Table des matières

1

Statistique univariée.

1.1 Les résultats d'une expérience.

Les résultats d'une expérience peuvent être qualitatives ou quantitatives.
Nous nous intéresserons ici aux résultats numériques. (souvent le résultat d'une mesure).

On note : $(\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n)$ la liste des données (n-uplets). Chaque $\tilde{x}_i$ est un nombre réel.

On commence souvent par rassembler les résultats identiques pour présenter les données sous la forme :

valeurs	x_1	x_2	$\cdots$	x_{p-1}	x_p
effectifs	e_1	e_2	$\cdots$	e_{p-1}	e_p

Les valeurs distinctes sont classées dans l'ordre croissant. $x_1 < x_2 < \cdots < x_{p-1} < x_p$
L'effectif e_i est le nombre de résultats égaux à x_i, $n = \sum_{i=1}^{p} e_i$ est l'effectif total.
Souvent on préfère souvent utiliser les fréquences plutôt que les effectifs.

valeurs	x_1	x_2	$\cdots$	x_{p-1}	x_p
fréquences	f_1	f_2	$\cdots$	f_{p-1}	f_p

La fréquence $f_i = \dfrac{e_i}{n}$ est la proportion de données dont la valeur est égale x_i.

Remarque : $\sum_{i=1}^{p} f_i = 1$

On présente aussi les résultats avec les effectifs cumulés.
L'effectif cumulé en un réel x est le nombre de données dont la valeur est inférieure ou égale à x.

valeurs	x_1	x_2	$\cdots$	x_{p-1}	x_p
effectifs cumulés	c_1	c_2	$\cdots$	c_{p-1}	c_p

pour i compris entre 1 et p, $\quad c_i = \sum_{k=1}^{i} e_k$

On utilise aussi les fréquences cumulées :

valeurs	x_1	x_2	$\cdots$	x_{p-1}	x_p
fréquences cumulées	F_1	F_2	$\cdots$	F_{p-1}	F_p

pour i compris entre 1 et p, $\quad F_i = \sum_{k=1}^{i} f_k$

c'est la proportion de données dont la valeur est inférieure ou égale à x_i et $F_p = 1$.

1.2 Des nombres caractéristiques.

1.2.1 Caractéristiques de position.

Moyenne :

$$\overline{x} = \frac{1}{n}\left(\sum_{k=1}^{n} \tilde{x}_k\right) = \frac{1}{\sum_{k=1}^{p} e_k}\left(\sum_{k=1}^{p} e_k\, x_k\right) = \sum_{k=1}^{p} f_k\, x_k$$

Médiane : *(Définition*

On suppose ici que : $\tilde{x}_1 \leqslant \tilde{x}_2 \leqslant \cdots \leqslant \tilde{x}_{n-1} \leqslant \tilde{x}_n$ les données brutes.

si n est impair alors on choisit pour médiane le réel $\quad \tilde{x}_{\frac{n+1}{2}}$

si n est pair alors on choisit pour médiane le réel $\quad \dfrac{1}{2}\left(\tilde{x}_{\frac{n}{2}} + \tilde{x}_{\frac{n}{2}+1}\right)$

Si on note $\text{med}(x)$ la médiane on a :

$$\text{card}\{\ i \mid x_i < \text{med}(x)\ \} = \text{card}\{\ i \mid X_i > \text{med}(x)\ \}$$

Si on change la numérotation comme avec Python : $\tilde{x}_0 \leqslant \tilde{x}_1 \leqslant \cdots \leqslant \tilde{x}_{n-2} \leqslant \tilde{x}_{n-1}$

si n est impair alors on choisit pour médiane le réel $\quad \tilde{x}_{\frac{n-1}{2}}$

si n est pair alors on choisit pour médiane le réel $\quad \dfrac{1}{2}\left(\tilde{x}_{\frac{n}{2}-1} + \tilde{x}_{\frac{n}{2}}\right)$

Premier et troisième quartiles. *(notées Q_1 et Q_3)*

Le **premier quartile** est l'unique valeur x_i où i est l'unique i de $[\![1, p]\!]$ vérifiant :

$$\sum_{k=1}^{i-1} e_k < \frac{n}{4} \qquad \sum_{k=1}^{i} e_k \geqslant \frac{n}{4}$$

Le **troisième quartile** est l'unique valeur x_i où i est l'unique i de $[\![1, p]\!]$ vérifiant :

$$\sum_{k=1}^{i-1} e_k < \frac{3n}{4} \qquad \sum_{k=1}^{i} e_k \geqslant \frac{3n}{4}$$

Premier et neuvième déciles *(notées D_1 et D_9)* :

Le **premier décile** est l'unique valeur x_i où i est l'unique i de $[\![1, p]\!]$ vérifiant :

$$\sum_{k=1}^{i-1} e_k < \frac{n}{10} \qquad \sum_{k=1}^{i} e_k \geqslant \frac{n}{10}$$

Le **neuvième décile** est l'unique valeur x_i où i est l'unique i de $[\![1, p]\!]$ vérifiant :

$$\sum_{k=1}^{i-1} e_k < \frac{9n}{10} \qquad \sum_{k=1}^{i} e_k \geqslant \frac{9n}{10}$$

Remarques :

• L'ensemble de ses valeurs se présente souvent sous la forme d'un **boxplot** ou encore **une boite à moustaches**.

• On rencontre une autre définition des fractiles : les quartiles sont alors les quatre intervalles séparant la population en quatre *(Voir sujet 0 de Biologie)*.

Modes ou valeurs modales :

L'ensemble des valeurs des valeurs modales :

$$\left\{ x_i \mid \quad 1 \leqslant i \leqslant p \quad \text{et} \quad e_i = \max_{1 \leqslant k \leqslant p} (e_k) \right\}$$

Remarques :

- C'est un x_i correspondant à la plus grande valeur des e_i. (*Il n'y a pas unicité de ce paramètre*).
- C'est une valeur qui apparaît le plus souvent.

1.2.2 Caractéristiques de dispersion.

L'écart-type est donné par chacune des formules :

$$\sigma_x = \sqrt{\frac{1}{n} \sum_{k=1}^{n} (\tilde{x}_k - \overline{x})^2} = \sqrt{\sum_{k=1}^{p} f_k (x_k - \overline{x})^2} = \sqrt{\frac{1}{\sum_{k=1}^{p} e_k} \left(\sum_{k=1}^{p} e_k (x_k - \overline{x})^2 \right)}$$

Remarques :

- On a toujours $\sigma_x \geqslant 0$.
- Si les x_i ont une unité alors σ_x a la même unité.
- $\sigma_x = 0$ si, et seulement si, $(x_1, ..., x_n) \in \text{vect} < (1, ..., 1) >$

On appelle **variance** (empirique) le carré de l'écart-type :

$$\sigma_x^2 = \frac{1}{n} \sum_{k=1}^{n} (\tilde{x}_k - \overline{x})^2 = \sum_{k=1}^{p} f_k (x_k - \overline{x})^2 = \frac{1}{\sum_{k=1}^{p} e_k} \left(\sum_{k=1}^{p} e_k (x_k - \overline{x})^2 \right)$$

Remarque :

Dans la suite du cours nous allons définir une variable aléatoire que l'on appellera aussi variance empirique.

2

Statistique bivariée.

2.1 Nuage de points.

Au cours d'une expérience on recueille des valeurs de deux caractères :

$$\tilde{x}_1\,,\,\tilde{x}_2\,,\,\ldots\,,\,\tilde{x}_n \qquad \text{et} \qquad \tilde{y}_1\,,\,\tilde{y}_2\,,\,\ldots\,,\,\tilde{y}_n$$

Lorsqu'on étudie la relation éventuelle entre les deux caractères on commence par représenter tous les points de coordonnées $(\tilde{x}_i, \tilde{y}_i)$. On obtient alors un nuage de points.

On place **le point moyen** du nuage : le point de coordonnées $(\overline{x}, \overline{y})$.

On peut organiser ces données dans le tableau suivant :

valeurs du premier caractère	x_1	x_2	$\cdots$	x_{p-1}	x_p
valeurs du deuxième caractère	y_1	y_2	$\cdots$	y_{p-1}	y_p
effectifs	e_1	e_2	$\cdots$	e_{p-1}	e_p

Remarque : $\quad n = \sum_{k=1}^{p} e_k$

2.2 Covariance et coefficient de corrélation linéaire.

Définition :

On appelle **covariance** des séries x et y le réel :

$$\text{cov}(x,y) = \frac{1}{n}\sum_{k=1}^{n}(\tilde{x}_k - \overline{x})(\tilde{y}_k - \overline{y}) \qquad \text{ou encore} \qquad \text{cov}(x,y) = \frac{1}{n}\sum_{k=1}^{p} e_k(x_k - \overline{x})(y_k - \overline{y})$$

Proposition :

$$\text{cov}(x,y) \;=\; \frac{1}{n}\left(\sum_{k=1}^{n}\tilde{x}_k\,\tilde{y}_k\right) - \overline{x}\,\overline{y} \;=\; \frac{1}{n}\left(\sum_{k=1}^{n} e_k\,x_k\,y_k\right) - \overline{x}\,\overline{y}$$

Définition

On appelle **coefficient de corrélation** (*linéaire*) le nombre :

$$r_{x\,y} = \frac{\text{cov}(x,y)}{\sigma_x\,\sigma_y}$$

Remarques :

- Ce coefficient est sans unité.
- Savoir calculer cette valeur avec votre calculatrice, un tableur et en faisant une fonction Python.

Propositions :

① $-1 \leqslant r_{x\,y} \leqslant 1$
② $r_{x\,y} = 1$ ou -1 si, et seulement si, les points M_1, M_2, … , M_n sont alignés.

2.3 Droite de régression linéaire

Définition :

La **droite de régression linéaire de y en x** est la droite d'équation $y = ax + b$ où : $$a = \frac{\text{cov}(x,y)}{\sigma_x^2} \qquad \text{et} \qquad b = \overline{y} - a\overline{x}$$

Démonstration :

Savoir faire : calculer a et b avec une calculatrice, un tableur **et** en faisant une fonction Python.

En donnant cette droite on dit que réalise un ajustement affine selon *la méthode des moindres carrés.*

Interprétation géométrique.

Remarques :

- (*Il existe une deuxième droite de régression linéaire :*)
  On définit la droite de **régression linéaire de x en y** qui est la droite d'équation $x = \alpha y + \beta$ où :
  $$\alpha = \frac{\text{cov}(x,y)}{\sigma_y^2} \qquad \text{et} \qquad \beta = \overline{x} - \alpha\overline{y}$$
- Ces deux droites de régression passent par le point milieu.

3

Loi faible des grands nombres.

3.1 Inégalité de Markov.

Théorème :

Soient X une variable aléatoire réelle et a un nombre réel,

$$\text{Si} \quad \begin{cases} X \text{ admet une espérance} \\ X \text{ est à valeurs positives} \\ a > 0 \end{cases} \quad \text{alors} \quad \mathbb{P}(X \geqslant a) \leqslant \frac{\mathbb{E}(X)}{a}.$$

Démonstration : On pose $Y = a\mathbb{1}_{X \geqslant a}$

• $\mathbb{1}_{X \geqslant a}$ suit la loi de Bernoulli de paramètre $\mathbb{P}(X \geqslant a)$ et ainsi son espérance est $\boxed{\mathbb{E}(\mathbb{1}_{X \geqslant a}) = \mathbb{P}(X \geqslant a)}$.

• On a : $\boxed{Y \leqslant X}$

En effet pour $\omega \in \Omega$,

si $X(\omega) \geqslant a$ alors $Y(\omega) = a$ donc on a bien $Y(\omega) \leqslant X(\omega)$

si $X(\omega) < a$ alors $Y(\omega) = 0$ et comme X est à valeurs positives ou nulles, ici aussi on a bien $Y(\omega) \leqslant X(\omega)$

• $Y \leqslant X$ et ces deux variables admettent une espérance donc $E(Y) \leqslant \mathbb{E}(X)$

or *(linéarité)* $\mathbb{E}(Y) = a\,\mathbb{E}(\mathbb{1}_{X \geqslant a})$ ou encore $\mathbb{E}(Y) = a\,\mathbb{P}(X \geqslant a)$

et comme $a > 0$ on peut en conclure que : $\boxed{\mathbb{P}(X \geqslant a) \leqslant \dfrac{\mathbb{E}(X)}{a}}$ ■

Interprétations :

• Quand X est à valeurs positives, il est peu probable que X prenne des valeurs grandes devant à $\mathbb{E}(X)$.

• Quand X est à valeurs positives, la probabilité que X dépasse largement son espérance est faible.

• Quand X est à valeurs positives, pour tout $k > 0$, $\quad P\big(X \geqslant k\,\mathbb{E}(X)\big) \leqslant \dfrac{1}{k}$

En particulier : $\quad P\big(X \geqslant 2\,\mathbb{E}(X)\big) \leqslant \dfrac{1}{2}, \quad P\big(X \geqslant 3\,\mathbb{E}(X)\big) \leqslant \dfrac{1}{3}, \quad P\big(X \geqslant 10\,\mathbb{E}(X)\big) \leqslant \dfrac{1}{10}$

Exemples :

• Si X est à valeurs positives et $\mathbb{E}(X) = 2$ alors $P(X \geqslant 10) \leqslant \dfrac{1}{5}$

• Si X est un temps d'attente de moyenne 10 minutes alors $\mathbb{P}(X \geqslant 25) \leqslant \dfrac{10}{25} = \dfrac{2}{5}$

En répétant cette expérience un grand nombre de fois,

la proportion d'attentes dépassant 25 minutes sera proche de $\mathbb{P}(X \geqslant 25)$ et donc inférieure à $\dfrac{2}{5}$.

Remarques :

Ce théorème donne des informations sur X sans connaître la loi de X, seul $\mathbb{E}(X)$ intervient.

Attention à bien rappeler les conditions d'application de ce théorème [1] :

$$a > 0, \quad X \text{ à valeurs dans } \mathbb{R}_+ \text{ et } X \text{ admet une espérance.}$$

1. Comme pour tous les théorèmes du cours

3.2 Inégalité de Bienaymé-Tchebychev.

Théorème :

Soit X une variable aléatoire réelle,

Si X admet une variance alors $\forall t > 0, \quad \mathbb{P}\left(|X - \mathbb{E}(X)| \geqslant t\right) \leqslant \dfrac{\mathbb{V}(X)}{t^2}$.

Démonstration :
X admet une variance donc $Y = (X - E(X))^2$ admet une espérance et et $Y = (X - E(X))^2$ est à valeurs positives ou nulles. La variable $Y = (X - E(X))^2$ vérifie les conditions d'utilisation de l'inégalité de Markov.

On en déduit avec $a = t^2 > 0$:

$$\mathbb{P}\left((X - \mathbb{E}(X))^2 \geqslant t^2 \right) \leqslant \frac{\mathbb{E}\left(X - \mathbb{E}(X)\right)^2}{t^2}$$

or $\left((X - \mathbb{E}(X))^2 \geqslant t^2\right) = (|X - \mathbb{E}(X)| \geqslant t)$ et $\mathbb{E}\left(X - \mathbb{E}(X)\right)^2 = V(X)$ donc

$$\mathbb{P}\left(|X - \mathbb{E}(X)| \geqslant t \right) \leqslant \frac{\mathbb{V}(X)}{t^2}$$

■

Interprétations :

- La variable aléatoire $|X - \mathbb{E}(X)|$ est l'écart de X à l'espérance $\mathbb{E}(X)$.
  On majore dans ce théorème la probabilité que X s'écarte de $\mathbb{E}(X)$ de plus de t.
- Plus t est grand et plus la probabilité est faible.
- $V(X)$ est un critère de dispersion : pour t fixé, plus $V(X)$ est faible et plus le majorant fournit ici est faible.

Exemples :

• Si $\mathbb{E}(X) = 10$ et $\sigma(X) = 2$ alors $\mathbb{P}(|X - 10| \geqslant 6) \leqslant \dfrac{4}{36} = \dfrac{1}{9}$.

• Si X est une mesure expérimentale de moyenne m et d'écart-type σ, alors la probabilité que la mesure s'écarte de plus de 3σ de la moyenne est inférieure à $\dfrac{1}{9}$.

Remarques :

Ce théorème donne des informations sur X sans connaître la loi de X, seules $\mathbb{E}(X)$ et $V(X)$ interviennent.

Bien rappeler les conditions d'application de ce théorème [2] : $t > 0$ et X admet une variance.

Corollaires.

Soit X une variable aléatoire réelle,

❶ Si X admet une variance σ^2 alors $\forall t > 0, \quad \mathbb{P}\left(|X - \mathbb{E}(X)| < t\right) \geqslant 1 - \dfrac{\sigma^2}{t^2}$.

❷ Si X admet une variance σ^2 alors $\forall k > 0, \quad \mathbb{P}\left(|X - \mathbb{E}(X)| \geqslant k\sigma\right) \leqslant \dfrac{1}{k^2}$.

En effet :

Lien avec les intervalles de confiance.

$$\mathbb{P}\Big(|X - m| < t\Big) = \mathbb{P}\Big(m \in \left]X - t\,;\, X + t\right[\Big) \qquad \Big(\leqslant \mathbb{P}\Big(m \in \left[X - t\,;\, X + t\right]\Big)\Big)$$

Une des applications importantes de ce théorème est la démonstration de **la loi faible des grands nombres**.

2. Comme pour tous les théorèmes du cours

3.3 Loi faible des grands nombres.

Théorème :

Soit $(X_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires réelles, on définit : $M_n = \dfrac{X_1 + X_2 + \cdots X_n}{n}$

Si les X_n sont mutuellement indépendantes, admettant la même espérance m et la même variance alors :

$$\text{pour tout } \varepsilon > 0, \quad \lim_{n\to+\infty} \mathbb{P}\big(|M_n - m| \geqslant \varepsilon\big) = 0$$

Démonstration :

• La linéarité de $E(.)$ donne $E(M_n) = m$ et l'indépendance des X_k donne $V(M_n) = \dfrac{\sigma^2}{n}$

En appliquant à M_n l'inégalité de Bienaymé-Tchebychev on obtient :

$$\forall \varepsilon > 0, \quad P\left(|M_n - m| \geqslant \varepsilon\right) \leqslant \frac{\sigma^2}{n\varepsilon^2}$$

Pour ε un réel strictement positif fixé, on a :

$$\forall n \in \mathbb{N}^*, \quad 0 \leqslant P\left(|M_n - m| \geqslant \varepsilon\right) \leqslant \frac{\sigma^2}{n\varepsilon^2}$$

or on sait que $\lim\limits_{n\to+\infty} \dfrac{\sigma^2}{n\varepsilon^2} = 0$, donc *(théorème des gendarmes)*

$$\boxed{\text{Quel que soit } \varepsilon > 0, \quad \lim_{n\to+\infty} P\left(|M_n - m| \geqslant \varepsilon\right) = 0}$$

Interprétation : $\big(|M_n - m| \geqslant \varepsilon\big)$ est l'événement : "M_n s'éloigne de m de plus de ε"

Quand n est suffisamment grand, $\big(|M_n - m| \geqslant \varepsilon\big)$ est négligeable, ceci même pour un petit $\varepsilon > 0$.

Remarques :

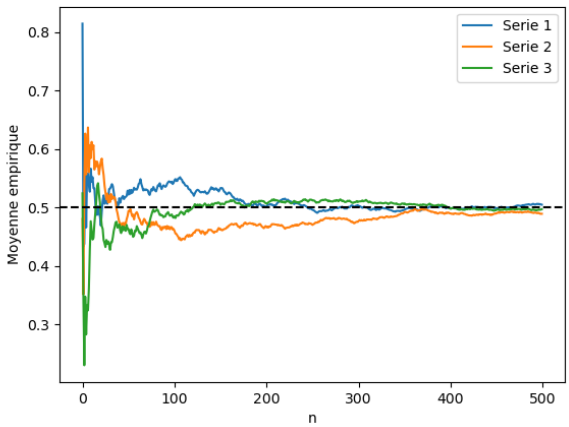
• Si $(X_1, \ldots, X_n)$ est un échantillon d'une loi mère X alors,
la moyenne empirique M_n *converge en probabilité* vers $E(X)$. *(Notion hors programme)*

• Quand on observe une expérience aléattoire et qu'on on obtient n valeurs numériques : $x_1, \ldots, x_n$,
pour n suffisamment grand, $\bar{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$ est une approximation de l'espérance m.

• C'est ce qui justifie qu'en informatique en réalisant un grand nombre de simulations on obtient une valeur approchée d'une espérance *(Une valeur moyenne)*. C'est le principe de la Méthode de Monte-Carlo.

```
def M_(n):
    M = [rd.random()]
    for k in range(1, n):
        M.append( (M[-1]*k+ rd.random())/(k+1)  )
    return M

for i in range(1, 4):
    E = M_(500)
    plt.plot(E, label = "Serie "+ str(i))
plt.xlabel("n")
plt.ylabel("Moyenne empirique")
plt.legend()
plt.axhline(0.5, color='black', linestyle='--')
plt.show()
```



Remarques :

Ce théorème donne des informations sur la suite (M_n) sans connaître la loi de X.

Bien rappeler les conditions d'applications de ce théorème [3] : X admet *une espérance et* une variance.

3. Comme pour tous les théorèmes du cours

Corollaire.

Soit $(X_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires suivant toutes la loi de Bernoulli de paramètre $p \in \left]0\,,1\right[$,
on définit : $F_n = \dfrac{X_1 + X_2 + \cdots X_n}{n}$
Si les X_n sont mutuellement indépendantes, alors

$$\text{pour tout } \varepsilon > 0, \quad \lim_{n\to+\infty} \mathbb{P}\big(|F_n - p| \geqslant \varepsilon\big) = 0$$

En effet :

(X_n) est une suite de variables aléatoires mutuellement indépendantes avec $E(X_k) = p$ et $V(X_k) = p(1-p)$.
On peut alors appliquer le théorème avec : $\boxed{E(F_n) = p}$

$$\text{Quel que soit } \varepsilon > 0, \quad \lim_{n\to+\infty} P\left(|F_n - p| \geqslant \varepsilon\right) = 0$$

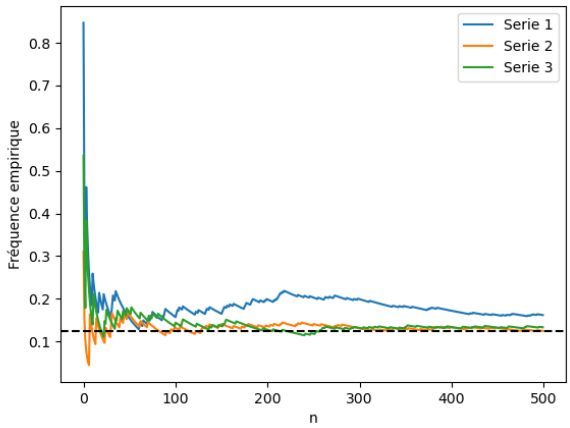
Interprétation : En considérant que l'événement $(X_k = 1)$ représente un succès, F_n est la proportion de succès.
$\big(|F_n - p| \geqslant \varepsilon\big)$ est l'événement : "F_n s'éloigne de p de plus de ε"
Quand n est suffisamment grand, $\big(|F_n - p| \geqslant \varepsilon\big)$ est négligeable, ceci même pour un petit $\varepsilon > 0$.

Remarques :

- Si $(X_1, ..., X_n)$ est un échantillon de la loi de Bernoulli $\mathscr{B}(p)$, alors
  la fréquence empirique F_n *converge en probabilité* vers p. *(Notion hors programme)*
- Si A est un événement d'une expérience que l'on répète n fois et $X_k = \mathbb{1}_A$ alors F_n est le nombre de réalisations de A divisé par le nombre total d'expériences *(la fréquence de réalisation de A)*
- Pour n suffisamment grand, la réalisation de F_n est une approximation de p.
- C'est ce qui justifie qu'en informatique on réalise un grand nombre de simulations pour obtenir une valeur approchée d'une probabilité. *(Méthode de Monte-Carlo)*

```
p = 1/8
def F_(n):
    f = [ bernoulli(p)]
    for k in range(1, n):
        f.append( (f[-1]*k+ bernoulli(p))/(k+1)  )
    return f

for i in range(1, 4):
    E = F_(500)
    plt.plot(E, label = "Serie "+ str(i))
plt.xlabel("n")
plt.ylabel("Fréquence empirique")
plt.legend()
plt.axhline(p, color='black', linestyle='--')
plt.show()
```



4

Echantillonnage. Estimation.

4.1 Echantillon

Définition. ***(complément)***

Soient $n \in \mathbb{N}^*$ et X une variable aléatoire,
On appelle n-échantillon de X un n-uplet $(X_1, ..., X_n)$ de variables aléatoires mutuellement indépendantes et de même loi que X.

Remarques :

- la loi de X est appelée la loi mère de l'échantillon.
- On dit que $X_1, ..., X_n$ sont n variables aléatoires indépendantes et identiquement distribuées. (i.i.d)
- On dit que le n-uplet $(x_1, ..., x_n) \in \mathbb{R}^n$ est une réalisation de l'échantillon de $(X_1, \ldots, X_n)$ lorsqu'il existe $\omega \in \Omega$ tel que :

$$(x_1, \ldots, x_n) = (X_1(\omega), \ldots, X_n(\omega))$$

Base de la statistique inférentielle :

En observant un phénomène aléatoire on obtient une série statistique $(x_1, \ldots, x_n)$ de n mesures.

On admet l'existence d'un espace probabilisé $(\Omega, \mathscr{T}, \mathbb{P})$ et d'une loi mère.

La statistique inférentielle a pour but de déterminer des informations sur cette loi mère à partir d'une réalisation $(x_1, \ldots, x_n)$.

4.2 Estimateur.

4.2.1 Définition. *Complément*

Définition :

Soit X une variable aléatoire dont la loi dépend d'un paramètre inconnu θ.
Soit $(X_1, \ldots, X_n)$ un échantillon de cette loi.
On appelle **estimateur** du paramètre θ toute variable aléatoire T_n qui est une fonction de l'échantillon $(X_1, \ldots, X_n)$:

$$T_n = g(X_1, \ldots, X_n).$$

Remarque : Un estimateur est une variable aléatoire, tandis qu'une estimation est la valeur obtenue lorsqu'on remplace l'échantillon aléatoire par des données observées.

4.2.2 Biais. *Complément*

Définition.

Soit T_n un estimateur d'un paramètre θ.
On appelle **biais de l'estimateur** T_n, le réel $\mathbb{E}(T_n) - \theta$

Remarques :

- Un estimateur T_n de θ est dit sans biais lorsque : $E(T_n) = \theta$
- Un estimateur T_n de θ est dit asymptotiquement sans biais lorsque : $\lim\limits_{n \to +\infty} E(T_n - \theta) = 0$

4.3 Moyenne empirique.

Définition.

Soient $n \in \mathbb{N}^*$ et $(X_1, ..., X_n)$ une liste de n variables aléatoires.

On appelle **moyenne empirique**, notée M_n, la variable aléatoire : $\quad M_n = \dfrac{1}{n}\sum_{k=1}^{n} X_k$

Remarques :

- Parfois cette variable est notée : $\overline{X_n}$ *(Cette notation est source d'erreurs)*
- Quand on observe une expérience on obtient n valeurs : $x_1, \ldots, x_n$, $\quad M_n$ prend alors la valeur $\bar{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$

Théorème.

Soient $n \in \mathbb{N}^*$ et $(X_1, ..., X_n)$, une liste de n variables aléatoires indépendantes de même loi, admettant une espérance μ et une variance σ^2 non nulle.

❶ M_n admet une espérance et $E(M_n) = \mu$,

❷ M_n admet une variance et $V(M_n) = \dfrac{\sigma^2}{n}$,

❸ La variable aléatoire centrée réduite associée à M_n, notée M_n^*, est égale à $\dfrac{M_n - \mu}{\frac{\sigma}{\sqrt{n}}}$

Démonstration. *(A savoir refaire)*

$M_n = \dfrac{1}{n}\sum_{k=1}^{n} X_k$ donc *(linéarité)* $E(M_n) = \dfrac{1}{n}\sum_{k=1}^{n} E(X_k)$ $\quad \boxed{E(M_n) = \mu}$

et *(indépendance)* $V(M_n) = \dfrac{1}{n^2}\sum_{k=1}^{n} V(X_k)$ $\quad \boxed{E(V_n) = \dfrac{\sigma^2}{n}}$

et comme M_n^*, est égale à $\dfrac{M_n - E(M_n)}{\sqrt{V(M_n)}}$ il vient : $\boxed{M_n^* = \dfrac{M_n - \mu}{\frac{\sigma}{\sqrt{n}}}}$

Remarques :

❶ M_n est un estimateur sans biais de μ.

❷ La variance de $M_n - \mu$ tend vers 0 quand n tend vers $+\infty$.

❸ On a vu que $\forall \varepsilon > 0, \quad \lim_{n \to +\infty} P(|M_n - \mu| \geqslant \varepsilon) = 0$ *(Loi Faible des grands nombres)*

Interprétations :

- Quand n est grand, la dispersion de M_n est petite.
- Quand n est grand, M_n fournit une approximation de μ.

4.4 Ecart-type empirique.

Définition.

Soient $n \in \mathbb{N}^*$ et $(X_1, ..., X_n)$, une liste de n variables aléatoires.

On appelle variance empirique la variable aléatoire : ${S_n}^2 = \dfrac{1}{n}\sum_{k=1}^{n}(X_k - M_n)^2$

On appelle **écart-type empirique** la variable aléatoire : $S_n = \sqrt{\dfrac{1}{n}\sum_{k=1}^{n}(X_k - M_n)^2}$

Proposition. *(Autre expression de la variance empirique).*

Soient $n \in \mathbb{N}^*$ et $(X_1, ..., X_n)$, une liste de n variables aléatoires.

$$\frac{1}{n}\sum_{k=1}^{n}(X_k - M_n)^2 = \frac{1}{n}\left(\sum_{k=1}^{n} X_k^2\right) - M_n^2$$

Démonstration. *(A savoir refaire)*

$$\begin{aligned} S_n^2 &= \frac{1}{n}\sum_{k=1}^{n}(X_k - M_n)^2 \\ &= \frac{1}{n}\sum_{k=1}^{n}\left(X_k^2 - 2M_nX_k + M_n^2\right) \\ &= \frac{1}{n}\sum_{k=1}^{n}X_k^2 - \frac{2M_n}{n}\sum_{k=1}^{n}X_k + \frac{1}{n}\sum_{k=1}^{n}M_n^2 \\ &= \frac{1}{n}\sum_{k=1}^{n}X_k^2 \ - \ 2M_n^2 \ + \ M_n^2 \end{aligned}$$

$$\boxed{S_n^2 = \frac{1}{n}\left(\sum_{k=1}^{n}{X_k}^2\right) \ - \ {M_n}^2}$$

Proposition. *(complément)*

Soient $n \in \mathbb{N}^*$ et $(X_1, ..., X_n)$, une liste de n variables aléatoires indépendantes de même loi, admettant une espérance μ et une variance σ^2 non nulle.

La variance empirique admet une espérance égale à $\dfrac{n-1}{n}\sigma^2$.

Démonstration. *(à savoir refaire)*

$S_n^2 = \dfrac{1}{n}\left(\sum_{k=1}^{n}{X_k}^2\right) \ - \ {M_n}^2$ donc *(linéarité)*

$$\begin{aligned} E(S_n^2) &= \frac{1}{n}\sum_{k=1}^{n}E({X_k}^2) \ - \ E(M_n^2) \\ &= \frac{1}{n}\sum_{k=1}^{n}(V(X_k) + E(X_k)^2) - (V(M_n) + E(M_n)^2) \qquad \text{En effet : } \underbrace{E(X^2) = V(X) + E(X)^2}_{\text{Koenig - Huygens}} \\ &= \frac{1}{n}\sum_{k=1}^{n}(\sigma^2 + \mu^2) - \left(\frac{\sigma^2}{n} + \mu^2\right) \\ &= \sigma^2 - \frac{\sigma^2}{n} \end{aligned}$$

$$\boxed{S_n^2 = \frac{(n-1)\sigma^2}{n}}$$

Remarques :

- $E(S_n^2 - \sigma^2) = -\dfrac{\sigma^2}{n}$ *(S_n^2 est un estimateur asymptotiquement sans biais de σ^2)*
- En posant ${S'_n}^2 = \dfrac{n}{n-1}S_n^2$, on a $E({S'_n}^2 - \sigma^2) = 0$ *(${S'_n}^2$ est un estimateur sans biais de σ^2)*

Définition.

On appelle **écart-type empirique corrigé** la variable aléatoire : $S'_n = \sqrt{\dfrac{1}{n-1}\sum_{k=1}^{n}(X_k - M_n)^2}$

5 Convergence en loi

Définition

Soient X une variable aléatoire réelle et $(X_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires réelles.
On note F la fonction de répartition de X et D l'ensemble des réels où F est discontinue.
Dire que (X_n) converge en loi vers X signifie que : $\forall x \in \mathbb{R}\setminus D, \quad \lim\limits_{n\to+\infty} \mathbb{P}(X_n \leqslant x) = F(x)$

Remarques :

1. On note : $X_n \xrightarrow{\mathscr{L}} X$
2. On doit pour chaque $x \in \mathbb{R}\setminus D$ calculer la limite de la suite $((\mathbb{P}(X_n \leqslant x))_{n\in\mathbb{N}}$.
3. On admet que l'ensemble D est au plus dénombrable.
4. La convergence en loi permet d'approcher la loi de X_n par celle de X lorsque n est assez grand.

Exemples : *Voir la feuille_info_22.*

Caractérisation *(quand X est à densité)*

Soient X une variable aléatoire réelle **à densité** et $(X_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires réelles.
(X_n) converge en loi vers X si, et seulement si, $\forall x \in \mathbb{R}, \quad \lim\limits_{n\to+\infty} \mathbb{P}(X_n \leqslant x) = \mathbb{P}(X \leqslant x)$

En effet :

Caractérisation *(quand X et X_n sont discrètes à valeurs dans $\mathbb{N}$)*

Soient X une variable aléatoire réelle à valeurs dans $\mathbb{N}$ et $(X_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires réelles à valeurs dans $\mathbb{N}$.
(X_n) converge en loi vers X si, et seulement si, $\forall k \in \mathbb{N}, \quad \lim\limits_{n\to+\infty} \mathbb{P}(X_n = k) = \mathbb{P}(X = k)$

Démonstration. *(admis)*

Exemple : Si $X_n \hookrightarrow \mathscr{B}\left(n, \dfrac{\lambda}{n}\right)$ alors $X_n \xrightarrow{\mathscr{L}} X$ avec $X \hookrightarrow \mathscr{P}(\lambda)$.

Remarque : Dans le théorème de la loi faible des grands nombres nous verrons une autre convergence.

Définition *(complément)*

Soient X une variable aléatoire réelle et $(X_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires réelles.
Dire que (X_n) **converge en probabilité** vers X signifie que,

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

6

Théorème central limite

6.1 Première forme.

6.1.1 Avec la moyenne empirique.

Théorème.

Soit $(X_n)_{n\geqslant 1}$ une suite de variables aléatoires, on définit pour tout $n \in \mathbb{N}^*$: $M_n = \dfrac{1}{n}\displaystyle\sum_{k=1}^{n} X_k$

Si les X_n sont indépendantes de même loi, admettant une espérance μ et une variance σ^2 non nulle

alors $(M_n^*)_{n\geqslant 1}$ converge en loi vers une variable suivant la loi normale centrée réduite.

Remarques :

❶ On rappelle que : $M_n^* = \dfrac{M_n - \mu}{\frac{\sigma}{\sqrt{n}}}$

❷ On note : $M_n^* \xrightarrow{\mathscr{L}} X$ avec $X \hookrightarrow \mathcal{N}(0,1)$ ou encore $M_n^* \xrightarrow{\mathscr{L}} \mathcal{N}(0,1)$

❸ Autrement dit :

❹ Dans le cas de gaussiennes identiques et indépendantes M_n^* suit exactement la loi normale centrée réduite.

❺ *(complément)*

$\left[M_n - u_{1-\frac{\alpha}{2}}\dfrac{\sigma}{\sqrt{n}}; M_n + u_{1-\frac{\alpha}{2}}\dfrac{\sigma}{\sqrt{n}}\right]$ est un intervalle de confiance asymptotique au niveau $1-\alpha$ pour μ.

Interprétation : La moyenne d'un grand nombre de variables aléatoires indépendantes devient approximativement gaussienne.

Théorème. *(Autre formulation)*

Si $(X_n)_{n\geqslant 1}$ est une suite de variables aléatoires indépendantes de même loi admettant une espérance μ et une variance σ^2 *(non nulle)* alors :

Pour tout (a,b) tel que $a \leqslant b$, $\displaystyle\lim_{n\to+\infty} \mathbb{P}\left(a \leqslant M_n^* \leqslant b\right) = \frac{1}{\sqrt{2\pi}}\int_a^b e^{-\frac{t^2}{2}}\,\mathrm{d}t$

6.1.2 Avec la somme.

Théorème.

Soit $(X_n)_{n\geqslant 1}$ une suite de variables aléatoires, on définit pour tout $n \in \mathbb{N}^*$: $Y_n = \displaystyle\sum_{k=1}^{n} X_k$

Si les X_n sont indépendantes de même loi, admettant une espérance μ et une variance σ^2 non nulle

alors $(Y_n^*)_{n\geqslant 1}$ converge en loi vers une variable suivant la loi normale centrée réduite.

En effet : En remarquant que $Y_n = nM_n$ on montre facilement que ${Y_n}^* = {M_n}^*$. *Remarque :* ${Y_n}^* = \dfrac{Y_n - n\mu}{\sigma\sqrt{n}}$

Interprétation : La somme de nombreuses contributions indépendantes devient approximativement gaussienne.

6.1.3 Un cas particulier

Théorème de Moivre-Laplace. *(corollaire du théorème précédent).*

Soient $p \in]0,1[$ et $(Y_n)_{n\geqslant 1}$ une suite de variables aléatoires réelles,
Si Y_n suit la loi binomiale de paramètres (n,p) alors
$\dfrac{Y_n - np}{\sqrt{np(1-p)}}$ converge en loi vers une variable aléatoire suivant la loi normale centrée réduite.

En effet : On applique le théorème précédent avec $X_k \hookrightarrow \mathscr{B}(p)$

Remarque : $E(Y_n) = np$, $V(Y_n) = np(1-p)$ donc $Y_n^* = \dfrac{Y_n - np}{\sqrt{np(1-p)}}$.

Autrement dit :

Si $Y_n \hookrightarrow \mathscr{B}(n,p)$ alors : pour tout $x \in \mathbb{R}$, $\lim\limits_{n\to+\infty} \mathbb{P}(Y_n^* \leqslant x) = \Phi(x)$
où Φ est la fonction de répartition de la loi $\mathcal{N}(0;1)$.

6.1.4 Approximations.

Pour $(X_1, ..., X_n)$ une liste de n variables aléatoires i.i.d. d'espérance μ et de variance $\sigma^2 \neq 0$.

- **Avec la moyenne empirique.**
  Pour n assez grand, la loi de M_n^* peut être approchée par la loi normale centrée réduite.
  Pour n assez grand, la loi de M_n peut être approchée par la loi $\mathcal{N}(\mu, \frac{\sigma^2}{n})$.
  Pour n assez grand , $\left[M_n \pm u_{1-\frac{\alpha}{2}} \dfrac{\sigma}{\sqrt{n}}\right]$ est un IC au niveau $1-\alpha$ pour μ.
- **Avec la somme Y_n.**
  Pour n assez grand, la loi de Y_n^* peut être approchée par la loi normale centrée réduite.
  Pour n assez grand, la loi de $Y_n = \sum\limits_{k=1}^{n} X_k$ peut être approchée par la loi $\mathcal{N}(n\mu, n\sigma^2)$.
- **Avec $Y_n \hookrightarrow \mathscr{B}(n,p)$.**
  Pour n assez grand, la loi de $\dfrac{Y_n - np}{\sqrt{np(1-p)}}$ peut être approchée par la loi $\mathcal{N}(0,1)$.
  Pour n assez grand, la loi de Y_n peut être approchée par la loi $\mathcal{N}(np, np(1-p))$.

Remarque : "La loi de X peut être approchée par la loi de Y" se note souvent $X \approx Y$ et signifie :

- si Y est à densité : $\forall x \in \mathbb{R},\ P(X \leqslant x) \approx P(Y \leqslant x)$ (au sens valeur approchée dans $\mathbb{R}$).
- si X et Y sont discrètes à valeurs dans $\mathbb{N}$: $\forall k \in \mathbb{N},\ P(X = k) \approx P(Y = k)$
  (au sens valeur approchée dans $\mathbb{R}$).

6.1.5 Critères d'approximation pour les lois usuelles.

Loi de départ	**Conditions**	**Approximation**
$\mathcal{B}(n,p)$	$n \geqslant 30$ et $np \leqslant 10$	$\mathcal{P}(np)$
$\mathcal{B}(n,p)$	$n \geqslant 20$ et $p \approx 0{,}5$	$\mathcal{N}(np,\ np(1-p))$
$\mathcal{B}(n,p)$	$np \geqslant 10$ et $n(1-p) \geqslant 10$	$\mathcal{N}(np,\ np(1-p))$
$\mathcal{P}(\lambda)$	$\lambda \geqslant 10$	$\mathcal{N}(\lambda,\ \lambda)$

Ces critères ne sont pas des théorèmes mais des règles pratiques. Dans un sujet ils devraient vous êtes rappelés.

6.2 Seconde forme

6.2.1 Avec l'écart-type empirique.

Théorème.

Soit $(X_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires, on définit pour tout $n \in \mathbb{N}^*$:

$$M_n = \frac{1}{n}\sum_{k=1}^{n} X_k \qquad \text{et} \qquad S_n = \sqrt{\frac{1}{n}\sum_{k=1}^{n}(X_k - M_n)^2}$$

Si les X_n sont indépendantes de même loi, admettant une espérance μ et une variance non nulle *(inconnue)*

alors $\left(\frac{M_n - \mu}{\frac{S_n}{\sqrt{n}}}\right)_{n\geqslant 1}$ converge en loi vers une variable suivant la loi normale centrée réduite.

Remarques :

- Ici on remplace l'écart-type par l'écart-type empirique S_n.
- Autrement dit $\forall x \in \mathbb{R}, \quad \lim_{n\to+\infty} \mathbb{P}\left(\frac{M_n - \mu}{\frac{S_n}{\sqrt{n}}} \leqslant x\right) = \Phi(x)$

où Φ est la fonction de répartition de la loi $\mathcal{N}(0;1)$.

Théorème. *(Autre formulation)*

Si $(X_n)_{n\in\mathbb{N}}$ est une suite de variables aléatoires indépendantes de même loi admettant une espérance μ et une variance σ^2 *(non nulle)* alors :

Pour tout a, b tel que $a \leqslant b$, $\lim_{n\to+\infty} \mathbb{P}\left(a \leqslant \frac{M_n - \mu}{\frac{S_n}{\sqrt{n}}} \leqslant b\right) = \frac{1}{\sqrt{2\pi}}\int_a^b e^{-\frac{t^2}{2}}\,\mathrm{d}t$

6.2.2 Avec une fréquence empirique *(complément)*

Corollaire.

Si $(X_n)_{n\in\mathbb{N}}$ est une suite de variables aléatoires indépendantes suivant la loi de Bernoulli de paramètre p alors :

en notant $F_n = \frac{1}{n}\sum_{k=1}^{n} X_k$ *(la fréquence empirique)*

$\frac{\sqrt{n}(F_n - p)}{\sqrt{F_n(1-F_n)}}$ converge en loi vers une variable aléatoire suivant la loi normale centrée réduite.

En effet :

Remarques :

- $\frac{\sqrt{n}(F_n - p)}{\sqrt{F_n(1-F_n)}} = \frac{F_n - p}{\frac{\sqrt{F_n(1-F_n)}}{\sqrt{n}}}$
- Ici $S_n^2 = F_n(1-F_n)$. *(Ici, il est plus simple de calculer S_n^2)*

6.2.3 Avec l'écart-type empirique corrigé.

Théorème.

Soit $(X_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires, on définit pour tout $n \in \mathbb{N}^*$:

$$M_n = \frac{1}{n}\sum_{k=1}^{n} X_k \qquad \text{et} \qquad S'_n = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(X_k - M_n)^2}$$

Si les X_n sont indépendantes de même loi, admettant une espérance μ et une variance *(inconnue)*

alors $\left(\frac{M_n - \mu}{\frac{S'_n}{\sqrt{n}}}\right)_{n\geqslant 1}$ converge en loi vers une variable suivant la loi normale centrée réduite.

Remarques :

• *(complément)*
Lorsque les X_k sont des lois normales, alors $\frac{M_n - \mu}{\frac{S'_n}{\sqrt{n}}}$ suit exactement la loi de Student à $n-1$ degrés de liberté.

• Les intervalle de confiance de la moyenne sont une application de ce théorème.

• La loi de Student n'est rigoureusement valable que dans le cas normal. Son utilisation dans les autres cas repose sur une approximation.

• En pratique, on utilise la loi de Student à $n-1$ degrés de liberté pour des petites valeurs de n.

7

Intervalle de confiance de la moyenne.

On observe une expérience aléatoire, plus particulièrement une variable aléatoire X qui admet une espérance μ et une variance σ^2 inconnues.

Proposition.

Soit $(X_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires, on définit pour tout $n \in \mathbb{N}^*$:

$$M_n = \frac{1}{n}\sum_{k=1}^{n} X_k \qquad \text{et} \qquad S_n = \sqrt{\frac{1}{n}\sum_{k=1}^{n}(X_k - M_n)^2}$$

Si les X_n sont indépendantes de même loi, admettant une espérance μ et une variance *(inconnues)* alors

$$\forall \alpha \in]0,1[, \quad \lim_{n\to+\infty}\left(\mathbb{P}\left(\left[|M_n - \mu| \leqslant u_{1-\frac{\alpha}{2}}\frac{S_n}{\sqrt{n}}\right]\right)\right) = 1-\alpha$$

Démonstration. *C'est une conséquence de la seconde forme du théorème central limite.*

Pour $(X_1, ..., X_n)$ un échantillon d'une loi d'espérance μ et de variance σ^2

Pour n assez grand ,

$\left[M_n - u_{1-\frac{\alpha}{2}}\frac{S_n}{\sqrt{n}}\,;\, M_n + u_{1-\frac{\alpha}{2}}\frac{S_n}{\sqrt{n}}\right]$ est un intervalle de confiance pour μ au niveau de confiance $1-\alpha$.

On parle aussi d'intervalle de confiance asymptotique.

En pratique : *(On ne connait pas μ et σ on cherche à estimer μ avec un intervalle de confiance)*

❶ On lit avec soin l'énoncé et on détermine :

n : la taille de l'échantillon et $(x_1, ..., x_n)$ les valeurs prises par l'échantillon.

(attention parfois les données sont sous la forme valeurs/effectifs)

❷ On calcule $u_{1-\frac{\alpha}{2}}$ avec une table ou avec Python ou avec la calculatrice.

❸ On calcule m_n la moyenne empirique, s_n^2 la variance empirique, $m_n - u_{1-\frac{\alpha}{2}}\frac{s_n}{\sqrt{n}}$ et $m_n + u_{1-\frac{\alpha}{2}}\frac{s_n}{\sqrt{n}}$

❹ On conclut :

Au seuil de α on estime que μ est dans l'intervalle de confiance $\left[m_n - u_{1-\frac{\alpha}{2}}\frac{s_n}{\sqrt{n}}\ ;\ m_n + u_{1-\frac{\alpha}{2}}\frac{s_n}{\sqrt{n}}\right]$.

Remarque. *Cette notion est souvent représentée dans les barres d'erreurs en sciences expérimentales.*

Exemple : Au risque de 5% la moyenne théorique est dans l'intervalle : $\left[m_n - 1{,}96\frac{s_n}{\sqrt{n}}\ ;\ m_n + 1{,}96\frac{s_n}{\sqrt{n}}\right]$.

Proposition. *(Avec des variables de Bernoulli)*

Avec des variables de Bernoulli de paramètre p *(inconnue)*, on obtient un intervalle de confiance de p :
(*au niveau de confiance* $1-\alpha$)

$$\forall \alpha \in]0,1[, \quad \mathbb{P}\left(p \in \left[F_n - u_{1-\frac{\alpha}{2}} \frac{\sqrt{F_n(1-F_n)}}{\sqrt{n}} \ ; \ F_n + u_{1-\frac{\alpha}{2}} \frac{\sqrt{F_n(1-F_n)}}{\sqrt{n}} \right]\right) \underset{n\to+\infty}{\longrightarrow} 1-\alpha$$

Démonstration. *l'idée c'est qu'ici* $S_n^2 = F_n(1-F_n)$*, vu dans la feuille_cours_14.*

En pratique : *(On ne connait pas p on cherche à estimer p avec un intervalle de confiance)*

❶ On lit avec soin l'énoncé et on détermine :
n : la taille de l'échantillon et $(x_1, ..., x_n)$ les valeurs prises par l'échantillon.

❷ On calcule $u_{1-\frac{\alpha}{2}}$ avec une table ou avec Python ou avec la calculatrice.

❸ On calcule f_n la fréquence empirique, on sait alors que la variance empirique est $f_n(1-f_n)$.

❹ On calcule $f_n - u_{1-\frac{\alpha}{2}} \frac{\sqrt{f_n(1-f_n)}}{\sqrt{n}}$ et $f_n + u_{1-\frac{\alpha}{2}} \frac{\sqrt{f_n(1-f_n)}}{\sqrt{n}}$

❺ On conclut :
Au risque de α on estime que p est dans l'intervalle de confiance $\left[f_n - u_{1-\frac{\alpha}{2}} \frac{\sqrt{f_n(1-f_n)}}{\sqrt{n}} \ ; \ f_n + u_{1-\frac{\alpha}{2}} \frac{\sqrt{f_n(1-f_n)}}{\sqrt{n}} \right]$.

Exemple. Pour $\alpha = 5\%$, on a : $u_{1-\frac{\alpha}{2}} \approx 2$ et comme $f_n(1-f_n) \leqslant \frac{1}{4}$ alors :

Au seuil de 5% on estime que p est dans l'intervalle de confiance $\left[f_n - \frac{1}{\sqrt{n}} \ ; \ f_n + \frac{1}{\sqrt{n}} \right]$

– – – – – – – – – – – – – – – – – –

Intervalle de confiance de Student.

Théorème *(Complément)*

Soit $(X_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires, on définit pour tout $n \in \mathbb{N}^*$:

$$M_n = \frac{1}{n}\sum_{k=1}^{n} X_k \qquad \text{et} \qquad S'_n = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n} (X_k - M_n)^2}$$

Si les X_n sont indépendantes suivant toute **la même loi normale** d'espérance μ et de variance *(inconnues)* alors

$$\forall \alpha \in]0,1[, \quad \mathbb{P}\left(\mu \in \left[M_n - t_{1-\frac{\alpha}{2}}^{(n-1)} \frac{S'_n}{\sqrt{n}} \ ; \ M_n - t_{1-\frac{\alpha}{2}}^{(n-1)} \frac{S'_n}{\sqrt{n}} \right]\right) = 1-\alpha$$

où $t_{1-\alpha}^{(n)}$ *est le quantile d'ordre* $1-\alpha$ *de la Loi de Student avec n degrés de liberté.*

Remarques :

- Lorsque les X_k sont des gaussiennes $\dfrac{M_n - \mu}{\frac{S'_n}{\sqrt{n}}}$ suit la loi de Student.
- Simulation des lois du khi-2 et de Student.
- Elaboration d'un test du khi-2.

8

Test de conformité à la moyenne.

On observe une expérience aléatoire, plus particulièrement une variable aléatoire X qui admet une espérance μ et une variance σ^2 inconnues.

On veut tester l'hypothèse (H_0) : $\mu = \mu_0$. *(Contre l'hypothèse alternative* (H_1) : $\mu \neq \mu_0$*)*

Pour cela on prend *(la réalisation)* d'un échantillon $(X_1, ..., X_n)$ de la variable aléatoire X.

Sous l'hypothèse H_0 on sait que : (*seconde forme du théorème central limite*)

$$\lim_{n\to+\infty} \mathbb{P}\left(\left[-u_{1-\frac{\alpha}{2}} \leqslant \frac{M_n - \mu_0}{\frac{S_n}{\sqrt{n}}} \leqslant u_{1-\frac{\alpha}{2}}\right]\right) = 1 - \alpha$$

ou encore

$$\lim_{n\to+\infty} \mathbb{P}\left(\left|\frac{M_n - \mu_0}{\frac{S_n}{\sqrt{n}}}\right| > u_{1-\frac{\alpha}{2}}\right) = \alpha$$

Donc pour n assez grand, $\dfrac{M_n - \mu_0}{\frac{S_n}{\sqrt{n}}}$ en dehors du segment $\left[-u_{1-\frac{\alpha}{2}}\,;u_{1-\frac{\alpha}{2}}\right]$ avec une probabilité proche de α.

Quand $\dfrac{M_n - \mu_0}{\frac{S_n}{\sqrt{n}}}$ prend une valeur en dehors du segment $\left[-u_{1-\frac{\alpha}{2}}\,;u_{1-\frac{\alpha}{2}}\right]$, on dira qu'on rejette l'hypothèse (H_0) au seuil de α.

Le plus souvent, on prend $\alpha = 5\%$,

Quand $\dfrac{M_n - \mu_0}{\frac{S_n}{\sqrt{n}}}$ prend une valeur en dehors de l'intervalle $[1,96; 1{,}96]$),

on prend la décision de rejetter l'hypothèse (H_0) au seuil de 5%.

En pratique :

❶ On lit avec soin l'énoncé et on détermine :

n : la taille de l'échantillon, $(x_1, ..., x_n)$ les valeurs de l'échantillon et μ_0 la valeur supposée de l'espérance.

❷ On calcule $m_n = \dfrac{1}{n}\sum_{k=1}^{n} x_k$ la moyenne empirique, $s_n^2 = \dfrac{1}{n}\sum_{k=1}^{n} x_k^2 - m_n^2$ la variance empirique et $\dfrac{m_n - \mu_0}{\frac{s_n}{\sqrt{n}}}$

❸ On calcule $u_{1-\frac{\alpha}{2}}$ avec une table ou avec Python ou avec la calculatrice.

❹ On teste l'hypothèse au seuil de α :

si $\dfrac{m_n - \mu_0}{\frac{s_n}{\sqrt{n}}} \notin \left[-u_{1-\frac{\alpha}{2}}\,;u_{1-\frac{\alpha}{2}}\right]$ alors : on rejette l'hypothèse (H_0) au seuil de α.

si $\dfrac{m_n - \mu_0}{\frac{s_n}{\sqrt{n}}} \in \left[-u_{1-\frac{\alpha}{2}}\,;u_{1-\frac{\alpha}{2}}\right]$ alors : on ne peut rien conclure de ce test.

Test de conformité d'une proportion.

Ici on note $\mu = p$ et $\mu_0 = p_0$,

et on veut tester l'hypothèse (H_0) : $p = p_0$. *(Contre l'hypothèse alternative (H_1) : $p \neq p_0$)*

En pratique :

❶ On lit avec soin l'énoncé et on détermine :
n : la taille de l'échantillon, $(x_1, ..., x_n)$ les valeurs de l'échantillon et p_0 la valeur supposée de la proportion.

❷ On calcule $f_n = \frac{1}{n}\sum_{k=1}^{n} x_k$ la fréquence empirique, on sait alors que $s_n^2 = f_n(1-f_n)$

❸ On calcule $\frac{f_n - p_0}{\frac{\sqrt{f_n(1-f_n)}}{\sqrt{n}}}$

❹ On calcule $u_{1-\frac{\alpha}{2}}$ avec une table ou avec Python ou avec la calculatrice.

❺ On teste l'hypothèse au seuil de α :

si $\frac{f_n - p_0}{\frac{\sqrt{f_n(1-f_n)}}{\sqrt{n}}} \notin \left[-u_{1-\frac{\alpha}{2}} ; u_{1-\frac{\alpha}{2}}\right]$ alors : on rejette l'hypothèse (H_0) au seuil de α.

si $\frac{f_n - p_0}{\frac{\sqrt{f_n(1-f_n)}}{\sqrt{n}}} \in \left[-u_{1-\frac{\alpha}{2}} ; u_{1-\frac{\alpha}{2}}\right]$ alors : on ne peut rien conclure de ce test.

Remarques :

- Ici on peut remplacer $\frac{f_n - p_0}{\frac{\sqrt{f_n(1-f_n)}}{\sqrt{n}}}$ par $\frac{f_n - p_0}{\frac{\sqrt{p_0(1-p_0)}}{\sqrt{n}}}$ *(Car on est sous l'hypothèse $p = p_0$)*

- Pour $\alpha = 5\%$ en utilisant la majoration $f_n(1-f_n) \leqslant \frac{1}{4}$ cela devient :

  si $\frac{f_n - p_0}{\frac{1}{\sqrt{n}}} \notin [-1; 1]$ alors : on rejette l'hypothèse (H_0) au seuil de α.

  si $\frac{f_n - p_0}{\frac{1}{\sqrt{n}}} \in [-1; 1]$ alors : on ne peut rien conclure de ce test.

– –