STATISTIQUE ÉLÉMENTAIRE

FRANÇOIS CHAPON

Université de Toulouse

2024–2025

Table des matières

Introduction

La statistique est une branche des mathématiques qui collecte, analyse, interprète et présente des données pour en extraire des informations utiles.

Elle se divise en *statistique descriptive* qui consiste en la description des données (résumé et représentation) et *statistique inférentielle* qui consiste à induire les caractéristiques d'un groupe général (la population) à partir d'un groupe particulier (l'échantillon).

1. Statistique élémentaire

1.1. **Statistiques descriptives.**

1.1.1. *Vocabulaire de base.*

- La **population** est l'ensemble que l'on observe et qui sera soumis à une étude statistique. (ex : la population française, les étudiant·e·s du cours de proba, une population de coccinelles, des objets inanimés, etc...) ;
- Chaque élément de la population est appelé un **individu** (ou une unité statistique) ;
- Un **échantillon** est un sous-ensemble de la population (typiquement, si la population est trop grande, on étudiera une partie de la population). Le nombre d'individus dans l'échantillon est la taille de l'échantillon ;
- Le **caractère**, ou **variable** statistique, est la propriété que l'on observe dans la population. Les valeurs prisent par un caractère sont appelées les **modalités** ;

On ne s'intéressera qu'aux variables quantitatives, c'est-à-dire dont les modalités sont à valeurs numériques (dans un sous-ensemble de $\mathbb{R}$).

Définition 1.1. *On appelle série statistique une suite d'observations d'un caractère relevées sur les individus d'une population ou d'un échantillon d'une population. On la notera souvent $x = (x_1, \dots, x_n)$, où n est la taille de l'échantillon et x_i la valeur du caractère correspondant au i^e individu.*

On supposera les séries statistiques ordonnées par ordre croissant. Si ce n'est pas le cas, on commencera par ordonner les valeurs par ordre croissant, et on notera

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

la série obtenue, appelée statistiques d'ordre. On a donc que $x_{(k)}$ est la k^{e} plus petite valeur de la série x.

1.1.2. *Traitement des données.*

Définition 1.2. *Soit $x = (x_1, \dots, x_n)$ une série statistique à valeurs dans E (un sous-ensemble de $\mathbb{R}$). La mesure empirique de x est la probabilité $\mathbb{P}_x$ sur E définie par :*

$$\mathbb{P}_x(\{k\}) = \frac{\text{Card}(\{i \in \{1, \dots, n\} \mid x_{(i)} = k\})}{n},$$

pour tout $k \in E$.

On a donc que $\mathbb{P}_x(\{k\})$ est donnée par la fréquence d'apparition de la modalité k dans x :

$$\mathbb{P}_x(\{k\}) = \frac{\text{effectifs de } k}{n} = \text{fréquence de } k.$$

La moyenne et la variance de la probabilité $\mathbb{P}_x$ sont alors données par :

Définition 1.3. *Soit $x = (x_1, \dots, x_n)$ une série statistique. La moyenne empirique de x est la quantité*

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

Remarque 1.1. La moyenne est un paramètre de position, fortement sensible aux valeurs extrêmes. Si par exemple, dans une classe de 10 élèves, 9 ont obtenu la note de 9 à l'examen, et 1 a obtenu la note 19, alors la moyenne de la classe est à 10, et pourtant une seule personne a validé l'examen...

Définition 1.4. *Soit $x = (x_1, \dots, x_n)$ une série statistique. La variance empirique de x est*

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

Son écart-type est la racine de la variance, i.e.

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

Remarque 1.2. L'écart-type est une mesure de dispersion, elle quantifie en quelle que sorte l'écart de x par rapport à sa moyenne. On prend la racine de la variance pour une question d'homogénéité (si par exemple x mesure des kg, la variance mesure des kg^2, et donc l'écart-type est une mesure en kg).

Pour des questions d'estimation, on utilise souvent une version corrigée de la variance et de l'écart-type :

Définition 1.5. *La variance empirique corrigée de x est*

$$\hat{\sigma}_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Son écart-type corrigée est la racine de $\hat{\sigma}_x^2$, i.e.

$$\hat{\sigma}_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Proposition 1.1. *(Formule de Koenig) La variance empirique σ_x^2 de x s'exprime aussi par la formule :*

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

(moyenne des carrées − carré de la moyenne).

Démonstration. Il suffit de développer le carré dans la définition de σ_x^2. □

1.1.3. *Fonction de répartition empirique et quantiles.* Soit $x = (x_1, \dots, x_n)$ une série statistique et $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ sa statistique d'ordre.

Définition 1.6. *On définit F_x la fonction de répartition empirique de x par : pour tout $t \in \mathbb{R}$,*

$$F_x(t) = \frac{\text{Card}(\{i \in \{1, \dots, n\} \mid x_{(i)} \leq t\})}{n}.$$

On a donc que F_x est une fonction en escalier, croissante, nulle pour $t < x_{(1)}$, valant 1 pour $t \geq x_{(n)}$, et telle que $F_x(t) = \frac{k}{n}$ si $x_{(k)} \leq t < x_{(k+1)}$.

Pour y un réel, on note $\lceil y \rceil$ le plus petit entier supérieur ou égale à y, i.e. l'unique entier vérifiant $\lceil y \rceil - 1 < y \leq \lceil y \rceil$.

Définition 1.7. *Soit $x = (x_1, \dots, x_n)$ une série statistique. On définit le quantile empirique d'ordre α (avec $\alpha \in]0, 1[$), notée q_α par :*

- *si $n\alpha$ n'est pas un entier, on prend $q_\alpha = x_{(\lceil n\alpha \rceil)}$;*
- *si $n\alpha$ est un entier, on prend $q_\alpha = \frac{x_{(n\alpha)} + x_{(n\alpha+1)}}{2}$ (convention, non unique) ;*

Les quantiles permettent de séparer une série statistique ordonnée en intervalles consécutif contenant le même nombre de données. Par exemple,

– si $\alpha = \frac{1}{2}$, le quantile $q_{\frac{1}{2}}$ est appelée la médiane ;

– si $\alpha = \frac{1}{4}$, le quantile $q_{\frac{1}{4}}$ est appelée le premier quartile ;

– si $\alpha = \frac{1}{4}$, le quantile $q_{\frac{3}{4}}$ est appelée le troisième quartile.

On a alors que le quantile d'ordre α vérifie :

$$F_x(q_\alpha) \geq \alpha \quad \text{et} \quad 1 - F_x(q_\alpha^-) \geq 1 - \alpha.$$

Exemple 1.1. Soit la série statistique $1, 1, 3, 5, 8, 12$, de taille $n = 6$. La médiane vaut $\frac{3+5}{2} = 4$, le premier quartile $x_2 = 1$, le dernier quartile $x_5 = 8$. On peut déterminer graphiquement les quantiles sur la fonction de répartition empirique. Pour cela, on trace une ligne horizontal à l'ordonnée correspond à l'ordre α : si on tombe sur un saut, on prend la valeur de la série immédiatement au dessus, si on tombe sur un palier, on prend la moyenne des deux valeurs, voir figure 1.1.

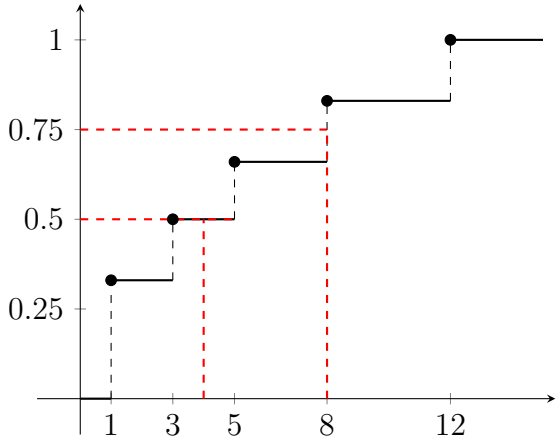

FIGURE 1.1. Fonction de répartition empirique de la série statistique $x = (1, 1, 3, 5, 8, 12)$ et quantiles $q_{1/2}$ et $q_{3/4}$.

1.1.4. *Représentations graphiques.* Il existe d'autres façons de représenter des données. Donnons les deux façons les plus usuelles.

1.1.5. *Boîtes à moustaches.* Une boîte à moustache est une représentation graphique résumant quelques indicateurs de position du caractère étudié. Il s'agit de tracer un rectangle allant du premier quartile au troisième quartile et coupé par la médiane. On ajoute alors des segments (les « moustaches ») partant des extrémités de la boîte jusqu'aux premier et neuvième décile. Les valeurs extrêmes en dehors de ces deux déciles sont représentées par des points.

Exemple 1.2. Une biologiste étudie la taille des pétales (en cm) d'une espèce de fleur rare. Elle prélève un échantillon de 12 fleurs et mesure la longueur de leurs pétales. Elle obtient les résultats suivants :

$$0.3, 1.2, 1.8, 2.0, 2.5, 2.8, 3.0, 3.1, 3.3, 3.5, 3.7, 3.9$$

On vérifie alors (exercice) que la médiane vaut 2.9, le 1^{er} quartile $q_{1/4} = 1.9$, le 3^{e} quartile $q_{3/4} = 3.4$, et les 1^{er} et 9^{e} déciles valent $q_{1/10} = 1.2$ et $q_{9/10} = 3.7$. Ces indicateurs peuvent alors être résumés par la boîte à moustaches ci-dessous.

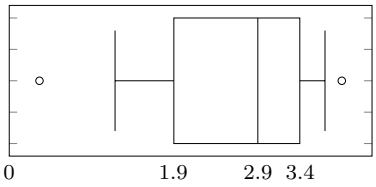

FIGURE 1.2. Boîte à moustache de la série statistique ci-dessus.

1.1.6. *Histogramme.* Un histogramme est une représentation graphique permettant de représenter la répartition empirique d'une série statistique. Le principe est de regrouper les observation proches en classes $(C_j)_{1\leq j\leq J}$ avec $C_0 = [b_0, b_1]$, et $C_j =]b_{j-1}, b_j]$, où $b_0 = x_{(1)}$, $b_J = x_{(n)}$ et $(C_j)_{1\leq j\leq J}$ formant une partition de $[x_{(1)}, x_{(n)}]$.

On représente alors la classe C_j par un rectangle de base $b_j - b_{j-1}$ et dont l'aire est égale à la fréquence de la classe C_j, c'est-à-dire

$$\frac{1}{n}\,\mathrm{Card}(\{i \in \{1,\dots,n\} \mid x_i \in C_j\}).$$

Par exemple, on souhaite analyser le temps de trajet domicile-université des étudiant·e·s du cours de probabilité. On a obtenu la série statistique de tableau des effectifs donné par :

Temps de trajet (min)	Effectif
$[5, 10]$	5
$]10, 15]$	35
$]15, 20]$	50
$]20, 25]$	30
$]25, 30]$	15

L'histogramme représentant cette série statistique est donné Fig. 1.3.

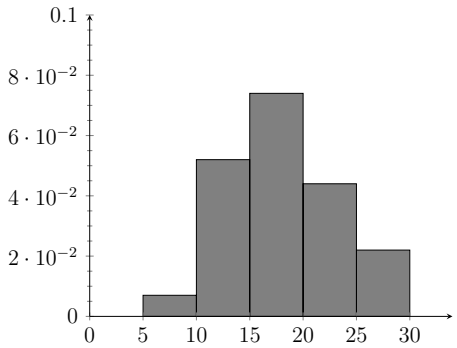

FIGURE 1.3. Exemple d'histogramme.

1.2. **Statistique inférentielle.** En statistique inférentielle, on suppose que les observations issues d'une série statistique $x = (x_1, \dots, x_n)$ sont en fait des réalisations de variables aléatoires $(X_1, \dots, X_n)$ définies sur un espace de probabilité $(\Omega, \mathbb{P})$, c'est-à-dire que $(x_1, \dots, x_n) = (X_1(\omega), \dots, X_n(\omega))$, pour un certain $\omega \in \Omega$. Le but est alors d'estimer certains paramètres d'intérêt inconnus.

Définition 1.8. *Soit X une variable aléatoire. Un échantillon de taille n de la loi de X est la donnée de n variables aléatoires $(X_1, \dots, X_n)$ définies sur le même espace de probabilité, indépendantes et de même loi que X.*

Un échantillon observé est une réalisation de l'échantillon $(X_1, \dots, X_n)$*, c'est-à-dire un* n*-uplet* $(x_1, \dots, x_n)$ *tel que* $(x_1, \dots, x_n) = (X_1(\omega), \dots, X_n(\omega))$ *pour un certain* $\omega \in \Omega$*.*

On appellera statistique toute fonction de l'échantillon $(X_1, \dots, X_n)$*.*

Une statistique n'est rien d'autre qu'une variable aléatoire, il s'agit essentiellement d'un changement de vocabulaire pour insister sur le fait que l'on ne connaît à priori pas la loi des variables X_i, et que l'on va se servir de ces variables pour estimer des paramètres d'intérêt.

La loi de X est donc supposée inconnue, et dépend d'un paramètre d'intérêt tel que la moyenne. En statistique inférentielle, on cherche à estimer un paramètre inconnu à l'aide d'un échantillon.

Commençons par quelques résultats.

2. Théorèmes limites

2.1. Inégalités de concentration.

Proposition 2.1 (Inégalité de Markov)**.** *Soit* X *une v.a. positive. Alors, pour tout* $a > 0$*, on a*

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

Démonstration. Rappelons la définition de la fonction indicatrice $\mathbb{1}_A$ d'un ensemble A :

$$\mathbb{1}_A(x) = \begin{cases} 1 & \text{si } x \in A \\ 0 & \text{sinon.} \end{cases}$$

On a alors :

$$a\mathbb{1}_{\{X \geq a\}} \leq X\mathbb{1}_{\{X \geq a\}} \leq X.$$

Donc en prenant l'espérance (qui est croissante), on obtient,

$$a\,\mathbb{E}\left(\mathbb{1}_{\{X \geq a\}}\right) \leq \mathbb{E}(X),$$

c'est-à-dire,

$$a\,\mathbb{P}(X \geq a) \leq \mathbb{E}(X). \qquad \square$$

Proposition 2.2 (Inégalité de Bienaymé-Tchebychev)**.** *Soit* X *une v.a. telle que* $\mathbb{E}(X^2) < +\infty$*. Alors, pour tout* $a > 0$*, on a*

$$\mathbb{P}\left(|X - \mathbb{E}(X)| \geq a\right) \leq \frac{\mathbb{V}\mathrm{ar}(X)}{a^2}.$$

Démonstration. Il suffit d'appliquer l'inégalité de Markov à la v.a. positive $|X - \mathbb{E}(X)|^2$, et de remarquer que

$$\{|X - \mathbb{E}(X)| \geq a\} = \left\{|X - \mathbb{E}(X)|^2 \geq a^2\right\}. \qquad \square$$

2.1.1. *Chernoff.* L'inégalité de Chernoff est de la forme suivante. Soit X une variable aléatoire, ayant des moments de tout ordre, c'est-à-dire $\mathbb{E}(|X|^k) < \infty$, pour tout $k \geq 0$. Soit $a > 0$. Alors pour tout $t > 0$, en utilisant le fait que l'exponentielle est bijective, on a

$$\mathbb{P}(X \geq a) = \mathbb{P}(tX \geq ta) = \mathbb{P}(e^{tX} \geq e^{ta}).$$

On peut alors appliquer l'inégalité de Markov, ce qui donne, que pour tout $t > 0$,

$$\mathbb{P}(X \geq a) \leq e^{-ta}\,\mathbb{E}\left(e^{tX}\right).$$

Le terme de gauche ne dépendant pas de t, on peut alors optimiser en t le terme de droite, ce qui donne une borne de type exponentielle à la probabilité que X dépasse la

valeur a. C'est donc une amélioration considérable par rapport à l'inégalité de Bienaymé-Tchebychev par exemple. En contre partie, ce n'est pas toujours facile d'expliciter une telle borne en fonction de quelle loi suit X, même le cas de la loi binomiale n'est pas des plus simple. On va montrer la proposition suivante :

Proposition 2.3. *Soit $X_1, \dots, X_n$ des v.a. indépendantes et identiquement distribuées de même loi de Bernoulli $\mathcal{B}(p)$. Soit $S_n = \sum_{i=1}^n X_i$, et posons $\mu = \mathbb{E}(S_n)$. Alors,*

(i) pour tout $\delta > 0$, on a

$$\mathbb{P}(S_n \geq (1+\delta)\mu) \leq e^{-\mu\delta^2/(2+\delta)},$$

(ii) pour tout $0 < \delta < 1$, on a

$$\mathbb{P}(S_n \leq (1-\delta)\mu) \leq e^{-\mu\delta^2/2}.$$

Corollaire 2.1. *Soit $X_1, \dots, X_n$ des v.a. indépendantes et identiquement distribuées de même loi de Bernoulli $\mathcal{B}(p)$. Soit $S_n = \sum_{i=1}^n X_i$, et posons $\mu = \mathbb{E}(S_n)$. Pour tout $0 < \delta < 1$, on a*

$$\mathbb{P}(|S_n - \mu| \geq \delta\mu) \leq 2e^{-\mu\delta^2/3}.$$

Démonstration. On a

$$\{|S_n - \mu| \geq \delta\mu\} = \{S_n \geq (1+\delta)\mu\} \cup \{S_n \leq (1-\delta)\mu\},$$

et comme une probabilité est croissante, on a donc

$$\mathbb{P}\left(|S_n - \mu| \geq \delta\mu\right) \leq \mathbb{P}\left(S_n \geq (1+\delta)\mu\right) + \mathbb{P}\left(S_n \leq (1-\delta)\mu\right).$$

Il suffit alors d'appliquer la proposition précédente (avec $\delta < 1$). □

Démonstration de la Proposition 2.3. On montre la borne supérieur (i), la preuve pour la borne inférieure étant analogue. Soit $t > 0$. Alors,

$$\mathbb{E}(e^{tS_n}) = \mathbb{E}\left(\prod_{i=1}^n e^{tX_i}\right) = \prod_{i=1}^n \mathbb{E}\left(e^{tX_i}\right) = \left(\mathbb{E}\left(e^{tX_1}\right)\right)^n,$$

car les variables $X_1, \dots, X_n$ sont indépendantes et identiquement distribuées. Comme X_1 est de loi $\mathcal{B}(p)$, on a $\mathbb{E}(e^{tX_1}) = pe^t + 1 - p$, ce qui donne :

$$\mathbb{E}(e^{tS_n}) = \left(pe^t + 1 - p\right)^n.$$

En utilisant l'inégalité de Markov, on obtient que pour tout $t > 0$,

$$\mathbb{P}(S_n \geq (1+\delta)\mu) = \mathbb{P}(e^{tS_n} \geq e^{t(1+\delta)\mu}) \leq e^{-t(1+\delta)\mu}\left(pe^t + 1 - p\right)^n.$$

On utilise alors l'inégalité $1 + x \leq e^x$, valable pour $x \geq 0$, appliqué à $x = p(e^t - 1)$. On obtient ainsi :

$$\mathbb{P}(S_n \geq (1+\delta)\mu) \leq e^{-t(1+\delta)\mu} e^{\mu(e^t-1)},$$

puisque $\mu = \mathbb{E}(S_n) = n\,\mathbb{E}(X_1) = np$. On cherche alors à minimiser en t l'expression de droite. En dérivant par rapport à t l'expression dans l'exponentielle, on obtient que le minimum est atteint en $t = \log(1+\delta)$. Ainsi,

$$\mathbb{P}(S_n \geq (1+\delta)\mu) \leq e^{\mu(\delta - (1+\delta)\log(1+\delta))}.$$

On utilise maintenant l'inégalité,

$$\log(1+x) \geq \frac{x}{1 + x/2},$$

valable pour $x > 0$ (exercice). Ceci donne

$$\delta - (1+\delta)\log(1+\delta) \leq -\frac{\delta^2}{2+\delta},$$

et finalement,

$$\mathbb{P}(S_n \geq (1+\delta)\mu) \leq e^{-\mu\delta^2/(2+\delta)},$$

ce qui est bien l'inégalité cherchée.

Pour la borne inférieur (ii), on fait essentiellement la même chose, cette fois le minimum est atteint pour $t = \log(1-\delta)$ et on utilise l'inégalité

$$\log(1-x) \geq -x + \frac{x^2}{2},$$

pour $0 < x < 1$. Les détails sont laissés en exercice. □

Exemple 2.1. On lance 100 fois une pièce de monnaie non truquée. Soit X le nombre de fois où on est tombé sur piles. La v.a. X suit donc la loi binomiale $\mathcal{B}(100, \frac{1}{2})$ avec $\mathbb{E}(X) = 50$ et $\mathbb{V}\mathrm{ar}(X) = 25$. On cherche à borner la probabilité d'obtenir au moins 75 piles. On a alors :

- Par l'inégalité de Markov, $\mathbb{P}(X \geq 75) \leq \frac{50}{75} = \frac{2}{3}$.
- Par l'inégalité de Bienaymé-Tchebchev, $\mathbb{P}(X \geq 75) = \mathbb{P}(X - 50 \geq 25) \leq \mathbb{P}(|X - 50| \geq 25) \leq \frac{25}{25^2} = 0.04$.
- Par l'inégalité de Chernoff, $\mathbb{P}(X \geq 75) = \mathbb{P}(X \geq (1+\frac{1}{2})50) \leq e^{-50/12} \approx 0.016$.

L'inégalité de Chernoff est donc meilleure (mais un logiciel de calcul donne, en utilisant la loi binomiale, que $\mathbb{P}(X \geq 75) \approx 2.8 \times 10^{-7}$...)

2.2. **Loi des grands nombres.**

Proposition 2.4 (Loi faible des grands nombres (LGN))**.** *Soit $(X_i)_{i\geq 1}$ une suite de v.a. i.i.d. telle que $\mathbb{E}|X_1|^2 < \infty$. On pose $\overline{X}_n = \frac{X_1+\cdots+X_n}{n}$. Alors, pour tout $\varepsilon > 0$,*

$$\mathbb{P}\left(\left|\overline{X}_n - \mathbb{E}(X_1)\right| > \varepsilon\right) \underset{n\to\infty}{\longrightarrow} 0.$$

On dit que $\overline{X}_n$ converge en probabilité vers $\mathbb{E}(X)$.

Démonstration. Par l'inégalité de Bienaymé-Tchebychev, on a

$$\mathbb{P}\left(\left|\overline{X}_n - \mathbb{E}(X_1)\right| > \varepsilon\right) \leq \frac{\mathbb{V}\mathrm{ar}(\overline{X}_n)}{\varepsilon^2}.$$

Or

$$\mathbb{V}\mathrm{ar}(\overline{X}_n) = \frac{1}{n^2}\sum_{i=1}^{n} \mathbb{V}\mathrm{ar}(X_i) = \frac{\mathbb{V}\mathrm{ar}(X_1)}{n},$$

la première égalité venant de l'indépendance des v.a., et la deuxième du fait qu'elles soient identiquement distribuées. Ainsi,

$$\mathbb{P}\left(\left|\overline{X}_n - \mathbb{E}(X_1)\right| > \varepsilon\right) \leq \frac{\mathbb{V}\mathrm{ar}(X_1)}{n\varepsilon^2} \underset{n\to\infty}{\longrightarrow} 0.$$ □

La loi des grands nombres nous dit donc que quand n est grand, la moyenne empirique $\overline{X}_n$ d'un échantillon $(X_1, \ldots, X_n)$ prend des valeurs qui ne s'écartent pas trop de la vraie moyenne $\mathbb{E}(X_1)$, avec grande probabilité. Ainsi, la moyenne empirique est un estimateur convergeant (en probabilité) de la moyenne. Si par exemple, $(X_1, \ldots, X_n)$ sont des v.a. de Bernoulli de paramètre p, représentant le résultat d'une suite de n lancers indépendants d'une pièce de monnaie truquée dont la probabilité de tomber sur pile est p, alors

$$\frac{X_1 + \cdots + X_n}{n} \approx p$$

avec grande probabilité. Estimer la probabilité de tomber sur pile, revient donc à lancer un grand nombre de fois la pièce, et de compter le nombre de fois où on est tombé sur pile.

2.3. **Théorème central limite.** On admettra le théorème suivant, qui se sera vu en toute généralité l'année prochaine.

Théorème 2.1 (Théorème central limite)**.** *Soit $(X_i)_{i\geq 1}$ une suite de variables aléatoires indépendantes et identiquement distribuées, de loi $\mathcal{B}(p)$. On pose $\overline{X}_n = \frac{X_1+\cdots+X_n}{n}$. Alors, pour tout $a < b$, on a*

$$\mathbb{P}\left(a \leq \frac{\sqrt{n}}{\sqrt{p(1-p)}}\left(\overline{X}_n - p\right) \leq b\right) \underset{n\to\infty}{\longrightarrow} \int_a^b \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx.$$

On dit que la v.a. $\frac{\sqrt{n}}{\sqrt{p(1-p)}}\left(\overline{X}_n - p\right)$ converge en loi vers une loi gaussienne (ou normale) centrée (c'est-à-dire de moyenne 0) réduite (c'est-à-dire de variance 1). Ce théorème porte en fait le nom de théorème de Moivre-Laplace, le théorème central limite en est sa généralisation à toute suite de v.a. i.i.d. ayant un moment d'ordre 2. On sait par la loi des grands nombres que $\overline{X}_n$ converge en probabilité vers p. Ce théorème nous dit alors que $\overline{X}_n$ converge vers p à la vitesse $\frac{1}{\sqrt{n}}$ et que les fluctuations de $\overline{X}_n$ autour de p sont données par une loi gaussienne.

Esquisse de la démonstration. La v.a. $S_n = X_1 + \ldots + X_n$ est une somme de n v.a. i.i.d. de loi de Bernoulli $\mathcal{B}(p)$, donc suit la loi $\mathcal{B}(n,p)$. Le terme de gauche dans l'énoncé du théorème est donc, en posant $q = 1-p$,

$$\begin{aligned}\mathbb{P}\left(a \leq \frac{\sqrt{n}}{\sqrt{p(1-p)}}\left(\overline{X}_n - p\right) \leq b\right) &= \sum_{a\leq x\leq b} \mathbb{P}\left(\frac{S_n - np}{\sqrt{npq}} = x\right) \\ &= \sum_{a\leq x_k\leq b} \mathbb{P}(S_n = k)\end{aligned}$$

où la somme porte sur les x_k de la forme

$$x_k = \frac{k-np}{\sqrt{npq}}.$$

Or $\mathbb{P}(S_n = k) = \binom{n}{k} p^k q^{n-k}$. On utilise alors la formule de Stirling pour donner un équivalent du coefficient binomial :

$$n! \sim \sqrt{2\pi n} n^n e^{-n}, \quad \text{quand } n \to \infty.$$

On a donc

$$\begin{aligned}\binom{n}{k} p^k q^{n-k} &\sim \frac{\sqrt{2\pi} n^{n+1/2} e^{-n}}{\sqrt{2\pi} k^{k+1/2} e^{-k} \sqrt{2\pi} (n-k)^{n-k+1/2} e^{-(n-k)}} p^k q^{n-k} \\ &= \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{n\frac{k}{n}\left(1-\frac{k}{n}\right)}} \left(\frac{p}{k/n}\right)^k \left(\frac{q}{1-k/n}\right)^{n-k} \\ &\sim \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{npq}} \varphi(n,k),\end{aligned}$$

avec

$$\varphi(n,k) = \left(\frac{np}{k}\right)^k \left(\frac{nq}{n-k}\right)^{n-k}$$

Rappelons le d.l. du logarithme : $\log(1+u) = u - \frac{u^2}{2} + o(u^3)$, quand $u \to 0$. Alors, quand $n \to \infty$,

$$\log\left[\left(\frac{np}{k}\right)^k\right] = k\log\left(\frac{np}{k}\right) = k\log\left(1 - \frac{\sqrt{npq}x_k}{k}\right)$$
$$= -\sqrt{npq}x_k - \frac{npqx_k^2}{2k} + \cdots$$

et de même,

$$\log\left[\left(\frac{nq}{n-k}\right)^{n-k}\right] = (n-k)\log\left(1 + \frac{\sqrt{npq}x_k}{n-k}\right)$$
$$= \sqrt{npq}x_k - \frac{npqx_k^2}{2(n-k)} + \cdots$$

Ceci donne que

$$\log\varphi(n,k) = -\frac{npqx_k^2}{2k} - \frac{npqx_k^2}{2(n-k)} + \cdots$$
$$= -\frac{x_k^2}{2}\frac{n^2pq}{k(n-k)} + \cdots$$

Or $\frac{n^2pq}{k(n-k)} \sim 1$, on a donc que

$$\log\varphi(n,k) = -\frac{x_k^2}{2} + \cdots,$$

et que

$$\binom{n}{k}p^kq^{n-k} = \frac{1}{\sqrt{2\pi npq}}e^{-x_k^2/2} + \cdots$$

En prenant la somme sur x_k, on remarque que l'on obtient une somme de Riemann (car $x_{k+1} - x_k = \frac{1}{\sqrt{npq}}$) qui converge, quand $n \to \infty$, vers

$$\int_a^b \frac{1}{\sqrt{2\pi}}e^{-x^2/2}dx.$$

Pour rendre rigoureuse cette preuve, il faut s'assurer que la somme des termes d'erreurs (les $\cdots$) donne une contribution nulle à la limite (ce qui est un peu pénible...). □

3. Estimation

3.1. **Estimation ponctuelle.** Soit $(X_1, \ldots, X_n)$ un échantillon d'une loi inconnue dépendant d'un paramètre θ (le plus souvent la moyenne).

Définition 3.1. *Soit $(X_1, \ldots, X_n)$ un échantillon d'une loi inconnue dépendant d'un paramètre θ. On appelle estimateur de θ toute fonction de l'échantillon $(X_1, \ldots, X_n)$. On le note usuellement $\widehat{\theta}$.*

Définition 3.2. *Soit $\hat{\theta}$ un estimateur d'un paramètre inconnu θ. On dit que $\widehat{\theta}$ est sans biais si $\mathbb{E}(\hat{\theta}) = \theta$. Dans le cas contraire, on dit que $\widehat{\theta}$ est biaisé.*

On cherche la plupart du temps à avoir un estimateur sans biais. Mais par exemple, $\frac{X_1+X_2}{2}$ ou encore $\frac{X_1+X_2+X_3}{3}$ sont des estimateurs sans biais de m. On a donc besoin de définir la qualité d'un estimateur. On introduit alors la notion suivante.

Définition 3.3. *Soit $\hat{\theta}$ un estimateur d'un paramètre inconnu θ, tel que $\widehat{\theta}$ admet un moment d'ordre 2. On appelle erreur quadratique moyenne ou risque quadratique, notée* $\mathrm{MSE}(\widehat{\theta})$*, la quantité*

$$\mathrm{MSE}(\widehat{\theta}) = \mathbb{E}\left[(\widehat{\theta} - \theta)^2\right].$$

Proposition 3.1. *On a*

$$\mathrm{MSE}(\widehat{\theta}) = \mathbb{V}\mathrm{ar}(\widehat{\theta}) + \left(\mathbb{E}(\widehat{\theta}) - \theta\right)^2,$$

autrement dit, l'erreur quadratique moyenne d'un estimateur $\hat{\theta}$ est la somme de sa variance et du carré de son biais.

Démonstration. Il suffit d'écrire :

$$\begin{aligned}\mathrm{MSE}(\widehat{\theta}) = \mathbb{E}\left[(\widehat{\theta} - \theta)^2\right] &= \mathbb{E}\left[(\widehat{\theta} - \mathbb{E}(\widehat{\theta}) + \mathbb{E}(\widehat{\theta}) - \theta)^2\right] \\ &= \mathbb{V}\mathrm{ar}(\widehat{\theta}) + 2\underbrace{\mathbb{E}\left(\widehat{\theta} - \mathbb{E}(\widehat{\theta})\right)}_{=0} + \left(\mathbb{E}(\widehat{\theta}) - \theta\right)^2 \\ &= \mathbb{V}\mathrm{ar}(\widehat{\theta}) + \left(\mathbb{E}(\widehat{\theta}) - \theta\right)^2. \qquad \square\end{aligned}$$

On a donc que pour un estimateur sans biais $\widehat{\theta}$, l'erreur quadratique moyenne est égale à sa variance.

3.1.1. *Estimation ponctuelle d'une moyenne.*

Définition 3.4. *Soit $(X_1, \ldots, X_n)$ un échantillon d'une loi de moyenne θ. On appelle moyenne empirique de l'échantillon, notée $\overline{X}_n$, la statistique*

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

On a alors que la moyenne empirique est un estimateur sans biais de θ. En effet,

$$\mathbb{E}\left(\frac{X_1 + \cdots + X_n}{n}\right) = \frac{1}{n}\left(\mathbb{E}(X_1) + \cdots + \mathbb{E}(X_n)\right) = \mathbb{E}(X_1) = \theta,$$

où l'on a utilisé la linéarité de l'espérance et le fait que les v.a. X_i sont toutes de même loi de moyenne θ. De plus, par la loi des grands nombres, c'est un estimateur convergeant en probabilité de θ (on dit que c'est un estimateur consistant).

Si $(X_1, \ldots, X_n)$ est un échantillon de la loi de Bernoulli de paramètre p, la moyenne étant égale à p, on parle aussi de fréquence empirique.

3.1.2. *Estimation ponctuelle de la variance.*

Définition 3.5. *Soit $(X_1, \ldots, X_n)$ un échantillon. On appelle variance empirique la statistique*

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

En développant le carré, on a immédiatement la formule alternative :

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

Remarque 3.1. Soit $(X_1, \ldots, X_n)$ un échantillon de la loi de Bernoulli $\mathcal{B}(p)$. Alors, la variance empirique est égale à

$$S_n^2 = \overline{X}_n(1 - \overline{X}_n).$$

En effet, comme X_i vaut 0 ou 1, on a $X_i^2 = X_i$, et donc

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

La formule se comprend aisément : $\overline{X}_n$ étant un estimateur de p, et la variance de la loi $\mathcal{B}(p)$ étant $p(1-p)$, la statistique $\overline{X}_n(1-\overline{X}_n)$ est naturellement un estimateur de $p(1-p)$.

Proposition 3.2. *Soit $(X_1, \dots, X_n)$ un échantillon. On note σ^2 la variance de X_1. La variance empirique est un estimateur biaisé de la variance :*

$$\mathbb{E}\left(S_n^2\right) = \frac{n-1}{n}\sigma^2.$$

Démonstration. On a :

$$\begin{aligned}\mathbb{E}\left(S_n^2\right) &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left(X_i^2\right) - \mathbb{E}\left(\overline{X}_n^2\right)\\ &= \mathbb{E}\left(X_1^2\right) - \mathbb{E}\left(\overline{X}_n^2\right)\end{aligned}$$

On calcule alors :

$$\mathbb{E}\left(\overline{X}_n^2\right) = \frac{1}{n^2}\sum_{i,j=1}^{n}\mathbb{E}(X_iX_j) = \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}\left(X_i^2\right) + \frac{1}{n^2}\sum_{1\le i\neq j\le n}\mathbb{E}(X_iX_j).$$

Comme X_i et X_j sont indépendantes pour $i \neq j$, on a $\mathbb{E}(X_iX_j) = \mathbb{E}(X_i)\,\mathbb{E}(X_j) = (\mathbb{E}(X_1))^2$. On obtient donc

$$\mathbb{E}\left(\overline{X}_n^2\right) = \frac{1}{n}\,\mathbb{E}(X_1^2) + \frac{n-1}{n}m^2,$$

et finalement,

$$\begin{aligned}\mathbb{E}\left(S_n^2\right) &= \mathbb{E}\left(X_1^2\right) - \frac{1}{n}\,\mathbb{E}\left(X_1^2\right) - \frac{n-1}{n}m^2\\ &= \frac{n-1}{n}\,\mathbb{V}\mathrm{ar}(X_1).\end{aligned}$$

□

On a donc que S_n^2 sous estime (légèrement) la variance. Ceci amène alors à définir l'estimateur suivant.

Définition 3.6. *Soit $(X_1, \dots, X_n)$ un échantillon. On appelle variance empirique corrigée la statistique*

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

C'est un estimateur sans biais de $\sigma^2 = \mathbb{V}\mathrm{ar}(X_1)$.

3.2. **Intervalles de confiance.** Un intervalle de confiance est un intervalle *aléatoire*, c'est-à-dire dont les bornes sont aléatoires et en général construites à partir d'un estimateur, et qui a une grande probabilité de contenir un paramètre que l'on cherche à estimer. On parle alors d'estimation par intervalle. La notion d'intervalle de confiance permet ainsi de définir une marge d'erreur plutôt que d'utiliser une simple estimation ponctuelle. Plus précisément :

Définition 3.7. *Soit $(X_1, \dots, X_n)$ un échantillon et θ un paramètre inconnu qu'on cherche à estimer. Un intervalle de confiance pour θ de niveau $1-\alpha$ (ou de risque α) est un intervalle de la forme $I = [a(X_1, \dots, X_n), b(X_1, \dots, X_n)]$ vérifiant*

$$\mathbb{P}(I \ni \theta) \ge 1-\alpha.$$

(Lire : la probabilité que I contienne θ est d'au moins $1-\alpha$.)

Le plus souvent, on recherche un intervalle de confiance de θ sous la forme d'un intervalle centré en une estimation ponctuelle de θ.

Par exemple, on cherche à estimer une proportion p. Soit $(X_1, \dots, X_n)$ un échantillon de la loi de Bernoulli $\mathcal{B}(p)$. On sait que la moyenne empirique $\overline{X}_n$ est un estimateur sans biais de p. On peut construire un intervalle de confiance pour p par l'inégalité de Bienaymé-Tchebychev. En effet, on a

$$\mathbb{P}\left(|\overline{X}_n - p| > a\right) \leq \frac{\mathbb{V}\mathrm{ar}(\overline{X}_n)}{a^2} = \frac{\mathbb{V}\mathrm{ar}(X_1)}{na^2}.$$

En posant $\alpha = \frac{\mathbb{V}\mathrm{ar}(X_1)}{na^2}$, on obtient

$$\mathbb{P}\left(|\overline{X}_n - p| \leq \sqrt{\frac{\mathbb{V}\mathrm{ar}(X_1)}{n\alpha}}\right) \geq 1-\alpha,$$

et donc

$$\mathbb{P}\left(\left[\overline{X}_n - \sqrt{\frac{\mathbb{V}\mathrm{ar}(X_1)}{n\alpha}}, \overline{X}_n + \sqrt{\frac{\mathbb{V}\mathrm{ar}(X_1)}{n\alpha}}\right] \ni p\right) \geq 1-\alpha.$$

On obtient ainsi un intervalle de confiance de p au niveau de confiance $1-\alpha$, non-asymptotique, c'est-à-dire valable pour tout $n \geq 1$. Le problème ici, est que $\mathbb{V}\mathrm{ar}(X_1)$ est à priori inconnu et dépend de p... Une première possibilité est de remarquer que comme $\mathbb{V}\mathrm{ar}(X_1) = p(1-p)$, on a $\mathbb{V}\mathrm{ar}(X_1) \leq \frac{1}{4}$ (une simple étude de fonction permet de le vérifier rapidement), on obtient donc un intervalle plus large en remplaçant $\mathbb{V}\mathrm{ar}(X_1)$ par $\frac{1}{4}$, ce qui donne

$$\mathbb{P}\left(\left[\overline{X}_n - \frac{1}{2\sqrt{n\alpha}}, \overline{X}_n + \frac{1}{2\sqrt{n\alpha}}\right] \ni p\right) \geq 1-\alpha.$$

Ainsi,

$$\left[\overline{X}_n - \frac{1}{2\sqrt{n\alpha}}, \overline{X}_n + \frac{1}{2\sqrt{n\alpha}}\right]$$

est un intervalle de confiance de p au niveau de confiance $1-\alpha$, valable pour tout $n \in \mathbb{N}$. Il est en général assez mauvais car bien trop large, et son risque réel peut être bien inférieur au risque α choisi.

Exemple 3.1. On chercher à estimer la proportion p d'électeurs du candidat A à la dernière élection présidentielle, avec une erreur d'au plus 1% et au risque 5%. On a donc que l'amplitude de notre intervalle de confiance doit être de 0.01 et donc la taille de notre échantillon doit satisfaire

$$\frac{1}{2\sqrt{0.05n}} \leq 0.01$$

ce qui donne $n \geq 50000$, ce qui est difficilement envisageable...

Une autre possibilité est d'utiliser un estimateur de la variance. On a vu que la variance empirique S_n^2 qui dans le cas d'une proportion se réécrit $S_n^2 = \overline{X}_n(1-\overline{X}_n)$ est un estimateur naturel de la variance. On peut même montrer par la loi des grands nombres que c'est un estimateur consistant. On remplace alors dans l'intervalle de confiance ci-dessus $\mathbb{V}\mathrm{ar}(X_1)$ par son estimateur S_n^2, et ainsi

$$\left[\overline{X}_n - \sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n\alpha}}, \overline{X}_n + \sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n\alpha}}\right]$$

est un intervalle de confiance de p au niveau $1-\alpha$. Il est cette fois-ci asymptotique, c'est-à-dire valable uniquement pour n assez grand.

Exemple 3.2. On effectue un sondage à la sortie des urnes sur 1000 électeurs, dont 520 on répondu ayant voté pour le candidat M. (pour simplifier, on ne tient pas compte des votes blancs). On a donc une réalisation d'un échantillon de taille 1000 de la loi de Bernoulli de paramètre p, la vraie proportion d'individus ayant voté pour M. Une observation de la moyenne empirique $\overline{X}_n$ est donc

$$\overline{x}_n = \frac{520}{1000} = 0.52.$$

Au risque 5%, un intervalle de confiance observé est donc

$$\left[0.52 - \sqrt{\frac{0.48 \times 0.52}{1000 \times 0.05}}, 0.52 + \sqrt{\frac{0.48 \times 0.52}{1000 \times 0.05}}\right] = [0.45, 0.59].$$

Peut-on prédire qui va gagner l'élection ?

Enfin, le théorème central limite nous donne de bien meilleurs intervalles de confiance, mais uniquement pour une taille n d'échantillon assez grande, on parle alors d'intervalle de confiance asymptotique. En prenant $a = -b$ dans le théorème central limite, on cherche donc z_α tel que

$$\int_{-z_\alpha}^{z_\alpha} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = 1 - \alpha.$$

Retenons que

$$\int_{-\infty}^{1.96} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx \approx 0.975,$$

ce qui donne que $z_{0.05} = 1.96$.

Un intervalle de confiance asymptotique de niveau 95% pour p est alors :

$$\left[\overline{X}_n - \frac{1.96}{2\sqrt{n}}, \overline{X}_n + \frac{1.96}{2\sqrt{n}}\right]$$

Dans l'exemple précédent, on obtient alors comme intervalle observé

$$[0.49, 0.55]$$

ce qui est bien meilleur ! Mais sans doute pas suffisant pour prédire l'issue de l'élection...

3.3. **Intervalles de fluctuation.** Pour finir ce chapitre, on donne la définition d'intervalles de fluctuations, ce qui va nous permettre de faire une petite introduction aux tests statistiques.

Un intervalle de fluctuation est un intervalle *non-aléatoire* dans lequel un estimateur a une grande probabilité de tomber. Plus précisément,

Définition 3.8. *Soit S une variable aléatoire et $\alpha \in [0, 1]$ Un intervalle de fluctuation de S au seuil $1 - \alpha$ est un intervalle $[a, b]$ (avec $a < b$) telle que*

$$\mathbb{P}(S \in [a, b]) \geq 1 - \alpha.$$

Un intervalle de fluctuation n'est bien entendu pas unique puisque si $[a, b] \subset [a', b']$ et que $[a, b]$ est un intervalle de fluctuation de S au seuil $1 - \alpha$, il en est de même pour l'intervalle $[a', b']$ par croissance d'une probabilité.

Remarque 3.2. Soit $\widehat{\theta}$ un estimateur d'un paramètre θ. Si $[\widehat{\theta} - \varepsilon, \widehat{\theta} + \varepsilon]$ est un intervalle de confiance de niveau $1-\alpha$ pour θ, alors il est évident que $[\theta-\varepsilon, \theta+\varepsilon]$ est un intervalle de fluctuation de $\widehat{\theta}$ de niveau $1 - \alpha$. On passe donc de l'un à l'autre dans cet exemple par un jeu de réécriture, mais ils ne représentent pas la même chose. Un intervalle de confiance est un intervalle aléatoire servant à donner une marge d'erreur dans l'estimation ponctuelle du paramètre d'intérêt θ, alors qu'un intervalle de fluctuation n'est pas aléatoire et est connu

à priori. Il va permettre de tester notre estimation, soit en testant la représentativité de notre échantillon, soit en testant une hypothèse de notre modélisation.

On peut ici encore utiliser l'inégalité de Bienaymé-Tchebychev pour donner un intervalle de fluctuation d'une proportion empirique, mais là encore, l'utilisation du théorème central limite est bien plus pertinente.

Définition 3.9. *Soit $(X_1, \dots, X_n)$ un échantillon de la loi $\mathcal{B}(p)$. Un intervalle de fluctuation asymptotique de $\overline{X}_n$ au niveau 95% est*

$$\left[p - 1.96\frac{\sqrt{p(1-p)}}{\sqrt{n}}, p + 1.96\frac{\sqrt{p(1-p)}}{\sqrt{n}} \right].$$

En majorant $p(1-p)$ par $\frac{1}{4}$, on pourra aussi utiliser l'intervalle simplifiée $[p-\frac{1}{\sqrt{n}}, p+\frac{1}{\sqrt{n}}]$, mais uniquement quand $p(1-p)$ n'est pas trop proche de 0 ou 1, on choisit usuellement $0.2 \leq p \leq 0.8$.

Exemple 3.3. On a lancé 1000 fois une pièce de monnaie supposée équilibrée, et a on obtenu 435 piles. La pièce est-elle truquée ? Faisons l'hypothèse que la pièce est équilibrée. Les résultats des lancers correspondent donc à un échantillon de 1000 v.a. de Bernoulli de paramètre $p = \frac{1}{2}$. Un intervalle de fluctuation de la proportion empirique de piles pour une pièce équilibrée est alors

$$[0.47, 0.53],$$

c'est-à-dire que la probabilité que $\overline{X}_{1000}$ soit dans cet intervalle est d'environ 0.95. Comme la valeur observée dans notre expérience est $\overline{x}_{1000} = 0.435$, il parait peu probable que la pièce soit réellement équilibrée ! On rejette donc l'hypothèse que la pièce est équilibrée, en prenant un risque de 5% de prendre la mauvaise décision.