Table des matières

1 Statistique univariée. 2
1.1 Les résultats d'une expérience. 2
1.2 Des nombres caractéristiques. 3
1.2.1 Caractéristiques de position. 3
1.2.2 Caractéristiques de dispersion. 4

2 Statistique bivariée. 5
2.1 Nuage de points. 5
2.2 Covariance et coefficient de corrélation linéaire. 5
2.3 Droite de régression linéaire 6

3 Loi faible des grands nombres. 7
3.1 Inégalité de Markov. 7
3.2 Inégalité de Bienaymé-Tchebychev. 7
3.3 Loi faible des grands nombres. 8

4 Echantillonnage. Estimation. 9
4.1 Echantillon 9
4.2 Estimateur. 9
4.2.1 Définition. *Complément* 9
4.2.2 Biais. *Complément* 9
4.3 Moyenne empirique. 10
4.4 Ecart-type empirique. 10

5 Convergence en loi 11

6 Théorème central limite 12
6.1 Première forme. 12
6.2 Seconde forme 13

7 Intervalle de confiance de la moyenne. 14

8 Test de conformité à la moyenne. 16

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

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, $\quad c_i = \displaystyle\sum_{k=1}^{i} e_k$

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, $\quad F_i = \displaystyle\sum_{k=1}^{i} f_k$

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 $\tilde{x}_{\frac{n+1}{2}}$

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

Si on note 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 $\tilde{x}_{\frac{n-1}{2}}$

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

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

2.2 Covariance et coefficient de corrélation linéaire.

Définition :

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

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

Proposition :

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

Définition

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

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

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 $M_1, M_2, \ldots, M_n$ 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 :

$$a = \frac{\mathrm{cov}(x, y)}{\sigma_x^2} \qquad \text{et} \qquad b = \overline{y} - a\overline{x}$$

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 :

$$\alpha = \frac{\mathrm{cov}(x, y)}{\sigma_y^2} \qquad \text{et} \qquad \beta = \overline{x} - \alpha\overline{y}$$

- 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 : *Voir feuille_cours_11.*

3.2 Inégalité de Bienaymé-Tchebychev.

Théorème :

Soit X une variable aléatoire réelle,

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

Démonstration : *Voir feuille_cours_11.*

Remarques :

- La variable aléatoire $|X - \mathbb{E}(X)|$ est l'écart de X à l'espérance $\mathbb{E}(X)$.
- On majore la probabilité que X s'écarte de $\mathbb{E}(X)$ de plus de t.
- Plus t est grand et plus la probabilité est faible.
- Seule condition d'application : X admet une variance.
- σ critère de dispersion : pour t fixé, si σ est faible alors la probabilité est faible.
- Parfois cette inégalité ne donne aucune information. $(t < \sigma)$

Corollaires.

Soit X une variable aléatoire réelle,

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

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

Démonstrations :

Pour ❶ il suffit de remarquer que l'événement contraire de $\Big[|X - \mathbb{E}(X)| < t\Big]$ est : $\Big[|X - \mathbb{E}(X)| \geqslant t\Big]$.

Pour ❷ il suffit de remplacer t par $k\sigma$ dans l'inégalité du théorème.

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**.

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 σ^2 alors :

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

Démonstration : *Voir feuille_Cours_11.*

Remarques :

- Si $(X_1, ..., X_n)$ est un échantillon d'une loi mère X alors la moyenne empirique M_n converge en probabilité vers $E(X)$. *(Notion hors programme)*
- Dans les conditions du théorèmes : $E(M_n) = m$ et $V(M_n) = \dfrac{\sigma^2}{n}$ *(Calculs à savoir refaire avec justifications)*
- En faisant la moyenne de n mesures faites dans les mêmes conditions on estime une moyenne théorique.
- $M_n = \dfrac{X_1 + X_2 + \cdots X_n}{n}$ est une variable aléatoire appelée moyenne empirique.
- Quand on observe l'expérience on obtient n valeurs : $x_1, x_2, \ldots, x_n$,

$$M_n \text{ prend alors la valeur } \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- Pour n "suffisamment grand", $\bar{x} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_i$ est une approximation de m.
- C'est ce qui justifie qu'en informatique on réalise un grand nombre de simulations pour trouver une valeur approchée d'une espérance *(Une valeur moyenne). (Méthode de Monte-Carlo)*

Corollaire.

Soit $(X_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires réelles et p un réel de $]0, 1[$,
on définit : $F_n = \dfrac{X_1 + X_2 + \cdots X_n}{n}$ *(fréquence empirique du succès)*
Si les X_n sont mutuellement indépendantes, et suivent toutes la loi de Bernoulli de paramètre p

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

Démonstration : *Voir feuille_Cours_11.*

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 répéte n fois et $X_k = \mathbb{1}_A$ alors F_n est le nombre de réalisations de A divisé par le nombre total *(la fréquence de réalisation de A)*
- Pour n "suffisamment grand", la fréquence empirique f_n est une approximation de p.
- C'est ce qui justifie qu'en informatique on réalise un grand nombre de simulations pour trouver une valeur approchée d'une probabilité. *(Méthode de Monte-Carlo)*
- Lors d'un sondage, une proportion sur l'échantillon permet d'estimer (plus ou moins bien) une proportion sur la population.

4

Echantillonnage. Estimation.

Soit un espace probabilisé $(\Omega, \mathscr{T}, \mathbb{P})$ quelconque.

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 identiques et indépendantes suivant la loi de X.
- 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 $(X_1, ..., X_n)$ lorsque :

$$\exists \omega \in \Omega, \quad (x_1, ..., x_n) = (X_1(\omega), X_2(\omega), \dots, X_n(\omega))$$

Base de la statistique inférentielle :

En observant un phénomène aléatoire on obtient une série statistique $(x_1, ..., x_n)$ de n mesures, on admet l'existence d'un espace probabilisé $(\Omega, \mathscr{T}, \mathbb{P})$ et d'une loi mère ; les statistiques inférentielles ont pour but *(entre autres)* de déterminer des informations sur les paramètres de la loi mère à partir d'une réalisation $(x_1, ..., x_n)$.

4.2 Estimateur.

4.2.1 Définition. *Complément*

Définition :

Soit X une variable aléatoire,
On appelle estimateur d'un paramètre θ de la loi de X est un suite de variables aléatoires réelles (T_n) telle que pour chaque n, T_n s'expriment en fonction de tout échantillon $(X_1, ..., X_n)$ de la loi de X "et *donne une information* sur le paramètre θ"

4.2.2 Biais. *Complément*

Définition.

Soit T_n un estimateur d'un paramètre θ.
On appelle **biais de l'estimateur** T_n, l'espérance de la variable aléatoire $T_n - \theta$.

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 : $M_n = \dfrac{1}{n}\displaystyle\sum_{k=1}^{n} X_k$

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) (Voir la feuille cours_14)*

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 $\forall \varepsilon > 0, \quad \lim\limits_{n \to +\infty} P(|M_n - \mu| \geqslant \varepsilon) = 0$ *(Loi Faible des grands nombres)*

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 : ${S_n}^2 = \dfrac{1}{n}\displaystyle\sum_{k=1}^{n}(X_k - M_n)^2$

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

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

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

Démonstration. *(A savoir refaire) (Voir la feuille cours_14)*

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. *(calcul classique fait dans la feuille cours_14)*

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. *Complément*

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

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 continue.
Dire que (X_n) converge en loi vers X signifie que : $\forall x \in D, \quad \lim\limits_{n\to+\infty} \mathbb{P}(X_n \leqslant x) = F(x)$

Remarques :

❶ On note : $X_n \xrightarrow{\mathscr{L}} X$

❷ On doit pour chaque $x \in D$ calculer la limite de la suite $((\mathbb{P}(X_n \leqslant x))_{n\in\mathbb{N}}$.

Exemples : *Voir la feuille_info_13.*

Proposition

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 :

Proposition

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 on a vu 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.

Théorème.

Soit $(X_n)_{n\in\mathbb{N}}$ 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 note : $M_n^* \xrightarrow{\mathscr{L}} X$ avec $X \hookrightarrow \mathcal{N}(0,1)$ ou encore $M_n^* \xrightarrow{\mathscr{L}} \mathcal{N}(0,1)$

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

❸ Autrement dit :

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

Illustration avec les lois usuelles. (voir les annexes de la feuille Cours_14)

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

Interprétation en terme d'approximation. *Pour une loi quelconque.*

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

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, la loi de $\displaystyle\sum_{k=1}^{n} X_k$ peut être approchée par la loi $\mathcal{N}(n\mu, n\sigma^2)$.

Pour n assez grand ,

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

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

Soient $p \in]0,1[$ et $(Y_n)_{n\in\mathbb{N}}$ 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.

Démonstration. (*Feuille cours_14*)

Autrement dit :

Si $Y_n \hookrightarrow \mathscr{B}(n,p)$ alors : pour tout $x \in \mathbb{R}$, $\lim\limits_{n\to+\infty} \mathbb{P}\left(Y_n^* \leqslant x\right) = \Phi(x)$

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

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

Interprétation en terme d'approximation.

Soit $p \in]0,1[$,

on note pour $n \in \mathbb{N}$, Y_n une variable aléatoire suivant la loi binomiale de paramètres (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 $\mathcal{N}(np, np(1-p))$.

6.2 Seconde forme

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 *(inconnue)*

alors $\left(\dfrac{M_n - \mu}{\frac{S_n}{\sqrt{n}}}\right)_{n\geqslant 1}$ 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 $\forall x \in \mathbb{R}$, $\lim\limits_{n\to+\infty} \mathbb{P}\left(\dfrac{M_n - \mu}{\frac{S_n}{\sqrt{n}}} \leqslant x\right) = \Phi(x)$

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

  Lorsque les X_k sont des lois normales et que l'on prend $S'_n = \sqrt{\dfrac{1}{n-1}\sum\limits_{k=1}^{n}(X_k - M_n)^2}$ à la place de S_n

  alors Y_n suit une loi de Student à $n-1$ degrés de liberté.

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 $a \leqslant b$, $\lim\limits_{n\to+\infty} \mathbb{P}\left(a \leqslant \dfrac{M_n - \mu}{\frac{S_n}{\sqrt{n}}} \leqslant b\right) = \dfrac{1}{\sqrt{2\pi}}\displaystyle\int_a^b e^{-\frac{t^2}{2}}\,\mathrm{d}t$

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 $F_n = \dfrac{1}{n}\displaystyle\sum_{k=1}^{n} X_k$ *(la fréquence empirique)*

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

Démonstration. (*Feuille cours_14*)

Remarque : $\dfrac{\sqrt{n}(F_n - p)}{\sqrt{F_n(1-F_n)}} = \dfrac{F_n - p}{\dfrac{\sqrt{F_n(1-F_n)}}{\sqrt{n}}}$

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

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

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

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

– – – – – – – – – – – – – – – – – – –

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 $m_n = \dfrac{1}{n}\displaystyle\sum_{k=1}^{n} x_k$ la moyenne empirique, $s_n^2 = \dfrac{1}{n}\displaystyle\sum_{k=1}^{n} x_k^2 - m_n^2$ 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 $f_n = \dfrac{1}{n}\displaystyle\sum_{k=1}^{n} x_k$ la fréquence empirique, on sait alors que $s_n^2 = f_n(1 - f_n)$

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

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

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

Remarques :

- Ici on peut remplacer $\dfrac{f_n - p_0}{\frac{\sqrt{f_n(1-f_n)}}{\sqrt{n}}}$ par $\dfrac{f_n - p_0}{\frac{\sqrt{p_0(1-p_0)}}{\sqrt{n}}}$ *(Car on est sous l'hypothèse $p = p_0$)*
- Pour $\alpha = 5\%$ en utilisant la majoration $f_n(1 - f_n) \leqslant \dfrac{1}{4}$ cela devient :

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

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