$L2$ Math-Éco, Probabilités-Statistique 2 - 2021 – 2022

Thibault Espinasse

Ce résumé a vocation à être un aide mémoire, à compléter, et éventuellement à corriger. Si vous trouvez des coquilles, merci de me contacter pour les corriger. Il s'agit évidemment d'une version provisoire. Quelques références en suppléments

- Le livre "Statistique mathématique, cours et exercices corrigés" de C. Vial et B. Cadre
- Le polycopié de M. Delarue de Statistique
- Le livre "Statistiques asyptotiques" de M. Van der Vaart [1]
- Le site internet Wikistat très riche : `http://wikistat.fr/`
- ...

En outre, pour celles et ceux que ça intéresse, voici quelques chaînes youtube de vulgarisation qui peuvent donner un éclairage supplémentaire sur le cours.

1. Science4all
2. ScienceEtonnante
3. Mickaël Launay
4. Hygiène Mentale
5. El Jj
6. Data Gueule
7. 3Blue1Brown
8. La statistique expliquée à mon chat
9. Le chat sceptique
10. François Husson
11. ...

Contents

1 Préambule

1.1 Notations et rappels

Lors d'une définition, il est possible de rencontrer le symbole := qui signifie seulement "égal par définition à".

Dans tout ce polycopié, nous utiliserons abondamment la notation $\mathbb{1}$ pour désigner la fonction indicatrice :

$$\mathbb{1}_{x\in A} := \begin{cases} 0 \text{ si } x \notin A \\ 1 \text{ si} x \in A \end{cases}$$

Un ensemble E est dit au plus dénombrable si l'on peut numéroter ses éléments, c'est à dire s'il existe une **injection** de E dans $\mathbb{N}$. Les ensembles finis, $\mathbb{N}, \mathbb{Z}$ et $\mathbb{Q}$ par exemple sont "au plus dénombrables", mais $\mathbb{R}$ n'est pas dénombrable.

2 Introduction et rappels de statistique

→ Plusieurs échelles : classification dans le tableau suivant

→ Omniprésence dans les champs d'application : Médecine, Energie, Fiabilité, Météo, Insee, Agronomie, Marketing...

→ Idée qu'on peut répondre à une question concernant une population globale, en observant seulement un sous échantillon.

La statistique						
Recueil de données *Choix des indices, collecte...*	⇒	Prétraitement *Quantités synthétiques, représentation...*	⇒	Analyse de données *ACP, traitement des valeurs extrêmes...* ⇓		Statistiques descriptives
				Modélisation		
Décisions *Éventuelle obtention de nouvelle données...*	⇐	Conclusions *Prédiction...*	⇐	⇓ Inférence *Esimation, tests...*		Statistique mathématique

L'objectif de la statistique est **d'extraire de l'information** de données. Que ce soit pour comprendre un caractère intrinsèque *(ex : estimation d'un paramètre ayant un sens physique)*, où pour prédire un phénomène, les résultats obtenus devraient toujours être **réplicables**, au sens où, en faisant la même expérience dans la même situation, on s'attend à obtenir les mêmes conclusions. Finalement, la grande majorité des cas peut être apparentée à l'expérience suivante :

- On collecte des données sélectivement *(ex : plan d'expérience)*, aléatoirement *(ex : sondage)*, ou sous contraintes *(ex : données par département)*
- De cette information partielle, on souhaite extraire une information globale *(ex : caractère chez population globale, temps de retour, séparation du "bruit" et du "signal"...)*
- On cherche à contrôler la pertinence du modèle, ou des conclusions *(ex : cross validation, test de normalité...)*

2.1 Définitions

- Population : Ensemble étudié (ex : La population de la région Rhône-Alpes)
- Individu : Un élément de la population (ex : Vous)
- Échantillon : Un sous-ensemble de la population (ex : 500 individus intérrogés)
- Caractère, variable : Une caractéristique mesurable chez chaque individu (ex : Niveau de revenu caractère quantitatif, pour ou contre... caractère qualitatif)

Remarque : À ce stade là, on a supposé que ce que l'on souhaitait quantifier se mesurait facilement (ex : température, taille, âge...) mais certains mots ne sont pas clairs à définir mathématiquement (chômage, réussite...). Attention donc aux choix des indices !

Le début du travail de modélisation est un travail de formalisme. La toute première étape est de **nommer les objets considérés**, c'est à dire de trouver des notations mathématiques pour désigner les objets d'interêt. Le premier réflexe est donc de donner un nom à chaque variable. Ensuite, il est nécessaire **d'observer** les données. Cela passe par des synthèses des variables en calculant des quantités empiriques, mais aussi par des représentations graphiques bien choisies. On appelle cela la **statistique descriptive**, qui fait aussi l'objet de la section suivante.

2.2 Les quantités empiriques

Nous allons voir en cours de Probabilités de nombreuses notions (espérance, variance, covariance, fonction de répartition...) Dans cette partie, nous allons considérer les pendants empiriques de toutes ces quantités.

Définition 2.1 *On définit les indices suivants :*

- *La moyenne empirique d'un échantillon $(x_i)_{i=1,\cdots,N}$ est définie par*

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i.$$

- *La variance empirique d'un échantillon $(x_i)_{i=1,\cdots,N}$ est définie par*

$$\bar{s}_X^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \bar{x}^2.$$

- *La covariance empirique entre x et y pour un échantillon $(x_i, y_i)_{i=1,\cdots,N}$ est définie par*

$$\bar{s}_{X,Y} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{N}\sum_{i=1}^{N} x_i y_i - \bar{x}\bar{y}.$$

- *La fonction de répartition empirique d'un échantillon $(x_i)_{i=1,\cdots,N}$ est définie par*

$$\bar{F}(t) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}_{x_i \leq t}.$$

- *Le quantile empirique d'ordre $p \in [0,1]$ est défini par*

$$\bar{q}_p \geq t \Leftrightarrow \bar{F}(t) \leq p.$$

  Notons qu'il se peut que plusieurs nombres vérifient cela, dans ce cas là, plusieurs conventions existent pour lever l'ambiguité.

  Dans le cadre de ce cours, nous utiliserons la convention suivante, dite de la moyenne pondérée :

$$\bar{q}_p = x_{j+1} + r(x_{j+2} - x_{j+1}),$$

  où j est la partie entière de $(N-1)p$ et r est la partie fractionnelle. Cette méthode est utilisée par défaut par plusieurs languages informatiques.

Parfois, on note toutes ces quantités avec un N en indice, pour spécifier le nombre d'observations

2.3 Représentation graphique

Avant de chercher à inférer des données, il est important d'avoir une idée de leur répartition. Pour cela, nous pouvons utiliser les indices définis ci-dessus pour avoir une première idée de la variabilité, de la concentration des données, et pouvoir discerner des valeurs suspectes (valeurs aberrantes ?), et effectuer éventuellement un prétraitement aux données (lissage, changement d'échelle, traitement des données manquantes...).

Pour cela, il existe plusieurs moyens de visualiser les données

- Tableau
- Diagramme en bâtons
- Fonction de répartition
- Diagramme-boîte ("boîte à moustaches"/*boxplot*)
- Histogramme
- Diagramme circulaire ("camembert")
- Nuages de points
- ACP(hors programme en $L2$)
- Q-Q plot

2.4 Premiers réflexes

La liste suivante n'est pas exhaustive, mais voilà quelques-unes des premières questions à se poser[1] :

1. Quel est le contexte ? Comment les données ont-elle été recueillies ? L'hypothèse d'échantillonage aléatoire est-elle revendiquée, si oui crédible ? Penser à lister les biais et incertitudes éventuel-e-s dues à la collecte de données.
2. Que signifient les différentes variables, les différents individus ?
3. Les questions posées/hypothèses à valider sont-elles claires ? *(Attention, l'approche sera différente en exploration de données, et lors d'une expérience statistique ayant pour but de confirmer une hypothèse).*
4. Y a-t-il des données manquantes ? Des erreurs de saisies ? Les données ont-elle été saisies uniformément *(ponctuation, mots clés...)* ?
5. En cas de données manquantes ou abhérantes, comment résout-on le problème *(suppression des lignes/colonnes incriminées, remplacement pas une valeur bien choisie...)* ?
6. Les variables sont-elles qualitatives, continues, ordonnées... ?
7. Quelles représentations pertinentes des données peut-on effectuer ? Comment peut on résumer l'information qui nous intéresse ?
8. Les données extrêmes sont-elles plausibles ? Observer l'étendue des données, calculer les différents indices.
9. La distribution des données est-elle plausible ?
10. Doit-on, ou souhaite-t-on modifier les données *(lissage, passage au log, discrétisation...)* avant de les traiter ?
11. Quel lien y-a-t il entre les variables ? *(penser aux régressions/ACP/Anova... et à observer les résidus !)*
12. Quels modèles peut-on mettre sur les données ? *(Loi de probabilité, modèles de regression...).*
13. Peut-on vérifier partiellement ces hypothèses (Q-Q plots, tests...)

[1] On pourra s'entraîner en récupérant des données ouvertes, par exemple des données historiques de température sur le site `https://www.ecad.eu//dailydata/predefinedseries.php`

2.5 Vers l'aléa : Pourquoi faire usage de la théorie des probabilités en statistique

Question :

- Pourquoi précise-t-on "choisi-e-s au hasard", et qu'est ce que cela signifie ?
- Quel impact aurait la suppression de ce critère de "sondage" ?
- Pour avoir une information sur toute la population, est-il nécessaire d'interroger tout le monde ? Pourquoi ?

→ Importance de *l'hypothèse d'aléa* dans le traitement statistique : vers l'échantillonnage et la correction des biais.

→ Rôle de la théorie des probabilités dans l'inférence statistique.

Se pose maintenant la question essentielle : comment peut-on affirmer quoique ce soit sur une population entière sans l'avoir mesurée intégralement.

Pour les modèles aléatoires, nous utiliserons la théorie des probabilités pour modéliser l'incertitude[2].

La prochaine étape de modélisation est la formulation d'une hypothèse fondamentale utilisée en statistique mathématique :

Hypothèse fondamentale : Les observations $(x_i)_{i=1,\cdots,N}$ sont des réalisations de variables aléatoires $X_1, \cdots, X_N$ (la plupart du temps supposées i.i.d.) de loi inconnue.

Exemples :

- On peut supposer que les x_i sont des réalisations i.i.d. de loi normale $\mathcal{N}(m, \sigma^2)$ où l'un ou les deux paramètres sont inconnus
- Dans un cadre de sondage (au sens large) on suppose que l'on est capable de "tirer au hasard" (échantillonner aléatoirement) dans la population. La loi de X (pour un tirage avec remise) est alors la loi empirique sur l'échantillon global :

  $$\forall i \in [1, n], \mathbb{P}(X = y_i) = \frac{1}{n},$$

  Cela revient à poser $X = y_U$, où $U \sim \mathcal{U}([\![1, n]\!])$ Cela n'empêche pas de mettre un modèle sur cette loi là, inconnue.
- Lorsque la famille de loi inconnues peut se décrire en fonction d'un nombre fini de **paramètres** réels, on parle de **famille paramétrique de lois**, et on s'intéresse aux statistiques paramétriques (cadre de ce cours)
- Il se peut aussi qu'on ne puisse pas paramétrer la famille avec un nombre fini de nombres réels (par exemple si on prend l'ensemble des densités de probabilité continues). Ces modèles sont l'objet des statistiques non-paramétriques, et dépasse le programme de ce cours.

L'étape de modélisation consiste donc à définir un cadre d'étude, et des hypothèses, qui permettront de répondre aux questions posées. Le type de réponse apportées prend souvent la forme *"Avec une certitude de* 95%*, on peut affirmer que ..."*. Cela signifie que il y a seulement 5% de chances que l'affirmation soit fausse, et due seulement à la variabilité des observations. Tout cela est relatif à la validité des hypothèses faites ici, rarement parfaites...

Avant d'introduire les premières méthodes d'estimation, il faut revoir quelques notions élémentaires de la théorie des probabilités.

[2] en notant que l'aléa permet de prendre en compte notre ignorance des facteurs déterminants)

3 Introduction et rappels de probabilités

3.1 Définitions générales

Définition 3.1 (Espace de probabilité) *On appelle espace de probabilité un triplet* $(\Omega, \mathcal{A}, \mathbb{P})$ *où*

1. Ω *est un ensemble quelconque*
2. $\mathcal{A}$ *un sous-ensemble de* $\mathcal{P}(\Omega)$ *qui vérifie*[3]
   - $\Omega \in \mathcal{A}$.
   - $\mathcal{A}$ *est stable par complémentarisation* ($A \in \mathcal{A} \Rightarrow A^C \in \mathcal{A}$).
   - $\mathcal{A}$ *est stable par union dénombrable* ($\forall n, A_n \in \mathcal{A} \Rightarrow \cup_n A_n \in \mathcal{A}$).
3. $\mathbb{P}$ *est une application* $\mathcal{A} \to \mathbb{R}^+$ *qui vérifie*
   - $\mathbb{P}(\emptyset) = 0$
   - $\mathbb{P}(\mathcal{A}) = 1$
   - *Pour toute famille* $(A_n)_{n\in\mathbb{N}} \in \mathcal{A}$ *disjointe,* $\mathbb{P}(\cup_n A_n) = \sum_n \mathbb{P}(A_n)$.

Remarque : Au passage, on remarque que si deux événements $A, B \in \mathcal{A}$ sont **incompatibles**, i.e. $A \cap B = \emptyset$, alors

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B).$$

Pour des événements exprimés en français, le $\cup$ *correspond au "ou" et le* $\cap$ *au "et". Cela signifie que lorsque deux événements ne peuvent advenir en même temps, alors la probabilité qu'un ou l'autre advienne est la somme des probabilités de ces deux événements.*

$*$ *Faire le lien avec votre intuition de* $\Omega, \mathcal{A}, \mathbb{P}$*, et les définitions de "Univers", "évènement", "evènement élémentaire"...*

Question : Et si on observe un seul dès dans un jet de plusieurs, ou la somme...

$\to$ Vers la notion de variables aléatoires

Définition 3.2 *Une variable aléatoire* X *sur un espace de probabilité* $(\Omega, \mathcal{A}, \mathbb{P})$ *et à valeur dans un espace*[4] $(E, \mathcal{B})$ *est une fonction*[5] *de* $(\Omega, \mathcal{A})$ *dans* $(E, \mathcal{B})$.

La loi de la variable X *est donnée par la "mesure image"* $\mathbb{P}_X$ *définie par*

$$\forall B \in \mathcal{B}, \mathbb{P}_X(B) := \mathbb{P}(X^{-1}(B)) =_{notation} \mathbb{P}(X \in B).$$

Remarque : En pratique, on ommet l'espace de probabilité la plupart du temps. En effet, des variables aléatoires de même loi peuvent être définies sur des espaces de probabilité différents. Par exemple, pour une variable de Bernoulli de paramètre $\frac{1}{2}$, on peut prendre

- $\Omega = \{0, 1\}, \mathcal{A} = \mathcal{P}(\Omega), \forall A \in \mathcal{A}, \mathbb{P}(A) = \frac{\sharp A}{2}, \forall \omega \in \Omega, X(\omega) = \omega$.
- $\Omega = [\![1, 6]\!], \mathcal{A} = \mathcal{P}(\Omega), \forall A \in \mathcal{A}, \mathbb{P}(A) = \frac{\sharp A}{6}, \forall \omega \in \Omega, X(\omega) = \mathbb{1}_{\omega \leq 3}$.
- $\Omega = [0, 1], \mathcal{A} = \mathcal{B}(\Omega), \forall A \in \mathcal{A}, \mathbb{P}(A) = \lambda(A), \forall \omega \in \Omega, X(\omega) = \mathbb{1}_{\omega \leq \frac{1}{2}}$.

[3] On appelle $(\Omega, \mathcal{A})$ un "espace mesurable", vu en L3

[4] mesurable

[5] que l'on exigera "mesurable", en L3, i.e. $\forall B \in \mathcal{B}, X^{-1}(B) \in \mathcal{A}$.

On notera $X \sim Y$ pour dire que X et Y ont même loi.

Remarque : On peut aisément définir une suite de variables aléatoires en considérant comme espace de probabilité (par exemple pour une suite de Bernoulli) :

- $\Omega = \{0,1\}^{\mathbb{N}}$.
- $\mathcal{A}$ est la tribu produit[6].
- $\mathbb{P}$ est déterminée par les valeurs $\mathbb{P}(\omega_1, \cdots, \omega_n, \{0,1\}^{\mathbb{N}}) = p^{\sum \omega_i}(1-p)^{\sum 1-\omega_i}$.

En général, la définition 3.2 signifie

$$\mathbb{P}(X \in A) = \int_{t \in A} \mathbb{P}_X(dt).$$

Définition 3.3 *On appelle la fonction de répartition de X la fonction*

$$t \mapsto F_X(t) := \mathbb{P}(X \leq t).$$

Cette fonction caractérise la loi de X !

Lorsque l'ensemble des valeurs que peut prendre X est dénombrable, on parle de variable aléatoire discrète[7]. L'équation précédente devient

$$\mathbb{P}(X \in A) = \sum_{k \in A} \mathbb{P}(X = k).$$

Lorsqu'on peut écrire

$$\forall A \in \mathcal{A}, \mathbb{P}(X \in A) = \int_A f(x)dx,$$

on parle de variable à densité. On dit alors que X a pour densité f[8].

Propriétés 3.4 *Si X admet pour densité f_X, par construction, F_X est la primitive de f_X qui s'annule en $-\infty$. On a donc F_X qui est une fonction dérivable, et on a*

$$F_X(t)' = f_X(t).$$

Proposition 3.5 *Si $U \sim \mathcal{U}([0,1])$, et si F_X est inversible*[9]*, alors $F_X^{-1}(U) \sim X$.*

Remarque : On peut dans ces cas là construire un espace de probabilité un peu "canonique" pour X, en prenant $\Omega = [0,1]$ muni de la mesure de Lebesgue, et $X(\omega) = F_X^{-1}(\omega)$ comme fonction mesurable pour construire explicitement X comme v.a.

Quelques exemples (lois classiques):

- Lois discrètes :

[6] cf le cours de théorie de la mesure, ceci n'est pas à connaître pour ce cours

[7] Dans ce cas, $\mathbb{P}_X$ est une somme pondérée de mesures de Dirac.

[8] Cela revient en fait à dire que $\mathbb{P}_X$ est absolument continue par rapport à la mesure de Lebesgue, f est la densité ou dérivée de Radon-Nikodym.

[9] Cela fonctionne aussi si F_X n'est pas inversible, en utilisant la pseudo-inverse, mais il faut faire un petit effort de rédaction, cela sera vu en TD

– Bernoulli $\mathcal{B}(p)$

Cette loi permet de modéliser une variable binaire (Pile ou Face, présence ou absence de maladie, caractère fumeur ou non...). Une variable X de Bernoulli de paramètre p est à valeur dans $\{0,1\}$, et vérifie

$$\forall k \in \{0,1\}, \mathbb{P}(X=k) = p^k(1-p)^{1-k}, i.e. \mathbb{P}(X=1) = p \quad \mathbb{P}(X=0) = 1-p$$

On parle souvent "suite d'épreuves de Bernoulli" pour désigner une suite de variables aléatoires de Bernoulli, de "succès" pour $X=1$, et d'échec pour $X=0$. **Remarque** : Pour tout événement A, la variable aléatoire $X = \mathbb{1}_A$ (i.e. $X(\omega) = \mathbb{1}_{\omega \in A}$) ne prend que les valeurs 0 et 1 par définition de l'indicatrice. C'est donc une variable aléatoire de loi de Bernoulli de paramètre $\mathbb{P}(A)$. Cette propriété est à retenir.

– Binomiale $\mathcal{B}in(N,p)$

Cette loi permet de modéliser le nombre de succès à une suite de N épreuves de Bernoulli indépendantes. Une variable X Binomiale de paramètre N,p est à valeur dans $\{0,\cdots,N\}$, et vérifie

$$\forall k \in \{0,\cdots,N\}, \mathbb{P}(X=k) = \binom{N}{k} p^k (1-p)^{N-k}.$$

Exercice : Calculer la limite de $\mathbb{P}(X_n = k)$ quand $X_n \sim \mathcal{B}in(N, \frac{p}{n})$.

– Poisson $\mathcal{P}(\lambda)$

Cette loi apparaît naturellement comme limite de la loi Binomiale, quand le nombre d'épreuves tends vers l'infini, à espérance constante. Une variable aléatoire X de loi de Poisson de paramètre λ permet de modéliser le nombre d'évenements ponctuels arrivant durant une certaine durée, ou dans une certaine zone spatiale, lorsqu'il y a stationnarité et indépendance des arrivées. Elle est à valeurs dans $\mathbb{N}$ et vérifie

$$\forall k \in \mathbb{N}, \mathbb{P}(X=k) = \frac{\lambda^k}{k!} e^{-\lambda}.$$

– Géométrique $\mathcal{G}(p)$

Une variable aléatoire X Géométrique de paramètre p représente l'instant du premier succès à une suite d'épreuves de Bernoulli indépendantes de paramètre p. Elle est à valeurs dans $\mathbb{N}^\star$[10], et vérifie

$$\forall k \in \mathbb{N}^\star, \mathbb{P}(X=k) = p(1-p)^{k-1}.$$

- Lois continues :

– Exponentielle $\mathcal{E}(\lambda)$

Une variable aléatoire X Exponentielle de paramètre λ représente l'instant inter-arrivée d'évenements ponctuels arrivant, lorsqu'il y a stationnarité et indépendance des arrivées.. Elle est à valeurs dans $\mathbb{R}^+$, et a pour densité la fonction définie par

$$f(x) = \lambda e^{-\lambda x} \mathbb{1}_{x \in \mathbb{R}^+}.$$

[10] Certains peuvent la définir dans $\mathbb{N}$, en considérant $X-1$

– Normale $\mathcal{N}(m, \sigma^2)$
Une variable aléatoire X Normale (ou gaussienne) de paramètres m et σ peut modéliser de nombreuses variables continues "unimodales". Cette loi apparaît en particulier dans le Théorème Central Limite. Elle est à valeurs dans $\mathbb{R}$, et a pour densité la fonction définie par

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

– Uniforme $\mathcal{U}([a, b])$
Une variable aléatoire X Uniforme sur $[a, b]$ est à valeurs dans $[a, b]$, et a pour densité la fonction définie par

$$f(x) = \frac{1}{b-a} \mathbb{1}_{x\in[a,b]}.$$

Ces lois seront étudiées en TD, en particulier leur calcul d'espérance et de variance.

Dans toute la suite, on considère une v.a. X à valeurs dans $\mathbb{R}$ définie sur un espace de probabilité $(\Omega, \mathcal{A}, \mathbb{P})$.

Définition 3.6 *Lorsque $\int_\Omega |X(\omega)|\mathbb{P}(d\omega) < +\infty$[11], on peut définir l'espérance de X par*

$$\mathbb{E}[X] := \int_\Omega X(\omega)\mathbb{P}(d\omega) = \int_{\mathbb{R}} t\mathbb{P}_X(dt).$$

En particulier, lorsque X est discrète, à valeurs dans E, cela devient

$$\mathbb{E}[X] := \sum_E k\mathbb{P}(X = k).$$

Et si X est continue, de densité f, on a

$$\mathbb{E}[X] := \int_{\mathbb{R}} tf(t)dt.$$

Propriétés 3.7 *L'espérance mathématique vérifie les propriétés suivantes :*

- $\forall a \in \mathbb{R}, \mathbb{E}[a] = a.$
- *Linéarité :* $\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y].$
- *Croissance : Si $X \geq Y$ p.s.,* $\mathbb{E}[X] \geq \mathbb{E}[Y].$
- $\forall A \in \mathcal{A}, \mathbb{E}[\mathbb{1}_A] = \mathbb{P}(A)$

Propriétés 3.8 *Si $X \geq 0$, p.s., on a*

$$\mathbb{E}[X] = \int_{\mathbb{R}^+} (1 - F_X(t))\mathrm{d}t = \int_{\mathbb{R}^+} \mathbb{P}(X > t)\mathrm{d}t.$$

Notons que nous avons introduit la notation "p.s.". Cela signifie "presque-sûrement". Dire qu'une propriété $\mathcal{P}$ est vraie p.s. signifie

$$\mathbb{P}\Big(\{\omega \in \Omega, \mathcal{P}(\omega)\}\Big) = 1.$$[12].

Remarque : On peut appliquer la définition précédente à une nouvelle variable aléatoire définie par $Y = g(X)$, dès lors que g est mesurable (i.e. dès lors que Y est bien une v.a., cf définition). Cela permet de faire la remarque suivante, citée comme proposition tant elle est utile.

[11] on note $X \in \mathrm{L}^1$

[12] Il se peut qu'il existe toutefois des $\omega \in \Omega$ pour lesquels la propriété est fausse, mais ils sont de probabilité nulle

Proposition 3.9 *Soit g une fonction* [13] *telle que $\int_{\mathbb{R}} |g(t)|\mathbb{P}_X(dt) < +\infty$, alors*

$$\mathbb{E}[g(X)] := \int_\Omega g(X(\omega))\mathbb{P}(d\omega) = \int_{\mathbb{R}} g(t)\mathbb{P}_X(dt).$$

En particulier, lorsque X est discrète, à valeurs dans E, cela devient

$$\mathbb{E}[g(X)] := \sum_E g(k)\mathbb{P}(X = k).$$

Et si X est continue, de densité f, on a

$$\mathbb{E}[X] := \int_{\mathbb{R}} g(t)f(t)dt.$$

Remarque : La définition de l'espérance, et la proposition précédente contient en réalité la même propriété, parfois appelé la "formule du transfert". Pour illustrer cette propriété, considérons par exemple $\Omega = \{1, 2, 3, 4, 5, 6\}$, avec $\mathbb{P}(\{\omega\}) = 1/6, \forall\omega \in \Omega$, et $X(\omega) = x_1 \mathbb{1}_{\omega\in\{2,3,6\}} + x_2\mathbb{1}_{\omega\in\{1,4,5\}}$ (avec $x_1 \neq x_2$)

On a alors, selon la définition

$$\mathbb{E}[g(X)] = \sum_{\omega\in\Omega} g(X(\omega))\mathbb{P}(\{\omega\}) = g\Big(x_1\mathbb{1}_{1\in\{2,3,6\}}+x_2\mathbb{1}_{1\in\{1,4,5\}}\Big)\mathbb{P}(\{1\})+\cdots+g\Big(x_1\mathbb{1}_{6\in\{2,3,6\}}+x_2\mathbb{1}_{6\in\{1,4,5\}}\Big)\mathbb{P}(\{6\})$$

ce qui donne

$$\mathbb{E}[g(X)] = \sum_{\omega\in\Omega} g(X(\omega))\mathbb{P}(\{\omega\}) = g(x_2)\mathbb{P}(\{1\})+g(x_1)\mathbb{P}(\{2\})+g(x_1)\mathbb{P}(\{3\})+g(x_2)\mathbb{P}(\{4\})+g(x_2)\mathbb{P}(\{5\})g(x_1)\mathbb{P}(\{6\}).$$

Mais on peut aussi regrouper par valeurs que prend X :

$$\mathbb{E}[g(X)] = g(x_1)\mathbb{P}\Big(\{2, 3, 6\}\Big) + g(x_2)\mathbb{P}\Big(\{1, 4, 5\}\Big) = g(x_1)\mathbb{P}(X = x_1) + g(x_2)\mathbb{P}(X = x_2).$$

Cela illustre cette propriété de "transfert" qui correspond en quelque sorte à "sommer les termes en les regrouppant".

Définition 3.10 *Le moment d'ordre r de X est définit lorsque $\int_{\mathbb{R}} |t^r|\mathbb{P}_X(dt) < +\infty$, et est donné par*

$$m_r = \mathbb{E}[X^r] := \int_\Omega X(\omega)^r\mathbb{P}(d\omega) = \int_{\mathbb{R}} t^r\mathbb{P}_X(dt).$$

En particulier, lorsque X est discrète, à valeurs dans E, cela devient

$$m_r := \sum_E k^r\mathbb{P}(X = k).$$

Et si X est continue, de densité f, on a

$$m_r := \int_{\mathbb{R}} t^r f(t)dt.$$

Si le moment d'ordre r de X est fini, on dira que $X \in \mathbb{L}^r$ [14]. Remarquons tout de suite que

[13] mesurable

[14] On devrait préciser $\mathbb{L}^r(\Omega, \mathcal{A}, \mathbb{P})$, mais on ommetra ces précisions dès qu'il n'y a aucune ambiguité.

Proposition 3.11 *Si $r_1 \leq r_2$, alors*

$$X \in \mathbb{L}^{r_2} \Rightarrow X \in \mathbb{L}^{r_1}$$

Définition 3.12 *Si $X \in \mathbb{L}^2$, on peut définir la* ***variance*** *de X par*

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

Propriétés 3.13 *La variance vérifie :*

- $\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$
- $\forall a \in \mathbb{R}, \mathrm{Var}(a) = 0$
- $\forall a, b \in \mathbb{R}, \mathrm{Var}(aX + b) = a^2 \,\mathrm{Var}(X)$

Définition 3.14 *On appelle* ***écart-type*** *de la v.a.r. X la racine carré de la variance :*

$$\sigma_X = \sqrt{\mathrm{Var}(X)}$$

Remarque : Si X est une v.a.r telle que $\mathbb{E}[X] = m$ et $\mathrm{Var}(X) = \sigma^2$, alors la variable $Y = \frac{X-m}{\sigma}$ est d'espérance nulle, et de variance 1. On dit que Y est centrée (d'espérance nulle) et réduite (de variance 1).

Théorème 3.15 *Inégalité de Markov : Si $X \geq 0$ et $a > 0$, alors*

$$\mathbb{P}(X \geq a) \leq \frac{1}{a}\mathbb{E}[X].$$

Inégalité de Bienaymé-Tchébychev : Si $X \in \mathbb{L}^2(\Omega)$.

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq a) \leq \frac{1}{a^2}\,\mathrm{Var}(X).$$

Inégalité de Cramèr[15] *:*

$$\mathbb{P}\Big(X > x\Big) \leq e^{-\psi_X^\star(x)},$$

où on note, pour $t > 0$, $\psi_X(t) = \log\big(\mathbb{E}[e^{tY}]\big)$, la log-Laplace de X, et $\psi_X^\star$ sa tranformée de Cramèr (ou de Fenchel-Legendre) définie par $\psi_X^\star(x) = \sup_{t>0}\{tx - \psi_Y(t)\}$.

3.2 Indépendance et conditionnement

Définition 3.16 *Deux événements $A, B \in \mathcal{A}$ sont dits indépendants si*

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

De même, on dit que des événements $A_1, \cdots A_n$ sont indépendants dans leur ensemble, si pour tout $j_1, \cdots, j_k$ on a

$$\mathbb{P}(A_{j_1} \cap \cdots \cap A_{j_k}) = \mathbb{P}(A_{j_1}) \cdots \mathbb{P}(A_{j_k}).$$

Deux événements sont donc indépendants s'ils n'ont pas plus de chances d'advenir ensemble ou séparément. Une façon d'intuiter ce résultat est de définir la probabilité conditionnelle.

[15] On parle plutôt de méthode de Cramèr dans la litterature

Définition 3.17 *Soient $A, B \in \mathcal{A}$ deux événements, tel que $\mathbb{P}(B) > 0$. La probabilité conditionnelle de A sachant B est définie par*

$$\mathbb{P}(A|B) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$$

Deux événements A et B sont donc indépendants lorsque $\mathbb{P}(A|B) = \mathbb{P}(A)$. Autrement dit, B n'apporte aucune information sur A.

Définition 3.18 *Un vecteur aléatoire (ou variable aléatoire multidimensionnelle) est une variable aléatoire à valeur dans $\mathbb{R}^n$. Une suite de variables aléatoires est une variables aléatoire à valeur dans $\mathbb{R}^{\mathbb{N}}$. Remarquons qu'il s'agit seulement de fonction de Ω dans $\mathbb{R}^n$ ou $\mathbb{R}^N$.*

Définition 3.19 *On dit que deux variables aléatoires X et Y définies sur le même espace de probabilité et à valeurs dans des ensembles respectifs $(E_1, \mathcal{B}_1)$ et $(E_2, \mathcal{B}_2)$ (i.e. (X, Y) est un vecteur aléatoire à valeur dans $(E_1 \times E_2, \mathcal{B}_1 \times \mathcal{B}_2)$) sont indépendantes si*

$$\forall B_1 \in \mathcal{B}_1, B_2 \in \mathcal{B}_2, \mathbb{P}(X \in B_1, Y \in B_2) = \mathbb{P}(X \in B_1)\mathbb{P}(Y \in B_2).^{16}$$

Cette définition s'étend naturellement au cas de n variables aléatoires.
Remarque : Si X et Y sont indépendantes, on a alors

$$\mathbb{P}_{X,Y}(x, y) = \mathbb{P}_X(x)\mathbb{P}_Y(y).$$

Si elles sont à densité, alors les densités se factorisent de la même manière

$$f_{X,Y}(x, y) = f_X(x)f_Y(y).$$

De plus, pour toutes fonction f, g mesurables telles que $\mathbb{E}[|f(X)|] < +\infty$ et $\mathbb{E}[|g(Y)|] < +\infty$, on a

$$\mathbb{E}[f(X)g(Y)] = \mathbb{E}[f(X)]\mathbb{E}[g(Y)].$$

En particulier, on a, pour des v.a. réelles

$$\forall x \in \mathbb{R}, y \in \mathbb{R}, F_{X,Y}(x, y) := \mathbb{P}(X \leq x, Y \leq y) = F_X(x)F_Y(y).$$

Ces trois propriétés caractérisent l'indépendance.

Définition 3.20 *Soit X, Y définies sur le même espace de probabilité. On suppose que $X, Y \in \mathbb{L}^2$. On peut définir la covariance entre X et Y par*

$$\mathrm{Cov}(X, Y) := \mathbb{E}\Big[\big(X - \mathbb{E}[X]\big)\big(Y - \mathbb{E}[Y]\big)\Big].$$

À comparer avec la définition de la variance...

Proposition 3.21 *Si X, Y sont indépendantes, alors*

$$\mathrm{Cov}(X, Y) = 0.$$

La réciproque n'est pas vraie !!!!

Propriétés 3.22 *Soient X, Y deux variables aléatoires*[17]

[16] Formellement, on note $\mathbb{P}(X \in B_1, Y \in B_2)$ pour $\mathbb{P}(\{X \in B_1\} \cap \{Y \in B_2\})$

[17] définies sur le même espace de probabilité, on risque de ne plus le préciser, mais c'est sous-entendu

1. $\mathrm{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$
2. *La fonction* $(X,Y) \mapsto \mathrm{Cov}(X,Y)$ *est une forme bilinéaire symmétrique :*
   - $\forall a_1, a_2 \in \mathbb{R}, \mathrm{Cov}(a_1X_1 + a_2X_2, Y) = a_1\,\mathrm{Cov}(X_1,Y) + a_2\,\mathrm{Cov}(X_2,Y)$
   - $\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$
3. *Conséquence :* $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X,Y)$.
4. *Lorsque* X, Y *sont indépendantes, l'égalité précédente devient :* $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$

Proposition 3.23 (Stabilité des lois normales) *Si* X *et* Y *sont deux variables aléatoires* ***indépendantes*** *de loi normales, alors toute combinaison affine de* X, Y *est aussi de loi normale :*

$$\forall a, b, c \in \mathbb{R}, aX + bY + c \text{ est de loi normale.}$$

Attention, cela ne suffit pas à caractériser la loi, puisqu'il reste à calculer espérance et variance. En particulier, en posant $b = 0$*, on voit que toute transformation affine d'une loi normale est aussi de loi normale*[18]

3.3 Convergence de v.a.

3.3.1 Différentes notions de convergence

Soit $(X_n)_{n\in\mathbb{N}}$ une suite de v.a. réelles et X une v.a. réelle[19] (définies sur le même espace de probabilité).

On va pouvoir citer plusieurs types de convergence :

Définition 3.24 *On dit que* $(X_n)_{n\in\mathbb{N}}$ *converge p.s. vers* X *si*

$$\mathbb{P}\Big(\big\{\omega \in \Omega, \lim_{n\to\infty} X_n(\omega) = X(\omega)\big\}\Big) = 1.$$

Définition 3.25 *On dit que* $(X_n)_{n\in\mathbb{N}}$ *converge vers* X *en probabilités (* $X_n \xrightarrow{\mathbb{P}} X$*) si*

$$\forall \epsilon > 0, \lim_{n\to\infty} \mathbb{P}(|X_n - X| > \epsilon) = 0.$$

Définition 3.26 *On dit que* $(X_n)_{n\in\mathbb{N}}$ *converge vers* X *en loi (*$X_n \xrightarrow{\mathcal{L}} X$*) si*

$$\forall f \text{ continue bornée}, \lim_{n\to\infty} \mathbb{E}[f(X_n)] = \mathbb{E}[f(X)].$$

Remarque : Il y a ici un abus de language à parler de convergence vers X, alors que seule la loi de X est importante !

Proposition 3.27 (Admis) *La convergence en loi peut-être caractérisée des façons suivantes :*

- $X_n \xrightarrow{\mathcal{L}} X$ *ssi* $F_{X_n} \to F_X$ *en tous points de continuité de la fonction de répartition* F_X *de* X*.*

Voilà en vrac quelques liens entre toutes ces convergences :

[18] en réalité cette propriété cache des propriétés beaucoup plus générale, de ce qu'on appelle les "vecteurs gaussiens" définis comme des familles de loi normales (non nécessairement indépendntes) telle que toute combinaison affine reste de loi normale. Mais c'est en dehors du programme de ce cours.

[19] En réalité, ça marche pour $\mathbb{R}^d$ aussi.

Proposition 3.28

- *Si* $X_n \xrightarrow{p.s.} X$*, alors* $X_n \xrightarrow{\mathbb{P}} X$*.*
- *Si* $X_n \xrightarrow{\mathbb{P}} X$*, il existe une sous suite* $(n_k)_{k\in\mathbb{N}}$ *telle que* $X_{n_k} \xrightarrow{p.s.} X$*.*
- *Si* $X_n \xrightarrow{\mathbb{P}} X$*, alors* $X_n \xrightarrow{\mathcal{L}} X$*.*
- *Si* $X_n \xrightarrow{\mathcal{L}} X$*, alors il existe un espace de probabilité sur lequel on peut définir* X'_n *de même loi que* X_n*,* X' *de même loi que* X*, telle que* $X'_n \xrightarrow{p.s.} X'$ *(Théorème de représentation de Skorokhod)*
- *Si* $X_n \xrightarrow{\mathbb{P}} X$*, et* $Y_n \xrightarrow{\mathbb{P}} Y$*, alors* $(X_n, Y_n) \xrightarrow{\mathbb{P}} (X, Y)$*.*
- *Si* $X_n \xrightarrow{\mathcal{L}} X$*, et s'il existe une constante* $C \in \mathbb{R}$*, telle que* $Y_n \xrightarrow{\mathbb{P}} C$*, alors le couple* (X_n, Y_n) *converge en loi vers le couple* (X, c) *(Théorème de Slutsky).*

3.3.2 Théorèmes limites

On peut maintenant citer

Théorème 3.29 (Loi forte des grands nombres) *Soit* $(X_n)_{n\in\mathbb{N}}$ *une suite i.i.d. de v.a, telles que* $\mathbb{E}[|X_1|] < +\infty$ *(i.e.* $X_1 \in \mathbb{L}^1$*). Alors*

$$\bar{X} := \frac{1}{n}\sum_{k=1}^{n} X_k \xrightarrow{p.s.} \mathbb{E}[X_1].$$

Ainsi que le

Théorème 3.30 (Théorème Central Limite) *Soit* $(X_n)_{n\in\mathbb{N}}$ *une suite i.i.d. de v.a de* $\mathbb{L}^2$*, telles que* $\mathbb{E}[X_1] = m$ *et* $\mathrm{Var}(X_1) = \sigma^2$*. On note* $\bar{X}_n = \frac{1}{n}\sum_{k=1}^{n} X_k$*. Alors*

$$\sqrt{n}\frac{\bar{X}_n - m}{\sigma} = \frac{\sum_{k=1}^{n} X_k - nm}{\sqrt{n}\sigma} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$$

4 Introduction à l'estimation et principe d'un intervalle de confiance

Exemple introductif :

On veut estimer la proportion d'étudiant-e-s/travailleu-r-ses à l'Université Lyon 1. On intérroge 100 personnes "au hasard". Parmi elles 40 affirment qu'elles travaillent à côté. Que peut-on dire ?

On doit faire des hypothèses

→ La notion "travailler" a été définie numériquement et les étudiant-e-s ont la même

→ Les étudiant-e-s n'ont pas menti

→ On a été capable de tirer les étudiant-e-s uniformément et indépendamment (par exemple)

Dans ce cadre, on peut formuler le modèle mathématique correspondant :

- L'ensemble les réponses des étudiants peut s'écrire $x_1, \cdots, x_N \in \{0,1\}$
- On note $p = \frac{1}{N}\sum_{i=1}^{N} x_i$
- On dispose d'une suite $U_1, \cdots, U_n$ de variable aléatoires indépendantes, uniformes sur $[\![0, N]\!]$
- Nos observations s'écrivent donc $(Y_i)_{i=1,\cdots,n}$ où $Y_i = x_{U_i}$ et vérifient $Y_i \sim \mathcal{B}er(p)$ et sont indépendantes

Dans ce cadre, on verra qu'on peut affirmer mathématiquement des propriétés de l'estimateur "naturel" que l'on voudrait construire. Intuitivement, on voudrait définir $\hat{P} = \frac{1}{n}\sum_{i=1}^{n} Y_i$, l'estimateur de p. On s'attend à ce que $\hat{P}$ soit "proche" de p.

Remarquons déjà que $\hat{P}$ est aléatoire, et que si nous intérrogeons de nouveau 100 personnes, alors nous aurons probablement une autre réponse. Si ici il y a 40 réponses positives, c'est le résultat d'une expérience aléatoire, et notre estimation sera de $\hat{p} = \frac{40}{100}$. Plus précisemment, réaliser une expérience revient à choisir un $\omega \in \Omega$ et on a $\hat{p} = \hat{P}(\omega) = \frac{1}{n}\sum_{i=1}^{n} Y_i(\omega) = \frac{1}{n}\sum_{i=1}^{n} y_i$, si $y_1, \cdots y_n$ désigne une réalisation de notre échantillon.

L'objet de cette partie est de formaliser cette intuition, d'obtenir des méthodes différentes pour construire des estimateurs, et d'en étudier les propriétés.

4.1 Statistique et Estimateur

Dans toute la suite, on va se placer dans une famille paramétrique de loi. Dans ce cours, on considèrera qu'une famille paramétrique de lois s'écrit de la forme $\mathcal{F} = \{p_\theta, \theta \in \Theta\}$ (où p_θ désigne une suite de probabilités) dans le cas discret et $\mathcal{F} = \{f_\theta, \theta \in \Theta\}$ (où f_θ désigne une densité) dans le cas de v.a. à densité[20].

Soit $X_1, \cdots, X_n$ un échantillon i.i.d. de loi inconnue dépendant d'un paramètre $\theta_0 \in \mathbb{R}^d$ que l'on souhaite estimer. On a donc $X_i \sim_{i.i.d.} f_{\theta_0}$, où $f_{\theta_0} \in \mathcal{F}$ une famille paramétrique de lois.

Définition 4.1 *On appelle statistique, ou estimateur*[21] *une variable aléatoire* $T = h(X_1, \cdots, X_n) = h \circ (X_1, \cdots, X_n)$ *où* h *est une fonction mesurable. Notons que cette statistique est une variable aléatoire, dont la loi dépend de* θ_0*. On appelle*[22] *estimation (ou estimation ponctuelle) la réalisation de notre estimateur* $T(\omega) = h(x_1, \cdots, x_n)$*.*

[20] En réalité, il suffit que le modèle soit dominé, i.e. que toutes les lois de la familles soient absolument continues par rapport à une même mesure de réference. Dans ce cas là, on considère la famille de densité de Radon Nikodim par rapport à cette mesure. En général, on prend comme mesure de référence la mesure de comptage, ou le mesure de Lebesgue.

[21] Dans de nombreux ouvrages, on demande de plus à ce qu'un estimateur soit à valeur dans le bon ensemble, par exemple $[0,1]$, si l'on veut estimer une probabilité.

[22] Les termes varies, mais il faut bien en choisir un !

Remarque : Pour bien comprendre quelles fonctions ne sont pas des statistique, considérons par exemple $s = \mathbb{E}[X_1] = \int_{\omega\in\Omega} X_1(\omega)\mathrm{d}\mathbb{P}(\omega)$. s n'est pas une statistique car elle est une fonction de la loi/variable aléatoire X_1, et non de sa réalisation. Ce n'est donc pas une statistique. De même, θ_0 n'est pas une statisique, au sens où il faut connaître la loi de X_i pour l'obtenir. $T = 5$ est une statistique (constante), car il peut bien s'obtenir à partir d'un échantillon.

Il faut donc retenir qu'on appelle *statistique* n'importe quelle variable que l'on peut construire à partir des observations. Remarquons aussi que l'on dispose de deux noms pour la même notion. L'utilisation dépendra du contexte. On attend d'un estimateur qu'il soit un "bon estimateur", pour cela, on va définir des critères de qualité, et on ne parlera jamais d'estimateur sans le qualifier.

Définition 4.2 (Premiers critères de qualité d'un estimateur) *Soit $T_n = T(X_1, \cdots, X_n)$ un estimateur de θ. On dit que*

- *T_n est sans biais si*
$$\mathbb{E}[T_n] = \theta.$$
- *T_n est convergent [23] (ou consistant) si*
$$T_n \xrightarrow{\mathbb{P}} \theta.$$
- *T_n est fortement convergent si*
$$T_n \xrightarrow{p.s.} \theta.$$

On définit donc le biais d'un estimateur T_n comme

$$Biais = \mathbb{E}[T] - \theta$$

Remarque : La "convergence" signifie que si vous disposez de très nombreuses observations, alors votre estimateur s'approche de la "vraie" valeur. Être sans biais signifie intuitivement (à cause de la LGN) tomber "en moyenne" sur la "vraie" valeur si vous faites de nombreuses fois l'expérience.

À partir de maintenant, on considèrera que $d = 1$, i.e. θ est un réel (on ne cherche à estimer qu'un unique paramètre). Les résultats suivant existent aussi pour $\mathbb{R}^d$, mais nécessitent un effort supplémentaire d'écriture (en particulier, la variance devient une matrice de Variance/Covariance...)

Proposition 4.3 *Un estimateur sans biais T_n vérifiant $\mathrm{Var}(T_n) \to 0$ est convergent. Un estimateur convergent, uniformément borné $\mathbb{L}^2$, est asymptotiquement sans biais.*

Remarque : On préfère donc des estimateurs avec une variance petite !

Proposition 4.4 (Décomposition biais-variance) *On a la décomposition suivante :*

$$MSE(T_n) := \mathbb{E}\big[(T_n - \theta)^2\big] = \underbrace{\mathbb{E}\big[(T_n - \mathbb{E}[T_n])^2\big]}_{Variance} + \underbrace{(\mathbb{E}[T_n] - \theta)^2}_{Biais^2}.$$

Définition 4.5
- *La moyenne empirique [24] d'un échantillon $(X_i)_{i=1,\cdots,n}$ est définie par*
$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

[23] Pour parler d'estimateur convergent, il faut en réalité une suite d'estimateurs. C'est toujours le cas lorsqu'on construit un estimateur sur un échantillon de taille quelconque

[24] prise en tant que variable atéatoire

- *La variance empirique non-biaisé est définie par*

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

Propriétés 4.6
- *Si* $\mathbb{E}[|X|] < +\infty$*, alors* $\bar{X}_n$ *est un estimateur non-biaisé, et fortement convergent, de l'espérance* $\mathbb{E}[X]$*.*
- *Si* $\mathbb{E}[|X|^2] < +\infty$ *alors* S_n^2 *est un estimateur non-biaisé, et fortement convergent, de l'espérance* $\mathbb{E}[X]$*.*

4.2 Construction d'estimateurs

On peut proposer quelques façons de construire un estimateur :
La plus simple est basée sur la loi des grands nombres.

Définition 4.7 (Estimateur empirique) *Soit* $X_1, \cdots X_n$ *un échantillon i.i.d. de loi* f_{θ_0}*,* θ_0 *inconnu. On veut estimer* $g(\theta_0)$ *(pour une certaine fonction* g*). S'il existe* h *telle que* $g(\theta) = \mathbb{E}_\theta[h(X)]$*, alors l'estimateur empirique de* $g(\theta)$ *s'écrit*

$$\hat{G} = \frac{1}{n}\sum_{i=1}^{n} h(X_i).$$

Le LGN nous donne directement le résultat suivant :

Propriétés 4.8 *Soit* h *telle que* $\forall\theta \in \Theta, \mathbb{E}[h(X)] < +\infty$*. Alors l'estimateur empirique de* $g(\theta) = \mathbb{E}[h(X)]$ *est un estimateur sans biais et fortement convergent.*

On se place maintenant dans le cadre d'un modèle dominé. Cela signifie que la loi de $X_1, \cdots, X_n$ admet une densité $f_\theta(x_1, \cdots, x_n)$ par rapport à une mesure de référence.
Dans le cadre de ce cours, on considère seulement les formes produit (le cas i.i.d.) $((f_\theta(x_1, \cdots, x_n) = \prod g(x_i))$.

Définition 4.9 *On appelle vraisemblance d'un échantillon la fonction* L *définie par* $L(\theta, x_1, \cdots, x_n) = f_\theta(x_1, \cdots, x_n)$*. Plus spécifiquement, on a*

- *Si* X *est discrète :*

$$L(\theta, x_1, \cdots, x_n) = \prod \mathbb{P}_\theta(X_i = x_i)$$

- *Si* X *a pour densité* f_θ *:*

$$L(\theta, x_1, \cdots, x_n) = \prod f_\theta(x_i)$$

Sous l'hypothèse de positivité de la vraisemblance*, on définit la log-vraisemblance par :*

$$l_\theta(x) := l(\theta, x_1, \cdots, x_n) := \ln\Big(L(\theta, x_1, \cdots, x_n)\Big).$$

Définition 4.10 (Estimateur par maximum de vraisemblance) *On appelle estimateur par maximum de vraisemblance la statistique* T_n *définie par :*

$$T_n = \arg\max_{t\in\Theta} l(t, X_1, \cdots, X_n).$$

Remarque : Intuitivement, lorsque Θ est discret, l'EMV minimise le risque quadratique $\mathbb{1}_{\hat{\theta}\neq\theta 0}$.

Propriétés 4.11 *L'EMV est un estimateur convergent de θ sous des hypothèses assez faibles (voir [1] pour les détails par exemple).*

Définition 4.12 (Estimateur par la méthode des moments) *On suppose que $\theta \in \mathbb{R}^k$, et on cherche à estimer θ par la méthode des moments. Pour cela, il faut être capable de calculer pour une famille de fonctions $(h_1, \cdots, h_k)$ (on prend la plupart du temps $h_i(x) = x^i$, d'où le nom de la méthode)*

$$g_i(\theta) = \mathbb{E}[h_i(X)].$$

En écrivant $g = \big(g_1, \cdots, g_k\big)$, on définit l'estimateur de θ par la méthode des moments T_n une statistique (à valeurs dans $\mathbb{R}^k$) vérifiant (si c'est possible) :

$$\forall i \in [1, k], g_i(T) = \frac{1}{n}\sum_{j=1}^{n} h_i(X_j).$$

Lorsque g est bijective, on a en particulier

$$T = g^{-1}\Big(\frac{1}{n}\sum_{j=1}^{n} h_1(X_j), \cdots, \frac{1}{n}\sum_{j=1}^{n} h_k(X_j)\Big)$$

Propriétés 4.13 *Cet estimateur est convergent sous des hypothèses assez faibles (voir [1] pour les détails par exemple).*

Par contre, cet estimateur n'est pas nécessairement sans biais !

À lire en complément pour les curieu-x-ses:

- Détails et preuves ?
- M-estimateurs et Z-estimateurs
- Estimation par noyaux
- Delta méthode
- Minimum de contraste/risque

4.3 Principe de construction d'un intervalle de confiance

Soit $\mathcal{F}$ une famille paramétrique de lois. On appelle intervalle de confiance pour θ au niveau de confiance p, un intervalle aléatoire construit à partir de deux statistiques $A_1(X_1, \cdots, X_n)$, et $A_2(X_1, \cdots, X_n)$ telle que

$$\mathbb{P}_{\theta_0}(\theta_0 \in [A_1, A_2]) = p.$$

On notera $IC_p(\theta) = [A_1, A_2]$. Remarquons que dans l'expression précédente, $A_1, A2$ sont aléatoire, mais pas θ_0. Cela signifie donc que l'intervalle ponctuel obtenu en prenant une réalisation de A_1 et A_2 ne permet plus de parler de probabilité[25].

Méthode usuelle de construction :

1. On construit un estimateur $\hat{T}$ de θ_0

[25] Le niveau de confiance d'un intervalle de confiance s'interprête donc comme le niveau de confiance de la démarche pour le construire : pour une proportion p des échantillons, la "vraie valeur" (inconnue) sera bien dans l'intervalle, et quelques fois (avec proba $1-p$) non.

2. On obtient une propriété permettant d'expliciter la loi de $\hat{T}$ en fonction de θ_0 (ou très souvent la loi d'une fonction liant les deux, ex $\hat{T} - \theta_0$, $\frac{\hat{T}}{\theta_0}$...

3. De cette loi connue, on extrait des quantiles permettant d'encadrer $\hat{T} - \theta_0$, $\frac{\hat{T}}{\theta_0}$...

4. On peut alors isoler $\hat{T}$ (paradigme des tests) ou θ_0 (paradigme des IC.)

Remarque :

- On peut aussi obtenir plutôt des lois asymptotiques pour obtenir des IC asymptotique. On demande alors $\limsup_{n\to+\infty} \mathbb{P}_{\theta_0}(\theta_0 \in [A_1(X_1, \cdots, X_n), A_2(X_1, \cdots, X_n)]) = p$
- On construira souvent des IC unilatéraux, en utilisant un seul sens de majoration (on a donc $A_1 = -\infty$, ou $A_2 = +\infty$)
- Pour trouver quelles estimateurs choisir, on pourra se référer au chapitre sur les Tests.

4.4 Variance, efficacité et borne de Cramer-Rao

Les preuves de cette partie seront a priori faites en TD
Un premier critère de qualité pour comparer des estimateurs convergent est d'utiliser le

$$MSE(T_n) := \mathbb{E}\big[(T_n - \theta)^2\big] = \underbrace{\mathbb{E}\big[(T_n - \theta)^2\big]}_{\text{Variance}} + \underbrace{(\mathbb{E}[T_n] - \theta)^2}_{\text{Biais}^2}.$$

Parmi les estimateurs sans biais, le meilleur est donc celui qui minimise la variance [26].
On peut aller plus loin, et se demander si on peut trouver une borne minimale pour la variance de *n'importe quel* estimateur sans biais. La réponse est oui, avec quelques bémols. Pour cela, il est nécessaire d'introduire quelques notations.
Sous l'hypothèse que la vraisemblance est $\mathcal{C}^1$, on définit la fonction **score** [27] par :

$$sc_\theta(x) := \partial_\theta l_\theta(x)$$

Cela permet de définir l'**Information de Fisher** du modèle [28], dès que la quantité suivante existe, par

$$I(\theta) := \mathbb{E}_\theta[sc_\theta(X)^2] = \mathbb{E}_\theta\Big[\big(\partial_\theta l_\theta(X)\big)^2\Big].$$

L'information de Fisher mesure à quel point le modèle change lorsque l'on varie le paramètre. Si le modèle "change" beaucoup, alors il sera facile d'estimer θ, et par conséquent, les données apporteront beaucoup d'information sur θ.
Il y a souvent une façon plus simple de calculer l'information de Fisher, comme donné par la formule suivante.

Proposition 4.14 *L'Information de Fisher vérifie l'égalité suivante :*

$$I_n(\theta) = -\mathbb{E}\Big[\partial_\theta^2 l_\theta(X)\Big] = -\mathbb{E}\Big[\partial_\theta^2 l(\theta, X_1, \cdots, X_n)\Big].$$

Lorsque $X_1, \cdots X_n$ désigne un échantillon i.i.d., on a alors

$$I_n(\theta) = -n\mathbb{E}\Big[\partial_\theta^2 \ln\big(f_\theta(X_1)\big)\Big]$$

[26] Mais le meilleurs estimateur au sens de l'erreur quadratique moyenne n'est pas nécessairement sans biais
[27] En dimension supérieure, la dérivée est remplacée par un gradient
[28] En dimension supérieure, l'information de Fisher est une matrice

Remarque : Remarquez bien que dans l'espérance, le $X = X_1, \cdots, X_n$ désigne un échantillon de loi donnée par f_θ. Lorsqu'on calculera l'Information de Fisher sur des exemples, il faudra, après avoir calculée la fonction $\partial_\theta^2 l_\theta(x)$, faire bien attention par l'appliquer en une variable aléaoire de loi donnée par f_θ.

Le résultat suivant permet de borner inférieurement la variance de n'importe quel estimateur sans biais.

Théorème 4.15 (Borne de Cramér-Rao) *Soit $T(X)$ un estimateur sans biais de θ. On suppose que la vraisemblance est $\mathcal{C}^1$, et que l'on peut dériver sous le signe $\mathbb{E}$. Alors*

$$\mathrm{Var}_\theta\big(T(X)\big) \geq \frac{1}{I_n(\theta)}.$$

Lorsque la borne est atteinte, on dit que T est un **estimateur efficace**. Lorsque l'on dispose d'une suite d'estimateurs T_n (sans biais [29] tels que $\lim n\,\mathrm{Var}_\theta\big(T(X)\big) = \lim \frac{n}{I_n(\theta)}$, on parle d'estimateur **asymptotiquement efficace**. En général, cette propriété prend plutôt la forme $\sqrt{n}\big(T_n - \theta\big) \xrightarrow{\mathcal{L}} \mathcal{N}\big(0, I_1^{-1}(\theta)\big)$, auquel cas, on parlera d'estimateur **asymptotiquement normal et efficace**.

Remarque : Ce resultat devrait au premier abord sembler surprenant et très fort. Il dit que, aussi inventif que vous soyez pour construire un estimateur sans biais, vous ne pourrez pas faire mieux que la borne de Cramér-Rao ! Il y a cependant quelques bémols :

- Il se peut que cela soit intéressant de rajouter du biais, au sens où un estimateur biaisé peut avoir un risque quadratique plus petit qu'un estimateur sans biais.
- Il n'existe pas nécessairement d'estimateur sans biais, encore moins nécessairement d'estimateur efficace.
- En réalité, un estimateur efficace existe essentiellement dans un cadre restreint, lorsque la densité prend la forme $f_\theta(x) = b(x)e^{\theta T(x) - \psi(\theta)}$.

À lire en complément pour les curieu-x-ses:

- Famille exponentielle
- Statistique Exhaustive/ stat complete
- Lien entre Information de Fisher et Entropy relative

[29] Il est en fait suffisant que le biais asymptotique décroisse plus rapidement que $\frac{1}{\sqrt{n}}$

5 Tests et statistiques à utiliser pour les intevalles de confiances

Dans toute cette section, on va seulement introduire des statistiques de test. On essaiera d'aller du plus simple au plus compliqué.

Dans ce cours, on n'attend pas de vous que vous connaissiez la théorie abstraite des test. Seulement de comprendre comment utiliser les différents résultats pour pratiquer un test.

5.1 Principe général et méthodologie

Pour faire les choses dans les règles de l'art, vous devriez écrire le protocole statistique **avant de recueillir les données**. Autrement dit, vous savez l'expérience que vous souhaitez réaliser, et vous détaillez les conclusions possibles en fonctions des données obtenues.

Avant de faire un test statistique, on dispose

- D'une question à laquelle on souhaiterait répondre (ex : ce dé est-il truqué en faveur du 6 ?)
- D'une expérience que l'on va réaliser (ex : on va lancer 600 fois le dé, et compter le nombre de 6).

Le raisonnement sera alors le suivant : "Sur les 600 lancers, **si le dé n'était pas truqué**, on s'attend à voir *en moyenne* 100 lancers ayant donné 6. Si on en obtient *beaucoup plus*, alors ces données sont tellement surprenantes que je concluerais que le dé est truqué"

Il reste évidemment à quantifier le *beaucoup plus*.

Plus précisemment, on peut suivre les étapes suivantes.

1. Modélisation : On va formaliser le problème dans un cadre mathématique, en explicitant la variable aléatoire étudiée, et le ou les paramètres inconnus. On formule dès lors des hypothèses sur notre expérience.

   Ex : On note $X_1, \cdots, X_{600}$ les variable aléatoires à valeurs dans $\{0,1\}$, telle que X_i vaut 1 si le $i$eme lancer donne un 6, et 0 sinon. Ces variables aléatoires sont supposées i.i.d., et donc de loi $Ber(p)$, avec un paramètre p inconnu.

2. Formulation de l'hypothèse alternative : On exprime mathématiquement (en fonction des paramètres inconnus) la conclusion que l'on souhaiterait obtenir de façon significative.

   Ex : $H_1 : p > \frac{1}{6}$

3. Formulation de l'hypothèse nulle : On exprime mathématiquement (en fonction des paramètres inconnus) l'hypothèse par défaut en l'absence de données particulières.

   Ex : $H_0 : p \leq \frac{1}{6}$

4. Identification du pire des cas pour H_0 : Dans le cadre du cours de cette année, on se restreindra à des hypothèses nulles simples (du type $H_0 : p = ...$). On recherche donc le cas le moins favorable pour différencier entre H_0 et H_1)

   Ex : Parmi tous les $p \leq \frac{1}{6}$, c'est $p = \frac{1}{6}$ qui nous rendra le plus difficile séparer H_0 et H_1.

5. Statistique de test et zone de rejet : L'objectif est maintenant d'identifier des évènement improbables si H_0 était vrai (et dont on peut quantifier la probabilité). Ces évènements prennent en général la forme $Z > a$ où Z sera une statistique de résumé.

   Ex : Si $p = \frac{1}{6}$, alors $S = \sum X_i \sim Binom(600, \frac{1}{6})$ est de loi connue, et les quantiles[30], nous donnent, pour un test à 1%, $\mathbb{P}_{H_0}(S > 122) \leq 0.01$.

[30] calculés par un logiciel par exemple

6. Expérience et conclusion : On conclut en fonction du résultat de l'expérience.

   Ex : L'expérience donne $s = 134$ (on a fait 134 lancers qui ont donné un 6), or $134 > 122$, donc on rejette H_0.

7. p-valeur : On peut aussi conclure en calculant la $p - valeur$, i.e. la probabilité, sous H_0, d'obtenir une statistique de résumé plus extrème que dans notre expérience.

   Ex : Ici, on calcule $\mathbb{P}_{H_0}(S > 134) = 0.0001363804$, on obtient donc une p-valeur de 10^{-4} environ. Ce nombre étant plus petit que 1%, on rejette H_0.

5.2 Quelques définitions

On appellera **échantillon** $X_1, \cdots, X_n$ une suite de v.a. i.i.d., et **statistique** une fonction mesurable d'un échantillon.

Définition 5.1 *Un test pur T est une statistique à valeur dans $\{0, 1\}$.*

Le but d'un test étant de tester, si possible avec qualité, on a besoin d'hypothèse nulle et d'hypothèse alternative. Pour un test paramétrique, on a un ensemble de paramètres $\Theta \subset \mathbb{R}^d, d > 1$ et les hypothèses se présentent sous la forme

$$H_0 : \theta \in \Theta_0$$
$$H_1 : \theta \in \Theta_1$$

Le test a pour but de "dire" 0 quand H_0 est vrai, et 1 quand H_0 est fausse

On peut alors définir les erreurs d'un test

Définition 5.2 *Le risque (ou erreur) de première espèce est défini par*

$$R_1 : \Theta_0 \to [0, 1]$$
$$\theta \mapsto \mathbb{E}_\theta[\mathbb{1}_{T=1}] = \mathbb{E}_\theta[T]$$

*Le niveau est le risque maximum de première espèce : $\sup_{\theta \in \Theta_0} R_1(\theta)$. C'est la probabilité maximale de **rejetter à tord** $->$ on le veut bas. Par abus de language, on parle de "test de niveau α" pour tous les tests de niveau plus petit que α.*

Le risque (ou erreur) de seconde espèce est défini par

$$R_2 : \Theta_1 \to [0, 1]$$
$$\theta \mapsto \mathbb{E}_\theta[\mathbb{1}_{T=0}] = 1 - \mathbb{E}_\theta[T]$$

La puissance est le contraire :

$$P : \Theta_1 \to [0, 1]$$
$$\theta \mapsto \mathbb{E}_\theta[\mathbb{1}_{T=1}] = \mathbb{E}_\theta[T]$$

*C'est la fonction qui donne pour tout θ où on aurait du rejetter la probabilité de **rejetter à raison**. $->$ on la veut haute.*

Définition 5.3 (p-valeur) *Soit $\mathcal{F} = (T_\alpha)_{\alpha \in]0,1[}$ une famille de tests de niveau α, et $x_1, \cdots, x_n$ une réalisation de $X_1, \cdots, X_n$. La p-valeur associée à cette famille et cette observation est le plus grand niveau tel que le test ne soit pas rejetté :*

$$p\text{-}valeur(\mathcal{F}, x) = \sup\{\alpha \in]0, 1[, T_\alpha(x) = 0\}.$$

Remarque : Le cas le plus classique est celui où les fonctions $\alpha \mapsto T_\alpha(x)$ sont croissantes, et on a aussi dans ce cas là :

$$p - \text{valeur}(\mathcal{F}, x) = \inf\{\alpha \in]0,1[, T_\alpha(x) = 1\}.$$

En français, cela signifie que c'est l'erreur qu'il faut accepter de faire pour pouvoir rejetter H_0. Si la p-valeur est très petite, cela signifie qu'on "choisit" H_1, au vue de nos observations, en ayant une très petite probabilité de se tromper. Si la p valeur est "grande",par exemple $\frac{3}{4}$, cela signifie qu'il faut accepter de se tromper 3 fois sur 4 pour pouvoir rejetter. **En réalité, la p-valeur mesure plus précisemment la probabilité d'observer des valeurs "plus improbables" (plus extrèmes dans la plupart des cas) que $(x_1, \cdots, x_n)$, sachant que H_0 est vraie.**

Cela signifie que lorsqu'on parle "d'accepter de se tromper", cela ne signifie pas "sur notre expérience particulière", mais bien sûr de nombreuses répétitions de l'expérience, si on pouvait la réaliser plusieurs fois dans les mêmes conditions.

Si on voulait avoir accès à la probabilité que H_0 soit vraie, sachant les observations, il faudrait connaître en plus le comportement de θ, pour savoir s'il est plus souvent dans Θ_0 ou dans Θ_1, voir le chapitre sur les statistiques Bayésiennes (...qui n'est pas encore écrit).

Au passage, dans tous les tests suivants, on controlera le niveau, mais pas la puissance. Un niveau bas, c'est bien, mais pas suffisant. En effet, si on accepte tout le temps (test constant $T = 0$), on a un niveau à 0 (on ne rejette jamais à tord, puisqu'on ne rejette jamais), mais une puissance nulle aussi (puisqu'on ne rejette jamais à raison non plus).

Donc dans la suite, même si on ne fait pas les calculs, ayez dans la tête que certains tests sont meilleurs que d'autres.

5.3 Test naif

Proposition 5.4 *Soit $X_1, \cdots, X_n$ un échantillon de loi $\mathcal{N}(m, \sigma^2)$. On introduit*

$$\bar{X} = \frac{1}{n}\sum_{i=0}^{n} X_i$$

La statistique T_1 définie par

$$T_1(X_1, \cdots, X_n) := \frac{1}{\sqrt{n}\sigma}\sum_{i=0}^{n}(X_i - m) = \sqrt{n}\frac{\bar{X} - m}{\sigma},$$

vérifie

$$T_1(X_1, \cdots, X_n) \sim \mathcal{N}(0, 1).$$

Utilisation :
Si σ^2 est connu et si on pose

$$H_0 : m = m_0$$
$$H_1 : m \neq m_0$$

On peut alors, sous H_0, connaître la loi de T_1 et en extraire les quantiles.
Si m est connu et si on pose

$$H_0 : \sigma = \sigma_0$$
$$H_1 : \sigma \neq \sigma_0$$

On peut alors, sous H_0, connaître la loi de T_1 et en extraire les quantiles. Par contre, celui là il est nul, donc on le l'utilise pas (mais on pourrait).

Proposition 5.5 (Version asymptotique) *Soit $X_1, \cdots, X_n$ un échantillon dans $\mathbb{L}^2$, d'espérance m et de variance σ^2). La statistique T_1 définie par*

$$T_1(X_1, \cdots, X_n) := \sqrt{n}\frac{\bar{X} - m}{\sigma}$$

vérifie

$$T_1(X_1, \cdots, X_n) \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1).$$

Remarque : C'est le TCL !
Utilisation : Idem, pour n assez grand.
Dans la suite, on notera T_i pour $T_i(X_1, \cdots, X_n)$ pour simplifier les notations.

5.4 Tests du χ^2

Définition 5.6 (Loi du χ^2, définition/propriété) *Soit $X_1, \cdots, X_n$ un échantillon $\mathcal{N}(0, 1)$. On pose*

$$Z = \sum_{i=1}^{n} X_i^2.$$

La v.a. Z suit la loi du "khi deux" à n degrés de libertés ($X \sim \chi^2(n)$) et admet pour densité

$$f_Z(t) = \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} t^{\frac{n}{2}-1} e^{-\frac{t}{2}} \mathbb{1}_{\mathbb{R}^+},$$

où Γ est la fonction gamma d'Euler :

$$\Gamma(z) = \int_0^{+\infty} t^{z-1}\, e^{-t} \mathrm{d}t.$$

Proposition 5.7 *Soit $X_1, \cdots, X_n$ un échantillon de loi $\mathcal{N}(m, \sigma^2)$. La statistique T_2 définie par*

$$T_2 := \sum_{i=0}^{n} \frac{1}{\sigma^2}(X_i - m)^2,$$

vérifie

$$T_2 \sim \chi^2(n).$$

On introduit

$$S_n = \sqrt{\frac{1}{n-1}\sum_{i=0}^{n}(X_i - \bar{X})^2}.$$

La statistique T_3 définie par

$$T_3 := \sum_{i=0}^{n} \frac{1}{\sigma^2}(X_i - \bar{X})^2 = (n-1)\frac{S_n^2}{\sigma^2},$$

(où $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$) vérifie

$$T_3 \sim \chi^2(n-1).$$

Utilisation : Si m est connu et si on pose

$$H_0 : \sigma = \sigma_0$$
$$H_1 : \sigma \neq \sigma_0$$

On peut alors, sous H_0, connaître la loi de T_2 et en extraire les quantiles. Si m est inconnu et si on pose

$$H_0 : \sigma = \sigma_0$$
$$H_1 : \sigma \neq \sigma_0$$

On peut alors, sous H_0, connaître la loi de T_3 et en extraire les quantiles.

Proposition 5.8 (Version asymptotique) *Soit $X_1, \cdots, X_n$ un échantillon (de loi discrète) à valeurs dans $\{e_1, \cdots, e_J\}$ et tel que $\forall j \in [\![1, J]\!], \mathbb{P}(X_1 = e_j) = p_j$. On pose*

$$\forall j \in [\![1, J]\!], n_j = \sum_{i=1}^{n} \mathbb{1}_{X_i = e_j}.$$

Alors la statistique T_4 définie par

$$T_4 := \sum_{j=1}^{J} \frac{\big(n_j - np_j\big)^2}{np_j},$$

vérifie

$$T_4 \xrightarrow{\mathcal{L}} \chi^2(J-1).$$

Lorsque l'échantillon est vectoriel ($X_i = (Y_i, Z_i)$), on peut se poser la question de l'indépendance entre Y et Z. pour cela, si Y est à valeurs dans $\{a_1, \cdots, a_I\}$, et Z à valeurs dans $\{b_1, \cdots, c_K\}$, on peut poser

$$\forall i \in [\![1, I]\!], \forall k \in [\![1, K]\!], m_{ik} = \sum_{i \in [\![1,I]\!], k \in [\![1,K]\!]} \mathbb{1}_{Y=a_i} \mathbb{1}_{Z=b_k}$$
$$M_i^{(Y)} = \sum_{i \in [\![1,I]\!]} \mathbb{1}_{Y=a_i}$$
$$M_k^{(Z)} = \sum_{k \in [\![1,K]\!]} \mathbb{1}_{Y=b_k}$$
$$\hat{p}_{ik} = \frac{M_i^{(Y)} M_k^{(Z)}}{n}$$

Et la statistique T_5 définie par

$$T_5 := \sum_{i \in [\![1,I]\!], k \in [\![1,K]\!]} \frac{\big(m_{ik} - \hat{p}_{ik}\big)^2}{\hat{p}_{ik}},$$

vérifie, lorsque Y et Z sont indépendantes,

$$T_5 \xrightarrow{\mathcal{L}} \chi^2\big((I-1)(K-1)\big).$$

Tout ceci est plus facile à comprendre visuelement.

Remarque : Vous verrez souvent les notations suivantes $O_{ik} = m_{ik}$ (O pour "observés") et $E_{ik} = \hat{p}_{ik}$ (E pour "Espérés").

Utilisation : Tout d'abord, si l'échantillon initial est continu, on le regroupe en classes, pour en faire un échantillon $X_1, \cdots, X_N$ discret.

On peut utiliser cette propriété pour

- Tester l'adéquation d'un échantillon à une loi donnée (donc on connaît les p_j sous H_0)
- Tester l'homogénéité entre deux échantillons
- Tester l'indépendance

Dans tous les cas, sous H_0, il suffit d'aller utiliser les quantiles du χ^2.

Remarque : En pratique, on dit souvent que pour que le test soit "valide", il faut $n > 30$, et $\hat{p}_{ik} > 5$. Evidemment, plus les effectifs sont grands, plus l'approximation est correcte.

5.5 Student

Définition 5.9 (Loi de Student, définition/propriété) *Soit $X \sim \mathcal{N}(0,1)$, et $U \sim \chi^2(n)$ deux v.a. indépendantes. On pose*

$$Z = \frac{X}{\sqrt{\frac{U}{n}}}.$$

La v.a. Z suit la loi de Student à n degrés de libertés ($X \sim \mathrm{St}(n)$) et admet pour densité

$$f_Z(t) = \frac{1}{\sqrt{n\pi}} \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})} \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}},$$

où Γ est la fonction Gamma d'Euler :

$$\Gamma(z) = \int_0^{+\infty} t^{z-1}\, e^{-t} \mathrm{d}t.$$

Proposition 5.10 *Soit $X_1, \cdots, X_n$ un échantillon de loi $\mathcal{N}(m, \sigma^2)$.*

La statistique T_6 définie par

$$T_6 := \frac{\sqrt{n}\big(\bar{X} - m\big)}{\sqrt{\frac{\sum_{i=1}^n \big(X_i - \bar{X}\big)^2}{n-1}}} = \sqrt{n}\frac{\bar{X} - m}{S_n}.$$

vérifie

$$T_6 \sim \mathrm{St}(n-1).$$

Utilisation : Ici, σ^2 et m sont inconnus. Si on pose

$$H_0 : m = m_0$$
$$H_1 : m \neq m_0$$

Alors, sous H_0, la loi de T_6 est connue, on peut conclure en allant chercher les quantiles.

Proposition 5.11 *Soit $X_1, \cdots, X_{n_1}$ un échantillon de loi $\mathcal{N}(m_1, \sigma^2)$, et $Y_1, \cdots, Y_{n_2}$ un échantillon de loi $\mathcal{N}(m_2, \sigma^2)$ (les deux échantillons ont donc la même variance). On suppose que les deux échantillons sont indépendants.*

La statistique T_7 définie par

$$T_7 := \frac{\bar{X} - m_1 - (\bar{Y} - m_2)}{s_{XY} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}.$$

où

$$s_X^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} \left(X_i - \bar{X}\right)^2.$$
$$s_Y^2 = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} \left(Y_i - \bar{Y}\right)^2.$$
$$s_{XY} := \sqrt{\frac{(n_1 - 1)s_X^2 + (n_2 - 1)s_Y^2}{n_1 + n_2 - 2}}.$$

vérifie

$$T_7 \sim \mathrm{St}(n_1 + n_2 - 2).$$

Utilisation : Ici, σ^2 est inconnu, mais on sait (?) que les deux échantillons X et Y ont même variance. Les espérances m_1 et m_2 sont inconnues. Si on pose

$$H_0 : m_1 - m_2 = k$$
$$H_1 : m_1 - m_2 \neq k$$

Alors, sous H_0, la loi de T_7 est connue, on peut conclure en allant chercher les quantiles.

5.6 Fisher

Définition 5.12 (Loi de Fisher, définition/propriété) *Soit $U_1 \sim \chi^2(k_1)$, et $U_2 \sim \chi^2(k_2)$ deux v.a. indépendantes. On pose*

$$Z = \frac{\frac{U_1}{k_1}}{\frac{U_2}{k_2}}.$$

La v.a. Z suit la loi de Fisher à (k_1, k_2) degrés de libertés ($X \sim F(k_1, k_2)$) et admet pour densité

$$f(x) = \frac{\left(\frac{k_1 x}{k_1 x + k_2}\right)^{\frac{k_1}{2}} \left(1 - \frac{k_1 x}{k_1 x + k_2}\right)^{\frac{k_2}{2}}}{x\beta(\frac{k_1}{2}, \frac{k_2}{2})} \mathbb{1}_{\mathbb{R}^+},$$

où β est la fonction bêta :

$$\beta(x, y) = \int_0^1 t^{x-1}(1 - t)^{y-1} \mathrm{d}t.$$

Proposition 5.13 *Soit $X_1, \cdots, X_{n_1}$ un échantillon de loi $\mathcal{N}(m_1, \sigma_1^2)$, et $Y_1, \cdots, Y_{n_2}$ un échantillon de loi $\mathcal{N}(m_2, \sigma_2^2)$ (les deux échantillons ont donc la même variance). On suppose que X et Y sont indépendants.*

La statistique T_8 définie par

$$T_8 := \frac{\frac{s_X^2}{\sigma_1^2}}{\frac{s_Y^2}{\sigma_2^2}}.$$

où

$$s_X^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} \big(X_i - \bar{X}\big)^2.$$
$$s_Y^2 = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} \big(Y_i - \bar{Y}\big)^2.$$

vérifie

$$T_8 \sim F(n_1 - 1, n_2 - 1).$$

Utilisation : Ici, m_1 et m_2 sont inconnus. Si on pose

$$H_0 : \frac{\sigma_1}{\sigma_2} = k$$
$$H_1 : \frac{\sigma_1}{\sigma_2} \neq k$$

Alors, sous H_0, la loi de T_8 est connue, on peut conclure en allant chercher les quantiles.

5.7 Test de Kolmogorov-Smirnov

Définition 5.14 (Loi de Kolmogorov-Smirnov, définition/propriété) *Soit $X_1, \cdots, X_n$ un échantillon de fonction de répartition F, que l'on suppose continue. On définit*

$$F_n(t) = \frac{1}{n} \sum_{k=1}^{n} \mathbb{1}_{X_i \leq t},$$

et on pose

$$D_n = \sup_{t \in \mathbb{R}} \big|F_n(t) - F(t)\big|.$$

On a alors

$$\sqrt{n} T_9 \to K.$$

La loi limite de $\sqrt{n}D$ (i.e. loi de K) ne dépend pas de F, est est appellée loi de Kolmogorov-Smirnov. Elle a pour fonction de répartition

$$\mathbb{P}\big(K \leq t\big) = 1 - 2 \sum_{r=1}^{+\infty} (-1)^{r-1} \exp(-2r^2 t^2).$$

Proposition 5.15 *Soit $X_1, \cdots, X_{n_1}$ un échantillon de fonction de répartition F continue*
La statistique T_9 définie par

$$T_9 := \sqrt{n} \sup_{t \in \mathbb{R}} \big|F_n(t) - F(t)\big|,$$

où

$$F_n(t) = \frac{1}{n} \sum_{k=1}^{n} \mathbb{1}_{X_i \leq t},$$

vérifie

$$T_9 \xrightarrow{L} K.$$

Utilisation : Ici, la loi de X_1 est inconnue, mais on suppose sa fonction de répartition F continue. On pose

$$H_0 : F = G$$
$$H_1 : F \neq G$$

Alors, sous H_0, la loi asymptotique de T_9 est connue, on peut conclure en allant chercher les quantiles.

5.8 Test du rapport de vraisemblance

Pour ce chapitre, on ne considère plus seulement des test purs, mais aussi des tests stochastiques :

Définition 5.16 *Un test stochastique T est une statistique à valeur dans* $[0, 1]$.

Comme pour un test "pur", lorsque $T = 0$, on ne rejettera pas H_0, lorsque $T = 1$, on rejettera H_0, et lorsque $T \in]0, 1[$ on acceptera H_0 avec probabilité égale à T. Les définitions de niveau et de puissance restent identiques.

Soit $X_1, \cdots, X_n$ un échantillon i.i.d. issu d'une loi f_θ appartenant à un modèle paramétrique $(f_\theta)_{\theta\in\Theta}, \Theta \subset \mathbb{R}^{d}$[31]

Soit $\theta_0, \theta_1 \in \Theta$, on appelle rapport de vraisemblance la statistique

$$R(\theta_0, \theta_1, X_1, \cdots, X_n) := \frac{L(\theta_1, X_1, \cdots, X_n)}{L(\theta_0, X_1, \cdots, X_n)},$$

où $\theta \mapsto L(\theta, X_1, \cdots, X_n)$ désigne la vraisemblance de l'échantillon.

On veut tester

$$H_0 : \theta = \theta_0$$
$$H_1 : \theta = \theta_1.$$

Lemma 5.17 (Neyman-Pearson) *Il existe $k_\alpha \in \mathbb{R}^+$ et $\gamma_\alpha \in]0, 1[$ tels que*

$$T_{10} = \mathbb{1}_{R>k_\alpha} + \gamma_\alpha \mathbb{1}_{R=k_\alpha}$$

vérifie

- $\mathbb{E}_{\theta_0}\big[T_{10}\big] = \alpha$ *(i.e. $\mathbb{1}_{T_{10}}$ est un test de niveau exactement α)*
- T_{10} *est le test est le plus puissant, parmis tous les tests de niveau α.*

Remarque :

- On peut aussi tester des "zones" $\Theta_0, \Theta_1 \subset \Theta$ (en prenant les sup), mais dans ce cas là, il faut des hypothèses supplémentaires pour conserver le résultat. On obtient alors un test "Uniformément plus puissant" (où l'uniformité est prise sur Θ_1.)
- Rien ne nous indique ici comment trouver k_α et γ_α. En réalité, pour le trouver, il suffit d'exprimer T_{10} (dont on ne connaît pas la loi) comme une fonction croissante d'une statistique dont la loi est connue. Puis de chercher $\tilde{k}_\alpha, \gamma_\alpha$ pour cette nouvelle statistique. On verra des exemples en exercice.

[31] On rappelle que cela signifie qu'il existe une mesure qui domine toutes les lois du modèle. En particulier, dans le cadre de ce cours, cela implique qu'on considère soit une famille de loi discrètes à valeurs dans un même ensemble E, soit une famille de lois continues (par rapport à la mesure de Lebesgue).

Preuve : On note $X = (X_1, \cdots, X_n)$ pour simplifier les notations. Pour prouver le premier point, il faut prouver qu'il existe k, γ tels que

$$\mathbb{P}_{\theta_0}\big(\frac{L(\theta_1, X)}{L(\theta_0, X)} > k\big) + \gamma \mathbb{P}_{\theta_0}\big(\frac{L(\theta_1, X)}{L(\theta_0, X)} = k\big) = \alpha.$$

Or la fonction $t \mapsto \mathbb{P}_{\theta_0}\big(\frac{L(\theta_1,X)}{L(\theta_0,X)} > t\big)$ est décroissante, il existe k_α tel que

$$\mathbb{P}_{\theta_0}\big(\frac{L(\theta_1, X)}{L(\theta_0, X)} > k_\alpha\big) \leq \alpha \leq \mathbb{P}_{\theta_0}\big(\frac{L(\theta_1, X)}{L(\theta_0, X)} \geq k_\alpha\big).$$

Si $\mathbb{P}_{\theta_0}\big(\frac{L(\theta_1,X)}{L(\theta_0,X)} = k_\alpha\big) = 0$, on peut poser $\gamma_\alpha = \frac{1}{2}$, sinon,

$$\gamma_\alpha = \frac{\alpha - \mathbb{P}_{\theta_0}\big(\frac{L(\theta_1,X)}{L(\theta_0,X)} > k_\alpha\big)}{\mathbb{P}_{\theta_0}\big(\frac{L(\theta_1,X)}{L(\theta_0,X)} = k_\alpha\big)},$$

convient.

Pour le second point, soit ψ un autre test de niveau α. Notons

- $T > \psi \Rightarrow T > 0 \Rightarrow L(\theta_1, X) \geq k_\alpha L(\theta_0, X)$
- $T > \psi \Rightarrow T < 1 \Rightarrow L(\theta_1, X) \leq k_\alpha L(\theta_0, X)$

D'où

$$(T - \psi)L(\theta_1, X) \geq k_\alpha (T - \psi) L(\theta_0, X).$$

On a

$$\mathbb{E}_{\theta_0}[\psi] \leq \alpha = \mathbb{E}_{\theta_0}[T].$$

On veut montrer que T est plus puissant que ψ, c'est à dire que

$$\mathbb{E}_{\theta_1}[T] \geq \mathbb{E}_{\theta_1}[\psi].$$

On calcule donc

$$\begin{aligned}
\mathbb{E}_{\theta_1}[T - \psi] &= \int (T - \psi)(X) L(\theta_1, X) \\
&\geq \int (T - \psi)(X) k_\alpha L(\theta_0, X) \\
&\geq k_\alpha \mathbb{E}_{\theta_0}[T - \psi] \\
&\geq 0
\end{aligned}$$

□

À lire en complément pour les curieu-x-ses:

- Voir les tests de signes et de rang : Wilcoxon et Mann Whitney.
- Théorème de Cochran pour les preuves
- Théorème de Donsker
- Test multiples : Bonferroni, FDR Benjamini-Hochberg

6 Bibliographie

References

[1] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.