Probabilités et statistique dans les nouveaux programmes de lycée

Valérie Robert [1]

[1] Docteure en statistique et professeure agrégée de mathématiques en CPGE

Plan

Sommaire

Qu'est-ce que la statistique ?

- Différence entre probabilités et statistique ?
  - probabilités : on se donne un modèle et on effectue des calculs à l'intérieur de ce modèle.
  - statistique : on observe des données et on tente de les expliquer en faisant éventuellement appel à un modèle probabiliste.
- Statistiques ou statistique ?
  - statistique : la "science" statistique, une branche des mathématiques.
  - statistiques : les résultats de calculs statistiques réalisés à partir d'un jeu de données.
- Société Française de Statistique (SFdS) : `www.sfds.asso.fr`
  - La compétition européenne de statistique à destination des lycéens

Vocabulaire utilisé en statistique (1/2)

- **Modélisation** :
  • "Art subtil de remplacer l'insaisissable réalité par un objet mathématique (loi de probabilité $\mathbb{P}_\theta$, ...) utilisable." C. Villani
  • "Tous les modèles sont faux, mais certains sont utiles." G. Box

- **Variable aléatoire** X : application définie sur l'ensemble des résultats possibles d'une expérience aléatoire. Loi de X : $\mathbb{P}_\theta(X = k \in \mathbb{Z})$ [discret]

- **Paramètre** θ : quantité intervenant dans la description d'un modèle (souvent espérance, variance).

- **Espérance de** X : la valeur que l'on s'attend à trouver, en moyenne, si l'on répète un grand nombre de fois la même expérience aléatoire. Elle se note $\mathbb{E}(X)$ ou μ. Indicateur de *position*.

- **Variance de** X : mesure de la *dispersion* des valeurs d'un échantillon ou d'une loi de probabilité. Se note $\mathbb{V}(X)$. L'écart-type $\sigma = \sqrt{\mathbb{V}(X)}$.

Vocabulaire utilisé en statistique (2/2)

- *n*-**échantillon** : ensemble de n individus ou observations représentatifs d'une population. Réalisation de n variables aléatoires $X_1, \cdots, X_n$ indépendantes et identiquement distribuées de même loi $\mathbb{P}_\theta$ que X.
  Ne pas confondre variable aléatoire (majuscule) et réalisation (minuscule) de cette variable.

- **Estimateur** $\widehat{\theta}$: une statistique basée sur des variables aléatoires permettant d'évaluer un paramètre θ inconnu. Une estimation est la réalisation de cet estimateur grâce au n-échantillon.
  Exemple 1. $\overline{X_n} = \frac{1}{n} \sum_{k=1}^{n} X_k$ est un estimateur de l'espérance. La fréquence observée f est une réalisation de cet estimateur.
  Exemple 2. $S^2 = \frac{1}{n} \sum_{k=1}^{n} (X_k - \overline{X_n})^2$ est un estimateur de la variance. La variance de la série observée est une réalisation de cet estimateur.

Démarche statistique

Une **étude statistique** a pour but de répondre à une question dans un domaine d'application particulier à partir d'un jeu de données :

- La réflexion sur le **protocole** à suivre pour le recueil des données (plan d'expériences, de sondage, élaboration d'un questionnaire ...) ;
- Le **recueil** et le codage des données (fichier informatique, base de données) ;
- L'**exploration des données** (statistique descriptive, analyse des données, fouille de données, ...), sans chercher à les modéliser ;
- Éventuellement, **prétraitement des données** (recodage, agrégation, transformations, création de nouvelles données .) ;
- Si les données sont issues d'un échantillon, **modélisation statistique** des données : inférence, appel à un modèle probabiliste ;

Sommaire

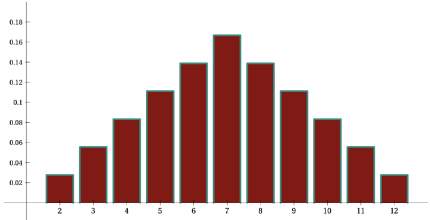
Loi d'une variable aléatoire

Soit $(\Omega, \mathbb{P})$) un espace probabilisé. Soit X une variable aléatoire. La loi de probabilité de X est la fonction souvent notée $\mathbb{P}_X$ définie par :

$$\mathbb{P}_X : \begin{cases} \mathcal{P}(X(\Omega)) & \to [0,1] \\ k & \mapsto \mathbb{P}_X(k) = \mathbb{P}(X = k) \end{cases}$$

On a nécessairement : $\displaystyle\sum_{k \in X(\Omega)} P(X = k) = 1.$

On lance deux dés bien équilibrés à 6 faces. Soit X la variable aléatoire donnant la somme des deux résultats obtenus. Représentation de la loi de X.



Loi de Bernoulli

- Une épreuve de Bernoulli est une expérience aléatoire qui n'admet que deux issues ; celles-ci sont généralement appelées ≪succès≫ (souvent notée S) et ≪échec≫ notée E ou $\overline{S}$.
- Une variable aléatoire X suit une loi de Bernoulli de paramètre $p \in [0, 1]$ si elle ne prenant que deux valeurs :
  • 1, le succès, avec une probabilité p ;
  • 0, l'échec, avec une probabilité $q = 1 - p$.
  On écrit alors : $X \hookrightarrow \mathcal{B}(p)$ ou $X \hookrightarrow \mathcal{B}(1, p)$
- Jeter une pièce et regarder si l'on obtient ≪ pile ≫ ou ≪ face ≫ est une épreuve de Bernoulli (la plus classique).
- Lancer un dé et regarder si l'on a obtenu un 1 ou non est également une épreuve de Bernoulli : il n'y a que deux issues : soit on a obtenu un 1, auquel cas, on a un succès, soit on n'a pas obtenu de 1.

Loi binomiale

- Un **schéma de Bernoulli** est la répétition indépendante d'une même épreuve de Bernoulli. On parle de schéma de Bernoulli de paramètres n et p où n désigne le nombre de répétitions de l'épreuve et p la probabilité de succès.

- Une variable aléatoire X suit une **loi binomiale de paramètre n et p** si elle compte le nombre de succès d'un schéma de Bernoulli. On écrit alors : $X \hookrightarrow \mathcal{B}(n, p)$.

  Remarque. La notation « $X \hookrightarrow \mathcal{B}(1, p)$ » pour dire qu'une variable aléatoire X suit une loi de Bernoulli vient du fait qu'une loi de Bernoulli est une loi binomiale de premier paramètre $n = 1$.

- On lance 50 fois un dé bien équilibré. Soit X le nombre de 3 obtenus. Déterminer la loi de X.

Indépendance deux à deux, mutuelle indépendance

Mutuelle indépendance. Soit $(\Omega, \mathbb{P})$ un espace probabilisé et $n \geq 2$. Soit $X_1, ..., X_n$ des variables aléatoires. Les assertions suivantes sont équivalentes :

(i) $X_1, ..., X_n$ sont mutuellement indépendantes ;

(ii) $\forall (k_1, ..., k_n) \in \mathbb{Z}^n$

$$\mathbb{P}(X_1 = k_1 \cap \cdots \cap X_n = k_n) = \mathbb{P}(X_1 = k_1) \times \cdots \times \mathbb{P}(X_n = k_n).$$

Indépendance deux à deux. Les assertions suivantes sont équivalentes :

(i) $X_1, ..., X_n$ sont deux à deux indépendantes ;
$\forall i \neq j$

$$\mathbb{P}(X_i = k_i \cap X_j = k_j) = \mathbb{P}(X_i = k_i) \times \mathbb{P}(X_j = k_j).$$

mutuelle indépendance $\underset{\not\Leftarrow}{\Rightarrow}$ indépendance deux à deux

Calculs d'espérance [Cas discret]

Espérance. Soit $(\Omega, \mathbb{P})$ un espace probabilisé fini. Soit X une variable aléatoire réelle sur Ω. L'espérance mathématique de X, notée $\mathbb{E}(X)$ ou $\mathbb{E}[X]$, est le réel défini par :

$$\mathbb{E}[X] = \sum_{x \in X(\Omega)} k\,\mathbb{P}(X = k).$$

Remarque. Une variable aléatoire réelle est centrée si $\mathbb{E}(X) = 0$.

(i) Si $a \in \mathbb{R}$, alors $\mathbb{E}[aX + bY] = a\mathbb{E}(X) + b\mathbb{E}(Y)$.

(ii) Si X et Y sont indépendantes, alors $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$.

(iii) [Théorème de transfert] Soit $f : X(\Omega) \to \mathbb{R}$ une fonction. On a :

$$\mathbb{E}[f(X)] = \sum f(k)\mathbb{P}(X = k).$$

Exemple. $\mathbb{E}(X^2) = ?$
$\mathbb{E}(a) = a$

Calculs d'espérance [Cas discret]

Espérance. Soit $(\Omega, \mathbb{P})$ un espace probabilisé fini. Soit X une variable aléatoire réelle sur Ω. L'espérance mathématique de X, notée $\mathbb{E}(X)$ ou $\mathbb{E}[X]$, est le réel défini par :

$$\mathbb{E}[X] = \sum_{x \in X(\Omega)} k\,\mathbb{P}(X = k).$$

Remarque. Une variable aléatoire réelle est centrée si $\mathbb{E}(X) = 0$.

(i) Si $a \in \mathbb{R}$, alors $\mathbb{E}[aX + bY] = a\mathbb{E}(X) + b\mathbb{E}(Y)$.

(ii) Si X et Y sont indépendantes, alors $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$.

(iii) [Théorème de transfert] Soit $f : X(\Omega) \to \mathbb{R}$ une fonction. On a :

$$\mathbb{E}[f(X)] = \sum f(k)\mathbb{P}(X = k).$$

Exemple. $\mathbb{E}(X^2) = \sum k^2\mathbb{P}(X = k)$
$\mathbb{E}(a) = a$

Calculs d'espérance [Cas discret]

Soit X une variable aléatoire réelle sur un espace probabilisé fini avec $X \hookrightarrow \mathcal{B}(1, p)$ avec $p \in [0, 1]$. Alors :

$$\mathbb{E}(X) = 0 \times (1 - p) + 1 \times p = p$$

Soit X une variable aléatoire réelle sur un espace probabilisé fini avec $X \hookrightarrow \mathcal{B}(n, p)$ avec $p \in [0, 1]$. Alors :

$$\mathbb{E}(X) = np$$

Soit $(X_1, \cdots, X_n)$ une suite de variables aléatoires indépendantes de même loi de Bernoulli de paramètre p.

$$\mathbb{E}\left(\frac{X_1 + \cdots + X_n}{n}\right) = \text{..........}$$

Calculs d'espérance [Cas discret]

Soit X une variable aléatoire réelle sur un espace probabilisé fini avec $X \hookrightarrow \mathcal{B}(1, p)$ avec $p \in [0, 1]$. Alors :

$$\mathbb{E}(X) = 0 \times (1 - p) + 1 \times p = p$$

Soit X une variable aléatoire réelle sur un espace probabilisé fini avec $X \hookrightarrow \mathcal{B}(n, p)$ avec $p \in [0, 1]$. Alors :

$$\mathbb{E}(X) = np$$

Soit $(X_1, \cdots, X_n)$ une suite de variables aléatoires indépendantes de même loi de Bernoulli de paramètre p.

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

Calculs de variance [Cas discret]

Soit $(\Omega, \mathbb{P})$ un espace probabilisé fini. Soit X une variable aléatoire sur Ω.

- La variance de X, notée $\mathbb{V}(X)$, est le réel défini par :

$$\mathbb{V}(X) = \mathbb{E}(X - \mathbb{E}(X))^2 = \sum (k - \mathbb{E}(X))^2 \, \mathbb{P}(X = k).$$

- L'écart-type de X, noté $\sigma(X)$, est défini par :

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

Remarque. Une variable aléatoire réelle est réduite si $\mathbb{V}(X) = 1$.

Égalité de Koenig-Huyens.

$$\mathbb{V}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2.$$

Soit a et $b \in \mathbb{R}$. Alors :
(i) $\mathbb{V}(aX + b) = a^2 \mathbb{V}(X)$
(ii) $\sigma(aX + b) = |a| \sigma(X)$
(ii) Si X et Y sont indépendantes, alors $\mathbb{V}[X + Y] = \mathbb{V}[X] + \mathbb{V}[Y]$

Calculs de variance [Cas discret]

Soit X une variable aléatoire réelle sur un espace probabilisé fini avec $X \hookrightarrow \mathcal{B}(1, p)$ avec $p \in [0, 1]$. Alors :

$$\mathbb{V}(X) = p(1 - p)$$

Soit X une variable aléatoire réelle sur un espace probabilisé fini avec $X \hookrightarrow \mathcal{B}(n, p)$ avec $p \in [0, 1]$. Alors :

$$\mathbb{V}(X) = np(1 - p)$$

Soit $(X_1, \cdots, X_n)$ une suite de variables aléatoires indépendantes de même loi de Bernoulli de paramètre p.

$$\mathbb{V}\left(\frac{X_1 + \cdots + X_n}{n}\right) = \ldots\ldots\ldots$$

Soit X une variable aléatoire non constante sur un espace probabilisé fini. Alors la variable Y $= \frac{X - \mathbb{E}[X]}{\sigma(X)}$ est ....

Calculs de variance [Cas discret]

Soit X une variable aléatoire réelle sur un espace probabilisé fini avec $X \hookrightarrow \mathcal{B}(1,p)$ avec $p \in [0,1]$. Alors :

$$\mathbb{V}(X) = p(1-p)$$

Soit X une variable aléatoire réelle sur un espace probabilisé fini avec $X \hookrightarrow \mathcal{B}(n,p)$ avec $p \in [0,1]$. Alors :

$$\mathbb{V}(X) = np(1-p)$$

Soit $(X_1, \cdots, X_n)$ une suite de variables aléatoires **indépendantes** de même loi de Bernoulli de paramètre p.

$$\mathbb{V}\left(\frac{X_1 + \cdots + X_n}{n}\right) = \frac{n\mathbb{V}(X_1)}{n^2} = \frac{\mathbb{V}(X_1)}{n}$$

Soit X une variable aléatoire non constante sur un espace probabilisé fini. Alors la variable $\mathrm{Y} = \frac{X - \mathbb{E}[X]}{\sigma(X)}$ est **centrée réduite**

Sommaire

Loi faible des grands nombres (1/3)

Loi faible des grands nombres. Soit (X_n) une suite de variables aléatoires réelles indépendantes, de même loi admettant une espérance μ.

$$\forall \varepsilon > 0, \quad \lim_{n \to +\infty} \mathbb{P}(|\overline{X_n} - \mu| > \varepsilon) = 0, \text{ où } \overline{X_n} = \frac{1}{n} \sum_{k=1}^{n} X_k.$$

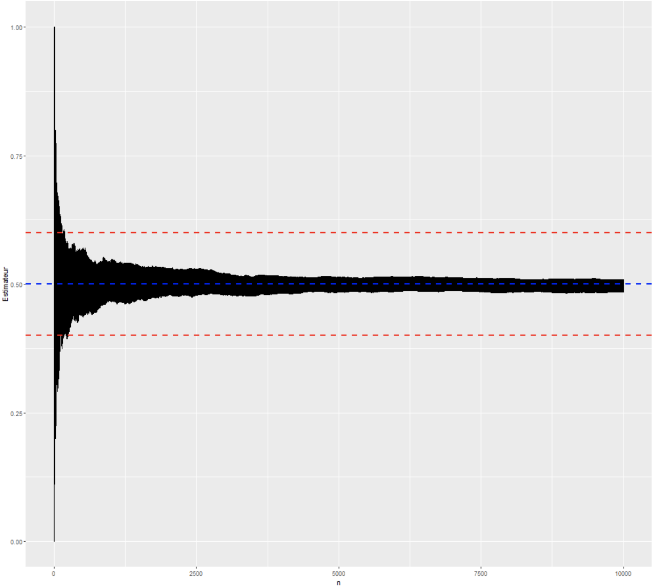
Version vulgarisée. Lorsque n est grand, sauf exception, la fréquence observée est proche de la probabilité.

Convergence en probabilité. Soit $(X_n)_{n \in \mathbb{N}}$ une suite de variables aléatoires réelles. On dit que X_n converge en probabilité vers X si

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

Autrement dit, la moyenne empirique converge en probabilité vers l'espérance.

Loi faible des grands nombres : N=100, n=10000, ε=0.1



Application : Variables de Bernoulli

Un caractère est en proportion p (paramètre) dans une population, on y prélève un échantillon aléatoire de taille n et on observe une fréquence f du caractère dans l'échantillon.

On modélise cette situation en considérant les variables aléatoires indépendantes $X_1, X_2, \ldots, X_n \hookrightarrow \mathcal{B}(p)$

La variable somme $S_n = X_1 + X_2 + \cdots + X_n$, correspondant au nombre d'occurrences du caractère sur un échantillon, suit la loi binomiale de paramètres n et p.
$\mathbb{E}(S_n) = np$ et $\sigma(S_n) = \sqrt{np(1-p)}$.

$\overline{X_n} = \frac{S_n}{n}$ avec écart type : $\sigma(\overline{X_n}) = \frac{\sigma(S_n)}{n} = \frac{1}{\sqrt{n}}\sqrt{p(1-p)}$.
L'écart type de la distribution d'échantillonnage des fréquences est donc de l'ordre de $\frac{1}{\sqrt{n}}$.

Écart type comme contrôle de cas rares

- L'étude de la fonction $p \mapsto p(1-p)$ sur $[0, 1]$ montre que

  $$p(1-p) \leq 1/4$$

  et que $p(1-p)$ est proche de $1/4$ lorsque p est proche de $1/2$.

  Lorsque p n'est ni trop proche de 0, ni trop proche de 1, l'écart type de la distribution des fréquences sur des échantillons de taille n est proche de $\frac{1}{2\sqrt{n}}$.

  Activité. Simuler N échantillons de taille n pour une expérience aléatoire à deux issues et comparer l'écart type de la série des N fréquences obtenues sur chacun des échantillons avec $\frac{1}{2\sqrt{n}}$.

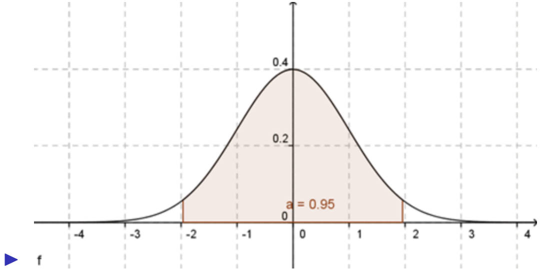
Écart type comme contrôle de cas rares

- **Théorème de Moivre-Laplace**. si la variable S_n suit une loi binomiale de paramètres n et de paramètre $p \in]0,1[$ alors la variable $Z_n = \frac{S_n - \mathbb{E}(S_n)}{\sigma(S_n)}$ converge en loi vers $\mathcal{N}(0,1)$.

  Remarque. Autrement dit, une loi $\mathcal{B}(n,p)$ peut être approché quand n est grand et p ni trop proche de 0 ou 1, par la loi Normale $\mathcal{N}(np, \sqrt{np(1-p)})$.

  Ainsi, pour n assez grand et p ni trop proche de 0, ni trop proche de 1, la variable d'échantillonnage $\overline{X_n}$ suit approximativement la loi normale de moyenne p et d'écart-type $\frac{1}{2\sqrt{n}}$.

Intervalle de fluctuation et estimation d'une probabilité



Quantile d'ordre α de la loi normale centrée réduite. Le quantile d'ordre α noté q_α est défini tel que $\mathbb{P}(X \leq q_\alpha) = \alpha$.

Remarque. Comme la loi normale est symétrique, alors $q_\alpha = -q_{1-\alpha}$. Ainsi on en déduit que

$$\begin{aligned} 0.95 = \mathbb{P}(Z_n \leq q_{95\%}) &= \mathbb{P}(Z_n \leq q_{97.5\%}) - \mathbb{P}(Z_n \leq q_{2.5\%}) \\ &= \mathbb{P}(q_{2.5\%} \leq Z_n \leq q_{97.5\%}) \\ &= \mathbb{P}(-q_{97.5\%}) \leq Z_n \leq q_{97.5\%}) \end{aligned}$$

Intervalle de fluctuation et estimation d'une probabilité

On a alors comme $p(1-p)$ au maximum égal à $1/4$ et $q_{97.5\%} = 1.96 \approx 2$

$$
\begin{aligned}
\mathbb{P}(-q_{97.5\%} \leq \frac{S_n - np}{\sqrt{np(1-p)}} \leq q_{97.5\%}) &= 0.95 \\
\mathbb{P}(-q_{97.5\%}\sqrt{\frac{p(1-p)}{n}} + p \leq \frac{S_n}{n} \leq q_{97.5\%}\sqrt{\frac{p(1-p)}{n}} + p) &= 0.95 \\
\mathbb{P}(-q_{97.5\%}\sqrt{\frac{p(1-p)}{n}} + p \leq \frac{S_n}{n} \leq q_{97.5\%}\sqrt{\frac{p(1-p)}{n}} + p) &= 0.95 \\
\mathbb{P}(-2\sqrt{\frac{1}{4n}} + p \leq \frac{S_n}{n} \leq 2\sqrt{\frac{1}{4n}} + p) &= 0.95 \\
\mathbb{P}(-\frac{1}{\sqrt{n}} \leq \frac{S_n}{n} - p \leq \frac{1}{\sqrt{n}}) &= 0.95
\end{aligned}
$$

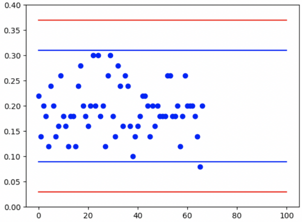
Ainsi, pour n assez grand et p ni trop proche de 0, ni trop proche de 1, la proportion des cas où l'écart entre p et f (réalisation de S_n/n) est inférieur ou égal à $\frac{1}{\sqrt{n}}$, soit environ deux écarts types, est de l'ordre de 95 %.
Comme $\mathbb{P}(\frac{S_n}{n} - \frac{1}{\sqrt{n}} \leq p \leq \frac{S_n}{n} + \frac{1}{\sqrt{n}}) = 0.95$, on peut **estimer** que dans 95 % des cas, environ, p est compris entre $f - \frac{1}{\sqrt{n}}$ et $f + \frac{1}{\sqrt{n}}$.

Intervalle de fluctuation en 1ère technologique

En 1ère technologique, on fait percevoir les interprétations, en termes de rareté, d'une fréquence observée dans le cadre du modèle.

Sur des simulations de N échantillons on évalue le pourcentage d'échantillons dont la fréquence observée des 1 se situe à une distance s, 2s ou 3s de p où s est **l'écart-type de la série des fréquences observées** (estimateur de l'écart-type).

Carte de contrôle



Intervalle de fluctuation en 1ère technologique

Si $X \hookrightarrow \mathcal{N}(\mu, \sigma)$, alors $\mathbb{P}(-k \leq \frac{X-\mu}{\sigma} \leq k) = 0.68$ $(k \approx 1)$, 0.95 $(k \approx 2)$, 0.997 $(k \approx 3)$. Comme la loi de $\overline{X_n}$ peut être approchée par une loi de normale de moyenne p et d'écart-type $\frac{\sigma}{\sqrt{n}}$ on a :

$$\mathbb{P}(p - k\frac{\sigma}{\sqrt{n}} \leq \overline{X_n} \leq p + k\frac{\sigma}{\sqrt{n}}) = 0.68\ (k \approx 1), 0.95\ (k \approx 2), 0.997 (k \approx 3).$$

Comme $\frac{\sigma}{\sqrt{n}}$ est inconnu on **l'estime** par s.

$$\mathbb{P}(-ks \leq \overline{X_n} - p \leq ks) = 0.68\ (k = 1),\ 0.95\ (k = 2),\ 0.997\ (k = 3)$$

Ainsi, pour n assez grand et p ni trop proche de 0, ni trop proche de 1, la proportion des cas où l'écart entre p et f (réalisation de $\overline{X_n}$) est inférieur ou égal à $2s$ (resp s, 3s), est de l'ordre de 95 % (resp 68%, 99,7%).

Sommaire

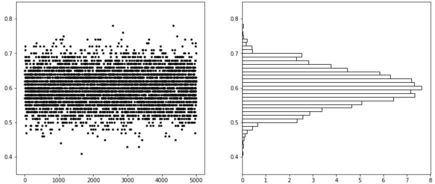
Théorème Central Limite (TCL)

Soit $(X_n)_{n\in\mathbb{N}^*}$ une suite de variables aléatoires variables aléatoires indépendantes de même loi, d'espérance μ et de variance σ^2 finies.

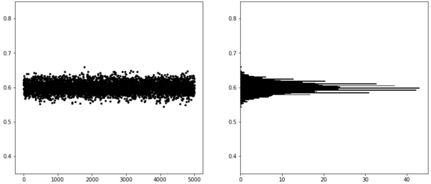
Posons : $\forall n \in \mathbb{N}^*$, $\quad \overline{X}_n = \dfrac{X_1 + \cdots + X_n}{n}$ et $\quad Z_n = \dfrac{\sqrt{n}}{\sigma}\left(\overline{X}_n - \mu\right)$.

La loi de Z_n converge vers la loi normale $\mathcal{N}(0,1)$,

Remarque. Généralisation du Théorème de Moivre Laplace.



Distribution d'échantillonnage des fréquences de 5 000 échantillons de taille n = 100 avec p = 0,6.



Estimation de la moyenne

Comme $q_{97.5\%} = 1.96 \approx 2$, on a alors

$$\begin{array}{rcl}
\mathbb{P}(q_{2.5\%} \leq Z_n \leq q_{97.5\%}) & = & 0.95 \\
\mathbb{P}(-q_{97.5\%} \leq Z_n \leq q_{97.5\%}) & = & 0.95 \\
\mathbb{P}(-q_{97.5\%} \leq \dfrac{\sqrt{n}}{\sigma}(\overline{X_n} - \mu) \leq q_{97.5\%}) & = & 0.95 \\
\mathbb{P}(-q_{97.5\%}\dfrac{\sigma}{\sqrt{n}} + \mu \leq \overline{X_n} \leq q_{97.5\%}\dfrac{\sigma}{\sqrt{n}} + \mu) & = & 0.95 \\
\mathbb{P}(-\dfrac{2\sigma}{\sqrt{n}} \leq \overline{X_n} - \mu \leq \dfrac{2\sigma}{\sqrt{n}}) & = & 0.95
\end{array}$$

Ainsi, pour n assez grand, pour environ 95 % des échantillons de taille n, la distance entre μ et la moyenne observée m (réalisation de $\overline{X_n}$) de l'échantillon est inférieure à deux écarts types soit environ $\frac{2\sigma}{\sqrt{n}}$.

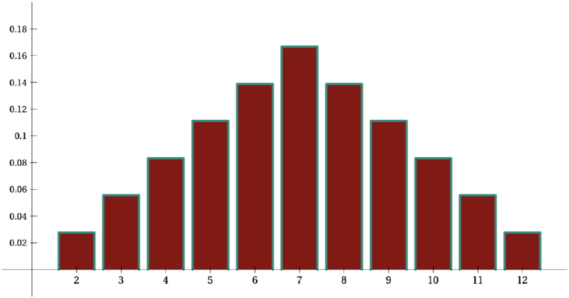
Comme $\mathbb{P}(\overline{X_n} - \frac{2\sigma}{\sqrt{n}} \leq \mu \leq \overline{X_n} + \frac{2\sigma}{\sqrt{n}}) = 0.95$, on peut **estimer** si σ connu que que dans 95 % des cas, environ, μ est compris entre $m - \frac{2\sigma}{\sqrt{n}}$ et $m + \frac{2\sigma}{\sqrt{n}}$

Inégalité de Bienaymé-Tchebychev : énoncé

Soit X une variable aléatoire d'espérance $\mathbb{E}(Z)$ et de variance finie $\mathbb{V}(Z)$. Alors pour tout réel strictement positif $a > 0$,

$$\mathbb{P}(|Z - \mathbb{E}(Z)| \geq a) \leq \frac{\mathbb{V}(Z)}{a^2}.$$

Remarques. Elle vise à montrer que la variable aléatoire X prend des valeurs proches de $\mathbb{E}(X)$ avec une grande probabilité, mais elle donne, en générale une majoration assez grossière. Probabilité d'autant plus faible que $\mathbb{V}(X)$ est petit et a grand.



On trouve que l'espérance de X est $E(X) = 7$ en faisant des calculs (en effet, on ne peut pas lire l'espérance sur l'histogramme). En effet, on voit que plus on s'éloigne de la valeur de l'espérance, plus les probabilités décroissent ; on peut juste « sentir » compte du phénomène, pour le vérifier, il faudrait faire les calculs.

Inégalité de Bienaymé -Tchebychev : preuve

Elle est basée sur :

Inégalite de Markov. Soit Y une variable aléatoire positive d'espérance $\mathbb{E}(Y)$. Alors pour tout réel strictement positif $a > 0$,

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

Preuve. On a alors
$$\begin{aligned}\mathbb{E}(Y) &= \sum_{k \in Y(\Omega)} k\mathbb{P}(Y = k) \\ &\geq \sum_{k \geq a} k\mathbb{P}(Y = k) \\ &\geq \sum_{k \geq a} a\mathbb{P}(Y = k) = a\mathbb{P}(Y \geq a).\end{aligned}$$

Preuve de l'inégalité de B-T. On utilise l'inégalité de Markov avec $Y = (Z - \mathbb{E}(Z))^2$. On a $\mathbb{P}(|Z - \mathbb{E}(Z)| \geq a) = \mathbb{P}((Z - \mathbb{E}(Z))^2 \geq a^2)$ et on remarque que $\mathbb{E}\left((Z - \mathbb{E}(Z))^2\right) = \mathbb{V}(Z)$.

Inégalité de Bienaymé-Tchebychev : application 1

Une usine fabrique des pièces dont une proportion inconnue p est défectueuse, et on souhaite trouver une valeur approchée de p. On effectue un prélèvement de n pièces. On suppose que le prélèvement se fait sur une population très grande, et donc qu'il peut s'apparenter à une suite de n tirages indépendants avec remise. On note S_n la variable aléatoire égale au nombre de pièces défectueuses et on souhaite quantifier le fait que $\frac{S_n}{n}$ approche p.

- Quelle est la loi de S_n ? Sa moyenne ? Sa variance ?
- Démontrer que, pour tout $\varepsilon > 0$, $\mathbb{P}(|\frac{S_n}{n} - p| \geq \varepsilon) \leq \frac{1}{4n\varepsilon^2}$.
- En déduire une condition sur n pour que $\frac{S_n}{n}$ soit une valeur approchée de p à 10^{-2} près avec une probabilité supérieure ou égale à 95%.

Inégalité de Bienaymé-Tchebychev : Correction

- S_n est la somme de n variables aléatoires de Bernoulli indépendantes de paramètre p. Ainsi, S_n suit une loi binomiale $\mathcal{B}(n,p)$. On en déduit que $\mathbb{E}(S_n) = np$ et $\mathbb{V}(S_n) = np(1-p)$.

- Appliquons l'inégalité de B-T à S_n. On a donc $\mathbb{P}(|\frac{S_n}{n} - p| \geq \varepsilon) \leq \frac{\mathbb{V}(S_n)}{\varepsilon^2}$.
  Or, $\mathbb{P}(|\frac{S_n}{n} - p| \geq \varepsilon) \Leftrightarrow \mathbb{P}(|S_n - np| \geq n\varepsilon) \Leftrightarrow \mathbb{P}(|S_n - \mathbb{E}(S_n)| \geq n\varepsilon)$.
  On applique donc l'inégalité de B-T avec $a = n\varepsilon$ et on trouve que $\mathbb{P}(|\frac{S_n}{n} - p| \geq \varepsilon) \leq \frac{p(1-p)}{n\varepsilon^2}$.
  Mais sur [0,1], la fonction $x \mapsto x(1-x)$ admet un maximum égal à $1/4$ (et atteint en $1/2$). On en déduit le résultat voulu.

- On cherche n tel que $\mathbb{P}(|\frac{S_n}{n} - p| < 10^{-2}) \geq 0,95$ soit encore, en passant à l'événement contraire :
  $1 - \mathbb{P}(|\frac{S_n}{n} - p| < 10^{-2}) = \mathbb{P}(|\frac{S_n}{n} - p| \geq 10^{-2}) \leq 1 - 0.95 = 0,05$. Il suffit donc de choisir n tel que $\frac{1}{4n10^{-4}} \leq 0.05$ soit $n \geq 5 \times 10^4$.

Inégalité de Bienaymé-Tchebychev : application 2

Rappel : loi faible des grands nombres. Soit (X_n) une suite de variables aléatoires réelles indépendantes, de même loi admettant une espérance μ.

$$\forall \varepsilon > 0, \quad \lim_{n \to +\infty} \mathbb{P}(|\overline{X_n} - \mu| > \varepsilon) = 0, \text{ où } \overline{X_n} = \frac{1}{n} \sum_{k=1}^{n} X_k.$$

Preuve de la loi faible des grands nombres. Soit $\overline{X_n} = \frac{X_1 + \cdots + X_n}{n}$. Son espérance vaut $\mathbb{E}(X)$ et sa variance vaut $\frac{\mathbb{V}(X)}{n}$.

L'inégalité de B.T appliqué à $\overline{X_n}$ donne avec $a = \varepsilon$:

$$\mathbb{P}\left(\left|\frac{X_1 + X_2 + \cdots + X_n}{n} - \mathbb{E}(X)\right| \geqslant \varepsilon\right) \leqslant \frac{\mathbb{V}(X)}{n\,\varepsilon^2} \underset{n \to +\infty}{\longrightarrow} 0 .$$

Démonstration du théorème central limite

Fonction caractéristique. La fonction caractéristique d'une variable aléatoire réelle X est la fonction à valeurs complexes définie sur $\mathbb{R}$ par

$$\phi_X(t) = \mathbb{E}\left[e^{\mathrm{i}tX}\right]$$

Propriétés.

- Elle détermine de façon unique la loi d'une variable aléatoire au sens où « $\phi_X = \phi_Y$ » équivaut à « X et Y ont la même loi. »
- Si X et Y sont deux variables aléatoires indépendantes, $\phi_{X+Y} = \phi_X \phi_Y$.
- Plus généralement, si $X_1, ..., X_n$ sont des variables aléatoires indépendantes dans leur ensemble, de même loi alors $\phi_{X_1+...+X_n} = \phi_{X_1}^n$.

Démonstration du théorème central limite

Pour une variable aléatoire Y d'espérance 0 et de variance 1, la fonction caractéristique de Y admet le développement limité :

$$\varphi_Y(t) = 1 - \frac{t^2}{2} + o(t^2), \quad t \to 0$$

Comme Z_n est une variable centrée réduite des observations $X_1, \cdots, X_n$,

$$Z_n = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} = \sum_{i=1}^{n} \frac{X_i}{\sqrt{n}},$$

on a alors, d'après les propriétés élémentaires des fonctions caractéristiques, la fonction caractéristique de Z_n est

$$\left[\varphi_Y\left(\frac{t}{\sqrt{n}}\right)\right]^n = \left[1 - \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right)\right]^n \underset{n\to\infty}{\longrightarrow} \mathrm{e}^{-t^2/2}.$$

Cette limite est la fonction caractéristique de la loi normale centrée réduite $\mathcal{N}(0,1)$ et on conclut grâce au théorème de Levy qui assure que la convergence simple des fonctions caractéristiques implique la convergence en loi.

Sommaire

Pour aller plus loin types de convergence

Loi forte des grands nombres. Convergence presque-sûre de $\frac{S_n}{n}$ vers son espérance.

Différents types de convergence.

Convergence presque-sûre $\Rightarrow$ convergence en probabilité $\Rightarrow$ convergence en loi

Autre inégalité de concentration

Inégalité de Hoeffding. Soit une suite $(X_k)_{1\le k\le n}$ de variables aléatoires réelles indépendantes vérifiant, pour deux suites $(a_k)_{1\le k\le n}$, $(b_k)_{1\le k\le n}$ de nombres réels tels que $a_k < b_k$,
$\forall k, \qquad \mathbb{P}(a_k \le X_k \le b_k) = 1$. On pose $S_n = X_1 + X_2 + \cdots + X_n$.
Alors, pour tout $t > 0$,

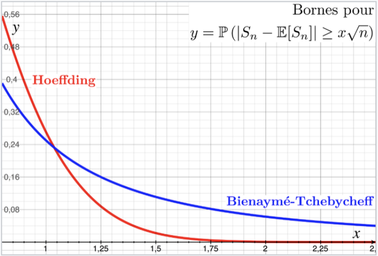
$$\begin{aligned}
\mathbb{P}\left(S_n - \mathbb{E}[S_n] \ge t\right) &\le \exp\left(-\frac{2\,t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right),\\
\mathbb{P}\left(S_n - \mathbb{E}[S_n] \le -t\right) &\le \exp\left(-\frac{2\,t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right),\\
\mathbb{P}\left(|S_n - \mathbb{E}[S_n]| \ge t\right) &\le 2\exp\left(-\frac{2\,t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).
\end{aligned}$$

Autre inégalité de concentration

Cas de la loi binomiale. Soit S_n suivant la loi binomiale de paramètres n et p. L'inégalité de Bienaymé-Tchebychev et l'inégalité Hoeffding donnent respectivement $\forall x > 0$

$$\begin{aligned} \mathbb{P}\left(|S_n - \mathbb{E}[S_n]| \geq x\sqrt{n}\right) &\leq \frac{p(1-p)}{x^2}, \\ \mathbb{P}\left(|S_n - \mathbb{E}[S_n]| \geq x\sqrt{n}\right) &\leq 2\exp\left(-2\,x^2\right). \end{aligned}$$

L'inégalité de Hoeffding est beaucoup plus précise pour x suffisamment grand (c'est assez représentatif de la situation générale).



Bornes pour la dispersion de la loi binomiale de paramètres n et p=0,5, obtenues respectivement à l'aide de l'inégalité de Bienaymé-Tchebychev et à l'aide de l'inégalité de Hoeffding.

Auteur : Chassain.

Pour aller plus loin : marche aléatoire en une dimension

- Soit $k \in \mathbb{N}^*$. Calculer la probabilité que le bonhomme soit à l'origine à l'instant $2k$.
- Soit $n \in \mathbb{N}^*$. On note N_n la variable aléatoire égale au nombre de passages à l'origine entre l'instant 0 et l'instant n. Calculer l'espérance de N_n et donner un équivalent de N_n quand n tend vers l'infini.