

Statistique

Université d'Angers – M1 Data Science

Support de cours

(F. Proïa)

Statistique

Université d'Angers – M1 Data Science

Support de cours

(F. Proïa)

Table des matières

1	Introduction à la statistique	2
1.1	Quelques outils essentiels	2
1.2	Quantiles	4
1.3	Statistique descriptive	5
1.4	Échantillonnage : l'exemple des sondages	8
1.5	Quatre distributions essentielles en statistique	9
1.6	Modélisation paramétrique	11
2	Estimation paramétrique	13
2.1	Estimation ponctuelle d'un paramètre	14
2.2	Principales propriétés des estimateurs	20
2.3	Estimation de variance minimale	25
2.4	Présentation de l'inférence bayésienne	31
3	Tests d'hypothèses	35
3.1	Tests construits sur des intervalles de confiance	36
3.2	Risques et puissance	40
3.3	Optimalité de Neyman-Pearson	40
3.4	Tests du khi-deux	43
3.5	Test de Kolmogorov-Smirnov	46
3.6	La p-valeur	47
3.7	Un bestiaire de tests...	50
4	Analyse des données	51
4.1	Analyse en composantes principales	51
4.2	Analyse factorielle des correspondances	58
	Formulaire – Lois de probabilité usuelles	63
	Formulaire – Espérance et variance/covariance	64

Rappel : • indépendant \Rightarrow non corrélée
~~cf~~

1 Introduction à la statistique

Nous supposons connus la théorie élémentaire de la mesure, le calcul des probabilités, ses axiomes et ses définitions ou encore la signification des convergences probabilistes (en loi, en probabilité, presque sûre, dans L^2). De plus, les lois de probabilités usuelles ainsi qu'un formulaire sur le comportement de l'espérance et de la variance sont disponibles en fin de poly. Nous rappelons simplement la loi des grands nombres, le théorème central limite et deux théorèmes de préservation des convergences qui nous suivront tout au long de ce module. Ensuite, nous passerons directement à la statistique...

cf formulaire

1.1 Quelques outils essentiels

1.1.1 Deux théorèmes fondamentaux

Théorème 1.1 (Lois des grands nombres) On a les deux lois des grands nombres suivantes :

→ **Loi faible (LfGN)** Soit $(X_n)_{n \geq 1}$ une suite de v.a.r. décorrélées, de même espérance μ et de même variance σ^2 finie. Alors,

$$\bar{X}_n = \frac{1}{n} \sum_{k=1}^n X_k \xrightarrow{\mathbb{P}} \mu.$$

→ **Loi forte (LFGN)** Soit $(X_n)_{n \geq 1}$ une suite de v.a.r. de même loi et indépendantes, d'espérance μ finie. Alors,

$$\bar{X}_n = \frac{1}{n} \sum_{k=1}^n X_k \xrightarrow{\text{p.s.}} \mu.$$

= i.i.d. = indépendantes et identiquement distribuées

<p>Preuve : Inégalité de Markov : Soit $Z \geq 0$ (Z v.a. $\mathbb{P}(Z=k) = 1$ avec $k \in \mathbb{R}$) Alors $\forall \varepsilon > 0 \quad \mathbb{P}(Z \geq \varepsilon) \leq \frac{\mathbb{E}(Z)}{\varepsilon}$</p> <p>On a $\mathbb{E}[\bar{X}_n] = \frac{1}{n} \sum_{k=1}^n \mathbb{E}[X_k] = \frac{n\mu}{n} = \mu$ <i>linéarité</i></p> <p>$\text{Var}(\bar{X}_n) = \frac{1}{n^2} \sum_{k=1}^n \text{V}(X_k) = \frac{1}{n^2} n \sigma^2 = \frac{\sigma^2}{n}$ <i>v.a.r. décorrélées.</i></p> <p>Soit $\varepsilon > 0$ $\mathbb{P}(\bar{X}_n - \mu \geq \varepsilon) = \mathbb{P}(\bar{X}_n - \mu ^2 \geq \varepsilon^2)$ <i>il y a une équivalence grâce à la valeur absolue</i> $\leq \frac{\mathbb{E}(\bar{X}_n - \mu ^2)}{\varepsilon^2} = \frac{\text{Var}(\bar{X}_n)}{\varepsilon^2} \leq \frac{\sigma^2}{n \varepsilon^2}$</p>	<p>En passant à la limite $n \rightarrow +\infty$ $0 \leq \lim_{n \rightarrow +\infty} \mathbb{P}(\bar{X}_n - \mu \geq \varepsilon) \leq 0$</p> <p>d'où $\bar{X}_n \xrightarrow{\mathbb{P}} \mu$</p>
---	--

Théorème 1.2 (Théorème central limite (TCL)) Soit $(X_n)_{n \geq 1}$ une suite de v.a.r. de même loi et indépendantes, d'espérance μ et de variance $\sigma^2 > 0$ finie. Alors,

$$\frac{\Sigma_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1) \quad \text{où} \quad \Sigma_n = \sum_{k=1}^n X_k.$$

En statistique, on rencontre plus fréquemment le TCL sous la forme

$$Z_n = \sqrt{n} \frac{\bar{X}_n - \mu}{\sigma} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$$

ou encore sous la forme

$$\forall a \leq b, \quad \lim_{n \rightarrow +\infty} \mathbb{P}(a \leq Z_n \leq b) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{s^2}{2}} ds.$$

$$Z_n \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$$

$$\forall x \in \mathbb{R} \quad \mathbb{P}(Z_n \leq x) = F_{Z_n}(x) \rightarrow F_Z(x)$$

où $Z \sim \mathcal{N}(0,1)$

$$\mathbb{P}(a \leq Z_n \leq b) = F_{Z_n}(b) - F_{Z_n}(a) + \mathbb{P}(Z_n = a)$$

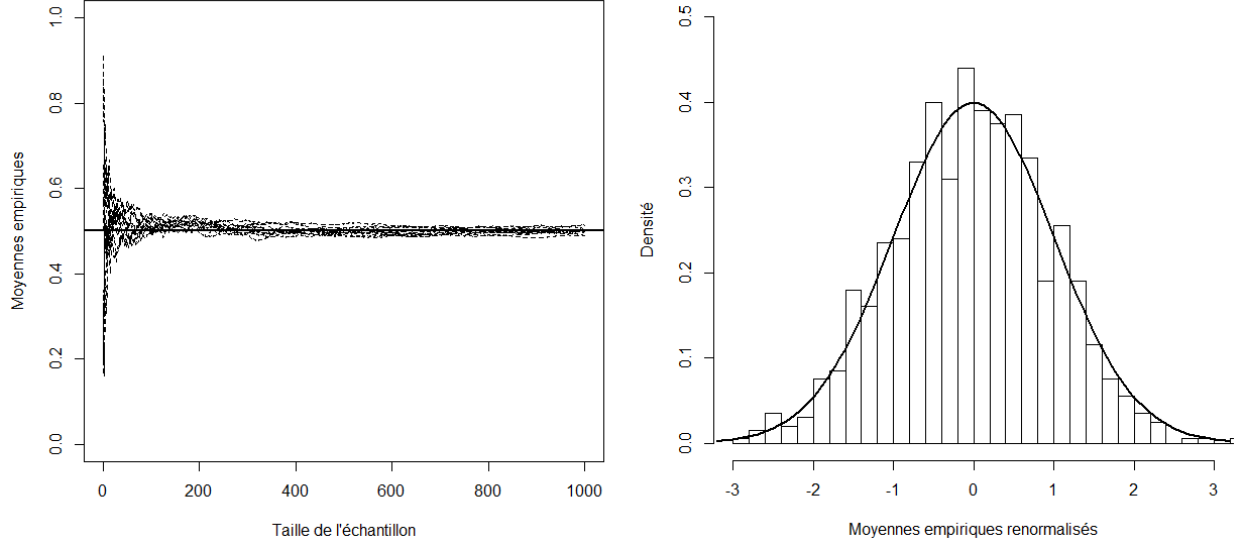
$$\xrightarrow{n \rightarrow +\infty} F_Z(b) - F_Z(a) + 0$$

$$= \int_{-\infty}^b \frac{1}{\sqrt{2\pi}} e^{-\frac{s^2}{2}} ds - \int_{-\infty}^a \frac{1}{\sqrt{2\pi}} e^{-\frac{s^2}{2}} ds$$

$$= \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{s^2}{2}} ds$$

Ci-dessous, deux graphes illustrant la LGN et le TCL par quelques simulations.

- Pour la LGN : on a simulé $N = 10$ échantillons de loi $\mathcal{U}([0, 1])$ de taille $n = 1000$, puis on a représenté l'évolution de la suite $(\bar{X}_k)_{1 \leq k \leq n}$ pour chaque échantillon.
- Pour le TCL : on a simulé $N = 1000$ échantillons de loi $\mathcal{U}([0, 1])$ de taille $n = 1000$, puis on a représenté l'histogramme des N valeurs de Z_n obtenues (ici, $\mu = \frac{1}{2}$ et $\sigma^2 = \frac{1}{12}$). La densité $\mathcal{N}(0, 1)$ est superposée.



1.1.2 Préservation des convergences

Proposition 1.1 (Théorème de continuité de Mann-Wald (CMT)) Soient $((X_n)_{n \geq 1}, X) \in \mathbb{R}^p$ une suite de vecteurs aléatoires et $h : \mathbb{R}^p \rightarrow \mathbb{R}^q$ une fonction dont l'ensemble D_h des discontinuités est tel que $\mathbb{P}(X \in D_h) = 0$. Alors,

$$X_n \xrightarrow{*} X \implies h(X_n) \xrightarrow{*} h(X)$$

où $\xrightarrow{*}$ désigne soit $\xrightarrow{\text{p.s.}}$, soit $\xrightarrow{\mathbb{P}}$, soit $\xrightarrow{\mathcal{L}}$.

Dans nos applications, on aura généralement $p = q = 1$ mais il est important de savoir que le résultat est vrai pour les vecteurs également, ne serait-ce que pour établir les corollaires suivants.

Corollaire 1.1 Si $U_n \xrightarrow{\text{p.s.}} U$ et $V_n \xrightarrow{\text{p.s.}} V$, alors

$$U_n + V_n \xrightarrow{\text{p.s.}} U + V \quad \text{et} \quad U_n V_n \xrightarrow{\text{p.s.}} UV.$$

Il en va de même si l'on remplace $\xrightarrow{\text{p.s.}}$ par $\xrightarrow{\mathbb{P}}$.

Attention à l'erreur classique : avec $\xrightarrow{\mathcal{L}}$, le résultat est potentiellement faux. On a besoin pour cela de l'hypothèse plus forte de convergence en loi du vecteur (U_n, V_n) vers le couple (U, V) . Tout cela nous conduit à cet autre résultat fondamental de préservation des convergences.

Proposition 1.2 (Lemme de Slutsky) Soient $((U_n)_{n \geq 1}, U) \in \mathbb{R}^p$ et $(V_n)_{n \geq 1} \in \mathbb{R}^q$ deux suites de vecteurs aléatoires. Alors, si les suites peuvent être additionnées et/ou multipliées,

$$U_n \xrightarrow{\mathcal{L}} U \quad \text{et} \quad V_n \xrightarrow{\mathbb{P}} c \implies U_n + V_n \xrightarrow{\mathcal{L}} c + U \quad \text{et/ou} \quad U_n V_n \xrightarrow{\mathcal{L}} cU$$

où $c \in \mathbb{R}^q$ est une constante.

Exemple. Soient $\mu \in \mathbb{R}$ et $\sigma > 0$ des paramètres. Supposons que $\sqrt{n} \frac{\bar{X}_n - \mu}{\sigma} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1)$, que $S_n^2 \xrightarrow{\text{p.s.}} \sigma^2$ et que $S_n^2 > 0$. On pose de plus $S_n^{*2} = \frac{n}{n-1} S_n^2$. Donner la limite de

$$\sqrt{n} \frac{\bar{X}_n - \mu}{\sqrt{S_n^2}} \quad \text{et celle de} \quad \sqrt{n} \frac{\bar{X}_n - \mu}{\sqrt{S_n^{*2}}}.$$

$U_n = \sqrt{n} \frac{\bar{X}_n - \mu}{\sigma} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1)$ $S_n^2 \xrightarrow{\text{p.s.}} \sigma^2 \Rightarrow S_n^2 \xrightarrow{\mathbb{P}} \sigma^2$ $\sqrt{n} \frac{\bar{X}_n - \mu}{\sqrt{S_n^2}} = \sqrt{n} \frac{\bar{X}_n - \mu}{\sigma} \cdot \frac{\sigma}{\sqrt{S_n^2}} = U_n \cdot \frac{\sigma}{\sqrt{S_n^2}}$ $h(S_n^2) = \frac{\sigma}{\sqrt{S_n^2}} \text{ et on pose } h(x) = \frac{\sigma}{\sqrt{x}} \quad x \in \mathbb{R}_+^*$ $h(S_n^2) \xrightarrow{\text{p.s.}} h(\sigma^2) \text{ par continuité}$ $\text{donc } \frac{\sigma}{\sqrt{S_n^2}} \xrightarrow{\text{p.s.}} \frac{\sigma}{\sqrt{\sigma^2}} = 1$	$\Rightarrow \frac{\sigma}{\sqrt{S_n^2}} \xrightarrow{\mathbb{P}} 1$ $\text{Donc } \sqrt{n} \frac{\bar{X}_n - \mu}{\sqrt{S_n^2}} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1) \text{ par Slutsky}$ <hr/> $\sqrt{n} \frac{\bar{X}_n - \mu}{\sqrt{S_n^{*2}}} = \underbrace{\sqrt{n} \frac{\bar{X}_n - \mu}{\sqrt{S_n^2}}}_{\xrightarrow{\mathcal{L}} \mathcal{N}(0, 1)} \times \underbrace{\sqrt{\frac{S_n^2}{S_n^{*2}}}}_{\sqrt{\frac{n-1}{n}} \rightarrow 1} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1)$ $\frac{S_n^2}{S_n^{*2}} = \frac{S_n^2}{\frac{n}{n-1} S_n^2} = \frac{n-1}{n}$
---	---

1.2 Quantiles

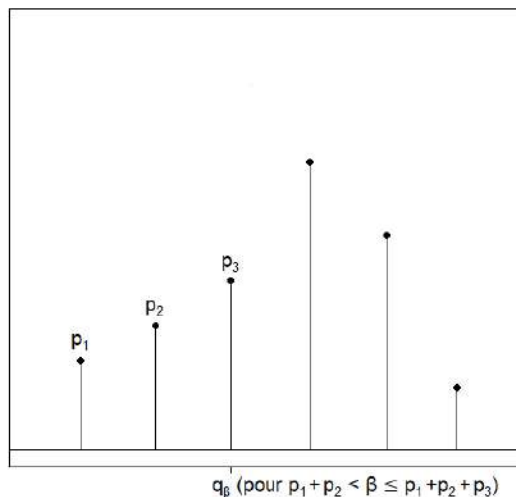
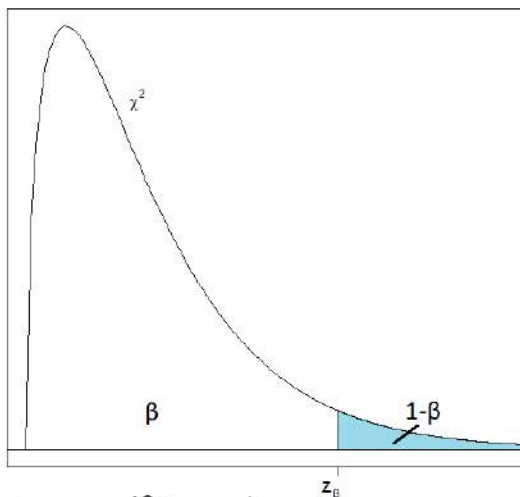
On aura souvent recours à la fonction quantile lors de la construction des intervalles de confiance. Ci-dessous, F_X désigne la fonction de répartition de X , variable aléatoire à valeurs dans un ensemble E .

Définition 1.1 La fonction quantile $Q_X : [0, 1] \rightarrow E$ de la v.a.r. X est définie par

$$Q_X(\beta) = \inf_{x \in E} \{x \mid F_X(x) \geq \beta\}.$$

$Q_X(\beta)$ est le quantile d'ordre β de la loi de X .

Lorsque F_X est strictement croissante, il s'agit simplement de son inverse F_X^{-1} sinon il s'agit d'une inverse généralisée à gauche, parfois notée F_X^- ou $F_X^<$. Lorsque $\beta = \frac{1}{2}$, le quantile obtenu est une médiane de X .



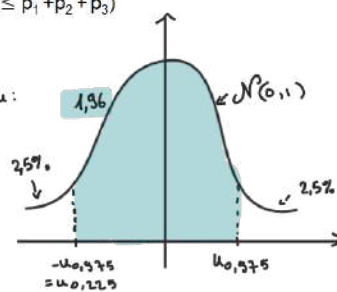
Rappel : si $X \sim \mathcal{N}(\mu, \sigma^2)$.

- * $\forall a \in \mathbb{R}^* \quad aX \sim \mathcal{N}(a\mu, a^2\sigma^2)$
- * $\forall b \in \mathbb{R} \quad X+b \sim \mathcal{N}(\mu+b, \sigma^2)$.

Quantile : loi continue (sur les densités)
 β est l'air sous la courbe

4

Le quantile le plus connu :



u_α : quantile d'ordre α

$Z \sim \mathcal{N}(0, 1)$

$P(Z \leq u_\alpha) = P(Z \geq u_{1-\alpha}) = P(-Z \leq -u_{1-\alpha})$

parité de la densité car Z et $-Z$ ont la m. loi

$= P(Z \leq -u_{1-\alpha}) \Rightarrow u_\alpha = u_{1-\alpha}$

$t(n)$ la loi de student

De même $-t_\alpha(n) = t_{1-\alpha}(n)$ par parité de la densité (voir section 1-5)

$$\mathbb{P}(|X| \geq u_{0.975}) = \mathbb{P}(X \leq -u_{0.975}) + \mathbb{P}(X \geq u_{0.975}) = 0.05.$$

Une étude statistique commence *nécessairement* par une étude descriptive des données à disposition. Avant de faire de l'inférence (modélisation, apprentissage, prédiction, etc.) sur une population, une étape préalable d'analyse descriptive est essentielle. Il s'agit de se poser des questions comme :

- On appelle *population* l'ensemble (généralement inobservable) sur lequel repose l'étude. Chaque élément de la population est un *individu* et l'on appelle *caractères* les caractéristiques qui le définissent.

Définition 1.2 Un caractère est **quantitatif** lorsque ses valeurs sont traitées en tant que valeurs numériques. Au contraire, le caractère est **qualitatif** lorsque les valeurs qu'il peut prendre ne sont pas traitées en tant que valeurs numériques, elles forment alors ses modalités. Une variable qualitative numérique ordonnée est qualifiée d'ordinaire. Le caractère est discret lorsque l'ensemble dans lequel il prend ses valeurs est dénombrable, sinon il est continu.

qualitatif : couleur de cheveux
quantitatif : la taille en cm
qualitatif ordinal : 0 an
1 an
⋮
âge.

Médiane meilleur que la moyenne si on a des valeurs aberrantes:
les tailles : 1,58 ; 1,65 ; ... ; 177
↑
faute de frappe

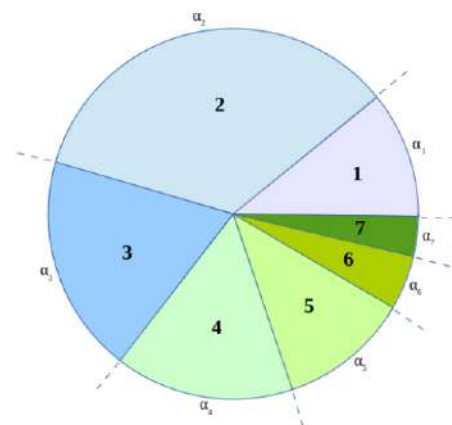
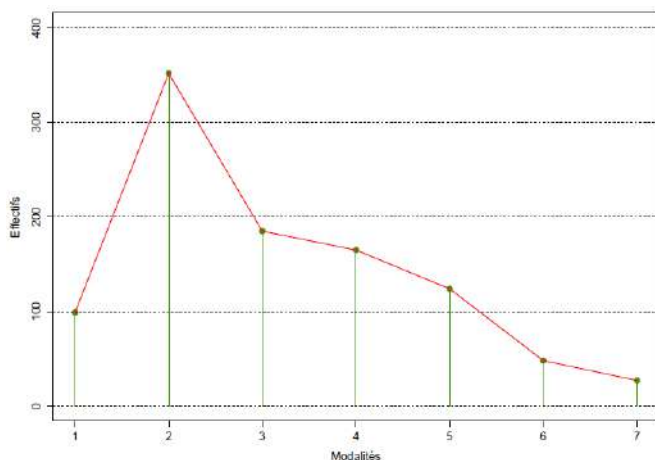
La médiane n'est pas impactée par cette faute tandis que la moyenne est grandement impactée

La méthode de collecte sert également à justifier rétrospectivement certaines hypothèses de modélisation (variables indépendantes et de même loi, par exemple).

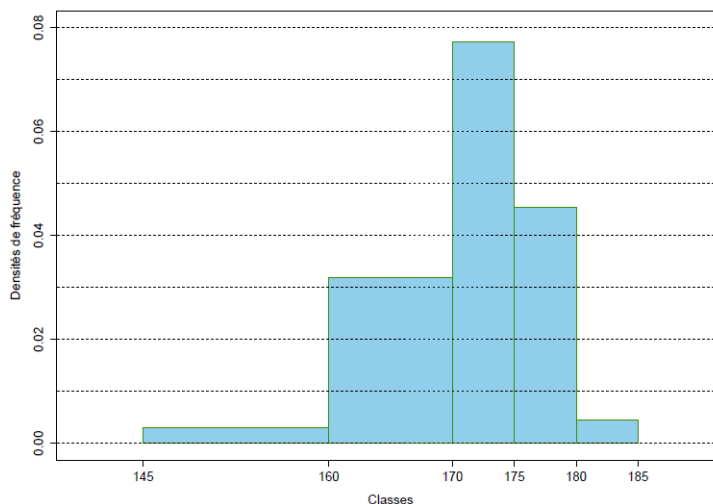
1.3.2 Visualisation des données

La représentation graphique permet d'avoir un premier aperçu de la nécessité éventuelle de transformer les données, de la présence de valeurs aberrantes, de la quantité de valeurs manquantes, de la loi de probabilité à choisir pour la modélisation, etc.

→ **Caractère discret.** On calcule les fréquences correspondant à chaque modalité, le *diagramme en bâtons* ou le *diagramme circulaire* permettent d'en avoir une bonne vision d'ensemble.



→ **Caractère continu.** On peut le discrétiser en classes, calculer les densités de fréquences (les fréquences renormalisées par la longueur des classes), et le mettre sous forme d'*histogramme*.

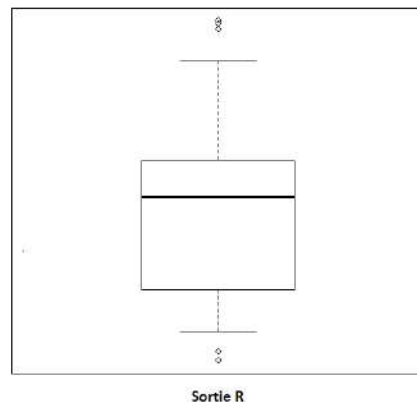
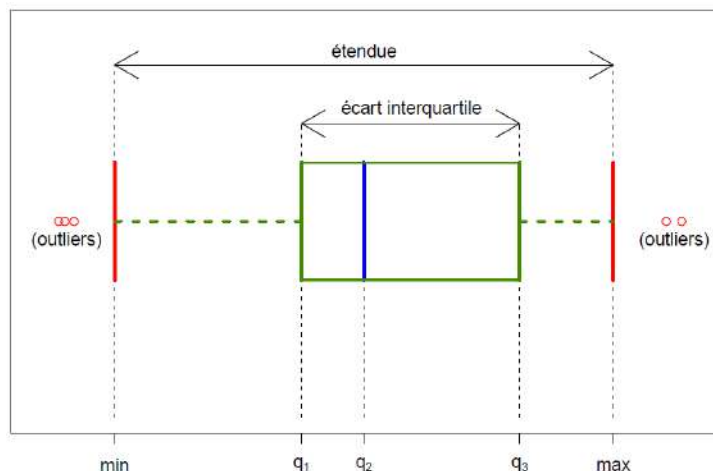


Dans le cas des caractères quantitatifs, les indicateurs de *centrage* sont la moyenne m ou la médiane q_2 (50/50) alors que, pour les indicateurs de *dispersion*, on peut penser aux extrêmes min et max, à la variance s^2 ou encore aux premier quartile q_1 (25/75) et dernier quartile q_3 (75/25). Ces données numériques se mettent généralement sous forme de *boxplot*.

1. pas de lien, on suppose corrélation linéaire
 2. exemple: $Y = aX + b + \varepsilon$, on suppose corrélation proche de 1 (ex 0,8 car trop de bruit)

3. exemple $Y = ax^2 + bx + c + \varepsilon$, on suppose corrélation ex -0,6 ($Y = -X$)

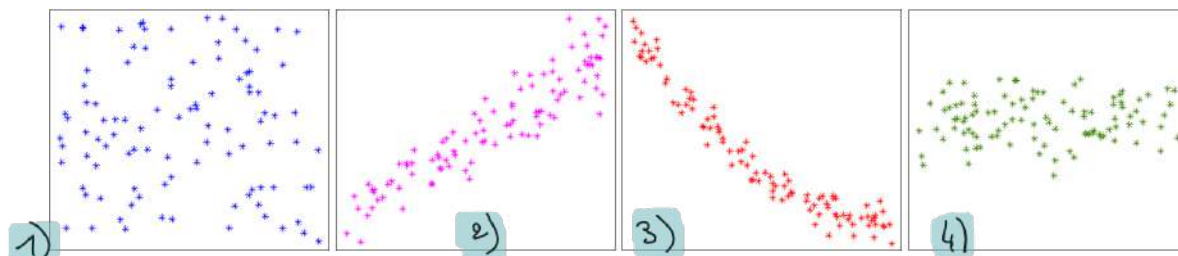
4. pas de lien entre y et x ($y = \text{constante}$). On suppose corrélation linéaire = 0



Ce dernier permet de mettre en évidence les *outliers* grâce à des seuils s_1 et s_3 spécifiquement calculés. Le logiciel R par exemple recommande par défaut dans sa commande `boxplot` de déceler comme outliers les observations extérieures aux bornes

$$s_1 = q_1 - \frac{3(q_3 - q_1)}{2} \quad \text{et} \quad s_3 = q_3 + \frac{3(q_3 - q_1)}{2}.$$

Ces représentations sont univariées, mais on peut encore travailler avec des représentations bivariées, comme des *nuages de points*, pour détecter l'existence d'une relation entre deux caractères quantitatifs ainsi que sa nature (linéaire, quadratique, etc.).



Les études descriptives multivariées sont plus délicates car l'outil visuel n'est pas toujours adéquat pour illustrer les relations corrélatives, mais une représentation graphique parallèle de chaque caractère est souvent utile, avec un récapitulatif des paramètres de centrage et de dispersion. Enfin, on peut être amené à transformer les données au préalable, par exemple lorsqu'elle changent d'échelle ou qu'elles doivent être linéarisées. Les *transformations de Box-Cox* sont un exemple courant de transformations, elles consistent à appliquer aux données, lorsque c'est possible, la fonction

$$f_\lambda(x) = \begin{cases} \frac{x^\lambda - 1}{\lambda} & \text{si } \lambda \neq 0 \\ \ln x = \lim_{\lambda \rightarrow 0^+} \frac{x^\lambda - 1}{\lambda} & \text{si } \lambda = 0 \end{cases}$$

pour une valeur de λ à affiner. On rencontre aussi des étapes de recentrage, de standardisation, de translation, etc. Avant de commencer la modélisation, il faudra également se poser la question du traitement des valeurs aberrantes et des valeurs manquantes. Ce problème est à examiner au cas par cas.

1.3.3 Différents types de corrélation

La présence de corrélation linéaire entre deux séries quantitatives $(x_i)_{1 \leq i \leq n}$ et $(y_i)_{1 \leq i \leq n}$ est classiquement évaluée grâce au *coefficient de Pearson*,

$$\rho_{x,y} = \frac{\sum_{k=1}^n (x_k - \bar{x}_n)(y_k - \bar{y}_n)}{\sqrt{\sum_{k=1}^n (x_k - \bar{x}_n)^2} \sqrt{\sum_{k=1}^n (y_k - \bar{y}_n)^2}}$$

$S_x^2 = \frac{1}{n} \sum_{k=1}^n (x_k - \bar{x})^2$: variance empirique de la série x

$S_y^2 = \frac{1}{n} \sum_{k=1}^n (y_k - \bar{y})^2$: variance empirique de la série y

$\text{Var}(X) = E[X^2] - (E[X])^2$

$\sigma_{x,y} = \frac{1}{n} \sum_{k=1}^n (x_k - \bar{x})(y_k - \bar{y})$: variance empirique entre les séries x et y

$\text{Cov}(X, Y) = E[XY] - E[X]E[Y]$

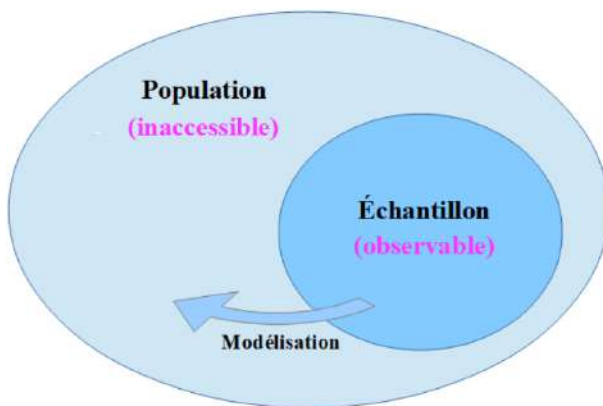
$$\rho_{x,y} = \frac{\sigma_{x,y}}{\sqrt{S_x^2} \sqrt{S_y^2}} \quad \left(\text{on a } \text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma(X) \sigma(Y)} \right)$$

où \bar{x}_n et \bar{y}_n désignent les moyennes empiriques. L'inégalité de Cauchy-Schwarz nous donne immédiatement $|\rho_{x,y}| \leq 1$ avec les valeurs extrêmes $\rho_{x,y} = \pm 1$ atteintes par les alignements parfaits. Le principal défaut de ce coefficient est qu'il ne permet de détecter que les relations linéaires et qu'il est peu sensible à la corrélation non linéaire. Le *coefficient de Spearman* consiste à calculer le coefficient de Pearson entre les rangs des variables dans les séries triées, il met donc en évidence les relations monotones. D'autres coefficients basés sur les rangs, tel celui de Kendall, permettent également de détecter les relations non linéaires.

Exemple. Dans les exemples présentés ci-dessus, le nuage rouge aura un coefficient de Spearman plus significatif que son coefficient de Pearson (environ -0.92 contre -0.77). À défaut de linéarité, la monotonie caractérise tout autant la présence de corrélation forte.

1.4 Échantillonnage : l'exemple des sondages

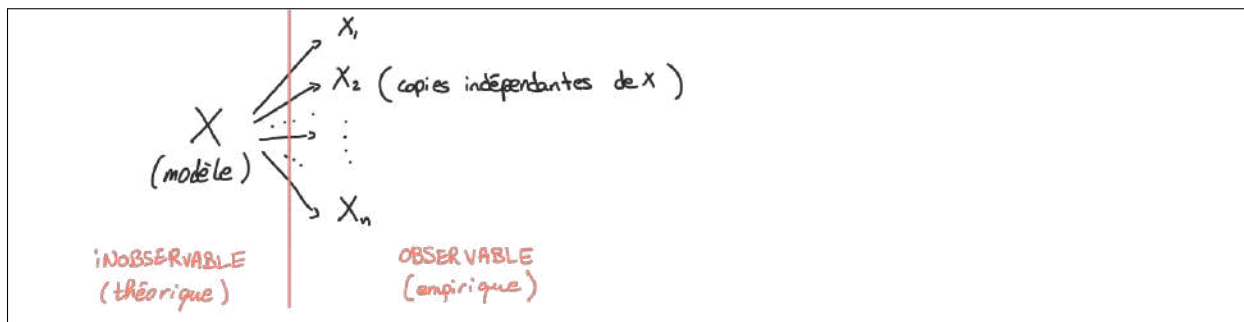
On aborde maintenant l'*inférence statistique*, et pour cela on suppose que l'on dispose d'un ensemble de données numériques, sur lequel tout le travail descriptif a été fait. On peut voir un *échantillon* comme un ensemble d'observations choisies dans une population avec un double objectif : mesurer plus rapidement et à moindre coût. L'*échantillonnage* doit par conséquent être pensé correctement pour que l'échantillon soit *représentatif* de la population. C'est une des problématiques majeures des études statistiques.



On va considérer que les données sont des réalisations de variables aléatoires dont la loi est représentative de la population. L'échantillon sert à fournir une approximation aussi crédible que possible de la loi en question, et les caractéristiques de cette dernière doivent ensuite permettre de tirer des conclusions à plus grande échelle sur la population (de type "grands nombres").

Définition 1.3 On appelle n -échantillon une suite X_1, \dots, X_n de v.a.r. indépendantes et identiquement distribuées. La v.a.r. parente du n -échantillon est notée X , il s'agit de la v.a.r. dont X_1, \dots, X_n sont des copies indépendantes.

Schématiquement, on cherche à expliquer l'observation des variables X_1, \dots, X_n par l'intermédiaire d'un *modèle* que l'on place sur la variable X inobservable :



Définition 1.4 Une statistique $T(X_1, \dots, X_n)$ sur un n -échantillon est une fonction des variables qui le composent.

Exemples. Quelques statistiques usuelles de description d'échantillon :

$$T(X_1, \dots, X_n) = \frac{1}{n} \sum_{k=1}^n X_k \quad \bar{X}_n$$

$$T(X_1, \dots, X_n) = \frac{1}{n} \sum_{k=1}^n (X_k - \bar{X}_n)^2 = S_n^2$$

$$T(X_1, \dots, X_n) = \min_{1 \leq k \leq n} X_k = X_{(1)} \rightarrow \begin{cases} \text{c'est la} \\ \text{plus petite} \\ \text{valeur} \end{cases}$$

statistique d'ordre si entre ()

Il faut que toutes les variables soient connues, si on travaille sur des variables (par exemple p) dont on ne connaît pas la valeur, ce n'est pas une statistique

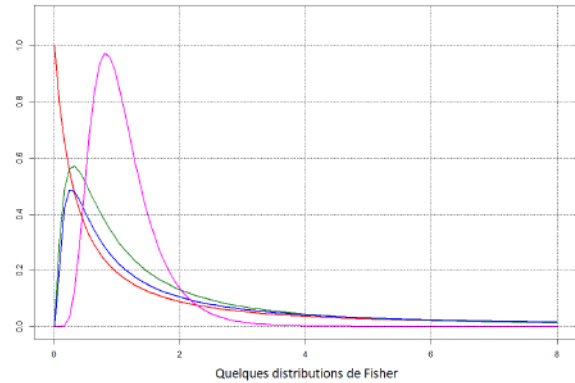
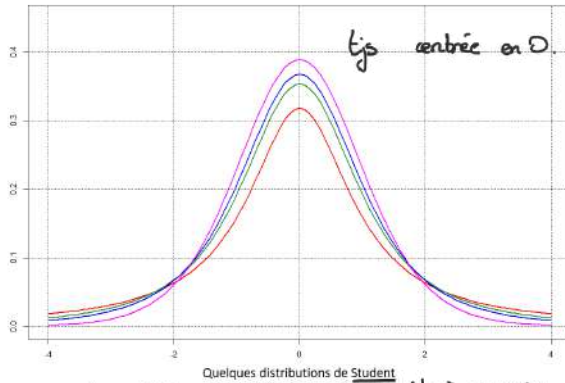
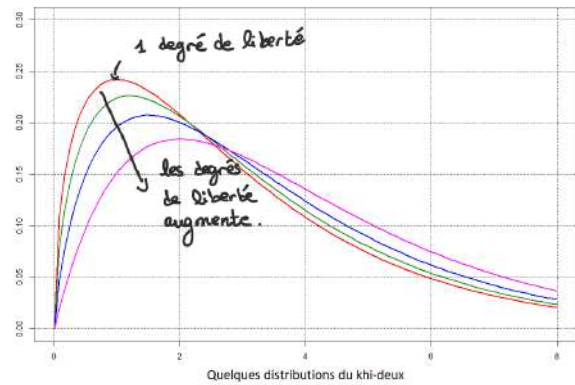
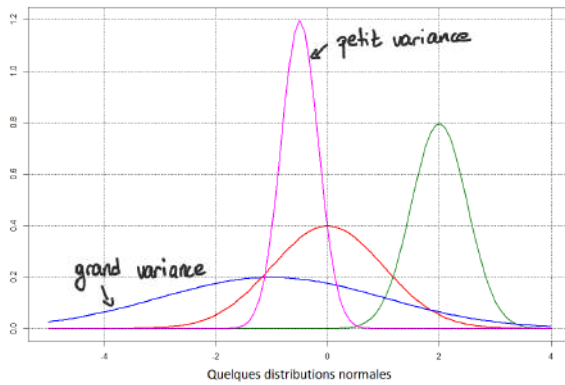
Un exemple très connu est le *sondage d'opinion*. Les problématiques usuelles de l'échantillonnage sont au cœur de l'expérience : la taille est-elle suffisante, l'échantillon peut-il être tenu pour représentatif de la population, quelle méthode d'enquête utiliser, quelle crédibilité accorder aux réponses, ... ? Si l'on veut compartimenter les mesures (par origine sociale, par sexe, par niveau d'instruction, etc.), on s'aperçoit qu'il faut vite un grand nombre de questionnaires pour pouvoir étudier l'influence de chaque croisement de facteurs sur le vote. Supposons que l'on dispose d'un panel (liste d'état civil, annuaire téléphonique, abonnés volontaires, ...). Voici quelques procédures :

- On fait confiance au hasard. On tire au sort dans le panel une fraction prédéfinie d'individus.
- On classe les individus, puis on choisit $1 \leq i_0 \leq k$. Ensuite on sonde de k en k : $i_0, i_0 + k, i_0 + 2k, \dots$
- On stratifie la population selon un ou plusieurs critères pertinents (tranches d'âge, situation socioprofessionnelle, niveau d'études, ...). Chaque strate est traitée séparément, en lui associant éventuellement un poids en lien avec sa taille.
- La population est clusterisée, le sondage est en grappe. On crée des groupes, puis on tire au sort les groupes sondés (soit en intégralité, soit de manière récursive en définissant des sous-groupes dans les groupes, et ainsi de suite).
- Lorsque l'on dispose d'une information très fiable sur la population, on peut décider de sonder par quotas (tranches d'âge, sexe, situation géographique, ...), pour former un échantillon aussi proche que possible de la population de référence.
- On questionne sans plan bien établi, comme dans une enquête de rue, les personnes présentes et qui acceptent de se prêter au jeu. On peut y ajouter un effet "boule de neige" lorsque les sondés sondent à leur tour leurs connaissances (méthode à la mode sur les réseaux sociaux).

La capacité d'extrapolation de l'échantillon à la population est un des objectifs de la statistique inférentielle, qui va permettre de mesurer la qualité de l'estimation et l'incertitude qui lui est associée.

1.5 Quatre distributions essentielles en statistique

Les quatre distributions présentées ci-dessous (normale, khi-deux, Student et Fisher) sont fondamentales en statistique, on les retrouve dans de nombreuses applications (tests et régression essentiellement). On se rapportera au formulaire pour comprendre comment ces distributions sont reliées entre elles. Mis à part les caractéristiques de la loi normale, les autres ne sont pas à connaître par cœur.



- expression des densités

→ **Loi normale** (en haut à gauche). Portée par \mathbb{R} , de paramètres $\mu \in \mathbb{R}$ et $\sigma^2 > 0$, de densité

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

avec espérance μ et variance σ^2 .

→ **Loi du khi-deux** (en haut à droite). Portée par \mathbb{R}^+ (ou \mathbb{R}^{+*}), à $k \in \mathbb{N}^*$ degrés de liberté, de densité

$$f(x) = \frac{1}{2^{\frac{k}{2}} \Gamma(\frac{k}{2})} x^{\frac{k}{2}-1} e^{-\frac{x}{2}} \mathbb{1}_{\{x \geq 0\}} \quad (\text{ou } \mathbb{1}_{\{x > 0\}} \text{ si } k = 1)$$

avec espérance k et variance $2k$.

→ **Loi de Student** (en bas à gauche). Portée par \mathbb{R} , à $k \in \mathbb{R}_+^*$ degrés de liberté, de densité

$$f(x) = \frac{1}{\sqrt{k\pi}} \frac{\Gamma(\frac{k+1}{2})}{\Gamma(\frac{k}{2})} \left(1 + \frac{x^2}{k}\right)^{-\frac{k+1}{2}}$$

avec espérance 0 (si $k > 1$) et variance $\frac{k}{k-2}$ (si $k > 2$).

→ **Loi de Fisher** (en bas à droite). Portée par \mathbb{R}^+ (ou \mathbb{R}^{+*}), à $k_1, k_2 \in \mathbb{R}_+^*$ degrés de liberté, de densité

$$f(x) = \frac{\left(\frac{k_1 x}{k_1 x + k_2}\right)^{\frac{k_1}{2}} \left(\frac{k_2}{k_1 x + k_2}\right)^{\frac{k_2}{2}}}{x B(\frac{k_1}{2}, \frac{k_2}{2})} \mathbb{1}_{\{x \geq 0\}} \quad (\text{ou } \mathbb{1}_{\{x > 0\}} \text{ si } k_1 = 1)$$

avec espérance $\frac{k_2}{k_2-2}$ (si $k_2 > 2$) et variance $\frac{2k_2^2(k_1+k_2-2)}{k_1(k_2-2)^2(k_2-4)}$ (si $k_2 > 4$).

1.6 Modélisation paramétrique

L'hypothèse fondamentale de la statistique paramétrique est qu'il existe un paramètre θ^* *inconnu* qui régit la loi de la variable parente X du n -échantillon. On va donc paramétrer la loi de X par θ , une variable d'ajustement à valeurs dans un espace supposé contenir le vrai θ^* et rechercher, en fonction des données disponibles, la meilleure valeur à donner à ce paramètre (en un certain sens). Pour simplifier, on notera $\underline{X} = (X_1, \dots, X_n)$ le vecteur formé par le n -échantillon et $\underline{x} = (x_1, \dots, x_n)$ sa réalisation. On notera aussi $f_X(\cdot; \theta)$ la densité de X et $F_X(\cdot; \theta)$ sa fonction de répartition. Seulement de manière générique (inutile dans les exemples...), un θ en indice viendra rappeler que la mesure de probabilité \mathbb{P}_θ est paramétrée, de même que $\mathbb{E}_\theta, \mathbb{V}_\theta, \text{Cov}_\theta, \dots$ et tout autre opérateur usuel calculé par rapport à \mathbb{P}_θ .

Θ peut être un couple, un vecteur....

Définition 1.5 On appelle **espace paramétrique** l'ensemble Θ dans lequel le paramètre θ prend ses valeurs.

Un modèle paramétrique pour X est donc caractérisé par la famille de fonctions

hypothèse $\theta^* \in \Theta$

$$\mathcal{F} = \{F_X(\cdot; \theta), \theta \in \Theta\}.$$

Procédures statistiques dont la fiabilité est uniforme en Θ
 $\forall \theta \in \Theta, \dots$ (ça marche pour tout θ y compris θ^* qui est dans Θ)

Remarque. Par *dimension* de Θ , on entend le nombre de composantes du vecteur θ . Par exemple, $\dim(\Theta) = 1$ pour $\Theta = \mathbb{R}^{+*}$ dans le cas du modèle $\mathcal{P}(\lambda)$, ou encore $\dim(\Theta) = 2$ pour $\Theta = \mathbb{R} \times \mathbb{R}^{+*}$ dans le cas du modèle $\mathcal{N}(\mu, \sigma^2)$. On rencontre rarement des modèles dont le paramètre contient plus de 2 composantes.

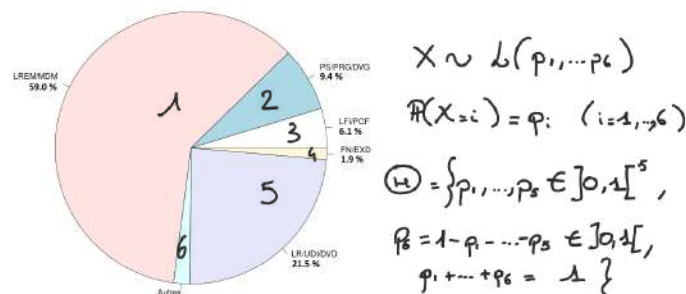
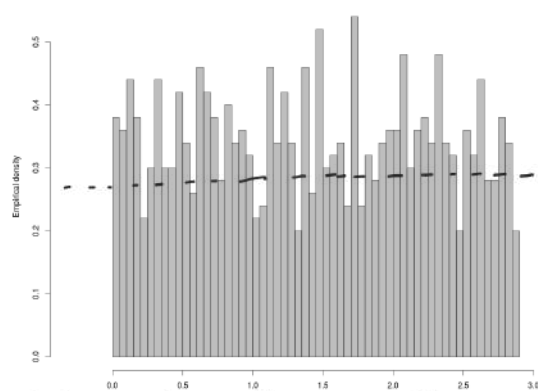
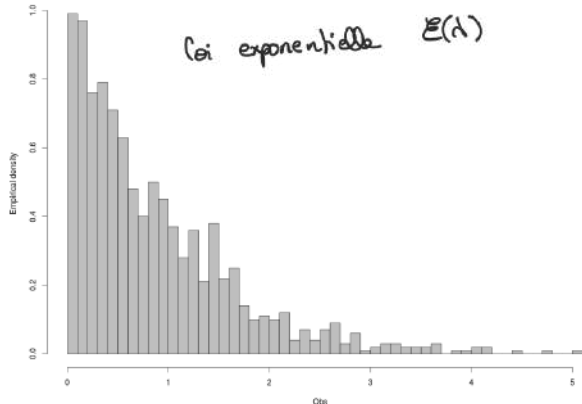
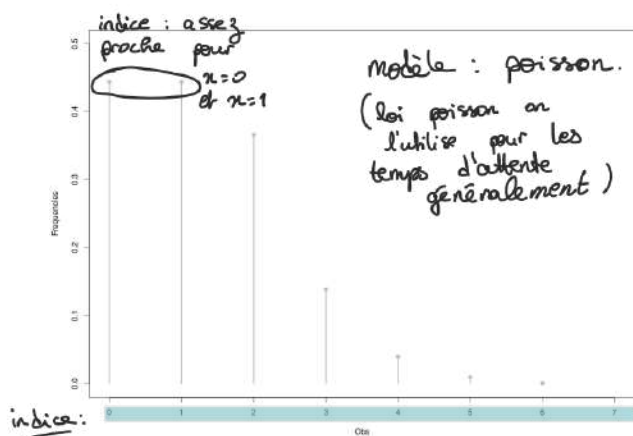
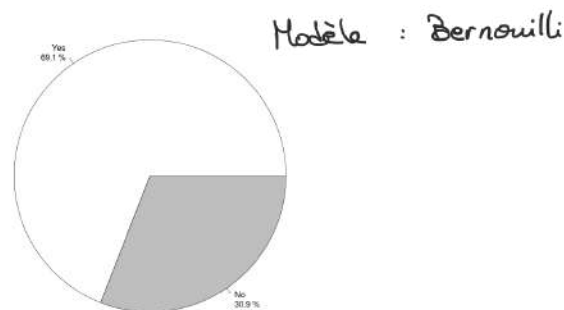
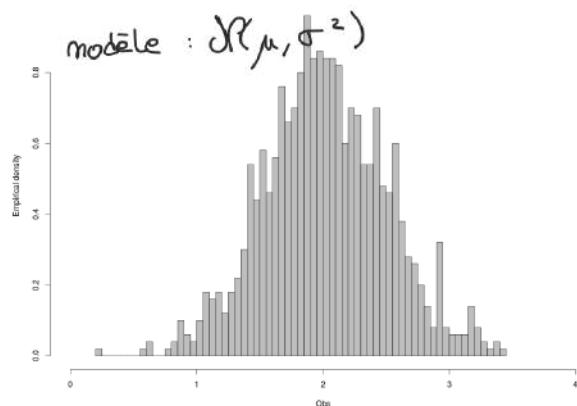
Exemples. Quelques modèles usuels :

* $X \sim \mathcal{B}(p)$, $p \in \Theta =]0, 1[$ (les cas 0 et 1 sont des cas Dirac, on ne les considère pas comme des cas de Bernoulli)
$\forall x \in E = \{0, 1\}$, $f_x(x; p) = p^x (1-p)^{1-x}$
* $X \sim \mathcal{P}(\lambda)$, $\lambda \in \Theta = \mathbb{R}^+$
$\forall x \in E = \mathbb{N}$, $f_x(x; \lambda) = e^{-\lambda} \frac{\lambda^x}{x!}$
* $X \sim \mathcal{N}(\mu, \sigma^2)$, $(\mu, \sigma^2) \in \Theta = \mathbb{R} \times \mathbb{R}^+$
$\forall x \in \mathbb{R}$, $f_x(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$
* $X \sim \mathcal{U}(0, \theta)$, $\theta \in \Theta = \mathbb{R}^+$
$\forall x \in \mathbb{R}$, $f_x(x; \theta) = \frac{1}{\theta} \mathbb{1}_{0 \leq x \leq \theta}$

1.6.1 Choisir un modèle

La question naturelle est donc : pourquoi choisir un modèle plutôt qu'un autre ? La statistique descriptive doit nous guider, comme nous l'avons vu dans la section précédente.

Exemples. Quel modèle choisir dans les exemples suivants ?



relativement uniforme : $\mathcal{U}(0, \theta)$

Pour des raisons d'interprétation du rôle joué par le paramètre, on choisira un modèle satisfaisant

$$\forall \theta_1, \theta_2 \in \Theta, \quad F_X(\cdot; \theta_1) = F_X(\cdot; \theta_2) \implies \theta_1 = \theta_2.$$

On parle d'identifiabilité du modèle pour qualifier cette relation injective entre le paramètre et la répartition. À titre d'exemple, le modèle $X \sim \mathcal{N}(\theta^2, 1)$ n'est pas identifiable sur $\Theta = \mathbb{R}$ car $F_X(\cdot; \theta) = F_X(\cdot; -\theta)$ mais il redevient identifiable sur $\Theta = \mathbb{R}^+$.

10 = 5+5 mais on veut un unique modèle et pas ça

10 = 4+6

10 = 3+7

⋮

1.6.2 Famille exponentielle

La *famille exponentielle*, recouvrant la majorité des modèles usuels (pas tous !), est une famille paramétrique proposant de nombreuses facilités calculatoires. La notation $\langle \cdot, \cdot \rangle$ fait ici référence au produit scalaire de \mathbb{R}^p .

Définition 1.6 La famille paramétrique dont les densités se mettent sous la forme

$$f_X(x; \theta) = a(x) b(\theta) e^{\langle \eta(\theta), t(x) \rangle}$$

avec $\eta(\theta) = (\eta_1(\theta), \dots, \eta_p(\theta))$, $t(x) = (t_1(x), \dots, t_p(x))$ et $p = \dim(\Theta)$, est appelée *famille exponentielle*. Si de plus $\eta_i(\theta) = \theta_i$ pour $i \in \{1, \dots, p\}$, la décomposition exponentielle est sous forme canonique.

Exemples. Quelques exemples et un contre-exemple :

* $X \sim \mathcal{B}(p)$, $f_X(x; p) = p^x (1-p)^{1-x}$ pour $x \in \{0, 1\}$
 $f_X(x; p) = e^{x \ln(p)} (1-p) e^{-x \ln(1-p)} = \underbrace{(1-p)}_{b(p)} e^{\underbrace{x \ln(p)}_{t_1(x)} \underbrace{\ln(p)}_{\eta_1(p)}} \quad \text{et } a(x) = 1$
 En posant $q = \ln \frac{p}{1-p}$, on obtient la forme canonique

* $X \sim \mathcal{P}(\lambda)$. Soit $x \in \mathbb{N}$, $f_X(x; \lambda) = e^{-\lambda} \frac{\lambda^x}{x!} = \underbrace{\frac{1}{x!}}_{a(x)} e^{\underbrace{-\lambda}_{b(\lambda)}} e^{\underbrace{x \ln \lambda}_{t_1(x)} \underbrace{\ln \lambda}_{\eta_1(\lambda)}}$
 → forme canonique $\mu = \ln p$.

* $X \sim \mathcal{N}(\mu, \sigma^2)$ $f_X(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
 $= \underbrace{1}_{a(x)} \times \underbrace{\frac{e^{-\frac{\mu^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}}_{b(\mu, \sigma^2)} e^{\underbrace{\frac{\mu^2}{\sigma^2} x}_{t_1(x)} \underbrace{-\frac{1}{2\sigma^2}}_{\eta_1(\mu, \sigma^2)} x^2}$

* Contre-exemple : $\mathcal{U}(0, \theta)$
 $f_X(x; \theta) = \frac{1}{\theta} \mathbb{1}_{\{0 \leq x \leq \theta\}}$
 On peut choisir des paramètres pour obtenir $e^0 = 1$ mais la fonction indicatrice pose problème ici car elle dépend de x et de θ .

Remarque. Quitte à faire un changement de paramètre, on peut chercher à se ramener à la forme canonique lorsque c'est nécessaire. Le paramètre naturel pour le modèle de Bernoulli est alors $q = \ln \frac{p}{1-p}$, celui du modèle de Poisson est $\mu = \ln \lambda$, etc. De même, on voit que la décomposition exponentielle n'est pas unique : dès qu'il existe une décomposition, il en existe une infinité.

Nous ne détaillerons pas plus dans ce module la notion de famille exponentielle, mais il semblait nécessaire d'en parler succinctement car c'est une source de nombreux résultats en statistiques.

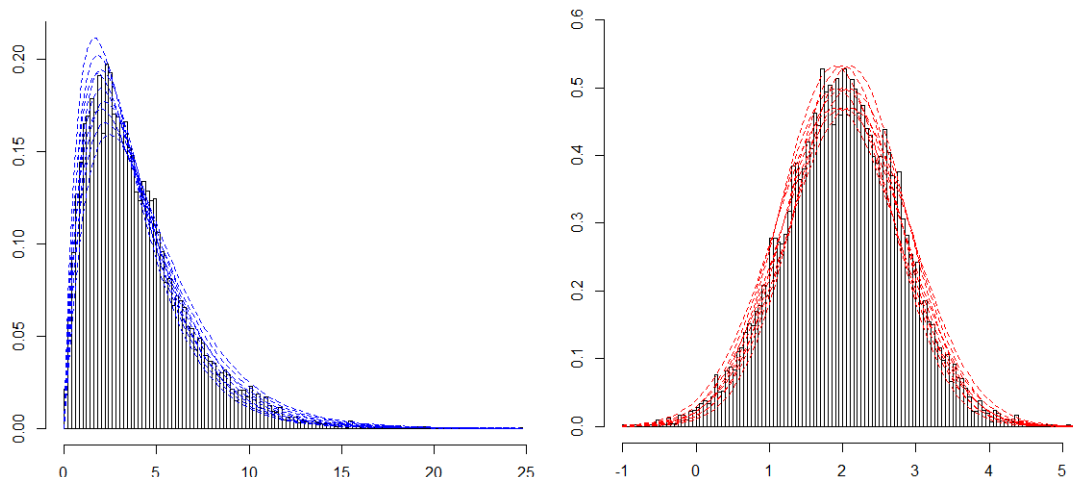
2 Estimation paramétrique

Définition 2.1 On appelle *estimateur de θ* toute statistique $\hat{\theta}_n = \hat{\theta}(X_1, \dots, X_n)$ sur un n -échantillon dont l'objectif est d'approximer la valeur de θ^* .

Un des enjeux principaux est donc de développer des méthodes dont la validité est uniforme en $\theta \in \Theta$: ainsi, la valeur θ^* est correctement traitée.

2.1 Estimation ponctuelle d'un paramètre

Pour comprendre le principe de l'estimation, on considère les deux exemples suivants. La forme de l'histogramme nous suggère une loi gamma $\Gamma(k, \lambda)$ à gauche et une loi normale $\mathcal{N}(\mu, \sigma^2)$ à droite. Comment déterminer les paramètres inconnus (k, λ) et (μ, σ^2) de sorte que les observations collent au modèle choisi ? La représentation des densités pour des paramètres variant légèrement montre que plusieurs solutions seraient envisageables et que l'on ne peut pas décider à l'œil nu. On va chercher à *estimer* ces paramètres, c'est-à-dire à les exprimer comme une fonction des observations seules.



2.1.1 Par la méthode des moments : moyenne (expérience) = espérance (théorique)

Dans la définition qui suit, on note $\hat{\mathbb{E}}[X^\ell]$ la valeur de $\mathbb{E}_\theta[X^\ell]$ dans laquelle on remplace θ par $\hat{\theta}_n$, c'est-à-dire par un estimateur.

Définition 2.2 Supposons que $\mathbb{E}_\theta[|X|^p] < +\infty$, et donc que X possède des moments d'ordre $p = \dim(\Theta)$. On appelle **estimateur des moments** de θ toute solution du système défini par

$$\forall \ell \in \{1, \dots, p\}, \quad \hat{\mathbb{E}}[X^\ell] = \frac{1}{n} \sum_{k=1}^n X_k^\ell. \quad \left. \begin{array}{l} \text{moment empirique} \\ \text{d'ordre } \ell \end{array} \right\}$$

Parfois les moments empiriques d'ordre $\{1, \dots, p\}$ ne permettent pas de conclure (par exemple si $p = 1$ et que $\mathbb{E}_\theta[X] = 0$), auquel cas il faut aller voir plus loin dans les moments (un exemple est proposé ci-dessous).

Exemples. Quelques exemples d'application de la méthode :

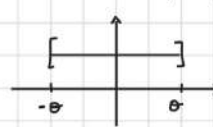
* $X \sim \mathcal{B}(p)$: $\mathbb{E}[X] = p$, $\hat{\mathbb{E}}[X] = \hat{p}_n$. Donc $\hat{p}_n = \frac{1}{n} \sum_{k=1}^n X_k = \bar{X}_n$ moment d'ordre 1, on a $\ell=1$ estimateur des moments de p : $\hat{p}_n = \bar{X}_n$
* $X \sim \mathcal{P}(\lambda)$: $\mathbb{E}[X] = \lambda$, $\hat{\mathbb{E}}[X] = \hat{\lambda}_n$. Donc $\hat{\lambda}_n = \bar{X}_n$ estimateur des moments de λ : $\hat{\lambda}_n = \bar{X}_n$
* $X \sim \mathcal{E}(\lambda)$: $\mathbb{E}[X] = \frac{1}{\lambda}$, $\hat{\mathbb{E}}[X] = \frac{1}{\hat{\lambda}_n}$. Donc $\frac{1}{\hat{\lambda}_n} = \bar{X}_n$ estimateur des moments de λ : $\hat{\lambda}_n = \frac{1}{\bar{X}_n}$
* $X \sim \mathcal{N}(\mu, \sigma^2)$: $\mathbb{E}[X] = \mu$, $\hat{\mathbb{E}}[X] = \hat{\mu}_n$, $\mathbb{E}[X^2] = \sigma^2 + \mu^2$, $\hat{\mathbb{E}}[X^2] = \hat{\sigma}_n^2 + \hat{\mu}_n^2$

Donc $\left\{ \begin{aligned} \hat{\mu}_n &= \bar{X}_n \\ \hat{\sigma}_n^2 &= \frac{1}{n} \sum_{k=1}^n X_k^2 \end{aligned} \right\} \Leftrightarrow \left\{ \begin{aligned} \hat{\mu}_n &= \bar{X}_n \\ \hat{\sigma}_n^2 &= \frac{1}{n} \sum_{k=1}^n X_k^2 - \bar{X}_n^2 = \frac{1}{n} \sum_{k=1}^n (X_k - \bar{X}_n)^2 = S_n^2 \end{aligned} \right.$

Estimateur des moments de (μ, σ^2) : $(\hat{\mu}_n, \hat{\sigma}_n^2) = (\bar{X}_n, S_n^2)$

* $\mathcal{U}(0, \theta)$: $\hat{\theta}_n = 2 \bar{X}_n$, $\mathbb{E}[X] = \frac{\theta}{2}$

* $X \sim \mathcal{U}(-\theta, \theta)$, $\theta > 0$ $\mathbb{E}[X] = 0$ $\hat{\mathbb{E}}[X] = 0 \dots ?$



$f_X(x) = \frac{1}{2\theta} \mathbb{1}_{\{|x| \leq \theta\}}$

$\mathbb{E}[X^2] = \text{Var}(X) + (\mathbb{E}[X])^2 = \frac{(2\theta)^2}{12} = \frac{\theta^2}{3}$, $\hat{\mathbb{E}}[X^2] = \frac{\hat{\sigma}_n^2}{3}$

estimateur des moments de θ : $\hat{\theta}_n = \pm \sqrt{\frac{3}{n} \sum_{k=1}^n X_k^2}$
car $\theta > 0$

En extrapolant la définition, la méthode des moments ne conduit généralement pas à une solution unique : lorsque $X \sim \mathcal{E}(\lambda)$ on vient de voir qu'un estimateur des moments est donné par $\hat{\lambda}_n = \frac{1}{\bar{X}_n}$ mais en considérant $\mathbb{E}[X^2] = \frac{2}{\lambda^2}$, on aurait tout aussi bien pu choisir

$$\hat{\lambda}_n = \sqrt{\frac{2n}{Q_n}} \quad \text{avec} \quad Q_n = \sum_{k=1}^n X_k^2.$$

C'est pour cela qu'en général on préfère parler d'un estimateur des moments.

Proposition 2.1 Supposons que $\mathbb{E}_\theta[|X|^q] < +\infty$. Alors pour tout $1 \leq \ell \leq q$, on a

$$\frac{1}{n} \sum_{k=1}^n X_k^\ell \xrightarrow{\text{p.s.}} \mathbb{E}_\theta[X^\ell].$$

Preuve. Par hypothèse, $\mathbb{E}_\theta[X^\ell]$ existe. Il suffit donc d'appliquer la LFGN au n -échantillon $X_1^\ell, \dots, X_n^\ell$.

Proposition 2.2 Soit X_1, \dots, X_n un n -échantillon dont l'espérance $\mathbb{E}_\theta[X]$ et la variance $\mathbb{V}_\theta(X)$ existent. Alors, la méthode des moments leur affecte respectivement comme estimateurs

$$\bar{X}_n = \frac{1}{n} \sum_{k=1}^n X_k \quad \text{et} \quad S_n^2 = \frac{1}{n} \sum_{k=1}^n (X_k - \bar{X}_n)^2.$$

Preuve :

$$l=1 \quad \hat{\mathbb{E}}[X] = \frac{1}{n} \sum_{k=1}^n X_k = \bar{X}_n$$

$$l=2 \quad \hat{\mathbb{E}}[X^2] = \frac{1}{n} \sum_{k=1}^n X_k^2$$

Donc $\widehat{\text{Var}}(X) = \hat{\mathbb{E}}[X^2] - (\hat{\mathbb{E}}[X])^2 = \frac{1}{n} \sum_{k=1}^n X_k^2 - (\bar{X}_n)^2$

Or $\frac{1}{n} \sum_{k=1}^n (X_k - \bar{X}_n)^2 = \frac{1}{n} \sum_{k=1}^n X_k^2 - \frac{2}{n} \sum_{k=1}^n X_k \cdot \bar{X}_n + \frac{n}{n} \bar{X}_n^2$

$$= \frac{1}{n} \sum_{k=1}^n X_k^2 - 2 \bar{X}_n + \bar{X}_n^2 = \frac{1}{n} \sum_{k=1}^n X_k^2 - \bar{X}_n^2$$

Remarque. Par la formule de König, on a la simplification (à savoir redémontrer),

$$\frac{1}{n} \sum_{k=1}^n (X_k - \bar{X}_n)^2 = \frac{1}{n} \sum_{k=1}^n X_k^2 - (\bar{X}_n)^2.$$

Définition 2.3 Les statistiques \bar{X}_n et S_n^2 sont appelées moyenne et variance empiriques de l'échantillon, respectivement. Elles jouent un rôle fondamental en statistique inférentielle. Pour des raisons qui seront explicitées par la suite, on définit également la variance empirique corrigée de l'échantillon par biais

$$S_n^{*2} = \frac{1}{n-1} \sum_{k=1}^n (X_k - \bar{X}_n)^2 = \frac{n}{n-1} S_n^2.$$

Remarque. Il peut arriver que $\mathbb{E}_\theta[X]$ n'existe pas, ce qui rend la méthode inapplicable. L'idée est alors de chercher une fonction h telle que $\mathbb{E}_\theta[h(X)]$ existe et de produire le même raisonnement, en moyennisant les $h(X_k)$ en lieu et place des X_k . On parle de méthode des moments généralisée (un exemple est donné en TD).

2.1.2 Par maximum de vraisemblance

Définition 2.4 Soit $\underline{X} = (X_1, \dots, X_n)$ un n -échantillon dont la v.a.r. parente X est de densité $f_X(\cdot; \theta)$ paramétrée par $\theta \in \Theta$. On définit sa fonction de vraisemblance au point \underline{X} par

$$\theta \mapsto \ell_X(\underline{X}; \theta) = \left[\prod_{k=1}^n f_X(X_k; \theta) \right] \quad \begin{array}{l} \rightarrow \underline{X} \text{ est fixé, seulement } \theta \text{ varie} \\ \text{cas particulier des variables indép. et de m loi} \end{array}$$

l pour likelihood

En fait, la vraisemblance désigne plus généralement la densité de l'échantillon vue comme une fonction de θ , mais la simplification ci-dessus est obtenue en faisant intervenir l'hypothèse que les variables sont i.i.d. (ce qui est rarement le cas dans les études concrètes). La présence du produit fait que l'on préfère généralement travailler sur la log-vraisemblance, lorsque c'est possible.

Définition 2.5 Soit $\underline{X} = (X_1, \dots, X_n)$ un n -échantillon de vraisemblance $\ell_X(\cdot; \theta)$. On définit sa fonction de log-vraisemblance (si elle existe) par

$$\theta \mapsto \ell\ell_X(\underline{X}; \theta) = \ln \ell_X(\underline{X}; \theta) = \left[\sum_{k=1}^n \ln f_X(X_k; \theta) \right] \quad \begin{array}{l} \leftarrow \text{somme plus facile à manipuler donc} \\ \text{on va plutôt utiliser celui-là} \end{array}$$

positive uniformément en θ

La vraisemblance étant ici un produit de densités, il est clair que $\ell_X(\cdot; \theta) \geq 0$. En pratique il faut vérifier que $\ell_X(\cdot; \theta) > 0$ pour tout $\theta \in \Theta$ pour pouvoir travailler avec $\ell_X(\cdot; \theta)$. Pour une modélisation, la fonction de vraisemblance est étroitement liée à la probabilité d'observer \underline{X} lorsque le paramètre prend la valeur θ . Il paraît donc naturel de chercher la valeur du paramètre qui maximise $\ell_X(\underline{X}; \theta)$. Si $\ell_X(\underline{X}; \theta_1) > \ell_X(\underline{X}; \theta_2)$, alors θ_1 est plus vraisemblable que θ_2 pour l'observation de \underline{X} .

Définition 2.6 Lorsque $\ell_X(\cdot; \theta)$ est dérivable par rapport à θ sur Θ , on appelle **équations de vraisemblance** les équations définies par le système

$$\nabla_{\theta} \ell_X(\underline{X}; \theta) = 0 \quad \text{ou de manière équivalente} \quad \nabla_{\theta} \log \ell_X(\underline{X}; \theta) = 0$$

si $\ell_X(\cdot; \theta)$ est bien définie. Toute statistique $\theta_n^{\nabla} = \theta^{\nabla}(X_1, \dots, X_n)$ satisfaisant

$$\nabla_{\theta} \ell_X(\underline{X}; \theta_n^{\nabla}) = 0$$

est une racine des équations de vraisemblance (REV).

rien ne garantit ni l'existence ni l'unicité

Définition 2.7 Tout estimateur θ_n^{∇} de θ est un estimateur du maximum de vraisemblance (EMV) s'il vérifie

$$\hat{\theta}_n \in \arg \max_{\theta \in \Theta} \ell_X(\underline{X}; \theta).$$

MLE en anglais

Il n'est donc pas clair que REV et EMV coïncident, une analyse plus fine est nécessaire en ce qui concerne le maximum global de $\ell_X(\underline{X}; \theta)$. Par exemple, lorsque c'est possible, il faudra vérifier que la hessienne $H_{\theta} \log \ell_X(\underline{X}; \theta_n^{\nabla})$ est définie négative pour s'assurer que le point critique θ_n^{∇} est bien un maximum. Rien ne garantit au préalable l'existence et l'unicité de l'EMV.

Exemples. Quelques exemples d'application de la méthode :

* $X \sim \mathcal{B}(p)$, $p \in]0, 1[$, $\ell_X(\underline{X}; p) = \prod_{k=1}^n p^{x_k} (1-p)^{1-x_k}$
 $\ell_X(\underline{X}; p) = n \bar{X}_n \ln(p) + n(1 - \bar{X}_n) \ln(1-p)$, ℓ_X dérivable sur $]0, 1[$
 $\frac{\partial}{\partial p} \log \ell_X(\underline{X}; p) = \frac{n \bar{X}_n}{p} - \frac{n(1 - \bar{X}_n)}{1-p} = \frac{n \bar{X}_n (1-p) - n p (1 - \bar{X}_n)}{p(1-p)} = \frac{n(\bar{X}_n - p)}{p(1-p)}$
 $= 0 \Leftrightarrow p_n^{\nabla} = \bar{X}_n$ (unique REV)
 $\frac{\partial^2}{\partial p^2} \log \ell_X(\underline{X}; p) = -\frac{n \bar{X}_n}{p^2} - \frac{n(1 - \bar{X}_n)}{(1-p)^2} = -\frac{n \bar{X}_n (1-p)^2 - n p^2 (1 - \bar{X}_n)}{(p(1-p))^2} < 0 \quad \forall p$
 L'unique EMV vaut $\hat{p}_n = \bar{X}_n$

* $X \sim \mathcal{P}(\lambda)$, $\lambda > 0$, $\ell_X(\underline{X}; \lambda) = \prod_{k=1}^n e^{-\lambda} \frac{\lambda^{x_k}}{x_k!} = e^{-n\lambda} \frac{\lambda^{n \bar{X}_n}}{P_n}$ avec $P_n = \prod_{k=1}^n (x_k!)$
 $\ell_X(\underline{X}; \lambda) = -n\lambda + n \bar{X}_n \ln(\lambda) - \ln P_n$
 $\frac{\partial}{\partial \lambda} \log \ell_X(\underline{X}; \lambda) = -n + \frac{n \bar{X}_n}{\lambda} = \frac{n(\bar{X}_n - \lambda)}{\lambda} \Leftrightarrow \lambda_n^{\nabla} = \bar{X}_n$ (unique REV)
 $\frac{\partial^2}{\partial \lambda^2} \log \ell_X(\underline{X}; \lambda) = -\frac{n \bar{X}_n}{\lambda^2} < 0 \quad \forall \lambda \quad (\exists k, x_k > 0)$
 L'unique EMV vaut $\hat{\lambda}_n = \bar{X}_n$

$$* X \sim \mathcal{N}(\mu, \sigma^2) \quad (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+^*$$

$$l_X(\underline{X}; \mu, \sigma^2) = \prod_{k=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(X_k - \mu)^2}{2\sigma^2}} = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{k=1}^n (X_k - \mu)^2\right) > 0 \quad \forall \mu, \sigma^2$$

$$ll_X(\underline{X}; \mu, \sigma^2) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} \sum_{k=1}^n (X_k - \mu)^2$$

l_X dérivable sur $\mathbb{R} \times \mathbb{R}_+^*$

$$\nabla_{\mu, \sigma^2} ll_X(\underline{X}; \mu, \sigma^2) = \begin{pmatrix} \frac{n}{\sigma^2} (\bar{X}_n - \mu) \\ -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^2} \sum_{k=1}^n (X_k - \mu) \end{pmatrix} = 0 \Leftrightarrow \begin{cases} \mu_n^\circ = \bar{X}_n \\ \frac{n(S_n^2 - \sigma^2)}{2\sigma^4} = 0 \end{cases}$$

$$\Leftrightarrow \begin{cases} \mu_n^\circ = \bar{X}_n \\ \sigma_n^{2\circ} = S_n^2 \end{cases} \quad \text{unique REV.}$$

$$H_{\mu, \sigma^2} ll_X(\underline{X}; \mu, \sigma^2) = \begin{pmatrix} -\frac{n}{\sigma^2} & -\frac{n(\bar{X}_n - \mu)}{\sigma^4} \\ -\frac{n(\bar{X}_n - \mu)}{\sigma^4} & \frac{n}{2\sigma^4} - \frac{1}{\sigma^6} \sum_{k=1}^n (X_k - \mu)^2 \end{pmatrix}$$

$$H_{\mu, \sigma^2} ll_X(\underline{X}; \mu_n^\circ, \sigma_n^{2\circ}) = \begin{pmatrix} -\frac{n}{S_n^2} & 0 \\ 0 & -\frac{n}{2(S_n^2)^2} \end{pmatrix} < 0 \quad (\text{def. nég}).$$

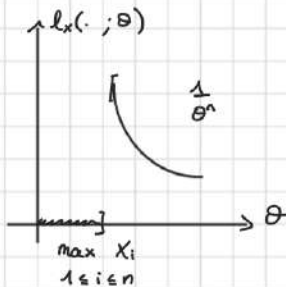
l'unique EMV vaut (\bar{X}_n, S_n^2)

$$* X \sim \mathcal{U}(0, \theta) \quad f_X(\theta) = \frac{1}{\theta} \mathbb{1}_{\{x \in [0, \theta]\}}$$

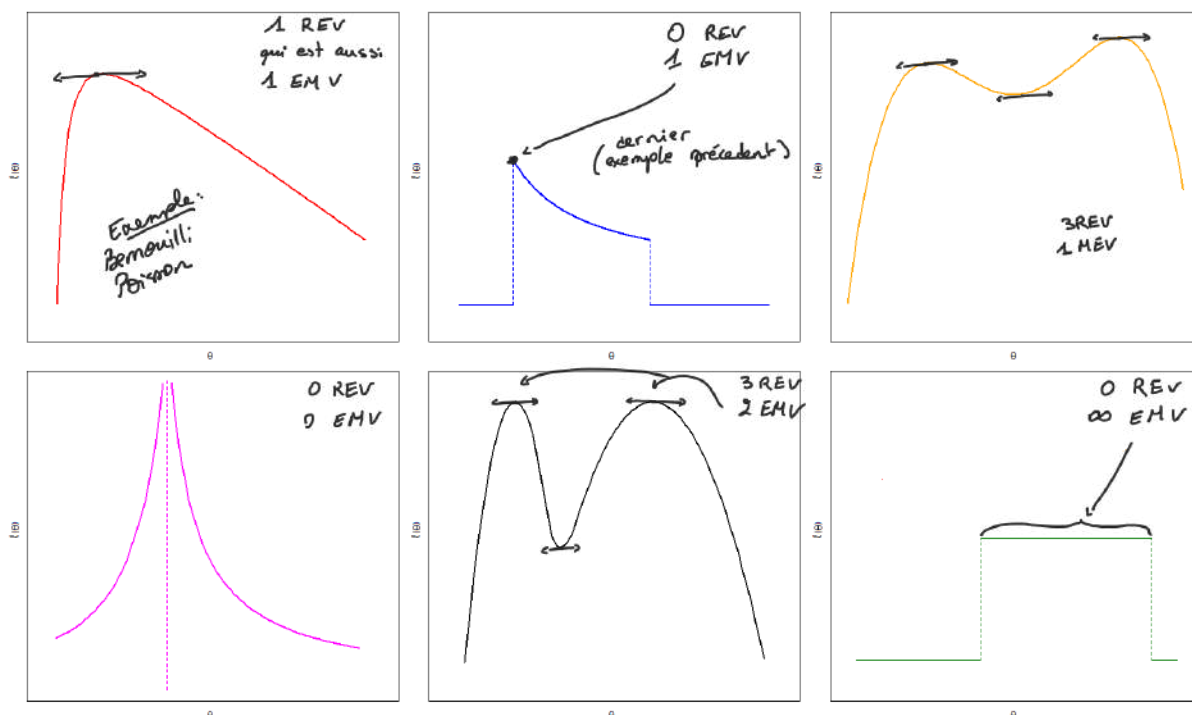
$$l_X(\underline{X}; \theta) = \prod_{i=1}^n \frac{1}{\theta} \mathbb{1}_{\{x \in [0, \theta]\}} = \begin{cases} \frac{1}{\theta^n} & \text{si } \theta \geq \max \{X_i; i = 1, \dots, n\} \\ 0 & \text{sinon} \end{cases}$$

Pour maximiser $l_X(\underline{X}; \theta)$ on doit prendre θ minimum sous contraintes $\theta \geq \max_{1 \leq i \leq n} X_i$

C'est-à-dire $\hat{\theta}_n = \max_{1 \leq i \leq n} X_i \rightarrow \text{EMV}$



Exemples. Commenter la présence de REV et d'EMV dans les exemples de vraisemblances (unidimensionnelles) suivants.



Remarque. L'EMV n'est que rarement accessible par le calcul direct comme dans les exemples précédents. L'algorithme EM est une variante ingénieuse qui permet, en introduisant une certaine information bien choisie dans le modèle, d'approximer par une succession d'étapes de calcul d'espérance (E) et de maximisation (M) un maximum local de la vraisemblance. En général, lorsque les données ne sont pas indépendantes, la vraisemblance n'a pas d'écriture exhaustive et des solveurs numériques se chargent d'approximer la valeur recherchée. L'essentiel est de montrer *a minima* l'existence (et, mieux, l'unicité).

On conclut cette section par la propriété d'invariance de l'EMV par les transformations bijectives (en fait pas nécessairement bijectives... mais la propriété devient plus compliquée à démontrer).

Proposition 2.3 Pour toute application bijective h définie sur Θ ,

$$\hat{\theta}_n \text{ est l'EMV de } \theta \implies h(\hat{\theta}_n) \text{ est l'EMV de } h(\theta).$$

Théorème de préservation

$$\begin{aligned} \text{Soit } \mu = h(\theta). \text{ Alors } l_{\theta}(x) &= l_x(x; \theta) = l_x(x; h^{-1}(\mu)) = (l_x \circ h^{-1})(x; \mu) = l_{\mu}(x) \\ \hat{\theta}_n \text{ maximise } l_x(x; \theta) &\iff \hat{\theta}_n = h^{-1}(\hat{\mu}_n) \text{ maximise } l_x(x; \theta) = (l_x \circ h^{-1})(x; \mu). \end{aligned}$$

Exemples. Déterminer l'EMV de $\mathbb{P}(X = 0)$ dans le modèle de Poisson $X \sim \mathcal{P}(\lambda)$. Déterminer de même l'EMV de $\mathbb{V}(X)$ dans le modèle $\mathcal{U}([0, \theta])$.

$$\begin{aligned} \mathbb{P}(X=0) &= e^{-\lambda} \text{ si } X \sim \mathcal{P}(\lambda) \quad \text{Soit } \mu = h(\lambda) \text{ avec } h: \lambda \mapsto e^{-\lambda} \\ \text{On a } \hat{\mu}_n &= e^{-\bar{x}_n} \text{ comme EMV de } \mathbb{P}(X=0) \\ \text{Var}(X) &= \frac{\theta^2}{12} \text{ si } X \sim \mathcal{U}(0, \theta) \quad \text{Soit } v = h(\theta) \text{ avec } h: \theta \mapsto \frac{\theta^2}{12} \end{aligned}$$

On a $\hat{\sigma}_n^2 = \frac{(\max_{1 \leq k \leq n} X_k)^2}{12}$ comme EMV de $\text{Var}(X)$.

2.2 Principales propriétés des estimateurs

Un estimateur n'étant jamais unique, il est indispensable de définir des critères de comparaison pour, à terme, privilégier un estimateur plutôt qu'un autre. On distinguera les critères exacts (valables pour tout n) et les critères asymptotiques (valables pour $n \rightarrow +\infty$).

2.2.1 Biais

Définition 2.8 Un estimateur $\hat{\theta}_n$ de θ est non biaisé (ou sans biais) si

$$\forall n, \forall \theta \in \Theta, \quad \mathbb{E}_\theta[\hat{\theta}_n] = \theta.$$

Il est asymptotiquement non biaisé (ou asymptotiquement sans biais) si

$$\forall \theta \in \Theta, \quad \lim_{n \rightarrow +\infty} \mathbb{E}_\theta[\hat{\theta}_n] = \theta.$$

Proposition 2.4 Si l'espérance existe, alors la moyenne empirique \bar{X}_n est un estimateur sans biais de $\mathbb{E}_\theta[X]$. Si de plus la variance existe, alors la variance empirique S_n^2 est un estimateur biaisé de $\mathbb{V}_\theta(X)$ mais asymptotiquement sans biais, et son biais vaut

$$\mathbb{E}_\theta[S_n^2] - \mathbb{V}_\theta(X) = -\frac{\mathbb{V}_\theta(X)}{n}.$$

La variance empirique corrigée S_n^{*2} est un estimateur sans biais de $\mathbb{V}_\theta(X)$. (X_k i.i.d dans tous les cours)

Preuve: Soit $e = \mathbb{E}_\theta[X]$ et $\sigma^2 = \text{Var}_\theta[X]$

On a $\mathbb{E}[\hat{\theta}_n] = \mathbb{E}[\bar{X}_n] = \mathbb{E}\left[\frac{1}{n} \sum_{k=1}^n X_k\right] \stackrel{\text{E linéaire}}{=} \frac{1}{n} \sum_{k=1}^n \mathbb{E}[X_k] \stackrel{X_k \text{ i.i.d}}{=} \frac{1}{n} \cdot n \cdot e = e \quad \forall n, \forall e.$

On a $\mathbb{E}[\hat{\sigma}_n^2] = \mathbb{E}[S_n^2] = \mathbb{E}\left[\frac{1}{n} \sum_{k=1}^n X_k^2\right] - \mathbb{E}[\bar{X}_n^2] = \frac{n}{n} \mathbb{E}[X^2] - \frac{n}{n^2} \mathbb{E}[X^2] - \frac{n^2 - n}{n^2} \mathbb{E}[X]^2$

$\mathbb{E}[\bar{X}_n^2] = \frac{1}{n^2} \sum_{k=1}^n \sum_{l=1}^n \mathbb{E}[X_k X_l] = \frac{1}{n^2} \mathbb{E}[X^2] + \frac{n^2 - n}{n^2} \mathbb{E}[X]^2$

$\mathbb{E}[\hat{\sigma}_n^2] = \left(1 - \frac{1}{n}\right) (\sigma^2 + e^2) - \frac{n-1}{n} e^2 = \frac{n-1}{n} \sigma^2 \neq \sigma^2$

Donc $\mathbb{E}[S_n^2] \neq \text{Var}(X)$ mais $\mathbb{E}[S_n^2] \xrightarrow{n \rightarrow +\infty} \text{Var}(X)$

En posant $S_n^{*2} = \frac{n}{n-1} S_n^2 = \frac{1}{n-1} \sum_{k=1}^n (X_k - \bar{X}_n)^2$, on obtient un estimateur sans biais de la variance

$$\triangle S_n^2 = \frac{1}{n} \left(\sum_{k=1}^n X_k \right) - \bar{X}_n^2$$

$$S_n^* \neq \frac{1}{n-1} \left(\sum_{k=1}^n X_k \right) - \bar{X}_n^2 = \frac{1}{n-1} \sum_{k=1}^n X_k - \frac{n}{n-1} \bar{X}_n^2$$

Exemples. Étudier quelques cas usuels, ainsi que le biais de l'estimateur $\hat{\lambda}_n = \frac{1}{\bar{X}_n}$ dans le modèle $X \sim \mathcal{E}(\lambda)$.

- * $X \sim \mathcal{B}(p)$, $\hat{p}_n = \bar{X}_n$ est sans biais pour $p = \mathbb{E}[X]$
- * $X \sim \mathcal{P}(\lambda)$, $\hat{\lambda}_n = \bar{X}_n$ est sans biais pour $\lambda = \mathbb{E}[X]$
- * $X \sim \mathcal{N}(\mu, \sigma^2)$ $(\hat{\mu}_n, \hat{\sigma}_n^2) = (\bar{X}_n, S_n^2)$ est biaisé pour $(\mu, \sigma^2) = (\mathbb{E}[X], \text{Var}(X))$
 $(\hat{\mu}_n, \hat{\sigma}_n^{*2}) = (\bar{X}_n, S_n^{*2})$ est sans biais

- * $X \sim \mathcal{E}(\lambda)$ $\mathbb{E}[X] = \frac{1}{\lambda}$ un estimateur des moments : $\hat{\lambda}_n = \frac{1}{\bar{X}_n}$

$$\triangle \mathbb{E} \left[\frac{1}{\bar{X}_n} \right] \neq \frac{1}{\mathbb{E}[\bar{X}_n]} \quad \mathbb{E} \left[\frac{1}{\bar{X}_n} \right] = n \mathbb{E} \left[\frac{1}{S_n} \right] \text{ avec } S_n = \sum_{k=1}^n X_k$$

Théorème de transfert $\rightarrow = n \int \frac{1}{s} f(s) ds$ avec f la densité de S_n

S_n suit la loi Gamma / exponentielle. Donc on prend la densité de la loi gamma : $\Gamma(n, \lambda)$.

$$\mathbb{E} \left[\frac{1}{\bar{X}_n} \right] = n \int_{\mathbb{R}} \frac{1}{s} f_n(s, \lambda) ds = n \int_0^{+\infty} \lambda^n \frac{s^{n-1} e^{-\lambda s}}{s \Gamma(n)} ds$$

$$\text{avec } \Gamma(n) = \int_0^{+\infty} s^{n-1} e^{-s} ds \stackrel{\text{i.f.p}}{=} (n-1) \Gamma(n-1) = \dots = (n-1)!$$

$$\mathbb{E} \left[\frac{1}{\bar{X}_n} \right] = \frac{n \lambda^n}{\Gamma(n)} \int_0^{+\infty} s^{n-2} e^{-\lambda s} ds = \frac{n \lambda^n}{\Gamma(n)} \int_0^{+\infty} \left(\frac{u}{\lambda} \right)^{n-2} e^{-u} \frac{du}{\lambda}$$

$$= \frac{n \lambda}{\Gamma(n)} \Gamma(n-1) = \frac{n(n-2)! \lambda}{(n-1)!} = \frac{n}{n-1} \lambda \quad \text{donc } \hat{\lambda}_n \text{ est biaisé mais asymptotiquement sans biais}$$

Mais $\lambda_n^* = \frac{n-1}{n} \hat{\lambda}_n = \frac{n-1}{S_n}$ (avec $S_n = \sum_{k=1}^n X_k$) est sans biais pour λ

2.2.2 Risque quadratique

On définit la fonction de risque quadratique d'un estimateur par

$$R(\hat{\theta}_n, \theta) = \mathbb{E}_\theta[\|\hat{\theta}_n - \theta\|^2].$$

Définition 2.9 Soient deux estimateurs $\hat{\theta}_n$ et $\tilde{\theta}_n$ de θ . Pour le risque quadratique, $\hat{\theta}_n$ est **préférable** à $\tilde{\theta}_n$ si

$$\forall n, \forall \theta \in \Theta, \quad R(\hat{\theta}_n, \theta) \leq R(\tilde{\theta}_n, \theta).$$

S'il existe $\theta' \in \Theta$ tel que l'inégalité soit stricte, alors $\hat{\theta}_n$ est **meilleur** que $\tilde{\theta}_n$. De plus, $\hat{\theta}_n$ est **admissible** si aucun autre estimateur de θ n'est meilleur que lui. Sinon, il est **inadmissible**.

Exemples. Sur la base du risque quadratique, comparer $\hat{p}_n = \bar{X}_n$ et $\tilde{p}_n = X_1$ dans le modèle $X \sim \mathcal{B}(p)$. Comparer également $\hat{\sigma}_n^2 = S_n^2$ et $\tilde{\sigma}_n^2 = S_n^{*2}$ dans le modèle $X \sim \mathcal{N}(0, \sigma^2)$.

* $X \sim \mathcal{B}(p)$ $\hat{p}_n = \bar{X}_n$ est sans biais \rightarrow plus pertinent
 $\tilde{p}_n = X_1$ $\mathbb{E}[\tilde{p}_n] = \mathbb{E}[X_1] = p$ \tilde{p}_n est sans biais \rightarrow exp. absurde

$$R(\hat{p}_n, p) = \mathbb{E}[(\bar{X}_n - p)^2] = \text{Var}(\bar{X}_n) = \frac{\text{Var}(X)}{n} = \frac{p(1-p)}{n}$$

car $\mathbb{E}[\bar{X}_n] = p$ cf. preuve LFGN

$$R(\tilde{p}_n, p) = \mathbb{E}[(X_1 - p)^2] = \text{Var}(X_1) = p(1-p)$$

\bar{X}_n est préférable à X_1 , meilleur que X_1 pour $n \geq 2$ (pour $n=1$, ils sont égaux).

* $X \sim \mathcal{N}(0, \sigma^2)$ $\mathbb{E}[S_n^2] = \frac{n-1}{n}\sigma^2$

$$\begin{aligned} R(S_n^2, \sigma^2) &= \mathbb{E}[(S_n^2 - \sigma^2)^2] = \mathbb{E}[(S_n^2 - \mathbb{E}[S_n^2] + \mathbb{E}[S_n^2] - \sigma^2)^2] \\ &= \text{Var}(S_n^2) - \left(\frac{n-1}{n} - 1\right)^2 \sigma^4 + 2 \underbrace{\mathbb{E}[S_n^2 - \mathbb{E}[S_n^2]]}_{=0} (\mathbb{E}[S_n^2] - \sigma^2) \\ &= \text{Var}(S_n^2) - \frac{\sigma^4}{n^2} = \frac{(2n-1)}{n^2} \sigma^4 \end{aligned}$$

$$\begin{aligned} R(S_n^{*2}, \sigma^2) &= \mathbb{E}[(S_n^{*2} - \sigma^2)^2] = \text{Var}(S_n^{*2}) \text{ car } S_n^{*2} \text{ est sans biais : } \mathbb{E}[S_n^{*2}] = \sigma^2 \\ &= \text{Var}\left(\frac{n}{n-1} S_n^2\right) = \left(\frac{n}{n-1}\right)^2 \text{Var}(S_n^2) = \frac{n^2}{(n-1)^2} \left[\frac{(2n-1)}{n^2} \sigma^4 - \frac{\sigma^4}{n^2}\right] \\ &= \frac{2\sigma^4(n-1)}{(n-1)^2} = \frac{2\sigma^4}{n-1} \end{aligned}$$

On remarquera que $\frac{R(S_n^2, \sigma^2)}{R(S_n^{*2}, \sigma^2)} = \frac{(2n-1)(n-1)}{2n^2} = \frac{2n-1}{2n} \cdot \frac{n-1}{n} < 1 \quad \forall n, \forall \sigma^2$

S_n^2 est donc meilleur que S_n^{*2} bien que biaisé

Proposition 2.5 (Décomposition biais/variance) On a la décomposition

$$\forall \theta \in \Theta, \quad R(\hat{\theta}_n, \theta) = \|\mathbb{E}_\theta[\hat{\theta}_n] - \theta\|^2 + \text{Tr}(\mathbb{V}_\theta(\hat{\theta}_n)).$$

En particulier si $p = \dim(\Theta) = 1$,

$$\forall \theta \in \Theta, \quad R(\hat{\theta}_n, \theta) = (\mathbb{E}_\theta[\hat{\theta}_n] - \theta)^2 + \mathbb{V}_\theta(\hat{\theta}_n).$$

Preuve pour $p=1$:

$$\begin{aligned} \forall \theta, R(\hat{\theta}_n, \theta) &= \mathbb{E}_\theta[(\hat{\theta}_n - \theta)^2] = \mathbb{E}_\theta[(\hat{\theta}_n - \mathbb{E}_\theta[\hat{\theta}_n] + \mathbb{E}_\theta[\hat{\theta}_n] - \theta)^2] \\ &= \mathbb{E}_\theta[(\hat{\theta}_n - \mathbb{E}_\theta[\hat{\theta}_n])^2] + (\mathbb{E}_\theta[\hat{\theta}_n] - \theta)^2 + 2(\mathbb{E}_\theta[\hat{\theta}_n] - \theta)(\underbrace{\mathbb{E}_\theta[\hat{\theta}_n - \mathbb{E}_\theta[\hat{\theta}_n]]}_{=0}) \\ &= \underbrace{\text{Var}_\theta(\hat{\theta}_n)}_{\text{variance}} + \underbrace{(\mathbb{E}_\theta[\hat{\theta}_n] - \theta)^2}_{\text{biais au carré}} \end{aligned}$$

Théorème
d'amélioration

Théorème 2.1 (Théorème de Rao-Blackwell) On considère un estimateur $\hat{\theta}_n$ de θ et une statistique $T_n = T(X_1, \dots, X_n)$ telle que la loi de $(X_1, \dots, X_n) | T_n$ ne dépend pas de θ . Alors,

donc cet estimateur est exhaustive

$$\tilde{\theta}_n = \mathbb{E}_\theta[\hat{\theta}_n | T_n]$$

est un estimateur de θ préférable à $\hat{\theta}_n$. De plus, $\hat{\theta}_n$ et $\tilde{\theta}_n$ ont le même biais.

Preuve pour $p=1$ (on admettra les propriétés élémentaires relatives à l'espérance conditionnelle).

$$\text{Ici } \tilde{\theta}_n \text{ préférable à } \hat{\theta}_n \Leftrightarrow \text{Var}(\tilde{\theta}_n) \leq \text{Var}(\hat{\theta}_n) \Leftrightarrow \mathbb{E}[(\tilde{\theta}_n - \theta)^2] \leq \mathbb{E}[(\hat{\theta}_n - \theta)^2]$$

→ Pour montrer que $\tilde{\theta}_n$ est non biaisé ; on a :

$$\cdot \mathbb{E}[\tilde{\theta}_n] = \mathbb{E}[\hat{\theta}_n | T_n] = \mathbb{E}[\hat{\theta}_n] = \theta$$

$$\begin{aligned} \cdot \text{Var}[\tilde{\theta}_n] &= \mathbb{E}[(\tilde{\theta}_n - \theta)^2] = \mathbb{E}[(\mathbb{E}(\hat{\theta}_n | T_n) - \theta)^2] = \mathbb{E}[(\mathbb{E}(\hat{\theta}_n | T_n) - \mathbb{E}(\theta | T_n))^2] \\ &= \mathbb{E}[\mathbb{E}[(\hat{\theta}_n - \theta)^2 | T_n]] \leq \mathbb{E}[\mathbb{E}[(\hat{\theta}_n - \theta)^2 | T_n]] = \mathbb{E}[(\hat{\theta}_n - \theta)^2] \end{aligned}$$

d'après l'inégalité de Jensen pour l'espérance conditionnelle

En conclusion $\tilde{\theta}_n$ est un meilleur estimateur que $\hat{\theta}_n$ en terme de variance

Exemple. Redémontrer que \bar{X}_n est préférable à X_1 dans le modèle $X \sim \mathcal{B}(p)$ en utilisant l'amélioration de Rao-Blackwell.

Posons $\hat{\theta}_n = X_1$, $\mathbb{E}[\hat{\theta}_n] = \mathbb{E}[X_1] = p$
On veut mq $T_n = \bar{X}_n$ alors T_n est exhaustive. On a tout d'abord :

$$\begin{aligned} \mathbb{P}(X_1 = 1 | \bar{X}_n = \frac{s}{n}) &= \frac{\mathbb{P}(\bar{X}_n = \frac{s}{n} | X_1 = 1) \mathbb{P}(X_1 = 1)}{\mathbb{P}(\bar{X}_n = \frac{s}{n})} = p \cdot \frac{\mathbb{P}(n\bar{X}_n = s | X_1 = 1)}{\mathbb{P}(n\bar{X}_n = s)} \\ &= p \frac{\mathbb{P}(n\bar{X}_{n-1} = s-1)}{\mathbb{P}(n\bar{X}_n = s)} \end{aligned}$$

$$\text{or } n\bar{X}_n \sim \mathcal{B}(n, p) \Leftrightarrow \mathbb{P}(n\bar{X}_n = k) = \binom{n}{k} p^k (1-p)^{n-k} \quad \forall k \in [0, n]$$

$$\text{Donc } \mathbb{P}(X_1 = 1 \mid \bar{X}_n = \frac{s}{n}) = p \frac{\binom{n-1}{s-1} p^{s-1} (1-p)^{n-s}}{\binom{n}{s} p^s (1-p)^{n-s}} = \frac{\binom{n-1}{s-1}}{\binom{n}{s}} = \frac{(n-1)!}{(s-1)!(n-s)!} \frac{(n-s)! s!}{n!} = \frac{s}{n}$$

$$\text{Donc } \tilde{\Theta}_n = \bar{X}_n$$

Pour montrer que T_n est exhaustive, on pose $(k_1, \dots, k_n) \in \{0, 1\}^n$ et on regarde :

$$\mathbb{P}(X_1 = k_1, \dots, X_n = k_n \mid \bar{X}_n = \frac{s}{n}) = \begin{cases} 0 & \text{si } \sum_{i=1}^n k_i \neq s \\ \frac{\mathbb{P}(X_1 = k_1, \dots, X_n = k_n)}{\mathbb{P}(T_n = \frac{s}{n})} = \frac{p^s (1-p)^{n-s}}{\binom{n}{s} p^s (1-p)^{n-s}} = \frac{1}{\binom{n}{s}} \quad \text{si } p \end{cases}$$

Voir l'exo du TD (loi géométrique)

$$n T_n = n \bar{X}_n \sim \mathcal{B}(n, p)$$

Remarque. Lorsque, comme dans le théorème précédent, la loi de $(X_1, \dots, X_n) \mid T_n$ ne dépend pas de θ , on dit que la statistique T_n est *exhaustive* pour θ . Il s'agit là encore d'une notion très importante en statistique théorique, source de nombreux travaux, sur laquelle nous n'avons pas le temps de nous attarder (un exercice de TD y est dévolu).

Remarque. On considère généralement le risque quadratique (dans L^2) car c'est probablement le plus simple à manipuler, mais on pourrait définir de même le risque L^p pour tout $p \geq 1$ donné par

$$R_p(\hat{\theta}_n, \theta) = \mathbb{E}_\theta[\|\hat{\theta}_n - \theta\|^p].$$

2.2.3 Consistance

Définition 2.10 Un estimateur $\hat{\theta}_n$ de θ est *faiblement consistant* si

$$\forall \theta \in \Theta, \quad \hat{\theta}_n \xrightarrow{\mathbb{P}} \theta. \quad \forall \varepsilon > 0 \quad \lim_{n \rightarrow +\infty} \mathbb{P}(|\hat{\theta}_n - \theta| \geq \varepsilon) = 0$$

Il est de plus *fortement consistant* si

$$\forall \theta \in \Theta, \quad \hat{\theta}_n \xrightarrow{\text{p.s.}} \theta. \quad \mathbb{P}(\omega \in \Omega : \lim_{n \rightarrow +\infty} \hat{\theta}_n(\omega) = \theta(\omega)) = 1$$

Exemples. Reprendre les cas étudiés dans la section sur le biais.

$$\begin{aligned} & * X \sim \mathcal{E}(\lambda) \quad \mathbb{E}[X] = \frac{1}{\lambda} \quad \bar{X}_n \xrightarrow[n \rightarrow +\infty]{\text{p.s.}} \frac{1}{\lambda} \quad \text{d'après LFGN} \\ & \text{Donc en particulier} \quad \bar{X}_n \xrightarrow[n \rightarrow +\infty]{\mathbb{P}} \frac{1}{\lambda} \\ & \text{D'après CMT} \quad \frac{1}{\bar{X}_n} \xrightarrow[n \rightarrow +\infty]{\text{p.s. et } \mathbb{P}} \lambda \end{aligned}$$

Continuous map theorem

En général, on établit la consistance forte par application de la LFGN et du CMT. Lorsque c'est impossible, on a le critère suivant souvent facile à appliquer et qui donne la consistance faible.

Proposition 2.6 Soit $\hat{\theta}_n$ un estimateur de θ (asymptotiquement) sans biais et tel que $\lim_{n \rightarrow +\infty} E[\hat{\theta}_n] = \theta$
 $\forall \theta \in \Theta, \quad \lim_{n \rightarrow +\infty} \nabla_{\theta}(\hat{\theta}_n) = 0.$

Alors $\hat{\theta}_n$ est faiblement consistant.

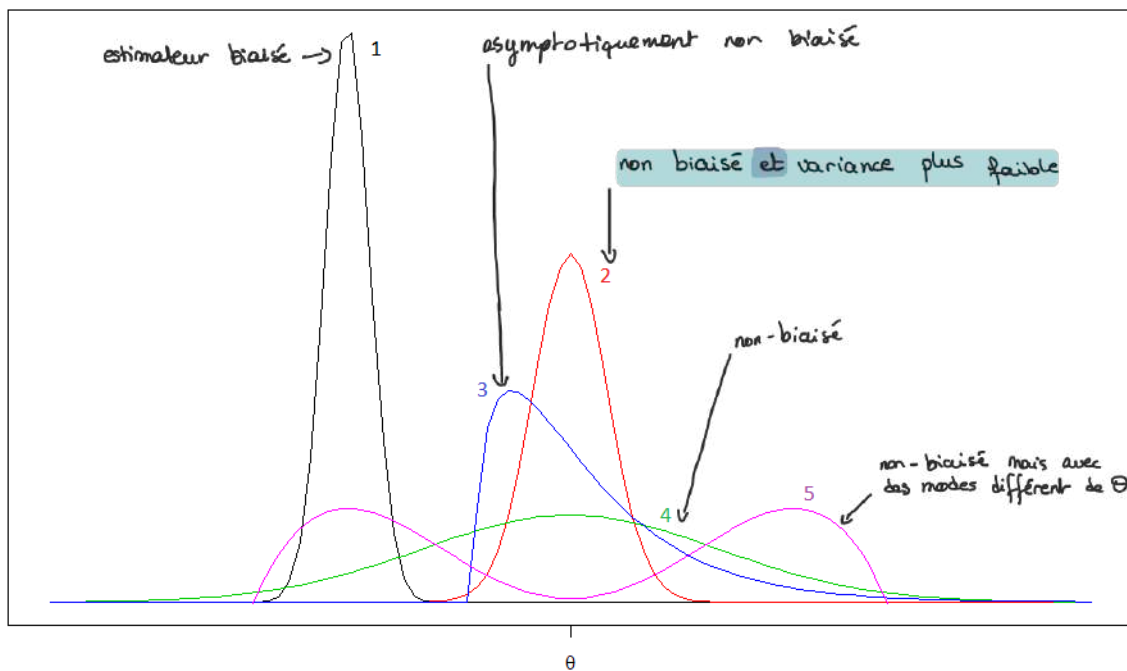
Preuve : Pour montrer la consistance faible, soit $\varepsilon > 0$

$$P(|\hat{\theta}_n - \theta| \geq \varepsilon) = P((\hat{\theta}_n - \theta)^2 \geq \varepsilon^2) \leq \frac{E[(\hat{\theta}_n - \theta)^2]}{\varepsilon^2} = \frac{\text{Var}(\hat{\theta}_n)}{\varepsilon^2} \xrightarrow[n \rightarrow +\infty]{P} 0$$
↑
Markov

Remarque. Il est évident qu'un estimateur fortement consistant l'est également au sens faible.

2.3 Estimation de variance minimale

On a représenté ci-dessous la distribution de cinq estimateurs $\hat{\theta}_n^1, \dots, \hat{\theta}_n^5$ autour du paramètre θ . Lequel devrait-on naturellement préférer et pourquoi ?



On utilisera dans ce qui suit les opérateurs

$$\nabla_{\theta} \equiv \begin{pmatrix} \frac{\partial}{\partial \theta_1} \\ \frac{\partial}{\partial \theta_2} \\ \vdots \\ \frac{\partial}{\partial \theta_p} \end{pmatrix} \quad \text{et} \quad H_{\theta} \equiv \begin{pmatrix} \frac{\partial^2}{\partial \theta_1^2} & \frac{\partial^2}{\partial \theta_1 \partial \theta_2} & \cdots & \frac{\partial^2}{\partial \theta_1 \partial \theta_p} \\ \frac{\partial^2}{\partial \theta_2 \partial \theta_1} & \frac{\partial^2}{\partial \theta_2^2} & \cdots & \frac{\partial^2}{\partial \theta_2 \partial \theta_p} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2}{\partial \theta_p \partial \theta_1} & \frac{\partial^2}{\partial \theta_p \partial \theta_2} & \cdots & \frac{\partial^2}{\partial \theta_p^2} \end{pmatrix}.$$

Les résultats de cette section sont soumis à de nombreuses hypothèses techniques. Lorsque par la suite on parlera de *modèle régulier*, cela sous-entendra qu'il satisfait les hypothèses ci-dessous. On appelle ν la mesure associée à la densité de X et l'on rappelle que $p = \dim(\Theta)$.

- (H₁). $E = X(\Omega)$ ne dépend pas de θ et pour tout $\theta \in \Theta$, $f_X(\cdot, \theta) > 0$.
- (H₂). L'espace Θ est un ouvert non vide de \mathbb{R}^p sur lequel les fonctions $\theta \mapsto f_X(x; \theta)$ et $\theta \mapsto \ln f_X(x; \theta)$ sont de classe \mathcal{C}^2 uniformément en x .
- (H₃). Pour tout $\theta \in \Theta$, les fonctions $x \mapsto \nabla_\theta f_X(x; \theta)$ et $x \mapsto H_\theta f_X(x; \theta)$ sont intégrables sur E et, pour tout $\mathcal{X} \subseteq E$,
 - (H_{3,1}). $\int_{\mathcal{X}} \nabla_\theta f_X(s; \theta) d\nu(s) = \nabla_\theta \int_{\mathcal{X}} f_X(s; \theta) d\nu(s)$.
 - (H_{3,2}). $\int_{\mathcal{X}} H_\theta f_X(s; \theta) d\nu(s) = H_\theta \int_{\mathcal{X}} f_X(s; \theta) d\nu(s)$.
- (H₄). La matrice de taille $p \times p$ donnée par

$$\mathcal{I}_X(\theta) = -\mathbb{E}_\theta[H_\theta \ln f_X(X; \theta)]$$

est inversible pour tout $\theta \in \Theta$.

Information de Fisher apportée par X sur son paramètre θ

Ces hypothèses sont avant tout techniques : elles permettent d'assurer que $X(\Omega)$ ne dépend pas de θ , que l'on peut chercher l'EMV en utilisant les équations de vraisemblance, que l'on peut permuter l'intégration par rapport à x et la dérivation jusqu'à l'ordre 2 par rapport à θ lors des calculs faisant intervenir $f_X(x; \theta)$, ou encore que l'information de Fisher (que l'on définira par la suite) existe et s'inverse.

Remarque. La plupart des modèles usuels étudiés dans le cadre de ce cours sont réguliers ($\mathcal{B}(p)$, $\mathcal{N}(\mu, \sigma^2)$, $\mathcal{P}(\lambda)$, etc.) Le modèle $\mathcal{U}([0, \theta])$ n'est pas régulier car (H₁) n'est pas satisfaite. En fait on peut même montrer que sous des hypothèses à peine plus fortes, un modèle dont la densité appartient à la famille exponentielle est régulier.

2.3.1 Information de Fisher

Définition 2.11 Soit X une v.a.r. dont la densité est paramétrée par $\theta \in \Theta$ dans un modèle régulier. On définit le *score* par

$$S_\theta(X) = \nabla_\theta \ln f_X(X; \theta).$$

Proposition 2.7 Si le modèle est régulier, le score est centré.

$$\mathbb{E}[S_\theta(X)] = 0$$

Preuve :

$$\mathbb{E}[S_\theta(X)] = \mathbb{E}\left[\nabla_\theta \ln f_X(X; \theta)\right] = \int_{\mathcal{X}} \nabla_\theta \ln(f_X(x; \theta)) f_X(x; \theta) d\nu(x) = \begin{pmatrix} \int_{\mathcal{X}} \frac{\partial}{\partial \theta_1} \ln(f_X(x; \theta)) f_X(x; \theta) d\nu(x) \\ \vdots \\ \int_{\mathcal{X}} \frac{\partial}{\partial \theta_p} \ln(f_X(x; \theta)) f_X(x; \theta) d\nu(x) \end{pmatrix}$$

$$\forall i \in \llbracket 1; p \rrbracket : \int_{\mathcal{X}} \frac{\partial}{\partial \theta_i} \ln(f_X(x; \theta)) f_X(x; \theta) d\nu(x) = \int_{\mathcal{X}} \frac{\frac{\partial}{\partial \theta_i} (f_X(x; \theta))}{f_X(x; \theta)} f_X(x; \theta) d\nu(x)$$

$$\Rightarrow \mathbb{E}[S_\theta(X)] = \int_{\mathcal{X}} \nabla_\theta f_X(x; \theta) d\nu(x) \stackrel{\text{d'après 4.3.1}}{=} \nabla_\theta \underbrace{\int_{\mathcal{X}} f_X(x; \theta) d\nu(x)}_1 = 0.$$

Définition 2.12 Soit X une v.a.r. dont la densité est paramétrée par $\theta \in \Theta$ dans un modèle régulier. Alors, l'information de Fisher apportée par X sur θ est définie par

$$\mathcal{I}_X(\theta) = \mathbb{E}_\theta[S_\theta(X) S_\theta^T(X)] = \left(\mathbb{E}_\theta \left[\frac{\partial}{\partial \theta_i} \ln f_X(X; \theta) \frac{\partial}{\partial \theta_j} \ln f_X(X; \theta) \right] \right)_{1 \leq i, j \leq p}$$

où $p = \dim(\Theta)$.

On remarque que, en vertu de la Prop. 2.7, on a aussi

$$\mathcal{I}_X(\theta) = \mathbb{V}_\theta(S_\theta(X)).$$

Proposition 2.8 Si le modèle est régulier, on a

$$\mathcal{I}_X(\theta) = -\mathbb{E}_\theta[H_\theta \ln f_X(X; \theta)] = \left(-\mathbb{E}_\theta \left[\frac{\partial^2}{\partial \theta_i \partial \theta_j} \ln f_X(X; \theta) \right] \right)_{1 \leq i, j \leq p}$$

où $p = \dim(\Theta)$.

Preuve: On veut mq $\forall (i, j) \in \llbracket 1, p \rrbracket^2$: $\mathbb{E} \left[\frac{\partial}{\partial \theta_i} \ln f_X(x; \theta) \frac{\partial}{\partial \theta_j} \ln f_X(x; \theta) \right] = \mathbb{E} \left[\frac{\partial^2}{\partial \theta_i \partial \theta_j} \ln f_X(x; \theta) \right]$

En effet $\forall x \in \mathcal{X}$ et $\forall \theta \in \Theta$:

$$\frac{\partial^2}{\partial \theta_i \partial \theta_j} \ln f_X(x; \theta) = \frac{\partial}{\partial \theta_i} \left[\frac{\frac{\partial}{\partial \theta_j} f_X(x; \theta)}{f_X(x; \theta)} \right] = \frac{\frac{\partial^2}{\partial \theta_i \partial \theta_j} f_X(x; \theta)}{f_X(x; \theta)} - \frac{\frac{\partial}{\partial \theta_i} f_X(x; \theta) \frac{\partial}{\partial \theta_j} f_X(x; \theta)}{f_X(x; \theta)^2}$$

Donc $\mathbb{E} \left[\frac{\partial^2}{\partial \theta_i \partial \theta_j} \ln f_X(x; \theta) \right] = \underbrace{\int_{\mathcal{X}} \frac{\partial^2}{\partial \theta_i \partial \theta_j} f_X(x; \theta) d\nu(x)}_{=0} - \mathbb{E} \left[\frac{\partial}{\partial \theta_i} \ln f_X(x; \theta) \frac{\partial}{\partial \theta_j} \ln f_X(x; \theta) \right]$

car $\int_{\mathcal{X}} H_\theta f_X(x; \theta)_{i,j} d\nu(x) = \left(H_\theta \int_{\mathcal{X}} f_X(x; \theta) d\nu(x) \right)_{i,j}$
d'après H.3.2 \uparrow $\underbrace{\int_{\mathcal{X}} f_X(x; \theta) d\nu(x)}_{=1}$ $\underbrace{\quad}_{=0}$

Proposition 2.9 Dans un modèle régulier, l'information de Fisher est une matrice symétrique définie positive qui possède la propriété d'additivité pour les v.a.r. indépendantes.

Preuve:

- $\mathcal{I}_X(\theta)$ est symétrique

$$\mathcal{I}_X(\theta)_{i,j} = \mathbb{E} \left[\frac{\partial}{\partial \theta_i} \ln f_X(x; \theta) \frac{\partial}{\partial \theta_j} \ln f_X(x; \theta) \right] = \mathbb{E} \left[\frac{\partial}{\partial \theta_j} \ln f_X(x; \theta) \frac{\partial}{\partial \theta_i} \ln f_X(x; \theta) \right] = \mathcal{I}_X(\theta)_{j,i}$$

- $\mathcal{I}_X(\theta)$ est définie positive: $\forall u \in \mathbb{R}^p$, $u \mathcal{I}_X(\theta) u^T \geq 0$

En effet $u \mathcal{I}_X(\theta) u^T = u \mathbb{E} [S_\theta(X) S_\theta(X)^T] u^T = \mathbb{E} [u S_\theta(X) (u S_\theta(X))^T]$
 $= \text{Var}(u S_\theta(X)) \geq 0$

- $\mathcal{I}_X(\theta)$ additivité (X, Y) couple v.a.r. II: $\mathcal{I}_{X,Y}(\theta) = \mathcal{I}_X(\theta) + \mathcal{I}_Y(\theta)$

or $f_{X,Y}(x, y; \theta) = f_X(x; \theta) f_Y(y; \theta) \Rightarrow \ln(f_{X,Y}(x, y; \theta)) = \ln f_X(x; \theta) + \ln f_Y(y; \theta)$

et $\mathcal{I}_{X,Y}(\theta) = -\mathbb{E} \left[\frac{\partial^2}{\partial \theta_i \partial \theta_j} \ln f_{X,Y}(x, y; \theta) \right] = -\mathbb{E} [H_\theta \ln f_{X,Y}(x, y; \theta)]$
 $= -\mathbb{E} [H_\theta (\ln f_X(x; \theta) + \ln f_Y(y; \theta))] = \mathcal{I}_X(\theta) + \mathcal{I}_Y(\theta)$

Par corollaire, on a

$$\mathcal{I}_{(X_1, \dots, X_n)}(\theta) = \sum_{k=1}^n \mathcal{I}_{X_k}(\theta) = n \mathcal{I}_X(\theta)$$

pour un n -échantillon X_1, \dots, X_n de v.a.r. parente X . Pour alléger les notations, on écrira par la suite

$$\mathcal{I}_n(\theta) = n \mathcal{I}_X(\theta)$$

pour désigner l'information de Fisher apportée par un n -échantillon sur son paramètre.

Exemples. Calculer l'information de Fisher apportée par X dans quelques modèles usuels :

$$\begin{aligned} X &\sim \mathcal{N}(\mu, \sigma^2) \quad p=2 \quad f_X(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}; \quad x \in \mathbb{R} \\ \Rightarrow \ln f_X(x; \mu, \sigma^2) &= -\frac{1}{2} \ln(2\pi) - \frac{1}{2} \ln(\sigma^2) - \frac{(x-\mu)^2}{2\sigma^2} \\ \text{Donc } \frac{\partial}{\partial \mu} \ln f_X(x; \mu, \sigma^2) &= \frac{x-\mu}{\sigma^2}; \quad \frac{\partial}{\partial \sigma^2} \ln f_X(x; \mu, \sigma^2) = -\frac{1}{2\sigma^2} + \frac{(x-\mu)^2}{2\sigma^4} \\ \frac{\partial^2}{\partial \mu^2} \ln f_X(x; \mu, \sigma^2) &= -\frac{1}{\sigma^2}; \quad \frac{\partial^2}{(\partial \sigma^2)^2} \ln f_X(x; \mu, \sigma^2) = \frac{1}{2\sigma^4} - \frac{(x-\mu)^2}{\sigma^6} \\ \frac{\partial^2}{\partial \mu \partial \sigma^2} \ln f_X(x; \mu, \sigma^2) &= -\frac{x-\mu}{\sigma^4} \\ \Rightarrow \mathcal{I}_X(\mu, \sigma^2)_{1,1} &= -\mathbb{E}\left[-\frac{1}{\sigma^2}\right] = \frac{1}{\sigma^2} \\ \mathcal{I}_X(\mu, \sigma^2)_{2,2} &= -\mathbb{E}\left[\frac{1}{2\sigma^4} - \frac{(X-\mu)^2}{\sigma^6}\right] = -\frac{1}{2\sigma^4} + \frac{\text{Var}(X)}{\sigma^6} = \frac{-\sigma^2 + 2\sigma^2}{2\sigma^6} = \frac{1}{2\sigma^4} \\ \mathcal{I}_X(\mu, \sigma^2)_{1,2} &= \mathcal{I}_X(\mu, \sigma^2)_{2,1} = -\mathbb{E}\left[-\frac{(X-\mu)}{\sigma^4}\right] = 0 \quad \mathbb{E}[X] = \mu. \\ \text{Donc } \mathcal{I}_X(\mu, \sigma^2) &= \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix} \end{aligned}$$

Soit $X \sim \mathcal{B}(p)$, on peut définir une densité contre la mesure de comptage:

$$v(\{0\}) = v(\{1\}) = 1 \text{ ou } 0 \text{ sinon. Dans ce cas:}$$

$$f_X(x|p) = P_p(X=x) = \begin{cases} p & \text{si } x=1 \\ 1-p & \text{si } x=0 \end{cases}$$

Puisque le support de v est $\{0, 1\}$ on peut supposer $f_X(x|p) = p^x(1-p)^{1-x}$

Donc $\ln f_X(x|p) = x \ln p + (1-x) \ln(1-p)$

$$\Rightarrow \frac{\partial}{\partial p} \ln f_X(x|p) = \frac{x}{p} - \frac{(1-x)}{1-p} =: S_p(x) \quad \text{le score de } x \text{ (son espérance vaut bien 0).}$$

$$\Rightarrow \frac{\partial^2}{\partial p^2} \ln f_X(x|p) = -\frac{x}{p^2} - \frac{(1-x)}{(1-p)^2}$$

On sait que: $\mathcal{I}_X(p) = -\mathbb{E}_p\left[\frac{\partial^2}{\partial p^2} \ln f_X(X|p)\right] = -\mathbb{E}_p\left[-\frac{X}{p^2} - \frac{(1-X)}{(1-p)^2}\right]$

$$= \frac{1}{p} + \frac{1}{1-p} = \frac{1}{p(1-p)} = \frac{1}{\text{Var}(X)}$$

Par ailleurs, l'information de Fisher est aussi étroitement liée au maximum de vraisemblance *via* la propriété suivante.

Proposition 2.10 Sous les conditions de régularité, l'EMV $\hat{\theta}_n$ d'un paramètre θ satisfait

$$\sqrt{n} (\hat{\theta}_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \mathcal{I}_X^{-1}(\theta)).$$

Exemples. Ce résultat n'est rien d'autre que le TCL dans les modèles usuels dont l'EMV est $\hat{\theta}_n = \bar{X}_n$.

i.e. $\forall u \in \mathbb{R}^k$ si $\Theta \subset \mathbb{R}^d$ alors $u \text{Var}_\theta(\hat{\theta}_n) u^T \geq u \mathcal{I}_n(\theta)^{-1} u^T$

2.3.2 Borne de Cramér-Rao

On va maintenant montrer que l'information de Fisher permet de construire une borne inférieure (FDCR pour Fréchet-Darmois-Cramér-Rao) pour la variance des estimateurs non biaisés.

Proposition 2.11 (Borne FDCR) Soit $\hat{\theta}_n$ un estimateur sans biais de θ dans un modèle régulier. Alors,

$$\forall n, \forall \theta \in \Theta, \quad \text{Var}(\hat{\theta}_n) \geq \mathcal{I}_n^{-1}(\theta)$$

où l'inégalité est à interpréter au sens des matrices semi-définies positives.

Preuve pour $p=1$:

Soit $\hat{\theta}_n$ un estimateur, par définition, $\hat{\theta}_n$ est une fonction de l'échantillon $\underline{X} = (X_1, \dots, X_n)$ avec $\hat{\theta}_n = h(\underline{X})$ où $h: \mathbb{R}^n \rightarrow \mathbb{R}$.
Puisque $\hat{\theta}_n$ est non biaisé, on a:

$$\mathbb{E}_\theta[\hat{\theta}_n] = \mathbb{E}[h(\underline{X})] = \theta \quad \text{et clairement} \quad \frac{\partial}{\partial \theta} \mathbb{E}[h(\underline{X})] = 1, \text{ et:}$$

$$\frac{\partial}{\partial \theta} \mathbb{E}_\theta[h(\underline{X})] = \int_{\mathbb{R}^n} h(\underline{x}) \frac{\partial}{\partial \theta} \underbrace{\prod_{i=1}^n f_x(x_i; \theta)}_{\substack{\text{par hypothèse de densité de } \underline{X}=(x_1, \dots, x_n) \\ \text{du modèle régulier}}} d\underline{x}$$

On note $g(\underline{x}; \theta) = \prod_{i=1}^n f_x(x_i; \theta)$, on a:

$$\begin{aligned} \frac{\partial}{\partial \theta} \mathbb{E}_\theta[h(\underline{X})] &= \int_{\mathbb{R}^n} h(\underline{x}) \frac{\partial}{\partial \theta} g(\underline{x}; \theta) d\underline{x} = \int_{\mathbb{R}^n} h(\underline{x}) \frac{\partial}{\partial \theta} \ln g(\underline{x}; \theta) \cdot g(\underline{x}; \theta) d\underline{x} \\ &= \int_{\mathbb{R}^n} h(\underline{x}) \frac{\partial}{\partial \theta} \left(\sum_{i=1}^n \ln f_x(x_i; \theta) \right) g(\underline{x}; \theta) d\underline{x} = \mathbb{E}_\theta \left[h(\underline{X}) \sum_{i=1}^n \left(\frac{\partial}{\partial \theta} \ln f_x(x_i; \theta) \right) \right] \end{aligned}$$

$$= \mathbb{E}_\theta \left[(h(\underline{X}) - \theta) \left(\sum_{i=1}^n \frac{\partial}{\partial \theta} \ln f_x(x_i; \theta) \right) \right] = \mathbb{E} \left[\underbrace{(h(\underline{X}) - \theta)}_{\text{sans biais}} \sum_{i=1}^n S_\theta(X_i) \right] \quad (1)$$

$S_\theta(X_i) \rightarrow$ le score va être égale à zéro donc ça ne change rien d'enlever θ .

$$\stackrel{\text{Cauchy-Schwarz}}{\leq} \frac{\sqrt{\mathbb{E}_\theta[(h(\underline{X}) - \theta)^2]}}{\text{Var}(h(\underline{X}))} \times \frac{\sqrt{\mathbb{E}_\theta \left[\left(\sum_{i=1}^n S_\theta(X_i) \right)^2 \right]}}{\text{Var}(S_\theta(X_i)) = n \mathcal{I}_x(\theta) = \mathcal{I}_n(\theta)}$$

Pour conclure $1 \leq \text{Var}(\hat{\theta}_n) \mathcal{I}_n(\theta)$

Remarque: si $\mathcal{I}_x(\theta) > 0$, alors: $\text{Var}(\hat{\theta}_n) \geq \frac{1}{n \mathcal{I}_x(\theta)} \xrightarrow{n \rightarrow \infty} 0$

2.3.3 Efficacité

On cherche à obtenir l'estimateur sans biais de variance minimale (ESBVM) dans la classe des estimateurs de θ : on parle d'estimateur efficace dès qu'un estimateur sans biais atteint la borne FDCR.

Définition 2.13 Un estimateur $\hat{\theta}_n$ de θ est efficace s'il est sans biais et si

$$\forall n, \forall \theta \in \Theta, \quad \mathbb{V}_{\theta}(\hat{\theta}_n) = \mathcal{I}_n^{-1}(\theta).$$

Il est asymptotiquement efficace s'il est (asymptotiquement) sans biais, et si

$$\forall \theta \in \Theta, \quad \lim_{n \rightarrow +\infty} \mathbb{V}_{\theta}(\hat{\theta}_n) \mathcal{I}_n(\theta) = I_p$$

où I_p est la matrice identité d'ordre $p = \dim(\Theta)$.

Exemples. Étudier l'efficacité de l'EMV dans les modèles pour lesquels l'information de Fisher a été calculée dans les sections précédentes.

* $X \sim \mathcal{B}(\theta)$ $\theta \in \Theta = [0, 1]$

On a $f_{(x_1, \dots, x_n)}(\underline{x}; \theta) = \theta^{\sum_{i=1}^n x_i} (1-\theta)^{n - \sum_{i=1}^n x_i}$ où $x_i \in \{0, 1\}^n$

$\Rightarrow \ln f_{(x_1, \dots, x_n)}(\underline{x}; \theta) = \left(\sum_{i=1}^n x_i \right) \ln \theta + \left(n - \sum_{i=1}^n x_i \right) \ln(1-\theta).$

$\Rightarrow \frac{\partial}{\partial \theta} \ln f_{(x_1, \dots, x_n)}(\underline{x}; \theta) = \frac{1}{\theta} \sum_{i=1}^n x_i - \frac{1}{1-\theta} \left(n - \sum_{i=1}^n x_i \right) = 0 \Leftrightarrow \hat{\theta}_n = \frac{1}{n} \sum_{i=1}^n x_i$

Ici l'EMV coïncide avec la moyenne empirique

Pour étudier l'efficacité ; ici on a bien que le modèle est régulier.

$\mathbb{E}[\hat{\theta}_n] = \theta$ et $\text{Var}(\hat{\theta}_n) = \frac{1}{n^2} \text{Var}\left(\sum_{i=1}^n X_i\right) = \frac{n}{n^2} \text{Var}(X_i) = \frac{\theta(1-\theta)}{n}$

On a montré que $\mathcal{I}_x(\theta) = \frac{1}{\theta(1-\theta)} \Rightarrow \mathcal{I}_n(\theta) = \frac{n}{\theta(1-\theta)} = \frac{1}{\text{Var}(\hat{\theta}_n)}$

* $X \sim \mathcal{N}(\mu, \sigma^2)$ $\mathcal{I}_n(\mu, \sigma^2) = \begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix} = n \mathcal{I}_x(\theta)$ voir p. 28

$\hat{\theta}_n = (\bar{X}_n, S_n^2)$ EMV

$\tilde{\theta}_n = (\bar{X}_n, S_n^{*2})$ sans biais.

voir chapitre 3
si $X \sim \mathcal{N}^p$, $\bar{X}_n \perp S_n^2$

$\text{Var}(\tilde{\theta}_n) = \begin{pmatrix} \text{Var}(\bar{X}_n) & \text{cov}(\bar{X}_n, S_n^{*2}) \\ \text{cov}(\bar{X}_n, S_n^{*2}) & \text{Var}(S_n^{*2}) \end{pmatrix} = \begin{pmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{2\sigma^4}{n-1} \end{pmatrix} \neq \mathcal{I}_n^{-1}(\mu, \sigma^2) = \begin{pmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{2\sigma^4}{n-1} \end{pmatrix}$

Donc $\tilde{\theta}_n$ n'est pas efficace.

Mais $\text{Var}(\tilde{\theta}_n) \mathcal{I}_n(\mu, \sigma^2) = \begin{pmatrix} 1 & 0 \\ 0 & \frac{n}{n-1} \end{pmatrix} \xrightarrow{n \rightarrow +\infty} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$

Donc $\tilde{\theta}_n$ est asymptotiquement efficace

2.3.4 Cas des estimateurs biaisés

Il est possible de généraliser cette borne inférieure de la variance à certains estimateurs biaisés. On en donne simplement un aperçu dans le cas unidimensionnel.

Proposition 2.12 (Borne FDCR généralisée) Soit $\hat{\theta}_n$ un estimateur de θ tel que $\mathbb{E}_\theta[\hat{\theta}_n] = \psi(\theta)$ dans un modèle régulier, pour $p = 1$. Alors, si ψ est dérivable sur Θ ,

$$\forall n, \forall \theta \in \Theta, \quad \mathbb{V}_\theta(\hat{\theta}_n) \geq \frac{(\psi'(\theta))^2}{\mathcal{I}_n(\theta)}.$$

Preuve. Identique à la preuve de la Prop. 2.11 pour laquelle on avait $\psi(\theta) = \theta$.

Et on remarque qu'on obtient ① $\psi(\hat{\theta}_n) - \psi(\theta) = (\hat{\theta}_n - \theta) \psi'(\theta) (1 + o_p(1))$

Notons qu'un estimateur biaisé peut atteindre une variance inférieure à la borne FDCR : on a vu dans l'exemple $X \sim \mathcal{N}(0, \sigma^2)$ que

$$\mathbb{V}(S_n^2) = \frac{2(n-1)\sigma^4}{n^2} < \frac{2\sigma^4}{n} = \frac{1}{\mathcal{I}_n(\sigma^2)} \quad \text{avec ici} \quad \psi(\sigma^2) = \frac{n-1}{n} \sigma^2.$$

La borne généralisée s'écrit alors

$$\frac{(\psi'(\sigma^2))^2}{\mathcal{I}_n(\sigma^2)} = \frac{2(n-1)^2 \sigma^4}{n^3} = \mathbb{V}(S_n^2) \frac{n-1}{n} < \mathbb{V}(S_n^2).$$

Le cas vectoriel ($p \geq 1$) se traite de manière similaire, avec intervention de la matrice jacobienne $J_\theta \psi(\theta)$ en lieu et place de $\psi'(\theta)$.

2.4 Présentation de l'inférence bayésienne

On va faire un détour rapide par le paradigme bayésien, sans pour autant le détailler (car cela demanderait un module complet...). Oublions quelques instants l'approche que l'on vient de développer, qualifiée de *fréquentiste* (on cherche la situation la plus vraisemblable en fonction des observations), pour discuter de l'approche *bayésienne* dans laquelle le paramètre θ n'est plus une quantité fixe à approximer, mais une variable aléatoire et Θ devient l'univers d'un espace probabilisé. Cette approche est adaptée lorsque le statisticien dispose d'une information sur le paramètre et qu'il veut l'exploiter. Par exemple, une étude préalable ou un expert ont déjà donné des indications spécifiques.

Définition 2.14 La distribution marginale de θ est sa loi *a priori*. Elle caractérise toute l'information disponible sur le paramètre non issue des observations. La distribution de θ conditionnellement aux observations est sa loi *a posteriori*.

La stratégie bayésienne est la suivante : poser un *a priori* sur le paramètre *avant* les observations, en déduire un *a posteriori* sur le paramètre *après* les observations.



Notons $\pi(\cdot)$ la densité *a priori* du paramètre et $\ell_X(\cdot | \theta)$ la densité des observations $\underline{X} = (X_1, \dots, X_n)$ lorsque le paramètre est connu (donc, la vraisemblance). La densité *a posteriori* du paramètre s'exprime, selon la formule de Bayes, par

$$\forall \theta \in \Theta, \quad \pi(\theta | \underline{X}) = \frac{\ell_X(\underline{X} | \theta) \pi(\theta)}{\int_{\Theta} \ell_X(\underline{X} | \theta) \pi(\theta) d\theta} = \frac{h(\theta, \underline{X})}{m(\underline{X})}.$$

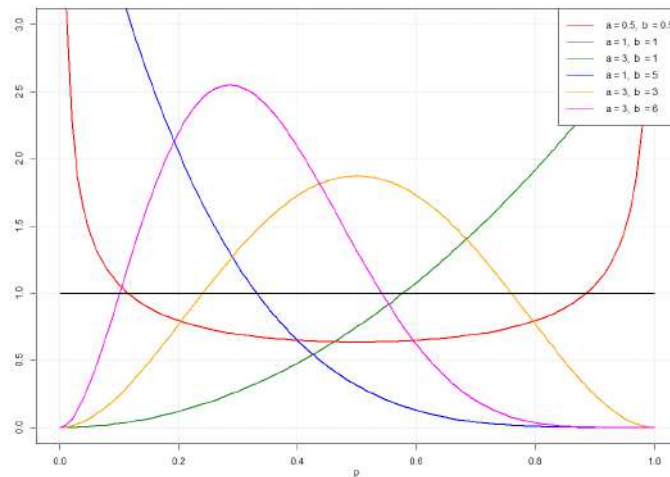
Fréquentiste	Bayésienne
θ paramètre fixe	θ aléatoire
EMV $\Rightarrow \hat{\theta}_n = \arg \max_{\theta \in \Theta} f_{\underline{X}}(\underline{x} \theta)$	EMVB $\Rightarrow \hat{\theta}_n = \arg \max_{\theta \in \Theta} \pi(\theta \underline{x})$

Au numérateur, $h(\cdot, \cdot)$ représente la densité jointe du couple (θ, \underline{X}) . Au dénominateur, $m(\cdot)$ caractérise la densité marginale des observations (on suppose pour simplifier que l'*a priori* est une densité continue). En fait, le raisonnement est grandement simplifié par le fait que le dénominateur ne nous intéresse pas car il ne dépend pas de θ : il s'agit d'une constante de normalisation qui garantit que $\int_{\Theta} \pi(\theta | \underline{X}) d\theta = 1$. Ainsi,

$$\pi(\theta | \underline{X}) \propto \ell_X(\underline{X} | \theta) \pi(\theta)$$

et cette **relation de proportionnalité** est fondamentale car elle montre que la loi *a posteriori* se déduit immédiatement de la loi *a priori* à l'aide d'une simple multiplication par la vraisemblance.

Remarque. Dans les modèles où le paramètre joue le rôle d'une probabilité (Bernoulli, géométrique, ...), il est fréquent qu'il soit muni d'un *a priori* bêta dans l'approche bayésienne. En effet, la distribution bêta est portée par $[0, 1]$ et ses deux paramètres lui offrent une grande flexibilité.



Dans ces exemples, choisir $\beta(1, 1)$ revient à mettre un *a priori non informatif* sur le paramètre : on ne présuppose rien, le paramètre vit uniformément dans $[0, 1]$. Au contraire, choisir $\beta(1, 5)$ revient à dire qu'on surcharge les petites valeurs : on apporte comme information initiale le fait que le paramètre est probablement proche de 0. Pour $\beta(3, 3)$, on a de bonnes raisons de penser que le paramètre n'est pas extrême (éloigné de 0 comme de 1), etc.

Exemple. Comme exemple typique, déterminer la loi *a posteriori* de p dans le modèle $X | p \sim \mathcal{B}(p)$ et $p \sim \beta(a, b)$, c'est-à-dire un modèle de Bernoulli dans lequel le paramètre est muni d'un *a priori* bêta.

$$f_p(x; a, b) = \mathbb{1}_{\{x \in [0, 1]\}} \frac{p^{a-1} (1-p)^{b-1}}{B(a, b)} =: \pi_{a,b}(p).$$

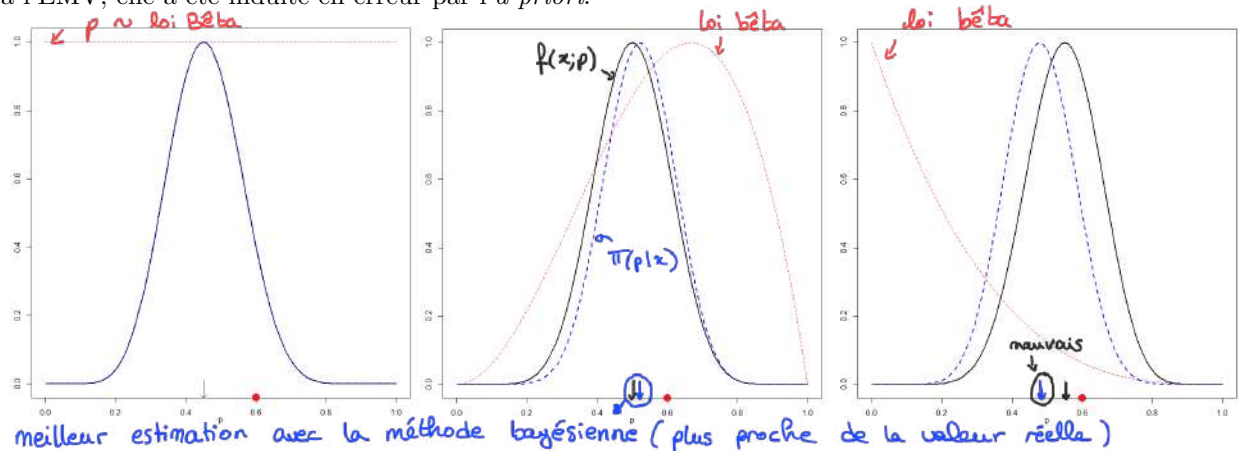
Escobar - Bach

Soi $p | X$:

$$\begin{aligned} \pi(p | x) &= \frac{f_X(x; p) \pi_{a,b}(p)}{\int_0^1 f_X(x; p) \pi_{a,b}(p) dp} = \frac{p^x (1-p)^{1-x} p^{a-1} (1-p)^{b-1}}{B(a, b) \int_0^1 p^x (1-p)^{1-x} p^{a-1} (1-p)^{b-1} dp} \\ &= \frac{p^{x+a-1} (1-p)^{b-x-1}}{\int_0^1 p^{x+a-1} (1-p)^{b-x-1} dp} = \frac{p^{a+x-1} (1-p)^{b+1-x-1}}{B(a+x, b+1-x)} \\ &\Leftrightarrow p | X \sim \mathcal{B}(a+x, b+1-x) \end{aligned}$$

$$\begin{aligned}
 p &\sim \beta(a, b) & \pi(p) &\propto p^{a-1} (1-p)^{b-1} & \text{a priori} & \text{Proia.} \\
 X|p &\sim \mathcal{B}(p) & \ell_n(\underline{X}|p) &= p^{n\bar{X}_n} (1-p)^{n(1-\bar{X}_n)} \\
 \Rightarrow p|\underline{X} &\propto p^{n\bar{X}_n+a-1} (1-p)^{n(1-\bar{X}_n)+b-1} \\
 &\sim \beta(a+n\bar{X}_n, b+n-n\bar{X}_n) & \text{a posteriori}
 \end{aligned}$$

On simule quelques occurrences de l'exemple précédent avec $n = 20$ et $p = 0.6$. La courbe noire représente la fonction de vraisemblance, la courbe rouge l'*a priori* et la courbe bleue l'*a posteriori* (elles sont renormalisées pour des raisons d'échelle). Dans le premier cas (à gauche), l'*a priori* est non informatif ($a = b = 1$) et, sans surprise, l'EMV est confondu avec le maximum *a posteriori*. Dans le second cas (au centre), on place un *a priori* informatif et correct ($a = 3, b = 2$) qui cible la bonne valeur, et l'estimation qui en découle est meilleure que l'EMV. Dans le dernier cas (à droite), on place un *a priori* informatif mais incorrect ($a = 1, b = 4$) qui cible des valeurs éloignées de la vraie valeur, l'estimation qui en découle est mauvaise comparée à l'EMV, elle a été induite en erreur par l'*a priori*.



La question a été abordée dans l'exemple ci-dessus, mais il reste à voir comment déduire un *estimateur bayésien* à partir de la loi *a posteriori*. On rencontre traditionnellement deux possibilités :

→ Le **maximum a posteriori** : c'est le choix retenu dans l'exemple précédent, il s'agit de l'abscisse (ou des abscisses) qui maximise(nt) la densité *a posteriori*,

$$\hat{\theta}_n \in \arg \max_{\theta \in \Theta} \pi(\theta | \underline{X}).$$

→ La **moyenne a posteriori** : il s'agit de l'espérance de la loi *a posteriori*, donnée par

$$\hat{\theta}_n = \mathbb{E}[\theta | \underline{X}] = \int_{\Theta} \theta \pi(\theta | \underline{X}) d\theta.$$

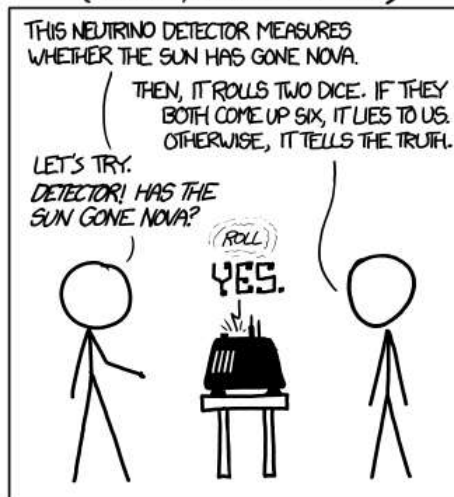
On pourrait aussi penser à la médiane *a posteriori* qui minimise le coût L^1 (l'espérance minimisant quant à elle le coût L^2).

Exemple. Que valent ces estimateurs bayésiens dans l'exemple précédent? À comparer avec l'approche fréquentiste...

$$\begin{aligned}
 X|p &\sim B(p) & p &\sim \beta(a,b) & p|\underline{X} &\sim \beta(a+n\bar{X}_n, b+n-n\bar{X}_n) \\
 \text{Le maximum a posteriori : } \ln(\pi(p|\underline{X})) &\propto \ln p^{n\bar{X}_n+a-1} (1-p)^{n(1-\bar{X}_n)+b-1} \\
 &\propto (n\bar{X}_n+a-1) \ln p + (n-n\bar{X}_n+b-1) \ln(1-p) \\
 \frac{\partial}{\partial p} \ln(\pi(p|\underline{X})) &\propto \frac{n\bar{X}_n+a-1}{p} - \frac{n-n\bar{X}_n+b-1}{1-p} = 0 \Leftrightarrow (1-p)(n\bar{X}_n+a-1) = p(n-n\bar{X}_n+b-1) \\
 \Leftrightarrow p(n-n\bar{X}_n+b-1) + p(n\bar{X}_n+a-1) &= n\bar{X}_n+a-1 \\
 \Leftrightarrow p(n+b+a-2) &= n\bar{X}_n+a-1 \Leftrightarrow \hat{p}_n = \frac{n\bar{X}_n+a-1}{n+b+a-2}
 \end{aligned}$$

On verra d'autres exemples en TD. Alors que préférer, l'approche bayésienne ou l'approche fréquentiste? En fait il n'existe pas de réponse universelle, tout dépend de la situation. Si l'information *a priori* est jugée fiable, on aurait tort de s'en priver.

DID THE SUN JUST EXPLODE?
(IT'S NIGHT, SO WE'RE NOT SURE.)



(xkcd)

la moyenne a posteriori :
Espérance d'une loi bēba $\frac{a}{a+b}$

$$\text{D'où } \hat{p}_n = \frac{a+n\bar{X}_n}{a+n\bar{X}_n+b+n-n\bar{X}_n}$$

$$\hat{p}_n = \frac{a+n\bar{X}_n}{a+b+n}$$

L'approche fréquentiste:

$$\hat{p}_n = \bar{X}_n$$

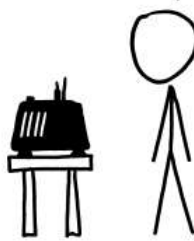
FREQUENTIST STATISTICIAN:

THE PROBABILITY OF THIS RESULT HAPPENING BY CHANCE IS $\frac{1}{36} = 0.027$. SINCE $p < 0.05$, I CONCLUDE THAT THE SUN HAS EXPLODED.



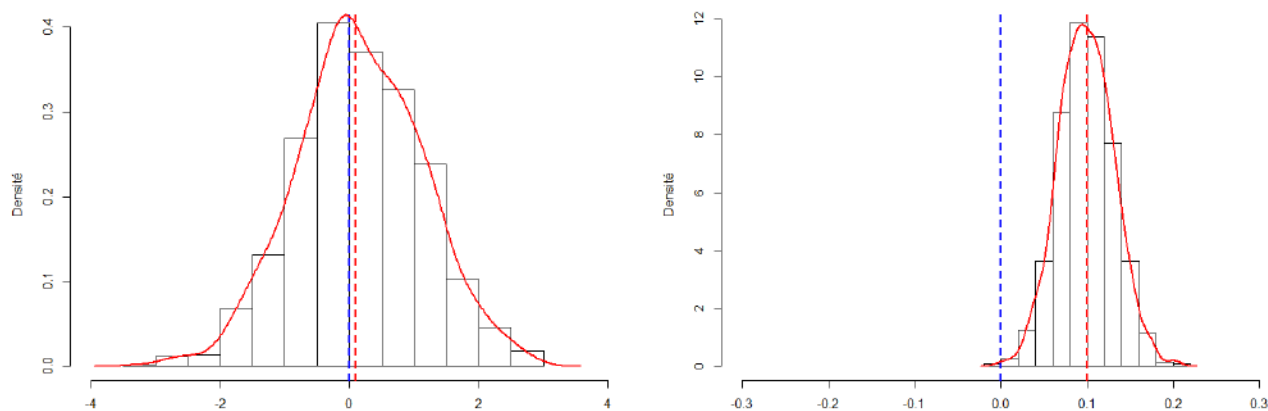
BAYESIAN STATISTICIAN:

BET YOU \$50 IT HASN'T.



3 Tests d'hypothèses

Avant toute chose, un exemple introductif est nécessaire pour bien appréhender la problématique des tests... On récolte des observations dont la moyenne empirique vaut 0.1. Peut-on les modéliser par des variables centrées? Certes 0.1 ce n'est pas 0, mais c'est quand même proche de 0. Comment déterminer la frontière à partir de laquelle la moyenne est trop éloignée de 0 pour pouvoir considérer que les variables sont centrées? Les graphes ci-dessous illustrent le fait qu'il n'y a pas de réponse objective mais que des décisions subjectives, qui dépendent du jeu de données. Avec une variance de 1 (à gauche), l'écart observé paraît insignifiant alors qu'avec une variance de 10^{-3} (à droite), ce même écart conduit sans ambiguïté à rejeter l'hypothèse de centrage. Le test d'hypothèses consiste à poser un cadre théorique à des règles de décision comme celle-ci.



Définition 3.1 Un test d'hypothèses est une procédure statistique ayant pour but de fournir une règle de décision permettant, sur la base des observations, de discriminer deux hypothèses statistiques.

L'hypothèse de départ se note généralement \mathcal{H}_0 , elle est qualifiée d'*hypothèse nulle*. Elle est mise en concurrence avec une *hypothèse alternative* \mathcal{H}_1 . Lorsque le test repose sur la valeur d'un paramètre $\theta \in \Theta$ (donc qu'il est paramétrique), on considère

$$\mathcal{H}_0 : " \theta \in \Theta_0 " \quad \text{contre} \quad \mathcal{H}_1 : " \theta \in \Theta_1 "$$

où Θ_0 et Θ_1 sont des sous-espaces paramétriques non vides de Θ , vérifiant $\Theta_0 \cap \Theta_1 = \emptyset$ (mais pas nécessairement $\Theta_0 \cup \Theta_1 = \Theta$). Si l'alternative \mathcal{H}_1 est orientée par rapport à \mathcal{H}_0 (par exemple pour les tests de la forme égal/égal, égal/supérieur ou inférieur/supérieur), le test est *unilatéral*. Il est *bilatéral* si \mathcal{H}_1 n'est pas orientée par rapport à \mathcal{H}_0 (par exemple de la forme égal/différent).

Exemples. Quelques exemples :

$\theta = \mathbb{E}[X] : \mathcal{H}_0 : " \theta = 0 " \text{ vs } \mathcal{H}_1 : " \theta \neq 0 "$ → le test de centrage donné en introduction	donc $\Theta_0 = \{0\}$ $\Theta_1 = \mathbb{R}^*$
$\theta = \text{Var}[X] : \mathcal{H}_0 : " \theta = 1 " \text{ vs } \mathcal{H}_1 : " \theta \neq 1 "$	donc $\Theta_0 = \{1\}$ $\Theta_1 = \mathbb{R}^+ \setminus \{1\}$
$X \sim \mathcal{B}(p) : \mathcal{H}_0 : " p = \frac{1}{2} " \text{ vs } \mathcal{H}_1 : " p \neq \frac{1}{2} "$ → test d'équilibre	donc $\Theta_0 = \{\frac{1}{2}\}$ $\Theta_1 =]0, 1[\setminus \{\frac{1}{2}\}$
C'est des test bilatéraux.	
$X \sim \mathcal{B}(p) : \mathcal{H}_0 : " p \geq 0,9 " \text{ vs } \mathcal{H}_1 : " p < 0,9 "$	donc $\Theta_0 =]0, 0,9]$ $\Theta_1 =]0,9, 1[$
C'est un test unilatéral.	

Définition 3.2 Un test de \mathcal{H}_0 contre \mathcal{H}_1 est une statistique $T = T(\underline{X})$ sur un échantillon $\underline{X} = (X_1, \dots, X_n)$, à valeurs dans $\{0\} \cup]0, 1[\cup \{1\}$. L'ensemble $T^{-1}(\{0\})$ conduisant à \mathcal{H}_0 est la zone d'acceptation du test, et l'ensemble $T^{-1}(\{1\})$ conduisant à \mathcal{H}_1 est la zone de rejet du test. L'ensemble $T^{-1}(]0, 1[)$ est la zone d'hésitation ou d'incertitude du test sur laquelle il est randomisé selon la loi $\mathcal{B}(T)$.

En pratique, la prise de décision sur la base des observations \underline{x} se fait selon les règles suivantes :

→ Si $T(\underline{x}) = 1$, on rejette \mathcal{H}_0 et l'on décide \mathcal{H}_1 .

→ Si $T(\underline{x}) = 0$, on ne rejette pas \mathcal{H}_0 au bénéfice du doute.

(→ Si $T(\underline{x}) \in]0, 1[$, les observations ne permettent pas de conclure. On décide \mathcal{H}_1 avec probabilité $T(\underline{x})$ et \mathcal{H}_0 avec probabilité $1 - T(\underline{x})$. On dit que le test est *randomisé*.)

Nous n'étudierons que des tests non randomisés dans le cadre de ce cours, donc des tests pour lesquels $T(\Omega) = \{0, 1\}$. Par la suite, on appellera $\mathcal{R} = \{T = 1\}$ l'évènement de rejet de \mathcal{H}_0 et l'on écrira \mathbb{P}_0 et \mathbb{E}_0 (resp. \mathbb{P}_1 et \mathbb{E}_1) lorsque les calculs seront menés sous \mathcal{H}_0 (resp. \mathcal{H}_1).

3.1 Tests construits sur des intervalles de confiance

L'intervalle de confiance donne plus de souplesse à l'estimation : plutôt que de produire une valeur approximée du paramètre, on propose un intervalle censé avoir une grande probabilité de le contenir. Cela permet de tenir compte des fluctuations, on fait une *estimation par intervalle* (par opposition à l'*estimation ponctuelle* du chapitre précédent). Le théorème ci-dessous sera particulièrement important dans la suite du cours.

Théorème 3.1 (Théorème de Cochran) Soient $\underline{Z} = (Z_1, \dots, Z_n)$ un vecteur de n v.a.r. indépendantes de même loi $\mathcal{N}(0, 1)$, F un s.e.v. de \mathbb{R}^n de dimension d et F^\perp son orthogonal dans \mathbb{R}^n de dimension $n - d$. On note P_F et P_{F^\perp} les matrices de projection orthogonale sur ces espaces. Alors :

→ $P_F \underline{Z}$ et $P_{F^\perp} \underline{Z}$ sont des vecteurs gaussiens indépendants de lois respectives $\mathcal{N}(0, P_F)$ et $\mathcal{N}(0, P_{F^\perp})$.

→ $\|P_F \underline{Z}\|^2$ et $\|P_{F^\perp} \underline{Z}\|^2$ sont des v.a.r. indépendantes et de lois respectives $\chi^2(d)$ et $\chi^2(n - d)$.

3.1.1 Intervalles de confiance exacts

Définition 3.3 Pour un risque $0 < \alpha < 1$, un intervalle de confiance pour le paramètre θ est un intervalle aléatoire $[B_{\inf}(\underline{X}), B_{\sup}(\underline{X})] = [b_1, b_2]$ construit sur un échantillon $\underline{X} = (X_1, \dots, X_n)$ de telle sorte que

$$\forall n, \forall \theta \in \Theta, \quad \mathbb{P}_\theta(b_1 \leq \theta \leq b_2) \geq 1 - \alpha.$$

les valeurs classiques :
 $\alpha = 5\%$ $\alpha = 0,1\%$
 $\alpha = 1\%$ $\alpha = 0,01\%$

On notera $IC_{1-\alpha}(\theta) = [b_1, b_2]$ cet intervalle pour la sécurité $1 - \alpha$.

Les intervalles exacts (valables pour tout n) sont en général inaccessibles, la principale exception concerne les échantillons gaussiens.

Proposition 3.1 (Échantillons gaussiens) Soit X_1, \dots, X_n un n -échantillon de loi parente $X \sim \mathcal{N}(\mu, \sigma^2)$. Alors, les moyenne et variance empiriques sont indépendantes et elles satisfont

voir efficacité
 (\bar{X}_n, S_n^{*2}) chap. 2

$$\sqrt{n} \frac{\bar{X}_n - \mu}{\sqrt{S_n^{*2}}} \sim t(n-1) \quad \text{et} \quad \frac{(n-1) S_n^{*2}}{\sigma^2} \sim \chi^2(n-1).$$

loi de Student à (n-1)
 degré de liberté

Preuve. C'est une application du théorème de Cochran et de la définition de la loi de Student. → voir exo 17 du TD.

On en déduit les intervalles de confiance

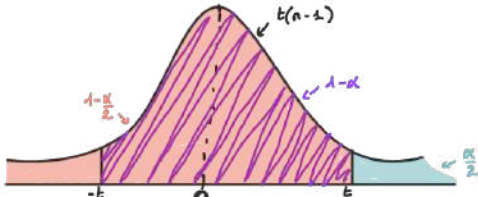
$$IC_{1-\alpha}(\mu) = \left[\bar{X}_n \pm \frac{\sqrt{S_n^{*2}}}{\sqrt{n}} t_{1-\frac{\alpha}{2}} \right] \quad \text{et} \quad IC_{1-\alpha}(\sigma^2) = \left[\frac{(n-1) S_n^{*2}}{z_{1-\frac{\alpha}{2}}}, \frac{(n-1) S_n^{*2}}{z_{\frac{\alpha}{2}}} \right]$$

estimation ponctuelle

fluctuations

Soit t le quantile de la loi $t(n-1)$ d'ordre $1-\frac{\alpha}{2}$
 On a $\forall n, \forall \mu, \quad \mathbb{P}\left(-t \leq \sqrt{n} \frac{\bar{X}_n - \mu}{\sqrt{S_n^{*2}}} \leq t\right) = 1 - \alpha$
 $\Leftrightarrow 1 - \alpha = \mathbb{P}\left(\bar{X}_n - \frac{\sqrt{S_n^{*2}}}{\sqrt{n}} t \leq \mu \leq \bar{X}_n + \frac{\sqrt{S_n^{*2}}}{\sqrt{n}} t\right)$

Soit z_β le quantile de la loi $\chi^2(n-1)$ d'ordre β
 On a $\forall n, \forall \sigma^2, \quad \mathbb{P}\left(z_{\frac{\alpha}{2}} \leq \frac{(n-1) S_n^{*2}}{\sigma^2} \leq z_{1-\frac{\alpha}{2}}\right) = 1 - \alpha$
 $\Leftrightarrow 1 - \alpha = \mathbb{P}\left(\frac{(n-1) S_n^{*2}}{z_{1-\frac{\alpha}{2}}} \leq \sigma^2 \leq \frac{(n-1) S_n^{*2}}{z_{\frac{\alpha}{2}}}\right)$



où ci-dessus, $t_{1-\frac{\alpha}{2}}$ désigne le quantile correspondant de la loi $t(n-1)$ alors que $z_{\frac{\alpha}{2}}$ et $z_{1-\frac{\alpha}{2}}$ désignent ceux de la loi $\chi^2(n-1)$.

Exemple. Un exemple en direct avec R...

<pre> n = 1000 X = rnorm(n, -2, sqrt(2)) hist(X) → l'histogramme est Mu = mean(X) est S2 = var(X) → il calcule la a = 0,5 variance corrigée directement </pre>	<pre> b1Mu = est Mu - sqrt(est S2) / sqrt(n) * qt(1 - alpha/2, n - 1) b2Mu = _____ + _____ </pre> <p> $qchisq(1 - \frac{\alpha}{2}) \rightarrow 3.1 - \frac{\alpha}{2}$ </p>
---	---

Pour tester $\mathcal{H}_0 : \mu = \mu_0$ contre $\mathcal{H}_1 : \mu \neq \mu_0$, l'idée est de construire un intervalle de confiance associé à μ et de vérifier si μ_0 tombe à l'intérieur. On peut donc poser

$$T = \mathbb{1}_{\{\mu_0 \notin \text{IC}_{1-\alpha}(\mu)\}} = \begin{cases} 1 & \text{on rejette } \mathcal{H}_0, \text{ on décide } \mathcal{H}_1 \\ 0 & \text{on ne rejette pas } \mathcal{H}_0 \end{cases}$$

ou comparer la *statistique de test*

$$Z = \sqrt{n} \frac{\bar{X}_n - \mu_0}{\sqrt{S_n^{*2}}}$$

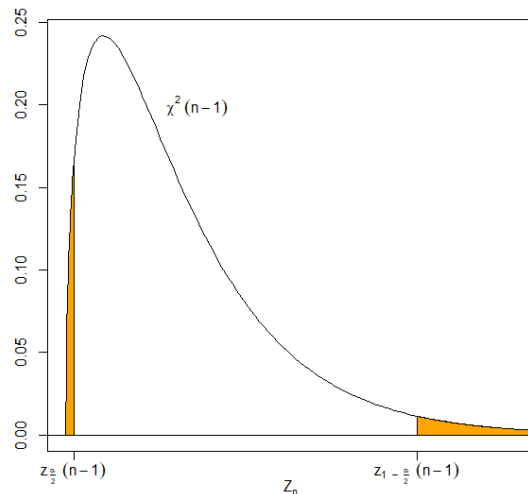
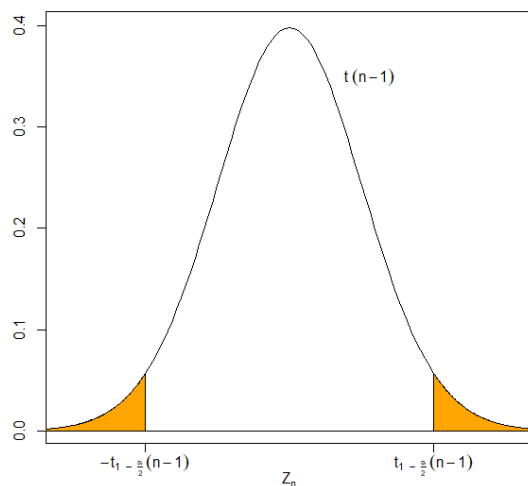
avec les quantiles de la loi $t(n-1)$.

Remarque. On peut montrer assez facilement que $t(n) \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1)$. Donc lorsque n est grand, on peut raisonnablement considérer que $t_\alpha(n-1) \approx u_\alpha$ et remplacer les quantiles de Student par les quantiles gaussiens.

Par analogie, le test de $\mathcal{H}_0 : \sigma^2 = \sigma_0^2$ contre $\mathcal{H}_1 : \sigma^2 \neq \sigma_0^2$ peut être conduit avec

$$T = \mathbb{1}_{\{\sigma_0^2 \notin \text{IC}_{1-\alpha}(\sigma^2)\}} \quad \text{ou} \quad Z = \frac{(n-1)S_n^{*2}}{\sigma_0^2}$$

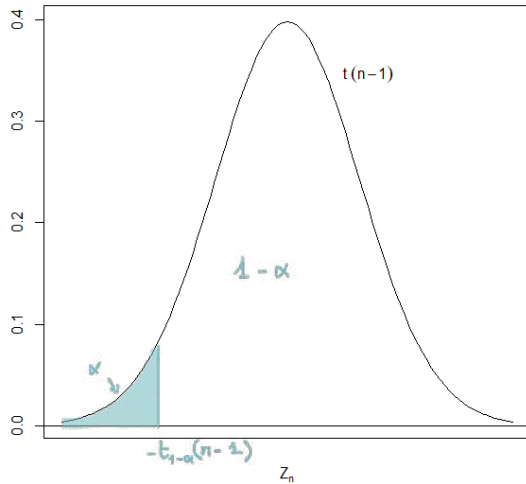
que l'on compare avec les quantiles de la loi $\chi^2(n-1)$. Les graphes ci-dessous représentent les zones de rejet associées à ces deux tests *bilatéraux*.



Exemple. Représenter les zones de rejet associées aux mêmes statistiques de test, respectivement pour les tests unilatéraux $\mathcal{H}_0 : \mu = \mu_0$ contre $\mathcal{H}_1 : \mu < \mu_0$ et $\mathcal{H}_0 : \sigma^2 = \sigma_0^2$ contre $\mathcal{H}_1 : \sigma^2 > \sigma_0^2$.

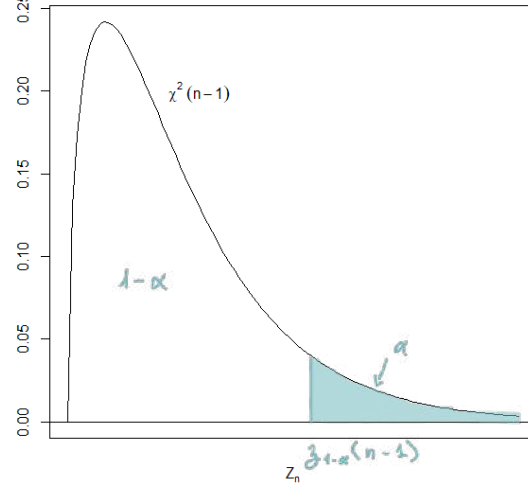
Sous $H_1: \mu < \mu_0$

$$Z = \frac{\bar{X}_n - \mu_0}{\sqrt{S_n^{*2}}} \rightarrow -\infty \quad \left. \begin{array}{l} \text{donc zone de} \\ \text{rejet à gauche} \end{array} \right\}$$



Sous $H_1: \sigma^2 > \sigma_0^2$

$$Z = \frac{(n-1) S_n^{*2}}{\sigma_0^2} \rightarrow +\infty \quad \left. \begin{array}{l} \text{donc zone de rejet à droite} \end{array} \right\}$$



Remarque. L'approche bayésienne (voir Sec. 2.4) permet également d'établir des *intervalles de crédibilité*. En reprenant les mêmes notations, tout intervalle I tel que $\mathbb{P}(\theta \in I | \underline{X}) = 1 - \alpha$ est un intervalle de crédibilité $1 - \alpha$ pour θ , c'est-à-dire tout intervalle I tel que

$$\int_I \pi(\theta | \underline{X}) d\theta = 1 - \alpha.$$

Il faut donc pour cela calculer la fonction quantile de la loi *a posteriori*. Mais il est important de noter que les intervalles fréquentistes et bayésiens ne sont pas directement comparables. Dans le second cas (bayésien), on peut vraiment dire que le paramètre a une probabilité $1 - \alpha$ de se trouver dans l'intervalle, alors que dans le premier cas (fréquentiste), le paramètre n'étant pas aléatoire, tout ce qu'on peut dire c'est qu'*avant* que les bornes soient observées, elles formaient des variables aléatoires qui avaient une probabilité $1 - \alpha$ d'encadrer la vraie valeur du paramètre.

3.1.2 Intervalles de confiance asymptotiques

Définition 3.4 Pour un risque $0 < \alpha < 1$, un *intervalle de confiance asymptotique* pour le paramètre θ est un *intervalle aléatoire* $[B_{\inf}(\underline{X}), B_{\sup}(\underline{X})] = [b_1, b_2]$ construit sur un échantillon $\underline{X} = (X_1, \dots, X_n)$ de telle sorte que

$$\forall \theta \in \Theta, \quad \lim_{n \rightarrow +\infty} \mathbb{P}_\theta(b_1 \leq \theta \leq b_2) \geq 1 - \alpha.$$

On notera $ICA_{1-\alpha}(\theta) = [b_1, b_2]$ cet intervalle pour la sécurité $1 - \alpha$.

Les intervalles asymptotiques ne sont pertinents que pour les grandes valeurs de n , mais ils sont plus faciles à obtenir que les intervalles exacts. L'outil essentiel est le TCL. Par exemple si l'espérance $\mathbb{E}_\theta[X] = m$ et la variance $\mathbb{V}_\theta(X) = v > 0$ existent, alors

$$\sqrt{n} \frac{\bar{X}_n - m}{\sqrt{v}} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1).$$

Si l'on sait estimer v de façon consistante par \hat{v}_n , on obtient par continuité et par le lemme de Slutsky,

$$\sqrt{n} \frac{\bar{X}_n - m}{\sqrt{\hat{v}_n}} = \sqrt{n} \frac{\bar{X}_n - m}{\sqrt{v}} \underbrace{\sqrt{\frac{v}{\hat{v}_n}}}_{\xrightarrow{\mathbb{P}} 1 \text{ (CMT)}} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1)$$

et donc, par définition de la convergence en loi,

$$\lim_{n \rightarrow +\infty} \mathbb{P}_\theta \left(-u_{1-\frac{\alpha}{2}} \leq \sqrt{n} \frac{\bar{X}_n - m}{\sqrt{\hat{v}_n}} \leq u_{1-\frac{\alpha}{2}} \right) = 1 - \alpha.$$

On en déduit l'intervalle de confiance asymptotique pour l'espérance

$$\text{ICA}_{1-\alpha}(m) = \left[\bar{X}_n \pm \frac{\sqrt{\hat{v}_n}}{\sqrt{n}} u_{1-\frac{\alpha}{2}} \right].$$

Parfois, la variance n'a pas besoin d'être estimée car elle est connue sous \mathcal{H}_0 (notons-la v_0). Dans ce cas, on a plus simplement

$$\text{ICA}_{1-\alpha}(m) = \left[\bar{X}_n \pm \frac{\sqrt{v_0}}{\sqrt{n}} u_{1-\frac{\alpha}{2}} \right].$$

Exemples. Pour le test de $\mathcal{H}_0 : "p = p_0"$ dans le modèle $\mathcal{B}(p)$, on peut utiliser directement $v_0 = p_0(1 - p_0)$. Par contre, dans le test de $\mathcal{H}_0 : "\mu = \mu_0"$ dans le modèle $\mathcal{N}(\mu, \sigma^2)$, on ne peut pas exprimer σ^2 en fonction de μ_0 et il faudra donc l'estimer, par exemple avec $\hat{v}_n = S_n^{*2}$.

Exemples. À partir du modèle $\mathcal{B}(p)$, construire le test de la proportion.

$X \sim \mathcal{B}(p)$ TCL : $\sqrt{n} \frac{\bar{X}_n - p}{\sqrt{p(1-p)}} \xrightarrow{L} \mathcal{N}(0, 1)$

Test de la proportion : " $\mathcal{H}_0 : p = p_0$ " vs " $\mathcal{H}_1 : p \neq p_0$ "

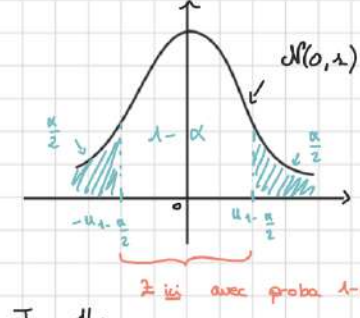
Si \mathcal{H}_0 est vraie : $Z = \sqrt{n} \frac{\bar{X}_n - p}{\sqrt{p(1-p)}} \xrightarrow{L} \mathcal{N}(0, 1)$.

- Si $Z \in \left[\pm u_{1-\frac{\alpha}{2}} \right]$: non rejet de \mathcal{H}_0
- Si $Z \notin \left[\pm u_{1-\frac{\alpha}{2}} \right]$: rejet de \mathcal{H}_0 au profit de \mathcal{H}_1

$T = \{ |Z| > u_{1-\frac{\alpha}{2}} \}$

Exemple : pièce équilibrée ? " $\mathcal{H}_0 : p_0 = \frac{1}{2}$ " vs " $\mathcal{H}_1 : p_0 \neq \frac{1}{2}$ " $n = 1000$ $\bar{X}_n = \frac{550}{1000}$

$Z = \sqrt{1000} \frac{\frac{550}{1000} - \frac{1}{2}}{\sqrt{\frac{1}{2} \cdot \frac{1}{2}}} \approx 3,16 \notin \left[\pm 1,96 \right]$. Donc \mathcal{H}_0 est rejeté au risque 5%



Exemple. On injecte un remède à un groupe de $n = 80$ souris contaminées. On pense que le remède est efficace dans 80 % des cas. Après l'expérience, on observe que 22 souris n'ont pas survécu. Vérifier ou infirmer l'hypothèse de départ au risque de 5 % à l'aide d'un test de la proportion. 58 souris ont survécu

Soit $X \sim \mathcal{B}(p)$: $X = 1$ (guérison), $X = 0$ (sinon). " $\mathcal{H}_0 : p = 0,8$ " vs " $\mathcal{H}_1 : p \neq 0,8$ "

Observations : $\bar{X}_n = \frac{58}{80}$ $Z = \sqrt{80} \frac{\frac{58}{80} - 0,8}{\sqrt{0,8 \cdot 0,2}} = -1,68 \notin \left[\pm 1,96 \right]$

$| -1,68 | < 1,96$

Donc non rejet de \mathcal{H}_0 , au risque 5%

Mais $0,8 \cdot 80 = 64$ souris aurait dû survivre. Et on a $58 < 64$

" $\mathcal{H}_0 : p = 0,8$ " vs " $\mathcal{H}_1 : p < 0,8$ "

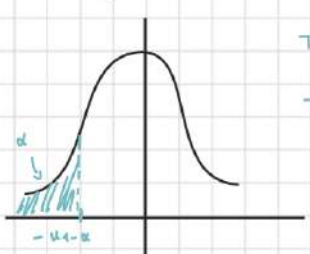
$T = \{ |Z| > u_{1-\frac{\alpha}{2}} \}$ (bilatéral)

$T = \{ Z < -u_{1-\alpha} \}$ (unilatéral à gauche)

$-u_{1-0,05} = -1,64$

Or $-1,68 < -1,64$

Donc **rejet** au risque 5%



3.2 Risques et puissance

La théorie des tests d'hypothèses possède un vocabulaire assez spécifique que l'on va passer en revue. Il existe deux façons de se tromper suite à un test statistique : décider \mathcal{H}_1 sachant que \mathcal{H}_0 est vraie (*condamner un innocent*) et symétriquement, ne pas décider \mathcal{H}_1 sachant que \mathcal{H}_1 est vraie (*acquitter un coupable*).

Définition 3.5 Soit $T = T(\underline{X})$ un test statistique de $\mathcal{H}_0 : \theta \in \Theta_0$ contre $\mathcal{H}_1 : \theta \in \Theta_1$. L'erreur de 1^{ère} espèce est l'application de Θ_0 dans $[0, 1]$ définie par

$$\forall \theta \in \Theta_0, \quad \alpha(\theta) = \mathbb{E}_0[T].$$

\uparrow
sous \mathcal{H}_0

Le test est de seuil α si

$$\sup_{\theta \in \Theta_0} \alpha(\theta) \leq \alpha.$$

Si de plus il existe une valeur $\theta_0 \in \Theta_0$ telle que $\alpha(\theta_0) = \alpha$, alors le test est de niveau α .

On voit que, dans le cas simplifié où $T(\Omega) = \{0, 1\}$,

$$\mathbb{P}_0(\mathcal{R}) = \mathbb{P}_0(T = 1) = \mathbb{E}_0[T].$$

$\mathcal{R} = \{T=1\}$: "événement du rejet de \mathcal{H}_0 "

L'erreur de 1^{ère} espèce est donc la probabilité de décider \mathcal{H}_1 sous \mathcal{H}_0 (la probabilité de condamner un innocent). On produit alors un faux-positif. On cherchera toujours à contrôler ce risque en fixant α au préalable (à 5 % ou 1 %, en général).

Définition 3.6 Soit $T = T(\underline{X})$ un test statistique de $\mathcal{H}_0 : \theta \in \Theta_0$ contre $\mathcal{H}_1 : \theta \in \Theta_1$. L'erreur de 2^{nde} espèce est l'application de Θ_1 dans $[0, 1]$ définie par

$$\forall \theta \in \Theta_1, \quad \beta(\theta) = 1 - \mathbb{E}_1[T] = 1 - \pi(\theta)$$

et $\pi(\theta) = \mathbb{E}_1[T]$ est la puissance du test.

On voit de même que

$$\mathbb{P}_1(\bar{\mathcal{R}}) = 1 - \mathbb{P}_1(T = 1) = 1 - \mathbb{E}_1[T].$$

$\bar{\mathcal{R}}$: "événement de non-rejet de \mathcal{H}_0 "

La puissance est donc la probabilité de décider \mathcal{H}_1 à raison (la probabilité de condamner un coupable) et l'erreur de 2^{nde} espèce est la probabilité de ne pas décider \mathcal{H}_1 sous \mathcal{H}_1 (la probabilité d'acquitter un coupable). On produit alors un faux-négatif.

	Décision \mathcal{H}_0 (on l'acquitte)	Décision \mathcal{H}_1 (on le condamne)
\mathcal{H}_0 est vraie (il est innocent)	$1 - \alpha(\theta) \geq 1 - \alpha$	$\alpha(\theta) \leq \alpha$
\mathcal{H}_1 est vraie (il est coupable)	$\beta(\theta)$	$\pi(\theta)$

3.3 Optimalité de Neyman-Pearson

Nous nous limiterons dans cette section aux hypothèses nulles de la forme $\Theta_0 = \{\theta_0\}$. On remarque tout d'abord qu'il n'est pas possible de minimiser simultanément les deux risques. En effet, si l'on choisit $T = 0$ alors le risque de 1^{ère} espèce est minimisé : $\alpha(\theta) = \mathbb{E}_0[T] = 0$. Par contre, $\beta(\theta) = 1 - \mathbb{E}_1[T] = 1$ et donc le risque de 2^{nde} espèce est maximisé. Si l'on décide de ne jamais rejeter \mathcal{H}_0 , alors on aura toujours raison sous \mathcal{H}_0 mais toujours tort sous \mathcal{H}_1 . On obtient une situation symétriquement comparable en posant $T = 1$. L'optimalité de Neyman-Pearson consiste à fixer le seuil α (la borne supérieure de l'erreur de 1^{ère} espèce) et à minimiser la borne supérieure de l'erreur de 2^{nde} espèce en contrepartie. Ainsi pour un risque contrôlé

sous \mathcal{H}_0 , on maximise la puissance sous \mathcal{H}_1 . On peut montrer qu'une telle stratégie repose sur le rapport entre les vraisemblances sous \mathcal{H}_1 et \mathcal{H}_0 . Pour simplifier, on notera par la suite

$$\underset{\text{Sous } \mathcal{H}_0}{\ell_0 = \ell_X(X; \theta_0)} \quad \text{et} \quad \underset{\text{Sous } \mathcal{H}_1}{\ell_1 = \ell_X(X; \theta_1)}.$$

Lorsque le rapport des vraisemblances,

$$R = \frac{\ell_1}{\ell_0}$$

tend à la hausse, c'est \mathcal{H}_1 qui gagne en vraisemblance : ℓ_1 augmente pendant que ℓ_0 diminue, et le phénomène inverse se produit lorsqu'il tend à la baisse. L'idée est donc de déterminer à partir de quelle frontière k on passe de \mathcal{H}_1 à \mathcal{H}_0 .

3.3.1 Hypothèse simple contre hypothèse simple

On veut tester $\mathcal{H}_0 : \theta = \theta_0$ contre $\mathcal{H}_1 : \theta = \theta_1$. Dans ce cas particulier,

$$\sup_{\theta \in \Theta_0} \alpha(\theta) = \alpha(\theta_0) \quad \text{et} \quad \sup_{\theta \in \Theta_1} \beta(\theta) = \beta(\theta_1).$$

$$\Theta_0 = \{\theta_0\}$$

$$\Theta_1 = \{\theta_1\}$$

Théorème 3.2 (Lemme de Neyman-Pearson (simplifié)) On peut construire un test qui, pour un seuil $0 < \alpha < 1$ fixé, est le plus puissant pour tester $\mathcal{H}_0 : \theta = \theta_0$ contre $\mathcal{H}_1 : \theta = \theta_1$. Ce test s'écrit

$$T = \begin{cases} 1 & \text{si } \ell_1 > k \ell_0 \\ 0 & \text{si } \ell_1 \leq k \ell_0. \end{cases}$$

Le problème réside dans le choix de k qui nécessite la connaissance de la loi de R sous \mathcal{H}_0 , puisqu'on veut que $\mathbb{P}_0(R > k) \leq \alpha$, ce qui peut être extrêmement compliqué. On simplifie le problème si l'on trouve une équivalence de la forme

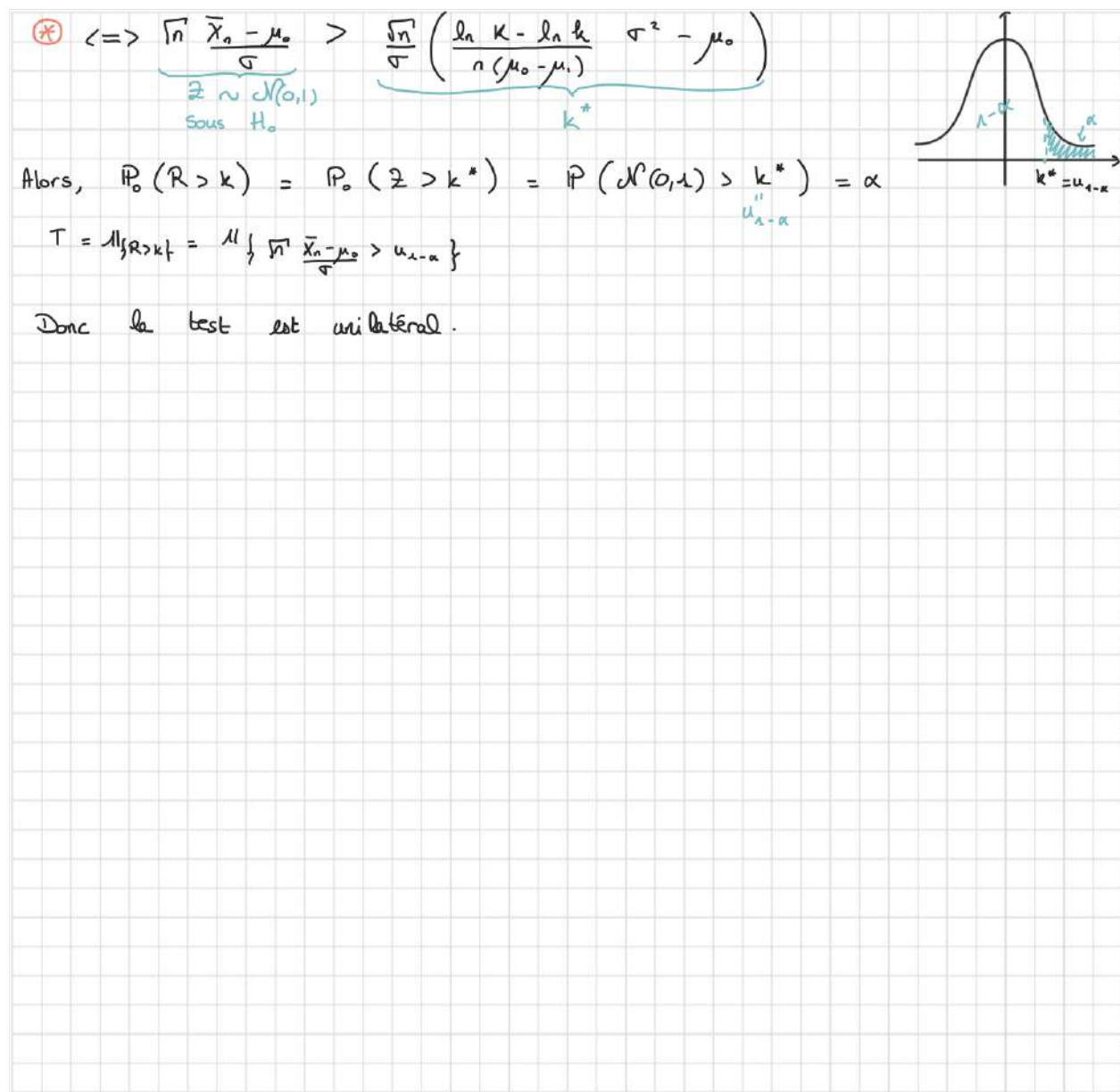
$$\{R > k\} = \{Z > k^*\}$$

pour une statistique Z dont on connaît la loi sous \mathcal{H}_0 . Une fois k déterminé, le test de Neyman-Pearson s'écrit simplement

$$T = \mathbb{1}_{\{R > k\}}.$$

Exemple. Soit $X \sim \mathcal{N}(\mu, \sigma^2)$ pour une variance σ^2 connue. On souhaite construire le test le plus puissant pour tester $\mathcal{H}_0 : \mu = \mu_0$ contre $\mathcal{H}_1 : \mu = \mu_1$, où $\mu_0, \mu_1 \in \mathbb{R}$ vérifient $\mu_1 > \mu_0$. Le test est-il unilatéral ou bilatéral ?

$$\begin{aligned} \ell_0 &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i - \mu_0)^2}{2\sigma^2}} = c e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu_0)^2} \quad (c > 0, \text{ ne dépend pas de } \mu_0) \\ \ell_1 &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i - \mu_1)^2}{2\sigma^2}} = c e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu_1)^2} \\ R = \frac{\ell_1}{\ell_0} &= e^{\frac{n\bar{X}_n}{\sigma^2}(\mu_0 - \mu_1) + \frac{n}{2\sigma^2}(\mu_0 - \mu_1)^2} = K e^{-\frac{n\bar{X}_n}{\sigma^2}(\mu_0 - \mu_1)} \quad K > 0 \text{ n'est pas aléatoire} \\ \{R \geq k\} &\text{ est de probabilité très compliquée à calculer ...} \\ \text{Mais : } R \geq k &\Leftrightarrow \ln K - \frac{n\bar{X}_n}{\sigma^2}(\mu_0 - \mu_1) > \ln k \quad (\ln \text{ strictement croissante}). \\ \Leftrightarrow \frac{n\bar{X}_n}{\sigma^2}(\mu_0 - \mu_1) < \ln K - \ln k &\Leftrightarrow \bar{X}_n > \frac{(\ln K - \ln k) \cdot \sigma^2}{n(\mu_0 - \mu_1)} \quad (*) \\ \text{Or sous } \mathcal{H}_0, \bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i &\sim \mathcal{N}(\mu_0, \frac{\sigma^2}{n}) \quad \text{d'où} \quad \frac{\bar{X}_n - \mu_0}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0, 1). \\ &\quad \triangle \text{ Ce n'est pas le TCL !!} \end{aligned}$$



3.3.2 Hypothèse simple contre hypothèse composite

Définition 3.7 Soit $T = T(\underline{X})$ un test statistique de $\mathcal{H}_0 : \theta \in \Theta_0$ contre $\mathcal{H}_1 : \theta \in \Theta_1$ de seuil α . Le test T est uniformément le plus puissant parmi tous les tests de seuil α si, pour tout autre test $T' = T'(\underline{X})$ de seuil α , on a

$$\forall \theta \in \Theta_1, \quad \mathbb{E}_1[T] \geq \mathbb{E}_1[T'].$$

On dit alors que T est UPP_α .

On veut tester $\mathcal{H}_0 : \theta = \theta_0$ contre $\mathcal{H}_1 : \theta \in \Theta_1$. Une fois que $\theta_1 \in \Theta_1$ est fixé, on sait construire par le lemme de Neyman-Pearson le test le plus puissant de $\mathcal{H}_0 : \theta = \theta_0$ contre $\mathcal{H}_1 : \theta = \theta_1$ au niveau $0 < \alpha < 1$. Le test est UPP_α pour $\mathcal{H}_0 : \theta = \theta_0$ contre $\mathcal{H}_1 : \theta \in \Theta_1$ si la conclusion reste valable pour toute valeur $\theta \in \Theta_1$ (et pas seulement pour la valeur θ_1 isolée). C'est en général le cas des tests unilatéraux.

Exemple. Montrer que le test construit dans la section précédente est UPP_α pour tester $\mathcal{H}_0 : \mu = \mu_0$ contre $\mathcal{H}_1 : \mu > \mu_0$.

$$T = \mathbb{1} \left\{ \sqrt{n} \frac{\bar{X}_n - \mu_0}{\sigma} > u_{1-\alpha} \right\} \quad \text{pas d'intervention directe de } \mu, \text{ mais indirecte } (> / < ; \mu_{1-\alpha} / -u_{1-\alpha}).$$

Valable $\forall \mu$ à condition que $\mu > \mu_0$
 \Rightarrow Valable sur $\Theta_1 =]\mu_0, +\infty[$

Pour les tests bilatéraux, il est rare d'obtenir l'uniformité. Si $\Theta_1 = \Theta \setminus \{\theta_0\}$ et que l'on teste $\mathcal{H}_0 : \theta = \theta_0$ contre $\mathcal{H}_1 : \theta \neq \theta_0$, alors, comme le lemme de Neyman-Pearson entraîne généralement que le test le plus puissant au niveau α pour $\mathcal{H}_0 : \theta = \theta_0$ contre $\mathcal{H}_1 : \theta = \theta_1$ possède une zone de rejet différente selon que $\theta_1 > \theta_0$ ou que $\theta_1 < \theta_0$, il est clair qu'un test bilatéral ne sera pas UPP_α . Dans ce cas, on propose deux zones de rejet d'aire $\frac{\alpha}{2}$ correspondant à chacune de ces alternatives.

Exemple. On reprend à nouveau l'exemple précédent, avec cette fois $\mathcal{H}_0 : \mu = \mu_0$ contre $\mathcal{H}_1 : \mu \neq \mu_0$. On choisit

$$T = \mathbb{1} \{ \{Z < -u_{1-\frac{\alpha}{2}}\} \cup \{Z > u_{1-\frac{\alpha}{2}}\} \} = \mathbb{1} \{|Z| > u_{1-\frac{\alpha}{2}}\} \quad \text{où} \quad Z = \sqrt{n} \frac{\bar{X}_n - \mu_0}{\sigma}.$$

Si l'on remarque que $\Theta_1 = \Theta_1^- \cup \Theta_1^+$ avec $\Theta_1^- =]-\infty, \mu_0[$ et $\Theta_1^+ =]\mu_0, +\infty[$, alors on a vu qu'on pouvait construire un test de $\mathcal{H}_0 : \mu = \mu_0$ contre $\mathcal{H}_1 : \mu = \mu_1$ différent de celui-ci et UPP_α pour ces alternatives, selon que $\mu_1 \in \Theta_1^-$ ou que $\mu_1 \in \Theta_1^+$. Ce test ne peut donc pas être UPP_α .

Remarque. Une thématique de la statistique moderne concerne les *tests multiples*. Sans entrer dans les détails, l'idée est que lorsqu'une étude demande beaucoup de tests simultanés, on ne contrôle plus du tout le risque de première espèce de la même façon. Pour s'en convaincre, supposons que l'on mesure l'expression d'un ensemble de gènes dans les conditions A (patients sains) et les conditions B (patients malades), pour détecter une corrélation éventuelle entre l'expression de ces gènes et la maladie. Si l'on teste, pour chaque gène, $\mathcal{H}_{0,g} : \text{"égalité des moyennes dans les conditions A et B pour le gène g"}$ contre $\mathcal{H}_{1,g} = \mathcal{H}_{0,g}$ à l'aide de l'une des procédures vues précédemment, alors la probabilité de considérer à tort qu'un gène est influent est de α . Pour $\alpha = 5\%$, si l'on teste 10000 gènes, on aura en moyenne 500 faux-positifs. C'est dramatique d'un point de vue économique et médical, puisque les expériences devront se poursuivre à tort sur tous ces gènes. Comment limiter les faux-positifs tout en ne passant pas à côté des vrais-positifs? C'est tout l'enjeu des tests multiples (que l'on ne développera pas plus dans ce module).

3.4 Tests du khi-deux → va sûrement bomber en exam, "3 ou 4 points offerts" selon PROIA.

Les tests du khi-deux, très répandus, s'intéressent à des caractéristiques globales d'une ou de plusieurs populations. Contrairement à ceux que l'on vient d'étudier, ce sont des tests *non paramétriques*.

3.4.1 Adéquation

On suppose ici que la v.a.r. parente X du n -échantillon X_1, \dots, X_n est *discrète*, à valeurs dans un ensemble de modalités $X(\Omega) = \{x_1, \dots, x_r\}$ avec $2 \leq r < +\infty$. Pour tout $i \in \{1, \dots, r\}$, on note

$$p_i = \mathbb{P}(X = x_i). \quad \sum_{i=1}^r p_i = 1$$

Le vecteur $p = (p_1, \dots, p_r)$ contient donc la loi de probabilité de X , et l'on suppose de plus que $\forall i, p_i > 0$. L'*effectif empirique* de la modalité x_i est la v.a.r. définie par

$$N^{(i)} = \sum_{k=1}^n \mathbb{1}_{\{X_k = x_i\}}. \quad \sum_{i=1}^r N^{(i)} = n$$

Pour une loi discrète $p^0 = (p_1^0, \dots, p_r^0)$, on s'intéresse au test de $\mathcal{H}_0 : "p = p^0"$ contre $\mathcal{H}_1 : "p \neq p^0"$: ainsi, on cherche l'adéquation entre la loi de X et une loi-test p^0 . La statistique de test du khi-deux prend la forme

$$D^2 = \sum_{i=1}^r \frac{(N^{(i)} - n p_i^0)^2}{n p_i^0} \quad \rightarrow \text{plus } D^2 \text{ est petit, plus ça donne des crédits à } \mathcal{H}_0.$$

Proposition 3.2 Supposons que $\mathcal{H}_0 : "p = p^0"$ est vraie. Alors,

$$D^2 \xrightarrow{\mathcal{L}} \chi^2(r-1).$$

Si \mathcal{H}_1 est vraie (et donc qu'il existe i tel que $p_i \neq p_i^0$), D^2 va avoir tendance à grandir très vite, c'est pourquoi le test à considérer est donné par

$$T = \mathbb{1}_{\{D^2 > z_{1-\alpha}\}} \quad \text{unilatéral (asymptotique)}$$

où $z_{1-\alpha}$ désigne le quantile correspondant de la loi $\chi^2(r-1)$. Cela revient à comparer la réalisation de D^2 avec les quantiles de la loi $\chi^2(r-1)$ pour obtenir la zone de rejet illustrée en fin de section.

Exemple. On cherche à savoir si les décimales de π sont uniformément réparties sur $\{0, \dots, 9\}$. Ses $n = 2400$ premières décimales donnent les effectifs suivants :

0	1	2	3	4	5	6	7	8	9
213	240	251	216	243	253	242	229	249	264

Effectuer le test avec un risque de 5 %.

X est une v.a.r par $\{0, 1, \dots, 9\}$ Ici $p_i = \mathbb{P}(X = i)$ avec $i \in \{0, 9\}$

Sous \mathcal{H}_0 : " les décimales de π sont uniformément réparties "

: " $p_i = \frac{1}{10} \forall i$ " d'où $p^0 = (\frac{1}{10}, \dots, \frac{1}{10})$.

($\mathcal{H}_1 = \bar{\mathcal{H}}_0$ ou on écrit \mathcal{H}_1 : " $\exists i$ tq $p_i \neq \frac{1}{10}$ ").

$D^2 = \frac{(213-240)^2}{240} + \frac{(240-240)^2}{240} + \dots + \frac{(264-240)^2}{240} \approx 9,94.$

$T = \mathbb{1}_{\{D^2 > z_{0,95}\}} = 0$ car $z_{0,95}(9) = qchisq(0,95, 9) \approx 16,92$ sur \mathbb{R}

On ne rejette pas \mathcal{H}_0 , les décimales de π peuvent être considérées comme uniformément réparties, avec un risque de 5%, sur la base de 2400 observations.

Remarque. Lorsque l'on souhaite tester si deux échantillons indépendants sont issus d'une même loi de probabilité discrète, on peut construire un test très similaire conduisant à la même zone de rejet (avec deux vecteurs de fréquences empiriques et non plus un vecteur de fréquences empiriques et une loi théorique).

Remarque. Lorsque la loi théorique est continue, il est nécessaire de la *discrétiser* en classes pour effectuer le test. La discrétisation doit être effectuée de façon pertinente en fonction des effectifs, car la procédure y est évidemment très sensible.

3.4.2 Indépendance

On suppose maintenant que le n -échantillon est formé de couples discrets $(X_1, Y_1), \dots, (X_n, Y_n)$. Le vecteur (X, Y) prend ses valeurs dans un ensemble de couples de modalités $(X, Y)(\Omega) = \{x_1, \dots, x_r\} \times \{y_1, \dots, y_s\}$ avec $2 \leq r, s < +\infty$. Sous la forme d'un *tableau de contingence*, on note

$$p_{i,j} = \mathbb{P}(X = x_i, Y = y_j) \quad \text{et} \quad N^{(i,j)} = \sum_{k=1}^n \mathbb{1}_{\{X_k = x_i, Y_k = y_j\}}$$

$$\sum_{i=1}^r \sum_{j=1}^s p_{i,j} = 1 \quad 44 \quad \sum_{i=1}^r \sum_{j=1}^s N^{(i,j)} = n$$

les proportions théoriques et les effectifs empiriques associés à chaque couple (x_i, y_j) de modalités. On s'intéresse au test de $\mathcal{H}_0 : "X \perp Y"$ contre $\mathcal{H}_1 : "X \not\perp Y"$. Sous \mathcal{H}_0 , l'indépendance conduit à

$$p_{i,j} = \mathbb{P}(X = x_i) \mathbb{P}(Y = y_j)$$

et les effectifs théoriques se calculent par proportionnalité,

$$E^{(i,j)} = \frac{1}{n} \sum_{v=1}^s N^{(i,v)} \sum_{u=1}^r N^{(u,j)} \quad \left(= \frac{\text{somme en ligne } i \times \text{somme en colonne } j}{\text{nombre d'observations}} \right).$$

La statistique de test du khi-deux prend la forme

$$D^2 = \sum_{i=1}^r \sum_{j=1}^s \frac{(N^{(i,j)} - E^{(i,j)})^2}{E^{(i,j)}}.$$

Proposition 3.3 Supposons que $\mathcal{H}_0 : "X \perp Y"$ est vraie. Alors,

$$D^2 \xrightarrow{\mathcal{L}} \chi^2((r-1)(s-1)).$$

Si \mathcal{H}_1 est vraie (et donc qu'il existe (i, j) tel que $p_{i,j} \neq \mathbb{P}(X = x_i) \mathbb{P}(Y = y_j)$), D^2 va avoir tendance à grandir très vite, c'est pourquoi le test à considérer est donné par

$$T = \mathbb{1}_{\{D^2 > z_{1-\alpha}\}}$$

où $z_{1-\alpha}$ désigne le quantile correspondant de la loi $\chi^2((r-1)(s-1))$. Cela revient à comparer la réalisation de D^2 avec les quantiles de la loi $\chi^2((r-1)(s-1))$ pour obtenir la zone de rejet illustrée en fin de section.

Exemple. Un clinicien cherche à savoir si la dose administrée au cours d'un traitement a une influence significative sur l'effet du traitement. Ses mesures sont les suivantes :

Dose administrée	Guérisons	Risques de rechute	Inefficacité	Décès	Total
20 mL	20	45	35	17	117
50 mL	111	129	25	55	320
70 mL	145	252 $N^{(3,2)}$	79	35	511
100 mL	63	82	52	2	199
Total	339	508	191	109	1147 $=n$

Effectuer le test avec un risque de 5 %.

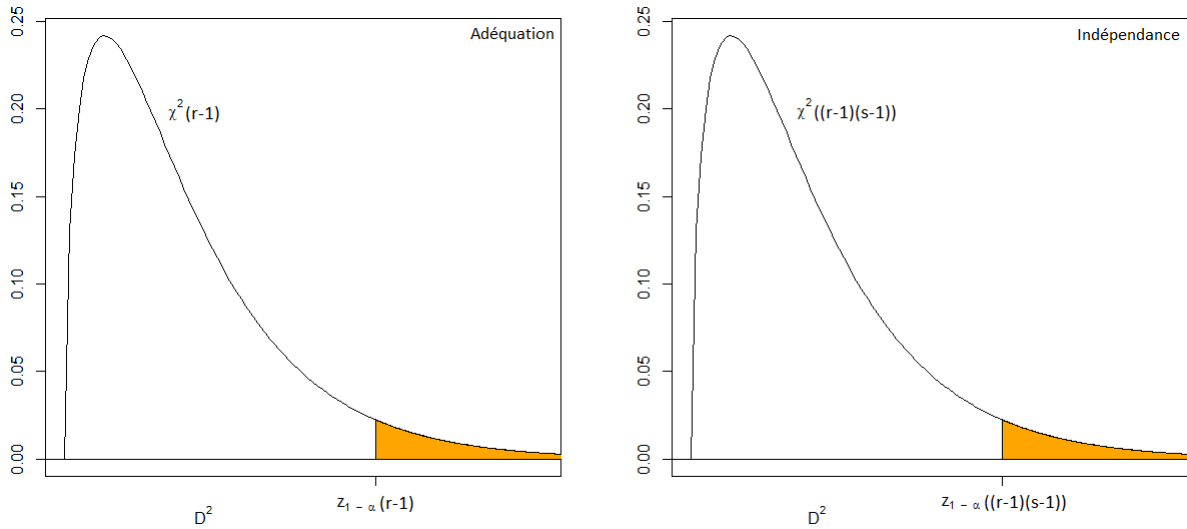
$H_0 : "X \perp Y"$ $X \in \{20 \text{ mL}, \dots, 100 \text{ mL}\}$ et $Y \in \{G, \dots, D\}$ $H_1 : "X \not\perp Y"$

$E^{(1,1)} = \frac{339 \times 117}{1147} \approx 34,58$, $E^{(3,2)} = \frac{508 \times 511}{1147} \approx 226,32$

$D^2 = \frac{\left(20 - \frac{339 \times 117}{1147}\right)^2}{\frac{339 \times 117}{1147}} + \frac{\left(252 - \frac{508 \times 511}{1147}\right)^2}{\frac{508 \times 511}{1147}} + \dots \approx 95,37$

$T = \mathbb{1}_{\{D^2 > z_{0,95}\}} = 1$ $z_{0,95} \approx 16,92$ quantile de $\chi^2(3.3)$.

On rejette H_0 avec un risque de 5% , il semble avoir un effet de la dose sur la réaction.

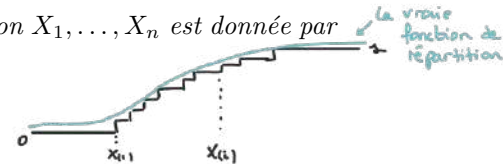


3.5 Test de Kolmogorov-Smirnov

Le *test de Kolmogorov-Smirnov* est probablement le plus utilisé en pratique pour tester l'adéquation d'un échantillon à une loi connue (par exemple sa normalité). Il s'agit de comparer une fonction de répartition donnée F_0 à la *répartition empirique* de l'échantillon. Plaçons-nous pour simplifier dans le cas continu avec aucun doublon dans les observations (sinon cela complique légèrement le déroulement du test).

Définition 3.8 La fonction de répartition empirique associée à un n -échantillon X_1, \dots, X_n est donnée par

$$\forall x \in \mathbb{R}, \quad \hat{F}_n(x) = \frac{1}{n} \sum_{k=1}^n \mathbb{1}_{\{X_k \leq x\}}.$$



C'est donc une fonction croissante, càdlàg, qui part de 0 pour atteindre 1 sous la forme d'un escalier (dont les sauts se situent sur les X_k). On verra en TD qu'elle possède de très bonnes propriétés d'estimation fonctionnelle pour la vraie répartition F_X de l'échantillon. Pour une fonction de répartition donnée F_0 , on peut tester

$$\mathcal{H}_0 : "F_X = F_0" \quad \text{contre} \quad \mathcal{H}_1 : "F_X \neq F_0"$$

à l'aide du résultat suivant.

Théorème 3.3 (Théorème de Kolmogorov) On a

$$\Delta_n = \sqrt{n} \|\hat{F}_n - F_X\|_\infty \xrightarrow{\mathcal{L}} \mathcal{K}$$

où \mathcal{K} est une v.a.r. distribuée selon la loi de Kolmogorov.

La loi de \mathcal{K} n'admet pas d'écriture explicite, mais on sait évaluer ses quantiles. On montre que

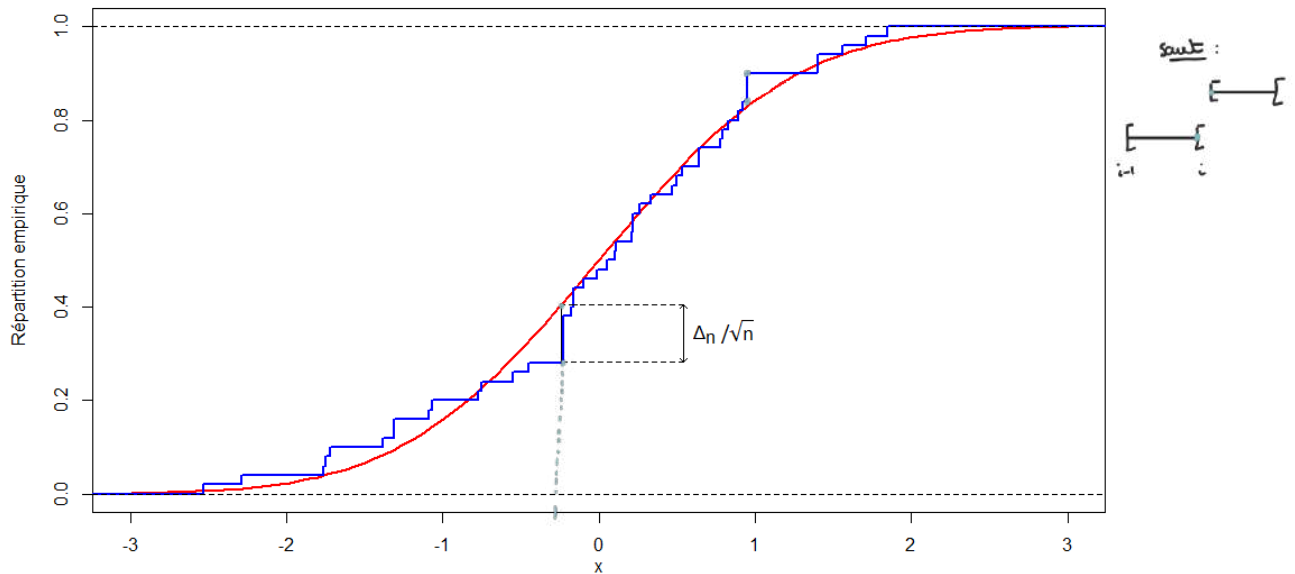
$$\forall t > 0, \quad \mathbb{P}(\mathcal{K} > t) = 2 \sum_{k=1}^{+\infty} (-1)^{k-1} e^{-k^2 t^2}$$

d'où l'on déduit les quantiles en résolvant approximativement l'équation $\mathbb{P}(\mathcal{K} > k_{1-\alpha}) = \alpha$. On constate que cette distribution limite ne dépend pas de F_0 . En pratique, à l'aide de l'échantillon trié $X_{(1)}, \dots, X_{(n)}$,

$$\Delta_n = \sqrt{n} \max_{i=1, \dots, n} \max \left\{ \left| F_0(X_{(i)}) - \frac{i}{n} \right|, \left| F_0(X_{(i)}) - \frac{i-1}{n} \right| \right\}$$

car on voit que $\hat{F}_n(X_{(i)}) = \frac{i}{n}$. On calcule donc le plus grand écart entre la courbe en escalier et la répartition théorique de la loi à tester. Le test est donné par

$$T = \mathbb{1}_{\{\Delta_n > k_{1-\alpha}\}} \quad \text{unilatéral}$$



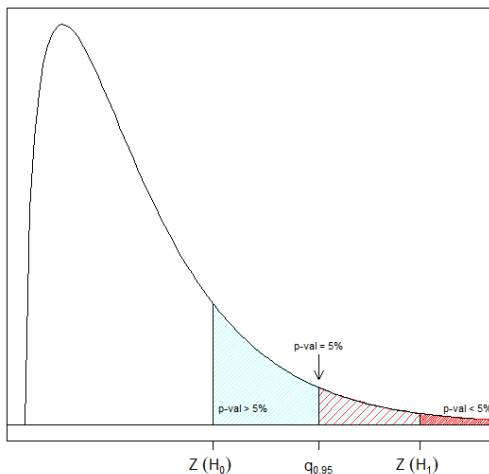
Il existe également une variante du test dans laquelle on compare l'égalité de deux répartitions empiriques (et donc l'égalité en loi de deux échantillons).

3.6 La p-valeur

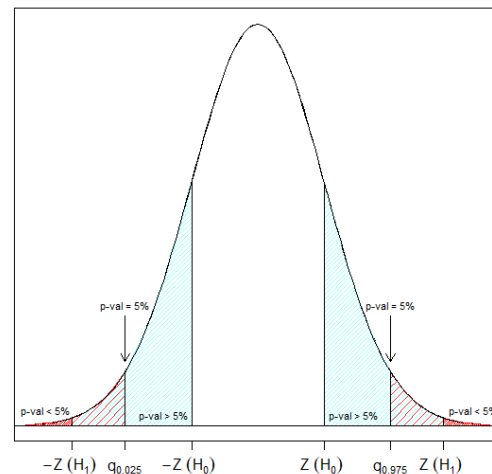
Lors de la mise en pratique d'un test, les logiciels de statistique fournissent généralement un indicateur appelé p-valeur. On définit

$$\alpha^*(\underline{X}) = \inf_{0 \leq \alpha \leq 1} \{ \alpha \mid \text{l'échantillon } \underline{X} \text{ conduit au rejet de } \mathcal{H}_0 \}.$$

La p-valeur $p\text{-val} = \alpha^*(x)$ calculée par les logiciels est la réalisation de $\alpha^*(\underline{X})$ sur les données observées \underline{x} . Ci-dessous, on représente schématiquement la p-valeur dans le cas d'un test unilatéral et d'un test bilatéral pour $\alpha = 5\%$: la courbe est la distribution d'une statistique de test Z sous \mathcal{H}_0 et la p-valeur est l'aire sous la courbe calculée à partir de l'abscisse Z en allant vers les extrêmes qui caractérisent \mathcal{H}_1 .



$z^{-1}(0.95)$
lorsque z est inversible



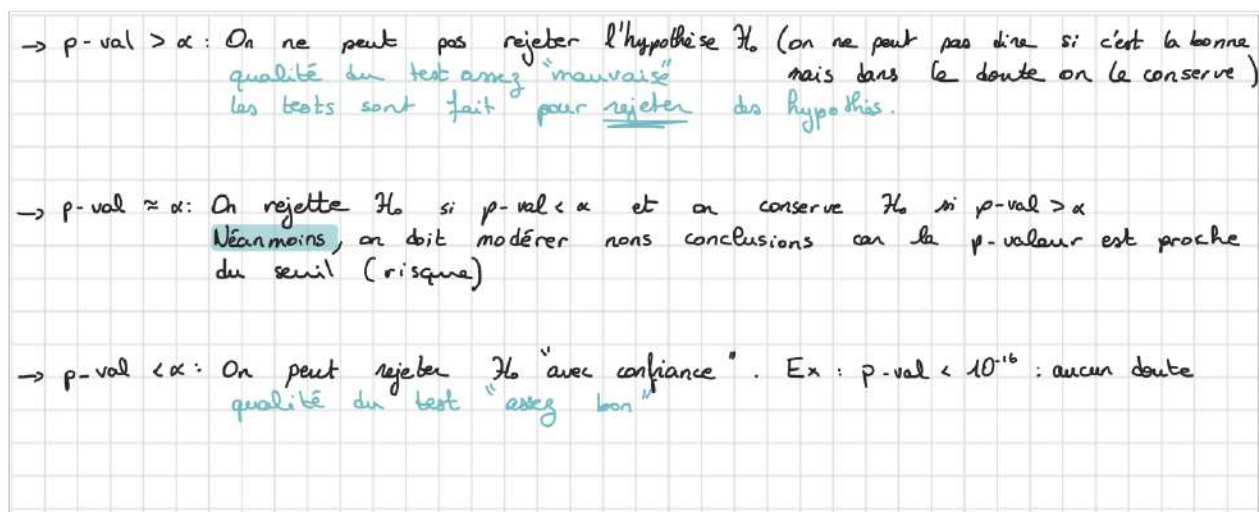
$$q_{1-\alpha} = F^{-1}(1-\alpha)$$

$$D_\alpha = \{x \in \mathbb{R}^n : h(x) > q_{1-\alpha}\}$$

$$\text{Soit } \underline{x}^{(n)} \text{ échantillon : } T = 1 \Leftrightarrow$$

$$\begin{aligned} h(\underline{x}^{(n)}) &> q_{1-\alpha} \\ \Leftrightarrow F(h(\underline{x}^{(n)})) &> 1-\alpha \\ \Leftrightarrow p\text{-val}(\underline{x}^{(n)}) &= 1 - F(h(\underline{x}^{(n)})) \end{aligned}$$

Pour mener un test de niveau α fixé, on comparera $p\text{-val}$ avec α selon la règle suivante :



Appelons z la valeur observée de la statistique de test Z et F_0 la fonction de répartition associée à la loi de Z sous \mathcal{H}_0 , que l'on suppose continue pour simplifier. Dans le cas d'un test unilatéral, on a donc

$$p\text{-val} = \mathbb{P}_0(Z < z) = F_0(z) \quad \text{ou} \quad p\text{-val} = \mathbb{P}_0(Z > z) = 1 - F_0(z)$$

selon que le test est unilatéral à gauche ou à droite. Si le test est bilatéral, on a

$$\begin{aligned} p\text{-val} &= \mathbb{P}_0(Z < -|z|) + \mathbb{P}_0(Z > |z|) = 1 + F_0(-|z|) - F_0(|z|) \\ &= 2(1 - F_0(|z|)) \text{ en cas de symétrie.} \end{aligned}$$

Exemples. On donne le code R de trois tests étudiés précédemment. Comprendre les commandes utilisées et interpréter les sorties du logiciel.

```
### Test de la moyenne dans un échantillon gaussien
```

```
### Cas bilatéral
```

```
> X = rnorm(1000, 1, sqrt(2))
> t.test(X, mu = 1, alternative = "two.sided", conf.level = 0.99)
```

One Sample t-test

data: X

$t = -0.09507$, $df = 999$, $p\text{-value} = 0.9243 \Rightarrow 1$ on ne rejette pas H_0
alternative hypothesis: true mean is not equal to 1

99 percent confidence interval: $IC_{99\%}$ de μ .
0.882255 1.109378

```
> sqrt(1000)*(mean(X) - 1)/sd(X)
[1] -0.09507028
> mean(X) - sd(X)/sqrt(1000)*qt(0.995, 999)
[1] 0.882255
> mean(X) + sd(X)/sqrt(1000)*qt(0.995, 999)
[1] 1.109378
> pt(-0.09507, 999) + 1 - pt(0.09507, 999)
```

```
[1] 0.9242783
```

```
### Test de la moyenne dans un échantillon gaussien
```

```
### Cas unilatéral
```

```
> t.test(X, mu = 0.5, alternative = "greater", conf.level = 0.95)
5%
```

One Sample t-test

```
data: X
```

```
t = 11.268, df = 999, p-value < 2.2e-16 <<  $\alpha$  → on rejette  $H_0$ 
```

```
alternative hypothesis: true mean is greater than 0.5
```

```
95 percent confidence interval:
```

```
0.9233707 Inf
```

```
> sqrt(1000)*(mean(X) - 0.5)/sd(X)
```

```
[1] 11.26775
```

```
> mean(X) - sd(X)/sqrt(1000)*qt(0.95, 999)
```

```
[1] 0.9233707
```

```
### Test du khi-deux d'adéquation
```

```
> Eff = c(213, 240, 251, 216, 243, 253, 242, 229, 249, 264)
```

```
> chisq.test(Eff, p = rep(1/10, 10))
```

Chi-squared test for given probabilities

```
data: Eff
```

```
X-squared = 9.9417, df = 9, p-value = 0.3552
```

```
> 1-pchisq(9.9417, 9)
```

```
[1] 0.3552255
```

```
### Test du khi-deux d'indépendance
```

```
> Tab = as.table(rbind(c(20, 45, 35, 17), c(111, 129, 25, 55),
```

```
  c(145, 252, 79, 35), c(63, 82, 52, 2)))
```

```
> dimnames(Tab) = list(Dose = c("20 mL", "50 mL", "70 mL", "100 mL"),
```

```
  Effet = c("Gué", "Rec", "Ine", "Dec"))
```

```
> chisq.test(Tab)
```

Pearson's Chi-squared test

```
data: Tab
```

```
X-squared = 95.37, df = 9, p-value < 2.2e-16 <<  $\alpha$  on rejette  $H_0$ 
```

```
### Test de Kolmogorov-Smirnov
```

```
> X = rnorm(100, 1, sqrt(2))
```

```
> ks.test(X, "pnorm", 1, sqrt(2))
```

One-sample Kolmogorov-Smirnov test

```
data: X
```

```
D = 0.12072, p-value = 0.1084
```

```
alternative hypothesis: two-sided
```

3.7 Un bestiaire de tests...

On donne, de manière évidemment non exhaustive, une liste de tests classiques regroupés par thèmes ainsi que, pour certains d'entre eux, le package et la fonction R permettant de les mettre en pratique. Ils peuvent être paramétriques, non paramétriques, exacts, asymptotiques, unilatéraux, bilatéraux, etc. Les étudiants intéressés chercheront d'eux-mêmes à en savoir plus, en commençant par la commande `help`.

- **Adéquation.** Adéquation entre observations et densité. Khi-deux (`stats::chisq.test`), Kolmogorov-Smirnov (`stats::ks.test`), Cramer-Von Mises.
- **Normalité.** Adéquation entre observations et loi normale. Shapiro-Wilk (`stats::shapiro.test`), Lilliefors (`nortest::lillie.test`), Jarque-Bera (`tseries::jarque.bera.test`), Anderson-Darling (`nortest::ad.test`).
- **Conformité.** Égalité entre moyennes, variances ou proportions. Student/Welch (`stats::t.test`), Fisher (`stats::var.test`), proportions (`stats::prop.test`).
- **Position.** Comparaison des rangs, des médianes, distributions symétriques. Wilcoxon/Mann-Whitney (`stats::wilcox.test`), signes (`stats::binom.test`), Kruskal-Wallis (`stats::kruskal.test`), Mood, Siegel-Tukey.
- **Significativité.** Significativité des paramètres dans les modèles de régression, modèles emboîtés, comparaison de modèles. Wald, Student, Fisher (dans le `summary`), Chow (`gap::chow.test`).
- **Homoscédasticité.** Homogénéité des variances entre groupes. Bartlett (`stats::bartlett.test`), Levene (`car::leveneTest`), Fligner-Killeen (`stats::fligner.test`), Breusch-Pagan (`lmtest::bptest`).
- **Corrélation.** Présence de corrélation entre séries ou d'autocorrélation dans une série. Coefficients de Pearson/Spearman/Kendall (`stats::cor.test`), Durbin-Watson (`lmtest::dwtest`), Breusch-Godfrey (`lmtest::bgtest`), H-test de Durbin, Box-Pierce ou Ljung-Box (`stats::Box.test`).
- **Indépendance.** Indépendance entre deux caractères couplés. Khi-deux (`stats::chisq.test`).

Exemple. Une course oppose une série de 15 lièvres à une série de 15 tortues. On note (L_i, T_i) les temps correspondant aux 15 duels, en unités de mesure adéquates : l'échantillon est *apparié*.

L_i	11	10	10	10	11	12	11	11	10	9	11	8	7	8	11
T_i	12	13	15	9	17	15	15	16	20	11	12	10	16	15	13

Par le test des signes, on regarde simplement le nombre de fois où $L_i < T_i$, ce qui arrive 14 fois ici. La commande `binom.test` donne $p\text{-val} \approx 4.88 \times 10^{-4}$ pour $\mathcal{H}_0 : \mathbb{P}(L < T) = \frac{1}{2}$ contre $\mathcal{H}_1 : \mathbb{P}(L < T) > \frac{1}{2}$, on admet que le lièvre court plus vite que la tortue. On dispose de plus d'informations si l'on tient compte des amplitudes $L_i - T_i$, qui devraient accentuer l'avantage des lièvres. Le test de Wilcoxon, effectué par la commande `wilcox.test` dans sa version appariée, confirme largement cette conclusion.

4 Analyse des données

Pour cette introduction à l'analyse des données (qui sera mise en pratique et largement complétée dans le module de *Datamining*), on se propose d'étudier en profondeur l'Analyse en Composantes Principales (ACP), probablement la méthode la plus célèbre et la plus utilisée pour traiter les tableaux de données quantitatives. Par la suite, on survolera l'Analyse Factorielle des Correspondances (AFC) pour comprendre comment extrapoler l'ACP à des tableaux de contingence résumant des données qualitatives.

4.1 Analyse en composantes principales

L'ACP est une méthode d'analyse **descriptive** multidimensionnelle très populaire, fréquemment rencontrée en analyse des données, de l'économie à la biologie en passant par le traitement de l'image et qui, comme

$$X = \begin{pmatrix} x_{11} & \dots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \dots & x_{np} \end{pmatrix} \in \mathbb{R}^{n \times p}.$$

- Décorrélér les données (les nouvelles variables construites sont orthogonales).
- Réduire la dimension de l'étude en considérant que certaines nouvelles variables ne jouent pas un rôle significatif.

[illegible]

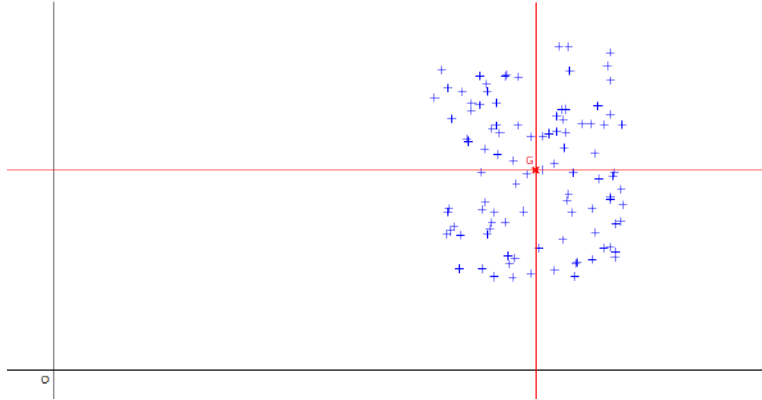
Dans la suite, on notera $X_i = (x_{i1}, \dots, x_{ip}) \in \mathbb{R}^p$ le i -ème individu et $V_k = (x_{1k}, \dots, x_{nk}) \in \mathbb{R}^n$ la k -ème variable. Le *nuage de points* des données est formé par les individus X_1, \dots, X_n , chacun représentant un point dans l'espace \mathbb{R}^p muni de la distance euclidienne. On rappelle que

1. http://www.sthda.com/english/wiki/wiki.php?id_contents=7851

Dans la version standard de l'ACP, la distance entre deux individus X_i et X_j est $d^2(X_i, X_j) = \|X_i - X_j\|^2$. Tout d'abord, il semble judicieux de chercher à placer l'origine du repère de \mathbb{R}^p au plus proche du nuage de points. Le *centre de gravité* est un choix naturel, il est donné par $G = (\bar{x}_1, \dots, \bar{x}_p)$ où

$$\bar{x}_k = \frac{1}{n} \sum_{i=1}^n x_{ik}$$

est la moyenne empirique des observations correspondant à la k -ème variable.



Par ailleurs, même si ce n'est pas systématique, il est usuel de *normaliser* les variables (cela caractérise l'ACP normée, sur laquelle on travaille ici). En effet, ces dernières ne s'expriment pas nécessairement dans la même unité de mesure et sans normalisation, on peut douter du sens à donner à la combinaison linéaire de deux grandeurs hétéroclites. Sans compter qu'un changement d'unités pourrait conduire à des résultats différents en sortie de l'ACP. Cela incite à considérer les données modifiées

$$X = \begin{pmatrix} \frac{x_{11}-\bar{x}_1}{s_1} & \dots & \frac{x_{1p}-\bar{x}_p}{s_p} \\ \vdots & \ddots & \vdots \\ \frac{x_{n1}-\bar{x}_1}{s_1} & \dots & \frac{x_{np}-\bar{x}_p}{s_p} \end{pmatrix} \quad \text{où} \quad s_k^2 = \frac{1}{n} \sum_{i=1}^n (x_{ik} - \bar{x}_k)^2$$

est la variance empirique des observations correspondant à la k -ème variable. La matrice appelée X ci-dessus est donc la matrice des observations centrées réduites. Les individus et les variables sont maintenant définis à l'aide de ces nouvelles observations, c'est-à-dire que pour toute la suite de l'étude,

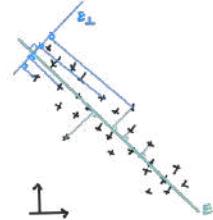
$$X_i = \begin{pmatrix} \frac{x_{i1}-\bar{x}_1}{s_1} \\ \vdots \\ \frac{x_{ip}-\bar{x}_p}{s_p} \end{pmatrix} \in \mathbb{R}^p \quad \text{et} \quad V_k = \begin{pmatrix} \frac{x_{1k}-\bar{x}_k}{s_k} \\ \vdots \\ \frac{x_{nk}-\bar{x}_k}{s_k} \end{pmatrix} \in \mathbb{R}^n$$

où l'on voit donc que X_i^T est la i -ème ligne de X et V_k la k -ème colonne de X . On remarque que la normalisation impose $\|V_k\|^2 = n$ et que le centrage impose $G = (0, \dots, 0)$.

4.1.2 Inertie et axes principaux

Par rapport au centre de gravité G qui est aussi l'origine du repère, l'*inertie* du nuage de points est donnée par :

$$I_G = \frac{1}{n} \sum_{i=1}^n d^2(X_i, G) = \frac{1}{n} \sum_{i=1}^n \|X_i\|^2 \quad \text{car} \quad G = (0, \dots, 0)$$



Cette quantité mesure la dispersion des individus autour du centre de gravité : plus l'inertie est faible, plus le nuage est concentré autour de G . Ce qui va également nous intéresser dans le cadre de l'ACP, c'est l'inertie définie par rapport à un sous-espace vectoriel E de \mathbb{R}^p passant par G , que l'on définit par :

$$I_E = \frac{1}{n} \sum_{i=1}^n d^2(X_i, P_E X_i) = \frac{1}{n} \sum_{i=1}^n \|X_i - P_E X_i\|^2 = \frac{1}{n} \sum_{i=1}^n \|(\mathbf{I} - P_E) X_i\|^2 = \frac{1}{n} \sum_{i=1}^n \|P_{E^\perp} X_i\|^2$$

P_E la projection orthogonale sur E

Par Pythagore, il est clair que $\|X_i\|^2 = \|P_E X_i\|^2 + \|P_{E^\perp} X_i\|^2$ de sorte que se dessine la relation fondamentale

$$I_G = I_E + I_{E^\perp}$$

de décomposition de l'inertie du nuage de points (à rapprocher du théorème de Huygens). Ainsi, en projetant les données sur E , on perd l'inertie I_E et il nous reste simplement I_{E^\perp} pour expliquer I_G . En conséquence, I_{E^\perp} est généralement qualifiée d'*inertie expliquée par E* et l'on a l'équivalence

$$\text{minimiser } I_E \iff \text{maximiser } I_{E^\perp}$$

que l'on va exploiter par la suite. Supposons maintenant que $E = \Delta_1$ est une droite passant par G . On peut se demander comment déterminer Δ_1 pour que I_{Δ_1} soit minimale, c'est-à-dire pour que l'on perde le moins d'inertie possible en résumant toute la complexité du nuage de points à l'aide d'une seule dimension. On va donc chercher à maximiser

$$\begin{aligned} I_{\Delta_1^\perp} &= \frac{1}{n} \sum_{i=1}^n \|P_{\Delta_1^\perp} X_i\|^2 \\ &= \frac{1}{n} \sum_{i=1}^n \langle X_i, u_1 \rangle^2 = \frac{1}{n} \sum_{i=1}^n u_1^T X_i X_i^T u_1 = u_1^T R u_1 \end{aligned}$$

où $u_1 \in \mathbb{R}^p$ est un vecteur directeur unitaire de Δ_1 , c'est-à-dire tel que $\|u_1\| = 1$, et où

$$R = \frac{1}{n} \sum_{i=1}^n X_i X_i^T = \frac{1}{n} \begin{pmatrix} \overset{\text{renormées}}{n} & \dots & r_{1p} \\ \vdots & \ddots & \vdots \\ r_{p1} & \dots & n \end{pmatrix} = \frac{X^T X}{n} \in \mathbb{R}^{p \times p}$$

est la matrice des corrélations empiriques $r_{kl} = \langle V_k, V_l \rangle$ entre les variables (car on a normalisé les données, sinon il s'agirait simplement des covariances empiriques). Mathématiquement parlant, on cherche donc le vecteur u_1 solution du système

$$\max_{u_1} u_1^T R u_1 \quad \text{sous la contrainte} \quad \|u_1\| = 1.$$

Proposition 4.1 Le problème ci-dessus est résolu lorsque u_1 est un vecteur propre unitaire associé à la plus grande valeur propre de R .

Preuve : Soit l'application $(u, \lambda) \mapsto u^T R u - \lambda(u^T u - 1)$

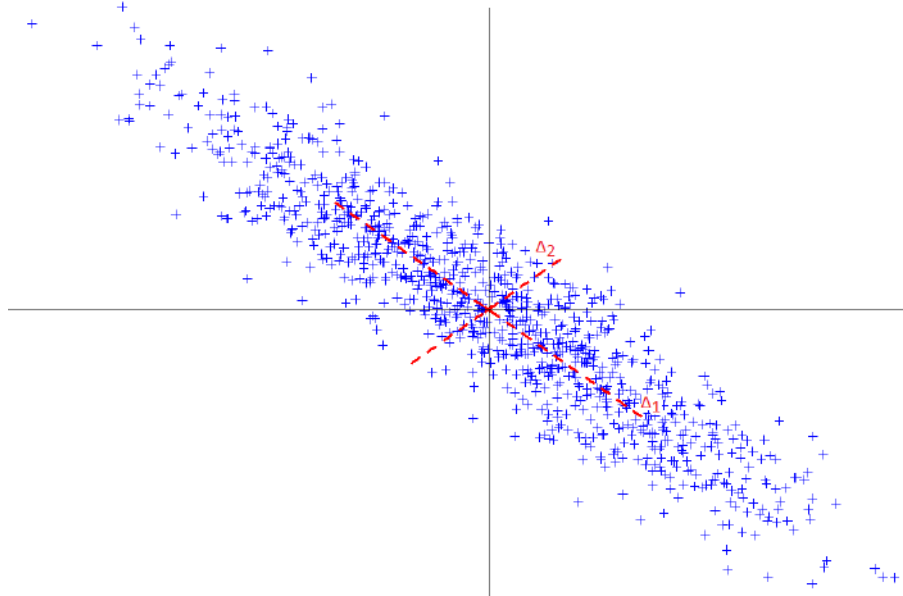
Son gradient s'annule si et seulement si $\|u\| = 1$ et $R u - \lambda u = 0 \iff R u = \lambda u$, d'où λ valeur propre de R associée au vecteur propre u .

On en tire aussi $u^T R u = \lambda u^T u = \lambda$, d'où le choix de la plus grande valeur propre de R .

Selon la même logique, on part à la recherche d'un second axe Δ_2 passant par G , orthogonal à Δ_1 et minimisant I_{Δ_2} . Cela se ramène à la résolution de

$$\max_{u_2} u_2^T R u_2 \quad \text{sous les contraintes} \quad \|u_2\| = 1 \quad \text{et} \quad \langle u_1, u_2 \rangle = 0.$$

Il s'ensuit que u_2 est un vecteur propre unitaire de R associé à la seconde plus grande valeur propre λ_2 et que $u_2^T R u_2 = \lambda_2$. En répétant cette stratégie, on détermine consécutivement les p axes principaux : on calcule le spectre de R en triant ses valeurs propres par ordre décroissant, et les vecteurs propres unitaires associés sont les vecteurs directeurs des axes principaux de l'ACP. Comme R est réelle, symétrique et (semi-)définie positive, toutes ses valeurs propres sont positives et tous ses vecteurs propres sont orthogonaux, ils forment donc à leur tour une base orthonormale de \mathbb{R}^p . Cela permet le passage des coordonnées initiales (dans la base canonique) aux nouvelles coordonnées (dans la base des vecteurs propres). On en déduit une particularité de l'ACP normalisée qui est que $I_G = p$. En effet, $I_G = I_{\Delta_1^\perp} + \dots + I_{\Delta_p^\perp} = \lambda_1 + \dots + \lambda_p = \text{Tr}(R)$, or R ne contient que des 1 dans la diagonale. Contrairement aux apparences dans le graphique ci-dessous, les axes principaux Δ_1 et Δ_2 sont bien orthogonaux (l'échelle n'est pas la même en abscisses et en ordonnées).



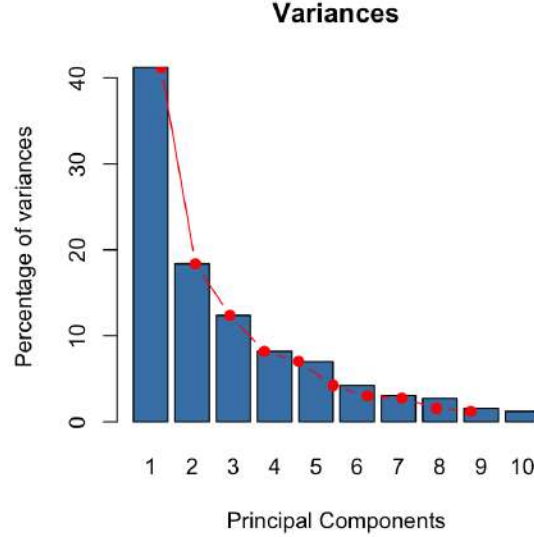
La *contribution* à l'inertie totale de l'axe Δ_k est $I_{\Delta_k^\perp} = \lambda_k$ et sa *contribution relative* vaut

$$\frac{I_{\Delta_k^\perp}}{I_G} = \frac{\lambda_k}{p}.$$

On définit de même la contribution jointe de plusieurs axes, par exemple

$$\frac{I_{\Delta_1^\perp} + \dots + I_{\Delta_k^\perp}}{I_G} = \frac{\lambda_1 + \dots + \lambda_k}{p}$$

est la contribution relative à l'inertie totale des k premiers axes de l'ACP. Cette mesure est fondamentale en réduction de dimension car sa représentation graphique nous permet de savoir à partir de quel axe I_G a été retrouvée de manière significative. À condition de réunir un bon pourcentage de l'inertie totale, le cas idéal consiste à se ramener à 2 axes. Dans l'exemple ci-dessous, avec 3 axes on récupère environ 75% de la variabilité initiale, on monte à plus de 80% avec un 4ème axe alors que les 2 seuls premiers axes ne permettent d'en récupérer que 60%. Il existe des critères empiriques pour décider de la dimension à retenir (méthode 'du coude', valeurs propres ≥ 1 pour l'ACP normée, etc.)



4.1.3 Individus

On sait maintenant représenter les individus dans la nouvelle base orthonormée $\mathcal{B} = \{u_1, \dots, u_p\}$, il suffit pour cela de calculer leurs nouvelles coordonnées. La k -ème coordonnée de X_i dans \mathcal{B} vaut donc $y_{ik} = \langle X_i, u_k \rangle$, ce qui se résume par

$$Y_i = U^T X_i \quad \text{où} \quad U = \begin{pmatrix} \vdots & & \vdots \\ u_1 & \dots & u_p \\ \vdots & & \vdots \end{pmatrix}$$

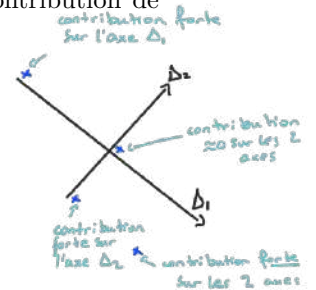
est la matrice des vecteurs propres (la matrice de passage). On adopte la convention suivante : X_i est le i -ème individu dans la base canonique, Y_i est ce même individu représenté dans \mathcal{B} . Comme les vecteurs propres, bien que normés, sont définis au signe près ($R u_k = \lambda_k u_k \Leftrightarrow R(-u_k) = \lambda_k(-u_k)$), on voit que l'orientation des axes principaux n'a aucune importance dans la représentation graphique. De même que l'on a caractérisé la contribution de l'axe Δ_k à l'inertie totale du nuage, on peut facilement caractériser la contribution de l'individu X_i à l'axe Δ_k à l'aide de la relation

$$I_{\Delta_k^\perp} = \frac{1}{n} \sum_{i=1}^n \|P_{\Delta_k} X_i\|^2.$$

On en tire naturellement que l'individu X_i apporte une contribution de

$$\frac{1}{n} \|P_{\Delta_k} X_i\|^2 = \frac{1}{n} d^2(P_{\Delta_k} X_i, G) = \frac{y_{ik}^2}{n}$$

contribution de l'individu i à l'axe Δ_k



à l'axe Δ_k . Ainsi, plus un individu sera projeté loin du centre de gravité sur un axe, et plus il contribuera à l'inertie portée par cet axe (en d'autres termes, plus il sera important dans la 'confection' de cet axe). En renormalisant par $I_{\Delta_k^\perp} = \lambda_k$, on en déduit sa contribution relative. En vertu des considérations précédentes et puisque

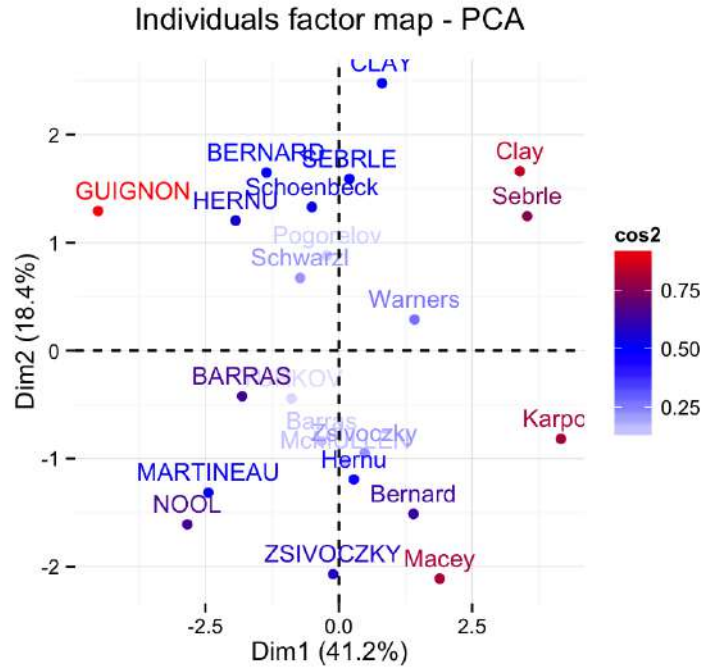
$$y_{ik}^2 = \langle X_i, u_k \rangle^2 = \|X_i\|^2 \cos^2(\alpha_{ik}) \quad \text{car } \|u_k\|^2 = 1$$



où α_{ik} est l'angle formé par Δ_k et X_i , on voit qu'un individu sera d'autant mieux décrit par l'axe Δ_k que $\cos^2(\alpha_{ik})$ sera proche de 1. D'où l'utilité de déterminer également pour chaque individu le cosinus carré de sa relation avec chaque axe. Typiquement, on a

$$\cos^2(\alpha_{ik}) = \frac{\langle X_i, u_k \rangle^2}{\|X_i\|^2} = \frac{u_k^T X_i X_i^T u_k}{\|X_i\|^2}$$

et pour étudier le cosinus carré de l'angle entre X_i et les k premiers axes (en fait l'hyperplan engendré par les k premiers axes), en raison de l'orthogonalité, on obtient directement $\cos^2(\alpha_{i1}) + \dots + \cos^2(\alpha_{ik})$. Si à l'issue de l'ACP on décide de ne retenir que ces k premiers axes, alors la somme précédente permettra de mettre en valeur la qualité de la représentation de l'individu X_i après la réduction de dimension. La projection des individus sur le plan formé par les deux axes principaux Δ_1 et Δ_2 peut révéler des comportements similaires et des regroupements, ou au contraire des antagonismes entre individus.



4.1.4 Variables

Pour l'instant nous n'avons exploré que les individus, il reste à réfléchir au traitement des variables. Lors du passage de la base canonique à \mathcal{B} , on a créé de nouveaux axes qui s'expriment comme des combinaisons linéaires des axes originaux ($y_{1k} = X_1^T u_k$, $y_{2k} = X_2^T u_k$, etc.). On remarque que

$$Z_k = \begin{pmatrix} y_{1k} \\ \vdots \\ y_{nk} \end{pmatrix} = X u_k = \begin{pmatrix} \vdots & \vdots \\ V_1 & \dots & V_p \\ \vdots & \vdots \end{pmatrix} u_k \in \mathbb{R}^n.$$

Les p nouvelles variables Z_1, \dots, Z_p sont les *composantes principales*, elles s'expriment comme des combinaisons linéaires des variables originales V_1, \dots, V_p . Plus généralement,

$$Z = XU$$

où l'on met dans $Z \in \mathbb{R}^{n \times p}$ les nouveaux individus en ligne et les composantes principales en colonnes. L'objectif de décorrélation des variables est atteint puisque :

$$\begin{aligned} \langle Z_k, Z_\ell \rangle &= \langle X u_k, X u_\ell \rangle = u_k^T X^T X u_\ell = n u_k^T \lambda_\ell u_\ell \quad \text{car} \quad \frac{X^T X}{n} u_\ell = \lambda_\ell u_\ell \\ &= n \lambda_\ell \langle u_k, u_\ell \rangle = \begin{cases} n \lambda_k & \text{si } k = \ell \\ 0 & \text{si } k \neq \ell \end{cases} \quad \text{car base orthonormale.} \end{aligned}$$

Alors que les composantes principales sont décorrélées comme on vient de le constater, il est intéressant de connaître le degré de corrélation qui lie ces nouvelles variables aux variables originales. Comme les variables initiales sont centrées, il est assez facile de voir que la covariance empirique entre la k -ème composante principale et la q -ème variable originale vaut

$$\begin{aligned}\frac{1}{n} \langle Z_k, V_q \rangle &= \frac{1}{n} u_k^T X^T X e_q \\ &= u_k^T R e_q = (R u_k)^T e_q = \lambda_k u_k^T e_q = \lambda_k u_{kq}\end{aligned}$$

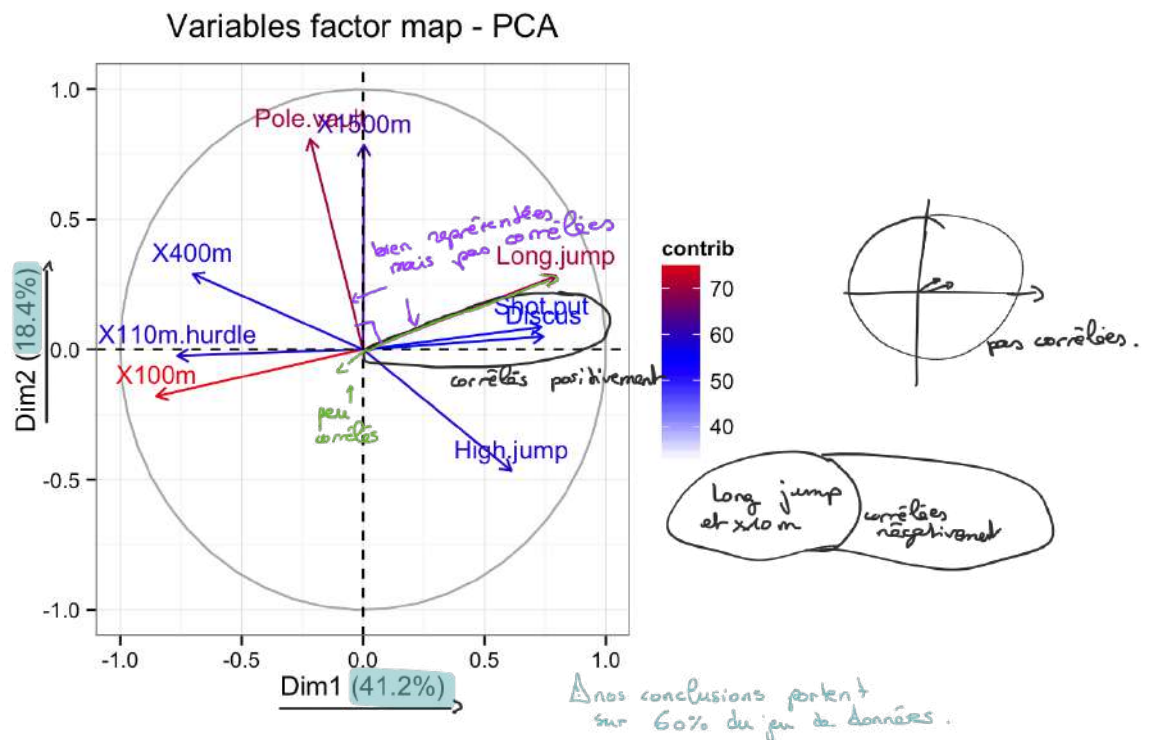
où $e_q = (0, \dots, 0, 1, 0, \dots, 0)$ est le q -ème vecteur de la base canonique de \mathbb{R}^p . On sait que $\frac{1}{n} \|V_q\|^2 = 1$ en raison de la normalisation et on a vu dans la section précédente que $\frac{1}{n} \|Z_k\|^2 = \lambda_k$. Combinant tout cela, la corrélation empirique entre Z_k et V_q vaut

$$c_{kq} = \sqrt{\lambda_k} u_{kq}$$

où l'on rappelle que u_{kq} est la q -ème coordonnée de u_k . Graphiquement, en positionnant dans le plan engendré par Δ_k et Δ_ℓ le point de coordonnées $(c_{kq}, c_{\ell q})$, on peut visualiser le degré de corrélation qui lie la q -ème variable originale aux k -ème et ℓ -ème composantes principales. Tous ces points ayant des corrélations comme coordonnées, ils sont situés dans un disque de rayon 1, d'où la dénomination de *cercle des corrélations*. Plus un point est proche du bord, et plus la variable d'origine projetée sur le plan est correctement expliquée par les composantes principales en question (en donnant un avantage à une composante par rapport à l'autre selon la proximité avec les axes). Les regroupements de variables dans le cercle formé par les deux axes principaux Δ_1 et Δ_2 révèlent en général beaucoup d'information sur la 'ressemblance' entre ces variables et sur l'interprétation à donner aux nouveaux axes. Une variable contribue d'autant à la 'confection' d'une composante principale qu'elle lui est corrélée. La *contribution relative* de la q -ème variable à la k -ème composante principale vaut

$$\frac{c_{kq}^2}{c_{k1}^2 + \dots + c_{kp}^2} = u_{kq}^2.$$

Ainsi pour une composante principale, on rapporte la contribution de la variable à la somme des contributions de toutes les variables. Elle est généralement représentée en pourcentage, comme dans l'exemple ci-dessous.



4.1.5 Individus et variables supplémentaires

On peut aussi vouloir faire apparaître sur les graphiques des individus ou des variables qui n'interviennent pas dans le calcul des composantes principales, on parle alors de données *supplémentaires* (voir par exemple le jeu de données décrit en première page). Ces dernières n'ont donc aucun impact sur les différents calculs de contribution aux axes. Représenter un individu supplémentaire est très facile, il suffit de calculer ses coordonnées dans la base \mathcal{B} . Les nouvelles coordonnées d'un individu X^* s'expriment par

$$Y^* = U^T X^*.$$

On pourra alors le superposer aux graphiques de représentation des individus de l'ACP pour en déduire son 'positionnement' par rapport au jeu de données. Il en va de même pour une variable supplémentaire, à condition qu'elle soit quantitative (et éventuellement normalisée). Ses coefficients de corrélation avec les composantes principales permettront de la situer dans les cercles de corrélation. Si cette dernière est qualitative, en général on se contentera d'utiliser des couleurs différentes pour chacune de ses modalités dans les représentations graphiques. On pourra ainsi détecter son influence éventuelle dans certains regroupements d'individus ou de variables.

4.2 Analyse factorielle des correspondances

L'ACP permet donc de visualiser les structures éventuelles cachées dans un tableau résumant des mesures quantitatives pour un ensemble d'individus. Qu'en est-il lorsque ces variables sont qualitatives ? Supposons maintenant que l'on croise deux variables qualitatives sous la forme d'un *tableau de contingence*. On considère une variable X à $r \geq 2$ modalités $\{x_1, \dots, x_r\}$ et une variable Y à $s \geq 2$ modalités $\{y_1, \dots, y_s\}$. On dispose de n_{ij} occurrences du couple (x_i, y_j) dans le jeu de données, soit

$$n = \sum_{i=1}^r \sum_{j=1}^s n_{ij}$$

observations en tout, ce que l'on résume dans le tableau de contingence ci-dessous (effectifs ou fréquences).

	y_1	\dots	y_s	Total		y_1	\dots	y_s	Total	
x_1	n_{11}	\dots	n_{1s}	$n_{1\bullet}$	\equiv	x_1	f_{11}	\dots	f_{1s}	$f_{1\bullet}$
\vdots	\vdots	\ddots	\vdots	\vdots		\vdots	\vdots	\ddots	\vdots	\vdots
x_r	n_{r1}	\dots	n_{rs}	$n_{r\bullet}$		x_r	f_{r1}	\dots	f_{rs}	$f_{r\bullet}$
Total	$n_{\bullet 1}$	\dots	$n_{\bullet s}$	n		Total	$f_{\bullet 1}$	\dots	$f_{\bullet s}$	1

avec $f_{ij} = \frac{n_{ij}}{n}$.

On sait tester l'existence d'une relation statistiquement significative entre X et Y par l'intermédiaire du test du khi-deux d'indépendance (voir Sec. 3.4.2). Mais lorsqu'un lien est détecté, le test ne donne aucune information supplémentaire sur les modalités de X fortement impactées par les modalités de Y (et réciproquement). On sait qu'il existe probablement un lien entre X et Y mais on ne sait pas le décrire. L'AFC va permettre d'apporter plus de précisions.

4.2.1 Profils

Définition 4.1 Les fréquences marginales des modalités de X et de Y sont respectivement définies, pour $1 \leq i \leq r$ et $1 \leq j \leq s$, par

$$f_{i\bullet} = \frac{1}{n} \sum_{j=1}^s n_{ij} = \frac{n_{i\bullet}}{n} \quad \text{et} \quad f_{\bullet j} = \frac{1}{n} \sum_{i=1}^r n_{ij} = \frac{n_{\bullet j}}{n}.$$

Le profil de la ligne i et le profil de la colonne j sont donnés par les vecteurs de fréquences conditionnelles

$$L_i = \left(\frac{f_{i1}}{f_{i\bullet}}, \dots, \frac{f_{is}}{f_{i\bullet}} \right) \in \mathbb{R}^s \quad \text{et} \quad C_j = \left(\frac{f_{1j}}{f_{\bullet j}}, \dots, \frac{f_{rj}}{f_{\bullet j}} \right) \in \mathbb{R}^r.$$

Si l'on appelle F la matrice des fréquences correspondant au tableau de contingence, alors on concentre les profils-ligne dans la matrice

$$L = D_r^{-1} F = \begin{pmatrix} \dots & L_1 & \dots \\ & \vdots & \\ \dots & L_r & \dots \end{pmatrix} \in \mathbb{R}^{r \times s} \quad \text{où} \quad D_r = \text{diag}(f_{1\bullet}, \dots, f_{r\bullet}).$$

On réunit de même les profils-colonne dans la matrice

$$C = D_s^{-1} F^T = \begin{pmatrix} \dots & C_1 & \dots \\ & \vdots & \\ \dots & C_s & \dots \end{pmatrix} \in \mathbb{R}^{s \times r} \quad \text{où} \quad D_s = \text{diag}(f_{\bullet 1}, \dots, f_{\bullet s}).$$

Proposition 4.2 *Le centre de gravité du nuage de points $\{L_1, \dots, L_r\}$ où L_i est muni du poids $f_{i\bullet}$ est donné par $(f_{\bullet 1}, \dots, f_{\bullet s})$. De même, le centre de gravité du nuage de points $\{C_1, \dots, C_s\}$ où C_j est muni du poids $f_{\bullet j}$ est donné par $(f_{1\bullet}, \dots, f_{r\bullet})$.*

Preuve :

$$\sum_{i=1}^r f_{i\bullet} L_i = \sum_{i=1}^r (f_{i1}, \dots, f_{is}) = (f_{\bullet 1}, \dots, f_{\bullet s})$$

$$\sum_{j=1}^s f_{\bullet j} C_j = \sum_{j=1}^s (f_{1j}, \dots, f_{rj}) = (f_{1\bullet}, \dots, f_{r\bullet})$$

4.2.2 Distance entre profils

Pour mesurer la 'ressemblance' entre deux profils-ligne de X ou entre deux profils-colonne de Y , on peut utiliser la distance euclidienne. Pour $1 \leq i, i' \leq r$ et $1 \leq j, j' \leq s$, on aurait

$$d^2(i, i') = \|L_i - L_{i'}\|^2 \quad \text{et} \quad d^2(j, j') = \|C_j - C_{j'}\|^2.$$

Dans une AFC, cette distance n'est pas recommandée car elle ne tient pas compte des croisements $X \times Y$: le terme $(L_{i1} - L_{i'1})^2$ et le terme $(L_{i2} - L_{i'2})^2$ jouent un rôle identique alors que $n_{\bullet 1}$ et $n_{\bullet 2}$ peuvent être très différents. Pour ramener les données à une échelle comparable, l'idée est d'affecter plus de poids aux fréquences sous-représentées (et inversement).

Définition 4.2 *La distance du khi-deux entre les profils-ligne i et i' est donnée par*

$$d_{\chi^2}^2(i, i') = \|L_i - L_{i'}\|_{D_s^{-1}}^2 = \sum_{j=1}^s \frac{1}{f_{\bullet j}} \left(\frac{f_{ij}}{f_{i\bullet}} - \frac{f_{i'j}}{f_{i'\bullet}} \right)^2.$$

De même, la distance du khi-deux entre les profils-colonne j et j' est donnée par

$$d_{\chi^2}^2(j, j') = \|C_j - C_{j'}\|_{D_r^{-1}}^2 = \sum_{i=1}^r \frac{1}{f_{i\bullet}} \left(\frac{f_{ij}}{f_{\bullet j}} - \frac{f_{ij'}}{f_{\bullet j'}} \right)^2.$$

Remarque. De manière générique, lorsque \mathbb{R}^q est muni de la métrique D_q (définie positive), on a

$$\forall x, x' \in \mathbb{R}^q, \quad \|x - x'\|_{D_q}^2 = (x - x')^T D_q (x - x').$$

Il s'agit de la distance euclidienne lorsque $D_q = I_q$.

4.2.3 Double ACP

L'AFC consiste alors à réaliser une double ACP : la première (directe) sur les profils-ligne et la seconde (duale) sur les profils-colonne. On rappelle que pour conduire une ACP sur une matrice X , on cherche en premier lieu à résoudre (voir Prop. 4.1)

$$\max_u u^T \frac{X^T X}{n} u \quad \text{sous la contrainte} \quad u^T u = 1.$$

→ **ACP sur L (analyse directe)**. Avec les métriques tenant compte des poids relatifs des lignes et des colonnes, on cherche

$$\max_{u \in \mathbb{R}^s} u^T ((L D_s^{-1})^T D_r (L D_s^{-1})) u \quad \text{sous la contrainte} \quad u^T D_s^{-1} u = 1.$$

Cela revient à dire que la distance du khi-deux est utilisée ($L D_s^{-1}$ au lieu de X) et que l'on tient compte d'une pondération spécifique aux lignes (D_r au lieu de $\text{diag}(\frac{1}{n}, \dots, \frac{1}{n})$) : plus une modalité i est représentée, plus le profil-ligne i prend de l'importance.

→ **ACP sur C (analyse duale)**. Avec les métriques tenant compte des poids relatifs des lignes et des colonnes, on cherche

$$\max_{v \in \mathbb{R}^r} v^T ((C D_r^{-1})^T D_s (C D_r^{-1})) v \quad \text{sous la contrainte} \quad v^T D_r^{-1} v = 1.$$

Cela revient à dire que la distance du khi-deux est utilisée ($C D_r^{-1}$ au lieu de X) et que l'on tient compte d'une pondération spécifique aux colonnes (D_s au lieu de $\text{diag}(\frac{1}{n}, \dots, \frac{1}{n})$) : plus une modalité j est représentée, plus le profil-colonne j prend de l'importance.

Proposition 4.3 Les problèmes ci-dessus sont résolus lorsque u est un vecteur propre de

$$A = L^T D_r L D_s^{-1}$$

associé à sa plus grande valeur propre, et lorsque v est un vecteur propre de

$$B = C^T D_s C D_r^{-1}$$

associé à sa plus grande valeur propre.

Preuve : $(u, \lambda) \mapsto u^T ((L D_s^{-1})^T D_r (L D_s^{-1})) u - \lambda (u^T D_s^{-1} u - 1)$

Le gradient s'annule ssi $u^T D_s^{-1} u = 1$ et $(L D_s^{-1})^T D_r (L D_s^{-1}) u = \lambda D_s^{-1} u$

$\Leftrightarrow L^T D_r L D_s^{-1} u = \lambda u$

(cf preuve proposition 4-1).

En itérant le processus, on construit les nouvelles bases $\mathcal{B}_L = \{u_1, \dots, u_s\}$ et $\mathcal{B}_C = \{v_1, \dots, v_r\}$. Il reste à placer les profils-ligne et les profils-colonne dans \mathcal{B}_L et dans \mathcal{B}_C , respectivement. Les coordonnées des r profils-ligne dans \mathcal{B}_L sont données (en lignes) par

$$L^* = L D_s^{-1} U \in \mathbb{R}^{r \times s} \quad \text{où} \quad U = \begin{pmatrix} \vdots & & \vdots \\ u_1 & \dots & u_s \\ \vdots & & \vdots \end{pmatrix}$$

alors que les coordonnées des s profils-colonne dans \mathcal{B}_C sont données (en lignes) par

$$C^* = C D_r^{-1} V \in \mathbb{R}^{s \times r} \quad \text{où} \quad V = \begin{pmatrix} \vdots & & \vdots \\ v_1 & \dots & v_r \\ \vdots & & \vdots \end{pmatrix}.$$

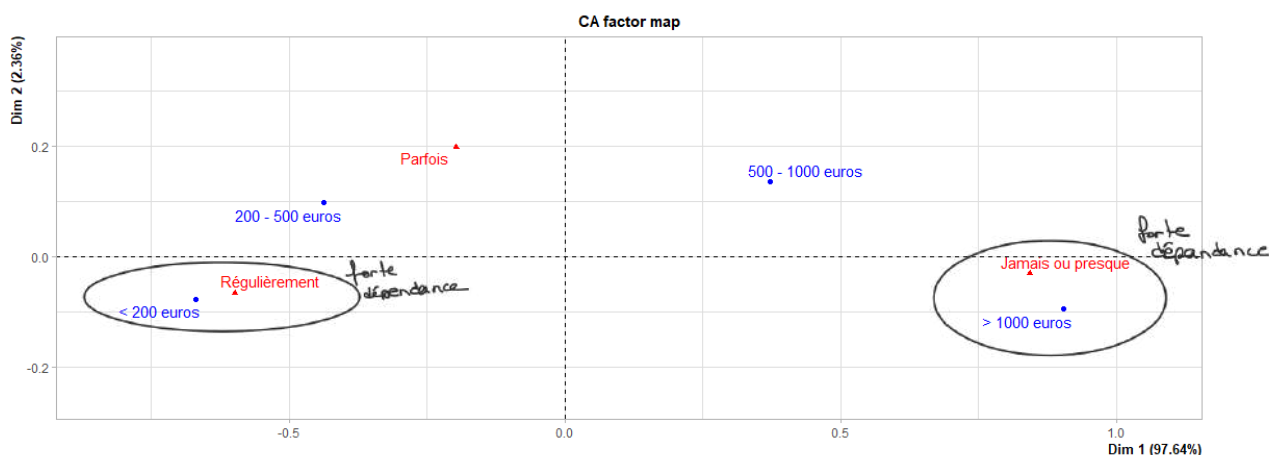
4.2.4 Superposition des plans factoriels

Nous ne le ferons pas dans le cadre de ce module, mais on peut montrer qu'il fait sens de superposer les projections des profils-ligne et des profil-colonne dans un même plan (après les opérations usuelles de recentrage et de renormalisation). De plus, on montre que les valeurs propres non nulles de A et de B coïncident, ce qui permet d'évaluer l'inertie relative portée par chaque axe dans les représentations graphiques superposées, et donc d'évaluer la qualité de la projection. La commande **CA** du package **FactoMineR** permet d'effectuer une AFC sur un tableau de contingence, et des outils supplémentaires de diagnostic sont accessibles dans le package **factoextra**. Pour un premier aperçu de la mise en pratique de l'AFC, prenons un exemple simple (factice). Dans une certaine population, on répartit les $n = 487$ individus (des ordinateurs) selon leur prix (X) et leur fréquence de pannes (Y). On obtient :

	Jamais ou presque	Parfois	Régulièrement	Total
< 200 €	9	35	121	165
200 – 500 €	13	24	53	90
500 – 1000 €	58	26	28	112
> 1000 €	95	13	12	120
Total	175	98	214	487

p-valeur :
 $P(\chi^2_6 > 209.78) < 2 \cdot 10^{-16}$
 $H_0 = X \perp Y$
 → On peut rejeter H_0 avec certitude.

Sur ce tableau de contingence, la statistique du khi-deux d'indépendance vaut $D^2 \approx 209.78$ ce qui, comparé au quantile $z_{0.95}(3 \times 2) \approx 12.6$, ne laisse planer aucun doute sur l'existence d'une dépendance entre X et Y . L'AFC donne la projection suivante sur le premier plan factoriel.



Le pourcentage de variabilité porté par le premier axe est prépondérant. Les modalités proches sont plus fréquentes que ce à quoi on se serait attendu sous l'hypothèse d'indépendance. On lit sur ce graphique la forte corrélation entre “Régulièrement” et “< 200 €” et celle de “Jamais ou presque” avec “> 1000 €”. On tire des conclusions similaires mais moins marquées pour “200 – 500 €” et “500 – 1000 €”. Tout comme pour l'ACP, il est possible d'inclure à cette étude des variables supplémentaires interprétables sur les graphiques (projections ou codes couleurs). Des extensions, comme l'Analyse des Correspondances Multiples (ACM), et une mise en pratique plus poussée seront proposées dans le module de *Datamining*.

$$\mathbb{E}[X] = \sum_{i=1}^{\infty} x_i p_i,$$

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f(x) dx,$$

D'après le théorème de transfert, elle est alors égale à

$$\mathbb{E}(\varphi(X)) = \int_F \varphi(x) d\mathbb{P}_X(x).$$

$$\mathbb{E}(X|Y)(y) \equiv \mathbb{E}(X|Y=y) \equiv \sum_x x \cdot \mathbb{P}(X=x|Y=y).$$

Formulaire – Loïs de probabilité usuelles

Loïs discrètes.

→ *Loi de Dirac.* $\mathbb{P}(X=c) = 1$ pour $c \in \mathbb{R}$.

→ *Loi uniforme.* $\forall k \in \{x_1, \dots, x_n\}, \mathbb{P}(X=k) = \frac{1}{n}$. $\mathbb{E} = \frac{x_1 + x_n}{2}$ $\text{Var} = \frac{n^2 - 1}{12}$

→ *Loi de Bernoulli.* $\forall k \in \{0, 1\}, \mathbb{P}(X=k) = p^k (1-p)^{1-k}$ pour $0 \leq p \leq 1$. $\rightarrow \mathbb{E} = p, \text{Var} = p(1-p)$

→ *Loi de Rademacher.* $\forall k \in \{-1, 1\}, \mathbb{P}(X=k) = p^{\frac{1+k}{2}} (1-p)^{\frac{1-k}{2}}$ pour $0 \leq p \leq 1$. $\mathbb{E}[X] = 2p-1$ et $\text{Var}(X) = 4p(1-p)$.
 $\mathbb{E} = 0$ $\text{Var} = 1$

→ *Loi binomiale.* $\forall k \in \{0, \dots, n\}, \mathbb{P}(X=k) = \binom{n}{k} p^k (1-p)^{n-k}$ pour $0 \leq p \leq 1$. $\mathbb{E} = np, \text{Var} = np(1-p)$

→ *Loi géométrique.* $\forall k \in \mathbb{N}^*, \mathbb{P}(X=k) = p(1-p)^{k-1}$ pour $0 < p \leq 1$. $\mathbb{E} = 1/p$ $\text{Var} = q/p^2$

→ *Loi de Poisson.* $\forall k \in \mathbb{N}, \mathbb{P}(X=k) = e^{-\lambda} \frac{\lambda^k}{k!}$ pour $\lambda > 0$. $\mathbb{E} = \lambda = \text{Var}$

→ *Loi de Pascal.* $\forall k \in \{n, n+1, \dots\}, \mathbb{P}(X=k) = \binom{k-1}{n-1} p^n (1-p)^{k-n}$ pour $0 < p \leq 1$. $\mathbb{E} = \frac{n}{p}$ $\text{Var} = \frac{nq}{p^2}$

Loïs continues.

→ *Loi uniforme.* $\forall x \in \mathbb{R}, f_X(x) = \frac{1}{b-a} \mathbb{1}_{\{a \leq x \leq b\}}$ pour $a < b$. $\mathbb{E} = \frac{a+b}{2}$, $\text{Var} = \frac{(b-a)^2}{12}$

→ *Loi exponentielle.* $\forall x \in \mathbb{R}, f_X(x) = \lambda e^{-\lambda x} \mathbb{1}_{\{x \geq 0\}}$ pour $\lambda > 0$. $\mathbb{E} = 1/\lambda$ $\text{Var} = 1/\lambda^2$

→ *Loi de Laplace.* $\forall x \in \mathbb{R}, f_X(x) = \frac{1}{2b} e^{-\frac{|x-\mu|}{b}}$ pour $\mu \in \mathbb{R}$ et $b > 0$. $\mathbb{E} = \mu$ $\text{Var} = 2b^2$

→ *Loi normale (gaussienne).* $\forall x \in \mathbb{R}, f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ pour $\mu \in \mathbb{R}$ et $\sigma^2 > 0$. $\mathbb{E} = \mu$ $\text{Var} = \sigma^2$

→ *Loi gamma.* $\forall x \in \mathbb{R}, f_X(x) = \frac{\lambda^k x^{k-1} e^{-\lambda x}}{\Gamma(k)} \mathbb{1}_{\{x > 0\}}$ pour $\lambda > 0, k > 0$ et $\Gamma(k) = \int_0^{+\infty} s^{k-1} e^{-s} ds$. $\mathbb{E} = \frac{k}{\lambda}$; $\text{Var} = \frac{k}{\lambda^2}$

→ *Loi bêta.* $\forall x \in \mathbb{R}, f_X(x) = \frac{x^{a-1} (1-x)^{b-1}}{\beta(a,b)} \mathbb{1}_{\{0 \leq x \leq 1\}}$ pour $a > 0, b > 0$ et $\beta(a,b) = \int_0^1 s^{a-1} (1-s)^{b-1} ds$.

$$\mathbb{E} = \frac{a}{a+b} \quad \text{Var} = \frac{ab}{(a+b)^2 (a+b+1)}$$

Loïs continues construites sur la loi normale.

→ *Loi de Cauchy.* $\frac{N_1}{N_2} \sim \mathcal{C}$, où $N_1 \sim \mathcal{N}(0,1)$, $N_2 \sim \mathcal{N}(0,1)$ et $N_1 \perp N_2$.

→ *Loi du khi-deux.* $N_1^2 + \dots + N_n^2 \sim \chi^2(n)$, où $N_1, \dots, N_n \stackrel{\text{iid}}{\sim} \mathcal{N}(0,1)$.

→ *Loi de Student.* $\frac{N}{\sqrt{Z_n/n}} \sim t(n)$, où $N \sim \mathcal{N}(0,1)$, $Z_n \sim \chi^2(n)$ et $N \perp Z_n$.

→ *Loi de Fisher.* $\frac{Z_{n_1}/n_1}{Z_{n_2}/n_2} \sim F(n_1, n_2)$, où $Z_{n_1} \sim \chi^2(n_1)$, $Z_{n_2} \sim \chi^2(n_2)$ et $Z_{n_1} \perp Z_{n_2}$.

Loïs de sommes de variables i.i.d.

→ *Bernoulli/Binomiale.* $X_1 + \dots + X_n \sim \mathcal{B}(n, p)$ si $X \sim \mathcal{B}(p)$.

→ *Poisson/Poisson.* $X_1 + \dots + X_n \sim \mathcal{P}(n\lambda)$ si $X \sim \mathcal{P}(\lambda)$.

→ *Géométrique/Pascal.* $X_1 + \dots + X_n \sim \mathcal{P}(n, p)$ si $X \sim \mathcal{G}(p)$.

→ *Normale/Normale.* $X_1 + \dots + X_n \sim \mathcal{N}(n\mu, n\sigma^2)$ si $X \sim \mathcal{N}(\mu, \sigma^2)$.

→ *Exponentielle/Gamma.* $X_1 + \dots + X_n \sim \Gamma(n, \lambda)$ si $X \sim \mathcal{E}(\lambda)$.

→ *Khi-deux/Khi-deux.* $X_1 + \dots + X_n \sim \chi^2(nd)$ si $X \sim \chi^2(d)$.

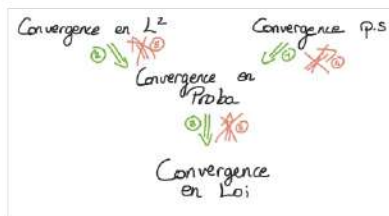
Formulaire – Espérance et variance/covariance

Variables aléatoires. Les X (ou X_i) sont des variables aléatoires réelles, les c (ou c_i) désignent des constantes réelles.

- $\mathbb{E}[c] = c.$
- $\mathbb{E}[c_1 X_1 + c_2 X_2] = c_1 \mathbb{E}[X_1] + c_2 \mathbb{E}[X_2].$
- $X \geq 0 \Rightarrow \mathbb{E}[X] \geq 0.$
- $\mathbb{V}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2.$
- $\mathbb{V}(X) \geq 0.$
- $\mathbb{V}(X) = 0 \Leftrightarrow X$ est constante.
- $\mathbb{V}(c_1 X + c_2) = c_1^2 \mathbb{V}(X).$
- $\text{Cov}(X_1, X_2) = \mathbb{E}[(X_1 - \mathbb{E}[X_1])(X_2 - \mathbb{E}[X_2])] = \mathbb{E}[X_1 X_2] - \mathbb{E}[X_1] \mathbb{E}[X_2] = \text{Cov}(X_2, X_1).$
- $\text{Cov}(X, c) = 0.$
- $\text{Cov}(X, X) = \mathbb{V}(X).$
- $\mathbb{V}(c_1 X_1 + c_2 X_2) = c_1^2 \mathbb{V}(X_1) + c_2^2 \mathbb{V}(X_2) + 2 c_1 c_2 \text{Cov}(X_1, X_2).$
- $\text{Cov}(c_1 X_1 + c_2 X_2, c_3 X_3) = c_1 c_3 \text{Cov}(X_1, X_3) + c_2 c_3 \text{Cov}(X_2, X_3).$

Compléments pour les vecteurs aléatoires. Les X (ou X_i) sont des vecteurs colonnes aléatoires réels de dimensions quelconques, les C (ou C_i) désignent des constantes matricielles dont les dimensions sont supposées adaptées aux opérations dans lesquelles elles interviennent.

- $\mathbb{E}[C_1 X C_2] = C_1 \mathbb{E}[X] C_2.$
- $\mathbb{E}[X^T] = \mathbb{E}[X]^T.$
- $\mathbb{V}(X) = \mathbb{E}[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^T] = \mathbb{E}[X X^T] - \mathbb{E}[X] \mathbb{E}[X]^T = \mathbb{V}(X)^T.$
- $\mathbb{V}(X) = \mathbb{V}(X^T)$ si X est un vecteur ligne.
- $\mathbb{V}(CX) = C \mathbb{V}(X) C^T.$
- $\text{Cov}(X_1, X_2) = \mathbb{E}[(X_1 - \mathbb{E}[X_1])(X_2 - \mathbb{E}[X_2])^T] = \mathbb{E}[X_1 X_2^T] - \mathbb{E}[X_1] \mathbb{E}[X_2]^T = \text{Cov}(X_2, X_1)^T.$
- $\mathbb{V}(C_1 X_1 + C_2 X_2) = C_1 \mathbb{V}(X_1) C_1^T + C_2 \mathbb{V}(X_2) C_2^T + C_1 \text{Cov}(X_1, X_2) C_2^T + C_2 \text{Cov}(X_2, X_1) C_1^T.$
- $\text{Cov}(C_1 X_1 + C_2 X_2, C_3 X_3) = C_1 \text{Cov}(X_1, X_3) C_3^T + C_2 \text{Cov}(X_2, X_3) C_3^T.$



$$\forall t \in \mathbb{R}, \quad F_X(t) = \int_{-\infty}^t f_X(s) ds.$$

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