm{E}(X_1 + \cdots + X_n) = n^{-1}np = p, \\ \mathrm{Var}(\overline{X}) &= n^{-2}\mathrm{Var}(X_1 + \cdots + X_n) = n^{-2}np(1-p) = p(1-p)/n \to 0, \end{aligned}$$

quand $n \to \infty$.

Loi des grands nombres

Exemple 36 *Soient $X_1, \ldots, X_n$ des variables indépendantes telles que $\mathrm{E}(X_i) = \mu < \infty$ et $0 < \mathrm{Var}(X_i) = \sigma^2 < \infty$, $i = 1, \ldots, n$. Trouver $\mathrm{E}(\overline{X})$ et $\mathrm{Var}(\overline{X})$, et montrer que $\mathrm{Var}(\overline{X}) \to 0$ pour $n \to \infty$.*

Solution Exemple 36
On a

$$\mathrm{E}(\overline{X}) = \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\mathrm{E}\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}(X_i) = \mu.$$

De plus, en utilisant l'indépendance des X_i,

$$\mathrm{Var}(\overline{X}) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) = \frac{\sigma^2}{n} \to 0.$$

Loi des grands nombres

Théorème 2 (Loi forte des grands nombres, LGN) *Soient $X_1, \ldots, X_n$ des variables aléatoires indépendantes et identiquement distribuées d'espérance μ finie, et soit*

$$\overline{X} = \frac{X_1 + \cdots + X_n}{n}.$$

On a

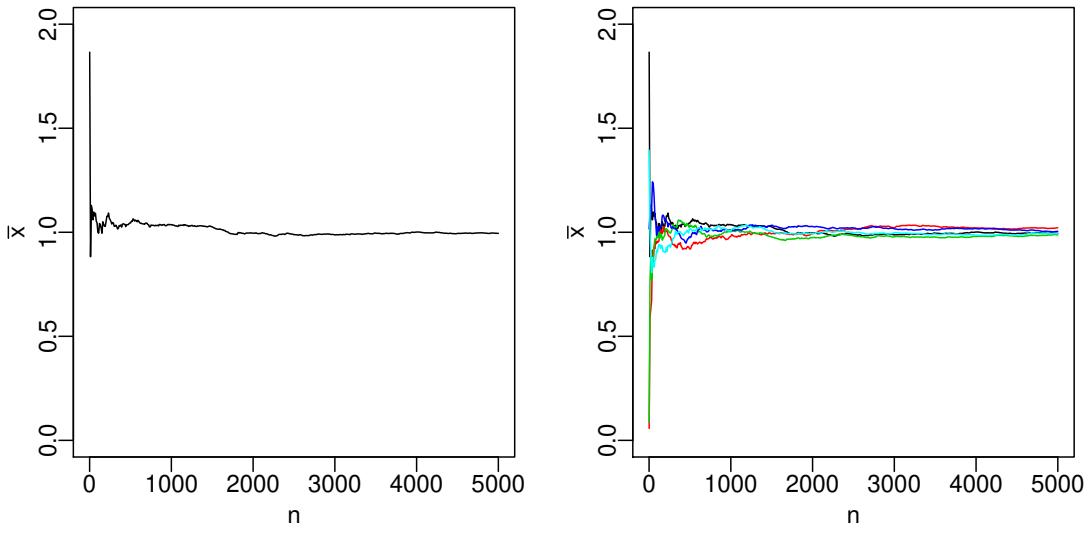
$$\Pr\left(\lim_{n\to\infty} \overline{X} = \mu\right) = 1.$$

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

De plus $\mathrm{Var}(\overline{X}) \to 0$ si les variances des $X_i, i = 1, \ldots, n$, sont finies.

Illustration de la LGN

Illustration pour des variables aléatoires distribuées selon $\exp(1)$.
A gauche : une simulation ; à droite : cinq simulations.



Théorème central limite

Supposons que les variables aléatoires $X_1, \ldots, X_n$ sont indépendantes et identiquement distribuées, d'espérance $\mu < \infty$ et variance $0 < \sigma^2 < \infty$. Soit

$$\overline{X} = \frac{X_1 + \cdots + X_n}{n}.$$

Il est facile de voir que $\mathrm{E}(\overline{X}) = \mu$ et $\mathrm{Var}(\overline{X}) = \sigma^2/n$. La version centrée réduite de $\overline{X}$ est donc

$$Z_n = \frac{\overline{X} - \mathrm{E}(\overline{X})}{\sqrt{\mathrm{Var}(\overline{X})}} = \sqrt{n}\left(\frac{\overline{X} - \mu}{\sigma}\right).$$

Théorème 3 (Théorème central limite, TCL) *Soient $X_1, \ldots, X_n$ des variables aléatoires indépendantes et identiquement distribuées telles que $\mathrm{E}(X_i) = \mu < \infty$ et $0 < \mathrm{Var}(X_i) = \sigma^2 < \infty$, $i = 1, \ldots, n$. Alors, pour tout $z \in \mathbb{R}$,*

$$\lim_{n\to\infty} \Pr(Z_n \leq z) = \Phi(z).$$

Donc pour n grand, on a $\overline{X} \dot{\sim} \mathcal{N}(\mu, \sigma^2/n)$, et $X_1 + \cdots + X_n \dot{\sim} \mathcal{N}(n\mu, n\sigma^2)$.
Une caractéristique remarquable du TCL réside dans le fait que l'approximation par la loi normale est vraie quelle que soit la loi des X_i dès lors qu'ils sont iid et ont une espérance finie et une variance finie et strictement positive.

Illustration du TCL

Illustration pour des variables aléatoires $\exp(1)$:

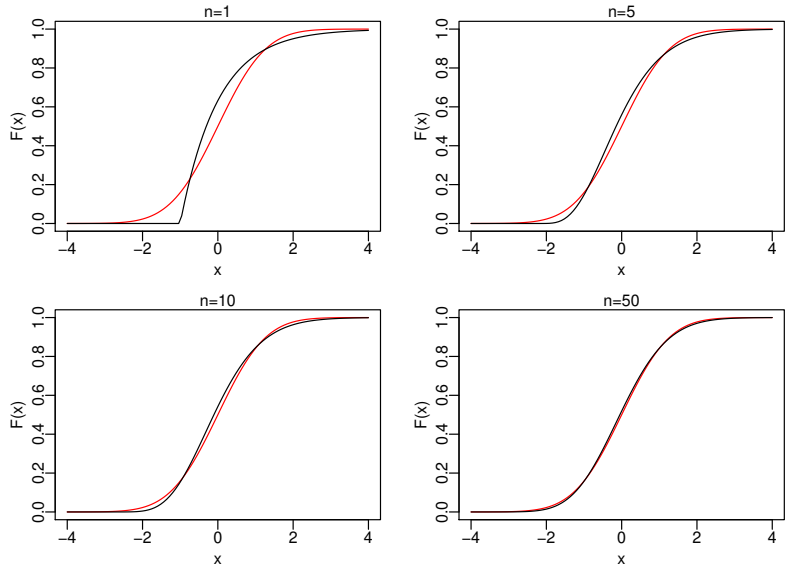
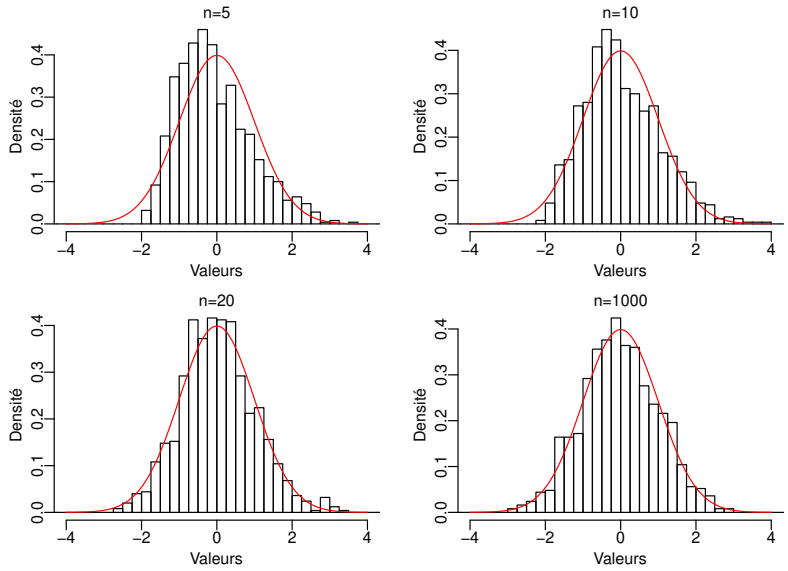


Illustration du TCL

Illustration pour des variables aléatoires $\exp(1)$:



Exemples

Exemple 37 *Soit $X \sim \mathcal{B}(m,p)$. Donner une approximation de $\Pr(X \leq r)$, pour $r \in \mathbb{R}$.*

Solution Exemple 37 :

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

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

Exemple

Exemple 38 *Soient* $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \exp(\lambda)$. *Donner une approximation de*

$$\Pr(X_1 + \cdots + X_n \leq x), \quad x \in \mathbb{R}.$$

Solution Exemple 38 :
Nous savons que $\mathrm{E}(X_1) = 1/\lambda$. De plus, il est possible de montrer que $\mathrm{Var}(\mathrm{X}_1) = 1/\lambda^2$. Ainsi, pour n grand, le TCL donne $S_n = X_1 + \ldots + X_n \dot{\sim} \mathcal{N}(n/\lambda, n/\lambda^2)$. Ainsi

$$\Pr(S_n \leq x) = \Pr\left(\frac{S_n - n/\lambda}{\sqrt{(n/\lambda^2)}} \leq \frac{x - n/\lambda}{\sqrt{(n/\lambda^2)}}\right) \approx \Phi\left(\frac{x - n/\lambda}{\sqrt{(n/\lambda^2)}}\right).$$

Modèles statistiques

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

- **modèle statistique** : on modélise la quantité étudiée (par exemple la taille de l'espèce humaine) par une variable aléatoire X dont la densité (on suppose qu'elle existe) f est supposée connue à l'exception d'un paramètre θ (vecteur de dimension finie) non-aléatoire ;
- **échantillon** (doit être représentatif de la population) : "données" $x_1, \ldots, x_n$, souvent supposées comme étant une réalisation de $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f$;
- **statistique** : une fonction $T = g(X_1, \ldots, X_n)$ des variables aléatoires $X_1, \ldots, X_n$;
- **estimateur** : une statistique utilisée pour estimer certains paramètres de f.
- **Notations** :

$T = g(X_1, \ldots, X_n)$	est la statistique (variable aléatoire);
$t = g(x_1, \ldots, x_n)$	est la **réalisation (valeur observée)** de T au moyen des x_i;
$\widehat{\theta}$	est un **estimateur** (variable aléatoire) d'un paramètre θ.

Commentaires

Exemple 39 *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*

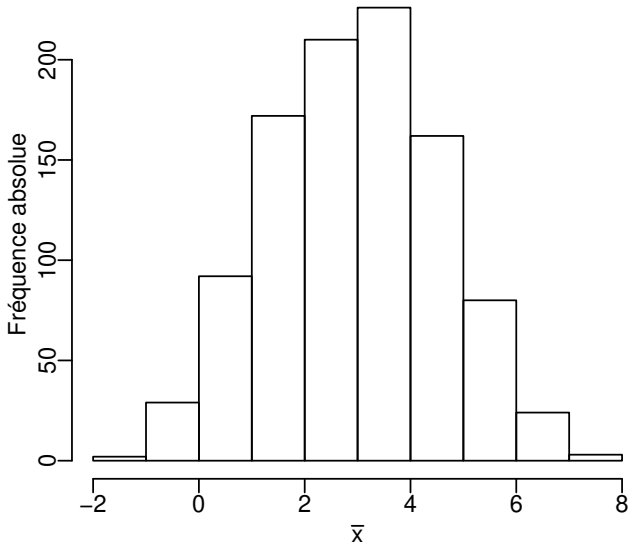
- ☐ $\widehat{\mu} = \overline{X}$ *est un estimateur de* μ *dont la réalisation est* $\overline{x}$ *;*
- ☐ $\widehat{\sigma}^2 = n^{-1} \sum_{i=1}^n (X_i - \overline{X})^2$ *est un estimateur de* σ^2 *dont la réalisation est* $n^{-1} \sum_{i=1}^n (x_i - \overline{x})^2$.

Remarques :

- ☐ Une statistique T étant fonction des variables aléatoires $X_1, \ldots, X_n$, c'est elle-même une variable aléatoire !
- ☐ La loi de T dépend de la loi des X_i et est appelée **distribution d'échantillonnage de** T.
- ☐ Si on ne peut pas déduire la loi exacte de T de celle des X_i, on doit parfois se contenter de la connaissance de $\mathrm{E}(T)$ et $\mathrm{Var}(T)$.
- ☐ $\mathrm{E}(T)$ et $\mathrm{Var}(T)$ fournissent une information partielle sur la loi de T et offrent parfois la possibilité (par exemple pour $T = \overline{X}$) d'utiliser une loi approximative de T (souvent grâce au théorème central limite).

Distribution d'échantillonnage : exemple

Soient $X_1, \ldots, X_{10} \overset{\text{iid}}{\sim} \mathcal{N}(3, 25)$ et $\overline{X} = \frac{1}{10}(X_1 + \ldots + X_{10})$. Histogramme de 1000 réalisations de $\overline{X}$:



3.1 Estimation de paramètres

slide 171

Questions d'intérêt et estimation

On suppose que l'on dispose d'un **modèle** (c'est-à-dire une famille de densités $f(x;\theta)$ indexée par θ). On souhaite, par exemple :

- **estimer** les paramètres de ce modèle ;
- répondre à des questions concernant la valeur de ces paramètres, par exemple **tester** si $\theta = 0$;
- **prédire** les valeurs des observations futures.

Il existe de nombreuses méthodes d'estimation des paramètres d'un modèle (le choix dépend de différents critères tels la précision, la robustesse et le temps de calcul). On va décrire les suivantes :

- **méthode des moments** (simple) ;
- **méthode des moindres carrés** (simple) ;
- **méthode du maximum de vraisemblance** (souvent utilisée car générale et optimale dans beaucoup de situations).

Méthode des moments

- Soient $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x;\theta)$.
- On considère le k-ème moment pour $k \geq 1$:
  - Moment "théorique" : $m_k = \mathrm{E}(X^k) = \int_{-\infty}^{\infty} x^k f(x;\theta)\mathrm{d}x$.
  - Moment "empirique" (calculé à partir de l'échantillon) : $\widehat{m}_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$.
- L'estimateur des moments de θ s'obtient en égalisant les moments "théoriques" et "empiriques" : $m_k = \widehat{m}_k$, pour k dans un ensemble de nombres entiers.
- On a besoin d'autant de moments (finis !) que de paramètres inconnus.

Exemple 40 *Soient $X_1, \ldots, X_n \overset{\text{iid}}{\sim} U(0,\theta)$. Trouver l'estimateur des moments de θ.*

Exemple 41 *Soient $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$. Quels sont les estimateurs des moments de μ et σ^2.*

Solution Exemple 40

On a

$$m_1 = \mathrm{E}(X) = \int_0^\theta \frac{x}{\theta} \mathrm{d}x = \theta/2,$$

On résout ensuite l'équation $\widehat{m}_1 = \overline{X} = \theta/2$, ce qui donne $\widehat{\theta} = 2\overline{X}$.

On peut se demander si, dans ce cas, il s'agit d'un bon estimateur. La réponse est non. Par exemple, si on observe les 5 valeurs

$$x_1 = 0, \quad x_2 = 0.5, \quad x_3 = 1.5, \quad x_4 = 2, \quad x_5 = 6,$$

alors $\overline{x} = 2$ et $\widehat{\theta} = 4$. Mais $x_5 = 6 > 4$, et donc l'échantillon ne peut pas provenir d'une loi uniforme sur $[0, 4]$ (on sait que $\theta \geq 6 = \max\{x_i\}$).

Solution Exemple 41

Moments théoriques :

$$m_1 = \mathrm{E}(X) = \mu \quad \text{et} \quad m_2 = \mathrm{E}(X^2) = \mathrm{Var}(X) + \mathrm{E}(X)^2 = \sigma^2 + \mu^2.$$

Moments empiriques :

$$\widehat{m}_1 = \tfrac{1}{n}\sum_{i=1}^n X_i = \overline{X} \quad \text{et} \quad \widehat{m}_2 = \tfrac{1}{n}\sum_{i=1}^n X_i^2.$$

Il faut donc résoudre $\begin{cases} \mu & = \overline{X} \\ \sigma^2 + \mu^2 & = \frac{1}{n}\sum_{i=1}^n X_i^2. \end{cases}$

D'où

$$\widehat{\mu} = \overline{X}, \quad \widehat{\sigma}^2 = \frac{1}{n}\left(\sum_{i=1}^n X_i^2 - n\overline{X}^2\right) = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2.$$

En effet

$$\sum_{i=1}^n (X_i - \overline{X})^2 = \sum_{i=1}^n (X_i^2 + \overline{X}^2 - 2X_i\overline{X}) = \left(\sum_{i=1}^n X_i^2\right) + n\overline{X}^2 - 2\overline{X}\sum_{i=1}^n X_i$$
$$= \left(\sum_{i=1}^n X_i^2\right) + n\overline{X}^2 - 2n\overline{X}^2 = \left(\sum_{i=1}^n X_i^2\right) - n\overline{X}^2.$$

Méthode des moindres carrés

☐ Soient $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f$, et supposons que le paramètre θ à estimer soit $\mathrm{E}(X_1)$. Alors :
- chaque X_i doit être "proche" de θ ;
- chaque différence $X_i - \theta$ doit être "assez petite".

☐ Donc une estimation raisonnable de θ est la valeur minimisant

$$S(\theta) = \sum_{i=1}^{n} (X_i - \theta)^2.$$

Exemple 42 *Soient $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f$ telles que $\mathrm{E}(X_i) = \theta$. Trouver l'estimateur des moindres carrés de θ.*

Solution Exemple 42

On a

$$S'(\theta) = \sum_{i=1}^{n} -2(X_i - \theta),$$

et donc

$$S'(\theta) = 0 \Leftrightarrow \sum_{i=1}^{n} (X_i - \theta) = 0 \Leftrightarrow \left(\sum_{i=1}^{n} X_i \right) - n\theta = 0 \Leftrightarrow \theta = \frac{1}{n} \sum_{i=1}^{n} X_i = \overline{X}.$$

De plus,

$$S''(\theta) = \left[-2 \sum_{i=1}^{n} (X_i) + 2n\theta \right]' = 2n > 0,$$

donc la valeur précédente correspond à un minimum. Finalement, $\widehat{\theta} = \overline{X}$.

Méthode du maximum de vraisemblance

Définition 20 *Soient $x_1, \ldots, x_n$ une réalisation de $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x;\theta)$. On appelle* **vraisemblance** *pour θ la fonction*

$$L(\theta) = f(X_1, \ldots, X_n; \theta) = f(X_1;\theta) \times f(X_2;\theta) \times \cdots \times f(X_n;\theta) = \prod_{i=1}^{n} f(X_i;\theta),$$

ou, plus souvent,

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

La vraisemblance est vue comme une fonction de θ.

Définition 21 *L'***estimateur du maximum de vraisemblance** $\widehat{\theta}_{\text{ML}}$ *d'un paramètre θ est celui qui maximise la fonction de vraisemblance parmi tous les θ possibles. Donc $\widehat{\theta}_{\text{ML}}$ satisfait*

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

Sa réalisation correspond à la valeur de θ qui maximise la probabilité d'observer les valeurs que l'on a effectivement observées.

Calcul de $\widehat{\theta}_{\text{ML}}$

On facilite les calculs en maximisant $\ell(\theta) = \log L(\theta)$ au lieu de $L(\theta)$. La démarche est la suivante :

1. calculer la vraisemblance $L(\theta)$;
2. en déduire la log-vraisemblance $\ell(\theta)$;
3. déterminer le $\widehat{\theta}_{\text{ML}}$ qui maximise $\ell(\theta)$. Il s'obtient souvent en résolvant $\mathrm{d}\ell(\theta)/\mathrm{d}\theta = 0$ puis en vérifiant qu'il s'agit bien d'un maximum, par exemple en montrant que $\mathrm{d}^2\ell(\theta)/\mathrm{d}\theta^2 < 0$.

Illustration : `https://rpsychologist.com/likelihood/`

Exemple 43 *Soient $x_1, \ldots, x_n$ une réalisation de $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \exp(\lambda)$. Trouver l'estimateur du maximum de vraisemblance de λ, $\widehat{\lambda}_{\text{ML}}$.*

Solution Exemple 43

La vraisemblance est

$$L(\lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda \sum_{i=1}^n x_i},$$

donc la log vraisemblance est

$$\ell(\lambda) = \log L(\lambda) = n \log \lambda - \lambda \sum_{i=1}^{n} x_i.$$

Ainsi

$$\ell'(\lambda) = 0 \Leftrightarrow \frac{n}{\lambda} - \sum_{i=1}^{n} x_i = 0 \Leftrightarrow \lambda = \frac{n}{\sum_{i=1}^n x_i} = \frac{1}{\overline{x}}.$$

De plus,

$$\ell''(\lambda) = -n/\lambda^2 < 0,$$

et donc la valeur ci-dessus correspond bien à un maximum. Finalement, $\widehat{\lambda}_{\mathrm{ML}} = 1/\overline{X}$.

Biais

Définition 22 *Le* **biais** *de l'estimateur $\widehat{\theta}$ de θ est défini par*

$$b(\widehat{\theta}) = \mathrm{E}(\widehat{\theta}) - \theta.$$

- ☐ Interprétation du biais :
  - si $b(\widehat{\theta}) < 0$, alors $\widehat{\theta}$ sous-estime θ en moyenne ;
  - si $b(\widehat{\theta}) > 0$, alors $\widehat{\theta}$ sur-estime θ en moyenne ;
  - si $b(\widehat{\theta}) = 0$, alors $\widehat{\theta}$ est dit **non-biaisé**.
- ☐ Le biais est indicateur de la qualité de $\widehat{\theta}$. Si $b(\widehat{\theta}) \approx 0$ alors $\widehat{\theta}$ fournit la vraie valeur du paramètre en moyenne.
- ☐ La variance de $\widehat{\theta}$ est aussi un indicateur important de la qualité de l'estimateur.

Exemple 44 *Soient $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$. Trouver le biais et la variance de $\widehat{\mu} = \overline{X}$ et le biais de $\widehat{\sigma}^2 = n^{-1} \sum_{i=1}^n (X_i - \overline{X})^2$.*

Solution Exemple 44

Pour $\widehat{\mu} = \overline{X}$ on a :

$$b(\widehat{\mu}) = \mathrm{E}(\widehat{\mu}) - \mu = \mathrm{E}(\overline{X}) - \mu = \mu - \mu = 0,$$
$$\mathrm{Var}(\widehat{\mu}) = \mathrm{Var}(\overline{X}) = \sigma^2/n.$$

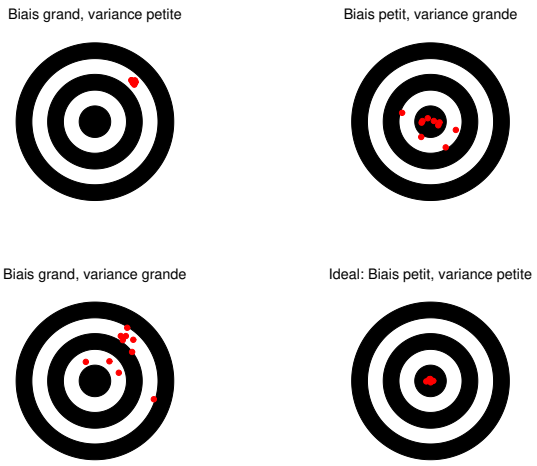
Pour $\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - \overline{X}^2$ on a

$$\mathrm{E}(\widehat{\sigma}^2) = \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^n X_i^2\right) - \mathrm{E}(\overline{X}^2) = \mathrm{E}(X_1^2) - \{\mathrm{Var}(\overline{X}) + \mathrm{E}(\overline{X})^2\}$$
$$= (\sigma^2 + \mu^2) - (\sigma^2/n + \mu^2) = \sigma^2(1 - 1/n) = \sigma^2\frac{n-1}{n}.$$

Ainsi le biais de $\widehat{\sigma}^2$ est $b(\widehat{\sigma}^2) = \sigma^2(1-1/n) - \sigma^2 = -\sigma^2/n$. Puisque $\mathrm{E}(\widehat{\sigma}^2) = \sigma^2 \times (n-1)/n$, on a $\mathrm{E}(\widehat{\sigma}^2) \times n/(n-1) = \sigma^2$ et on définit un estimateur non biaisé de σ^2 par

$$S^2 = \widehat{\sigma}^2 \times n/(n-1) = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2.$$

Biais et variance



- ☐ θ = centre de la cible, supposé être la vraie valeur.
- ☐ Réalisations de $\widehat{\theta}$ = fléchettes rouges, valeurs estimées à l'aide de différents échantillons.

Erreur quadratique moyenne

Définition 23 *L'***erreur quadratique moyenne** *de l'estimateur $\widehat{\theta}$ de θ est*

$$\mathrm{EQM}(\widehat{\theta}) = \mathrm{E}\{(\widehat{\theta} - \theta)^2\} = \cdots = \mathrm{Var}(\widehat{\theta}) + b(\widehat{\theta})^2.$$

Si $\widehat{\theta}$ est un estimateur sans biais du paramètre θ, alors $\mathrm{EQM}(\widehat{\theta}) = \mathrm{Var}(\widehat{\theta})$.

Définition 24 *Soient $\widehat{\theta}_1$ et $\widehat{\theta}_2$ deux estimateurs sans biais du même paramètre θ. On dit que $\widehat{\theta}_1$ est* **plus efficace** *que $\widehat{\theta}_2$ si*

$$\mathrm{Var}(\widehat{\theta}_1) \leq \mathrm{Var}(\widehat{\theta}_2).$$

On préfère alors $\widehat{\theta}_1$ à $\widehat{\theta}_2$.

Exemple 45 *Soient $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$. La médiane M suit une loi $\mathcal{N}(\mu, \sigma^2\pi/(2n))$ pour n grand. Lequel des estimateurs $\overline{X}$ et M de μ est préférable ? Et si des valeurs aberrantes peuvent apparaître ?*

Solution Exemple 45

On a

$$\mathrm{Var}(M) = \sigma^2\pi/(2n) > \sigma^2/n = \mathrm{Var}(\overline{X}).$$

Ainsi, étant donné que les deux estimateurs sont non biaisés, on préfère utiliser $\overline{X}$ pour estimer μ (il est plus précis au sens de l'EQM).

En revanche, en présence de valeurs aberrantes (ne provenant pas de la loi normale), la médiane est plus robuste et peut donc être préférable.

3.2 Intervalles de confiance

Intervalles de confiance : définition

Une manière de rapporter l'information qui permet de prendre en compte la variabilité de l'estimation est d'utiliser un **intervalle de confiance (IC)**.

- ☐ Puisqu'une erreur se produit vraisemblablement lors de l'estimation de la moyenne de notre population, il est très informatif de fournir une indication de l'importance de cette erreur.
- ☐ On pourrait ainsi spécifier une marge d'erreur, ce qui donne une estimation par intervalle du paramètre d'intérêt

Intervalles de confiance : définition

Soient $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x;\theta)$.

- ☐ Au lieu d'une estimation ponctuelle ($\widehat{\theta}$) du paramètre θ, on préfère un intervalle aléatoire contenant θ avec une grande probabilité.
- ☐ Soit $\alpha \in (0,1)$. Un **intervalle de confiance** (IC) à $100(1-\alpha)\%$ pour θ est un intervalle aléatoire $[I, S]$ tel que

  $$\Pr(I \leq \theta \leq S) = 1 - \alpha,$$

  et les bornes I et S sont des variables aléatoires qui ne dépendent pas de θ. Elles sont appelées borne inférieure et supérieure de l'intervalle de confiance, respectivement. Le **niveau de confiance** est $1-\alpha$.
- ☐ La quantité α est choisie de sorte à ce que $1-\alpha$ soit grand : des valeurs typiques pour α sont 0.1, 0.05 et 0.01, la plus courante étant 0.05.

Intervalles de confiance : méthode

- ☐ La première étape est de trouver un pivot, c'est-à-dire une fonction $T = p((X_1, \ldots, X_n), \theta)$ **dont la loi est connue et ne dépend pas de** θ.
- ☐ Il s'agit ensuite de choisir $\alpha \in (0,1)$ ainsi que $\alpha_I, \alpha_S \in (0,1)$ tels que $\alpha_I + \alpha_S = \alpha$ (on choisit souvent le cas symétrique où $\alpha_I = \alpha_S = \alpha/2$). Puisque la loi de T est connue et ne dépend pas de θ, on peut facilement trouver les quantiles $q_T(\alpha_I)$ et $q_T(1-\alpha_S)$. Par définition, ils vérifient

$$\alpha_I = \Pr(T < q_T(\alpha_I)) \quad \text{et} \quad 1 - \alpha_S = \Pr(T \leq q_T(1-\alpha_S)),$$

  et on a donc

$$\begin{aligned}\Pr(q_T(\alpha_I) \leq T \leq q_T(1-\alpha_S)) &= \Pr(T \leq q_T(1-\alpha_S)) - \Pr(T < q_T(\alpha_I)) \\ &= (1-\alpha_S) - \alpha_I = 1 - \alpha.\end{aligned}$$

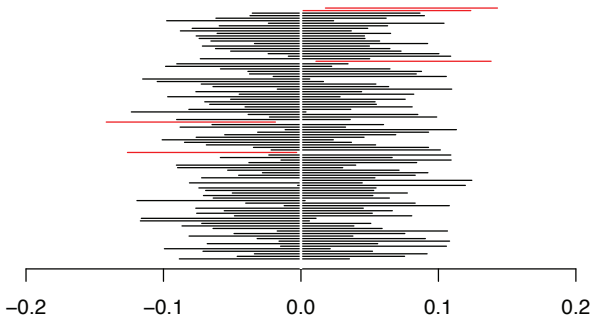
- ☐ La dernière étape consiste à isoler θ (si possible), ce qui permet de trouver des variables aléatoires I, S (fonctions de $X_1, \ldots, X_n$, $q_T(\alpha_I)$ et $q_T(1-\alpha_S)$ mais pas de θ) telles que

$$\Pr(I \leq \theta \leq S) = 1 - \alpha.$$

- ☐ On constate que $[I, S]$ est bien un IC à $100(1-\alpha)\%$ (ou encore au niveau de confiance $1-\alpha$) pour θ.

Interprétation

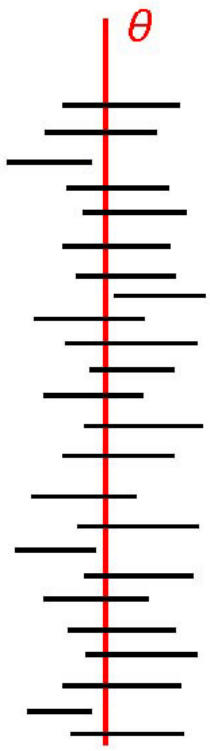
Si on recommence l'expérience dans les mêmes conditions un grand nombre de fois avec un échantillon de taille n à chaque fois et qu'on calcule l'IC au niveau 95% pour chacun d'eux, **une proportion de 95% de ces intervalles va contenir la vraie valeur de** μ.



Intervalle de confiance pour la moyenne d'échantillons aléatoires de taille $n = 1000$ d'une loi $\mathcal{N}(0,1)$ (100 réplications) ; les lignes rouges indiquent les intervalles qui ne couvrent pas la vraie valeur, ici zéro.

Interprétation

- ☐ $[I, S]$ est un intervalle aléatoire qui contient le vrai paramètre θ avec une probabilité ("confiance") $1 - \alpha$.
- ☐ La probabilité que la variable aléatoire I soit inférieure à θ **et** que la variable aléatoire S soit supérieure à θ est égale à $1 - \alpha$.
- ☐ Il est (en théorie) incorrect de dire que la probabilité que $\theta \in [I, S]$ est égale à $1 - \alpha$. En effet, ce sont les quantités I et S qui sont aléatoires et non θ.
- ☐ Attention à la différence entre l'intervalle de confiance (aléatoire) et sa réalisation ! Souvent, le terme "intervalle de confiance" est utilisé dans les deux cas.



IC pour l'espérance d'une loi normale de variance connue

Soient $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, avec σ^2 **connu** et soit $\alpha \in (0, 1)$. On se place dans le cas $\alpha_I = \alpha_S = \alpha/2$. On a (admis)

$$T = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1).$$

On prend T comme pivot. Soit z_α le quantile au niveau α de la loi $\mathcal{N}(0, 1)$. On sait que

$$\Pr(z_{\alpha/2} \leq T \leq z_{1-\alpha/2}) = 1 - \alpha.$$

Par symétrie de la loi normale, $z_{\alpha/2} = -z_{1-\alpha/2}$. Ainsi,

$$\Pr\left(-z_{1-\alpha/2} \leq \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \leq z_{1-\alpha/2}\right) = 1 - \alpha,$$

i.e.,

$$\Pr\left(-\overline{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \leq -\mu \leq -\overline{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha.$$

IC pour l'espérance d'une loi normale de variance connue

On obtient donc

$$\Pr\left(\overline{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha.$$

On en déduit qu'un IC pour μ au niveau $1-\alpha$ est

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

Il s'agit d'un IC bilatéral.

Exemple 46 *On suppose que la résistance X d'un certain type d'équipement électronique suit une loi normale telle que $\sigma = 0.12$ ohm. On a obtenu sur un échantillon de taille $n = 64$ la moyenne empirique $\overline{x} = 5.34$ ohm. Trouver un IC pour μ au niveau 95%.*

Solution Exemple 46

On veut que $100(1-\alpha)\% = 95\%$, i.e., $1-\alpha = 0.95$ et donc $\alpha = 0.05$. Ainsi, $z_{1-\alpha/2} = z_{0.975} = 1.96$ et la réalisation sur ces données de l'IC pour μ obtenu précédemment est

$$\left[5.34 - 1.96 \times \tfrac{0.12}{8}, 5.34 + 1.96 \times \tfrac{0.12}{8}\right] = [5.31, 5.37].$$

Loi de Student

Définition 25 *Soient ν un entier positif et $X_1, \ldots, X_\nu \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$. La variable aléatoire*

$$U = \sum_{i=1}^{\nu} X_i^2$$

suit **la loi du khi-deux à ν degrés de liberté**. *On note $U \sim \chi^2_\nu$.*

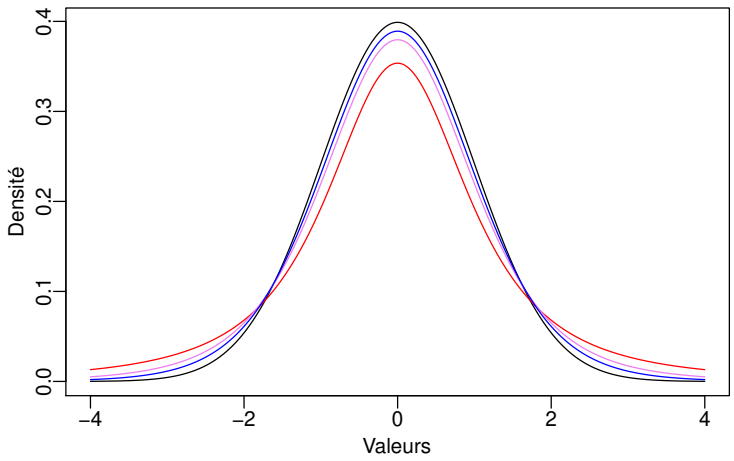
Définition 26 *Soit $Z \sim \mathcal{N}(0,1)$ et $U \sim \chi^2_\nu$ indépendante de Z. La variable aléatoire*

$$T = \frac{Z}{\sqrt{U/\nu}}$$

suit **la loi de Student t à ν degrés de liberté**. *On note $T \sim t_\nu$.*

Remarque : Les queues de la loi de Student sont plus lourdes que celles de la loi normale centrée réduite. Ainsi, une variable de Student a plus de chance de prendre des valeurs extrêmes qu'une variable normale.

Représentation de la loi de Student



Densité de la loi $\mathcal{N}(0,1)$ (en noir) et densités des lois t_ν pour $\nu = 2$ (rouge), $\nu = 5$ (violet) et $\nu = 10$ (bleu).

IC pour l'espérance d'une loi normale de variance inconnue

Soient $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$ avec σ^2 **inconnu**, et soit

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}.$$

Soit $\alpha \in (0,1)$. On se place dans le cas $\alpha_I = \alpha_S = \alpha/2$. On a (admis)

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}.$$

On prend T comme pivot. Soit $t_{n-1,\alpha}$ le quantile au niveau α de la loi t_{n-1}. On sait que

$$\Pr(t_{n-1,\alpha/2} \leq T \leq t_{n-1,1-\alpha/2}) = 1 - \alpha.$$

Par symétrie de la loi de Student, $t_{n-1,\alpha/2} = -t_{n-1,1-\alpha/2}$. Ainsi,

$$\Pr\left(-t_{n-1,1-\alpha/2} \leq \frac{\overline{X} - \mu}{S/\sqrt{n}} \leq t_{n-1,1-\alpha/2}\right) = 1 - \alpha.$$

IC pour l'espérance d'une loi normale de variance inconnue

On obtient donc

$$\Pr\left(-\overline{X} - t_{n-1,1-\alpha/2}\frac{S}{\sqrt{n}} \leq -\mu \leq -\overline{X} + t_{n-1,1-\alpha/2}\frac{S}{\sqrt{n}}\right) = 1 - \alpha,$$

i.e.,

$$\Pr\left(\overline{X} - t_{n-1,1-\alpha/2}\frac{S}{\sqrt{n}} \leq \mu \leq \overline{X} + t_{n-1,1-\alpha/2}\frac{S}{\sqrt{n}}\right) = 1 - \alpha.$$

On en déduit qu'un IC pour μ au niveau $1-\alpha$ est

$$\left[\overline{X} - t_{n-1,1-\alpha/2}\frac{S}{\sqrt{n}}, \overline{X} + t_{n-1,1-\alpha/2}\frac{S}{\sqrt{n}}\right].$$

Cet IC est appelé intervalle de Student.

Exemple 47 *On suppose que le point de fusion d'un certain alliage suit une loi normale d'espérance μ et variance σ^2 inconnues. On a obtenu $n = 9$ observations qui ont donné une moyenne $\overline{x} = 1040°C$ et un écart-type $s = 16°C$. Construire un IC pour μ à 95%.*

Solution Exemple 47

On choisit $\alpha = 0.05$, ce qui nous donne à l'aide des tables $t_{n-1,1-\alpha/2} = t_{8,0.975} = 2.306$. Ainsi la réalisation sur ces données de l'IC pour μ obtenu précédemment est

$$\left[1040 - 2.306 \times \tfrac{16}{3}, 1040 + 2.306 \times \tfrac{16}{3}\right] = [1027.8, 1052.2]\,.$$

Remarques

- ☐ Il est souvent possible d'obtenir des ICs approchés grâce au théorème central limite. Cependant, dans certains cas (notamment la loi normale), on peut obtenir des ICs exacts.
- ☐ Un IC n'indique pas seulement où un paramètre inconnu est situé. Sa largeur donne une idée de la précision de l'estimation ponctuelle.
- ☐ Si on diminue α, i.e., si on augmente $1 - \alpha$ (c'est-à-dire que l'on augmente la probabilité que l'IC contienne le paramètre θ), l'IC devient plus large.
- ☐ Les ICs bilatéraux symétriques pour μ sont tous de la forme

  $$\left[\overline{X} - \frac{c}{\sqrt{n}}, \overline{X} + \frac{c}{\sqrt{n}}\right].$$

  Ainsi, augmenter n permet d'avoir un IC plus étroit.
- ☐ On peut définir des IC unilatéraux. Par exemple, soient $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, avec σ^2 connu. Les ICs pour μ de la forme $(-\infty, \overline{X} + z_{1-\alpha}\sigma/\sqrt{n}]$ et $[\overline{X} - z_{1-\alpha}\sigma/\sqrt{n}, \infty)$ sont des ICs **unilatéraux à gauche et à droite**, respectivement, qui contiennent μ avec une probabilité $1 - \alpha$.

Estimateur du maximum de vraisemblance et IC

Théorème 4 *Soit $\widehat{\theta}_{\mathrm{ML}}$ l'estimateur du maximum de vraisemblance du paramètre θ pour un modèle "régulier". Alors*

$$\widehat{\theta}_{\mathrm{ML}} \dot{\sim} \mathcal{N}\left\{\theta, J(\widehat{\theta}_{\mathrm{ML}})^{-1}\right\} \quad \textit{pour } n \textit{ grand},$$

où $J(\theta) = -\mathrm{d}^2\ell(\theta)/\mathrm{d}\theta^2$ est appelé **l'information observée** *pour θ. Donc l'IC bilatéral symétrique pour θ au niveau $1-\alpha$ a pour bornes $\widehat{\theta}_{\mathrm{ML}} \pm z_{1-\alpha/2} J(\widehat{\theta}_{\mathrm{ML}})^{-1/2}$.*

La plupart des modèles rencontrés dans la pratique sont réguliers.

Un résultat similaire est valable quand θ est un vecteur : dans ce cas $J(\widehat{\theta}_{\mathrm{ML}})$ est la matrice Hessienne de $-\ell(\theta)$ évaluée en $\theta = \widehat{\theta}_{\mathrm{ML}}$.

Exemple 48 *Soient $X_1, \ldots, X_n \overset{\mathrm{iid}}{\sim} \exp(\lambda)$. Trouver un intervalle de confiance à $100(1-\alpha)\%$ pour λ. Sachant que l'on a les données $n = 25$ et $\overline{x} = 40$, trouver un IC à 95% pour λ.*

Solution Exemple 48

On utilise les résultats de l'exemple 43 :

$$\widehat{\lambda}_{\mathrm{ML}} = 1/\overline{x} \quad \text{et} \quad \ell''(\lambda) = -n/\lambda^2.$$

Ainsi $J(\widehat{\lambda}_{\mathrm{ML}}) = -\ell''(\widehat{\lambda}_{\mathrm{ML}}) = n\overline{x}^2$, et

$$\widehat{\lambda}_{\mathrm{ML}} \dot{\sim} \mathcal{N}\{\lambda, (n\overline{x}^2)^{-1}\}.$$

Un IC au niveau $1-\alpha$ pour λ a donc pour limites $\widehat{\lambda}_{\mathrm{ML}} \pm z_{1-\alpha/2}(\sqrt{n}\overline{x})^{-1}$. La réalisation de cet IC à 95% sur ces données est $1/40 \pm 1.96(5 \times 40)^{-1}$, i.e., environ $[0.0152, 0.0348]$.

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 (directement observées ou résultant d'une expérience).
- ☐ 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 (directement observées ou résultant d'une expérience).
- ☐ **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 49 *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*

Question : L'alcool ralentit-il les réflexes ?

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 :

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

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 $T = g(X_1, ..., X_n)$ prenant typiquement des valeurs "petites" sous l'hypothèse nulle H_0 (i.e., si H_0 est vraie) et "grandes" ("grandes" dans la direction de l'hypothèse alternative H_1) sous H_1, ou en tous cas plus petites sous H_0 que 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.

Cadre statistique : [2] Statistique de test

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 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 $T = \dfrac{\overline{X} - \mu_0}{S/\sqrt{n}}$.

- ☐ Si H_0 est vraie, alors $T \sim t_{n-1}$ (donc si H_0 est vraie, T prend typiquement des valeurs "petites" au sens 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.

Cadre statistique : [2] Statistique de test

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 $T = \dfrac{\overline{X} - \mu_0}{S/\sqrt{n}}$.

- ☐ 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 / Verité	H_0	H_1
Non-rejet de H_0	☺ (Vrai négatif)	**Erreur de Type II** (Faux négatif)
Rejet de H_0	**Erreur de Type I** (Faux positif)	☺ (Vrai positif)

Cadre statistique : [3] Significativité statistique

- ☐ Les valeurs critiques dépendent de l'erreur que l'on considère comme la plus grave. Si l'on souhaite une probabilité d'erreur de type I faible (on rejette seulement pour des valeurs très extrêmes de la statistique de test), celle d'erreur de type II est élevée. Si l'on souhaite une probabilité d'erreur de type II moins élevée (on rejette pour des valeurs moins élevées), il faut accepter une probabilité d'erreur de type I moins faible. Il y a un compromis à effectuer.
- ☐ En général, il existe une asymétrie naturelle entre les deux hypothèses : l'erreur de type I est considérée comme étant la plus grave (exemple des filtres de spams). Ainsi, on fixe un seuil que l'on ne souhaite pas dépasser (tout en ayant conscience que plus ce seuil est faible, plus la probabilité d'erreur de type II est élevée) pour la probabilité d'erreur de type I et les valeurs critiques en découlent.
- ☐ De toute façon, la loi de T étant souvent inconnue sous H_1, il serait difficile de déduire des valeurs critiques d'une borne supérieure sur la probabilité d'erreur de type II.

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 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

$$\Pr[\text{Rejet de } H_0 | H_0 \text{ est vraie}] = \alpha.$$

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

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

$$\Pr[T < \text{ valeur critique} | H_0 \text{ est vraie}] = \alpha \quad \text{(cas unilatéral à gauche)},$$

$$\Pr[T > \text{ valeur critique} | H_0 \text{ est vraie}] = \alpha \quad \text{(cas unilatéral à droite)}.$$

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

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

$$\Pr[|T| > v^* | H_0 \text{ est vraie}] = \alpha,$$

i.e.,

$$\Pr[T < -v^* \text{ ou } T > v^* | H_0 \text{ est vraie}] = \alpha.$$

ce qui implique

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

où $t_{n-1,1-\alpha/2}$ est le quantile au niveau $100(1 - \alpha/2)\%$ 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 : $p_{\text{obs}} = \Pr[|T| \geq t_{\text{obs}}|H_0]$,
  - Cas unilatéral à gauche : $p_{\text{obs}} = \Pr[T \leq t_{\text{obs}}|H_0]$,
  - Cas unilatéral à droite : $p_{\text{obs}} = \Pr[T \geq t_{\text{obs}}|H_0]$.
- Des valeurs p_{obs} "assez petites" s'opposent à H_0 car elles démontrent que la realité observée serait très improbable si l'hypothèse nulle H_0 était vraie.
- Quelles valeurs de p_{obs} peuvent être considérées comme "assez petites" pour justifier le rejet de H_0 ?

Cadre statistique : [4] La p-valeur

Comment définir la notion d' "assez petite" ? Souvent, nous suivons la même approche que celle décrite précédemment, i.e., nous fixons le **niveau de significativité** α.

- Nous choisissons la valeur maximale que l'on tolère pour la probabilité d'erreur de type I, α. On veut donc

$$\Pr[\text{Rejet de } H_0|H_0 \text{ est vraie}] = \alpha.$$

  Typiquement, $\alpha = 0.1, 0.05, 0.01$; le plus souvent, $\alpha = 0.05$.
- Notre règle de décision sera : **rejeter** H_0 **si** $p_{\text{obs}} < \alpha$.
- La probabilité d'erreur de type I en utilisant cette règle de décision est exactement α.
- Cette approche 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 est 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 $p_{\text{obs}} < \alpha$.

Choix de la statistique de test T

- ☐ On est libre de choisir T comme on le souhaite dès l'instant que plus sa valeur est grande, plus l'indication contre H_0 est forte.
- ☐ Le choix de T dépend de l'**hypothèse alternative** H_1 — ce que l'on imagine possible si H_0 est fausse. Plus H_1 est précise, plus on peut choisir une statistique T appropriée.
- ☐ On souhaite, pour un α donné, utiliser la statistique qui minimise la probabilité d'erreur de type II (ou maximise la puissance du test, cf ci-après).

Détermination de H_0 parmi deux hypothèses

Supposons que l'on veuille choisir entre deux hypothèses A et B (par exemple $A : \theta = \theta_0$ et $B : \theta \neq \theta_0$).
Comment choisir si l'on prend A ou B comme hypothèse nulle H_0, i.e., si l'on teste "$H_0 : A$ contre $H_1 : B$" ou "$H_0 : B$ contre $H_1 : A$" ?
Il y a deux critères de choix principaux :

- ☐ Souvent, la loi de statistique de test n'est pas connue sous l'une des deux hypothèses (exemple de $\theta \neq \theta_0$). On prend alors pour H_0 l'hypothèse sous laquelle la loi de la statistique de test est connue.
- ☐ Si l'on a de bonnes raisons de croire que l'une des deux hypothèses est clairement vraie, on choisit si possible cette hypothèse pour H_1. En effet, rejeter H_0 en faveur de H_1 est un résultat plus fort (concernant H_1) que de ne pas rejeter H_0 (concernant H_0).

Exemple 50 *On a contrôlé 10 compteurs d'électricité nouvellement fabriqués et obtenu les valeurs suivantes (en MW) :*
$983 \quad 1002 \quad 998 \quad 996 \quad 1002 \quad 983 \quad 994 \quad 991 \quad 1005 \quad 986.$
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* $\overline{x} = 994 < 1000$. *S'agit-il d'un hasard ou une faute de production ?*

Solution Exemple 50

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

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

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

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

donc $t_{\text{obs}} = -2.35$.
On rejette H_0 si et seulement si $t_{\text{obs}} < -t_{n-1,1-\alpha/2}$ ou $t_{\text{obs}} > t_{n-1,1-\alpha/2}$. Si l'on choisit $\alpha = 5\%$, $t_{n-1,1-\alpha/2} = 2.262$ (voir les tables), et comme $t_{\text{obs}} = -2.35 < -2.262$, on rejette l'hypothèse H_0.

Tests et ICs

De nombreux tests statistiques concernent la valeur d'un paramètre θ (d'une densité par exemple). Il y a un lien entre de tels tests et les intervalles de confiance pour θ. En particulier, les tests statistiques peuvent être basés sur les intervalles de confiance.

Supposons que l'on veuille tester l'hypothèse $H_0 : \theta = \theta_0$. Soit T un pivot défini par

$$T = \frac{\widehat{\theta} - \theta_0}{\mathrm{sd}(\widehat{\theta})},$$

où $\mathrm{sd}(\widehat{\theta})$ est la déviation standard de $\widehat{\theta}$. Sa réalisation est $t_{\mathrm{obs}} = \dfrac{\widehat{\theta}_{\mathrm{obs}} - \theta_0}{\mathrm{sd}(\widehat{\theta})}$.

Alors les procédures de test suivantes sont équivalentes :

- ☐ Si θ_0 n'appartient pas à la réalisation d'un IC pour θ au niveau de confiance $1 - \alpha$, on rejette H_0 au niveau α ; si la réalisation de l'IC contient θ_0, on ne rejette pas H_0.
- ☐ La stratégie de test traditionnelle décrite dans les slides précédents en utilisant comme statistique de test le pivot T défini ci-dessus.

Tests et ICs

Plus précisément, si $[I, S]$ désigne l'intervalle de confiance bilatéral symétrique au niveau de confiance $1 - \alpha$, i.e., $[I, S] = [\widehat{\theta} - q_T(1 - \alpha/2)\mathrm{sd}(\widehat{\theta}), \widehat{\theta} - q_T(\alpha/2)\mathrm{sd}(\widehat{\theta})]$:

- ☐ Dans le cas d'un test bilatéral ($H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$) au niveau de significativité α, l'approche de test traditionnelle est équivalente à rejeter H_0 en faveur de H_1 si et seulement si

$$\theta_0 \notin (I, S).$$

- ☐ Dans le cas d'un test unilatéral à gauche ($H_0 : \theta = \theta_0$ vs $H_1 : \theta < \theta_0$) au niveau de significativité $\alpha/2$, l'approche de test traditionnelle est équivalente à rejeter H_0 si et seulement si

$$\theta_0 \notin (-\infty, S).$$

- ☐ Dans le cas d'un test unilatéral à droite ($H_0 : \theta = \theta_0$ vs $H_1 : \theta > \theta_0$) au niveau de significativité $\alpha/2$, l'approche de test traditionnelle est équivalente à rejeter H_0 si et seulement si

$$\theta_0 \notin (I, \infty).$$

3.4 Tests du khi-deux

slide 222

Test d'adéquation du khi-deux

- **Test d'adéquation** d'une distribution théorique (spécifiée) à des données.
- Soit H_0 : "les observations proviennent de la loi théorique spécifiée".
- Supposons que l'on observe n valeurs tombant dans k classes disjointes. Soient $o_1, \ldots, o_k$ (réalisations de variables aléatoires notées $O_1, \ldots, O_k$) les **fréquences observées** dans chacune des classes et soient $E_1, \ldots, E_k$ les **fréquences théoriques** correspondantes sous H_0.
- Une mesure de l'écart entre la distribution théorique et les données (distribution empirique) est fournie par la **statistique du khi-deux** (ou statistique de Pearson)

$$T = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}.$$

Notons que $\sum_{i=1}^{k} O_i = \sum_{i=1}^{k} E_i = n$.
Sous H_0, T suit approximativement (pour n grand) une distribution χ^2_r, où

- $r = k - 1$ si les E_i peuvent être calculés sans avoir à estimer de paramètres inconnus ;
- $r = k - 1 - c$ si les E_i sont calculés après avoir estimé c paramètres.

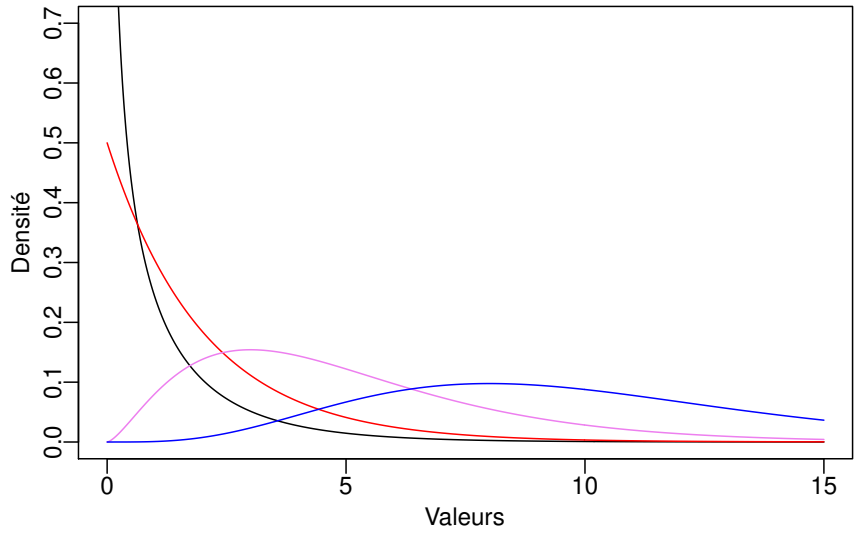
Remarques

- Pour assurer la convergence de T vers la loi du khi-deux, regrouper si besoin les données de façon à ce que $E_i > 5$ pour $i = 1, \ldots, k$.
- Pas d'hypothèse alternative spécifique : le choix se fait entre "rejet de H_0" ou "non-rejet de H_0".
- On rejette H_0 si la valeur observée

$$t_{\text{obs}} = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i} = \cdots = \sum_{i=1}^{k} \frac{o_i^2}{e_i} - n$$

est suffisamment élevée, i.e., au-dessus d'une valeur critique. Plus précisément, pour un test au niveau de significativité α, on rejette H_0 si $t_{\text{obs}} > \chi^2_{r,1-\alpha}$ (quantile au niveau $1 - \alpha$ de la loi du khi-deux à r degrés de liberté) ; sinon on ne la rejette pas.

Représentation de la loi du khi-deux



Densité de la loi χ^2_r pour $r = 1, 2, 5, 10$ (noir, rouge, violet, bleu).

Exemples

Exemple 51 *(Equilibre d'un dé)* 60 *lancers 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*

Tester l'hypothèse H_0 *"le dé est équilibré" au niveau de significativité* $\alpha = 5\%$.

Exemple 52 1000 *personnes ont passé un test de quotient intellectuel (QI) et les résultats suivants ont été obtenus :*

QI (X)	*[0, 70[*	*[70, 85[*	*[85, 100[*	*[100, 115[*	*[115, 130[*	*[130,* ∞*[*
Nombre o_i	*34*	*114*	*360*	*344*	*120*	*28*

Tester l'hypothèse H_0 *"*$X \sim \mathcal{N}(100, 15^2)$*" au niveau de significativité* $\alpha = 5\%$.

Solution Exemple 51

L'hypothèse H_0 est équivalente à $\Pr(X = x_i) = 1/6, i = 1, \ldots, 6$. Ainsi,

Valeur x_i	1	2	3	4	5	6	
$f_X(x_i) = \Pr(X = x_i)$	1/6	1/6	1/6	1/6	1/6	1/6	
$e_i = n \times \Pr(X = x_i)$	10	10	10	10	10	10	60

où X est le numéro obtenu. Donc

$$t_{\text{obs}} = \sum_{i=1}^{6} \frac{(o_i - e_i)^2}{e_i} = 8.5$$

et $T \overset{H_0}{\dot{\sim}} \chi^2_r$ avec $r = k - 1 = 6 - 1 = 5$ où $k = 6$ classes (faces). On a $\chi^2_{5,0.95} = 11.1 > 8.5 = t_{\text{obs}}$ donc on ne rejette pas H_0.

Solution Exemple 52

Sous H_0 les répartitions théoriques sont

$$e_i \mid 22.75 \mid 135.91 \mid 341.34 \mid 341.34 \mid 135.91 \mid 22.75$$

Ainsi,

$$\begin{aligned}
e_1 &= n \times \Pr(0 \leq X \leq 70) \\
&= n \times \Pr\left(-\tfrac{100}{15} \leq \tfrac{X-100}{15} \leq -\tfrac{30}{15}\right) \\
&= n \times \left\{\Phi(-2) - \Phi\left(-\tfrac{20}{3}\right)\right\} \\
&= n \times \left\{(1 - \Phi(2)) - \left(1 - \Phi\left(\tfrac{20}{3}\right)\right)\right\} = n \times \left\{\Phi\left(\tfrac{20}{3}\right) - \Phi(2)\right\} \\
&\approx n \times (1 - 0.97725) = n \times 0.02275 = 1000 \times 0.02275 = 22.75.
\end{aligned}$$

On obtient

$$t_{\text{obs}} = \sum_{i=1}^{6} \frac{(o_i - e_i)^2}{e_i} = 13.21,$$

et on a $T \overset{H_0}{\dot{\sim}} \chi^2_r$, avec $r = 6 - 1 = 5$. Puisque $\chi^2_{5,0.95} = 11.1 < 13.21 = t_{\text{obs}}$ on rejette H_0.

Tableaux de contingence

On considère n individus (ou objets) et on s'intéresse à **l'indépendance** de deux caractéristiques relatives à ces individus.

- ☐ Supposons que l'on observe pour chaque individu deux caractéristiques : A (pouvant appartenir à h classes) et B (pouvant appartenir à k classes).
- ☐ Soit n_{ij} le nombre de personnes se trouvant dans la classe i de la caractéristique A et dans la classe j de la caractéristique B, et soient

$$n_{i\cdot} = \sum_{j=1}^{k} n_{ij}, \quad n_{\cdot j} = \sum_{i=1}^{h} n_{ij}, \quad \text{et} \quad n_{\cdot\cdot} = \sum_{j=1}^{k}\sum_{i=1}^{h} n_{ij} = n.$$

- ☐ Le tableau de contingence est :

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

Indépendance

- ☐ On souhaite tester si les deux caractéristiques A et B sont indépendantes. Ainsi, on considère H_0 : "A et B **sont indépendantes**".
- ☐ On va utiliser un test du khi-deux afin de comparer les observations du tableau de contingence avec les valeurs théoriques sous l'hypothèse H_0 d'indépendance.
- ☐ On doit donc construire le tableau des fréquences théoriques (ou plutôt de leurs valeurs estimées) sous H_0, i.e.,

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

Estimation des fréquences théoriques sous H_0

☐ Sous H_0 (indépendance entre A et B) on a, pour $i = 1, \ldots, h$ et $j = 1, \ldots, k$,

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

☐ Les lois marginales de A et de B sont inconnues et il faut donc les estimer. On a, pour $i = 1, \ldots, h$,

$$\widehat{\Pr}(A = i) = \frac{\text{Nombre de cas favorables}}{\text{Nombre total de cas possibles}} = \frac{\sum_{j=1}^{k} n_{ij}}{\sum_{i=1}^{h}\sum_{j=1}^{k} n_{ij}} = \frac{n_{i\cdot}}{n_{\cdot\cdot}} = \frac{n_{i\cdot}}{n},$$

et, de même, pour $j = 1, \ldots, k$,

$$\widehat{\Pr}(B = j) = n_{\cdot j}/n.$$

☐ On en déduit

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

Test d'indépendance

☐ On utilise un test du khi-deux dont la valeur observée de la statistique de test T s'écrit

$$t_{\text{obs}} = \sum_{i=1}^{h}\sum_{j=1}^{k} \frac{(n_{ij} - e_{ij})^2}{e_{ij}} = \sum_{i=1}^{h}\sum_{j=1}^{k} \frac{(n_{ij} - n_{i\cdot}n_{\cdot j}/n)^2}{n_{i\cdot}n_{\cdot j}/n}.$$

☐ Sous H_0 et pour n grand, la statistique T suit une distribution χ^2_r où

$$r = hk - 1 - c,$$

où c est le nombre de paramètres estimés pour calculer les e_{ij}.

☐ Les lois marginales de A et B ont été estimées à l'aide de $h - 1$ et $k - 1$ paramètres (proportions), respectivement. Au total on a donc estimé $c = (k - 1) + (h - 1)$ paramètres, ce qui donne $r = (h - 1)(k - 1)$.

☐ Pour un test au niveau de significativité α, on rejette H_0 si et seulement si $t_{\text{obs}} > \chi^2_{(h-1)(k-1),1-\alpha}$.

Exemple

Exemple 53 *On a relevé chez 95 personnes la couleur des yeux (caractéristique A) ainsi que celle des cheveux (caractéristique B) et on a obtenu les résultats suivants :*

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

Tester au niveau de significativité $\alpha = 0.05$ si la couleur des cheveux est indépendante de celle des yeux.

Solution Exemple 53

On a

$$t_{\text{obs}} = \frac{\left(32 - \frac{44 \times 52}{95}\right)^2}{\frac{44 \times 52}{95}} + \cdots + \frac{\left(9 - \frac{43 \times 15}{95}\right)^2}{\frac{43 \times 15}{95}}$$
$$= 2.59 + 3.14 + 1.65 + 1.99 + 0.59 + 0.71 = 10.67.$$

De plus, $T \dot{\sim} \chi^2_\nu$, où $\nu = (3-1)(2-1) = 2$, et $\chi^2_{2,0.95} = 5.99$. Comme $5.99 < 10.67 = t_{\text{obs}}$, on rejette donc H_0, i.e., l'indépendance.

3.5 Comparaison de tests

slide 235

Tests paramétriques et non-paramétriques

Il existe une grande variété de tests différents pour des hypothèses plus ou moins complexes. Deux types importants de tests sont :

- ☐ les tests **paramétriques**, fondés sur un modèle statistique paramétrique (i.e., entièrement déterminé par un nombre fini de paramètres)—par exemple, $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$ et $H_0 : \mu = 0$;
- ☐ les tests **non-paramétriques**, fondés sur un modèle statistique plus général—par exemple, $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f$ et $H_0 : \Pr(X > 0) = \Pr(X < 0) = 1/2$, i.e., la médiane associée à f vaut 0.

L'avantage principal des tests paramétriques réside dans la possibilité de trouver un test (presque) optimal si les suppositions sous-jacentes sont correctes. En revanche, un tel test peut être mauvais en présence d'outliers (par exemple de valeurs aberrantes).
Les tests non-paramétriques sont souvent plus robustes mais en général moins **puissants** que les tests paramétriques si ces derniers sont utilisés de manière appropriée.

Puissance

Les deux types d'erreur possible lors d'un test statistique sont rappelées dans le tableau ci-dessous :

Décision / Verité	H_0	H_1
Non-rejet de H_0	🙂 (Vrai négatif)	**Erreur de Type II** (Faux négatif)
Rejet de H_0	**Erreur de Type I** (Faux positif)	🙂 (Vrai positif)

La région de rejet est déterminée de sorte à ce que $\Pr(\text{ Erreur de Type I }) = \alpha$, où α est le niveau de significativité choisi par la personne effectuant le test. Ainsi, la probabilité d'erreur de type I est contôlée mais pas celle d'erreur de type II. Cette dernière (probabilité de ne pas rejeter une fausse hypothèse H_0) dépend de H_1.

Définition 27 *La* **puissance** *d'un test est*

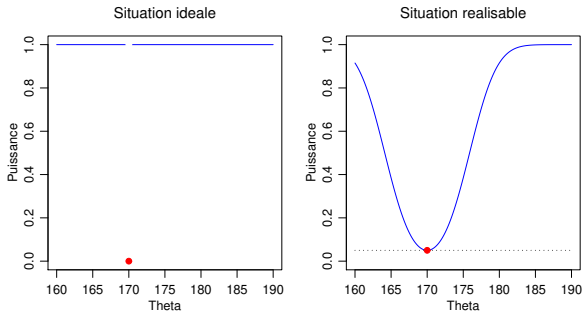
$$\beta(H_1) = \Pr_{H_1}(\textit{Rejet de } H_0) = 1 - \Pr(\textit{Erreur de Type II}) = 1 - \Pr_{H_1}(\textit{Non-rejet de } H_0),$$

où $\Pr_{H_1}$ *désigne la probabilité sous* H_1*. Ainsi, dans le cas où* $H_0 : \theta = \theta_0$ *et* H_1 *dépend de* θ*, la puissance peut s'écrire* $\beta(\theta)$*.*

Puissance

- ☐ A α fixé, on souhaite la plus grande puissance ($\beta(\theta)$) possible.
- ☐ Généralement, $\beta(\theta)$ est difficile à calculer.
- ☐ Plus la réalité sous H_1 est éloignée de H_0, plus la puissance est grande car les écarts importants ont plus de chance d'être détectés.
- ☐ La puissance augmente avec la taille de l'échantillon, n.

Illustration dans le cas d'un test $H_0 : \theta = 170$ contre $H_1 : \theta \neq 170$. Gauche : cas idéal (en général irréalisable). Droite : un cas plus réaliste ($\alpha = 0.05$).



4. Régression linéaire slide 239

4.1 Introduction slide 240

Régression en général

La **régression** concerne la relation entre une variable d'intérêt que l'on cherche à expliquer et une ou plusieurs autres variables dont on se sert pour expliquer la variable d'intérêt.

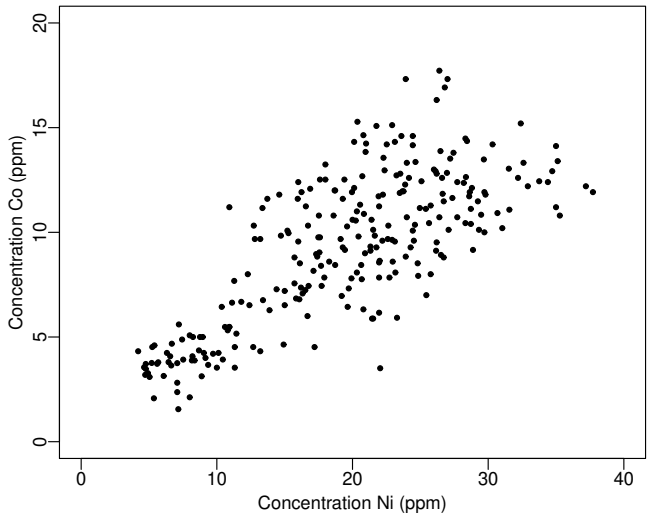
Variables et notations :

- y : la variable d'interêt, appelée **réponse** (ou encore variable expliquée ou variable dépendante) ;
- $x^{(1)}, \ldots, x^{(d)}$: les autres variables, appelées **covariables** (ou encore variables explicatives, variables indépendantes ou prédicteurs), considérées comme fixes (i.e., non-aléatoires).

Estimation et prédiction :

- Il faut **estimer** une relation éventuelle entre y et les $x^{(j)}$, $j = 1, \ldots, d$, appelée fonction de régression ;
- L'un des buts principaux de la régression est la **prédiction** des valeurs futures de y connaissant les valeurs des $x^{(j)}$.

Exemple : concentrations de cobalt et de nickel



Quelle est la relation entre les concentrations de Co et de Ni ? Celle-ci peut-elle être approximée par une droite ?

Problème d'ajustement

- ☐ On considère une variable de réponse y que l'on cherche à expliquer par une covariable x.
- ☐ Supposons que l'on dispose de n observations concomitantes de x et y, notées $x_1, \ldots, x_n$ et $y_1, \ldots, y_n$, respectivement. On dispose donc de l'ensemble de points $(x_1, y_1)', \ldots, (x_n, y_n)'$, où $'$ désigne la transposition. On peut représenter ces points graphiquement, ce qui donne lieu à un "scatter plot".
- ☐ Le **problème d'ajustement** consiste à trouver une courbe $\mu(\cdot)$ qui passe le mieux possible par l'ensemble des points. On suppose ici que la fonction $\mu(\cdot)$ est déterminée par un nombre fini de paramètres. **Comment les calculer/estimer** ?
- ☐ S'il existe une **relation approximativement linéaire** entre les x_i et les y_i (détectable sur un scatter plot), on souhaite résumer celle-ci par une simple droite. On peut utiliser la corrélation pour mesurer la dépendance linéaire entre les deux variables correspondantes.

Estimation par moindres carrés

- But : estimer les paramètres de la fonction $\mu(\cdot)$.
- Les écarts verticaux entre les y_i (observations de la variable de réponse y) et les valeurs ajustées $\mu(x_i)$ sont

$$y_i - \mu(x_i), \quad i = 1, \ldots, n.$$

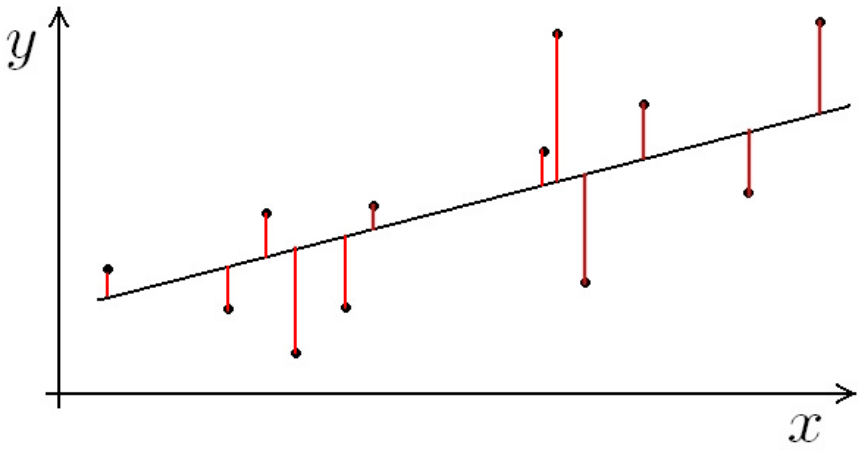
- On cherche les paramètres de la fonction $\mu(.)$ tels que la **somme des carrés** des écarts verticaux,

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

  soit minimale.
- L'ajustement est dit **linéaire** simple si $\mu(x) = \beta_0 + \beta_1 x$, $x \in \mathbb{R}$, où $\beta_0, \beta_1 \in \mathbb{R}$. Dans ce cas, il faut minimiser

$$\mathrm{SC}(\beta_0, \beta_1) = \sum_{i=1}^{n} \{y_i - \mu(x_i)\}^2 = \sum_{i=1}^{n} \{y_i - (\beta_0 + \beta_1 x_i)\}^2.$$

Estimation par moindres carrés : illustration



Estimateurs des moindres carrés

Théorème 5 *Supposons que $x_1, \ldots, x_n$ sont tels que au moins deux des x_i soient différents. Si l'on souhaite ajuster une relation du type $\mu(x) = \beta_0 + \beta_1 x$, alors les réalisations des* **estimateurs des moindres carrés** *de β_0 et β_1 sont*

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n x_i(y_i - \overline{y})}{\sum_{i=1}^n (x_i - \overline{x})^2} \quad \textit{et} \quad \widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x}.$$

Il est facile de voir que l'on a également

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n y_i(x_i - \overline{x})}{\sum_{i=1}^n (x_i - \overline{x})^2} = \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^n (x_i - \overline{x})^2}.$$

Définition 28 *La quantité $\widehat{\beta}_0 + \widehat{\beta}_1 x$ s'appelle la* **droite des moindres carrés**, *$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$ est la* **valeur ajustée** *correspondant à (x_i, y_i), et*

$$r_i = y_i - \widehat{y}_i = y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_i)$$

est le **résidu** *associé à y_i.*

Quelques propriétés

- ☐ La droite des moindres carrés passe par $(\overline{x}, \overline{y})$;
- ☐ $\sum_{i=1}^n r_i = 0$;
- ☐ $\sum_{i=1}^n x_i r_i = 0$;
- ☐ $\sum_{i=1}^n \widehat{y}_i r_i = 0$.

Décomposition de la somme totale des carrés

On déduit de la première et dernière égalité précédente que

$$\sum_{i=1}^{n}(y_i-\overline{y})^2=\sum_{i=1}^{n}(y_i-\widehat{y}_i+\widehat{y}_i-\overline{y})^2=\cdots=\sum_{i=1}^{n}(\widehat{y}_i-\overline{y})^2+\sum_{i=1}^{n}r_i^2.$$

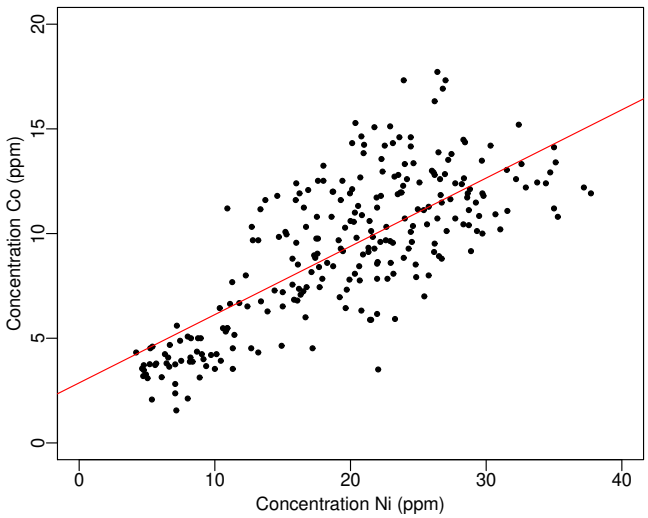
Ainsi,

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

où :

- $\mathrm{SC}_{\mathrm{Total}}=\sum_{i=1}^{n}(y_i-\overline{y})^2$ est la somme totale des carrés des écarts à la moyenne (variation totale).
- $\mathrm{SC}_{\mathrm{R}}=\sum_{i=1}^{n}(\widehat{y}_i-\overline{y})^2$ est la somme des carrés due à la régression (variation expliquée par la régression).
- $\mathrm{SC}_{\mathrm{E}}=\sum_{i=1}^{n}r_i^2$ est la somme des carrés due à l'erreur (variation non-expliquée par le modèle).

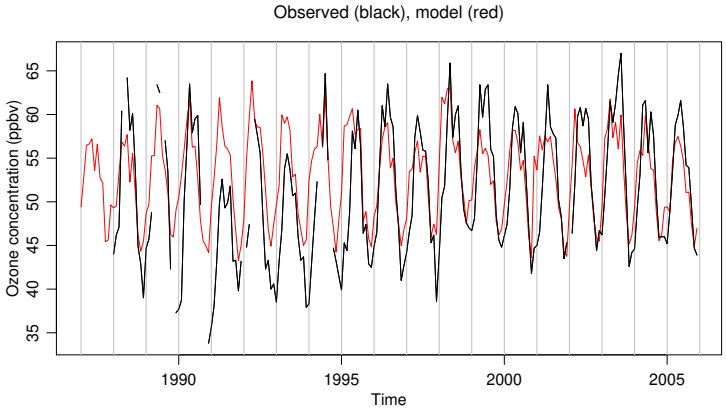
Concentration de cobalt et de nickel : régression linéaire



Droite des moindres carrés : $\widehat{\beta_0}+\widehat{\beta_1}x=2.59+0.33x$.

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.



Soient $y_1, \ldots, y_n$ les données observées et $x_1, \ldots, x_n$ les résultats du modèle.

Exemple : ozone atmosphérique (régression linéaire)

- ☐ Il y a 207 paires "(observation, résultat du modèle) $= (y_i, x_i)$" complètes ainsi que 21 paires pour lesquelles la valeur y_i est manquante.
- ☐ On estime une relation linéaire entre les x_i et les y_i.
- ☐ A partir des paires complètes, on obtient la droite des moindres carrés

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x = -5.511 + 1.069x.$$

La décomposition de la variation totale donne

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

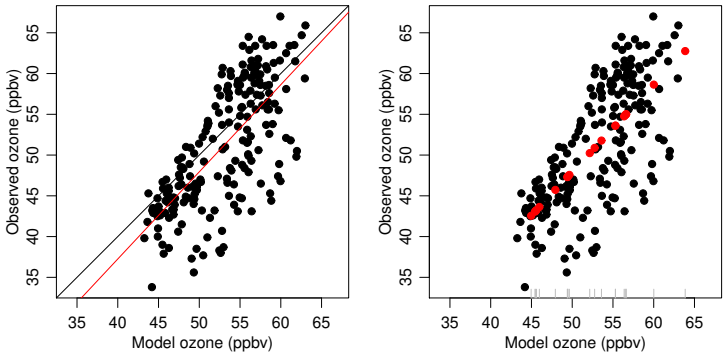
Ainsi, la régression explique environ la moitié de la somme des carrés totale.

- ☐ Pour une paire "(observation, modèle) $= (\,?, x_k)$", on peut remplacer la valeur manquante par la valeur ajustée correspondante

$$\widehat{y}_k = \widehat{\beta}_0 + \widehat{\beta}_1 x_k.$$

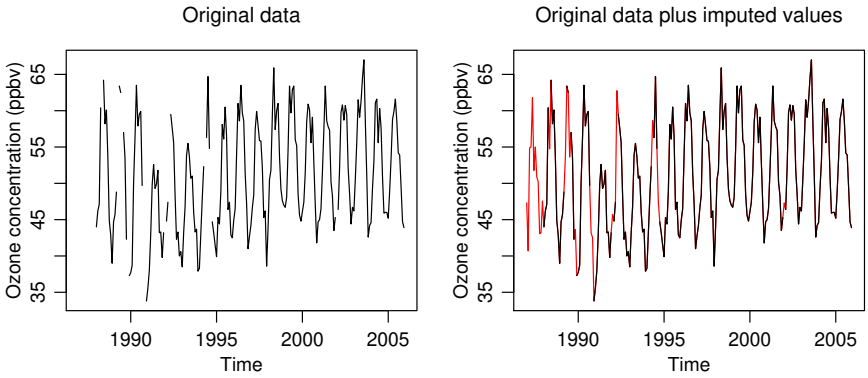
On parle d'imputation de donnée.

Exemple : ozone atmosphérique (modèle ajusté)



- Gauche : droite $y = x$ (noir) et droite ajustée $\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x = -5.511 + 1.069x$ (rouge).
- Droite : valeurs ajustées pour certaines valeurs manquantes y_i (rouge).

Exemple : ozone atmosphérique (valeurs imputées)



Gauche : données originales.

Droite : données originales (noir) et valeurs imputées (rouge).

4.2 Modèle statistique

Régression linéaire simple

- ☐ On rappelle que Y est la variable de réponse et que x est la covariable. En pratique, on n'a jamais exactement $Y = \mu(x)$, et c'est d'ailleurs pour cela que l'on considère Y comme une variable aléatoire.
- ☐ Pour modéliser ceci, on introduit un terme d'erreur (ou de bruit) aléatoire. Ici, comme souvent, ce dernier est supposé gaussien.
- ☐ On suppose que les $y_1, \dots, y_n$ sont des réalisations de variables aléatoires indépendantes $Y_1, \dots, Y_n$ telles que

$$Y_i \sim \mathcal{N}\left(\mu(x_i), \sigma^2\right), \quad i = 1, \dots, n.$$

Cela se réécrit

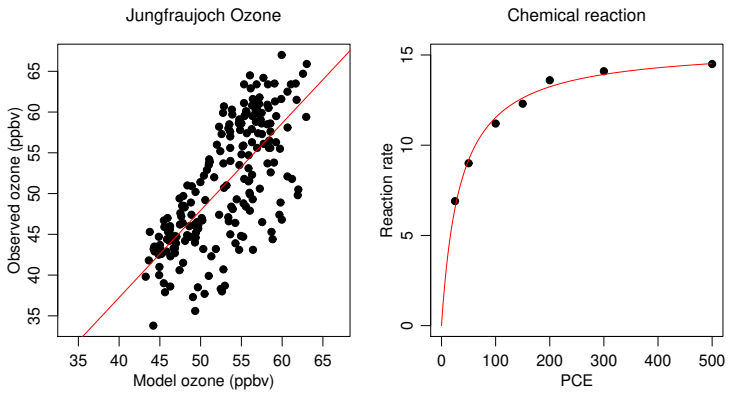
$$Y_i = \mu(x_i) + \varepsilon_i, \quad i = 1, \dots, n,$$

où $\varepsilon_1, \dots, \varepsilon_n \overset{\text{iid}}{\sim} \mathcal{N}\left(0, \sigma^2\right)$.

- ☐ Ainsi la relation entre Y et x est donnée par $\mathrm{E}(Y) = \mu(x)$. Le bruit autour de cette moyenne est caractérisé par σ^2.

Exemples

A gauche : $\mu(\cdot)$ linéaire, σ^2 grand. A droite : $\mu(\cdot)$ non-linéaire, σ^2 petit.



Linéarité

Quand on parle de régression linéaire ou de modèle linéaire, la linéarité s'entend par rapport aux paramètres (et non aux covariables). Par exemple :

☐ Le modèle

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n,$$

où $\varepsilon_1, \ldots, \varepsilon_n \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$, est linéaire (car linéaire en β_0 et β_1, i.e., par rapport au vecteur $(\beta_0, \beta_1)'$).

☐ Le modèle

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \varepsilon_i, \quad i = 1, \ldots, n,$$

où $\varepsilon_1, \ldots, \varepsilon_n \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$, est linéaire (car linéaire en β_0, β_1, β_2 et β_3).

Linéarité

☐ Le modèle

$$Y_i = \gamma_0 x_i^{\gamma_1} \eta_i, \quad i = 1, \ldots, n,$$

où $\eta_1, \ldots, \eta_n \overset{\text{iid}}{\sim} \exp(1)$, devient linéaire après transformation logarithmique. En effet,

$$\ln Y_i = \ln \gamma_0 + \gamma_1 \ln x_i + \ln \eta_i = \beta_0 + \beta_1 \tilde{x}_i + \ln \eta_i, \quad i = 1, \ldots, n,$$

où $\beta_0 = \ln \gamma_0$, $\beta_1 = \gamma_1$ et $\tilde{x} = \ln x$, est linéaire par rapport à β_0 et β_1.

☐ Le modèle

$$Y_i = \frac{\gamma_0 x_i}{\gamma_1 + x_i} + \varepsilon_i, \quad i = 1, \ldots, n,$$

où $\varepsilon_1, \ldots, \varepsilon_n \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$, n'est pas linéaire (car non-linéaire en γ_0 et γ_1).

Estimation des paramètres du modèle linéaire simple

Nous supposons que $\mu(x) = \beta_0 + \beta_1 x$, $x \in \mathbb{R}$, où $\beta_0, \beta_1 \in \mathbb{R}$.

- ☐ Il y a trois paramètres inconnus : l'ordonnée à l'origine β_0, la pente β_1 et la variance de l'erreur σ^2. Ainsi, $\theta = (\beta_0, \beta_1, \sigma^2) \in \mathbb{R}^2 \times \mathbb{R}_+$.
- ☐ Nous les estimons par la méthode du maximum de vraisemblance.
- ☐ Il est facile de voir que la log-vraisemblance (version variable aléatoire) s'écrit

$$\ell(\theta) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\{Y_i - (\beta_0 + \beta_1 x_i)\}^2.$$

  En maximisant ℓ par rapport à θ, nous obtenons (après calculs)

$$\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1\overline{x}, \quad \widehat{\beta}_1 = \frac{\sum_{i=1}^n x_i(Y_i - \overline{Y})}{\sum_{i=1}^n (x_i - \overline{x})^2}, \quad \widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2 = \frac{1}{n}\sum_{i=1}^{n} R_i^2.$$

- ☐ On observe que les estimateurs $\widehat{\beta}_0$ et $\widehat{\beta}_1$ sont les estimateurs des moindres carrés. Par ailleurs, ils sont sans biais. En revanche, $\mathrm{E}(\widehat{\sigma}^2) < \sigma^2$ et on préfère l'estimateur non biaisé S^2, où

$$S = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n} R_i^2} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2}.$$

Inférence pour les paramètres du modèle linéaire simple

Le coefficient β_1 (pente) est plus intéressant que β_0 (ordonnée à l'origine). On se concentre donc ici sur l'inférence concernant β_1.

- ☐ La "standard error" (notée sde) d'un estimateur (parfois appelée erreur type en français) correspond à sa déviation standard. Il s'agit d'un bon indicateur de précision dans le cas d'un estimateur sans biais. Celle-ci est en général inconnnue mais il est possible de l'estimer.
- ☐ On peut montrer que

$$\mathrm{Var}(\widehat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \overline{x})^2}.$$

  Ainsi, un estimateur sans-biais de la "standard error" de $\widehat{\beta}_1$ est

$$\widehat{\mathrm{sd}}(\widehat{\beta}_1) = \frac{S}{\sqrt{\sum_{i=1}^n (x_i - \overline{x})^2}},$$

  et sa valeur estimée est obtenue en remplaçant S par sa valeur observée s.

Inférence pour les paramètres du modèle linéaire simple

☐ Il est possible d'établir (admis) que

$$T = \frac{\widehat{\beta}_1 - \beta_1}{S/\sqrt{\sum_{i=1}^n (x_i - \overline{x})^2}} \sim t_{n-2}.$$

Notons que les résultats de la slide précédente nous donnent que

$$T = \frac{\widehat{\beta}_1 - \beta_1}{\widehat{\mathrm{sd}}(\widehat{\beta}_1)}.$$

☐ En choisissant T comme pivot et statistique de test respectivement, nous pouvons appliquer les idées du chapitre précédent pour obtenir des intervalles de confiance et effectuer des tests statistiques à propos de β_1.

Intervalles de confiance pour β_1

On en déduit des intervalles de confiance pour β_1 au niveau de confiance $1-\alpha$, pour $\alpha \in (0,1)$:

☐ Intervalle de confiance bilatéral symétrique :

$$\left[\widehat{\beta}_1 - t_{n-2,1-\alpha/2}\frac{S}{\sqrt{\sum_{i=1}^n (x_i - \overline{x})^2}}, \widehat{\beta}_1 + t_{n-2,1-\alpha/2}\frac{S}{\sqrt{\sum_{i=1}^n (x_i - \overline{x})^2}}\right].$$

☐ Intervalle de confiance unilatéral à gauche :

$$\left(-\infty, \widehat{\beta}_1 + t_{n-2,1-\alpha}\frac{S}{\sqrt{\sum_{i=1}^n (x_i - \overline{x})^2}}\right].$$

☐ Intervalle de confiance unilatéral à droite :

$$\left[\widehat{\beta}_1 - t_{n-2,1-\alpha}\frac{S}{\sqrt{\sum_{i=1}^n (x_i - \overline{x})^2}}, \infty\right).$$

Tests pour β_1

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

- ☐ Test bilatéral $H_0 : \beta_1 = \beta_1^{(0)}$ contre $H_1 : \beta_1 \neq \beta_1^{(0)}$. On rejette H_0 si et seulement si $|t_{\text{obs}}| > t_{n-2,1-\alpha/2}$.
- ☐ Test unilatéral à gauche $H_0 : \beta_1 = \beta_1^{(0)}$ contre $H_1 : \beta_1 < \beta_1^{(0)}$. On rejette H_0 si et seulement si $t_{\text{obs}} < t_{n-2,1-\alpha}$.
- ☐ Test unilatéral à droite $H_0 : \beta_1 = \beta_1^{(0)}$ contre $H_1 : \beta_1 > \beta_1^{(0)}$. On rejette H_0 si et seulement si $t_{\text{obs}} > t_{n-2,1-\alpha}$.

Exemple : données d'ozone

Affichage des données d'ozone à l'aide du logiciel R :

```
> 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
...
```

Exemple : données d'ozone (inférence)

Résultat de l'ajustement du modèle linéaire aux données d'ozone, effectué à l'aide du logiciel R :

```
> 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_1 au niveau de confiance $1-\alpha$ est
$$\left[\widehat{\beta}_1 - t_{n-2,1-\alpha/2}\widehat{\mathrm{sd}}(\widehat{\beta}_1), \widehat{\beta}_1 + t_{n-2,1-\alpha/2}\widehat{\mathrm{sd}}(\widehat{\beta}_1)\right].$$
- ☐ Ainsi, en lisant les sorties du logiciel, on obtient qu'une réalisation de l'IC précédent pour β_1 au niveau de confiance 95% est donnée par
$$1.06903 \pm t_{205,0.975} \times 0.07479 \doteq 1.07 \pm 1.97 \times 0.07 = [0.93, 1.21].$$
- ☐ Souvent, on veut tester si le terme impliquant la covariable est significatif. Cela revient à tester $H_0 : \beta_1 = 0$.
- ☐ Ici, le scatter plot semble clairement indiquer que β_1 est différent de 0 et on effectue donc plutôt le test $H_0 : \beta_1 = 1$. 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
$$T = \frac{\widehat{\beta}_1 - 1}{\widehat{\mathrm{sd}}(\widehat{\beta}_1)}$$
est strictement supérieure à $t_{n-2,1-\alpha/2} = t_{205,0.975} \doteq 1.97$. On a $t_{\mathrm{obs}} \doteq 0.92$ et on ne rejette donc pas H_0.

Coefficient de détermination

- Nous avons déjà vu la **décomposition de la somme totale des carrés**

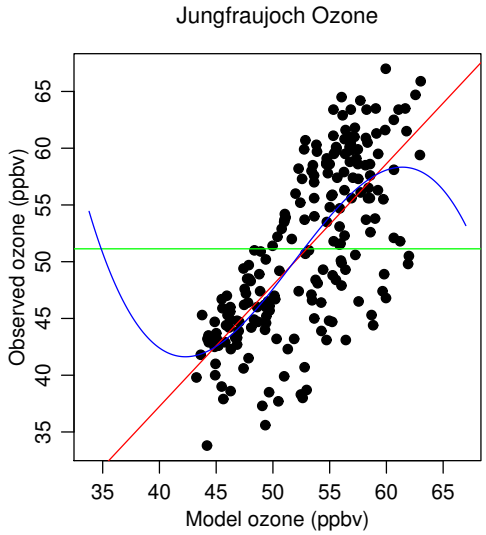
$$\sum_{i=1}^{n}(Y_i-\overline{Y})^2=\sum_{i=1}^{n}(\widehat{Y}_i-\overline{Y})^2+\sum_{i=1}^{n}R_i^2, \quad \text{soit} \quad \mathrm{SC}_{\mathrm{Total}}=\mathrm{SC}_{\mathrm{R}}+\mathrm{SC}_{\mathrm{E}},$$

  en une partie expliquée par la régression (SC_{R}) et une partie due à l'erreur (SC_{E}).
- La proportion 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 appelée **coefficient de détermination**. On a $0\leq R^2\leq 1$.
- $R^2\approx 1$ implique $\widehat{y}_i\approx y_i$ et donc $r_i\approx 0$ pour tout $i=1,\ldots,n$: le modèle explique très bien les données ; $R^2\approx 0$ implique $\widehat{\beta}_1\approx 0$: la covariable n'explique presque rien de la variation des Y_i.
- Données d'ozone : $R^2=0.5$, donc la moitié de la variation est expliquée par le modèle ; Données chimiques : $R^2=0.99$, donc le modèle explique presque la totalité de la variation.

Comparaison de modèles



- Nous souhaitons comparer les modèles

$$\begin{aligned} Y_i &= \beta_0+\varepsilon_i,\\ Y_i &= \beta_0+\beta_1 x_i+\varepsilon_i,\\ Y_i &= \beta_0+\beta_1 x_i+\beta_2 x_i^2+\beta_3 x_i^3+\varepsilon_i, \end{aligned}$$

  où $\varepsilon_1,\ldots,\varepsilon_n \overset{\text{iid}}{\sim} \mathcal{N}(0,\sigma^2)$.
- Le modèle rouge semble être bien meilleur que le vert, mais le rouge et le bleu semblent avoir une performance similaire. Comment tester ces constats ?

Loi de Fisher

Définition 29 *Soient U_1 et U_2 des variables aléatoires indépendantes telles que $U_1 \sim \chi^2_{d_1}$ et $U_2 \sim \chi^2_{d_2}$, où d_1 et d_2 sont des entiers positifs. La variable aléatoire*

$$X = \frac{U_1/d_1}{U_2/d_2}$$

suit la **loi de Fisher (ou de Fisher-Snedecor ou encore F de Snedecor) à d_1 et d_2 degrés de liberté, notée F_{d_1,d_2}.**

Remarque : Il est facile d'établir le lien suivant entre la loi de Student et la loi de Fisher : si $Y \sim t_\nu$ alors $Y^2 \sim F_{1,\nu}$.

Comparaison de modèles (régression linéaire simple)

- ☐ On souhaite comparer le modèle sans covariable et le modèle linéaire avec une covariable, i.e.,

  $$Y_i = \beta_0 + \varepsilon_i \quad \text{et} \quad Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n,$$

  où $\varepsilon_1, \ldots, \varepsilon_n \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$.
- ☐ Pour tester s'il vaut la peine d'ajouter le terme $\beta_1 x$, on considère l'hypothèse nulle $H_0 : \beta_1 = 0$. Sous H_0, on a

  $$F_{\text{s}} = \frac{\text{SC}_{\text{R}}/1}{\text{SC}_{\text{E}}/(n-2)} \sim F_{1,n-2},$$

  et on peut donc fonder un test sur la statistique F_{s}. Soit $\alpha \in (0,1)$ le niveau de significativité α. On rejette H_0 au si et seulement si $f_{\text{s,obs}} > F_{1,n-2,1-\alpha/2}$, où $F_{1,n-2,1-\alpha/2}$ est le quantile au niveau $1-\alpha/2$ de la loi de Fisher à 1 et $n-2$ degrés de liberté.
- ☐ Ce test de $H_0 : \beta_1 = 0$ est parfaitement équivalent au test décrit précédemment.
- ☐ Sur les données d'ozone, on obtient $f_s = 204.3$. Sachant que $F_{1,205,0.95} = 3.887$, on rejette $H_0 : \beta_1 = 0$. La p-valeur correspondante est inférieure à 2.2×10^{-16}.

Comparaison de modèles (régression linéaire multiple)

- Considérons le modèle linéaire, pour $q < p$,

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

  où $\varepsilon_1, \ldots, \varepsilon_n \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$.
- Afin de tester s'il est utile de prendre en compte les covariables $x^{(q+1)}, \ldots, x^{(p)}$, on considère $H_0 : \beta_{q+1} = \cdots = \beta_p = 0$.
- Pour ce test, on utilise les sommes des carrés dues aux erreurs suivantes : $\mathrm{SC}_{\mathrm{E},p}$ qui correspond au modèle avec l'ensemble des p covariables $x^{(1)}, \ldots, x^{(p)}$ et $\mathrm{SC}_{\mathrm{E},q}$ qui correspond au modèle réduit impliquant seulement les q premières covariables $x^{(1)}, \ldots, x^{(q)}$. On a $\mathrm{SC}_{\mathrm{E},p} \leq \mathrm{SC}_{\mathrm{E},q}$ et l'idée est de rejeter H_0 si l'ajout de $x^{(q+1)}, \ldots, x^{(p)}$ diminue substantiellement la somme des carrés due aux erreurs. Sous H_0 on a

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

  On peut donc fonder un test sur la statistique F_{m}. Soit $\alpha \in (0,1)$ le niveau de significativité. On rejette H_0 si et seulement si $f_{\mathrm{m,obs}} > F_{p-q,n-p-1,1-\alpha}$.

Application aux données d'ozone

Dans le cas des données d'ozone, on s'intéresse au modèle (modèle bleu présenté précédemment) :

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \varepsilon_i, \quad i = 1, \ldots, n,$$

où $\varepsilon_1, \ldots, \varepsilon_n \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$. Afin d'évaluer une potentielle évidence du fait que le modèle bleu est meilleur que le rouge, on teste $H_0 : \beta_2 = \beta_3 = 0$. On a $n = 207$, $p = 3$, $q = 1$, et

$$f_{\mathrm{m,obs}} = \frac{(5831.9 - 5712.2)/(3-1)}{5712.2/(207-3-1)} = 2.13.$$

Sachant que $F_{3-1,207-3-1,0.95} = F_{2,203,0.95} = 3.04$, on ne rejette pas H_0. Il n'y a pas assez d'évidence dans les données pour préférer le modèle bleu au modèle rouge.

Validation du modèle de régression linéaire simple

A posteriori, il faut vérifier que les hypothèses sous-jacentes sont appropriées. Le modèle linéaire simple gaussien est fondé sur quatre hypothèses principales :

- ☐ Linéarité : $\mathrm{E}(Y)$ est correctement spécifiée, i.e., $\mu(x) = \beta_0 + \beta_1 x$ est adaptée.
- ☐ Homoscédasticité (variance constante) des erreurs : pour tout $i = 1, \ldots, n$, $\mathrm{Var}(\varepsilon_i) = \sigma^2$.
- ☐ Normalité des erreurs.
- ☐ Indépendance des erreurs : pour tout $i, j = 1, \ldots, n$, ε_i et ε_j sont indépendantes.

La normalité des erreurs implique que

$$\frac{Y_i - (\beta_0 + \beta_1 x_i)}{\sigma} \sim \mathcal{N}(0,1), \quad i = 1, \ldots, n,$$

et donc que les **résidus standardisés**

$$\tilde{R}_i = \frac{Y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_i)}{S}$$

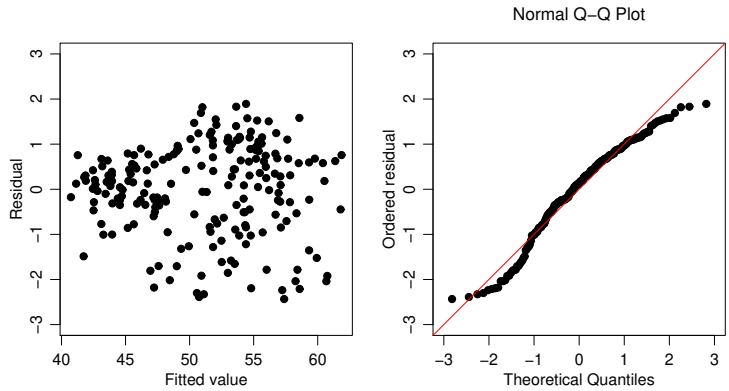
vérifient

$$\tilde{R}_i \dot{\sim} \mathcal{N}(0,1), \quad i = 1, \ldots, n.$$

Validation du modèle de régression linéaire simple

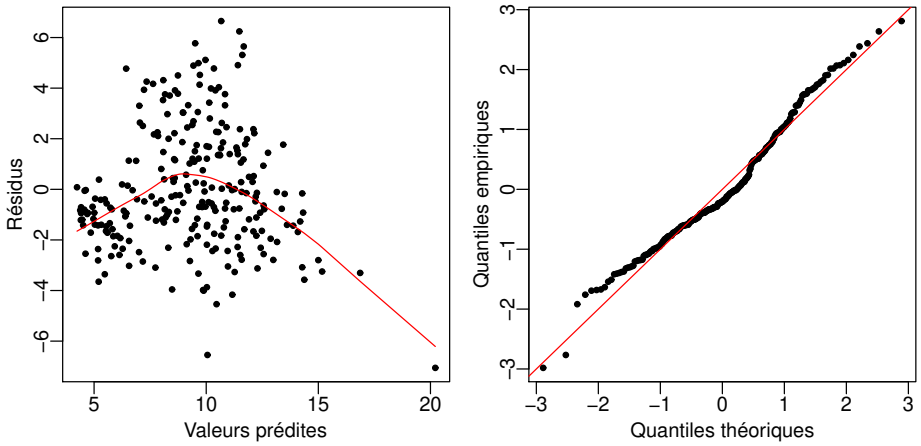
- ☐ Afin d'analyser si $\mathrm{E}[Y]$ est bien spécifiée, on peut tracer le scatter plot des résidus r_i en fonction des x_i. Aucun pattern particulier ne devrait apparaître. Tout pattern systématique (par exemple une parabole) indique que μ est inadéquat.
- ☐ Pour vérifier que l'hypothèse d'homoscédasticité est acceptable, on trace le scatter plot des résidus r_i en fonction des $\widehat{y}_i$. On s'attend à un nuage de points sans variation de la dispersion. La présence de patterns spécifiques (tels un élargissement du nuage de points) indique une violation de l'hypothèse.
- ☐ Pour évaluer l'hypothèse de normalité des erreurs, on utilise un quantile-quantile plot (Q-Q plot) visant à vérifier la normalité des résidus standardisés. Un Q-Q plot normal est un graphique des quantiles empiriques des données (ici les résidus standardisés) contre les quantiles théoriques de la loi $\mathcal{N}(0,1)$. Si les $\tilde{r}_i$ 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.
- ☐ Afin de juger l'hypothèse d'indépendance, il convient d'utiliser des outils de la théorie des séries temporelles qui vont au-delà de ce cours.

Exemple : données d'ozone



- Gauche : scatter plot des r_i contre les $\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$. On constate un élargissement modéré du nuage de points, qui indique que l'hypothèse d'homoscédasticité n'est pas parfaitement vérifiée.
- Droite : Q-Q plot normal des $\tilde{r}_i$. On observe des écarts non négligeables par rapport à la diagonale (en rouge). La loi des erreurs n'est pas normale. Dans le cas présent, elle est même asymétrique.

Exemple : concentration de métaux



- Gauche : scatter plot des r_i contre les $\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$. La dispersion varie en fonction des $\widehat{y}_i$ et l'hypothèse d'homoscédasticité n'est donc pas parfaitement vérifiée.
- Droite : Q-Q plot normal des $\tilde{r}_i$. On observe des écarts non négligeables par rapport à la diagonale (en rouge). La loi des erreurs n'est donc pas normale.