HAL
archives-ouvertes.fr

Cours de Statistiques niveau L1-L2

Kévin Polisano

▶ **To cite this version:**

Kévin Polisano. Cours de Statistiques niveau L1-L2. Licence. France. 2018. <cel-01787365>

HAL Id: cel-01787365
https://hal.archives-ouvertes.fr/cel-01787365

Submitted on 7 May 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Cours de Statistiques

(L1 – MAP 201)

Kévin Polisano
(kevin.polisano@univ-grenoble-alpes.fr)

Université Grenoble Alpes

14 février 2018

Déroulement du cours

- 4 séances de 1h30 en amphithéâtre (cours)
  1. Statistiques descriptives
  2. Introduction à la théorie des probabilités
  3. Estimation paramétrique
  4. Introduction aux tests d'hypothèse
- 4 séances de 3h en salle informatique (TP)
  1. Prise en main du logiciel R et statistique descriptive univariée
  2. Loi binomiale, loi normale et théorèmes limites (CC1)
  3. Estimation ponctuelle, loi du χ^2 et de student
  4. Applications des intervalles de confiance et tests statistiques (paranormal, cryptographie, adéquation de loi, etc.) (CC2)
- 1 langage et logiciel de programmation dédié aux statistiques : R
- 1 note de contrôle continu (CC) moyenne de CC1 et CC2
- 1 note d'examen écrit (EX)

$$\textbf{Note finale} = \max(\text{EX}, \text{moyenne}(\text{CC}, \text{EX}))$$

Supports de cours

Des notes de cours

- Notes de cours d'Élise Arnaud

  https://team.inria.fr/steep/files/2015/03/cours.pdf

- Notes de cours d'Olivier Gaudoin

  https://www-ljk.imag.fr/membres/Olivier.Gaudoin/PMS.pdf

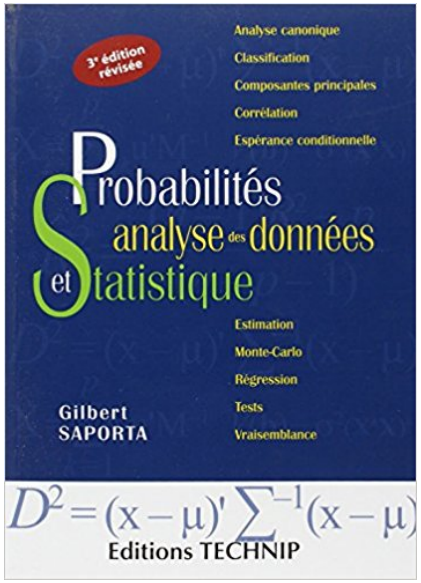
- Notes de cours de Bernard Ycart

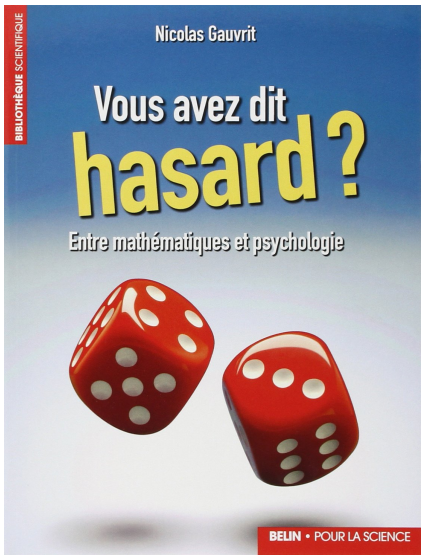
  https://toltex.u-ga.fr/SPLS

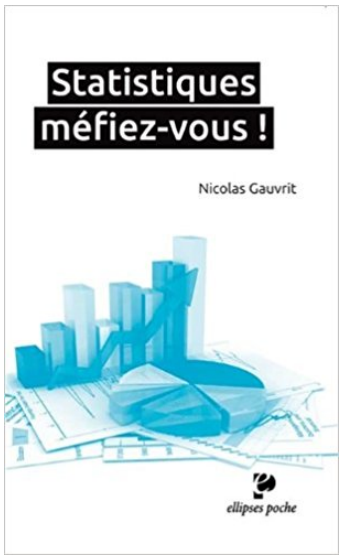
Supports de cours

Des livres

Gilbert Saporta, *Probabilités, analyse des données et statistique*, Editions Technip, 2006.







Nicolas Gauvrit,
Vous avez dit hasard ?, 2014.
Statistiques, méfiez-vous !, 2007.

Partie I : Statistiques descriptives

1. Introduction
2. Bases de la statistique descriptive
   - Vocabulaire
   - Tableaux statistiques
   - Méfiez-vous des statistiques ! Le paradoxe de Simpson
3. Représentations graphiques
   - Histogrammes
   - Fonction de répartition empirique
4. Indicateurs statistiques
   - Indicateurs de localisation ou de tendance centrale
   - Indicateurs de dispersion ou de variabilité
5. Corrélation et causalité
   - Régression linéaire
   - Exemples de corrélations

Définition : « Le mot statistique désigne à la fois un ensemble de données d'observations et l'activité qui consiste dans leur recueil, leur traitement et leur interprétation » (Encyclopedia Universalis)

Étymologie : « De l'allemand *Staatskunde*, dérivé de l'italien *statista* (homme d'État, statiste), la statistique représentant l'ensemble des connaissances que doit posséder un homme d'État. » (1785)

Introduction

Histoire de la statistique

- Recensements en Chine au XXIII^{e} siècle av. J.-C. ou en Égypte au XVIII^{e} av. J.-C, système de recueil se poursuivant jusqu'au XVII^{e}.
- Rôle prévisionnel des statistiques au XVIII^{e} siècle avec la construction des premières tables de mortalité avec Antoine Deparcieux, *l'Essai sur les probabilités de la durée de vie humaine* (1746).
- Rôle démographique au XIX^{e} siècle, le Baron de Reiffenberg présentait en 1842 à l'Académie ses calculs rétrospectifs de population chez des peuples gaulois, d'après des chiffres donnés par Jules César dans sa conquête des gaules.

Introduction

Histoire de la statistique mathématique

- Premiers textes connus sur le calcul des hasards (ou des chances) au XVI^{e} siècle avec Cardan et au XVII^{e} siècle avec Galilée.
- Début officiel avec Pascal, Fermat et Huyguens au XVII^{e} siècle.
- Tournant au XVIII^{e} siècle avec Montmort (combinatoire), Bernoulli (loi des grands nombres) puis De Moivre et Laplace (traitement analytique des probabilités et théorèmes limites).
- Théorie des ensembles et de la mesure par Borel et Lebesgue et calcul des probabilités par Lévy au XX^{e} siècle
- Axiomatisation de la théorie des probabilités par Kolmogorov (1933).

Pour aller plus loin :

- Brigitte Chaput et al., *Autour de la modélisation en probabilités*, Histoire 81, 2005.
- Ian Hacking, *The emergence of probability : A philosophical study of early ideas about probability, induction and statistical inference*, Cambridge University Press, 2006.

Introduction

Objectifs du cours

But du cours :

- faire quelques rappels et connaître le vocabulaire
- savoir décrire et représenter un ensemble de données
- vous réconcilier avec les probabilités et les statistiques ... ?
- comprendre le lien entre les deux

Introduction

Divers domaines d'application

- **Economie, assurance, finance** : études quantitatives de marchés, prévisions économétriques, analyse de la consommation des ménages, taxation des primes d'assurances et de franchises, gestion de portefeuille, évaluation d'actifs financiers, ...
- **Biologie, médecine** : essais thérapeutiques, épidémiologie, dynamique des populations, analyse du génôme, ...
- **Sciences de la terre** : prévisions météréologiques, exploration pétrolière, ...
- **Sciences humaines** : enquêtes d'opinion, sondages, étude de population, ...
- **Sciences de l'ingénieur** : contrôle qualité, sûreté de fonctionnement, évaluation des performances, ...
- **Sciences de l'information** : traitement des images et des signaux, reconnaissance de forme et parole, *machine learning*, ...

Crédits : O. Gaudoin

Introduction

But de la Statistique

Les données sont entâchées d'incertitudes et présentent des variations pour plusieurs raisons :

- le déroulement des phénomènes observés n'est pas prévisible à l'avance avec certitude
- toute mesure est entâchée d'erreur
- seuls quelques individus sont observés
- ...

$\Rightarrow$ données issues de phénomènes aléatoires
$\Rightarrow$ intervention du hasard et des probabilités

Objectifs : maîtriser au mieux cette incertitude pour **extraire des informations utiles des données**, par l'intermédiaire de **l'analyse des variations** dans les observations.

Introduction

Deux classes de méthodes statistiques

1. **Statistique descriptive** : elle a pour but de **résumer l'information** contenue dans les données de façon synthétique et efficace par :
   - Représentations graphiques
   - Indicateurs de position, de dispersion et de relation
   - Régression linéaire

   $\Rightarrow$ permet de dégager les caractéristiques essentielles du phénomène étudié et de suggérer des hypothèses pour une étude ultérieure plus poussée. Les probabilités n'ont ici qu'un rôle mineur.
2. **Statistique inférentielle** : elle a pour but de faire des **prévisions** et de **prendre des décisions** au vu des observations par :
   - Estimation paramétrique
   - Intervalles de confiance, tests d'hypothèse

   $\Rightarrow$ Nécessite de définir des **modèles probabilistes** du phénomène aléatoire et savoir gérer les risques d'erreurs.

Probabilité vs. Statistique

- la statistique repose sur l'observation de phénomènes concrets et utilise les probabilités comme **outils d'analyse et de généralisation**
- la théorie des probabilités permet de modéliser efficacement certains phénomènes aléatoires et d'en **faire l'étude théorique.**

Probabilité vs. Statistique

Le calcul des probabilités propose des modèles simplificateurs du comportement d'un phénomène

- les données observées sont souvent imprécises. Le modèle probabiliste permet de représenter comme des variables aléatoires les déviations entre “vraies" valeurs et valeurs observées.
- la répartition statistique d'une variable au sein de la population est souvent voisine de modèles mathématiques proposés par le calcul des probabilités (ex : supposer que la durée de vie d'un composant electronique suit une loi exponentielle).

Le calcul des probabilités fournit des théorèmes si le processus d'échantillonnage equiprobable des individus parmi la population est respecté.

Résumé de la démarche statistique

1. **Recueil des données** $\Rightarrow$ construction d'un échantillon
2. **Statistique exploratoire** $\Rightarrow$ formulation d'hypothèses sur la nature du phénomène aléatoires
3. **Choix d'un modèle probabiliste** $\Rightarrow$ test d'adéquation
4. **Estimation des paramètres inconnus du modèle** $\Rightarrow$ construction d'estimateurs
5. **Prévision sur les observations futures** $\Rightarrow$ associer un degré de confiance

1 Introduction

2 Bases de la statistique descriptive
- Vocabulaire
- Tableaux statistiques
- Méfiez-vous des statistiques ! Le paradoxe de Simpson

3 Représentations graphiques
- Histogrammes
- Fonction de répartition empirique

4 Indicateurs statistiques
- Indicateurs de localisation ou de tendance centrale
- Indicateurs de dispersion ou de variabilité

5 Corrélation et causalité
- Régression linéaire
- Exemples de corrélations

Vocabulaire

- Faire de la statistique suppose que l'on étudie un ensemble d'objets équivalents sur lesquels on observe des caractéristiques appelées **variables**.
- Le groupe ou l'ensemble d'objets équivalents est appelé la **population**.
- Les objets sont appelés des **individus**.
- En général, la population et trop vaste pour pouvoir être observée exhaustivement. On étudie alors la variable sur une sous partie de la population. On étudie alors un **échantillon**.

Vocabulaire

On souhaite étudier un caractère X prenant ses valeurs dans Ω, sur une population $\mathcal{P}$.

Exemple : si l'échantillon est un groupe de TD de MAP 201 ...

- **un individu** est un étudiant
- **la population** peut être l'ensemble de étudiants de MAP 201, des L1, de Grenoble, de France etc.
- **les variables** étudiées peuvent être le sexe, la taille, la moyenne d'année, le nombre de cafés consommés, etc.

Vocabulaire

En général, on ne peut pas observer ce caractère sur tous les individus d'une grande population, mais seulement sur une sous-population de $\mathcal{P}$ de taille n. On notera alors :

- la sous population : $\{i_1, ..., i_j, ..., i_n\}$ un ensemble de n individus choisis au hasard dans $\mathcal{P}$.
- l'échantillon de données : $x_1, ..., x_j, ..., x_n$ les n valeurs observées du caractère X sur les individus de la sous-population.

Deux problèmes se posent alors :

1. **Quelles informations sur le caractère X** peut-on tirer de l'échantillon ?
2. **Quelle prévision** pourrait on faire sur un individu non observé de $\mathcal{P}$ à partir des données observées $x_1, ..., x_j, ..., x_n$?

Vocabulaire

Chaque **individu** est décrit par un ensemble de **variables** X. Ces variables peuvent être classées selon leur nature :

- variable qualitative s'exprimant par l'appartenance à une modalité. Ω={Homme, Femme} ; Ω={Rap, chanson française, classique, etc.}
- variable quantitative, s'exprimant par des nombres réels, par exemple la taille des individus ou les résultats d'un examen.
- On distingue les variables quantitatives discrètes lorsque Ω est une suite finie ou infinie d'éléments de $\mathbb{N}$ (ex : $\Omega = \{1, 2, 3\}$; $\Omega = \mathbb{N}$) des variables quantitatives continues si toutes les valeurs d'un intervalle de $\mathbb{R}$ sont acceptables.

Vocabulaire

Concept clé en statistique : la **variabilité**, qui signifie que des individus en apparence semblables peuvent prendre des valeurs différentes.

Exemple :

Un processus industriel de fabrication ne fournit jamais des caractéristiques parfaitement constantes.

L'analyse statistique a pour but d'étudier cette variabilité

- en tenir compte pour prévoir de façon probabiliste le comportement d'individus non observés,
- chercher à l'expliquer à l'aide de variables extérieures,
- chercher à l'augmenter dans le but de distinguer les individus entre eux.

Tableaux statistiques - variables discrètes ou quantitatives

- n la taille de l'échantillon
- k le nombre de modalités.
- $m_i, i \in [1, k]$ les modalités
- n_i le nombre d'occurence (l'effectif) de m_i dans l'échantillon
- et f_i la fréquence correspondante.

on a $\sum n_i = n$; $f_i = n_i/n$; $\sum f_i = 1$.

Exemple du lancer d'un dé

$$x_i = \{2, 5, 6, 4, 5, 4, 2, 1, 6, 5, 1, 2\},\ n = 12$$

m_i	1	2	3	4	5	6
n_i	2	3	0	2	3	2

$\Rightarrow$ Vérification empirique qu'un dé est équilibré ?
$\Rightarrow$ Simulation informatique d'une loi uniforme ?

À vos boitiers de vote !

Mettez le boitier sur le canal 41 :
Pressez le bouton « Channel » puis tapez « 41 »



À vos boitiers de vote !

Choisissez au hasard un chiffre entre 1 et 9

1, 2, 3, 4, 5, 6, 7, 8, 9

À vos boitiers de vote !

Résultats du diagramme en bâtons

- Faible choix des valeurs extrêmes 1 et 9 ?
- Choix majoritaire du chiffre 7 ?

$\Rightarrow$ L'être humain est en général un piètre générateur de hasard.

Autre test : donnez au hasard une série de 200 « zéro ou un » à la main puis avec l'aide d'une pièce par « pile ou face ».

$\Rightarrow$ Repérable au nombre de « pile » (ou « face ») consécutifs et au biais d'alternance sous-jacent.

Tableaux statistiques - variables discrètes ou quantitatives

modalité m_i nb personnes	effectif n_i	fréquence f_i en pourcentage
1	7 381 150	31.0
2	7 404 960	31.1
3	3 857 246	16.2
4	3 285 802	13.8
5	1 309 559	5.5
6 et plus	571 444	2.4

Figure: Le recensement de 1999 donne la répartition des $n = 23810161$ ménages, selon la variable X nombre de personnes du ménage

Tableaux statistiques - variables continues

On regroupe les valeurs en k classes d'extrémité $a_0, a_1, \ldots, a_k$, et on note pour chaque classe $[a_{i-1}, a_i]$ l'effectif n_i, la fréquencef_i.

modalité classes d'âge	effectif n_i	fréquence f_i en pourcent
[0,4]	2 986 925	20.77
[5,9]	3 629 294	25.24
[10,14]	3 833 120	26.65
[15,19]	3 932 101	27.34

Figure: Le recensement de 1999 donne la répartition des $n = 14381440$ personnes moins de 20 ans, selon la classe d'âge

Tableaux statistiques - fréquences cumulées

modalité x_i	fréquence f_i	fréquence cumulée F_i
1	31.0	31.0
2	31.1	62.1
3	16.2	78.3
4	13.8	92.1
5	5.5	97.6
6 et plus	2.4	100

Figure: recensement de 1999 (a) répartition des ménages, selon le nombre de personnes du ménage (b) fonction de répartition empirique

Tableaux statistiques - tableaux de contingences

	femmes	**hommes**	total
agriculteurs exploitant	204 209	437 958	1.3 %
artisants, commerçants	484 443	1 174 609	3.4 %
cadres et professions intel. sup.	1 101 537	2 063 798	6.6 %
professions intermédiaires	2 771 948	2 990 937	11 %
employés	5 973 956	1 835 135	16.2 %
ouvriers	1 426 472	5 635 270	15.8 %
retraités	5 434 200	5 200 243	22.1 %
autres sans activité prof.	7 593 554	3 740 108	23.6 %
total	52 %	48 %	48 068 377

Figure: recensement de 1999 - population de 15 ans ou plus par sexe et catégorie socioprofessionnelle

À vos boitiers de vote !

Votre pouvoir d'achat a diminué de 12% en 2017 mais remontera de 12% en 2018. Votre pouvoir d'achat en 2018 sera :

A) Plus important qu'en 2017
B) Identique à 2017
C) Plus faible qu'en 2017
D) La réponse D

Méfiez-vous des statistiques !

Variations relatives

Votre pouvoir d'achat a diminué de 12% en 2017 mais remontera de 12% en 2018. Votre pouvoir d'achat en 2018 sera :

A) Plus important qu'en 2017
B) Identique à 2017
C) **Plus faible qu'en 2017**
D) La réponse D

Une quantité x subissant une diminution de $p\%$ puis une augmentation de $p\%$ s'écrit

$$x \times (1-p) \times (1+p) = x \times (1-p^2) \leq x$$

$\Rightarrow$ Pour $p = 12\%$ on obtient une baisse d'environ $1,5\%$.

Méfiez-vous des statistiques !

Variations relatives et absolues

La dette de la France, qui avait augmenté de 15% l'an passé, n'a augmenté cette année que de 14%.
Le gouvernement se félicite de sa gestion exemplaire.

Crédits : N. Gauvrit

Méfiez-vous des statistiques !

Variations relatives et absolues

La dette de la France, qui avait augmenté de 15% l'an passé, n'a augmenté cette année que de 14%.
Le gouvernement se félicite de sa gestion exemplaire.

- Dette de départ : 100 M€
- **Déficit 1ère année** : $15\% \times 100 = 15$ M€ $\Rightarrow$ dette $= 115$ M€
- **Déficit 2ème année** : $14\% \times 115 = 16,1$ M€ > 15 M€

Augmentation du déficit : de 15 milliards d'euros l'an passé il dépasse cette année 16 milliards d'euros !
L'opposition déplore la gestion du gouvernement.

« *Les statistiques ont une particularité majeure : elles ne sont jamais les mêmes selon qu'elles sont avancées par un homme de gauche ou par un homme de droite* » – Jacques Maillot.

Méfiez-vous des statistiques !

Variations relatives et absolues

Le syndicat d'une entreprise déclare :

Les ouvriers touchaient 200€ mensuels en 2017, on leur offre désormais 180€, soit une baisse de 10%. Les cadres gagnaient l'an dernier 2000€ mensuels, et aujourd'hui 1800€, soit là encore une baisse de 10%

La patron de l'entreprise affirme :

L'an dernier, le salaire mensuel moyen était de 363,64€. Il passe cette année à 916,34€, soit une augmentation de 152% !

« *Il y a trois sortes de mensonges : les mensonges, les sacrés mensonges et les statistiques* » – Mark Twain.

Crédits : N. Gauvrit

Méfiez-vous des statistiques !

Variations relatives et absolues

		employés	
		ouvriers	cadres
2017	salaire	200€	2000€
	effectif	1000	100
2018	salaire	180€	1800€
	effectif	600	500

$$\begin{cases} 200 \to 180 \\ 2000 \to 1800 \end{cases} \Rightarrow \text{baisse de } 10\%$$

Méfiez-vous des statistiques !

Variations relatives et absolues

		employés	
		ouvriers	cadres
2017	salaire	200€	2000€
	effectif	1000	100
2018	salaire	180€	1800€
	effectif	600	500

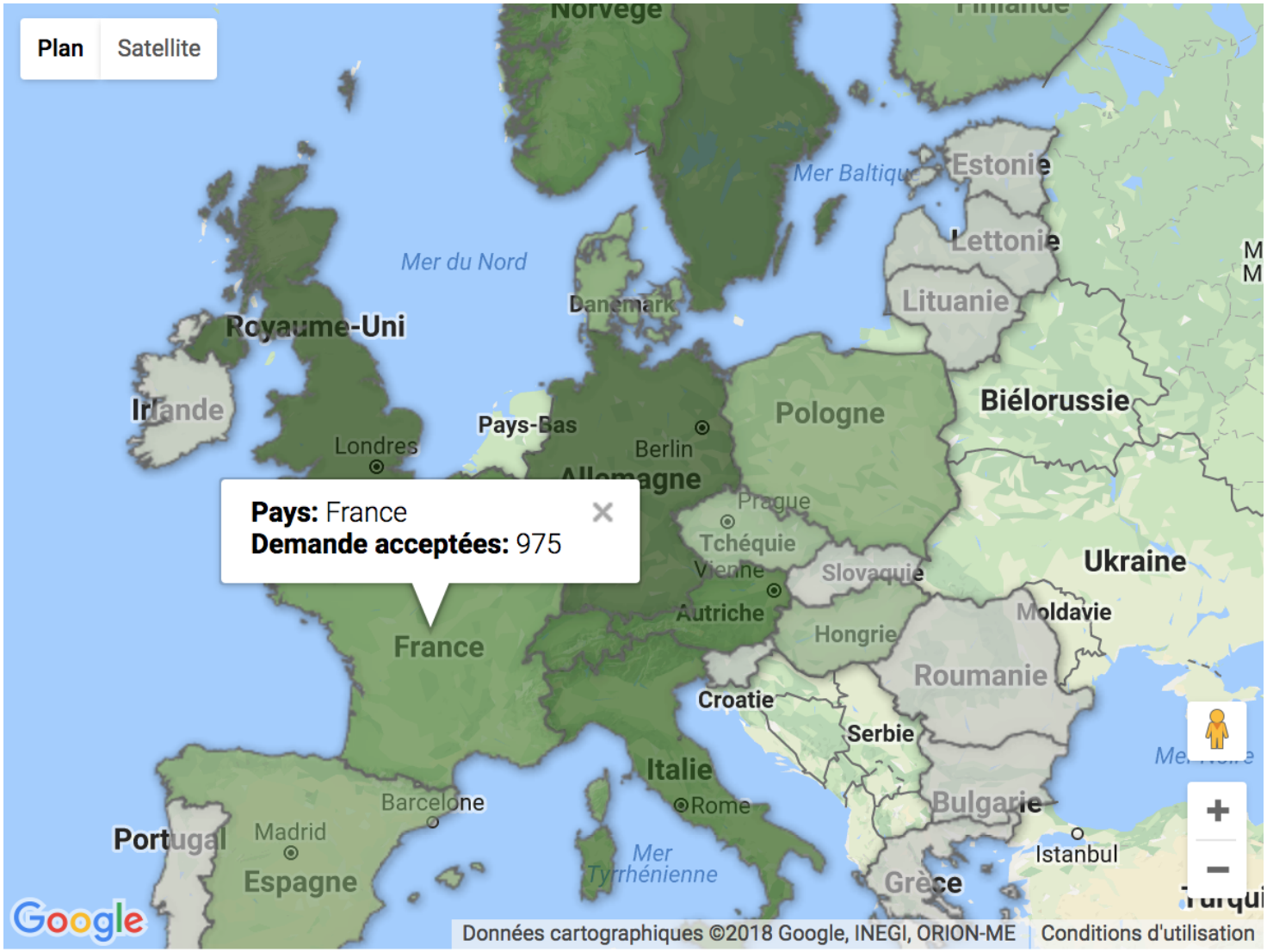
$$\begin{cases} 200 \to 180 \\ 2000 \to 1800 \end{cases} \Rightarrow \text{baisse de } 10\%$$

$$\begin{cases} \dfrac{200 \times 1000 + 2000 \times 100}{1100} = 363,64 \\ \dfrac{180 \times 600 + 1800 \times 500}{1100} = 916,34 \end{cases} \Rightarrow \text{augmentation de } 152\%$$

évolution du salaire moyen $\neq$ évolution moyenne du salaire

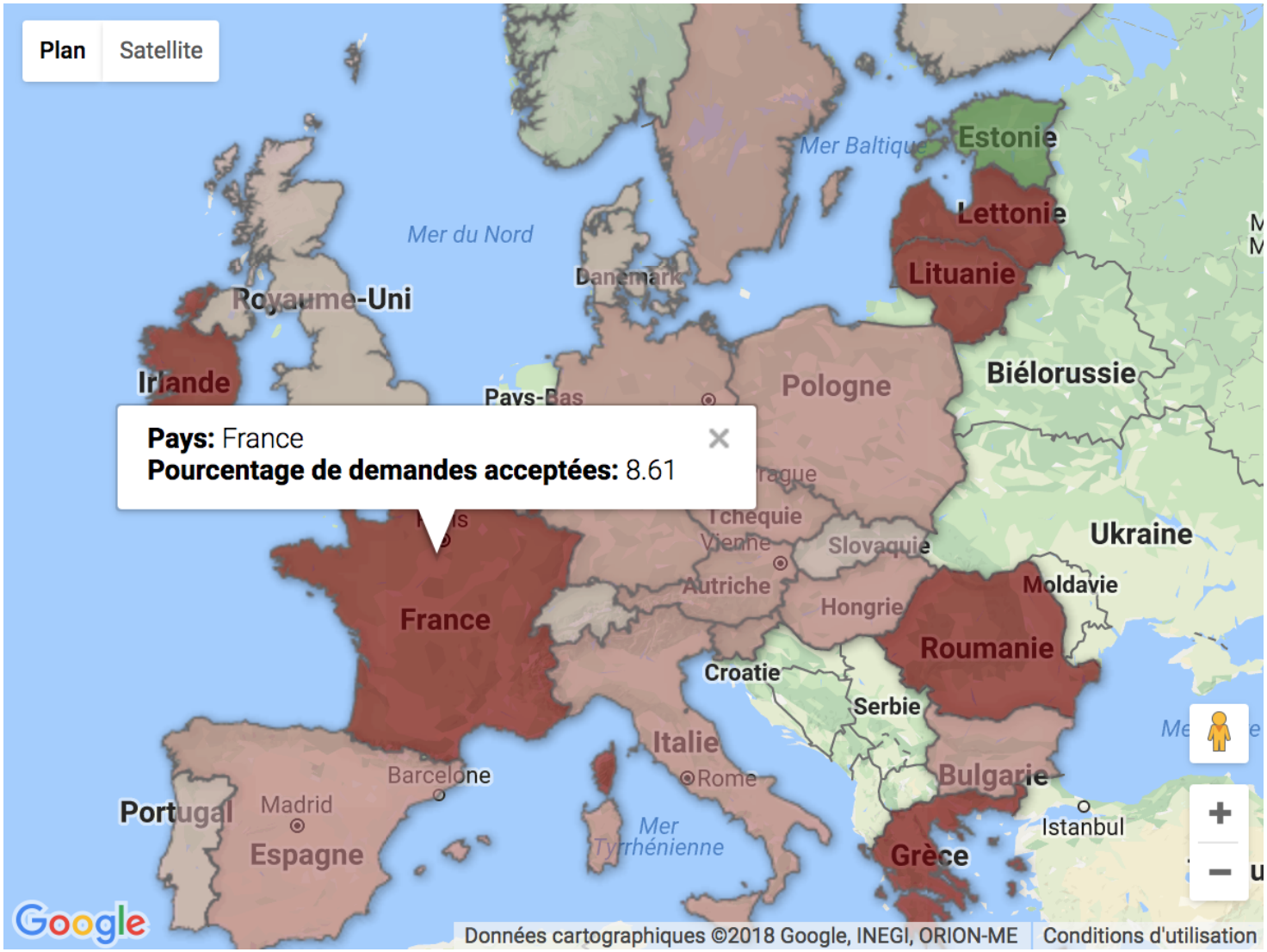
Méfiez-vous des statistiques !

Variations relatives et absolues : comparer des carottes à des potirons



Méfiez-vous des statistiques !

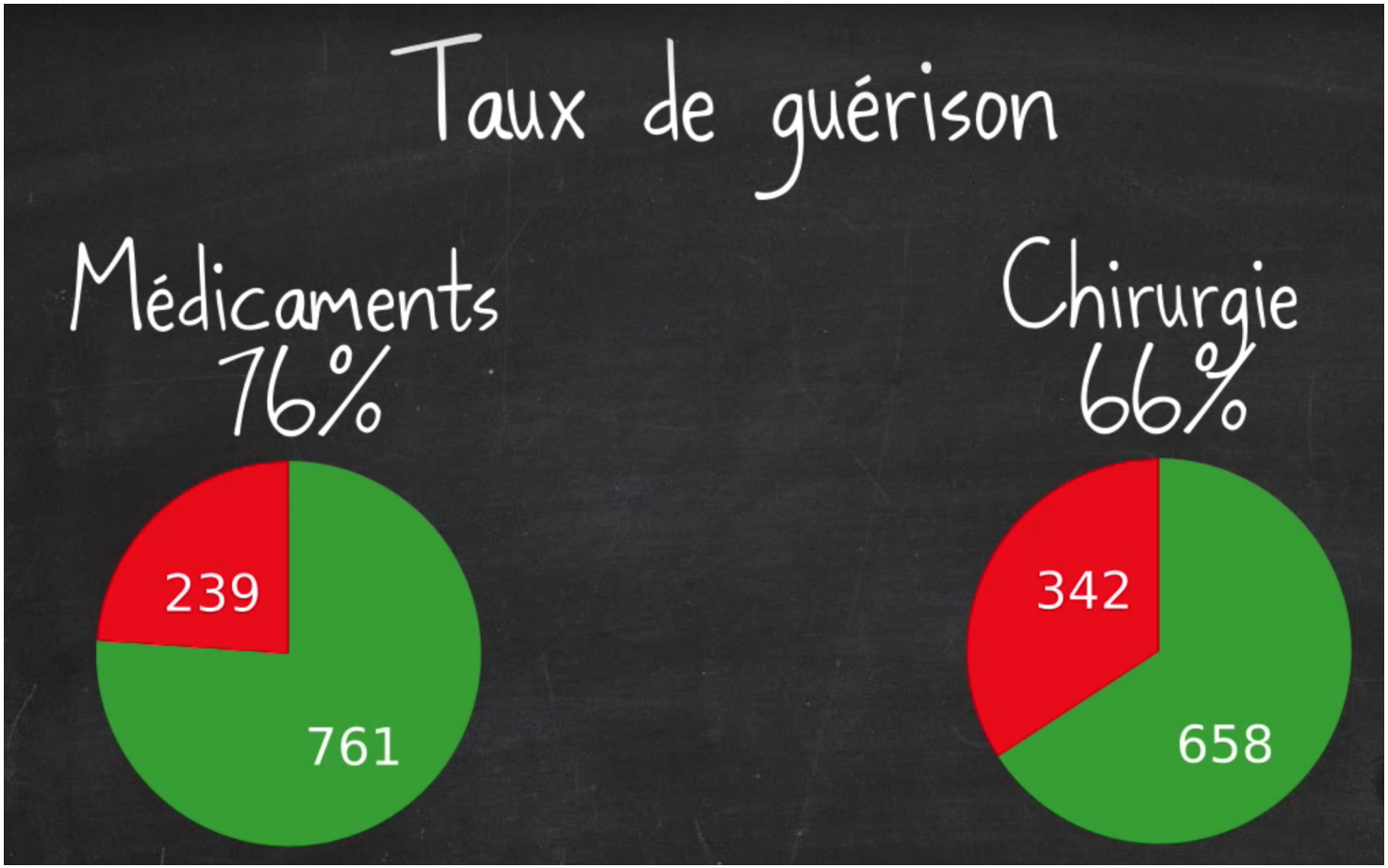
Variations relatives et absolues : comparer des carottes à des potirons



Paradoxe de Simpson

Taux de guérison moyen d'une tumeur : médicaments vs. chirurgie



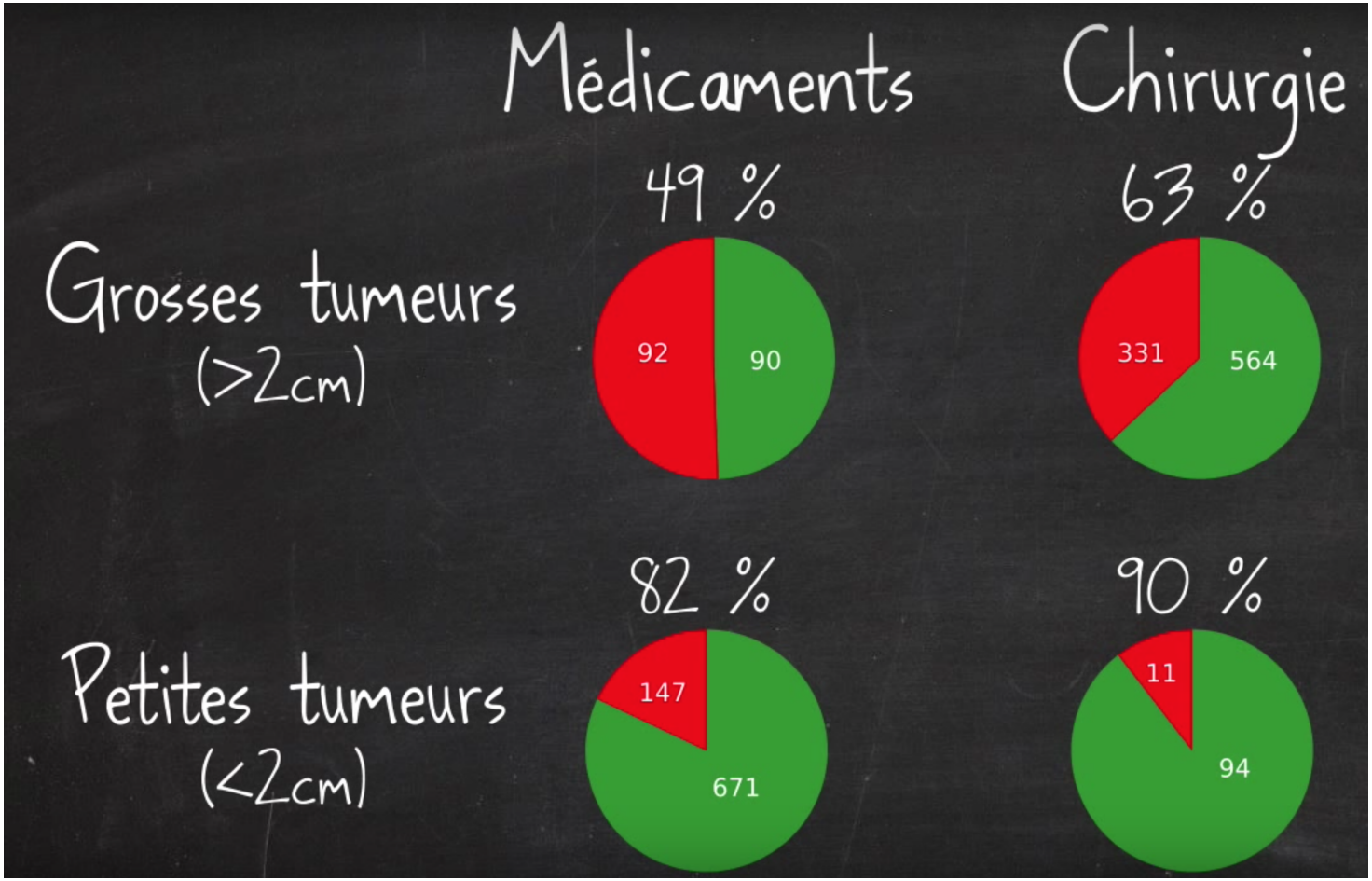


Crédits : D. Louapre (ScienceEtonnante)

Paradoxe de Simpson

Taux de guérison de la méthode suivant la taille de la tumeur



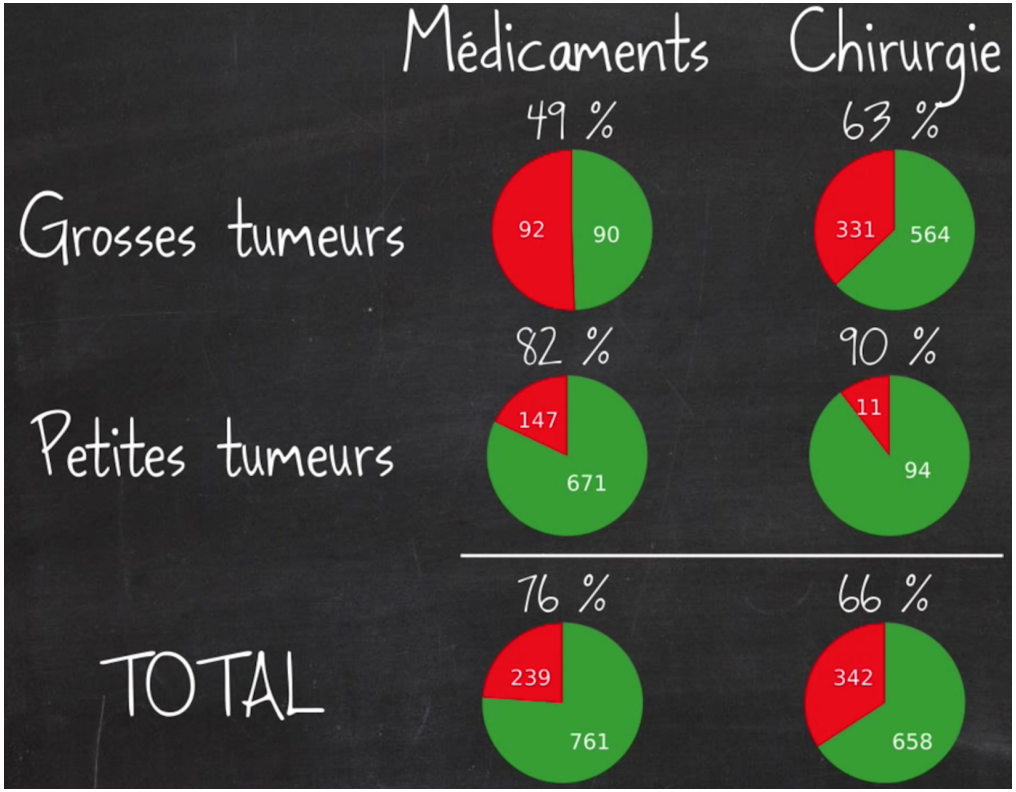


Crédits : D. Louaprre

Paradoxe de Simpson

À vos boitiers de vote !





Selon vous, quel traitement marche le mieux ?

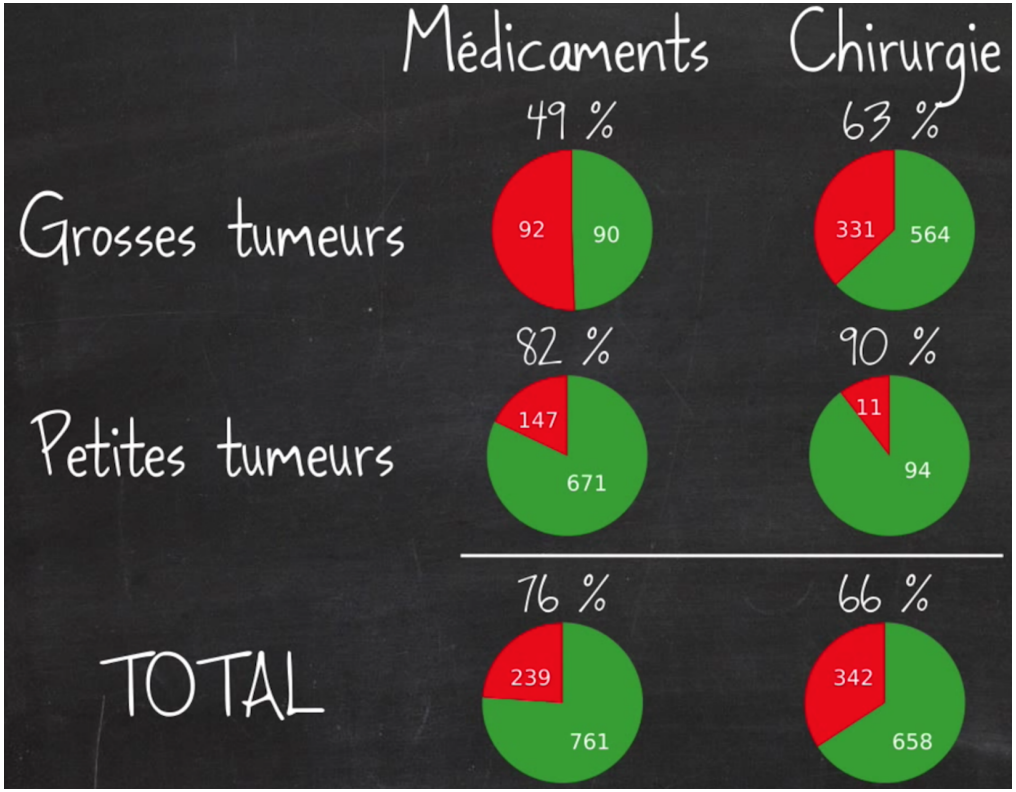
A) Médicaments
B) Chirurgie

Crédits : D. Louaprre

Paradoxe de Simpson

Résultats





Selon vous, quel traitement marche le mieux ?

A) Médicaments

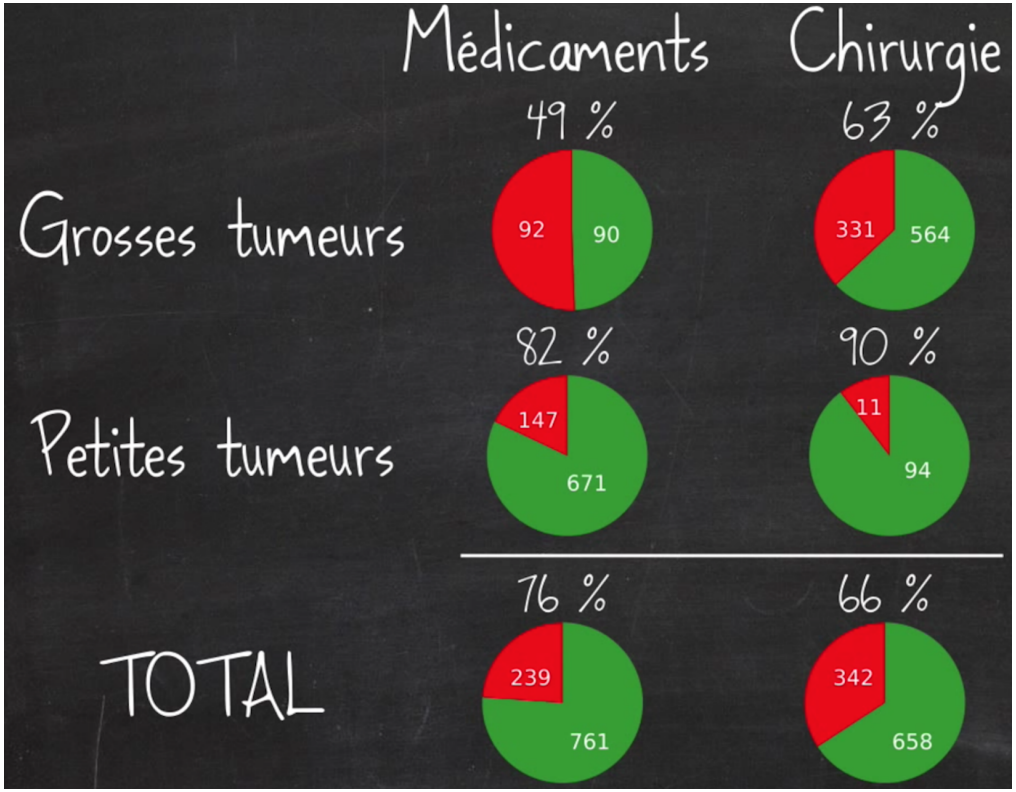
B) **Chirurgie**

Crédits : D. Louaprre

Paradoxe de Simpson

Kesako ?





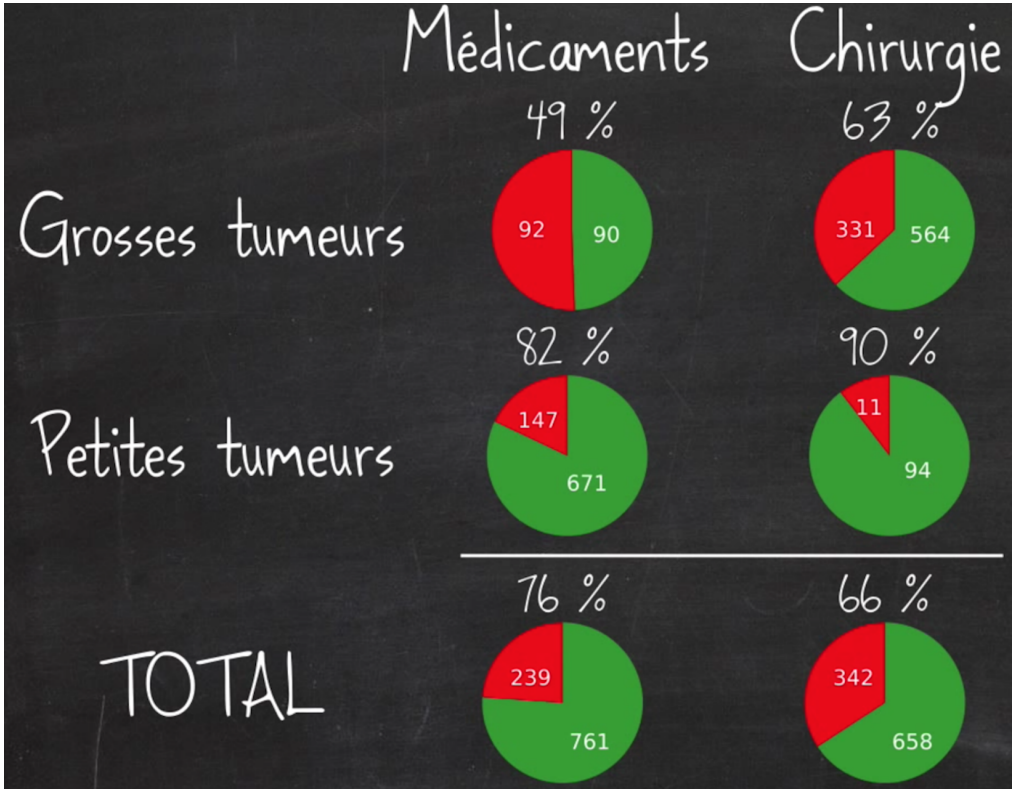
Deux observations importantes sur les grosses tumeurs :

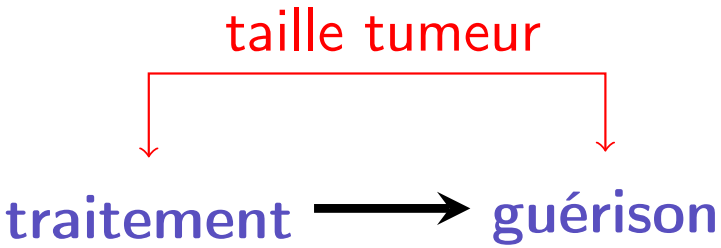
1. Elles ont des taux de guérison plus faible que les petites tumeurs
2. Elles donnent plus souvent lieu à une intervention chirurgicale

Paradoxe de Simpson

Gare aux facteurs de confusions !







Paradoxe de Simpson

En résumé



Pour que le paradoxe se produise, il faut 2 ingrédients :

- **Une variable qui influe sur le résultat final** (le groupe), et qui n'est pas forcément explicitée au départ. On appelle cela un **facteur de confusion**. Il s'agit de la taille des tumeurs dans cet exemple.
- **Une distribution non homogène** de l'échantillon. Dans cet exemple la chirurgie est plus souvent adoptée sur les grosses tumeurs, et les médicaments sur les petites.

Crédits : D. Louaprre

1 Introduction

2 Bases de la statistique descriptive
- Vocabulaire
- Tableaux statistiques
- Méfiez-vous des statistiques ! Le paradoxe de Simpson

3 Représentations graphiques
- Histogrammes
- Fonction de répartition empirique

4 Indicateurs statistiques
- Indicateurs de localisation ou de tendance centrale
- Indicateurs de dispersion ou de variabilité

5 Corrélation et causalité
- Régression linéaire
- Exemples de corrélations

Représentations graphiques

Aperçu des méthodes abordées

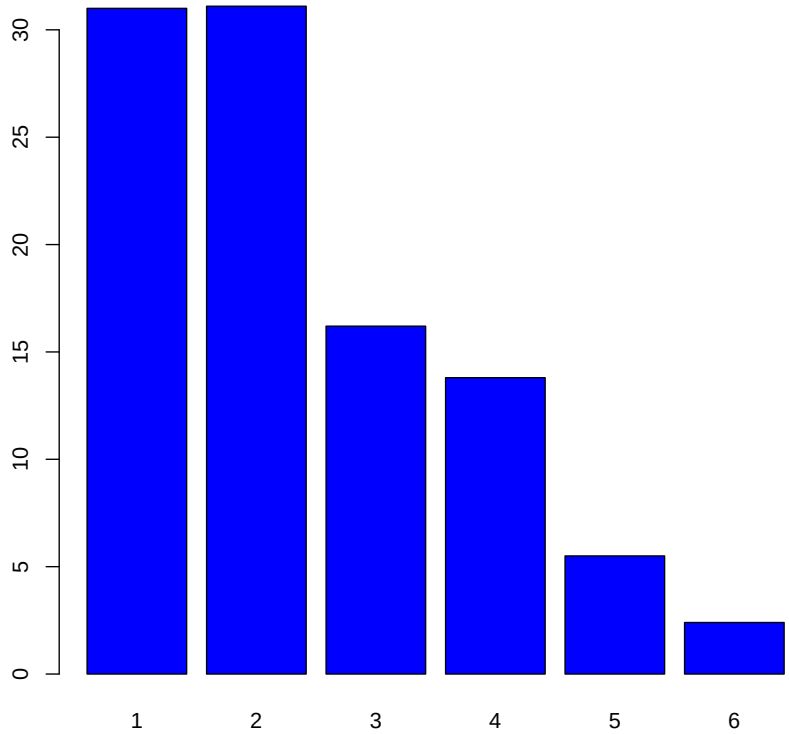
Différents type de représentations graphiques :

- Diagramme en bâtons et en camembert
- Histogramme des fréquences
- Graphique des fréquences cumulées (= fonction de répartition)
- Boite à moustache
- ...

Représentations graphiques

Variables discrètes – Diagrammes en bâtons

modalité m_i nb pers.	fréquence f_i (en %)
1	31.0
2	31.1
3	16.2
4	13.8
5	5.5
6 et plus	2.4

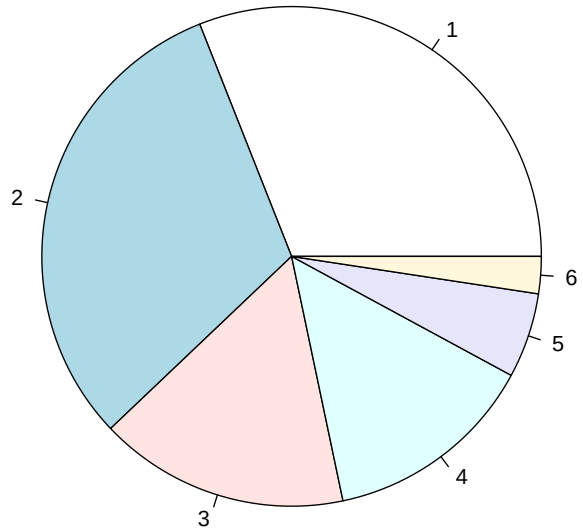


Diagrammes en bâtons : à chaque modalité correspond un rectangle vertical dont la hauteur est proportionnelle à la fréquence relative de la modalité.

Représentations graphiques

Variables discrètes – Diagrammes sectoriels (ou en camemberts)

modalité m_i nb pers.	fréquence f_i (en %)
1	31.0
2	31.1
3	16.2
4	13.8
5	5.5
6 et plus	2.4



Diagrammes sectoriels (ou en camemberts) : à chaque modalité correspond un secteur de disque dont l'aire est proportionnelle à la fréquence relative de la modalité.

Histogramme

Variable continue

Quand la variable étudiée est **continue**, les représentations du type **diagramme en bâtons sont sans intérêt**, car les données de x sont en général toutes distinctes, donc les effectifs tous égaux à 1.

$\Rightarrow$ La représentation par **histogramme** consiste à **regrouper les observations « proches » en classes :**

On trie le vecteur x (noté alors x^*), et on partitionne l'intervalle $]a_0, a_k]$ $(a_0 < x_1^*, a_k > x_n^*)$ en k intervalles $]a_{i-1}, a_i]$ appelés **classes**. La largeur de la classe i est notée $h_i = a_i - a_{i-1}$ (et $h = (a_k - a_0)/k$ si pas fixe).

Définition de l'histogramme

L'histogramme est la figure constituée de rectangles dont les bases sont les classes et dont les aires sont égales aux fréquences de ces classes. Autrement dit, la hauteur du $i^{\text{ème}}$ rectangle est n_i/nh_i.

Histogramme

Variable continue : durée de vie d'ampoules

$x = 91.6, 35.7, 251.3, 24.3, 5.4, 67.3, 170.9, 9.5, 118.4, 57.1$

$x^* = 5.4, 9.5, 24.3, 35.7, 57.1, 67.3, 91.6, 118.4, 170.9, 251.3$

- Choix du nombre de classes k : $k \approx 1 + \log_2 n$ (règle de Sturges)
- Choix des bornes pour a_0 et a_k : $x_1^* \pm 0.025(x_n^* - x_1^*)$
- Largeur des classes (fixe) $h = (a_k - a_0)/k$

$n = 10$, $k = 5$, $a_0 = -0.74 \approx 0$ et $a_k = 257.4 \approx 260$, $h = 260/5 = 52$.

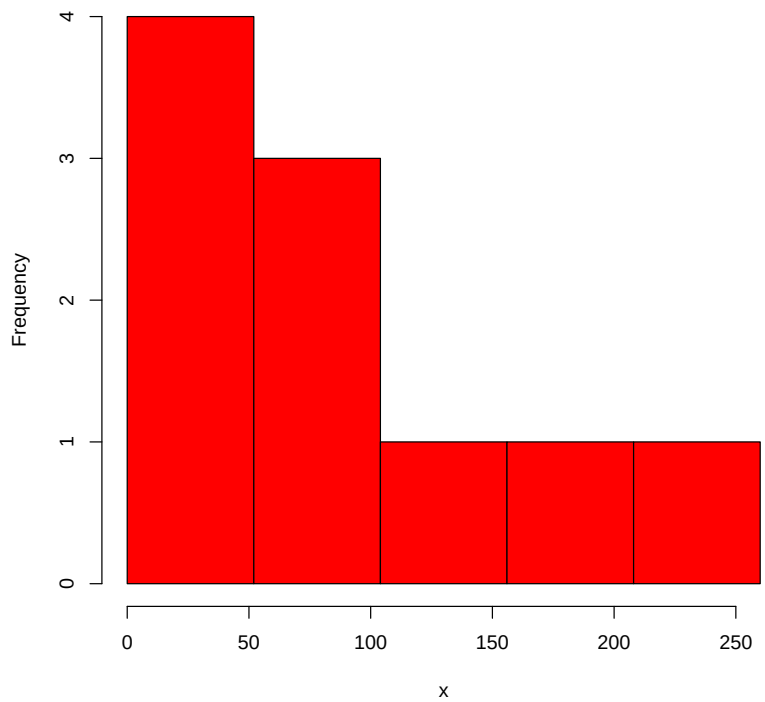
classes $]a_{i-1}, a_i]$	$]0, 52]$	$]52, 104]$	$]104, 156]$	$]156, 208]$	$]208, 260]$
effectifs n_i	4	3	1	1	1
fréquences n_i/n	40%	30%	10%	10%	10%
hauteurs n_i/nh	0.0077	0.0058	0.0019	0.0019	0.0019

Crédits : O. Gaudoin

Histogramme

Variable continue : durée de vie d'ampoules

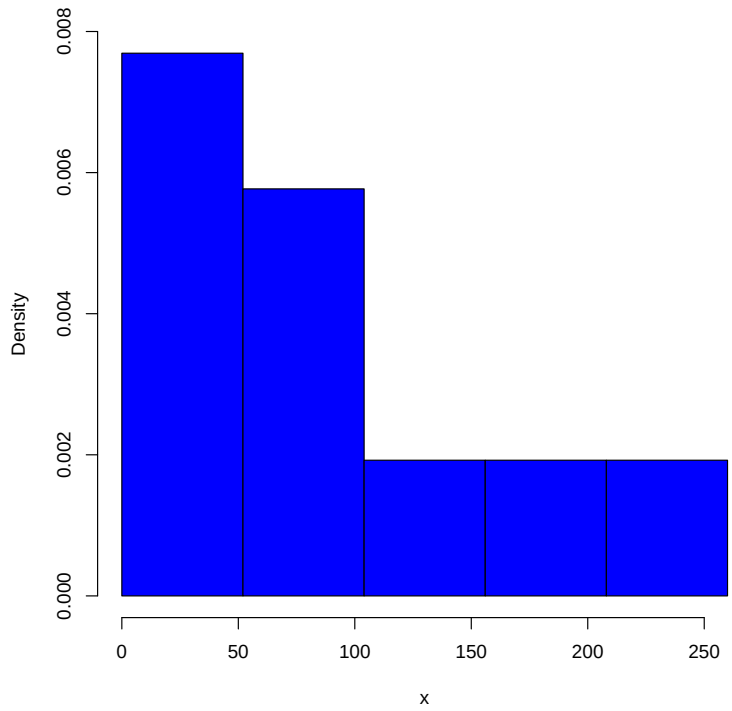
classes $]a_{i-1}, a_i]$	$]0, 52]$	$]52, 104]$	$]104, 156]$	$]156, 208]$	$]208, 260]$
effectifs n_i	4	3	1	1	1
fréquences n_i/n	40%	30%	10%	10%	10%
hauteurs n_i/nh	0.0077	0.0058	0.0019	0.0019	0.0019



Histogramme

Variable continue : durée de vie d'ampoules

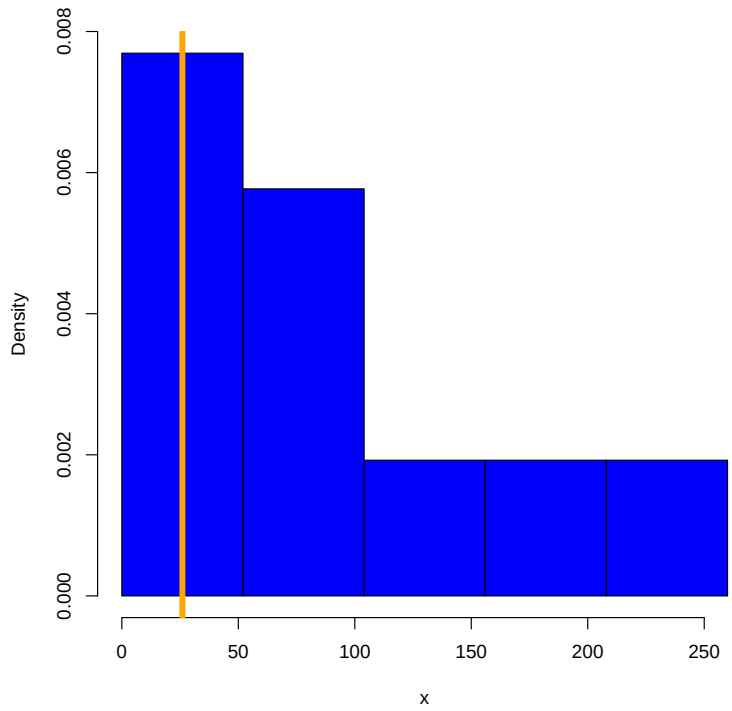
classes $]a_{i-1}, a_i]$	$]0, 52]$	$]52, 104]$	$]104, 156]$	$]156, 208]$	$]208, 260]$
effectifs n_i	4	3	1	1	1
fréquences n_i/n	40%	30%	10%	10%	10%
hauteurs n_i/nh	0.0077	0.0058	0.0019	0.0019	0.0019



Histogramme

Mode de l'histogramme

classes $]a_{i-1}, a_i]$	$]0, 52]$	$]52, 104]$	$]104, 156]$	$]156, 208]$	$]208, 260]$
effectifs n_i	4	3	1	1	1
fréquences n_i/n	40%	30%	10%	10%	10%
hauteurs n_i/nh	0.0077	0.0058	0.0019	0.0019	0.0019



Histogramme

Approximation de la densité

classes $]a_{i-1}, a_i]$	$]0, 52]$	$]52, 104]$	$]104, 156]$	$]156, 208]$	$]208, 260]$
effectifs n_i	4	3	1	1	1
fréquences n_i/n	40%	30%	10%	10%	10%
hauteurs n_i/nh	0.0077	0.0058	0.0019	0.0019	0.0019

fonction escalier : $\hat{f}_{|]a_{i-1},a_i]} = n_i/nh$

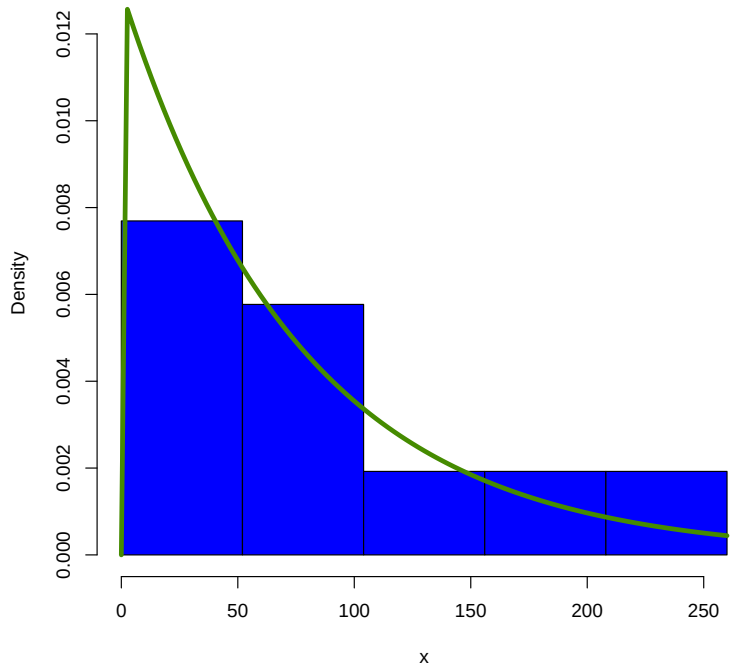
aire rect. $i = n_i/n = \int_{a_{i-1}}^{a_i} \hat{f}(x)\mathrm{d}x$

$n_i/n =$ % obs. dans $]a_{i-1}, a_i]$

$\Longleftrightarrow$

proba qu'une obs. soit dans $]a_{i-1}, a_i]$

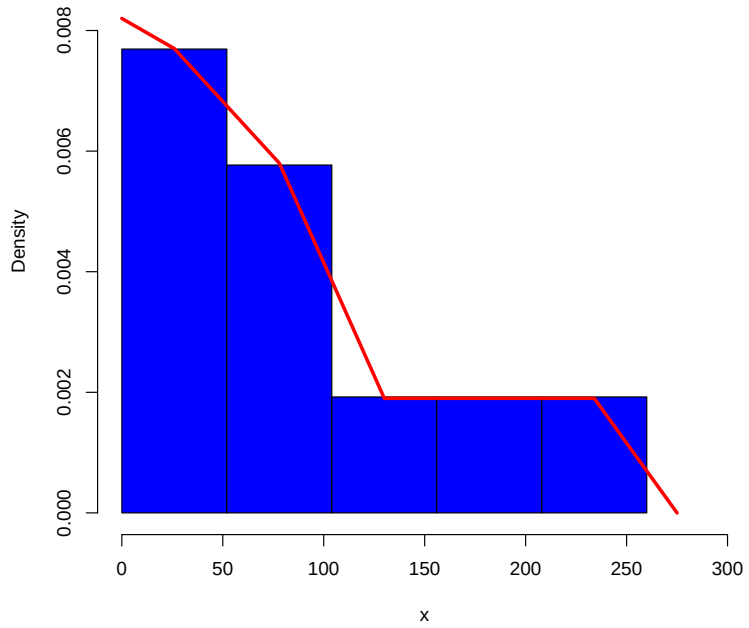
$\mathbb{P}(a_{i-1} \leq X \leq a_i) = \int_{a_{i-1}}^{a_i} f(x)\mathrm{d}x$



Histogramme

Polygone des fréquences

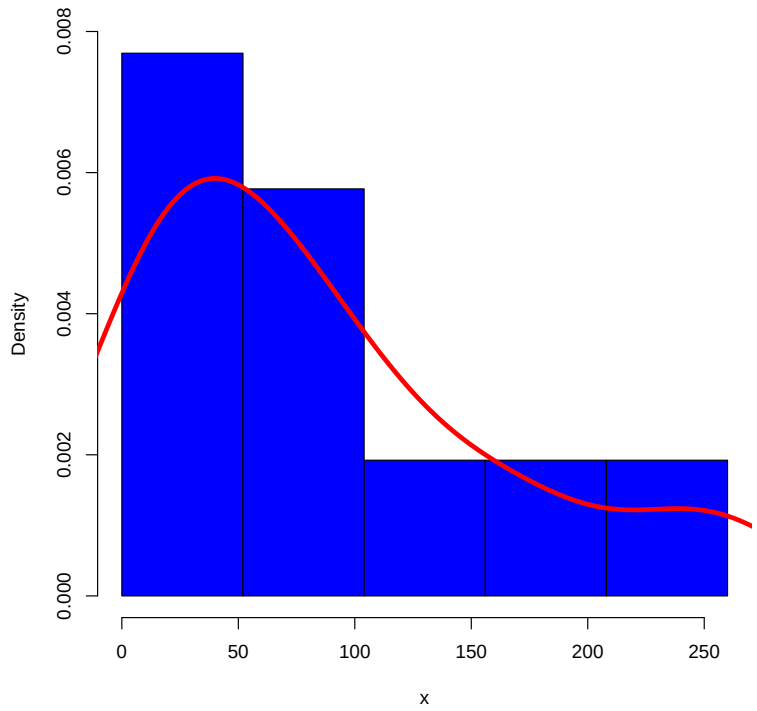
classes $]a_{i-1}, a_i]$	$]0, 52]$	$]52, 104]$	$]104, 156]$	$]156, 208]$	$]208, 260]$
effectifs n_i	4	3	1	1	1
fréquences n_i/n	40%	30%	10%	10%	10%
hauteurs n_i/nh	0.0077	0.0058	0.0019	0.0019	0.0019



Histogramme

Densité (continue) approchant l'histogramme : R density.

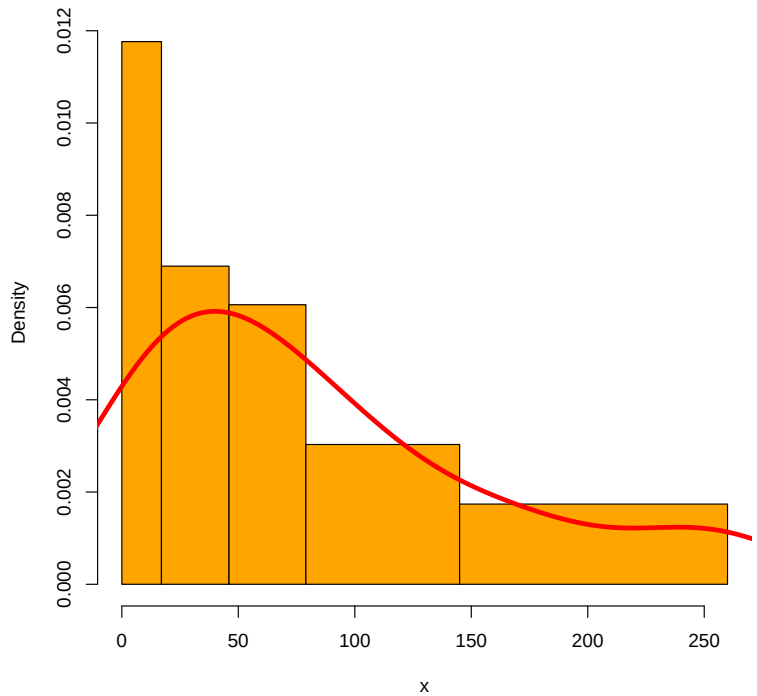
classes $]a_{i-1}, a_i]$	$]0, 52]$	$]52, 104]$	$]104, 156]$	$]156, 208]$	$]208, 260]$
effectifs n_i	4	3	1	1	1
fréquences n_i/n	40%	30%	10%	10%	10%
hauteurs n_i/nh	0.0077	0.0058	0.0019	0.0019	0.0019



Histogramme

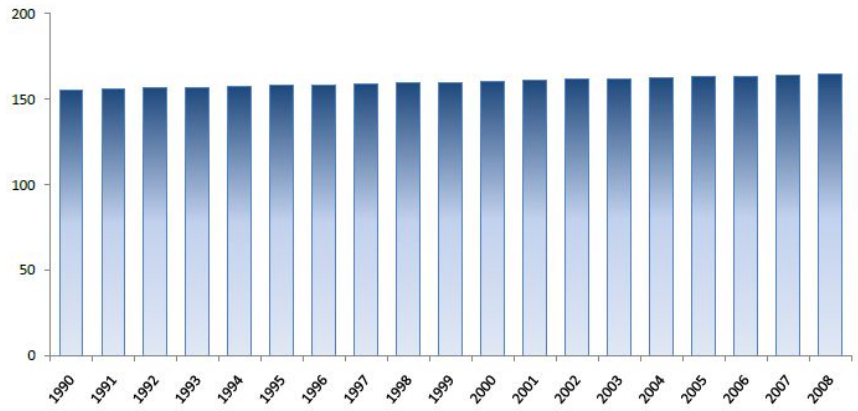
Classes de même effectif

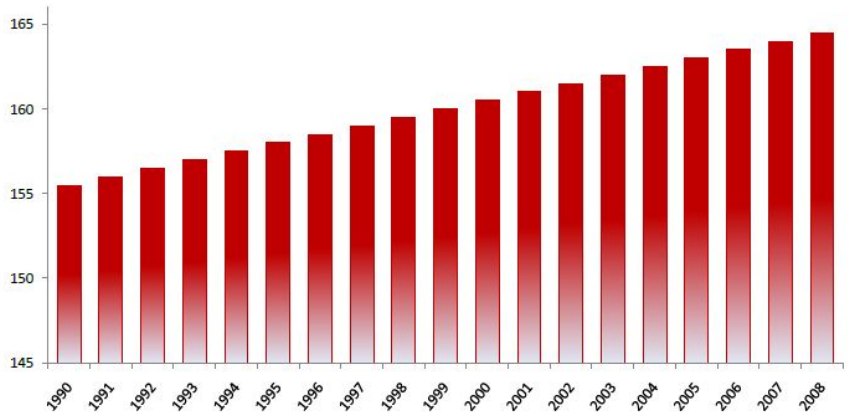
classes $]a_{i-1}, a_i]$	$]0, 17]$	$]17, 46]$	$]46, 79]$	$]79, 145]$	$]145, 260]$
effectifs n_i	2	2	2	2	2
fréquences n_i/n	20%	20%	20%	20%	20%
hauteurs n_i/nh	0.0118	0.0069	0.0061	0.0030	0.0017



Ne passez pas sous les échelles

Graphique à la loupe

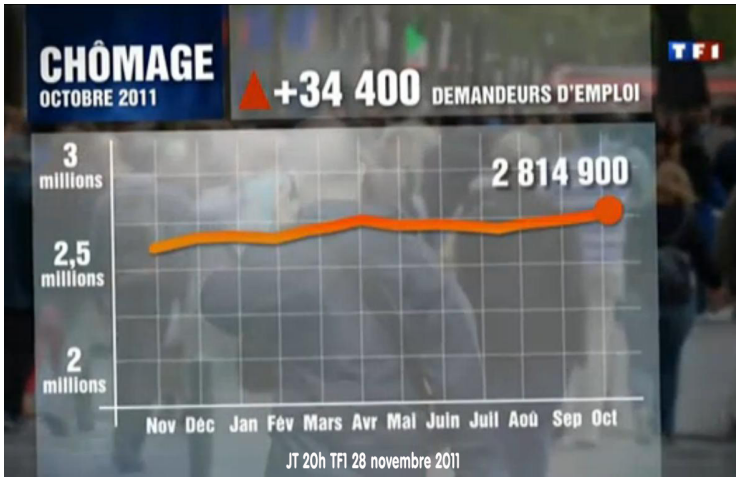


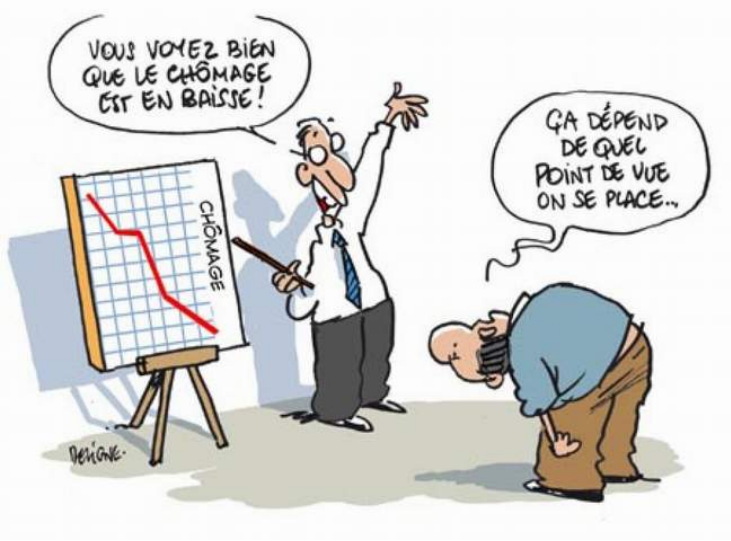


Crédits : Cortecs (zététique)

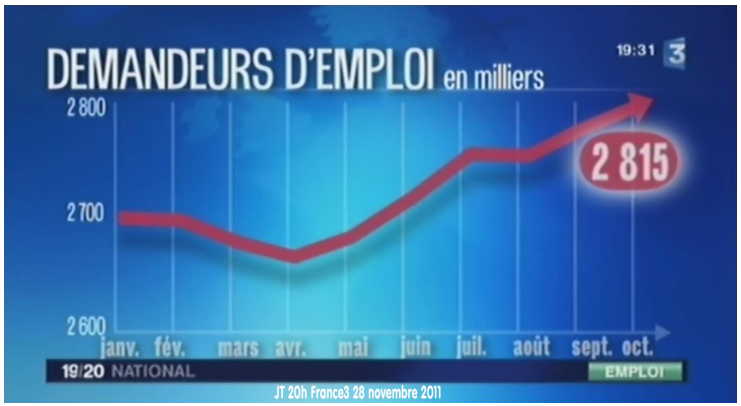
Ne passez pas sous les échelles

Quand les journaux télévisés présentent leurs courbes du chômage...







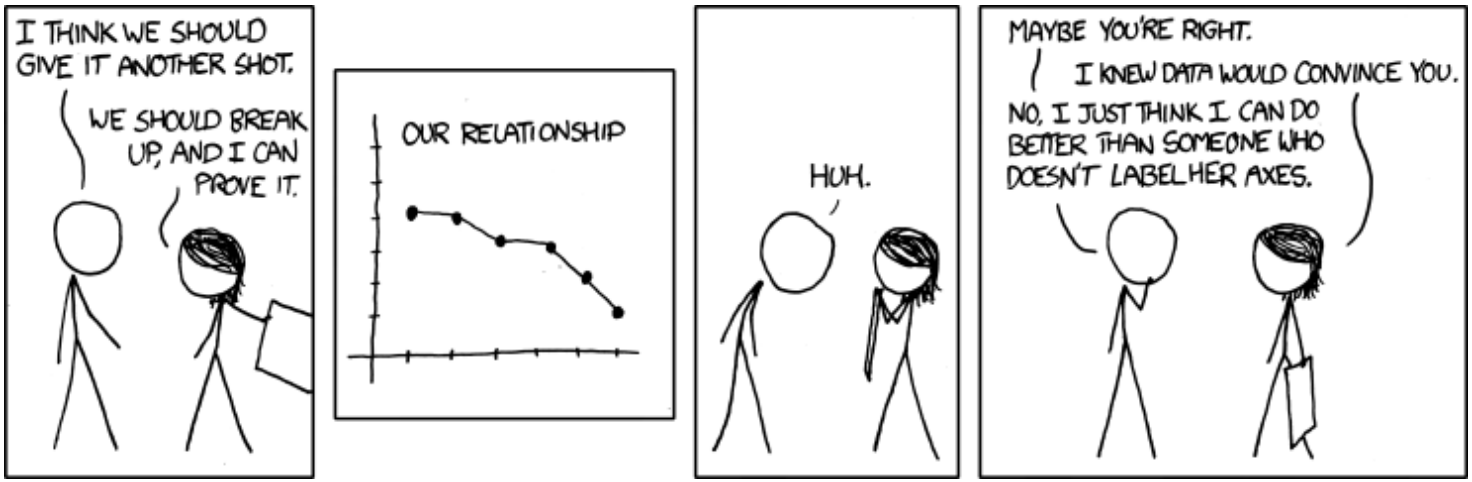


Crédits : Cortecs (zététique)

Ne passez pas sous les échelles

Surtout quand il n'y a pas d'échelle !



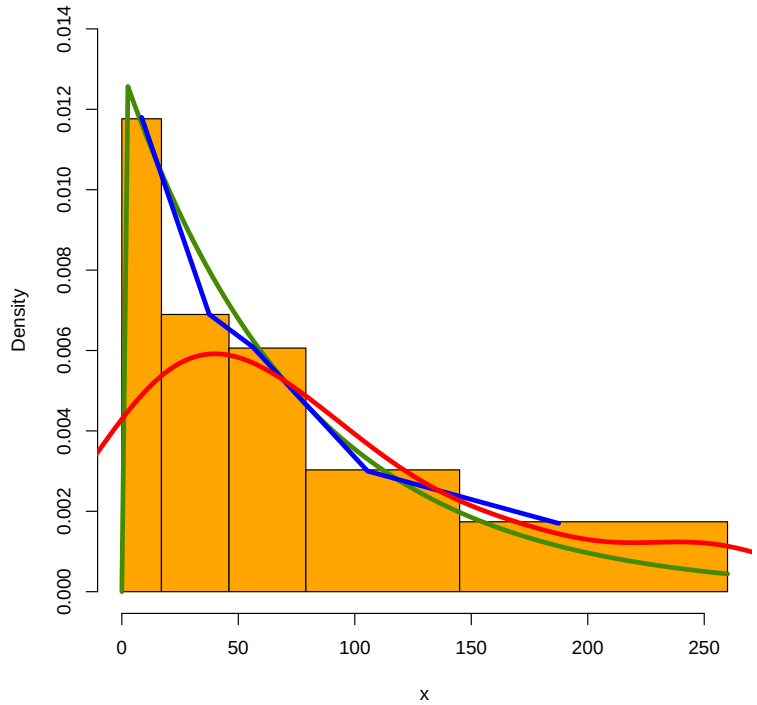


Crédits : Cortecs (zététique) + xkcd – https://xkcd.com/833/

Histogramme et densité

Approximation de la densité

- L'histogramme de même effectif approche une densité exponentielle $f(x) = \lambda \mathrm{e}^{-\lambda x}$ (en vert)
- Il en va de même pour son polygone des fréquences (en bleu)
- L'approximation continue de la densité (en rouge) n'est pas efficace sur peu de données



Histogramme cumulé et fonction de répartition

Approximation de la fonction de répartition

- Au lieu des effectifs n_i considérer les effectifs cumulés $m_i = \sum_{l=1}^{i} n_l$
- L'histogramme cumulé construit approche la fonction de répartition

$$F(x) = 1 - \mathrm{e}^{-\lambda x}$$

- Il en va de même pour son polygone des fréquences cumulées

Histogramme cumulé

Quand Tim Cuisine ses graphiques...

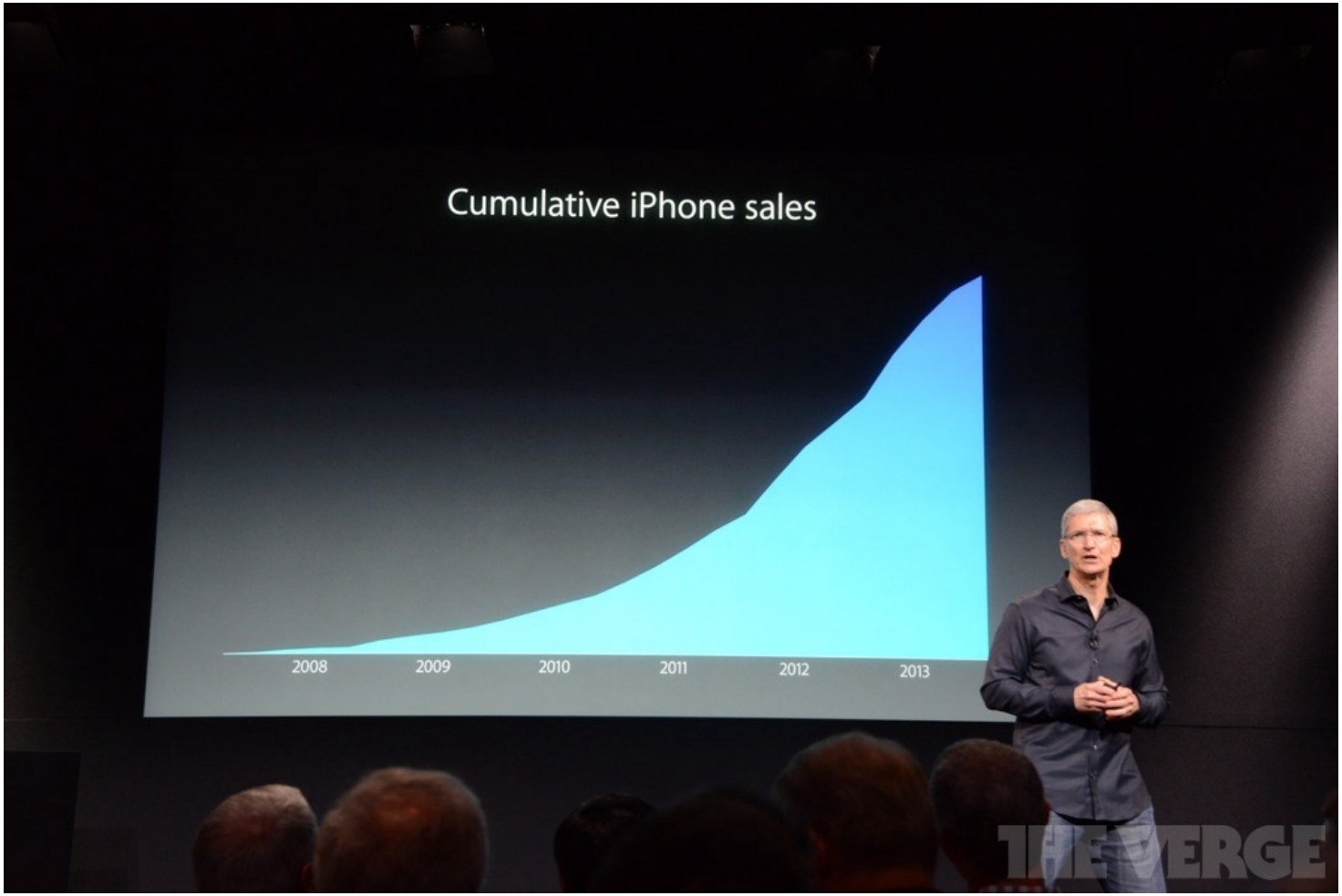


Figure: Tim Cook présente les ventes d'iPhone cumulées

Histogramme cumulé

Quand Tim Cuisine ses graphiques...

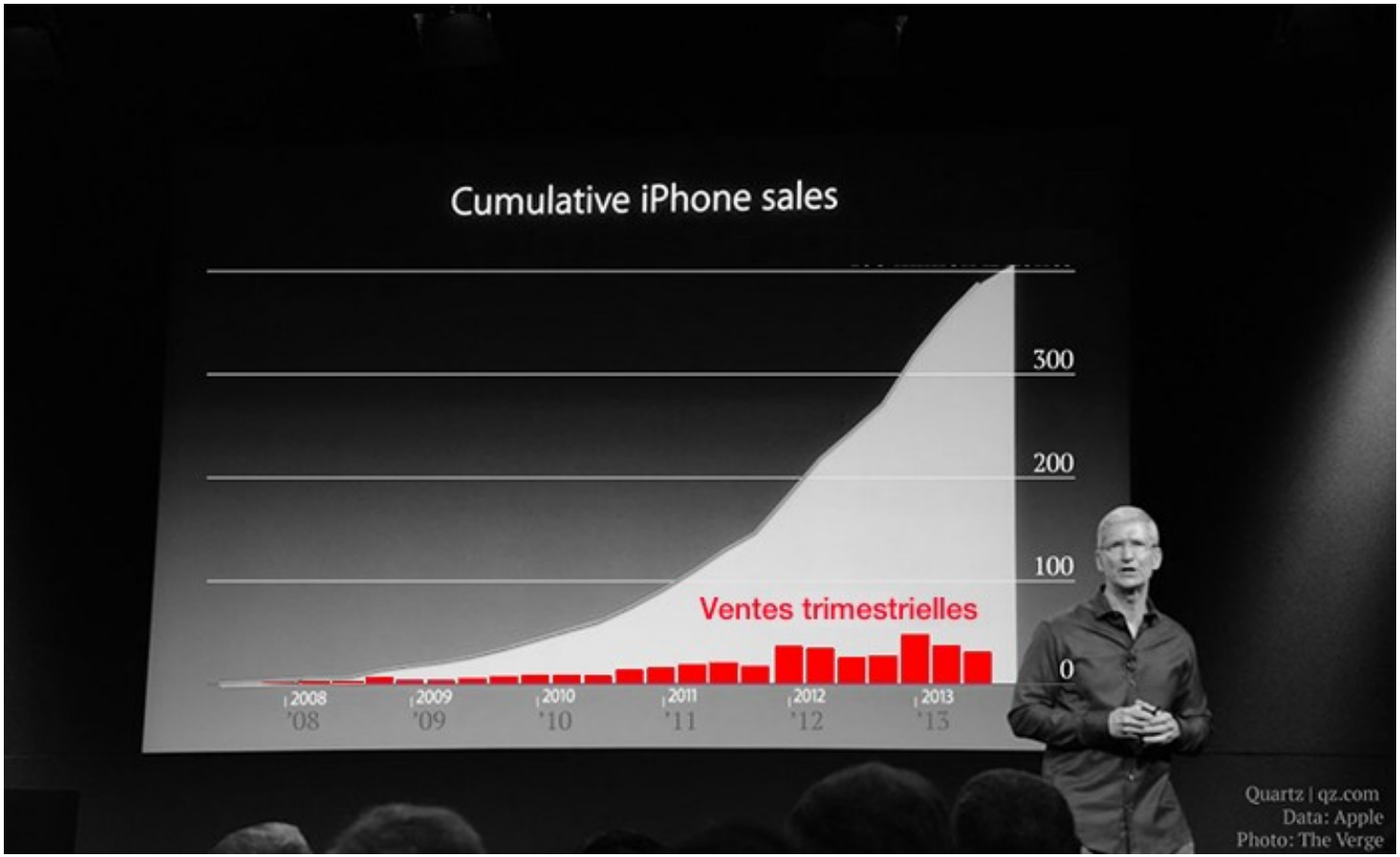


Figure: Si Tim Cook présentait les ventes d'iPhone **trimestrielles**

Fonction de répartition empirique

La fonction de répartition empirique F_n associée à un échantillon $x_1, \dots x_n$ est la fonction définie par :

$$\forall x \in \mathbb{R}, \quad F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{x_i \leq x\}} = \begin{cases} 0 \text{ si } x < x_1^* \\ \frac{i}{n} \text{ si } x_i^* \leq x \leq x_{i+1}^* \\ 1 \text{ si } x > x_n^* \end{cases}$$

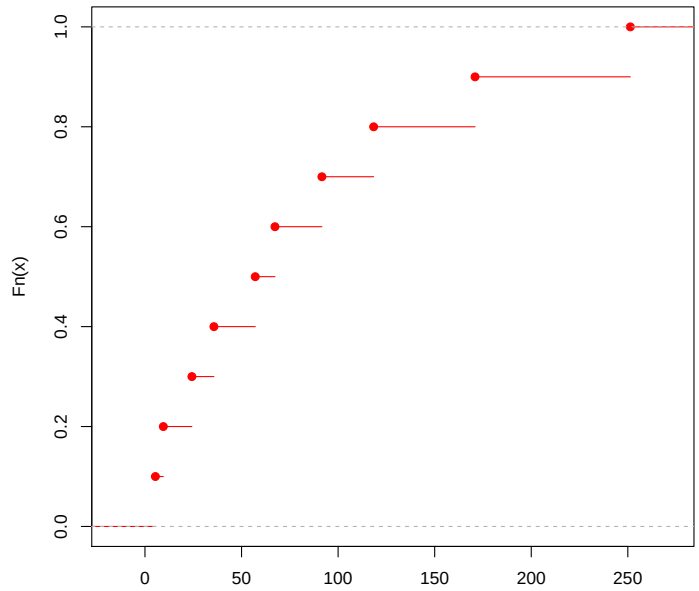


Figure: Fonction de répartition empirique de x (ampoules) approchant F

Crédits : O. Gaudoin

Fonction de répartition candidate ?

Proposition

Soit F la fonction de répartition d'une loi de probabilité, dépendant d'un paramètre inconnu θ. S'il existe des fonctions h, g, α et β telles que

$$\forall x \in \mathbb{R}, \quad h[F(x)] = \alpha(\theta)g(x) + \beta(\theta)$$

alors le nuage des points

$$(g(x_i^*), h(i/n)), \quad i \in \{1, \ldots, n\}$$

est le **graphe de probabilités** pour la loi de fonction de répartition F. Si les points du nuage sont approximativement alignés, on admettra que F est une fonction de répartition plausible pour les observations.

Preuve : $h[F_n(x_i^*)] = h(i/n) \approx h[F(x_i^*)] = \alpha(\theta)g(x_i^*) + \beta(\theta)$

Crédits : O. Gaudoin

Test sur la durée de vie des ampoules

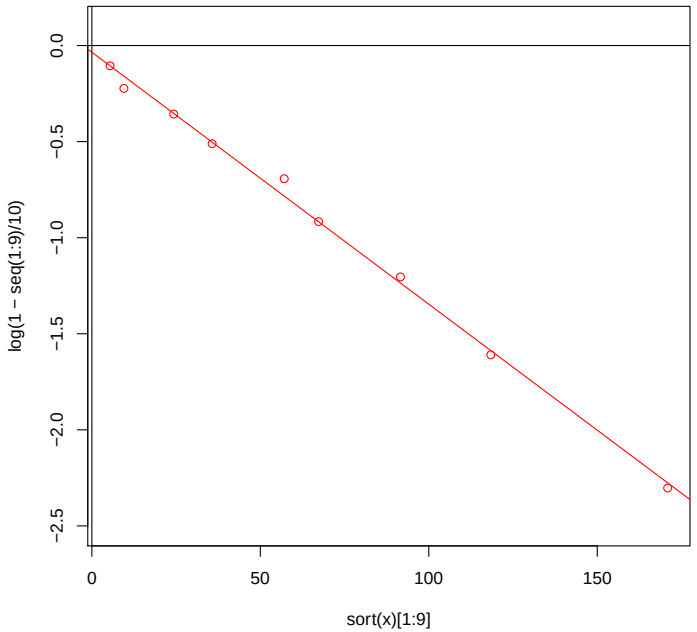
- On suppose $F(x) = 1 - \mathrm{e}^{-\lambda x}$. En considérant $h(y) = \ln(1-y)$:

$$h[F(x)] = \ln(1 - F(x)) = -\lambda x$$

- Le **graphe de probabilité** pour F est le nuage de points

$$(x_i^*, \ln(1 - i/n)), \quad i \in \{1, \ldots, n-1\}$$

- La droite qui approche ce nuage de point est $y = -\lambda x$



Crédits : O. Gaudoin

1 Introduction

2 Bases de la statistique descriptive
- Vocabulaire
- Tableaux statistiques
- Méfiez-vous des statistiques ! Le paradoxe de Simpson

3 Représentations graphiques
- Histogrammes
- Fonction de répartition empirique

4 Indicateurs statistiques
- Indicateurs de localisation ou de tendance centrale
- Indicateurs de dispersion ou de variabilité

5 Corrélation et causalité
- Régression linéaire
- Exemples de corrélations

Indicateurs de localisation (ou de tendance centrale)

La moyenne empirique

Définir une valeur autour de laquelle se repartissent les observations

- **Moyenne empirique**

  Valeur qu'auraient tous les individus s'ils prennaient la même valeur

$$\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\sum_{i=1}^{k} n_i\, m_i$$

Durée de vie moyenne des ampoules

On trouve $\bar{x}_{10} = 83.15$ heures en moyenne.
De plus pour une loi exponentielle $\lambda \approx \frac{1}{\bar{x}_{10}} = 0.012$.

Indicateurs de localisation (ou de tendance centrale)

Valeurs extrêmes

- **Valeurs extrêmes**

  Un indicateur de localisation à partir de $x_1^* = \min x_i$ et $x_n^* = \max x_i$ est

  $$\frac{x_1^* + x_n^*}{2}$$

  Exemple des ampoules : on trouve 128.35 heures.

- **Mode**

  Valeur pour laquelle l'histogramme des fréquences presente un maximum. Modalité la plus représentée dans l'échantillon.

Valeurs aberrantes

Des valeurs exagéréments grandes ou petites par rapport aux autres valeurs de l'échantillon peuvent fortement influer sur **la moyenne qui est sensibles aux extrêmes**.

Indicateurs de localisation (ou de tendance centrale)

La moyenne salariale dans certains médias

Jobat.be
Trouvez le job de votre vie

Le Belge gagne en moyenne 3 200 euros bruts par mois

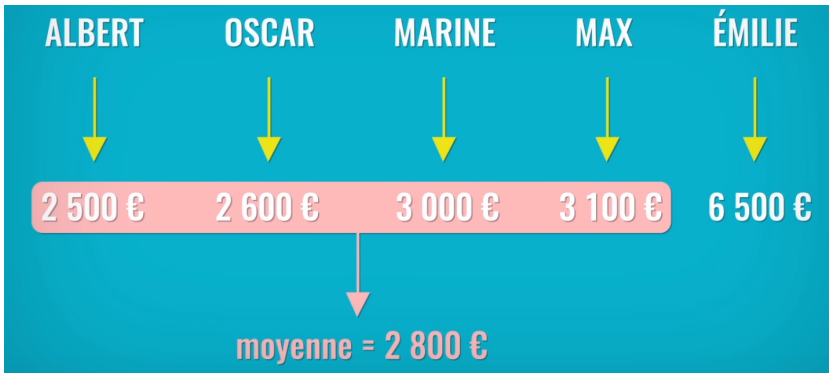
RTL INFO ECONOMIE LE REVENU MOYEN DES BELGES EST EN AUGMENTATION

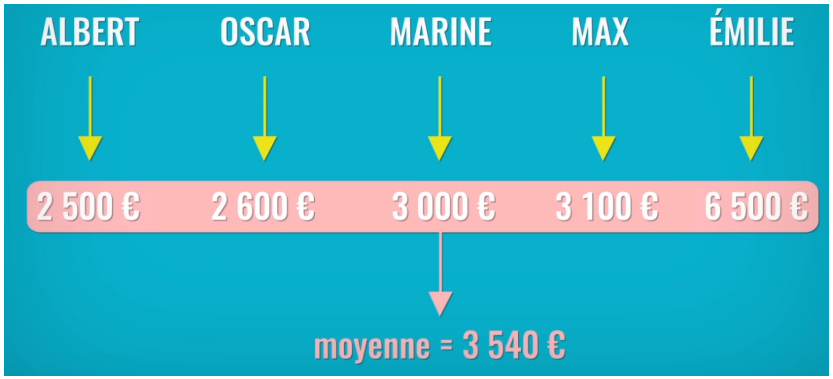
Le revenu moyen net annuel des Belges s'élevait à 17.684 euros en 2014, soit 665 euros de plus que l'année précédente, ressort-il vendredi des dernières données disponibles du SPF Economie. Par mois, cela fait dont un salaire de 1.473 euros.

La moyenne seule ne permet pas de résumer correctement la distribution des salaires à l'échelle d'un pays !

Indicateurs de localisation (ou de tendance centrale)

La moyenne salariale sensible aux valeurs extrêmes

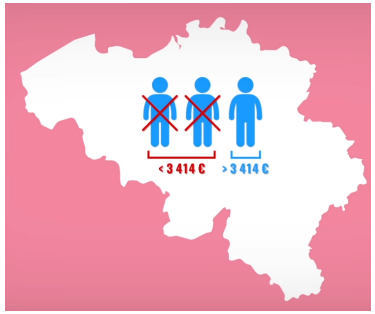






Crédits : Les statistiques expliquées à mon chat

La médiane



- **Médiane** : valeur **partageant la population en 2 effectifs égaux**.

$$\tilde{x}_n = \begin{cases} x^*_{(n+1)/2} \text{ si } n \text{ impair} \\ (x^*_{n/2} + x^*_{n/2+1})/2 \text{ si } n \text{ pair} \end{cases}$$

- **Graphiquement** peut se lire sur la courbe de F :
  - Variable continue

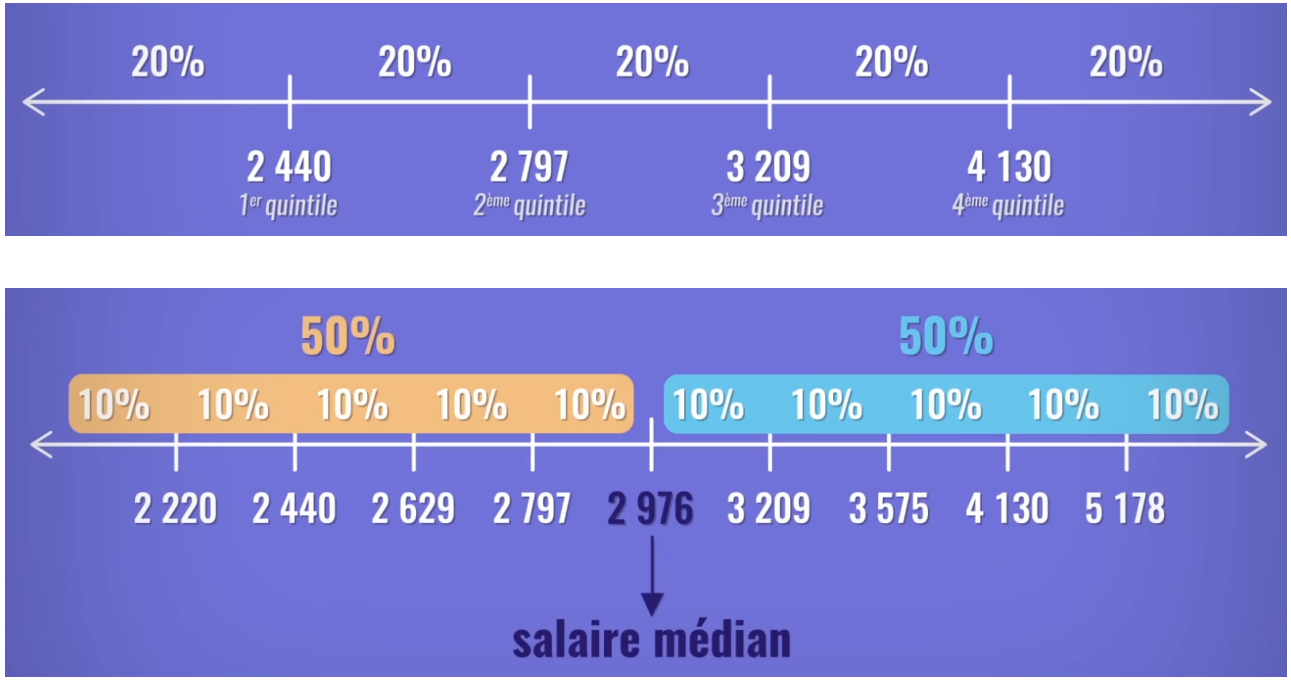
$$q_{0.5} : F(q_{0.5}) = 0.5$$

  - Variable discrète : plus petite valeur où F franchit le palier 50%

$$q_{0.5} : F(q^-_{0.5}) < 0.5, \quad F(q^+_{0.5}) \geq 0.5$$

Exemple de l'ampoule : $\tilde{x}_{10} = (57.1 + 67.3)/2 = 62.2$ heures.

Les quantiles empiriques

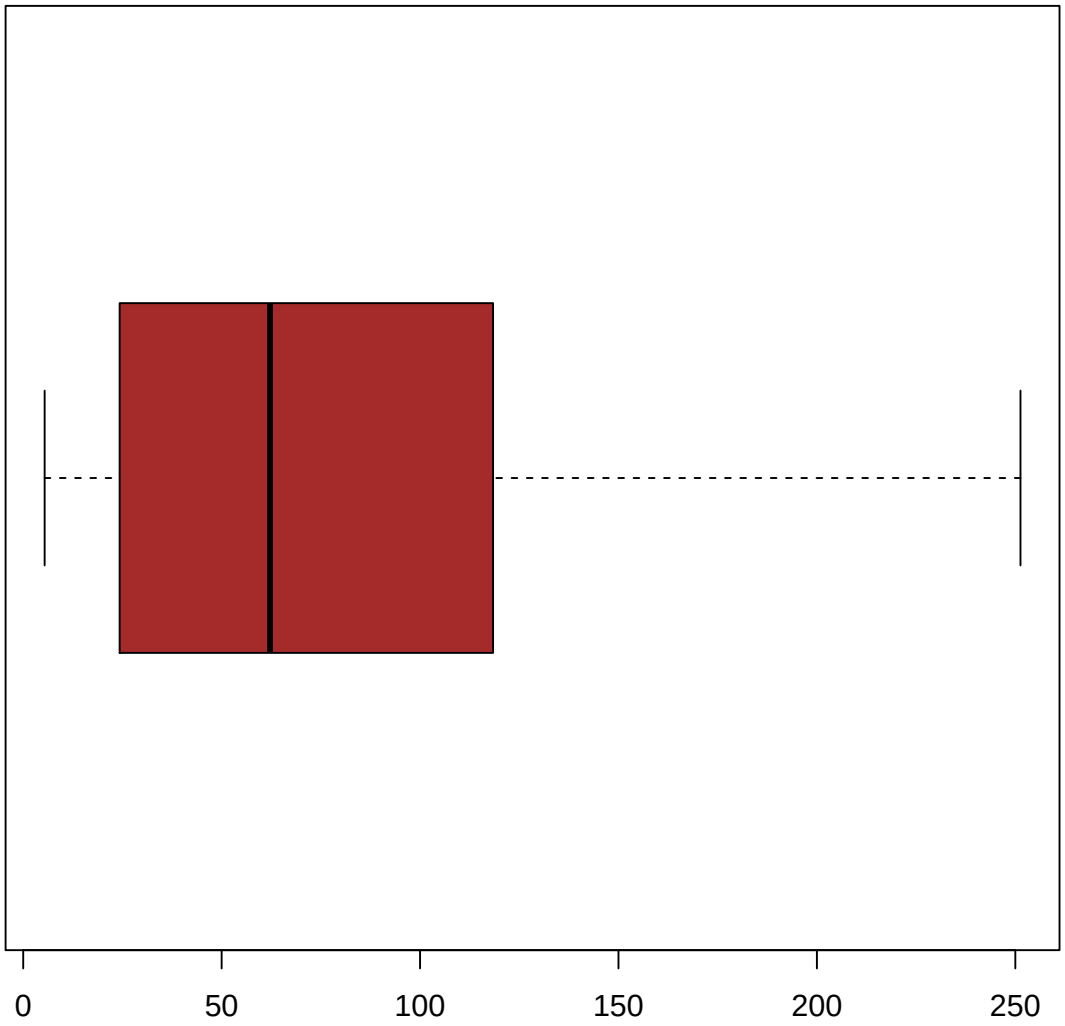


- **Quantiles empiriques** : **valeurs partageant l'échantillon ordonné en un certain nombre de parties de même effectif**.

$$\forall p \in [0,1], \quad \tilde{q}_{n,p} = \begin{cases} (x^*_{np} + x^*_{np+1})/2 \text{ si } np \text{ entier} \\ x^*_{\lfloor np \rfloor + 1} \text{ sinon} \end{cases}$$

Exemple des ampoules : $\tilde{q}_{n,1/4} = x^*_3 = 24.3$, $\tilde{q}_{n,3/4} = x^*_8 = 118.4$ (quartiles).
Pour $p = 1/2$ on retrouve la médiane empirique $\tilde{x}_n = \tilde{q}_{n,1/2}$.

Boite à moustache



Indicateurs de dispersion (ou de variabilité)

La variance empirique

- **Variance empirique**

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2 = \frac{1}{n}\sum_{i=1}^{k} n_i(m_i - \bar{x}_n)^2$$

- $\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}_n^2$ (moyenne du carré - carré de la moyenne)
- **Écart type empirique :** σ_x (racine de la variance)
- **Coefficient de variation empirique :** $cv_n = \frac{\sigma_x}{\bar{x}_n}$ (sans dimension)
- Dans R var(x) donne $\sigma_x'^2 = \frac{n}{n-1}\sigma_x^2$ (variance sans biais)

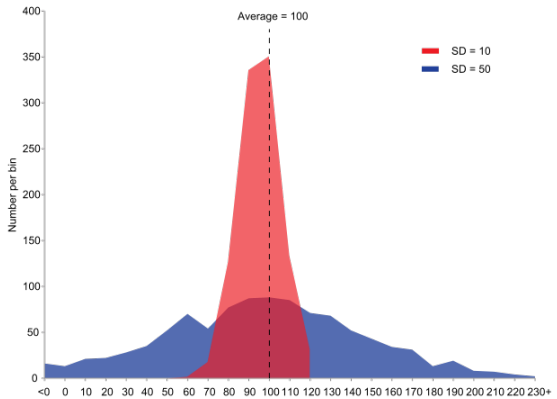
Indicateurs de dispersion (ou de variabilité)

L'écart type empirique

- **Écart type empirique :** σ_x (racine de la variance)
  L'écart type sert à mesurer la dispersion d'un ensemble de données. Plus il est faible, plus les valeurs sont regroupées autour de la moyenne.

Répartition des notes d'une classe

Plus l'écart type est faible, plus la classe est homogène. À l'inverse, s'il est plus important, les notes sont moins resserrées. Dans le cas d'une notation de 0 à 20, l'écart type minimal est 0 (notes toutes identiques), et peut valoir jusqu'à 10 si la moitié de la classe a 0 et l'autre moitié 20



Caractérisation des indicateurs

L'erreur commise en résumant l'observation x_i par c peut être quantifiée par une distance (ou écart) entre ces deux valeurs $d(x_i, c)$.
Un bon indicateur doit minimiser l'erreur moyenne $e = \frac{1}{n}\sum_{i=1}^{n} d(x_i, c)$.

- **Écart quadratique** : $e = \frac{1}{n}\sum_{i=1}^{n}(x_i - c)^2$, minimal quand

$$\frac{\partial e}{\partial c} = 0 \Leftrightarrow -\frac{2}{n}\sum_{i=1}^{n}(x_i - c) = 0 \Leftrightarrow c = \bar{x}_n$$

- **Écart absolu** : $e = \frac{1}{n}\sum_{i=1}^{n}|x_i - c|$, minimal quand $c = \tilde{x}_n$
- **Écart sup** : $e = \frac{1}{n}\sup_{i=1}^{n}|x_i - c|$, minimal quand $c = (x_1^* + x_n^*)/2$

1. Introduction
2. Bases de la statistique descriptive
   - Vocabulaire
   - Tableaux statistiques
   - Méfiez-vous des statistiques ! Le paradoxe de Simpson
3. Représentations graphiques
   - Histogrammes
   - Fonction de répartition empirique
4. Indicateurs statistiques
   - Indicateurs de localisation ou de tendance centrale
   - Indicateurs de dispersion ou de variabilité
5. Corrélation et causalité
   - Régression linéaire
   - Exemples de corrélations

Rappels : indices de localisation, dispersion, relation

Pour un nuage de points (x_i, y_i), $\forall i \in \{1, \dots, n\}$ on définit :

- Les moyennes empiriques (localisation)

$$\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \bar{y}_n = \frac{1}{n}\sum_{i=1}^{n} y_i$$

- Les variances empiriques (dispersion)

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}_n^2$$

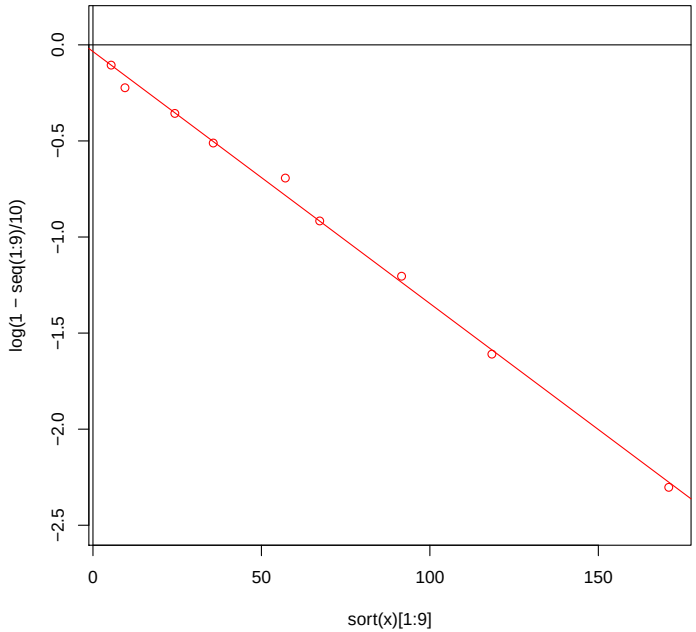
$$\sigma_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y}_n)^2 = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - \bar{y}_n^2$$

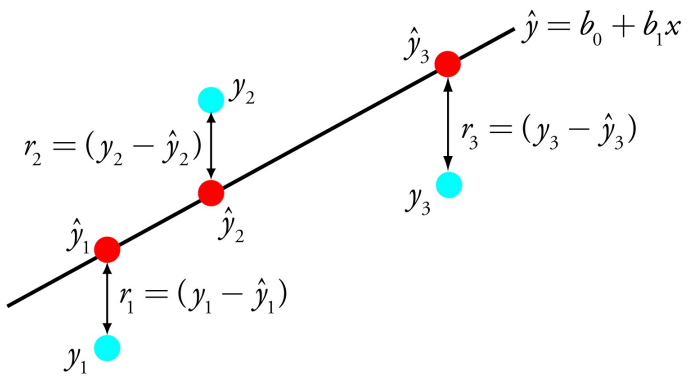
- La covariance et corrélation empirique entre les x_i et y_i (relation)

$$\sigma_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)(y_i - \bar{y}_n) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}_n \bar{y}_n, \quad r_{xy} = \frac{c_{xy}}{s_x s_y}$$

Régression linéaire

Le but est de trouver la droite « la plus proche » d'un nuage de points





L'idée consiste à estimer y_i par $\hat{y}_i = \beta_1 x_i + \beta_0$ en choisissant β_1 et β_0 qui **minimise l'erreur quadratique moyenne** :

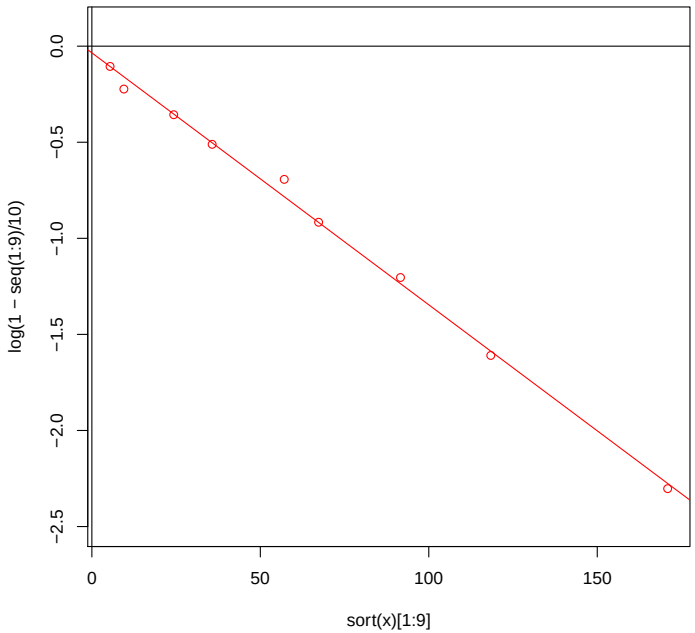
$$\delta^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \beta_1 x_i - \beta_0)^2$$

Crédits : Parag Radke

Régression linéaire

La droite $y = \hat{\beta}_1 x + \hat{\beta}_0$ qui **minimise l'erreur quadratique moyenne** :

$$\hat{\beta}_1 = \frac{\sigma_{xy}}{\sigma_x^2}, \quad \hat{\beta}_0 = \bar{y}_n - \bar{x}_n \frac{\sigma_{xy}}{\sigma_x^2}$$



Taux de décroissance de la loi exponentielle des ampoules

La régression linéaire sur le graphe de probabilité fournit :

$$\hat{\beta}_1 = -0.01311, \quad \hat{\beta}_0 = -0.03484$$

Crédits : O. Gaudoin

Cum hoc ergo propter hoc

Corrélation

Deux événements (appelons les X et Y) sont corrélés si l'on observe une dépendance, une relation entre les deux. Par exemple, le nombre de cheveux d'un homme a tendance à diminuer avec l'âge : âge et nombre de cheveux sont donc corrélés.

Corrélation ou causalité ?

Une erreur de raisonnement courante consiste à dire : « X et Y sont corrélés, donc X cause Y ». On confond alors corrélation et causalité car en réalité, il se pourrait aussi que :

- Y cause X
- X et Y aient une cause commune Z
- X et Y soient accidentellement liés mais n'aient aucun lien de causalité.

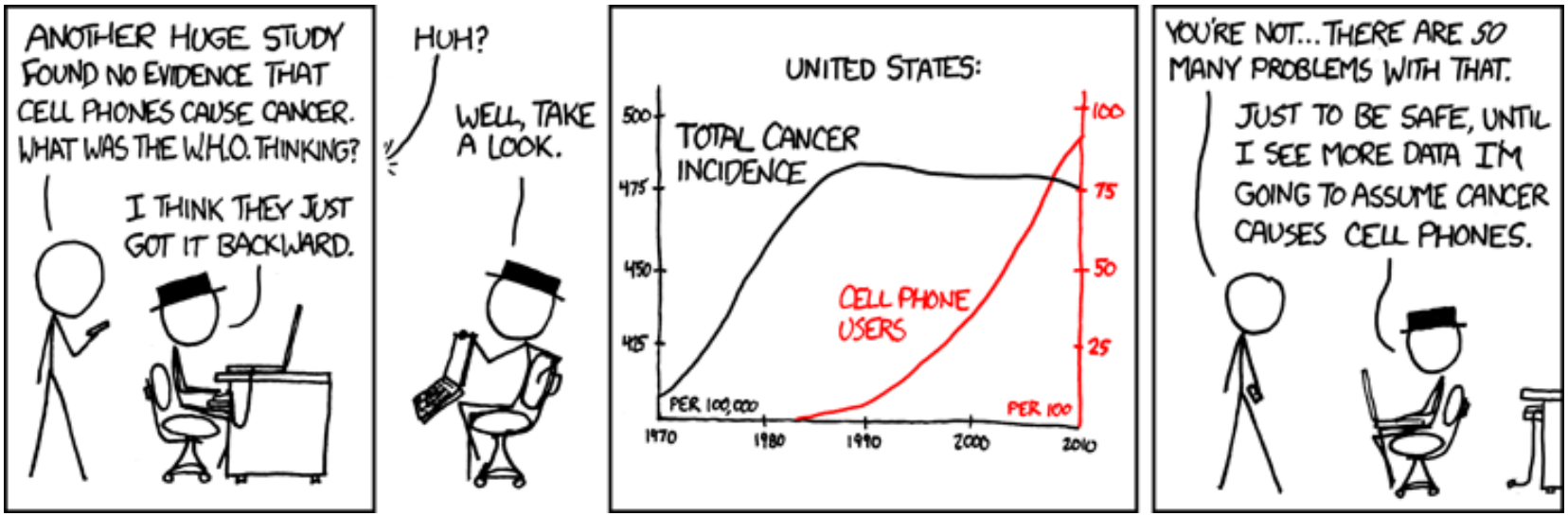
Crédits : Cortecs (zététique)

Cum hoc ergo propter hoc



Effet cigogne

Par exemple, dans les communes qui abritent des cigognes, le taux de natalité est plus élevé que dans l'ensemble du pays. Conclusion : les cigognes apportent les bébés ! Voici une explication plus probable : les cigognes nichent de préférence dans les villages plutôt que dans les grandes agglomérations, et il se trouve que la natalité est plus forte en milieu rural que dans les villes.



Crédits : Cortecs (zététique) + xkcd – https://xkcd.com/925/

Bonne Saint-Valentin...

FOOD LIFE

Saint Valentin 2017 : Il paraît que le fromage est aphrodisiaque !



« Après le gingembre et le chocolat, un petit nouveau vient d'entrer dans le cercle très prisé des aliments aphrodisiaques : le fromage. Oui, vous avez bien lu. »

Bonne Saint-Valentin...

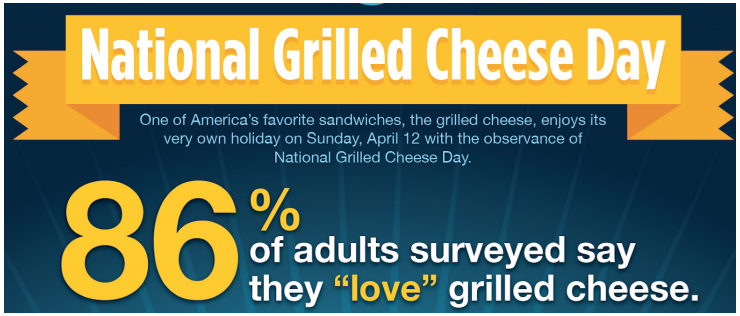
FOOD LIFE

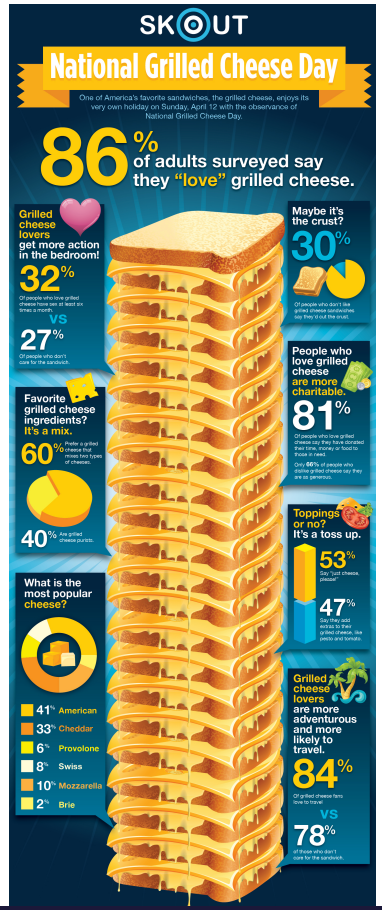
Saint Valentin 2017 : Il paraît que le fromage est aphrodisiaque !

« c'est bien ce que révèle le récent sondage réalisé par le réseau social Skout, également site de rencontres, mené sur 4600 personnes. Interrogées sur leur consommation de fromage et la fréquence de leurs rapports sexuels, l'étude aurait démontré une forte corrélation. Oui, 32% des mangeurs de Grilled Cheese (ce sandwich grillé au fromage dont raffolent les Américains) feraient l'amour en moyenne 6 fois par mois. »

Bonne Saint-Valentin...

Exercice : quelle(s) critique(s) formuleriez-vous à l'égard de ces statistiques ?









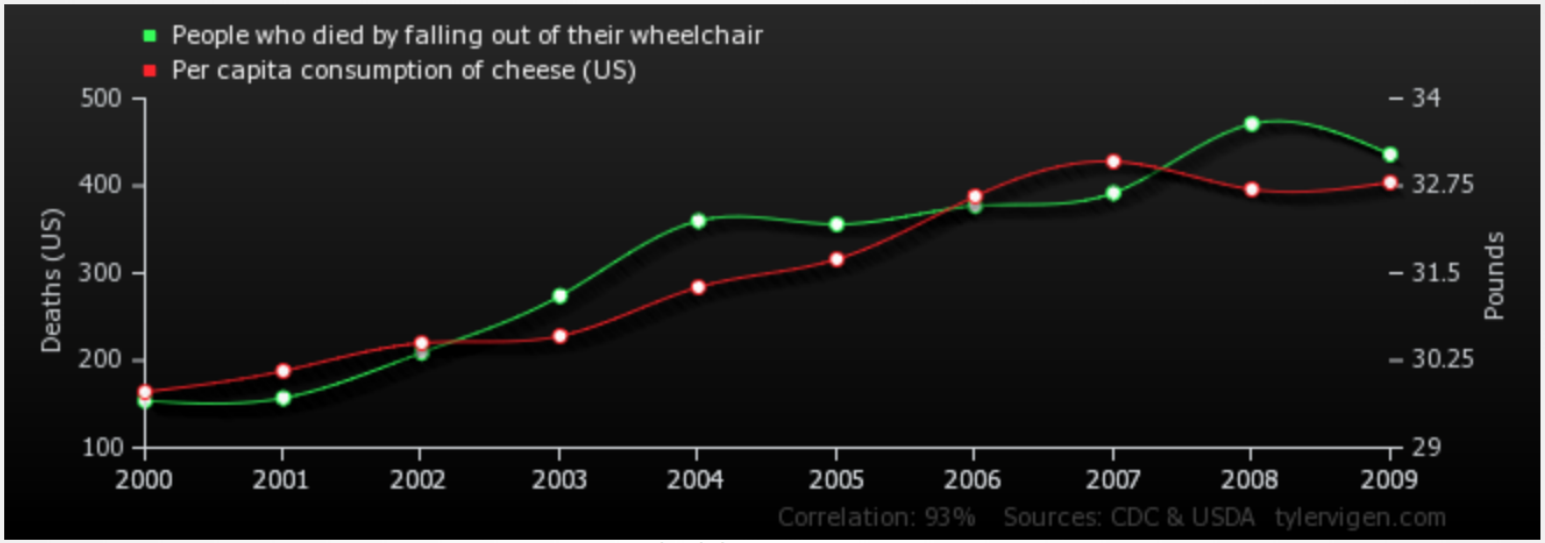


Le fromage, aphrodisiaque... mais dangereux !

Nombre de personnes handicapées décédées d'une chute de leur fauteuil

corrélé avec

La consommation de fromage par habitant



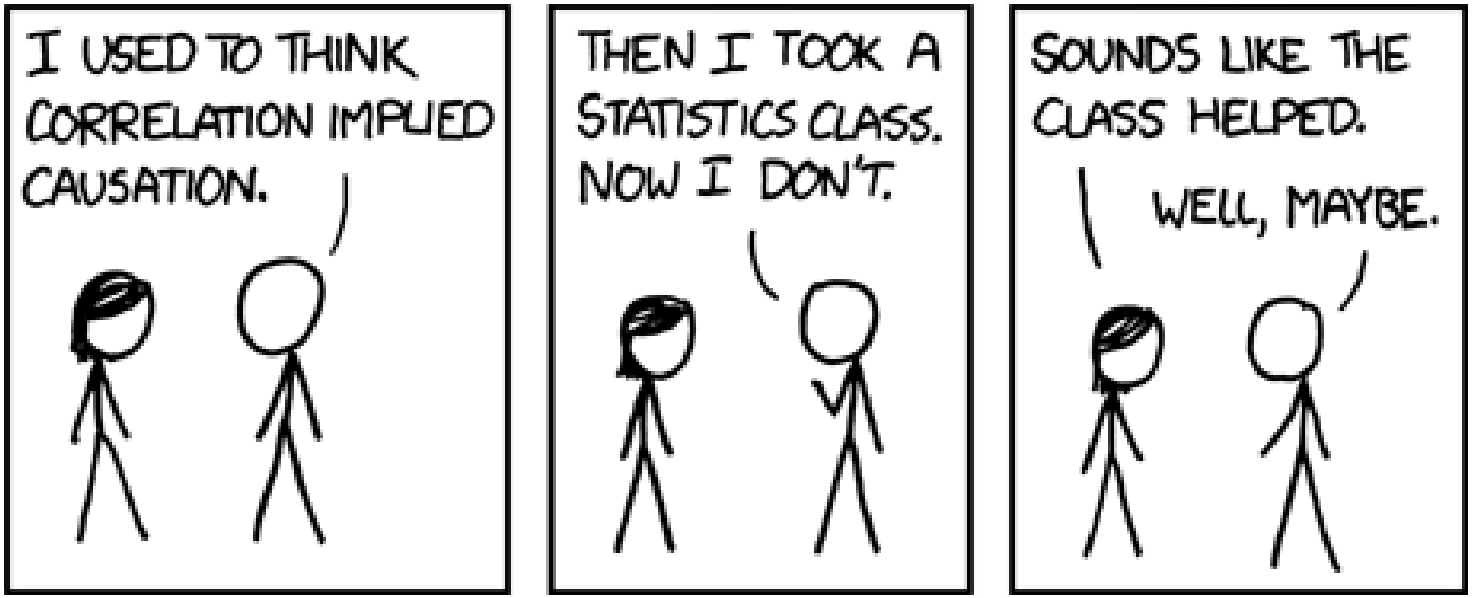
Upload this image to imgur

	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
People who died by falling out of their wheelchair *Deaths (US) (CDC)*	154	157	209	274	360	356	377	392	471	436
Per capita consumption of cheese (US) *Pounds (USDA)*	29.8	30.1	30.5	30.6	31.3	31.7	32.6	33.1	32.7	32.8

Correlation: 0.931497

Crédits : Tyler Vigen – http://www.tylervigen.com/spurious-correlations

Hope it helps



Crédits : xkcd – https://xkcd.com/552/

CQFR : **C**e **Q**u'il **F**aut **R**etenir

CQFR

- Corrélation $\neq$ causalité
- La statistique peut être comme la langue d'Esope la meilleure ou la pire des choses. Il convient de se méfier des pièges qu'elle recèle tout en se servant de ses résultats.
- Les représentations graphiques des données statistiques permettent une analyse visuelle de la répartition des données.
- Les indicateurs de localisation, de dispersion et de relation permettent de les quantifier et de résumer l'information.
- Ces deux outils suggèrent une caractérisation de la loi statistique sous-jacente et donc des modèles théoriques plausibles.

Et après ?

Pour aller plus loin dans l'analyse et la généralisation

on a besoin d'outils probabilistes

Suite au prochain épisode ...

Partie II : Introduction à la théorie des probabilités

1. Introduction

2. Rappels de probabilité
   - Axiomatique
   - Conditionnement et indépendance

3. Variables aléatoires
   - Définitions discrètes et continues
   - Fonction de densité et de répartition
   - Espérance et variance
   - Lois usuelles

4. Théorème limites
   - La loi des grands nombres
   - Théorème limite central

1 Introduction

2 Rappels de probabilité
- Axiomatique
- Conditionnement et indépendance

3 Variables aléatoires
- Définitions discrètes et continues
- Fonction de densité et de répartition
- Espérance et variance
- Lois usuelles

4 Théorème limites
- La loi des grands nombres
- Théorème limite central

1 Introduction

2 Rappels de probabilité
- Axiomatique
- Conditionnement et indépendance

3 Variables aléatoires
- Définitions discrètes et continues
- Fonction de densité et de répartition
- Espérance et variance
- Lois usuelles

4 Théorème limites
- La loi des grands nombres
- Théorème limite central

Modéliser une expérience aléatoire

1) Choix de l'univers

Expérience : On lance deux dés, non pipés et identiques.

- On note l'univers des possibles Ω l'ensemble des résultats possibles de l'expérience. Ω dépend de l'usage de l'expérience.

  Exemples : on note les 2 chiffres obtenus
  $\Omega_1 = \{(1,1),(1,2),(1,3),\ldots\}$; $\Omega_2 = \{2,3,\ldots 12\}$ si on s'intéresse à la somme des chiffres des 2 dés, ...

- On note évènement A une proposition relative au résultat de l'expérience

  Exemples : j'ai tiré un 3, la somme des points est égale à 7, la somme est supérieur a 10, etc.

- On note une tribu $\mathcal{A}$ l'ensemble des évènements ($\Omega \in \mathcal{A}$, $\varnothing \in \mathcal{A}$) inclus dans l'ensemble des parties de $\mathcal{P}(\Omega)$, possédant une certaine structure (d'algèbre) : pour tout $A \in \mathcal{A}$ le complémentaire $A^c \in \mathcal{A}$ et $\mathcal{A}$ est stable par réunion finie ou dénombrable.

Modéliser une expérience aléatoire

2) Choix de la mesure de probabilité sur cet univers/cette tribu

- On appelle loi de probabilité sur $(\Omega, \mathcal{A})$ l'application :

$$\mathbb{P} : \mathcal{A} \longmapsto [0, 1] \text{ telle que}$$

$$\mathbb{P}(\Omega) = 1$$

$$\mathbb{P}\left(\bigcup_{i \geq 0} A_i\right) = \sum_{i=0}^{+\infty} \mathbb{P}(A_i) \text{ pour des évènements incompatibles}$$

Modéliser une expérience aléatoire

Espace probabilisé

Une probabilité en mathématiques présuppose :

Deux ingrédients indispensables

1. Le choix de l'univers et d'une tribu modélisant l'expérience aléatoire
2. Le choix de la mesure de probabilité sur cet univers/tribu

L'espace $(\Omega, \mathcal{A}, \mathbb{P})$ est appelé **espace probabilisé**.

Nota Bene : en faisant ces choix on « modélise » le phénomène aléatoire, c'est-à-dire qu'on émet des hypothèses quant à sa nature, permettant ensuite de calculer la probabilité d'évènements sur cette base. La branche des mathématiques dévouée au choix de la mesure de probabilité, donc au choix du modèle, est la *statistique*.

Modéliser une expérience aléatoire

Le cas du lancé de 2 dés

Supposés non pipés chaque face des dés a même chance d'apparition, mais selon le choix de l'univers le calcul de la probabilité de l'évènement $A =$ « la somme des dés est égale à 4 » va être différent.

- Si $\Omega_1 = \{(1,1),(1,2),(1,3),\ldots\}$, on est en situation d'**équiprobabilité** et la probabilité $\mathbb{P}_1$ sur Ω_1 est prise uniforme :

$$\mathbb{P}_1(\{i,j\}) = \frac{1}{36}, \quad \forall (i,j) \in \{1,\ldots,6\}^2$$

$$\mathbb{P}_1(A) = \mathbb{P}_1[\{(1,3),(2,2),(3,1)\}] = \frac{\mathrm{Card}(A)}{\mathrm{Card}(\Omega_1)} = \frac{3}{36} = \frac{1}{12}$$

- Si $\Omega_2 = \{2,3,\ldots 12\}$, on est plus en situation d'équiprobabilité, en effet $\mathbb{P}(\{2\}) = \frac{1}{36}$ (cas unique où on obtient deux 1) et

$$\mathbb{P}(A) = \mathbb{P}(\{4\}) = \frac{1}{12} \neq \frac{\mathrm{Card}(A)}{\mathrm{Card}(\Omega_2)} = \frac{1}{11}$$

$\Rightarrow$ *Cf.* Paradoxe de Bertrand pour le choix crucial de l'univers.

Modéliser une expérience aléatoire

L'interprétation de la probabilité est ensuite suspendue

La probabilité d'un évènement peut se comprendre soit :

- De manière « fréquentiste » : c'est le pourcentage de fois où l'évènement se produit si on répète indéfiniment la même expérience.
- De manière « subjectiviste » : c'est alors une mesure subjective (un degré de croyance) dépendant du contexte et de la vraisemblance de l'évènement.

« On ne peut guère donner une définition satisfaisante de la probabilité. La définition complète de la probabilité est donc une sorte de pétition de principe » (Henri Poincaré)

Nota Bene : Les mathématiques s'exonèrent de ces considérations métaphysique par l'axiomatisation de Kolmogorov. Néanmoins la théorie des probabilité est « robuste » au sens où ces axiomes permettent de démontrer la loi des grands nombres, qui fait le lien avec l'approche fréquentiste, ce qui justifie *a posteriori* le cadre développé.

Propriétés découlant des axiomes des probabilités

Une probabilité $\mathbb{P}(A)$ est définie sur un évènement A

- $\mathbb{P}(\Omega) = 1$
- $\mathbb{P}(\varnothing) = 0$
- $0 < \mathbb{P}(A) < 1$
- $\mathbb{P}(\bar{A}) = 1 - \mathbb{P}(A)$
- $\mathbb{P}(A) \leq \mathbb{P}(B)$ si $A \subset B$
- $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$
- $\mathbb{P}(\cup A_i) \leq \sum \mathbb{P}(A_i)$

Probabilités conditionnelles

La **probabilité conditionnelle** d'un évènement A sachant un évènement B dénote la probabilité de A **dans le cas où B est réalisé**, notée $\mathbb{P}(\mathbf{A} \,|\, \mathbf{B})$, et définie par :

$$\mathbb{P}_{\mathbf{B}}(\mathbf{A}) := \mathbb{P}(\mathbf{A} \,|\, \mathbf{B}) := \frac{\mathbb{P}(\mathbf{A} \cap \mathbf{B})}{\mathbb{P}(\mathbf{B})}$$

On a alors :

$$\mathbb{P}(\mathbf{A} \cap \mathbf{B}) = \mathbb{P}(\mathbf{A} \,|\, \mathbf{B})\ \mathbb{P}(\mathbf{B}) = \mathbb{P}(\mathbf{B} \,|\, \mathbf{A})\ \mathbb{P}(\mathbf{A})$$

Remarque : $\mathbb{P}(A \cap B)$ est symétrique, $\mathbb{P}(A \,|\, B)$ ne l'est pas. On vérifie que $\mathbb{P}_B$ définit bien une probabilité.

Évènements indépendants

- A est indépendant de B si $\mathbb{P}(A \mid B) = \mathbb{P}(A)$

  c'est-à-dire la connaissance de B ne change pas les "chances" de réalisation de A

- A est indépendant de $B \Rightarrow B$ est indépendant de A

- Si A and B sont **indépendants**, alors :

$$\mathbb{P}(\mathbf{A} \cap \mathbf{B}) = \mathbb{P}(\mathbf{A})\ \mathbb{P}(\mathbf{B})$$

Théorème de Bayes

$$\boxed{\mathbb{P}(\mathbf{A} \mid \mathbf{B}) = \frac{\mathbb{P}(\mathbf{A} \cap \mathbf{B})}{\mathbb{P}(\mathbf{B})} = \frac{\mathbb{P}(\mathbf{B} \mid \mathbf{A})\ \mathbb{P}(\mathbf{A})}{\mathbb{P}(\mathbf{B})}}$$

Par la **formule des probabilités totales** :

$$\mathbb{P}(B) = \mathbb{P}(A \cap B) + \mathbb{P}(\bar{A} \cap B) = \mathbb{P}(B \mid A)\ \mathbb{P}(A) + \mathbb{P}(B \mid \bar{A})\ \mathbb{P}(\bar{A})$$

D'où :

$$\mathbb{P}(\mathbf{A} \mid \mathbf{B}) = \frac{\mathbb{P}(\mathbf{B} \mid \mathbf{A})\ \mathbb{P}(\mathbf{A})}{\mathbb{P}(\mathbf{B} \mid \mathbf{A})\ \mathbb{P}(\mathbf{A}) + \mathbb{P}(\mathbf{B} \mid \overline{\mathbf{A}})\ \mathbb{P}(\overline{\mathbf{A}})}$$

Thomas Bayes (1701-1761) mathématicien et pasteur britannique
Pierre-Simon Laplace (1749-1827) mathématicien, astronome, physicien français

À retenir

- **Théorème de Bayes**

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B \mid A)\ \mathbb{P}(A)}{\mathbb{P}(B)}$$

- **Règle de la somme**

$$\mathbb{P}(A) = \sum_{B} \mathbb{P}(A \cap B)$$

- **Règle du produit**

$$\mathbb{P}(A \cap B) = \mathbb{P}(A \mid B)\mathbb{P}(B)$$

- **D'où**

$$\mathbb{P}(A) = \sum_{B} \mathbb{P}(A \mid B)\mathbb{P}(B)$$

Vocabulaire : probabilité conditionnelle, probabilité jointe, probabilité marginale

Théorème de Bayes

À vos boitiers !

Vous venez de passer un test pour le dépistage du cancer.
Le médecin vous convoque pour vous annoncer le résultat : mauvaise nouvelle, **il est positif**. Pas de chance, alors que ce type de cancer **ne touche que 0.1% de la population**.

Vous lui demandez si le test est fiable. Sa réponse est sans appel :
« Si vous avez le cancer, le test sera positif dans 90% des cas ; alors que si vous ne l'avez pas, il sera négatif dans 97% des cas ».

Selon vous, après le résultat d'un tel test,
quelle est la probabilité que vous ayez le cancer ?

A) $> 90\%$
B) $= 90\%$
C) $= 9\%$
D) $< 5\%$

Théorème de Bayes

À vos boitiers !

Vous venez de passer un test pour le dépistage du cancer.
Le médecin vous convoque pour vous annoncer le résultat : mauvaise nouvelle, **il est positif**. Pas de chance, alors que ce type de cancer **ne touche que 0.1% de la population**.

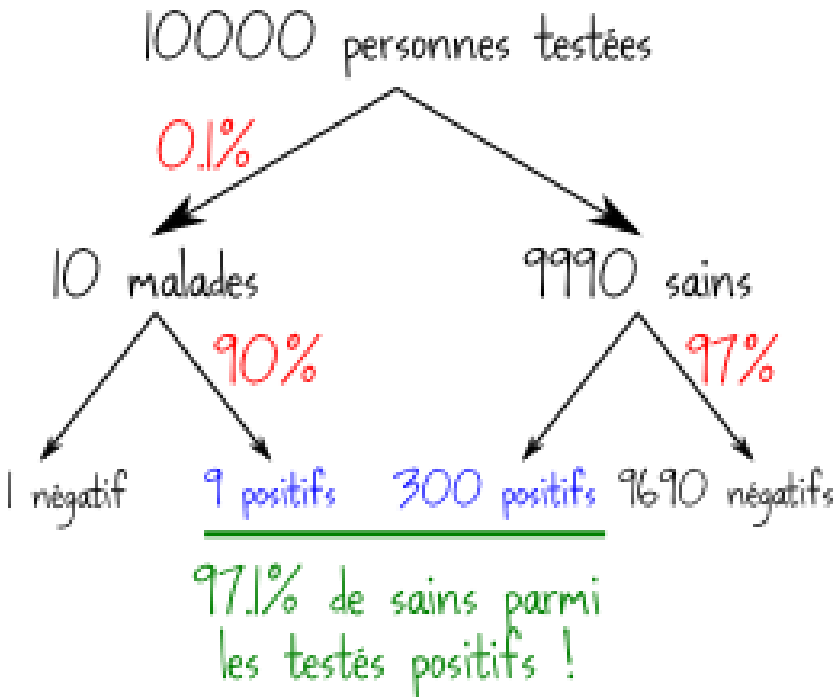
Vous lui demandez si le test est fiable. Sa réponse est sans appel :
« Si vous avez le cancer, le test sera positif dans 90% des cas ; alors que si vous ne l'avez pas, il sera négatif dans 97% des cas ».

Selon vous, après le résultat d'un tel test,
quelle est la probabilité que vous ayez le cancer ?

A) $> 90\%$
B) $= 90\%$
C) $= 9\%$
D) $< 5\%$

Théorème de Bayes

Faux positifs



Explications :

Sur les 309 personnes qui sont testées positives, 9 seulement sont réellement malades, et **300 sont saines** : ces 300 sont ce qu'on appelle des **faux positifs**. Si vous êtes positif, vous n'avez donc que

$$\boxed{\frac{9}{309} = 2.9\%}$$

de risque d'être réellement malade, et donc 97.1% de chance d'être un faux positif, et donc d'être sain.

Crédits : D. Louapre (Science Étonnante)

Vous avez répondu 90% ou plus ?

- Si vous êtes testé positif et que vous vous demandez si vous avez le cancer, vous cherchez :

  « la probabilité d'être malade sachant que le test est positif »

- Quand le médecin vous dit que « Si vous avez le cancer, le test sera positif dans 90% des cas », il s'agit de :

  « la probabilité d'être testé positif sachant que l'on est malade »

Théorème de Bayes

Formuler le problème

$H =$ "je suis malade" *(hypothèse à tester)*
$O =$ "le test est positif" *(l'observation)*

- Si vous êtes testé positif et que vous vous demandez si vous avez le cancer, vous cherchez $\mathbb{P}(H \mid O)$:

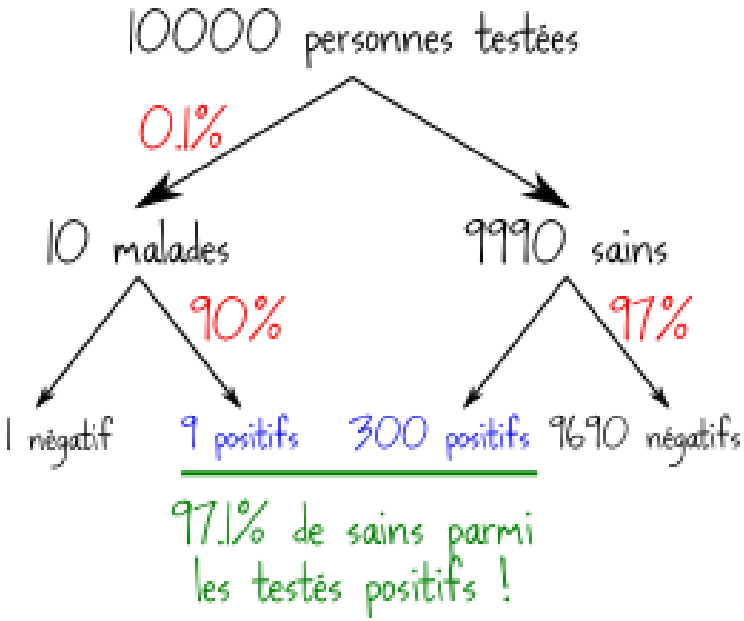
« la probabilité d'être malade sachant que le test est positif »

- Quand le médecin vous dit que « Si vous avez le cancer, le test sera positif dans 90% des cas », il s'agit de $\mathbb{P}(O \mid H)$:

« la probabilité d'être testé positif sachant que l'on est malade »

Théorème de Bayes

Écrire la formule de Bayes



$H =$ "je suis malade" *(hypothèse à tester)*
$O =$ "le test est positif" *(l'observation)*

- $\mathbb{P}(H) = 0.001$
- $\mathbb{P}(O \mid H) = 0.9$
- $\mathbb{P}(O) = \mathbb{P}(O \cap H) + \mathbb{P}(O \cap \bar{H})$

$$\begin{aligned}\mathbb{P}(O) &= \mathbb{P}(O \mid H)\mathbb{P}(H) + \mathbb{P}(O \mid \bar{H})\mathbb{P}(\bar{H}) \\ &= 0.9 \times 0.001 + 0.03 \times 0.999 \\ &= 0.03087\end{aligned}$$

$$\begin{aligned}\mathbb{P}(H \mid O) &= \frac{\mathbb{P}(O \mid H)}{\mathbb{P}(O)}\mathbb{P}(H) \\ &= \frac{0.9}{0.03087} \times 0.001 \approx 2.9\%\end{aligned}$$

Crédits : D. Louapre

1. **Ne pas se tromper de question : $\mathbb{P}(A \mid B) \neq \mathbb{P}(B \mid A)$**

« la probabilité d'être malade sachant que le test est positif »

$\neq$

« la probabilité d'être testé positif sachant que l'on est malade »

Raisonnement inductif

Erreurs à ne pas commettre

1. **Ne pas se tromper de question** : $\mathbb{P}(A \mid B) \neq \mathbb{P}(B \mid A)$
2. **Ne pas négliger le taux de base** : la fiabilité d'un test n'est pas suffisant, il faut s'intéresser à la probabilité *a priori*
   - Manger un aliment X augmente de 300% le risque de cancer C
   - Ne pas manger X augmente de 30% le risque d'anémie A

   Oui **MAIS** :
   - Si $\mathbb{P}(C) = 0,0001\%$ alors l'augmentation du risque de 300% donne $\mathbb{P}(C \mid X) = 0,0004\% \Rightarrow$ augmentation brute de $0,0003\%$
   - Si $\mathbb{P}(A) = 0,1\%$ alors l'augmentation du risque de 30% donne $\mathbb{P}(A \mid \bar{X}) = 0,13\% \Rightarrow$ augmentation brute du risque de $0,03\%$ soit **100 fois plus importante**.

Raisonnement inductif

Erreurs à ne pas commettre

1. **Ne pas se tromper de question** : $\mathbb{P}(A \mid B) \neq \mathbb{P}(B \mid A)$
2. **Ne pas négliger le taux de base** : la fiabilité d'un test n'est pas suffisant, il faut s'intéresser à la probabilité *a priori*
3. **Prendre en considération les faux positifs et faux négatifs** : la fiabilité d'un test s'évalue au regard de ces types d'erreurs I & II.
   - Sensibilité, spécificité et valeurs prédictives d'un test de dépistage : http://www.adeca68.fr/prevention_et_depistage/performances_dun_test_de_depistage.166.html
   - Important pour apprécier la pertinence de dépistages systématiques



Raisonnement inductif

Erreurs à ne pas commettre

1. **Ne pas se tromper de question** : $\mathbb{P}(A \mid B) \neq \mathbb{P}(B \mid A)$
2. **Ne pas négliger le taux de base** : la fiabilité d'un test n'est pas suffisant, il faut s'intéresser à la probabilité *a priori*
3. **Prendre en considération les faux positifs et faux négatifs** : la fiabilité d'un test s'évalue au regard de ces types d'erreurs I & II.
4. Une affirmation extraordinaire requiert une preuve extraordinaire
   - **Rappel** : Test fiable à 99% $\neq$ 99% de chance que le test soit vrai !
   - Exemple : Une maladie touche 0,1% de gens, vous passez un test qui est négatif et fiable à 99%. Il serait étrange que votre probabilité d'être sain soit passée à 99% alors qu'initialement elle était de 99,9%.
   - Au contraire le résultat négatif augmente votre probabilité d'être sain
   - **Dans quelle mesure ?** Si la preuve est fiable et notre degré de croyance initiale est fort alors on y croira encore plus à l'issue du test. Si on y croyait presque pas avant le test, on y croira plus que « presque pas », ce qui ne signifie pas forcément d'y croire tout à fait ! Si vous partez d'une croyance *a priori* extrêmement faible il vous faudra une preuve extrêmement fiable pour passer la barre des 50%.

Raisonnement inductif

Erreurs à ne pas commettre

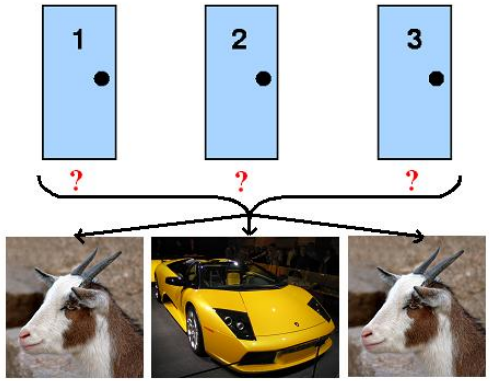
1. **Ne pas se tromper de question** : $\mathbb{P}(A \mid B) \neq \mathbb{P}(B \mid A)$
2. **Ne pas négliger le taux de base** : la fiabilité d'un test n'est pas suffisant, il faut s'intéresser à la probabilité *a priori*
3. **Prendre en considération les faux positifs et faux négatifs** : la fiabilité d'un test s'évalue au regard de ces types d'erreurs I & II.
4. **Une affirmation extraordinaire requiert une preuve extraordinaire** : Si vous partez d'une croyance *a priori* extrêmement faible il vous faudra une preuve extrêmement fiable pour passer la barre des 50%.

Monty Hall

À vos boitiers !

Vous êtes candidat à un jeu télévisé animé par un présentateur.

- Soit **trois portes**, l'une cache **une voiture**, les **deux autres une chèvre**, répartis par tirage au sort et connu du présentateur.
- Vous choisissez une des portes, mais rien n'est révélé.
- Le présentateur ouvre une autre porte ne révélant pas la voiture.
- **Et vous propose avant d'ouvrir d'échanger votre choix.**



Beeeeh que faites vous ? Est-il préférable :

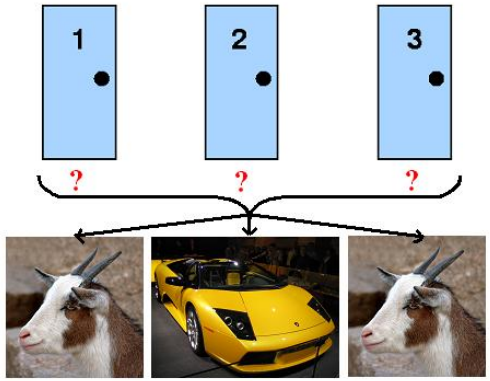
A) De conserver son choix
B) De changer son choix

Le paradoxe de Monty Hall

Le premier choix n'est pas toujours le bon...

Vous êtes candidat à un jeu télévisé animé par un présentateur.

- Soit **trois portes**, l'une cache **une voiture**, les **deux autres une chèvre**, répartis par tirage au sort et connu du présentateur.
- Vous choisissez une des portes, mais rien n'est révélé.
- Le présentateur ouvre une autre porte ne révélant pas la voiture.
- **Et vous propose avant d'ouvrir d'échanger votre choix.**



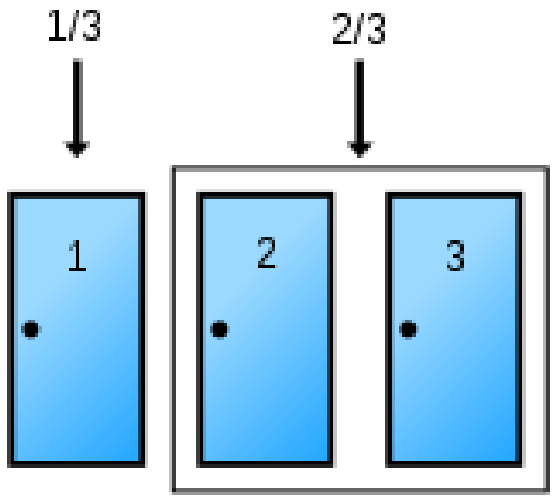
Beeeeh que faites vous ? Est-il préférable :

A) De conserver son choix
B) **De changer son choix**

Le paradoxe de Monty Hall

Explications

Vous choisissez la porte 1.



- $\mathbb{P}(O_i)$ = probabilité que le présentateur ouvre la porte i
- $\mathbb{P}(H_i)$ = probabilité que la voiture soit derrière la porte i

Ce que l'on sait *a priori*

Avant ouverture d'une porte :

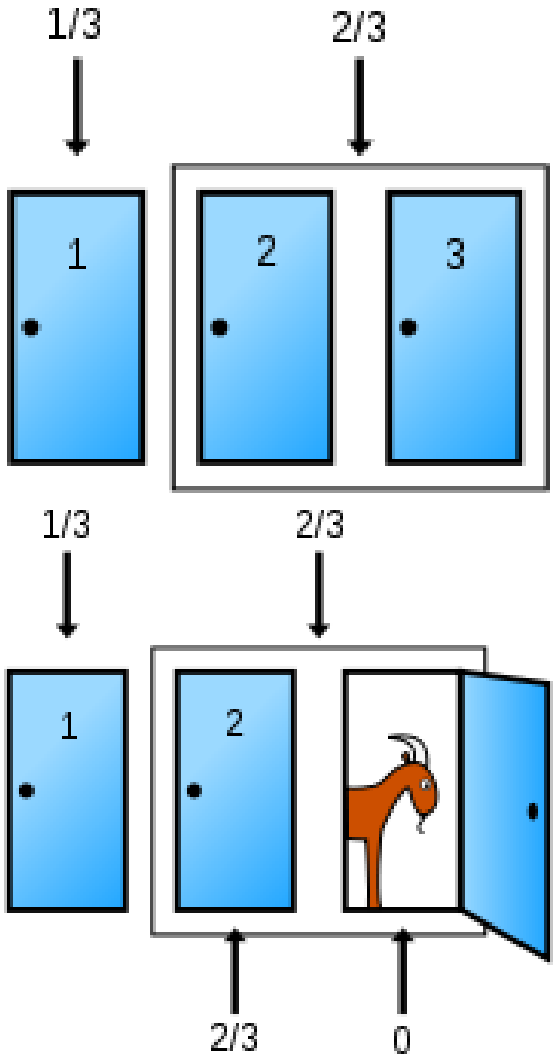
- $\mathbb{P}(H_1) = \mathbb{P}(H_2) = \mathbb{P}(H_3) = \frac{1}{3}$
- $\mathbb{P}(O_3 \mid H_1) = \frac{1}{2}$
- $\mathbb{P}(O_3 \mid H_3) = 0$
- $\mathbb{P}(O_3 \mid H_2) = 1$
- Formule des probabilités totales :

$$\begin{aligned}\mathbb{P}(O_3) &= \sum_{i=1}^{3} \mathbb{P}(O_3 \mid H_i)\mathbb{P}(H_i) \\ &= \frac{1}{2}\end{aligned}$$

Le paradoxe de Monty Hall

Explications

Le présentateur ouvre la porte 3.



Ce que l'on sait *a posteriori*

Après ouverture de la porte 3 :

$$\begin{aligned}\mathbb{P}(H_2 \mid O_3) &= \frac{\mathbb{P}(O_3 \mid H_2)}{\mathbb{P}(O_3)} \times \mathbb{P}(H_2) \\ &= \frac{1}{\frac{1}{2}} \times \frac{1}{3} \\ &= \frac{2}{3}\end{aligned}$$

Vous devez choisir la porte 2 !

Anatomie du raisonnement bayésien

Un modèle d'apprentissage

$$\underbrace{\mathbb{P}(H \mid O)}_{\text{probabilité a posteriori}} = \underbrace{\frac{\overbrace{\mathbb{P}(O \mid H)}^{\text{vraisemblance}}}{\mathbb{P}(O)}}_{\text{apport des observations}} \underbrace{\mathbb{P}(H)}_{\text{probabilité a priori}}$$

- $\mathbb{P}(H)$ degré de confiance que l'on a vis-à-vis de l'hypothèse H avant de prendre en compte les observations
- $\mathbb{P}(H \mid O)$ degré de confiance après prise en compte des observations
- Le terme $\mathbb{P}(O \mid H)$ s'appelle la vraisemblance, et quantifie le degré de compatibilité de l'hypothèse H et des observations O

La formule de Bayes est alors un moyen de relier la probabilité *a posteriori*, et la probabilité *a priori*. C'est donc une formule qui permet de réviser nos degrés de confiance en fonction des observations et de rendre quantitatif le raisonnement inductif.

Anatomie du raisonnement bayésien

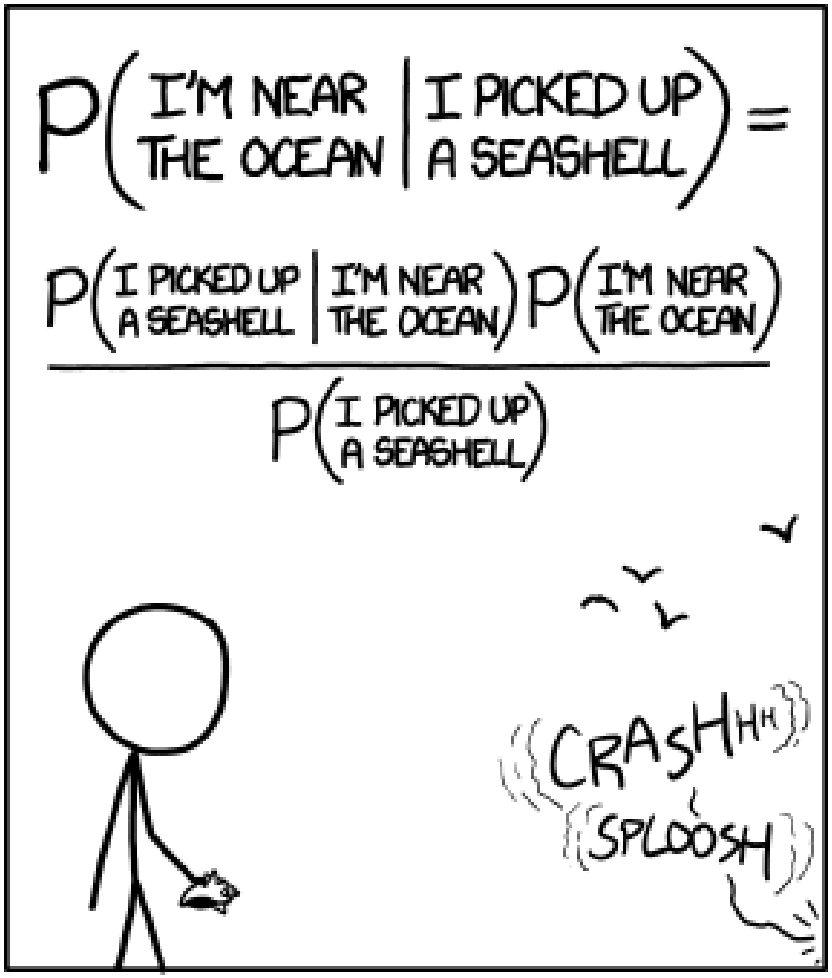
Un modèle d'apprentissage

$$\underbrace{\mathbb{P}(H \mid O)}_{\text{probabilité a posteriori}} = \underbrace{\frac{\overbrace{\mathbb{P}(O \mid H)}^{\text{vraisemblance}}}{\mathbb{P}(O)}}_{\text{apport des observations}} \underbrace{\mathbb{P}(H)}_{\text{probabilité a priori}}$$

Exemples de champs d'application :

- Sciences cognitives bayésiennes : cerveau statisticiens des bébés
- Inférence bayésienne en perception visuelle
- Filtrage de spam
- Réseaux bayésiens en apprentissage machine
- Justice
- ...

Formule de Bayes par xkcd



Crédits : xkcd – https://xkcd.com/1236/

1 Introduction

2 Rappels de probabilité
- Axiomatique
- Conditionnement et indépendance

3 Variables aléatoires
- Définitions discrètes et continues
- Fonction de densité et de répartition
- Espérance et variance
- Lois usuelles

4 Théorème limites
- La loi des grands nombres
- Théorème limite central

Variables aléatoires

- Le concept de **variable aléatoire** formalise la notion de grandeur variant selon le résultat d'une expérience aléatoire.
- Une variable aléatoire X est une **fonction** $X : \Omega \to \mathbb{R}$ qui permet de passer d'une sortie d'une expérience aléatoire vers un nombre de $\mathbb{R}$.
- On distingue les **variables aléatoires discrètes** et les **variables aléatoires continues**.

  Exemples : La variable qui donne la somme des deux valeurs obtenus par le lancé de deux dés (discrète), la variable qui donne la taille d'un étudiant du DLST (continue).
- Pour tout ensemble (borélien) $A \subset \mathbb{R}$, cet ensemble est un **évènement**

$$\{X \in A\} = \{\omega \in \Omega : X(\omega) \in A\}$$

Exercice : Pour une variable discrète $A = \{k\}$ et on note l'ensemble $[X = k]$, par ex. l'ensemble des cas où la somme des dés vaut 4. Pour une variable continue on a par exemple un intervalle $A = [a, b]$ et $\{X \in [a, b]\}$ représente par ex. l'ensemble des étudiants dont la taille se situe entre 1m70 et 1m80.

Loi d'une variables aléatoires

L'application $\mathbb{P}_X : \mathcal{B} \subset \mathbb{R} \longmapsto [0,1]$ définie par :

$$\mathbb{P}_X(A) = \mathbb{P}(X \in A) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \in A\}), \quad \forall A \in \mathcal{B}\ ,$$

est une mesure de probabilité sur $(\mathbb{R}, \mathcal{B})$, appelée **loi de X**.

Loi d'une variables aléatoires discrète

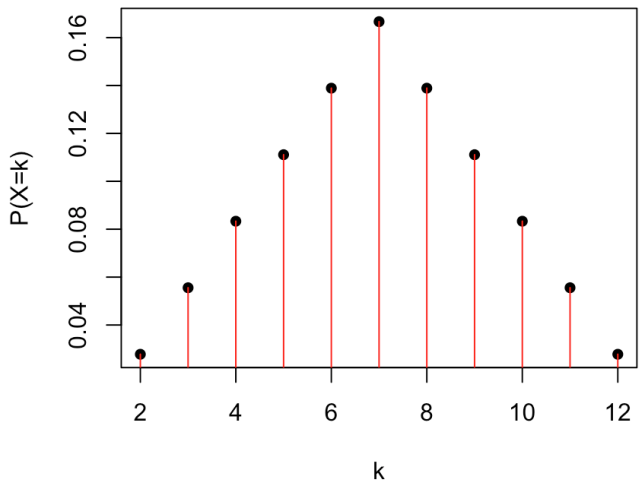
Exemple avec la variable somme des dés

Soit le cas considéré avec l'univers $\Omega = \{(1,1),(1,2),\ldots,(6,5),(6,6)\}$ et la variable aléatoire

$$\begin{array}{rccc} X: & \Omega & \to & \mathbb{R} \\ & (\omega_1,\omega_2) & \mapsto & \omega_1+\omega_2 \end{array}$$

L'ensemble des valeurs possibles de X est $\{2,3,\ldots 12\}$. La loi de X, ou encore distribution de probabilité, est $k \mapsto \mathbb{P}_X(k) = \mathbb{P}(X=k)$ dont les valeurs sont données par le tableau suivant :

k	2	3	4	5	6	7	8	9	10	11	12
$\mathbb{P}(X=k)$	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36



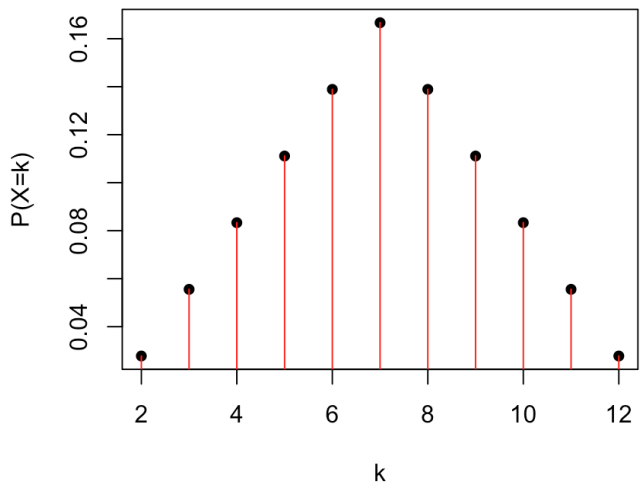
Densité et fonction de répartition d'une variables discrète

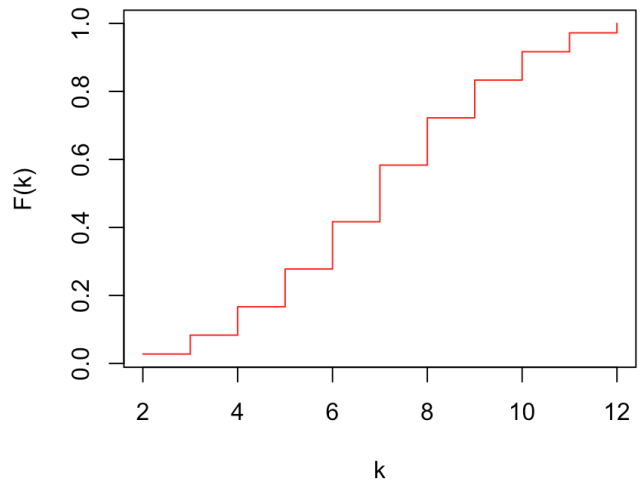
Exemple avec la variable somme des dés

- La loi de X est $k \mapsto \mathbb{P}_X(k) = \mathbb{P}(X = k) \equiv p_k$
- Le graphe (k, p_k) (à gauche) représente la fonction densité
- La fonction constante par morceau (à droite) représente la fonction de répartition $F : \mathbb{R} \to [0, 1]$ définie par

$$F(x) = \mathbb{P}_X(]-\infty, x]) = \mathbb{P}(X \leq x) = \sum_{i=1}^{k} \mathbb{P}(X = x_i), \quad k \text{ t.q } x_k \leq x < x_{k+1}$$

k	2	3	4	5	6	7	8	9	10	11	12
$\mathbb{P}(X = k)$	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36





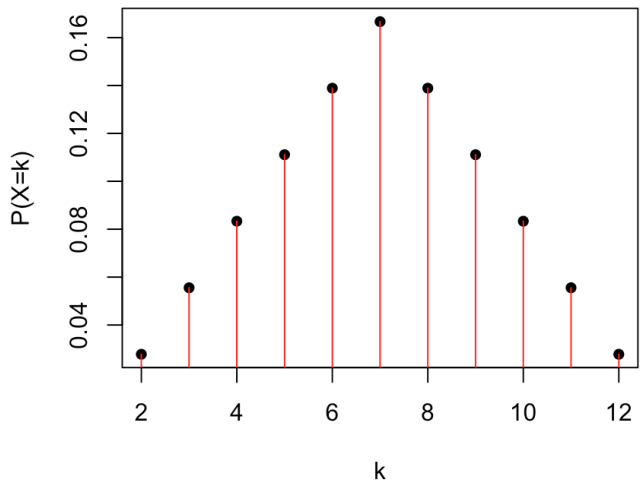
Densité et fonction de répartition d'une variables discrète

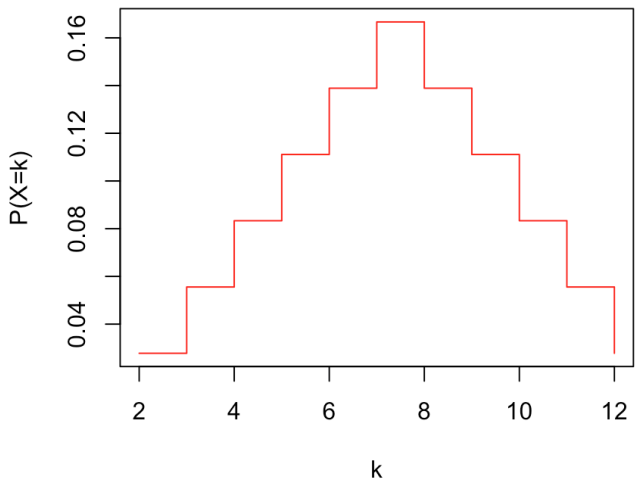
Exemple avec la variable somme des dés

- La loi de X est $k \mapsto \mathbb{P}_X(k) = \mathbb{P}(X = k) \equiv p_k$
- Le graphe (k, p_k) (à gauche) représente la fonction densité
- À droite sa représentation sous forme d'histogramme de largeur 1.

$$\mathbb{P}_X([a, b]]) = F(b) - F(a) = \sum_{i \in I} 1 \cdot \mathbb{P}(X = x_i), \quad i \in I \text{ tel que } a \leq x_i \leq b$$

Il s'agit de **l'aire des rectangles entre les abscisses a et b.**





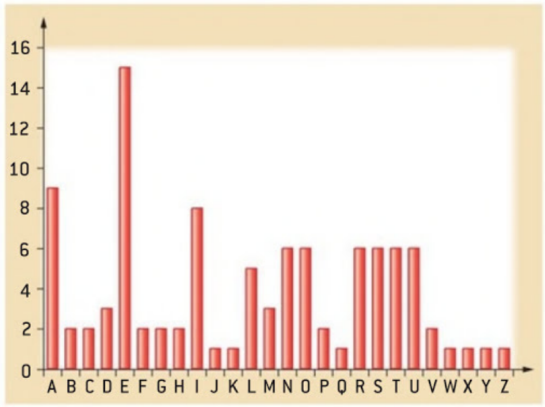
Mesure de probabilité vue comme une mesure d'aire

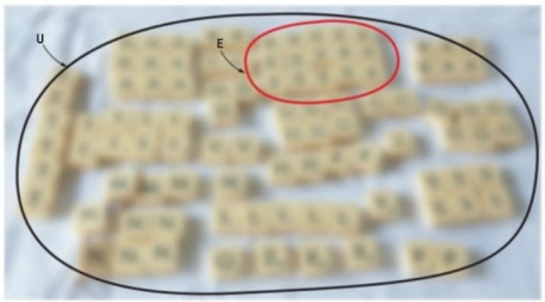
Exemple avec le jeu de scrabble

- Dans le cas du scrabble francophone (à gauche)
- La probabilité de tirer la lettre A est $\mathbb{P}(\{A\}) = \frac{9}{100}$, la lettre B est $\mathbb{P}(\{B\}) = \frac{2}{100}$, ..., $\mathbb{P}(\{Z\}) = \frac{1}{100}$ (fréquences indiquées au centre)
- On peut représenter cette probabilité sous la forme d'une aire, par exemple pour la lettre E (à droite), traduisant la proportion de E par rapport à l'ensemble de l'univers (d'aire 1).

$\Rightarrow$ Ceci explique pourquoi la théorie des probabilité est une branche de la **théorie de la mesure** !







Crédits : N. Gauvrit

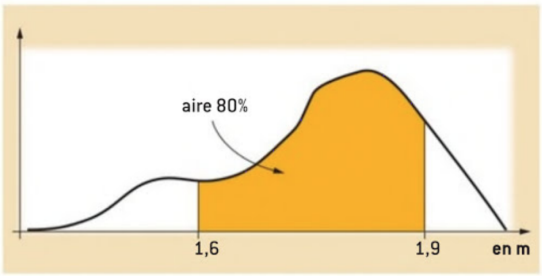
Densité et fonction de répartition d'une variables continue

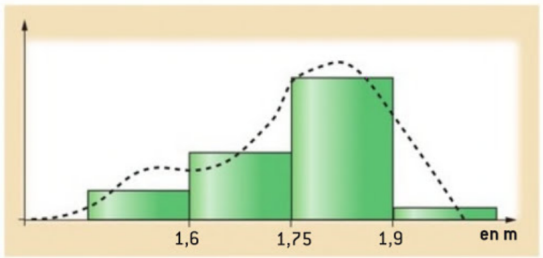
Exemple avec la taille des étudiants

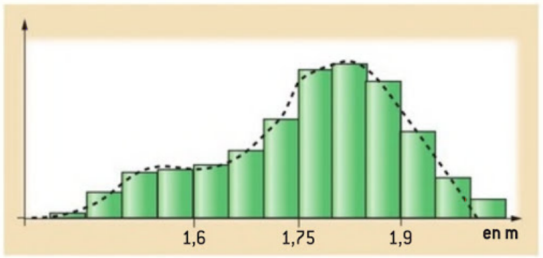
- La loi de X : $\mathbb{P}_X([a,b]) = \mathbb{P}(a \leq X \leq b)$ probabilité que la taille d'un individu soit comprise entre a et b.
  Exemple : $\mathbb{P}_X(1,6 \leq X \leq 1,9) = 0,8$.
- Fonction de répartition
  $F(x) = \mathbb{P}_X(]-\infty, x]) = \mathbb{P}(X \leq x) = \int_{-\infty}^{x} f(t)dt$ où f est la **densité de probabilité** de X.
- À droite des histogrammes approchant la densité.

$$\mathbb{P}_X([a,b]) = \mathbb{P}(a \leq X \leq b) = F(b) - F(a) = \int_a^b f(t)\mathrm{d}t$$

Il s'agit de **l'aire de la fonction f entre les abscisses a et b.**







Crédits : N. Gauvrit

CQFR : Fonction de répartition et densité de probabilité

Fonction de répartition d'une variable discrète

soit X une v.a. discrète pouvant prendre les valeurs $x_1, x_2, \dots x_n$ avec les probabilités $p_i = \mathbb{P}(X = x_i)$

$$F_X(x) = \mathbb{P}(X \leq x) = \sum_{x_i \leq x} \mathbb{P}(X = x_i) = \sum_{x_i \leq x} p_i$$

Fonction de répartition d'une variable continue

soit X une v.a. continue, alors elle est caractérisée par une **densité de probabilité** f_X telle que :

$$F_X(x) = \int_{-\infty}^{x} f_X(u) \, \mathrm{d}u$$

CQFR : Fonction de répartition et densité de probabilité

- **X variable aléatoire discrète $\rightarrow$ distribution de probabilités**

  l'ensemble des m probabilités associés aux m modalités de X

$$p_i = \mathbb{P}(X = i)$$

- **X variable aléatoire continue $\rightarrow$ densité de probabilités**

  Pour simplifier, on suppose f_X continue et définie sur $]-\infty, +\infty[$

  - $f_X(x) \geq 0, \forall x \in \mathbb{R}$
  - $\int_{-\infty}^{+\infty} f_X(x)\, \mathrm{d}x = 1$
  - $\mathbb{P}(a \leq X \leq b) = F_X(b) - F_X(a) = \int_a^b f_X(x)\, \mathrm{d}x$

CQFR : Densité de probabilité

Pour simplifier, on suppose f_X continue et définie sur $]-\infty, +\infty[$

- $f_X(x) \geq 0, \forall x \in \mathbb{R}$
- $\int_{-\infty}^{+\infty} f_X(x)\,\mathrm{d}x = 1$
- $\mathbb{P}(a \leq X \leq b) = F_X(b) - F_X(a) = \int_a^b f_X(x)\,\mathrm{d}x$

Espérance

Discrète vs. continue

$$\mathbb{E}(X) = \sum_i x_i \, \mathbb{P}(X = x_i) \qquad \mathbb{E}(X) = \int_{-\infty}^{+\infty} x \, f_X(x) \, \mathrm{d}x$$

Quelques propriétés : si a est une constante alors

- $\mathbb{E}(a) = a$
- $\mathbb{E}(a\,X) = a\,\mathbb{E}(X)$
- $\mathbb{E}(X + a) = \mathbb{E}(X) + a$

Propriété d'additivité : l'espérance d'une somme de variables aléatoires (indépendantes ou non) est égale à la somme de leur espérance

$$\mathbb{E}(X_1 + X_2) = \mathbb{E}(X_1) + \mathbb{E}(X_2)$$

Indépendance des variables :

$$X \text{ et } Y \text{ indépendantes} \Rightarrow \mathbb{E}(XY) = \mathbb{E}(X)\,\mathbb{E}(Y)$$

Variance et écart-type

Définition variance :

$$\mathbb{V}(X) = \mathbb{E}[(X - \mathbb{E}(X))^2] = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = \int_{-\infty}^{+\infty} (x - \mathbb{E}(X))^2 \, f_X(x) \, \mathrm{d}x$$

Exercice : montrer que $\mathbb{V}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$

Quelques propriétés : si a est une constante alors

- $\mathbb{V}(X + a) = \mathbb{V}(X)$
- $\mathbb{V}(aX) = a^2 \, \mathbb{V}(X)$
- $\mathbb{V}(X + Y) = \mathbb{V}(X) + \mathbb{V}(Y) + 2\mathrm{Cov}(X, Y)$
  $\mathrm{Cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))]$
- X et Y indépendantes, alors $\mathbb{V}(X + Y) = \mathbb{V}(X) + \mathbb{V}(Y)$

Définition écart type :

$$\sigma(X) = \sqrt{\mathbb{V}(X)}$$

Variable centrée réduite, médiane et mode

- On appelle **variable aléatoire centrée réduite** la v.a. Y construite telle que

$$Y = \frac{X - \mathbb{E}(X)}{\sigma(X)}$$

  C'est le moyen le plus classique pour normaliser une v.a. Par construction, on obtient $\mathbb{E}(Y) = 0$ et $\mathbb{V}(Y) = 1$.

- La **médiane** est la valeur correspondant au milieu de la fonction de répartition

$$\tilde{x} : F_X(\tilde{x}) = \frac{1}{2}$$

  Si la loi de la v.a. est symétrique, alors la médiane = l'espérance.

- Le **mode** d'une variable aléatoire est sa valeur la plus probable.

Fonction de répartition et fonction quantile

Le cas continu

- Soit $F : \mathbb{R} \to [0, 1]$ fonction croissante et continue avec $F(+\infty) = 1$. Son **inverse généralisée** noté F^{-1} est défini par :

$$\mathbf{F^{-1}(p) = \inf\{x \in \mathbb{R} : F(x) \geq p\}, \quad p \in [0, 1]}$$

- La fonction de répartition d'une variable aléatoire X est est croissante continue et définie par :

$$F_X(x) = \mathbb{P}(X \leq x)$$

- Son inverse généralisée F_X^{-1} s'appelle la **fonction quantile** de X et la quantité $F_X^{-1}(p)$ s'appelle le **quantile** ou **fractile** d'ordre p de X.

Lois usuelles

Lois discrètes

- loi uniforme discrète (le loto)
- loi de Bernoulli (le tirage d'une pièce)
- loi binomiale (plusieurs tirages d'une même pièce)
- loi géométrique (temps d'attente d'un premier succès)

× loi de Poisson (nombre d'éléments dans une file d'attente)

Lois continues

- loi uniforme continue
- loi normale ou gaussienne (Saint-Graal des statisticiens)
- loi exponentielle (durée de vue de circuits electroniques)

* loi du χ^2 (adéquation d'une distribution empirique à une loi donnée)

* loi de Student (tests de comparaisons, intervalles de confiance)

* *(abordées dans la Partie III)*, × *(non abordée dans ce cours)*

Loi discrète uniforme

La loi discrète uniforme est la loi qui décrit le fait que chaque valeur d'un ensemble finie de valeurs possibles a la même probabilité de se réaliser.

loi	proba.	espérance	variance
$\mathcal{U}[a,b]$	$\mathbb{P}(X) = \frac{1}{b-a+1}$ si $a \leq x \leq b$; 0 sinon	$\frac{a+b}{2}$	$\frac{(b-a+1)^2-1}{12}$

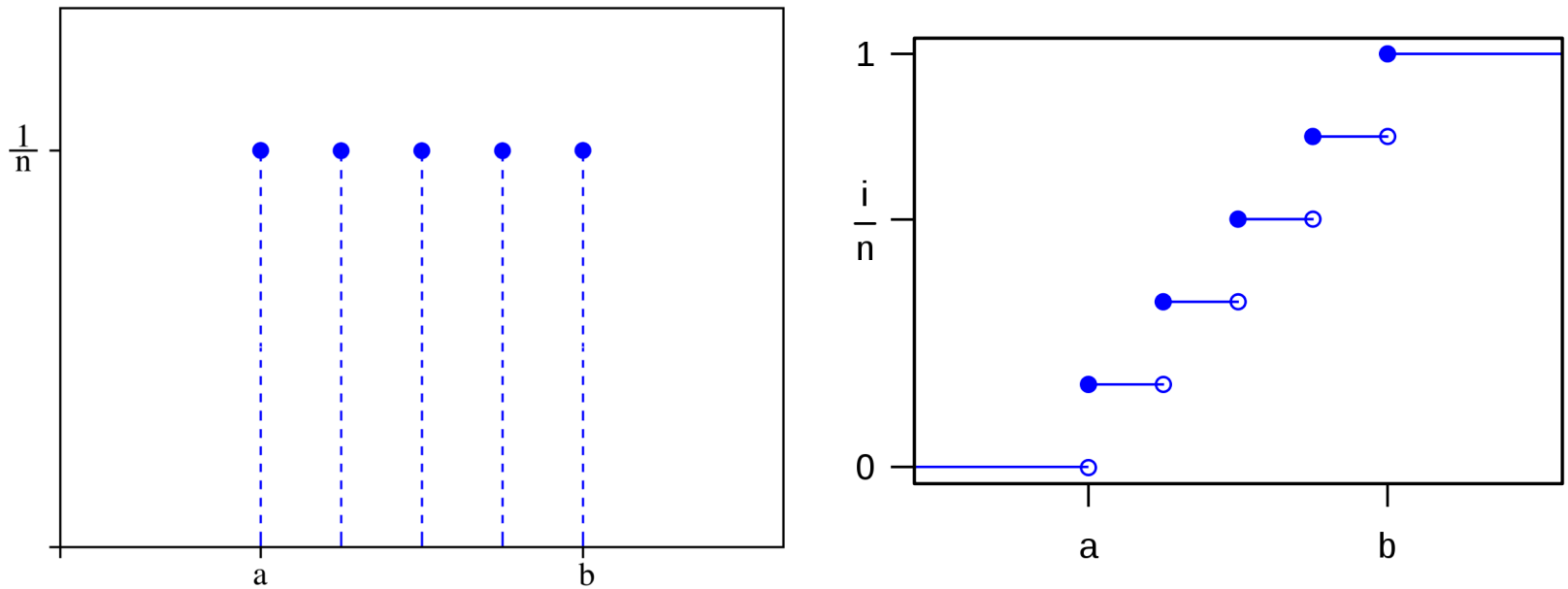


Figure: Probabilité et fonction de répartition de la loi uniforme discrète ($n = b - a + 1$)

Loi uniforme continue

loi	densité	espérance	variance
$\mathcal{U}[a,b]$	$f_X(x) = \frac{1}{b-a}$ avec $a \leq x \leq b$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$

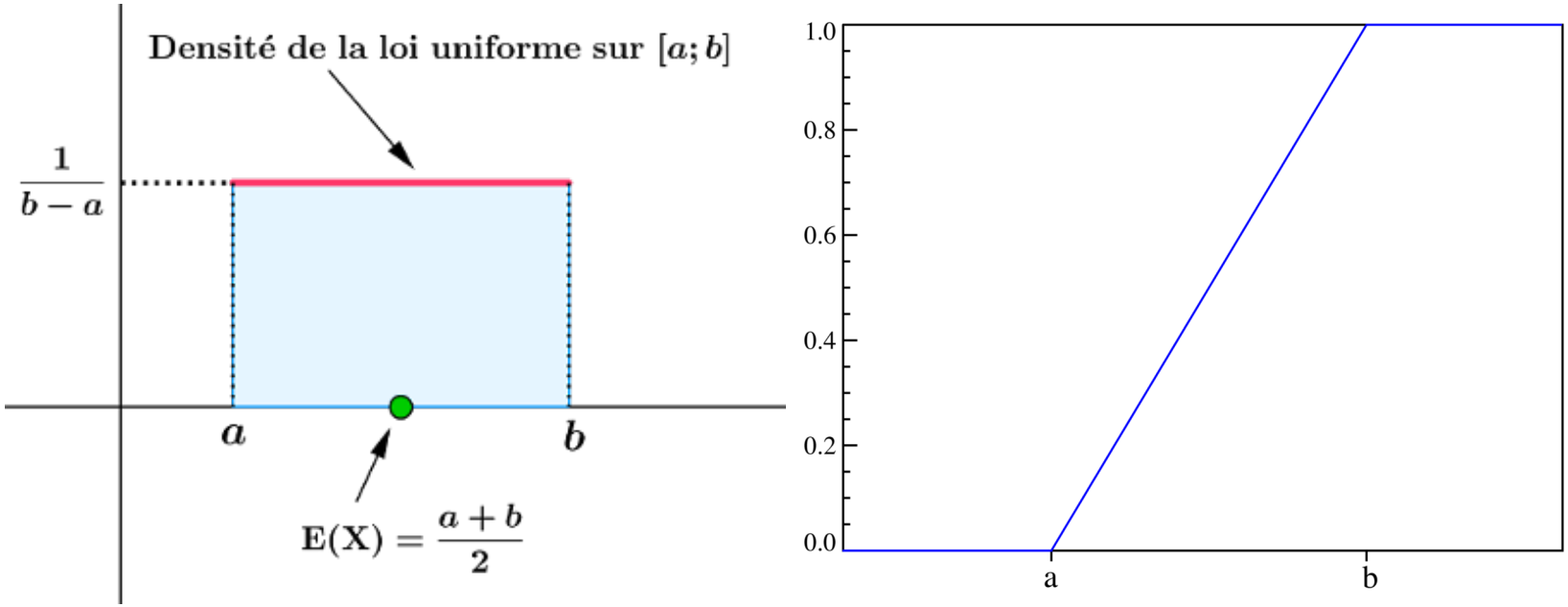


Figure: Densité de probabilité et fonction de répartition de la loi uniforme continue

Loi de Bernoulli : on tire une pièce une fois

Une épreuve de Bernoulli de paramètre p (réel compris entre 0 et 1) est une expérience aléatoire comportant deux issues :

- le succès, avec la probabilité p
- l'échec, avec la probabilité $1 - p$

loi	proba.	espérance	variance
$\mathcal{B}(p)$	$P(X = 0) = 1 - p$; $P(X = 1) = p$	p	$p(1 - p)$

Exemple 1 : Le lancer d'une pièce équilibrée est une expérience de Bernoulli de paramètre $p = 0.5$.

Exemple 2 : On tire au hasard une boule dans une urne contenant 7 boules blanches et 3 boules noires. On considère comme un succès le fait de tirer une boule noire. Cette expérience est une expérience de Bernoulli de paramètre $p = 0.3$ car la probabilité de tirer une boule noire est de 3/10.

Loi binomiale : on tire la même pièce n fois

La loi binomiale est la loi associée à n répétitions, dans des conditions iid, d'une expérience aléatoire dont l'issue est l'apparition ou non d'un évènement.

Si $X_1 .... X_n$ sont n variables de Bernoulli de paramètre p alors

$$X = \sum_{i=1}^{n} X_i \rightsquigarrow \mathcal{B}(n, p)$$

loi	proba.	espérance	variance
$\mathcal{B}(n,p)$	$P(X=k) = C_n^k \ p^k (1-p)^{(n-k)}$	np	$np(1-p)$

Exemple : On admet qu'un étudiant prend au plus un café par jour, que chaque jour sa proba de prendre un café vaut p, et qu'il y a indépendance entre ses choix quotidiens. La variable X décrivant le nombre de cafés pris par l'étudiant en une semaine est une variable aléatoire de loi $\mathcal{B}(5, p)$

Rappel : $C_n^k = \frac{n!}{k!(n-k)!}$

Loi géométrique : temps d'attente du premier succès

On considère une épreuve de Bernoulli dont la probabilité de succès est p et celle d'échec $1-p$. On renouvelle cette épreuve de manière indépendante jusqu'au premier succès. On appelle X la variable aléatoire donnant le rang du premier succès.

loi	proba.	espérance	variance
$\mathcal{G}(p)$	$P(X=k)=(1-p)^{k-1}p$	$\frac{1}{p}$	$\frac{1-p}{p^2}$

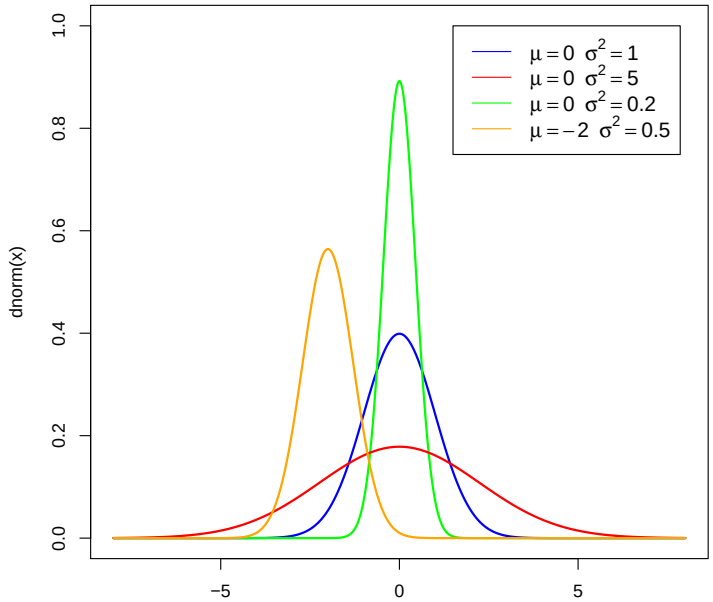
Exemple : On lance successivement un dé équilibré jusqu'à l'obtention d'un six. On pose X le nombre de lancers nécessaires. On a :

$$\mathbb{P}(X=n)=\left(\frac{5}{6}\right)^{n-1}\left(\frac{1}{6}\right),$$

et donc $X \sim \mathcal{G}(p)$ avec $p=1/6$.

Loi normale = loi gaussienne (Saint Graal des statisticiens)

loi	densité de proba.	espérance	variance
$X \sim \mathcal{N}(\mu, \sigma^2)$	$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$	μ	σ^2



La loi normale est l'une des lois de probabilité les plus adaptées pour modéliser des phénomènes naturels issus de plusieurs événements aléatoires. Elle est en lien avec de nombreux objets mathématiques dont le mouvement brownien, le bruit blanc gaussien ou d'autres lois de probabilité qu'elle approche par le théorème limite central. Elle permet la mesure d'erreur ou tests statistiques.

Loi normale = loi gaussienne

Quelques propriétés

- Loi normale centrée réduite : $\mathcal{N}(0,1)$
- Si $X \sim \mathcal{N}(\mu, \sigma^2)$ alors $\frac{X-\mu}{\sigma} \sim \mathcal{N}(0,1)$
- Si $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ et si $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$ alors

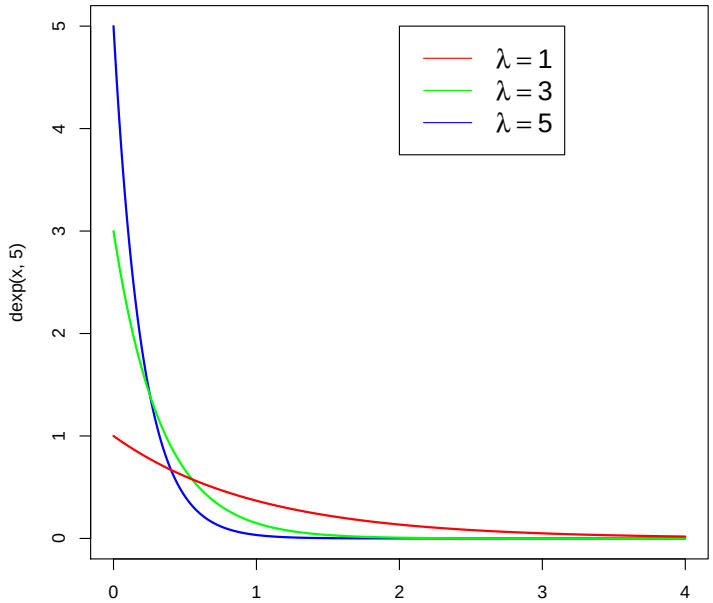
$$aX_1 + bX_2 \sim \mathcal{N}(a\mu_1 + b\mu_2, a^2\sigma_1^2 + b^2\sigma_2^2)$$

Propriété : conservation vis à vis de l'addition

Soit $\{X_i\}$ un ensemble de p v.a. normales de paramètres (μ_i, σ_i^2) indépendantes. Alors leur somme est une v.a. normale de paramètres $(\sum \mu_i, \sum \sigma_i^2)$

Loi exponentielle

loi	densité de proba.	espérance	variance
$X \sim \mathcal{E}(\lambda)$	$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & \text{si } x \geqslant 0 \\ 0 & \text{si } x < 0 \end{cases}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$



Cette loi permet entre autres de modéliser la durée de vie de la radioactivité ou d'un composant électronique. Elle peut aussi être utilisée pour décrire par exemple le temps écoulé entre deux coups de téléphone reçus au bureau, ou le temps écoulé entre deux accidents de voiture dans lequel un individu donné est impliqué.

1. Introduction

2. Rappels de probabilité
   - Axiomatique
   - Conditionnement et indépendance

3. Variables aléatoires
   - Définitions discrètes et continues
   - Fonction de densité et de répartition
   - Espérance et variance
   - Lois usuelles

4. Théorème limites
   - La loi des grands nombres
   - Théorème limite central

La loi faible des grands nombres

On considère une suite $(X_n)_{n\in\mathbb{N}^*}$ de variables aléatoires non corrélées définies sur un même espace probabilisé, ayant même variance finie $\mathbb{V}(X)$ et même espérance $\mathbb{E}(X)$. La loi faible des grands nombres stipule que, pour tout réel ε strictement positif, la probabilité que la moyenne empirique

$$Y_n \equiv \bar{x} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

s'éloigne de l'espérance d'au moins ε tend vers 0 quand $n \to +\infty$.

Théorème (loi faible des grandes nombres)

$$\forall \varepsilon > 0, \quad \lim_{n\to+\infty} \mathbb{P}\left(\left|\frac{X_1 + X_2 + \cdots + X_n}{n} - \mathbb{E}(X)\right| \geqslant \varepsilon\right) = 0.$$

Autrement dit, $(Y_n)_{n\in\mathbb{N}^*}$ converge en probabilité vers $\mathbb{E}(X)$.

Inégalité de Markov

Soit Z une variable aléatoire réelle définie sur un espace probabilisé $(\Omega, \mathcal{A}, \mathbb{P})$ et supposée presque sûrement positive ou nulle. Alors

$$\forall a > 0, \qquad \mathbb{P}(Z \geqslant a) \leqslant \frac{\mathbb{E}[Z]}{a}$$

Démo : On a l'inégalité

$$\forall \omega \in \Omega, \quad Z(\omega) \geqslant a\, 1_{\{Z(\omega) \geqslant a\}} \ ,$$

dès que $a \geqslant 0$. On en déduit que

$$\mathbb{E}[Z] \geqslant \mathbb{E}[a\, 1_{\{Z \geqslant a\}}] = a\, \mathbb{P}(Z \geqslant a)$$

Inégalité de Bienaymé–Tchebychev

Pour tout réel strictement positif α,

$$\mathbb{P}\left(|X - \mathbb{E}[X]| \geq \alpha\right) \leq \frac{\mathbb{V}(X)}{\alpha^2}$$

Démo : simple application de l'inégalité de Markov à la variable $(X - \mathbb{E}[X])^2$ et au réel α^2 strictement positif compte tenu du fait que $\{|X - \mathbb{E}[X]| \geq \alpha\} = \{(X - \mathbb{E}[X])^2 \geq \alpha^2\}$.

La loi faible des grands nombres

Preuve : de l'inégalité de Bienaymé–Tchebychev à la loi faible des grands nombres

Théorème (loi faible des grandes nombres)

$$\forall \varepsilon > 0, \quad \lim_{n \to +\infty} \mathbb{P}\left(\left| \frac{X_1 + X_2 + \cdots + X_n}{n} - \mathbb{E}(X) \right| \geqslant \varepsilon \right) = 0.$$

Démo : On a d'après l'inégalité de Bienaymé–Tchebychev :

$$\mathbb{P}(|Y - E(Y)| \geqslant \varepsilon) \leqslant \frac{V(Y)}{\varepsilon^2}$$

On remarque que la variable aléatoire $Y_n = \dfrac{X_1 + X_2 + \cdots + X_n}{n}$ a pour espérance $\mathbb{E}(X)$ et pour variance $\dfrac{\mathbb{V}(X)}{n}$. Ainsi, pour tout n :

$$\mathbb{P}\left(\left| \frac{X_1 + X_2 + \cdots + X_n}{n} - \mathbb{E}(X) \right| \geqslant \varepsilon \right) \leqslant \frac{\mathbb{V}(X)}{n\,\varepsilon^2}$$

Théorème de la limite centrée

La variable aléatoire $Y_n = \dfrac{X_1 + X_2 + \cdots + X_n}{n}$ a pour espérance $\mathbb{E}(X)$ et pour variance $\dfrac{\mathbb{V}(X)}{n}$, donc la variable aléatoire

$$Z_n = \frac{\sqrt{n}}{\sqrt{\mathbb{V}(X)}}(Y_n - \mathbb{E}(X))$$

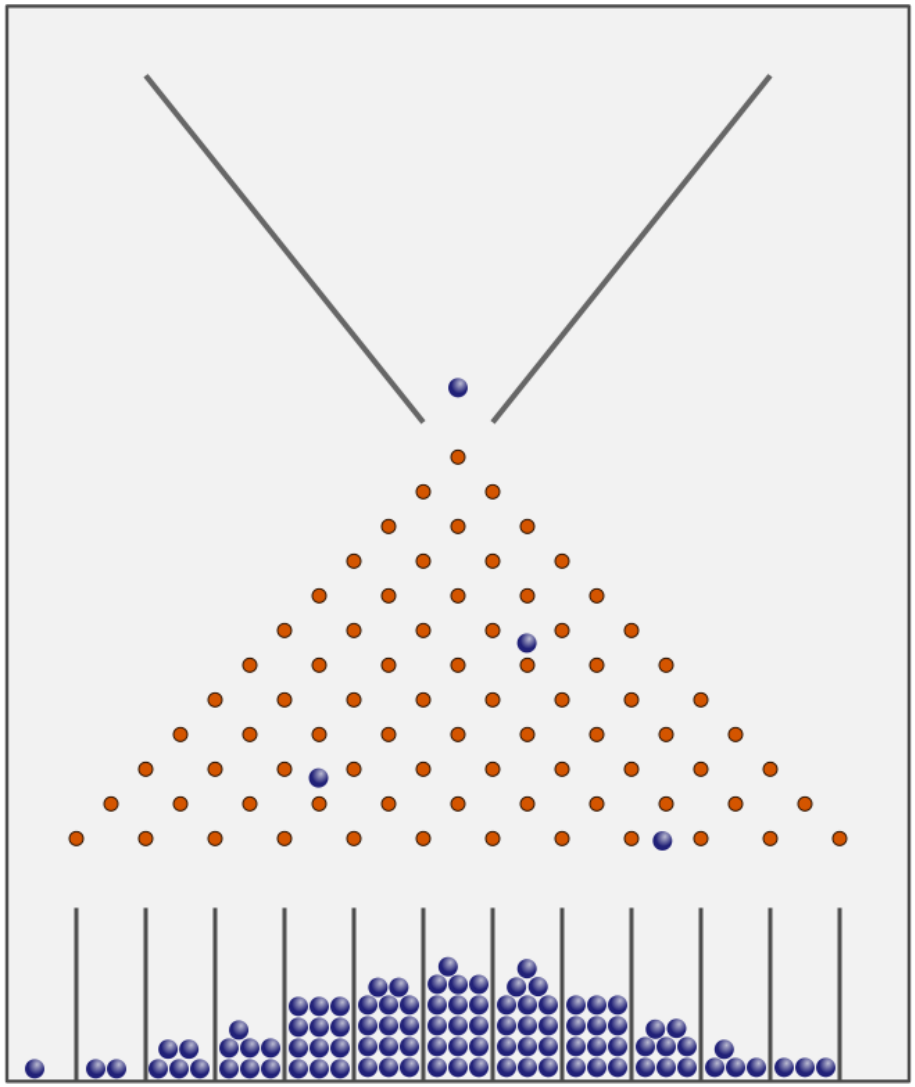
est d'espérance nulle et de variance 1.

Théorème de la limite centrée

La suite (Z_n) converge en loi vers une loi normale centrée réduite $Z \sim \mathcal{N}(0, 1)$.

Théorème de la limite centrée

Planche de Galton



Théorème de la limite centrée

Planche de Galton

Partie III : Estimation paramétrique

1. Introduction

2. Estimation ponctuelle
   - Estimateur statistique
   - Qualité d'un estimateur
   - Estimateur de l'espérance, d'une proportion et de la variance

3. Estimation par intervalle de confiance
   - Notion d'intervalle de confiance
   - Estimation par intervalle de confiance d'une proportion

1 Introduction

2 Estimation ponctuelle
- Estimateur statistique
- Qualité d'un estimateur
- Estimateur de l'espérance, d'une proportion et de la variance

3 Estimation par intervalle de confiance
- Notion d'intervalle de confiance
- Estimation par intervalle de confiance d'une proportion

Introduction

Questions abordées dans cette partie

On sait calculer des indicateurs numériques à partir d'un échantillon de données ...

- Mais comment généraliser à la population entière ?
- Quelles informations sur la population obtient-on en étudiant l'échantillon ?
- Quelle confiance peut-on accorder à ces informations ?

Introduction

Statistique inférentielle

L'idée : à partir d'échantillons représentatifs, on va introduire des résultats sur la population.

On étudie une variable X, dont on observe des réalisations. On suppose que X suit une loi connue, i.e. on choisit parmi les modèles existants la loi la plus appropriée au phénomène observé. Seule la valeur numérique du paramètre θ intervenant dans cette loi de probabilité est inconnue.

$$X \sim \mathcal{P}(\theta), \quad \theta \text{ inconnu}$$

Exemple : soit X la taille des habitants de Grenoble. On suppose que X suit une loi normale, de moyenne inconnue θ et de variance connue. On va donc chercher à estimer θ à partir d'un échantillon de données.

Introduction

Deux types d'estimation

On considère généralement deux types d'estimation :

- **L'estimation ponctuelle** : on cherche à calculer une *unique* valeur $\hat{\theta}$ estimant au mieux θ.

- **L'estimation par intervalle de confiance** : on estime la probabilité que la valeur vraie d'un paramètre appartienne à un intervalle donné, on a ainsi un ensemble de valeurs vraisemblables donc une estimation *ensembliste* ou région de confiance. Typiquement, on cherche a et b tel que, par exemple

$$\mathbb{P}(\mathbf{a} \leq \theta \leq \mathbf{b}) = \mathbf{0.95}$$

On note les données $x_1, x_2, \dots x_n$.

- On regardera x_i comme le i-ème tirage d'une variable aléatoire X :

$$X \sim \mathcal{P}(\theta)$$

- Ou de façon équivalente comme une réalisation d'une variable X_i de même loi que X. De plus, les X_i sont supposées indépendantes :

$$X_i \sim \mathcal{P}(\theta)$$

1 Introduction

2 Estimation ponctuelle
- Estimateur statistique
- Qualité d'un estimateur
- Estimateur de l'espérance, d'une proportion et de la variance

3 Estimation par intervalle de confiance
- Notion d'intervalle de confiance
- Estimation par intervalle de confiance d'une proportion

Estimation ponctuelle

Soient $x_1, x_2, \ldots, x_n$ les n valeurs prises par la v.a. X dans un échantillon de taille n prélevé dans la population-mère.

- Une statistique t est une fonction des observations $x_1, x_2, \ldots, x_n$:

$$\begin{array}{rccc} t : & \mathbb{R}^n & \to & \mathbb{R}^m \\ & (x_1, \ldots, x_n) & \mapsto & t(x_1, \ldots, x_n) \end{array}$$

- Un estimateur de θ est une fonction construite à l'aide des $\{X_i\}$:

$$T_n = t(X_1, \ldots, X_n)$$

- Une estimation est une réalisation $t_n = t(x_1, \ldots, x_n)$ de l'estimateur T_n. On note la valeur numérique de cette estimation par

$$\hat{\theta} = t(x_1, \ldots, x_n)$$

Estimation ponctuelle

Exemple

1. On observe un phénomène de production de pièces manufacturées. Chaque pièce est associée à une mesure (un indicateur de qualité par exemple). Comme on ne peut pas vérifier chaque mesure, on procède à un échantillonnage qui nous fournit donc un échantillon.
2. Supposons que la connaissance de la nature de cet indicateur nous permet de faire l'hypothèse qu'il obéit à une loi de probabilité normale.
3. Le problème est maintenant, au vue de l'échantillon $\{x_i\}$, de proposer une valeur pour la moyenne de cette loi normale. Il faut procéder à une estimation du paramètre vrai θ qui se traduit par la valeur $\hat{\theta}$. Il y a une infinité de manière possible parmi lesquelles :
   - $\hat{\theta} =$ la moyenne,
   - $\hat{\theta} =$ la médiane,
   - $\hat{\theta} =$ le mode,
   - $\hat{\theta} = x_{29}$, ...

   $\Rightarrow$ Quel est le meilleur estimateur de la moyenne ? Existe-t-il ? Qu'est ce que cela veut dire **meilleur** ?

Qualité d'un estimateur

Il n'existe pas de "meilleur estimateur" ! mais il existe des critères de comparaison :

Biais	On souhaite que l'estimation ne soit pas systématiquement décalée par rapport à la valeur vraie.
Précision	Si l'on répète l'estimation sur un autre échantillon, on souhaite obtenir une estimation cohérente, donc peu de variation d'un échantillon à l'autre. On parlera aussi d'efficacité.
Convergence	Si l'on peut estimer la valeur du paramètre sur toute la population-mère, la valeur de l'estimation obtenue doit être la valeur vraie du paramètre.
Complexité	Toute estimation nécessite un calcul donc un temps. On s'attachera donc à évaluer la complexité du calcul en fonction de la taille des données.
Robustesse	Dans tout cas concret, il existe des sources de perturbations. On souhaite que l'estimation ne soit pas sensible à la présence de valeurs abérantes.

Estimation ponctuelle

- **Estimateur sans biais**

  Un estimateur T_n est dit sans biais lorsque son espérance mathématique est égale à la valeur vraie du paramètre.

$$\text{Biais} = \mathbb{E}(T_n) - \theta = 0$$

- **Précision d'un estimateur**

  On mesure généralement la précision d'un estimateur par l'erreur quadratique moyenne :

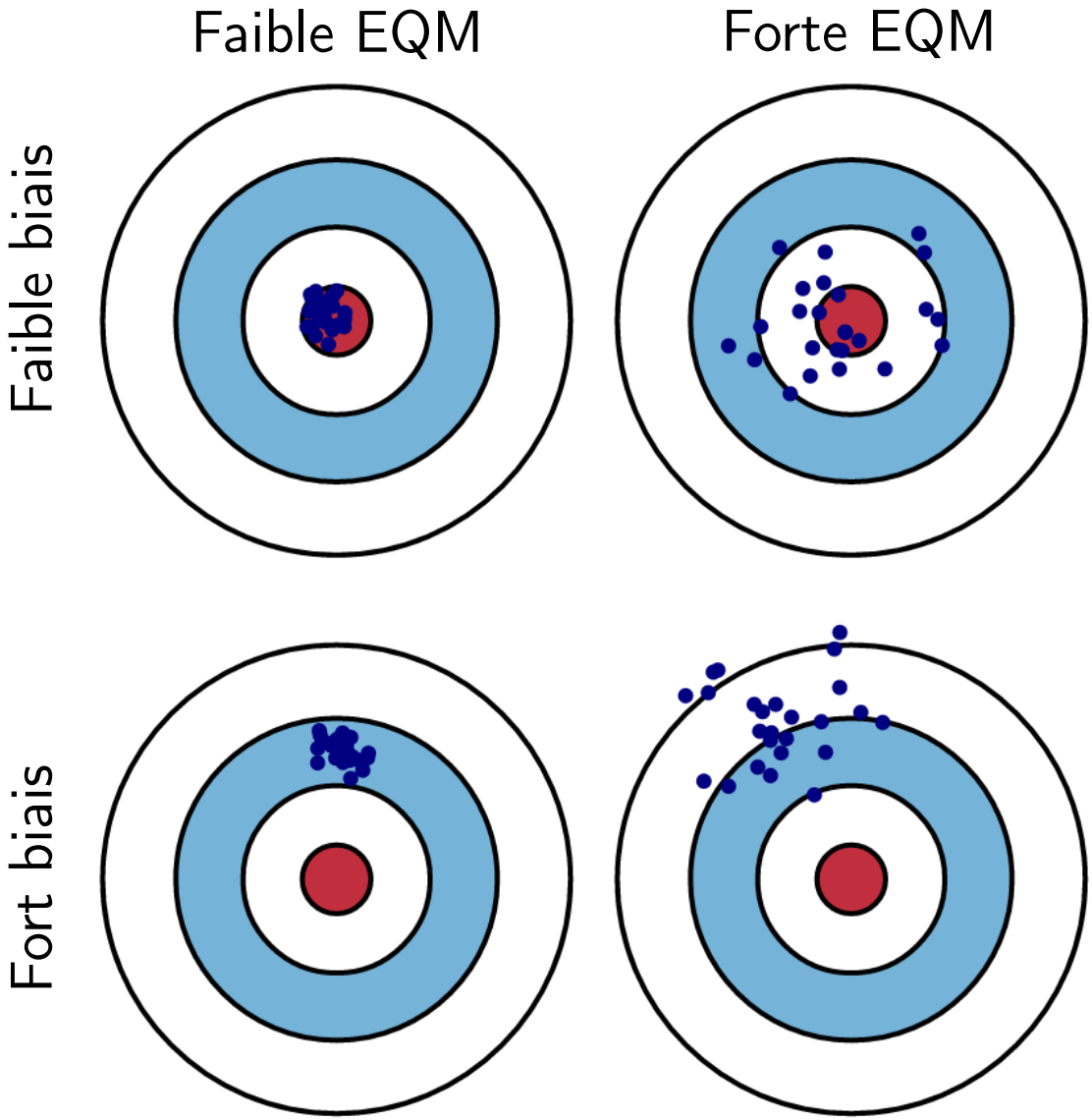
$$\text{EQM}(T_n) = \mathbb{E}((T_n - \theta)^2) = \mathbb{V}(T_n) + [\mathbb{E}(T_n) - \theta]^2$$

- **Estimateur convergent**

  Un estimateur T_n de θ est convergent en moyenne quadratique si

$$\mathbb{E}((T_n - \theta)^2) \to 0 \textbf{ quand } n \to \infty$$

Illustration



Crédits : http://www.cs.cornell.edu/courses/cs4780/2015fa/web/lecturenotes/lecturenote12.html

Estimation d'une espérance

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

est un estimateur sans biais et convergent de l'espérance

$$\theta = \mathbb{E}[X]$$

- La moyenne est un estimateur sans biais :

$$\mathbb{E}[\bar{X}_n] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X] = \theta$$

- La moyenne est un estimateur convergent :

$$\mathbb{V}[\bar{X}_n] = \mathbb{V}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\mathbb{V}\left[\sum_{i=1}^{n} X_i\right] \overset{(*)}{=} \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}[X_i] = \frac{1}{n}\mathbb{V}[X] \xrightarrow[n\to\infty]{} 0$$

$(*)$ car les X_i sont indépendants.

Estimation d'une proportion

Un estimateur fournit par la loi des grands nombres

La fréquence empirique d'un évènement
est une estimation de sa probabilité

Preuve : Soit $f_n(i)$ la fréquence de la valeur x_i dans l'échantillon de taille n $(X_1, \dots, X_n)$, $B_k = \mathbb{1}_{(X_k = x_i)}$ et $p_i = \mathbb{P}(X = x_i)$. Ainsi la suite (B_k) est constituée de loi de Bernoulli indépendantes de paramètre p_i, de variance finie et d'espérance commune $\mathbb{E}(B_k) = p_i$, d'où d'après la loi des grands nombres :

$$f_n(i) = \frac{B_1 + \cdots + B_n}{n} \xrightarrow[n \to +\infty]{} p_i$$

$\Rightarrow$ L'estimation d'une proportion peut être vu comme un problème d'estimation de moyenne.

Estimation d'une proportion

Un estimateur sans biais convergent

La fréquence empirique d'un évènement
est une estimation de sa probabilité

- f_n est un estimateur sans biais :

$$\mathbb{E}[f_n] = \mathbb{E}\left[\frac{1}{n}\sum_{k=1}^{n} B_k\right] = \frac{1}{n}\mathbb{E}\left[\sum_{k=1}^{n} B_k\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[B_k] = \frac{1}{n}\sum_{i=1}^{n} p_i = p_i$$

- f_n est un estimateur convergent :

$$\mathbb{V}[f_n] = \mathbb{V}\left[\frac{1}{n}\sum_{k=1}^{n} B_k\right] = \frac{1}{n^2}\mathbb{V}\left[\sum_{k=1}^{n} B_k\right] \overset{(*)}{=} \frac{1}{n^2}\sum_{k=1}^{n}\mathbb{V}[B_k] = \frac{p_i(1-p_i)}{n} \to 0$$

$(*)$ car les X_i sont indépendants.

Estimation d'une variance

On définit :

$$V_n = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 = \frac{1}{n}\sum X_i^2 - \mu^2$$

Si μ est connue, alors V_n est un estimateur **sans biais** de $\mathbb{V}[X]$

Preuve :

$$\begin{aligned}\mathbb{E}[\mathbb{V}] &= \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n} X_i^2\right] - \mathbb{E}[\mu^2] \\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i^2] - \mu^2 \\ &= \mathbb{E}[X_i^2] - \mu^2 \equiv \mathbb{V}[X]\end{aligned}$$

Estimation d'une variance

On définit :

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

Si μ est inconnue, alors S_n^2 est un estimateur **biaisé** de $\mathbb{V}(X)$

Preuve :

$$\begin{aligned}
\mathbb{E}[S_n^2] &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i^2] - \mathbb{E}[\bar{X}^2] \\
&= \frac{1}{n}\sum_{i=1}^{n}(\mathbb{V}[X_i] + \mathbb{E}[X_i]^2) - \mathbb{V}[\bar{X}] - \mathbb{E}[\bar{X}]^2 \\
&= \frac{1}{n}(n\mathbb{V}[X] + n\mathbb{E}[X_i]^2) - \frac{1}{n}\mathbb{V}[X] - \mathbb{E}[X_i]^2 \\
&= \mathbb{V}[X] - \frac{1}{n}\mathbb{V}[X] = \frac{n-1}{n}\mathbb{V}[X]
\end{aligned}$$

Estimation d'une variance

On définit :

$$S_n'^2 = \frac{n}{n-1} S_n^2$$

Si μ est inconnue, alors $S_n'^2$ est un estimateur **sans biais** de $\mathbb{V}(X)$

Preuve :

$$\begin{aligned}
\mathbb{E}[S_n'^2] &= \frac{n}{n-1}\mathbb{E}[S_n^2] \\
&= \frac{n}{n-1}\frac{n-1}{n}\mathbb{V}[X] \\
&= \mathbb{V}[X]
\end{aligned}$$

Estimation d'une variance

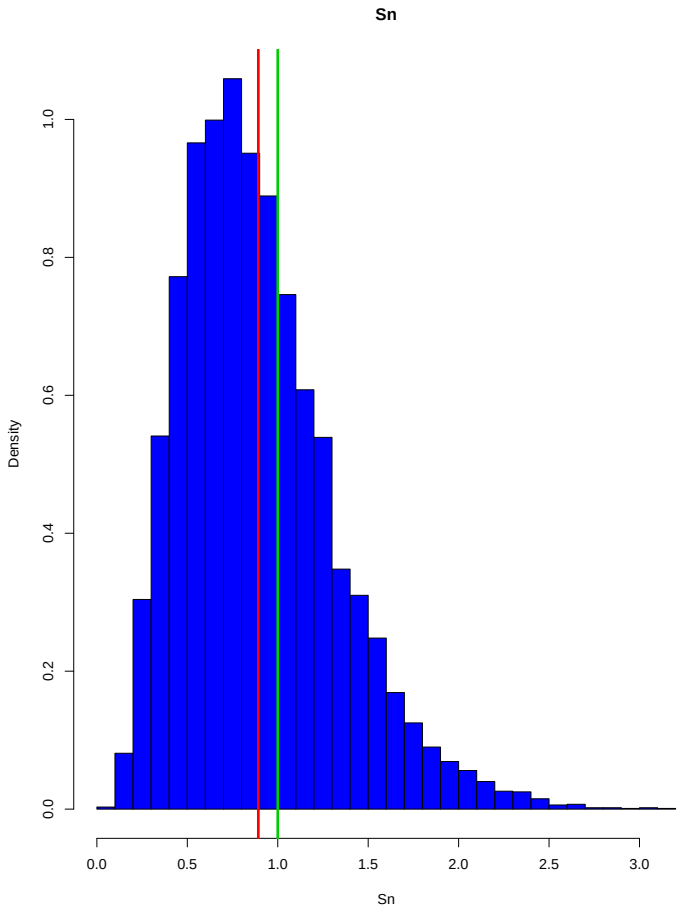
$$S_n'^2 = \frac{n}{n-1} S_n^2$$

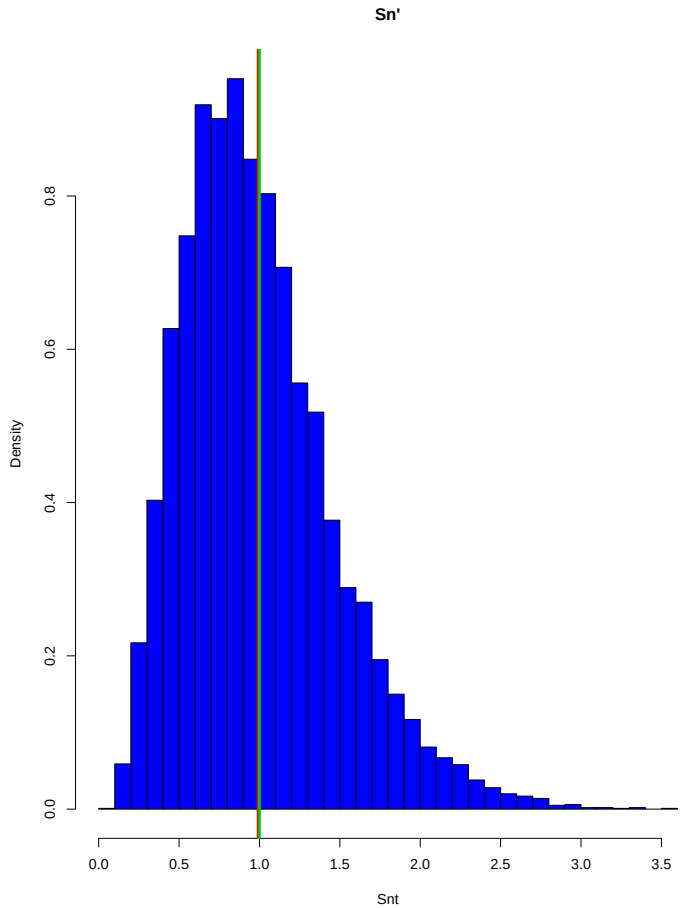
```
K <- 10000
n <- 10
mu=0
sigma2=1
# On considère K échantillons composés de n valeurs issues de N(mu,sigma2)
# réparties sur chaque ligne de la matrice Xkn suivante
Xkn <- matrix(rnorm(n*K,mean=mu,sd=sqrt(sigma2)),ncol=n)
# On applique la variance selon les lignes, donnant K variances
Snt <- apply(Xkn,FUN=var,MARGIN=1)
Sn <- (n-1)/n*Snt
# S est la version biaisée (car var utilise la variance débiaisée)
hist(Sn,nclass=30,probability=TRUE,main="Sn",col="blue")
# la variance est décalée à n/(n-1)*sigma2=0.9
abline(v=mean(Sn),col="red") ; abline(v=sigma2,col=3)
hist(Snt,nclass=30,probability=TRUE,main="Sn'",col="blue")
abline(v=mean(Snt),col="red") ; abline(v=sigma2,col=3)
```

Estimation d'une variance

Illustration du biais

$$S_n'^2 = \frac{n}{n-1} S_n^2, \quad \mathbb{E}[S_n^2] = \frac{n-1}{n}\mathbb{V}[X], \quad \mathbb{E}[S_n'^2] = \mathbb{V}[X]$$





Exercice

On a mesuré le poids de raisins produits par souche sur 10 souches prises au hasard dans la vigne. On a obtenu les résultats suivants exprimés en kilogrammes :

$$2.4 \quad 3.4 \quad 3.6 \quad 4.1 \quad 4.3 \quad 4.7 \quad 5.4 \quad 5.9 \quad 6.5 \quad 6.9$$

1. Calculer la moyenne et la variance de l'échantillon.
2. En déduire les estimation ponctuelles non biaisées de la moyenne et de la variance de la population dont sont extraites les souches.

1 Introduction

2 Estimation ponctuelle
- Estimateur statistique
- Qualité d'un estimateur
- Estimateur de l'espérance, d'une proportion et de la variance

3 Estimation par intervalle de confiance
- Notion d'intervalle de confiance
- Estimation par intervalle de confiance d'une proportion

Estimation par intervalle de confiance

La "fourchette"

Considérons un vote avec un assez grand nombre d'électeurs. Quand le scrutin est clot, on commence à dépouiller les bulletins. Assez vite on est en mesure de donner une estimation du résultat final. En pratique, on ne donne pas une estimation numérique (telle liste obtient 18% des votes) mais une fourchette, c'est-à-dire un petit intervalle dans lequel on estime que le pourcentage exact figure.

- La **taille de la fourchette** dépend de la confiance qu'on souhaite avoir dans l'estimation.
- On peut vouloir que la probabilité que le pourcentage exact d'une liste soit bien dans la fourchette dépasse 0.95 (**le niveau de confiance**).
- Plus on exige un haut niveau de confiance, plus la fourchette sera large.

Estimation par intervalle de confiance

Notion d'intervalle de confiance

Il est souvent plus réaliste et plus intéressant de fournir un renseignement de type $a < \theta < b$ plutôt que de calculer $\hat{\theta}$.

On cherche à determiner l'intervalle $[a, b]$, centré sur la valeur numérique estimée du paramètre inconnu θ, contenant la valeur vraie avec une probabilité $1 - \alpha$ $(0 \leq \alpha \leq 1)$:

$$\mathbb{P}(a < \theta < b) = 1 - \alpha$$

- L'intervalle $[\mathbf{a}, \mathbf{b}]$ est appelé intervalle de confiance, α le risque et $\mathbf{1 - \alpha}$ le niveau de confiance.

- Données de depart : **l'échantillon** et la connaissance de **la loi de probabilité du paramètre** à estimer.

Estimation par intervalle de confiance d'une proportion

Soit une population dont les individus possèdent un caractère A avec une probabilité p. On dispose d'un échantillon de taille n, dont x individus possèdent le caractère A.

- On sait maintenant que la proportion $\mathbf{f_n = x/n}$ est une estimation de la valeur vraie $\mathbf{p}$ ...
- Mais avec quelle confiance ?
- On cherche donc à construire un intervalle de confiance de l'estimateur.

Crédits : O. Gaudoin

Estimation par intervalle de confiance d'une proportion

Soit une population dont les individus possèdent un caractère A avec une probabilité $\mathbf{p}$. On dispose d'un échantillon de taille $\mathbf{n}$, dont $\mathbf{x}$ individus possèdent le caractère A. La proportion $\mathbf{f_n = x/n}$ est une estimation de la valeur vraie $\mathbf{p}$.

Principe

- Soit $\mathbf{F_n} = \frac{1}{\mathbf{n}} \sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{X_i}$. $\mathbf{F_n}$ est une v.a. construite comme somme de n v.a indépendantes de type Bernoulli et de paramètre $\mathbf{p}$, i.e $\mathbf{X_i} \sim \mathcal{B}(\mathbf{p})$.
- La loi de $\mathbf{T_n = nF_n}$ suit une loi binomiale $\mathcal{B}(\mathbf{n}, \mathbf{p})$ ...
- La loi de $\mathbf{T_n}$ tend vers une loi normale de moyenne $\mathbf{np}$ et de variance $\mathbf{np(1-p)}$ (*cf.* TP 2, approximation valide si $np > 10$ et $n(1-p) > 10$)
- La variable renormalisée approche une loi normale centrée réduite :

$$\frac{\mathbf{T_n - np}}{\sqrt{\mathbf{np(1-p)}}} \sim_\infty \mathcal{N}(\mathbf{0}, \mathbf{1})$$

Crédits : O. Gaudoin

Estimation par intervalle de confiance d'une proportion

On écrit alors

$$\mathbb{P}\left(\left|\frac{\mathbf{T_n - np}}{\sqrt{\mathbf{np(1-p)}}}\right| \leq \mathbf{u}_\alpha\right) \approx \mathbf{1} - \alpha \ ,$$

où u_α est une valeur (se lisant dans la table de la loi normale $\mathcal{N}(0,1)$) qui vérifie :

$$\mathbb{P}(|U| > u_\alpha) = \alpha, \quad U \sim \mathcal{N}(0,1) \ .$$

Pour en déduire un intervalle de confiance, il suffit d'écrire $\left|\frac{T_n - np}{\sqrt{np(1-p)}}\right| \leq u_\alpha$ sous la forme $Z_1 \leq p \leq Z_2$:

$$\left|\frac{T_n - np}{\sqrt{np(1-p)}}\right| \leq u_\alpha \iff \frac{(T_n - np)^2}{np(1-p)} \leq u_\alpha^2$$

$$\iff p^2(n + u_\alpha^2) - p(2T_n + u_\alpha^2) + \frac{T_n^2}{n} \leq 0$$

Crédits : O. Gaudoin

Estimation par intervalle de confiance d'une proportion

Intervalle de confiance asymptotique

Le trinôme $p^2(n+u_\alpha^2) - p(2T_n + u_\alpha^2) + \frac{T_n^2}{n}$ est toujours positif sauf entre ses racines. Donc ses racines sont les bornes de l'intervalle de confiance recherché :

$$\left[\frac{\frac{T_n}{n} + \frac{u_\alpha^2}{2n} - u_\alpha \sqrt{\frac{u_\alpha^2}{4n^2} + \frac{T_n(n-T_n)}{n^3}}}{1 + \frac{u_\alpha^2}{n}}, \frac{\frac{T_n}{n} + \frac{u_\alpha^2}{2n} + u_\alpha \sqrt{\frac{u_\alpha^2}{4n^2} + \frac{T_n(n-T_n)}{n^3}}}{1 + \frac{u_\alpha^2}{n}} \right]$$

Pour les valeurs usuelles de α et pour n grand, on peut négliger u_α^2 par rapport à n. D'où, avec $F_n = \frac{T_n}{n}$ et une réalisation f_n de F_n, on obtient **l'intervalle de confiance asymptotique** suivant :

$$\left[f_n - u_\alpha \sqrt{\frac{f_n(1-f_n)}{n}}, f_n + u_\alpha \sqrt{\frac{f_n(1-f_n)}{n}} \right]$$

Crédits : O. Gaudoin

Estimation par intervalle de confiance d'une proportion

Fourchette du sondage

Une élection oppose deux candidats A et B. Un institut de sondage interroge **800 personnes** sur leurs intentions de vote :

- 420 déclarent voter pour A
- 380 déclarent voter pour B

Estimer le résultat de l'élection, c'est estimer le pourcentage p de voix qu'obtiendra A le jour de l'élection, en inférant sur l'ensemble de la population. L'estimateur de p est la proportion $f_n = \frac{420}{800} = 52.5\%$. L'institut de sondage estime donc que le candidat A va gagner l'élection. Mais pour évaluer l'incertitude, on a besoin d'un intervalle de confiance de seuil disons 5% pour p. On obtient alors l'intervalle de confiance asymptotique suivant

$$[0.4904, 0.5596]$$

Conclusion : on a une confiance de 95% dans le fait que le pourcentage de voix qu'obtiendra le candidat A sera compris entre 49% et 56%.

Crédits : O. Gaudoin

Estimation par intervalle de confiance d'une proportion

Obtenir une prédiction plus précise

À quelle condition l'intervalle de confiance pour p sera entièrement situé au dessus de 50% ?

$\Rightarrow$ Il s'agit donc de réduire l'intervalle de confiance, de largeur :

$$\ell = 2u_\alpha \sqrt{\frac{f_n(1-f_n)}{n}}$$

Pour diminuer cette largeur ℓ, on peut :

- Diminuer u_α, c'est-à-dire augmenter α, donc augmenter la probabilité de se tromper en affirmant que le candidat est élu ;
- Augmenter n, c'est-à-dire augmenter le nombre de personnes interrogées.

Crédits : O. Gaudoin

Estimation par intervalle de confiance d'une proportion

Taille de l'échantillon minimum

À un seuil de confiance $\alpha = 5\%$ fixé, combien de personnes n doit-on interroger pour que l'intervalle de confiance n'excède pas une largeur ℓ ?

On sait que $\forall p \in [0,1], p(1-p) \leq \frac{1}{4}$, donc

$$2u_\alpha \sqrt{\frac{f_n(1-f_n)}{n}} \leq \frac{u_\alpha}{\sqrt{n}}$$

Ainsi il suffit de déterminer n tel que

$$\frac{\mathbf{u_\alpha}}{\sqrt{\mathbf{n}}} < \ell \Longleftrightarrow \mathbf{n} > \frac{\mathbf{u_\alpha^2}}{\ell^2}$$

Pour $n = 800$, on a $\frac{u_\alpha}{\sqrt{n}} = \frac{1.96}{\sqrt{800}} \approx 7.5\%$ $\Rightarrow$ la précision sur l'estimation de p est donc avec une confiance de 95% de plus ou moins 3.5%, ce qu'on a constaté avec l'intervalle $[49\%, 56\%]$.

Crédits : O. Gaudoin

Estimation par intervalle de confiance d'une proportion

Taille de l'échantillon minimum

À un seuil de confiance $\alpha = 5\%$ fixé, combien de personnes n doit-on interroger pour que l'intervalle de confiance n'excède pas une largeur ℓ ?

On sait que $\forall p \in [0,1], p(1-p) \leq \frac{1}{4}$, donc

$$2u_\alpha \sqrt{\frac{f_n(1-f_n)}{n}} \leq \frac{u_\alpha}{\sqrt{n}}$$

Ainsi il suffit de déterminer n tel que

$$\frac{\mathbf{u_\alpha}}{\sqrt{\mathbf{n}}} < \ell \iff \mathbf{n} > \frac{\mathbf{u_\alpha^2}}{\ell^2}$$

Si on veut, avec le même niveau de confiance, avoir une précision $<$ à 1%, il faudra interroger au moins :

$$n = \frac{u_\alpha^2}{\ell^2} = \frac{1.96^2}{0.01^2} = 38\,416 \text{ personnes}$$

Crédits : O. Gaudoin

Exercice

Afin d'étudier l'influence des rayons X sur la spermatogénèse de Bombyx Mori (vers à soie sous sa forme papillon), on a irradié des mâles au 2ème jour et au 4ème jour du stade larvaire. Ces mâles ont été accouplés avec des femelles non irradiées. On a compté le nombre d'œufs fertiles dans la ponte des femelles, et on a obtenu :

nombre d'œufs totals	nombre d'œufs fertiles
5646	4998

1. Donner l'estimation du pourcentage d'œufs fertiles
2. Calculer un intervalle de confiance approximatif du pourcentage d'œufs fertiles au niveau de confiance 0.9. On donne $u_{0.1} = 1.6448$.

Rappel théorème de la limite centrée

- Si $X_1, \ldots, X_n$ sont indépendantes et de même loi partageant d'espérance μ et de variance σ^2, l'estimateur sans biais de l'espérance μ est la moyenne empirique $\bar{X}_n$
- Le théorème central limite assure que pour n suffisamment grand :

$$\overline{\mathbf{X}}_{\mathbf{n}} \simeq \mathcal{N}\left(\mu, \frac{\sigma^2}{\mathbf{n}}\right)$$

Rappel théorème de la limite centrée

Pour une loi normale

- Si $X_1, \ldots, X_n$ sont indépendantes et de même loi normale $\mathcal{N}(\mu, \sigma^2)$, l'estimateur sans biais de variance minimale (ESBVM) de μ est la moyenne empirique $\bar{X}_n$
- Les propriétés élémentaires de la loi normale permettent d'établir :

$$\sum_{i=1}^{n} X_i \sim \mathcal{N}(n\mu, n\sigma^2) \Rightarrow \mathbf{\overline{X}_n} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{\mathbf{n}}\right)$$

Caractérisation de la loi de l'estimateur $\bar{X}_n$

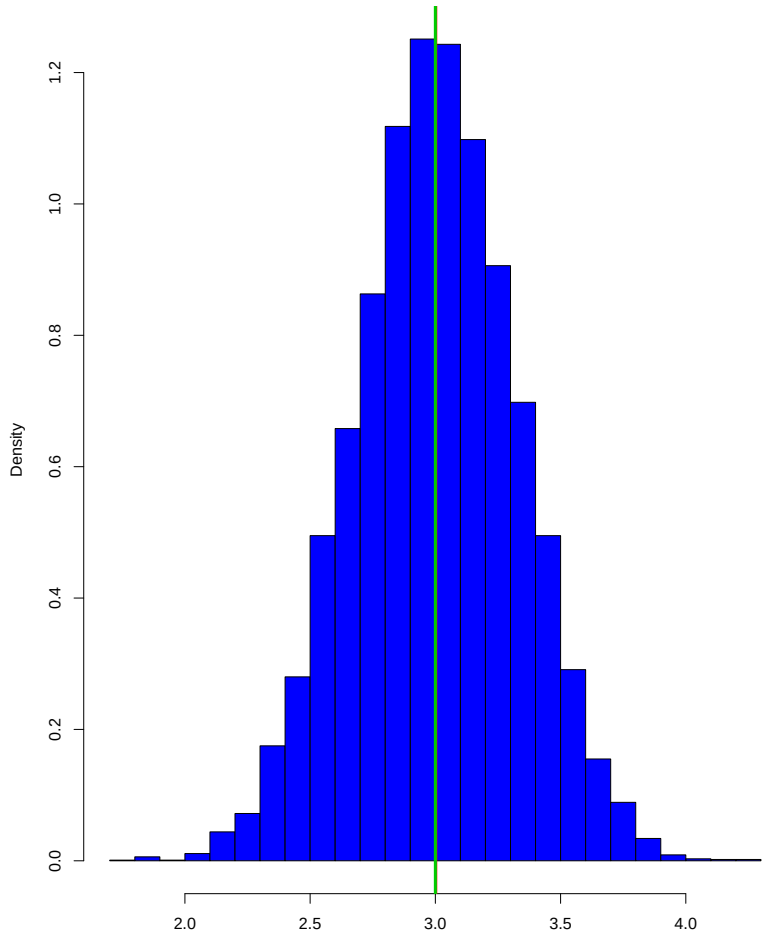
Pour une loi normale

```
K <- 10000
n <- 10
mu=3
sigma2=1
# On considère K échantillons composés de n valeurs issues de N(mu,sigma2)
# réparties sur chaque ligne de la matrice Xkn suivante
Xkn <- matrix(rnorm(n*K,mean=mu,sd=sqrt(sigma2)),ncol=n)
# On applique la moyenne selon les lignes, donnant K moyennes
Mn <- apply(Xkn,FUN=mean,MARGIN=1)
hist(Mn,nclass=30,probability=TRUE,main="Mn",col="blue")
abline(v=mean(Mn),col="red") ; abline(v=mu,col=3)
```

Caractérisation de la loi de l'estimateur $\bar{X}_n$

Pour une loi normale, effet de la taille de l'échantillon $n = 10$

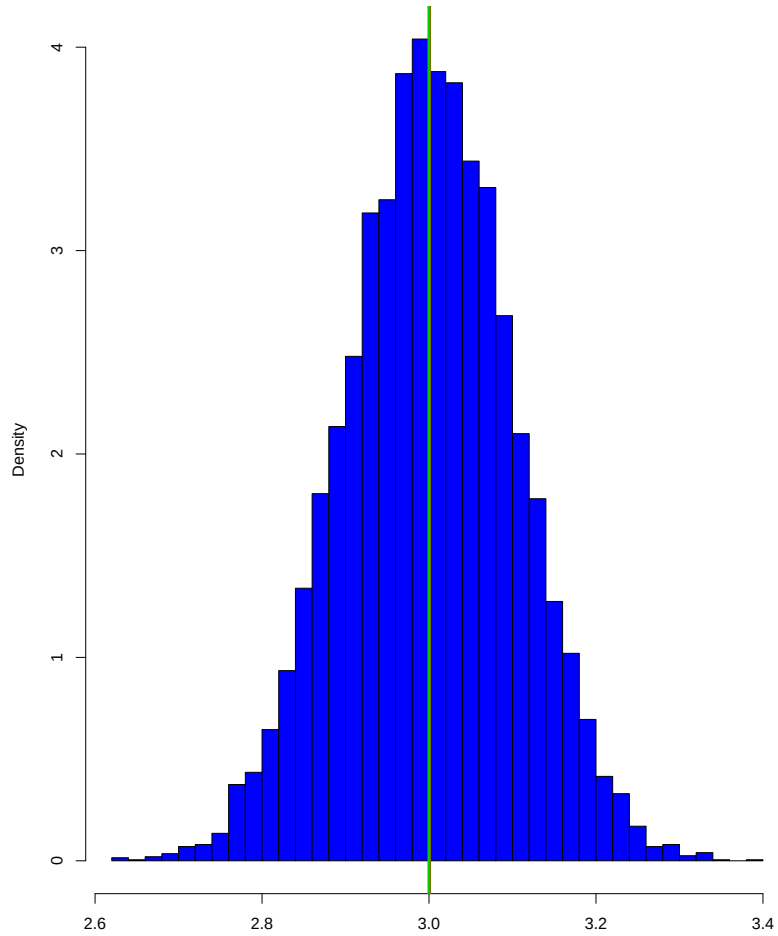
$$\sum_{i=1}^{n} X_i \sim \mathcal{N}(n\mu, n\sigma^2) \Rightarrow \overline{\mathbf{X}}_\mathbf{n} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{\mathbf{n}}\right)$$



Caractérisation de la loi de l'estimateur $\bar{X}_n$

Pour une loi normale, effet de la taille de l'échantillon $n = 100$

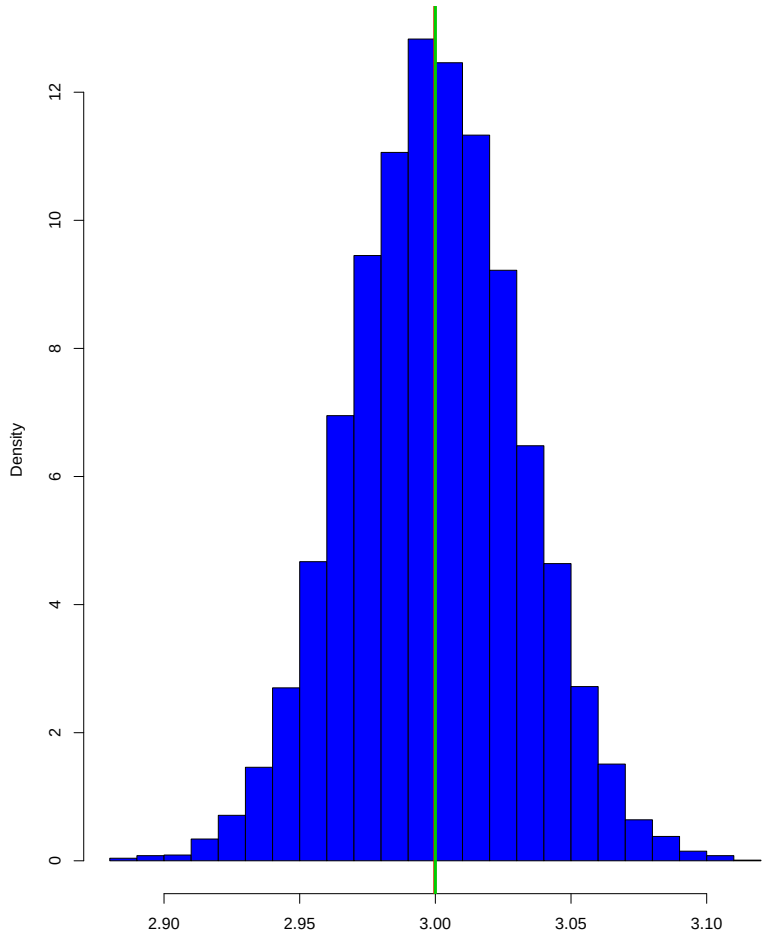
$$\sum_{i=1}^{n} X_i \sim \mathcal{N}(n\mu, n\sigma^2) \Rightarrow \overline{\mathbf{X}}_{\mathbf{n}} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{\mathbf{n}}\right)$$



Caractérisation de la loi de l'estimateur $\bar{X}_n$

Pour une loi normale, effet de la taille de l'échantillon $n = 1000$

$$\sum_{i=1}^{n} X_i \sim \mathcal{N}(n\mu, n\sigma^2) \Rightarrow \overline{\mathbf{X}}_\mathbf{n} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{\mathbf{n}}\right)$$



Estimation par intervalle de confiance de la moyenne

Pour une loi normale, lorsque la variance σ est connue

- Si $X_1, \ldots, X_n$ sont indépendantes et de même loi normale $\mathcal{N}(\mu, \sigma^2)$, l'estimateur sans biais de variance minimale (ESBVM) de μ est la moyenne empirique $\bar{X}_n$
- Les propriétés élémentaires de la loi normale permettent d'établir :

$$\sum_{i=1}^{n} X_i \sim \mathcal{N}(n\mu, n\sigma^2) \Rightarrow \mathbf{\overline{X}_n} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{\mathbf{n}}\right)$$

$$\Rightarrow \mathbf{U} = \frac{\mathbf{\overline{X}_n} - \mu}{\sqrt{\sigma^2/\mathbf{n}}} \sim \mathcal{N}(\mathbf{0}, \mathbf{1})$$

- On cherche un intervalle de confiance pour μ de la forme $[\bar{X}_n - \epsilon, \bar{X}_n + \epsilon]$, soit pour α fixé :

$$\mathbb{P}(|\bar{X}_n - \mu| \leq \epsilon) = 1 - \alpha$$

Crédits : O. Gaudoin

Estimation par intervalle de confiance de la moyenne

Pour une loi normale, lorsque la variance σ est connue

- Si $X_1, \ldots, X_n$ sont indépendantes et de même loi normale $\mathcal{N}(\mu, \sigma^2)$, l'estimateur sans biais de variance minimale (ESBVM) de μ est la moyenne empirique $\bar{X}_n$
- On cherche un intervalle de confiance pour μ de la forme $[\bar{X}_n - \epsilon, \bar{X}_n + \epsilon]$, soit pour α fixé :

$$1 - \alpha = \mathbb{P}(|\bar{X}_n - \mu| \leq \epsilon)$$

$$1 - \alpha = \mathbb{P}\left(|U| \leq \frac{\sqrt{n}\epsilon}{\sigma}\right)$$

$$1 - \mathbb{P}\left(|U| > u_\alpha\right) = 1 - \mathbb{P}\left(|U| > \frac{\sqrt{n}\epsilon}{\sigma}\right)$$

- On en déduit $u_\alpha = \frac{\sqrt{n}\epsilon}{\sigma}$ soit $\epsilon = \frac{\sigma}{\sqrt{n}} u_\alpha$, d'où l'intervalle de confiance :

$$\left[\overline{\mathbf{X}}_\mathbf{n} - \frac{\sigma}{\sqrt{\mathbf{n}}}\mathbf{u}_\alpha, \overline{\mathbf{X}}_\mathbf{n} + \frac{\sigma}{\sqrt{\mathbf{n}}}\mathbf{u}_\alpha\right]$$

Crédits : O. Gaudoin

Loi du χ^2

Définition :

Une v.a réelle suit une loi du χ^2_p (chi-deux) à p degrés de liberté si elle se réalise comme une somme

$$\chi^2_p = \sum_{i=1}^{p} U_i^2 \ ,$$

où $U_1, \ldots, U_p$ sont p variables normales $\mathcal{N}(0,1)$ indépendantes.

La densité de χ^2_p est donnée par

$$f_p(t) = \frac{1}{2^{p/2}\Gamma(p/2)} \exp(-t/2) t^{p/2-1} \ ,$$

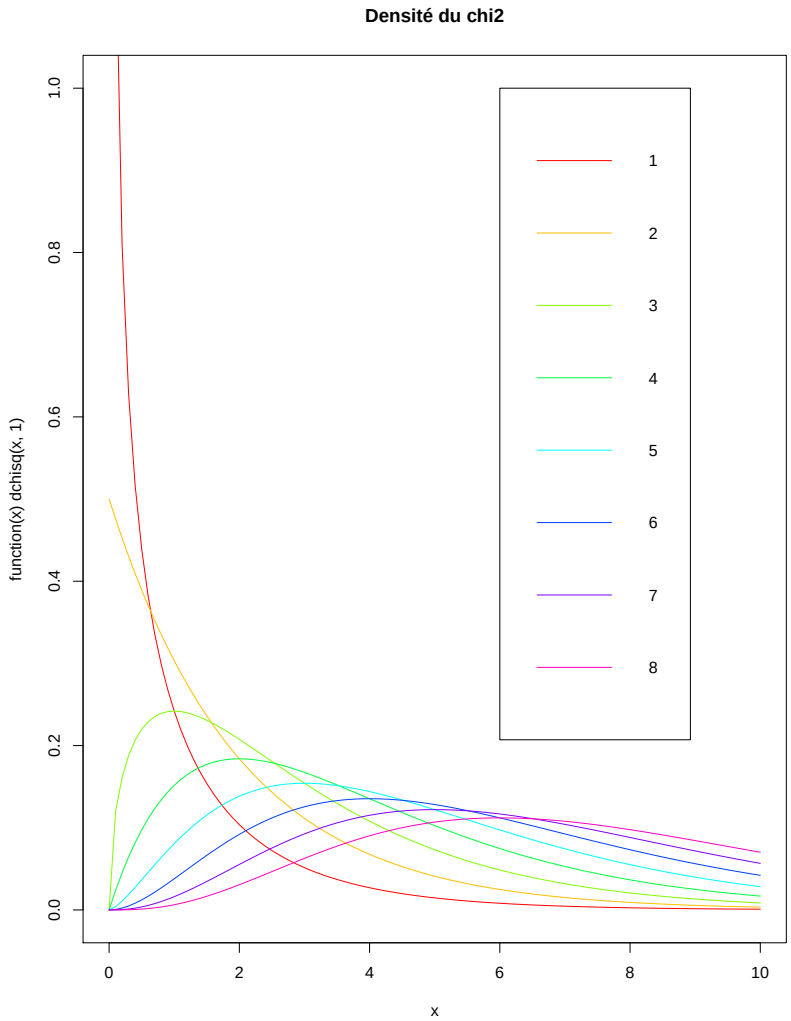
où

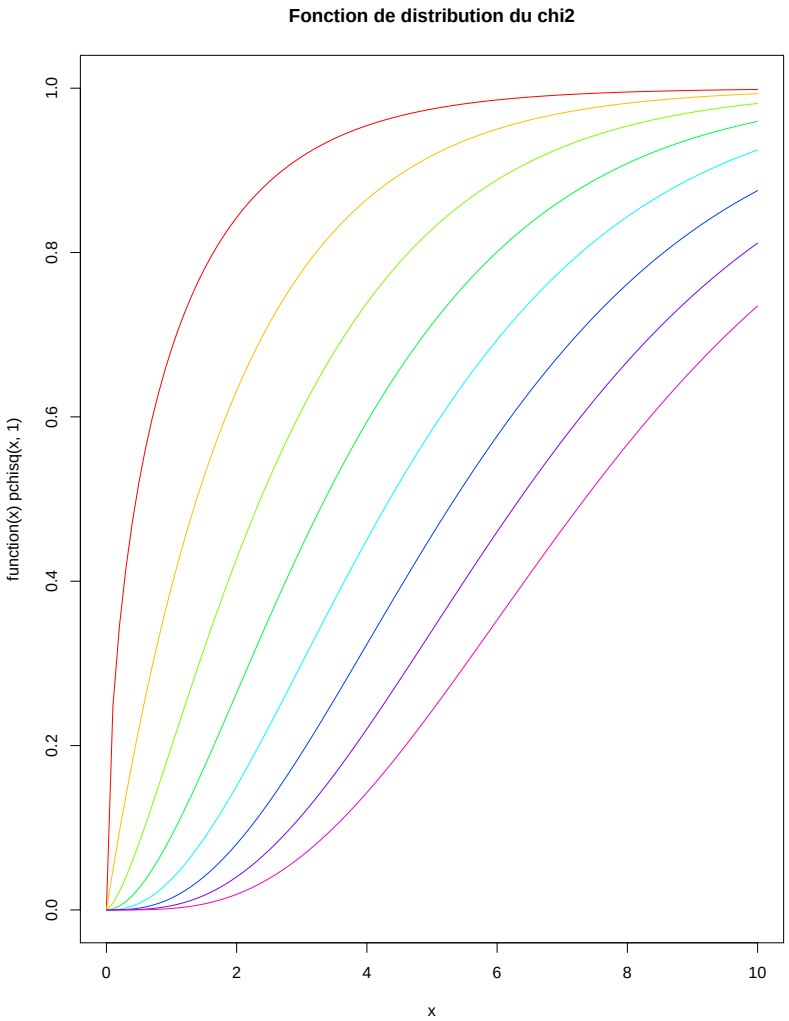
$$\Gamma : x \in \mathbb{R}^+ \mapsto \int_0^{+\infty} t^{x-1} \exp(-t) \mathrm{d}t \ ,$$

et son espérance vaut p et sa variance $2p$.

Loi du χ^2_p

Illustration de la densité et de la fonction de répartition selon p





Théorème de Fischer

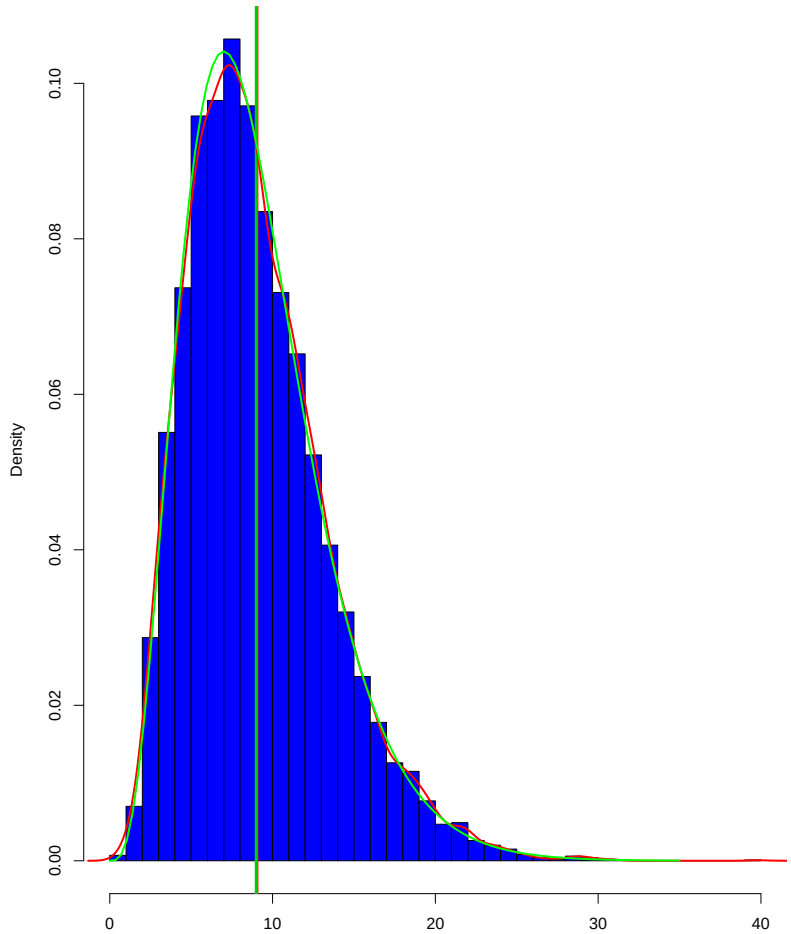
Si $X_1, \ldots, X_n$ sont indépendantes et de même loi normale $\mathcal{N}(\mu, \sigma^2)$, alors

$$\frac{\mathbf{nS_n^2}}{\sigma^2} = \frac{\mathbf{(n-1)S_n'^2}}{\sigma^2} \sim \chi^2_{\mathbf{n-1}}$$

Caractérisation de la loi de l'estimateur S'_n

```r
K <- 10000
n <- 10
mu=0
sigma2=1
# On considère K échantillons composés de n valeurs issues de N(mu,sigma2)
# réparties sur chaque ligne de la matrice Xkn suivante
Xkn <- matrix(rnorm(n*K,mean=mu,sd=sqrt(sigma2)),ncol=n)
# On applique la variance selon les lignes, donnant K variances
St <- apply(Xkn,FUN=var,MARGIN=1)
Ki <- (n-1)/sigma2*St
# On calcule l'histogramme qui doit approcher une loi chi2(n-1)
hist(Ki,nclass=30,probability=TRUE,main="Chi")
# On affiche les moyennes empiriques et vraies
abline(v=mean(Ki),col="red",lwd=2) ;
abline(v=n-1,col=3,lwd=2)
# On affiche les courbes de densités approchées et vraies
lines(density(Ki),col="red")
plot(function(x) dchisq(x,n-1),0,35,main="Chi square density
    ", col="green",ylim=c(0,1),xlim=c(0,10),add=TRUE)
```

Caractérisation de la loi de l'estimateur S'_n



Estimation par intervalle de confiance de la variance

Pour une loi normale

$$\mathbb{P}\left(a \leq \frac{nS_n^2}{\sigma^2} \leq b\right) = \mathbb{P}\left(\frac{nS_n^2}{b} \leq \sigma^2 \leq \frac{nS_n^2}{a}\right)$$
$$= F_{\chi^2_{n-1}}(b) - F_{\chi^2_{n-1}}(a)$$

$\Rightarrow$ Il y a une infinité de façons possibles de choisir a et b de sorte à ce que cette probabilité soit égale à $1-\alpha$. Si on choisit :

$$F_{\chi^2_{n-1}}(b) = 1 - \frac{\alpha}{2}, \quad F_{\chi^2_{n-1}}(a) = \frac{\alpha}{2}$$

alors, avec $z_{n-1,\alpha}$ dans la table de χ^2_{n-1}, on a

$$\mathbb{P}(Z > z_{n-1,\alpha}) = 1 - F_{\chi^2_n}(z_{n-1,\alpha}) = \alpha \ ,$$

et ainsi les valeurs suivantes conviennent :

$$b = z_{n-1,\alpha/2}, \quad a = z_{n-1,1-\alpha/2}$$

Crédits : O. Gaudoin

Estimation par intervalle de confiance de la variance

Pour une loi normale

Un intervalle de confiance de seuil α pour le paramètre σ^2 de la loi $\mathcal{N}(\mu, \sigma^2)$ est :

$$\left[\frac{nS_n^2}{z_{n-1,\alpha/2}}, \frac{nS_n^2}{z_{n-1,1-\alpha/2}}\right] = \left[\frac{(n-1)S_n'^2}{z_{n-1,\alpha/2}}, \frac{(n-1)S_n'^2}{z_{n-1,1-\alpha/2}}\right]$$

Loi de Student

Définition :

Une v.a réelle suit une loi de Student à p degrés de liberté si elle se réalise comme sous la forme

$$\frac{\mathbf{U}}{\sqrt{\frac{\chi^2_{\mathbf{p}}}{\mathbf{p}}}} \sim \mathbf{St(p)}\,,$$

où U suit une loi $\mathcal{N}(0,1)$ et χ^2_p indépendantes.

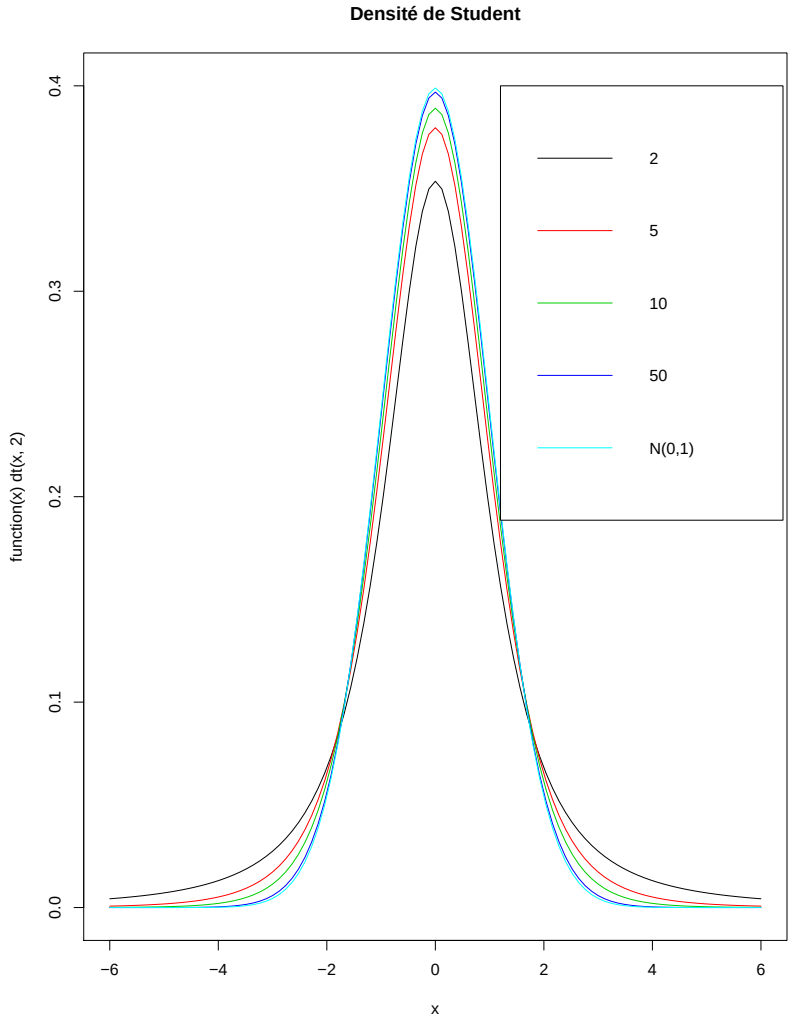
La densité de $\mathrm{St}(p)$ est donnée par

$$f_{\mathrm{St}(p)}(t) = \frac{1}{\sqrt{k\pi}} \frac{\Gamma(\frac{k+1}{2})}{\Gamma(\frac{k}{2})} \left(1 + \frac{t^2}{k}\right)^{-\frac{k+1}{2}}\,,$$

son espérance n'est pas définie pour $p = 1$ et est nulle pour $p > 1$, sa variance est infinie pour $p \leq 2$ et vaut $\frac{p}{p-2}$ pour $p > 2$

Loi de Student

Illustration de la densité et normalité asymptotique



Estimation par intervalle de confiance de la moyenne

Pour une loi normale, lorsque la variance σ est connue

- Si $X_1, \ldots, X_n$ sont indépendantes et de même loi normale $\mathcal{N}(\mu, \sigma^2)$, l'ESBVM de μ est la moyenne empirique $\bar{X}_n$
- Un intervalle de confiance pour μ de la forme $[\bar{X}_n - \epsilon, \bar{X}_n + \epsilon]$, est pour un risque α fixé :

$$\left[\overline{\mathbf{X}}_\mathbf{n} - \frac{\sigma}{\sqrt{\mathbf{n}}}\mathbf{u}_\alpha, \overline{\mathbf{X}}_\mathbf{n} + \frac{\sigma}{\sqrt{\mathbf{n}}}\mathbf{u}_\alpha\right]$$

- **Une idée naturelle est de remplacer σ par son estimateur S'_n**
- On utilise non plus $\mathbf{U} = \frac{\overline{\mathbf{X}}_\mathbf{n} - \mu}{\sqrt{\sigma^2/\mathbf{n}}} \sim \mathcal{N}(\mathbf{0}, \mathbf{1})$, mais :

$$\frac{\overline{\mathbf{X}}_\mathbf{n} - \mu}{\sqrt{\mathbf{S}'^2_\mathbf{n}/\mathbf{n}}} = \sqrt{\mathbf{n}}\frac{\overline{\mathbf{X}}_\mathbf{n} - \mu}{\mathbf{S}'_\mathbf{n}} \not\sim \mathcal{N}(\mathbf{0}, \mathbf{1})$$

Crédits : O. Gaudoin

Théorème de Fischer

Si $X_1, \ldots, X_n$ sont indépendantes et de même loi normale $\mathcal{N}(\mu, \sigma^2)$, alors

$$\sqrt{\mathbf{n}}\frac{\overline{\mathbf{X}}_\mathbf{n} - \mu}{\mathbf{S}'_\mathbf{n}} \sim \mathbf{St(n-1)}$$

Estimation par intervalle de confiance de la moyenne

Pour une loi normale, lorsque la variance σ est inconnue

- Si $X_1, \ldots, X_n$ sont indépendantes et de même loi normale $\mathcal{N}(\mu, \sigma^2)$, les estimateurs sans biais de variance minimale (ESBVM) de (μ, σ^2) sont $(\bar{X}_n, S'^2_n)$.
- On cherche un intervalle de confiance pour μ de la forme $[\bar{X}_n - \epsilon, \bar{X}_n + \epsilon]$, soit pour α fixé :

$$1 - \alpha = \mathbb{P}(|\bar{X}_n - \mu| \leq \epsilon)$$

$$1 - \alpha = \mathbb{P}\left(\left|\sqrt{n}\frac{\overline{X}_n - \mu}{S'_n}\right| \leq \frac{\sqrt{n}\epsilon}{S'_n}\right)$$

$$1 - \mathbb{P}\left(\left|\sqrt{n}\frac{\overline{X}_n - \mu}{S'_n}\right| > t_{n-1,\alpha}\right) = 1 - \mathbb{P}\left(\left|\sqrt{n}\frac{\overline{X}_n - \mu}{S'_n}\right| > \frac{\sqrt{n}\epsilon}{S'_n}\right)$$

- On en déduit $t_{n-1,\alpha} = \frac{\sqrt{n}\epsilon}{S'_n}$ soit $\epsilon = \frac{S'_n}{\sqrt{n}} t_{n-1,\alpha}$, d'où l'intervalle de confiance :

$$\left[\overline{\mathbf{X}}_\mathbf{n} - \frac{\mathbf{S}'_\mathbf{n}}{\sqrt{\mathbf{n}}}\mathbf{t}_{\mathbf{n}-\mathbf{1},\alpha}, \overline{\mathbf{X}}_\mathbf{n} + \frac{\mathbf{S}'_\mathbf{n}}{\sqrt{\mathbf{n}}}\mathbf{t}_{\mathbf{n}-\mathbf{1},\alpha}\right]$$

Crédits : O. Gaudoin

Liens avec les fonctions de distributions et quantiles

Pour une loi normale

Pour une variable aléatoire $U \sim \mathcal{N}(0,1)$ et un réel $\alpha \in [0,1]$ on a défini la valeur u_α vérifiant :

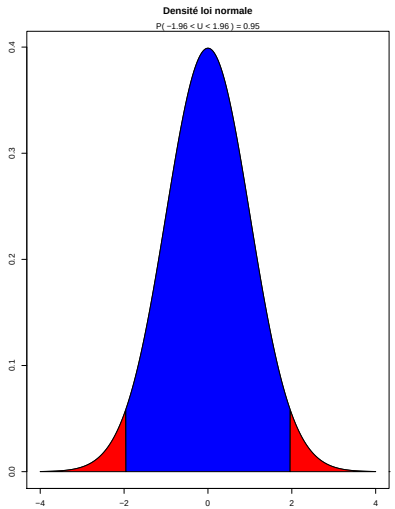
$$\mathbb{P}(|U| > u_\alpha) = \alpha$$

Par symétrie de la loi on a alors $\mathbb{P}(U > u_\alpha) = \frac{\alpha}{2}$ et $\mathbb{P}(U < -u_\alpha) = \frac{\alpha}{2}$.

$$F_U(u_\alpha) = \mathbb{P}(U \leq u_\alpha) = 1 - \mathbb{P}(U > u_\alpha) = 1 - \frac{\alpha}{2}\ ,$$

d'où $u_\alpha = F_U^{-1}(1 - \frac{\alpha}{2}) = \mathrm{qnorm}(1 - \frac{\alpha}{2})$. On a ainsi

$$\mathbb{P}(-u_\alpha \leq U \leq u_\alpha) = 1 - \alpha$$



Liens avec les fonctions de distributions et quantiles

Pour une loi du χ^2

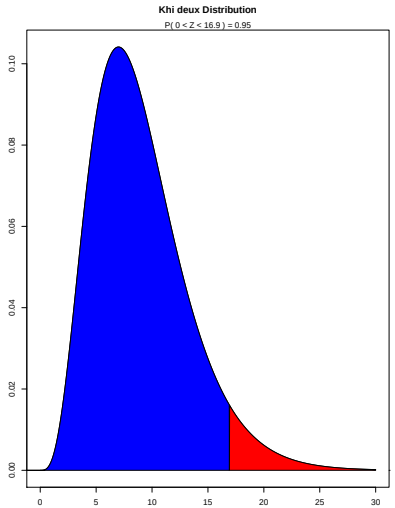
Pour une variable aléatoire $Z \sim \chi^2_n$ et un réel $\alpha \in [0,1]$ on a défini la valeur $z_{n,\alpha}$ vérifiant :

$$\mathbb{P}(Z > z_{n,\alpha}) = \alpha$$

$$F_Z(z_{n,\alpha}) = \mathbb{P}(Z \leq z_{n,\alpha}) = 1 - \mathbb{P}(Z > z_{n,\alpha}) = 1 - \alpha \ ,$$

d'où $z_{n,\alpha} = F_Z^{-1}(1-\alpha) = \text{qchisq}(1-\alpha)$. On a ainsi

$$\mathbb{P}(0 \leq Z \leq z_{n,\alpha}) = 1 - \alpha$$



Liens avec les fonctions de distributions et quantiles

Pour une loi de Student

Pour une variable aléatoire $T \sim \mathrm{St}(n)$ et un réel $\alpha \in [0,1]$ on a défini la valeur $t_{n,\alpha}$ vérifiant :

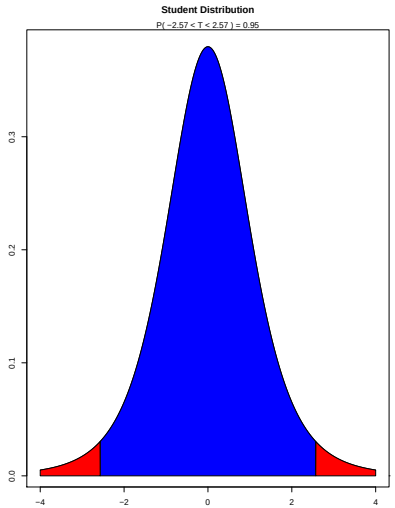
$$\mathbb{P}(|T| > t_{n,\alpha}) = \alpha$$

Par symétrie de la loi on a alors $\mathbb{P}(T > t_{n,\alpha}) = \frac{\alpha}{2}$ et $\mathbb{P}(T < -t_{n,\alpha}) = \frac{\alpha}{2}$.

$$F_T(t_{n,\alpha}) = \mathbb{P}(T \leq t_{n,\alpha}) = 1 - \mathbb{P}(T > t_{n,\alpha}) = 1 - \frac{\alpha}{2},$$

d'où $t_{n,\alpha} = F_T^{-1}(1 - \frac{\alpha}{2}) = \mathrm{qt}(1 - \frac{\alpha}{2})$. On a ainsi

$$\mathbb{P}(-t_{n,\alpha} \leq T \leq t_{n,\alpha}) = 1 - \alpha$$



Partie IV : Introduction aux tests d'hypothèse

Introduction

Jeu de Pile ou Face et triche

Karl et Ronald jouent à Pile ou Face.
Karl parie systématiquement sur Pile et Ronald sur Face.

- Au bout de 6 lancers :
  - Karl obtient 1 fois Pile
  - Ronald obtient 5 fois Face

  $\Rightarrow$ Cela vous semble-t-il suspect ?
- Ils continuent. Au bout de 18 lancers :
  - Karl obtient 4 fois Pile
  - Ronald obtient 14 fois Face

  $\Rightarrow$ Cela vous semble-t-il suspect ?

Crédits : A. Jebrane (cours statistiques L2 psychologie)

Introduction

Jeu de Pile ou Face et triche

- Au bout de 6 lancers :
  - Karl obtient 1 fois Pile
  - Ronald obtient 5 fois Face

  $\Rightarrow$ Cela vous semble-t-il suspect ? Si $X \sim \mathcal{B}(6, \frac{1}{2})$ alors

$$\mathbb{P}(\mathbf{X \geq 5}) = \mathbf{1 - pbinom(4, 6, 0.5) \approx 0.109}$$

- Ils continuent. Au bout de 18 lancers :
  - Karl obtient 4 fois Pile
  - Ronald obtient 14 fois Face

  $\Rightarrow$ Cela vous semble-t-il suspect ? Si $X \sim \mathcal{B}(18, \frac{1}{2})$ alors

$$\mathbb{P}(\mathbf{X \geq 14}) = \mathbf{1 - pbinom(13, 18, 0.5) \approx 0.015}$$

Karl a 985 chances sur 1000 de ne pas se tromper en refusant d'attribuer au hasard seul sa perte au jeu.

Crédits : A. Jebrane (cours statistiques L2 psychologie)

Formulation des hypothèses

Jeu de Pile ou Face et triche

On souhaite déterminer si Ronald est un tricheur ou est honnête. On confronte alors ces deux hypothèses :

- H_0 : Ronald est honnête, chaque lancer a une chance sur deux de faire Face. (Hypothèse nulle)
- H_1 : Ronald est un tricheur, il utilise une pièce qui a plus de chances de faire Face. (Hypothèse alternative)

Ces deux hypothèses ne jouent pas des rôles symétriques :

- la première suppose que Ronald n'a pas d'effet sur le jeu, que seul le hasard intervient ;
- tandis que la second considère qu'un processus supplémentaire (par exemple la triche, utilisation d'une pièce truquée) modifie les résultats par rapport au premier cas de figure.

Crédits : A. Jebrane (cours statistiques L2 psychologie)

Formulation mathématiques des hypothèses

Jeu de Pile ou Face et triche

Le modèle probabiliste doit permettre de voir si l'échantillon observé est une « exception » ou s'il ne diffère pas significativement de la **majorité** des autres échantillons choisis au hasard.

Soit X la v.a comptant le nombre de Face obtenues après n lancers (ici $n = 18$), les hypothèses H_0 et H_1 peuvent se réécrire comme des hypothèses sur la loi de X, ce qui se traduit par un **test paramétrique** :

- $H_0 : X \sim \mathcal{B}(n, p)$ où $p = \frac{1}{2}$. (Hypothèse nulle)
- $H_1 : X \sim \mathcal{B}(n, p)$ où $p > \frac{1}{2}$. (Hypothèse alternative)

$\Rightarrow$ p n'est connue que dans le cas H_0 où Ronald est honnête, on ne peut donc calculer explicitement de probabilité que dans le cas de l'hypothèse H_0. On dit que H_0 est **testable** et que H_1 ne l'est pas directement.

Crédits : A. Jebrane (cours statistiques L2 psychologie)

Principe du test d'hypothèse

Analogie avec un procès en justice

- On se place sous l'hypothèse H_0 (présomption d'innocence) pour voir s'il est raisonnable de maintenir cette hypothèse au vu des données observées (éléments de l'enquête).
- À l'issue du test statistique (après enquête), on pourra prendre la décision de **rejeter l'hypothèse** H_0 (de condamner Ronald) si l'on considère les résultats de l'expérience comme incompatibles avec cette l'hypothèse, jugée fortement improbable au vu des données.
- Si au contraire les résultats sont compatibles avec l'hypothèse, on dira que l'on ne rejette pas H_0 (Ronald est acquitté).

$\Rightarrow$ Cela ne signifie pas que l'on ait la certitude que H_0 soit vrai (être acquitté est différent que d'être innocent), mais que l'**on ne dispose pas d'assez de preuves pour la rejeter** (c'est-à-dire ici pour accuser Ronald).

Crédits : A. Jebrane (cours statistiques L2 psychologie)

Risques d'erreur

Quand on prend des décisions en se basant sur des tests statistiques, on n'est pas à l'abri de commettre des erreurs. Elles sont de deux types :

- **Erreur de type I** : Rejeter à tord H_0, cela revient à accuser un innocent (erreur judiciaire).
- **Erreur de type II** : Accepter à tord H_0, cela revient à innocenter un coupable.

Etat \ Décision	Accepter H_0	Rejeter H_0
H_0 vraie	Pas d'erreur	Erreur de type I
H_1 vraie	Erreur de type II	Pas d'erreur

Crédits : A. Jebrane (cours statistiques L2 psychologie)

Risques d'erreur

- **Erreur de type I** : Probabilité de rejeter H_0 alors que H_0 est vraie :

$$\alpha = \mathbb{P}(\text{rejeter } H_0 \,|\, H_0 \text{ est vraie})$$

Proposition contraire : accepter H_0 alors que H_0 est vraie (vraisemblance) :$1 - \alpha = \mathbb{P}(\text{accepter } H_0 \,|\, H_0 \text{ est vraie})$

- **Erreur de type II** : Probabilité d'accepter H_0 alors que H_1 est vraie :

$$\beta = \mathbb{P}(\text{accepter } H_0 \,|\, H_1 \text{ est vraie})$$

Proposition contraire : rejeter H_0 lorsque que H_1 est vraie (puissance) :$1 - \beta = \mathbb{P}(\text{rejeter } H_0 \,|\, H_1 \text{ est vraie})$

Etat \ Décision	Accepter H_0	Rejeter H_0
H_0 vraie	Pas d'erreur (proba $1-\alpha$)	Erreur (proba α)
H_1 vraie	Erreur (proba β)	Pas d'erreur (proba $1-\beta$)

Crédits : A. Jebrane (cours statistiques L2 psychologie)

Région de rejet

La prise de décision se fera en fonction de l'appartenance des données observées à une certaine région de valeurs. Ici, on a envie :

- D'accuser Ronald de tricherie si le nombre de Faces est très élevé.
- De ne pas l'accuser si le nombre de Faces est raisonnable.

On cherche donc une région, que l'on notera W_α appelée **région critique**, composée de valeurs élevées, dans laquelle on a peu de chances de tomber si jamais H_0 est vraie :

$$\mathbb{P}(X \in W_\alpha \,|\, H_0) \leq \alpha$$

On choisit de **rejeter** H_0 dans cette région.

Crédits : A. Jebrane (cours statistiques L2 psychologie)

Région de rejet

Dans notre exemple, on prend par exemple $\alpha = 0.05$ et on rejette H_0 si le nombre de Faces observé est trop grand au niveau α, c'est-à-dire s'il est plus grand qu'une valeur seuil k_α qui dépend du risque d'erreur que l'on est prêt à accepter.

Pour trouver cette région W_α la plus grande possible, on doit chercher tous les k tels que

$$\mathbb{P}(X \geq k) \leq \alpha$$

et prendre la plus petite parmi elles. Par exemple pour $X \sim \mathcal{B}(6, \frac{1}{2})$ on a

$$\mathbb{P}(X \geq 6) = 0.0156, \quad \mathbb{P}(X \geq 5) = 0.109$$

donc $W_\alpha = \{6\}$, tandis que pour $X \sim \mathcal{B}(18, \frac{1}{2})$ on a

$$W_\alpha = \{13, 14, 15, 16, 17, 18\}$$

Crédits : A. Jebrane (cours statistiques L2 psychologie)

Notion de p-valeur

Si on prend un risque $\alpha = 0.01$ alors la région critique sera

$$W_\alpha = \{15, 16, 17, 18\}$$

qui ne contient pas la valeur observée 14. Donc **entre les deux niveaux de risques il y a une valeur α où on change de décision**, **cette valeur s'appelle la p-valeur.**

Dans notre exemple, pour quel niveau α a-t-on

$$W_\alpha = \{14, 15, 16, 17, 18\} \ ?$$

On obtient

$$\boxed{\alpha = \mathbb{P}(X \geq 14) = 0.015}$$

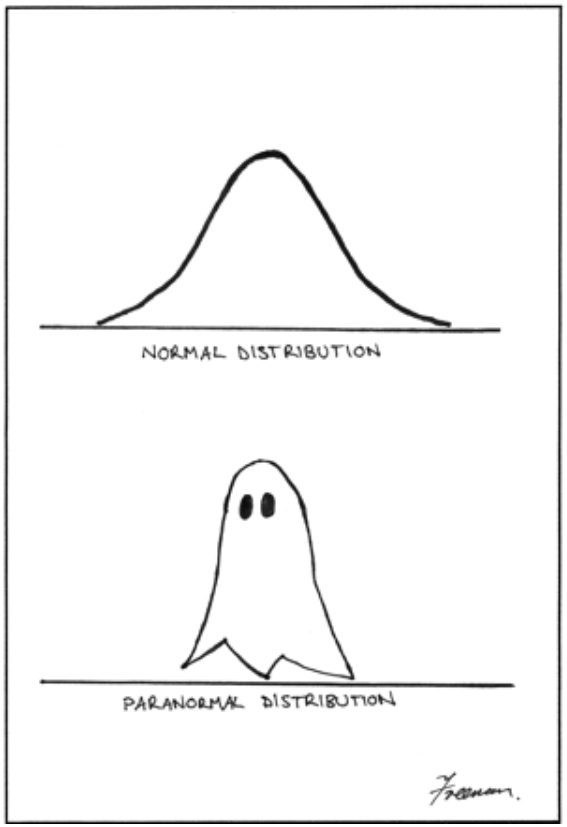
La p-valeur est donc par définition la probabilité sous l'hypothèse nulle d'observer des données au moins aussi grandes que la donnée observée.

Crédits : A. Jebrane (cours statistiques L2 psychologie)

Notion de p-valeur

Un domaine d'application : l'étude du paranormal

Afin de tester un potentiel don d'un individu, on peut soumettre ce dernier à une épreuve aléatoire (par exemple : le test des cartes de Zener vu en TP), et quantifier *via* la p-valeur si le résultat obtenu peut être considéré comme extraordinaire sous l'hypothèse nulle.



Notion de p-valeur en vidéo

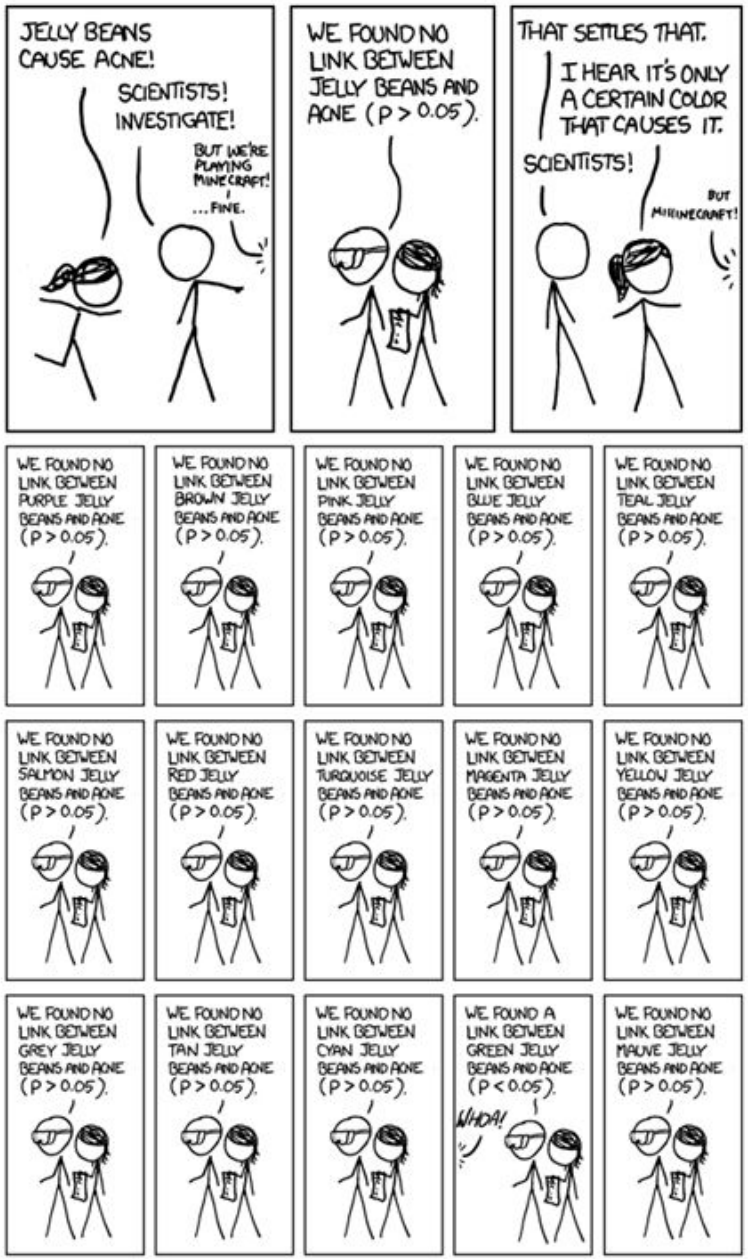


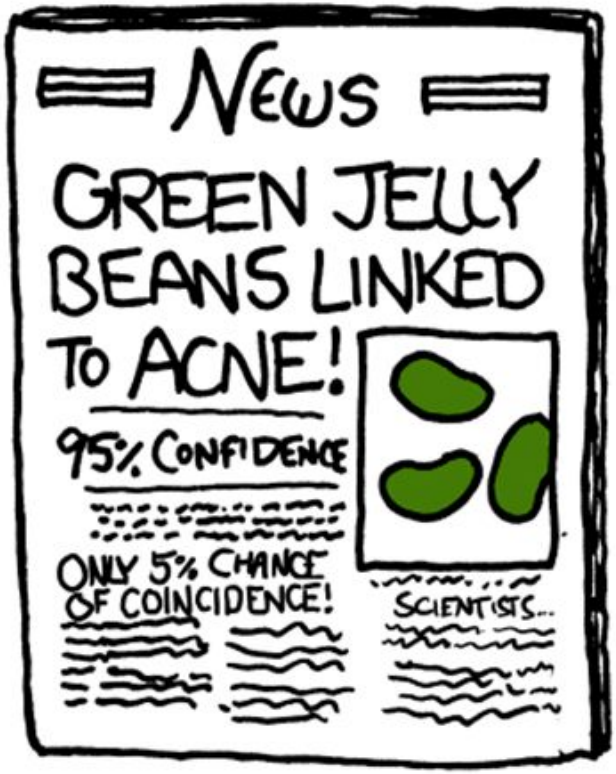
https://www.youtube.com/watch?v=xVIt51ybvu0



https://www.youtube.com/watch?v=PRtwo1j0y2I

Attention aux tests multiples





http://xkcd.com/882/

Quelques mots de conclusion pour ce cours

Ce qu'il faut retenir de tout cela

- Savoir résumer et présenter des données.
- Qu'est qu'une variable aléatoire ? une densité de probabilité ? une fonction de répartition ?
- Savoir manipuler la règle de la somme, du produit et de Bayes.
- Connaitre et reconnaitre la loi normale et autres lois usuelles .
- Étudier les propriétés des estimateurs (biais, variance, etc.) ainsi que leur loi.
- Comprendre et expliciter les intervalles de confiance, tests d'hypothèse et calcul de p-valeur.
- Interpréter des résultats statistiques et éviter les écueils mentionnés dans ce cours.