Chapitre 15 Théorèmes limites

Commentaires :

Globalement, en statistiques, on peut chercher à faire plusieurs types de choses :

1. ***un résumé des phénomènes observés*** *(moyenne d'un échantillon, médiane, etc...) Ce sont les statistiques dites "descriptives" ;*
2. ***Utiliser la théorie pour prévoir les observations*** *(fréquence d'un événement, moyenne d'un échantillon, etc...). C'est ce qu'on a fait par exemple dans les chapitres précédents de probabilités ;*
3. ***Se servir des observations pour estimer les données théoriques*** *(probabilité, moyenne, etc...). Ce sont les statistiques dites "inférentielles" ;*
4. ***Utiliser les observations pour élaborer des modèles théoriques et "prédire l'avenir !"****. Ce sont les statistiques dites "prédictives".*

(1) : Le principe des statistiques descriptives est, comme son nom l'indique, de "décrire" les phénomènes constatés sur l'ensemble de la population en les résumant avec différents outils potentiels de description (mode, moyenne, écart-type, médiane, quantiles, etc...) On appelle ceci généralement des résultats "empiriques", c'est-à-dire ceux qui sont réellement obtenus en collectant les données sur la population.

(2) : Le principe des probabilités de manière générale est d'établir un comportement typique partant de données théoriques. Elles permettent entre autre de :
- prévoir un comportement général
- prévoir les résultats empiriques que l'on est censé obtenir
- d'évaluer dans une certaine mesure l'écart à ce comportement. (On étudiera dans ce chapitre divers résultats dans ce sens.)

(3) : L'étude des probabilités nous permet de plus d'étudier des comportements dits "limites", qui se produisent lors d'un grand nombre de répétitions d'une même expérience. Les outils et attentes sont multiples, mais dans le cadre des statistiques inférentielles, nous étudierons par exemple dans ce chapitre une approche des "tests de conformité", qui permettent dans certains cas, de prévoir si un échantillon a une chance de faire partie d'une population donnée.

(4) : Dès lors, en revanche, que l'on souhaite faire des prévisions sur les réalisations futures de la variable, il faut aller un peu plus loin dans l'analyse. En effet, on a généralement à faire à des quantités dont on ne connait pas les caractéristiques théoriques (espérance, variance, ou loi). Or, ce sont des informations capitales si on veut pouvoir établir des tendances et faire des prévisions avec plus ou moins de précision.
C'est là que les statistiques prédictives prennent le relais, mais mis à part des cas d'études très simples et relativement naïfs, ceci ne fait pas partie du champs d'étude de notre chapitre.

Notation :

Dans tout le chapitre, sauf précision, le terme *variable aléatoire* désignera une variable aléatoire réelle finie, ou discrète, ou à densité. De plus les suites de variables aléatoires (X_n) seront systématiquement construites sur un même espace probabilisé. La notation F_T désignera la fonction de répartition de la variable aléatoire T et de même, la notation f_T désignera la densité de T si celle-ci existe.

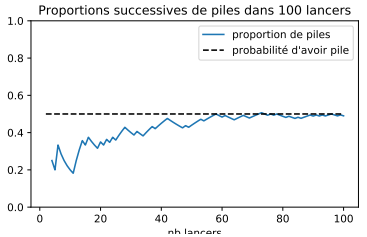
I Loi faible des grands nombres et approximation de la moyenne, d'une probabilité et de la variance

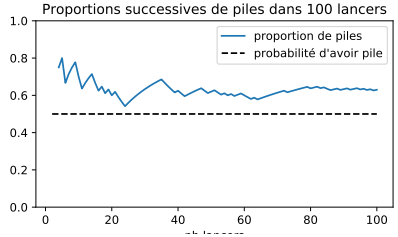
I-1 Problématique et vocabulaire sur un exemple :

Exemple 1 :

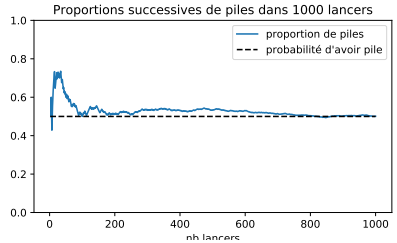
On s'intéresse à une pièce **déséquilibrée** et on se demande quelle est la probabilité d'obtenir Pile.

L'instinct nous dit que si on lance un grand nombre n de fois la pièce, la proportion observée de "Piles" sur une suite de lancers devrait être assez proche de la probabilité théorique qui serait ici $\frac{1}{2}$. Observons des résultats obtenus sur des simulations graphiquement grâce à Python :





La graphique de gauche nous confirme l'instinct. Néanmoins, le graphique de droite semble ne pas corroborer cette affirmation, étant donné que les proportions semblent encore relativement "loin" de la probabilité $\frac{1}{2}$!
Néanmoins, on peut observer qu'en augmentant le nombre de tentatives, le phénomène fini inexorablement par se produire :



Il nous reste donc à démontrer tout ceci. De plus, la suite du chapitre nous permettra de caractériser le caractère plus ou moins "proche", en fonction du nombre de tentatives !

Mettons en place un peu de vocabulaire afin de simplifier les énoncés qui vont suivre :

Définition

Si X est une variable aléatoire, on appelle *n-échantillon* de X une n-liste $(X_1, \dots, X_n)$ de variables aléatoires mutuellement indépendantes et de même loi que X.

■ Exemple 2 :

Soit X le résultat d'un lancer de dés. Si on fait n lancers et qu'on note $X_1, \dots, X_n$ les résultats des n lancers successifs, $(X_1, \dots, X_n)$ est un n-échantillon de X. Un résultat de ce n-échantillon avec $n = 5$ pourrait donc être $(1, 6, 4, 2, 4)$.
Ce résultat désigne le fait que le premier lancer donne 1, le deuxième lancer donne 6, etc...

Dans la suite, si X est une variable aléatoire, on posera donc $(X_1, \dots, X_n)$ un n-échantillon de X.

Commentaires :

Afin de prouver nos dires sur l'exemple 1, il nous faut tout d'abord transformer un calcul de proportion en un calcul de moyenne. L'exemple ci-dessous nous montre comment :

■ Exemple 3 :

Soit X la variable de Bernoulli donnant le nombre de succès "obtenir Pile" dans **un seul** lancer d'une pièce non nécessairement équilibrée :

$$X = \begin{cases} 1 & \text{si on obtient Pile} \\ 0 & \text{sinon} \end{cases}, \qquad \text{avec } X \hookrightarrow \mathcal{B}\left(\frac{1}{2}\right)$$

Alors, dans une série de n lancers, si on note X_i la valeur de X obtenue pour le $i^{\text{ème}}$ lancer, le nombre total de Piles correspond à $X_1 + \dots + X_n$, et la proportion de Piles dans l'échantillon est

$$\frac{X_1 + \dots + X_n}{n}$$

Par exemple, si $n = 5$, voici un exemple de n-échantillon possible :

$$(\underbrace{0}_{X_1}, \underbrace{1}_{X_2}, \underbrace{0}_{X_3}, \underbrace{0}_{X_4}, \underbrace{1}_{X_5}) \text{ désigne } FPFFFP$$

Ici, le nombre de Piles obtenus est $X_1 + \dots + X_n = 2$ et $\frac{2}{5} = \frac{X_1 + \dots + X_n}{n}$ est la proportion de Piles obtenus.
De plus, $\mathbb{E}[X] = P(\text{"obtenir un Pile"})$. On la note
Ainsi, justifier la véracité de l'intuition de l'exemple 1 reviendrait donc à démontrer l'approximation suivante :

$$\underbrace{\mathbb{E}[X]}_{\substack{\text{i.e. } P(\text{"obtenir un Pile"}) \\ \textit{(valeur théorique)}}} \simeq \underbrace{\frac{X_1 + \dots + X_n}{n}}_{\substack{\text{proportion de Piles} \\ \textit{(valeur effective)}}}$$

Commentaires :

On différencie ainsi les paramètres théoriques (que l'on cherche souvent à déterminer) et les résultats empiriques observés. On distinguera donc en particulier :

- *la moyenne théorique $\mathbb{E}[X]$ et la moyenne empirique notée $\overline{X_n} = \frac{X_1 + \dots + X_n}{n}$.*
- *la variance théorique $V(X)$ et la variance empirique $S_n^2 = \frac{1}{n} \sum_{i=1}^{n} \left((X_i - \overline{X_n})^2\right)$.*
- *la loi et l'histogramme d'un n-échantillon (point et explications sur lequels nous reviendrons plus tard)*

En Python :

À partir d'une liste de données du nom de `Donnees`, pour obtenir les caractéristiques de base d'un échantillon avec Python, les commandes issues de la bilbiothèque `numpy` sont les suivantes :

```
  # moyenne empirique X_n :
np.mean(Donnees)
  # variance empirique S_n^2 :
np.var(Donnees)
```

On peut également obtenir les médianes, quartiles, etc...avec d'autres commandes.

I-2 Inégalités

Cette partie va permettre de mettre en place les résultats préliminaires permettant la conjecture avancée dans la partie précédente. (Re)-voyons un premier résultat général permettant de décrire le comportement théorique d'une variable :

Lemme 1 *(Inégalité de Markov)*

Si X est une variable aléatoire réelle **positive** *admettant une espérance, alors elle vérifie l'inégalité*

$$P(X \geqslant a) \leqslant \frac{E(X)}{a} \qquad \forall a > 0$$

Remarque :

Ce résultat confirme en particulier que la probabilité que X prenne des valeurs très grandes est forcément petite. (Ceci est d'ailleurs trivial sur les variables finies...)

Démonstration :

• Cas d'une variable discrète : On note $Supp(X) = \{x_i \mid i \in \mathbb{N}\}$ les valeurs de X. (Toutes positives par hypothèse.) Alors,

$$\begin{aligned} E(X) &= \sum_{i=0}^{+\infty} x_i P(X = x_i) = \sum_{\substack{i=0 \\ x_i < a}}^{+\infty} \underbrace{x_i P(X = x_i)}_{\geqslant 0} + \sum_{\substack{i=0 \\ x_i \geqslant a}}^{+\infty} x_i P(X = x_i) \\ &\geqslant \sum_{\substack{i=0 \\ x_i \geqslant a}}^{+\infty} \underbrace{x_i}_{\geqslant a} P(X = x_i) \geqslant a \underbrace{\sum_{\substack{i=0 \\ x_i \geqslant a}}^{+\infty} P(X = x_i)}_{P(X \geqslant a)} = a\, P(X \geqslant a) \end{aligned}$$

• Cas d'une variable à densité : On note f une densité de X. Notons que, comme X est positive, on peut supposer que f est nulle sur $]-\infty; 0[$.

$$\begin{aligned} E(X) &= \int_{-\infty}^{+\infty} x f(x)\, dx = \int_{0}^{+\infty} x f(x)\, dx \qquad (X \geqslant 0) \\ &= \int_0^a \underbrace{x f(x)}_{\geqslant 0}\, dx + \int_a^{+\infty} \underbrace{x}_{\geqslant a} f(x)\, dx \\ &\geqslant \int_0^{+\infty} a f(x)\, dx = a \underbrace{\int_a^{+\infty} f(x)\, dx}_{P(X \geqslant a)} \geqslant a\, P(X \geqslant a) \end{aligned}$$

□

Passons maintenant à l'écart à la moyenne. On peut la majorer grâce à une formule simple :

Théorème 2 *(Inégalité de Bienaymé-Tchebychev)*

Si X est une variable aléatoire admettant un moment d'ordre 2, alors elle vérifie l'inégalité

$$P(|X - E(X)| \geqslant \epsilon) \leqslant \frac{V(X)}{\epsilon^2} \qquad \forall \epsilon > 0$$

Démonstration :

On pose $Y = |X - E(X)|^2$. Y est une variable aléatoire positive.
Comme X admet un moment d'ordre 2, alors $E(|X - E(X)|^2)$ existe (c'est la variance de X) et donc, d'après l'inégalité de la proposition précédente, on a

$$P(|X - E(X)|^2 \geqslant \epsilon^2) \leqslant \frac{E(|X - E(X)|^2}{\epsilon^2} \qquad \forall \epsilon > 0$$

i.e.

$$P(|X - E(X)| \geqslant \epsilon) \leqslant \frac{V(X)}{\epsilon^2} \qquad \forall \epsilon > 0$$

□

Commentaires :

On constate que plus ϵ est grand ou plus la variance est petite, plus la probabilité devient faible. Autrement dit, ceci confirme notre éventuelle intuition que X ne peut s'éloigner de manière trop importante de son espérance qu'avec une probabilité faible. Cette inégalité servira par exemple à démontrer le théorème de la loi faible des grands nombres qui suit dans le paragraphe suivant.

I-3 Loi faible : approximation de la moyenne

Commentaires :

Le théorème qui suit est fondamental en statistiques (même s'il n'est pas le seul !) Il permet de commencer à prouver que l'intuition est bien suivie d'un résultat théorique avéré : le fait qu'une moyenne empirique se rapproche de la moyenne théorique et en conséquence, que la proportion se rapproche en effet inexorablement du paramètre théorique (exemple 1)

Théorème 3 *(Loi faible des grands nombres)*

Soit $(X_1, \ldots, X_n)$ un n-échantillon d'une variable aléatoire X d'espérance μ admettant une variance.

Alors la moyenne empirique $\overline{X_n} = \dfrac{X_1 + \ldots + X_n}{n}$ vérifie

$$\lim_{n \to +\infty} P\Big(|\overline{X_n} - \underbrace{\mu}_{\substack{\uparrow \\ =\mathbb{E}[X]=\mathbb{E}(\overline{X_n})}}| \geqslant \epsilon\Big) = 0 \qquad \forall \epsilon > 0$$

Démonstration :

C'est une application du théorème de Bienaymé-Tchebychev. On note

$$S_n = X_1 + \ldots + X_n.$$

Par linéarité de E, on a $\mathbb{E}(S_n) = n\mu$, d'où

$$\mathbb{E}(\overline{X_n}) = \mu.$$

Comme les variables sont non corrélées, on a également

$$V(S_n) = V(X_1) + \ldots + V(X_n) = n\sigma^2,$$

d'où $V(\overline{X_n}) = V\left(\dfrac{S_n}{n}\right) = \dfrac{1}{n}\sigma^2$.
D'après l'inégalité de Bienaymé-Tchebytchev, on a

$$P(|\overline{X_n} - \mu| \geqslant \epsilon) \leqslant \frac{\sigma^2}{n\epsilon^2}$$

Le passage à la limite achève la démonstration. □

Commentaires :

La loi faible des grands nombres signifie qu'en quelque sorte, $\overline{X_n}$ converge vers son espérance, d'où,

$$\boxed{\text{pour } n \text{ grand, } \overline{X_n} \simeq \mathbb{E}[X]}$$

On note de plus dans la démonstration que

$$\boxed{\mathbb{E}[\overline{X_n}] = \mu = \mathbb{E}[X]}$$

et on peut montrer que (exercice)

$$V(\overline{X_n} - \mu) \xrightarrow[n\to+\infty]{} 0$$

ce qui confirme que pour n grand, $\overline{X_n}$ s'éloigne très peu de μ.

I-4 Application : approximation des probabilités

La loi faible vaut en particulier pour des variables X_k qui suivent une loi de Bernoulli. Dans ce cas, on rappelle que

$$\boxed{\text{Si } X \hookrightarrow \mathcal{B}(p), \text{ alors } \mu = \mathbb{E}[X] = p}$$

Ainsi, le théorème de la loi faible se traduit comme ci-dessous :

Théorème 4 *(de Bernoulli)*

Soit $(X_n)_{n\in\mathbb{N}^*}$ une suite de variables aléatoires **indépendantes** *suivant une même* **loi de Bernoulli** *$\mathcal{B}(p)$. Alors, si on pose $\overline{X_n} = \dfrac{X_1 + \ldots + X_n}{n}$, on a*

$$\lim_{n\to+\infty} P\Big(|\overline{X_n} - p| \geqslant \epsilon\Big) = 0 \qquad \forall \epsilon > 0$$

`Démonstration :`

C'est exactement la loi faible appliquée aux variables de Bernoulli.
En effet, avec les notations de la loi faible, on a $\mu = \mathbb{E}[X] = p$. □

Remarque :

Traduit en "variable binomiale", cela donne :

Si $(S_n)_{n\in\mathbb{N}^*}$ est une suite de variable suivant respectivement une **loi binomiale** $\mathcal{B}(n,p)$, alors

$$\lim_{n\to+\infty} P\Big(\Big|\frac{S_n}{n} - p\Big| \geqslant \epsilon\Big) = 0 \qquad \forall \epsilon > 0$$

Commentaires :

Cet exemple justifie en particulier le fait suivant :

$$\boxed{\text{Si } X \hookrightarrow \mathcal{B}(p), \text{ alors } \overline{X_n} \simeq p \text{ pour } n \text{ grand.}}$$

i.e. : si on répète un grand nombre de fois une épreuve de Bernoulli $\mathcal{B}(p)$,

la fréquence du nombre de succès se rapproche fatalement de la probabilité de succès p.

(cf exemple sur les n lancers de pièce déséquilibrée du début du chapitre.) La théorie confirme donc encore une fois l'intuition.

I-5 Application : approximation de la variance

On s'intéresse à la variance empirique $S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X_n})^2$ d'un échantillon dont on ne connait pas la variance théorique. On aimerait justifier que

$$\boxed{S_n^2 \text{ est une approximation de } \sigma^2.}$$

Nous allons justifier ce résultat en 2 étapes.

La loi faible donne directement le résultat suivant :

Propriété 5

Si X est une variable admettant un moment d'ordre 4, alors, pour $Y = (X-\mu)^2$, on a

$$\lim_{n\to+\infty} P\Big(\big|\overline{Y_n} - \sigma^2\big| \geqslant \epsilon\Big) = 0 \qquad \forall \epsilon > 0$$

d'où, pour n grand :

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 \simeq V(X).$$

Remarque :

On observe d'ailleurs la relation (vraie dès que X admet un moment d'ordre 2) :

$$\boxed{\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2\right] = \sigma^2 = V(X)}$$

il suffit, pour le prouver, d'appliquer la linéarité de l'espérance à la somme pour l'obtenir, en sachant que $V(X) = \mathbb{E}[(X - \mathbb{E}[X])^2]$.

Commentaires :

Sachant que d'après la loi faible, on a pour n grand : $\mu \simeq \overline{X_n}$, si on admet que l'on peut faire les deux approximations successives (attention, la somme a un grand nombre de termes, on pourrait donc peut être avoir beaucoup de perte de précision avec cette approximation...), on pourrait dire que

$$\sigma^2 \simeq \overline{Y_n} \simeq S_n^2$$

ainsi

$$\boxed{S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X_n})^2 \simeq \sigma^2.}$$

Néanmoins, contrairement à ce qui se passe pour $\mu = \mathbb{E}[\overline{X_n}]$, on a $\mathbb{E}[S_n^2] \neq \sigma^2$. En effet :

Propriété 6

Sous réserve d'existence (X admettant au moins un moment d'ordre 4), on a

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

`Démonstration :`

$$\begin{aligned}
\mathbb{E}[S_n^2] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X_n})^2\right] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu + \mu - \overline{X_n})^2\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\underbrace{\mathbb{E}\left[(X_i-\mu)^2\right]}_{=\sigma^2} + \frac{1}{n}\sum_{i=1}^{n}\underbrace{\mathbb{E}\left[(\mu - \overline{X_n})^2\right]}_{V(\overline{X_n})=\frac{\sigma^2}{n}} - 2\underbrace{\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)(\overline{X_n}-\mu)\right]}_{\alpha} \\
&= \frac{1}{n}\cdot n\cdot\sigma^2 + \frac{1}{n}\cdot n\cdot\frac{\sigma^2}{n} - 2\alpha
\end{aligned}$$

Or,

$$\begin{aligned}
\alpha &= \mathbb{E}\left[(\overline{X_n}-\mu)\left(\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)\right)\right] = \mathbb{E}\left[(\overline{X_n}-\mu)\left(\frac{1}{n}\sum_{i=1}^{n}X_i - \frac{1}{n}\cdot n\mu\right)\right] \\
&= -\mathbb{E}\left[(\overline{X_n}-\mu)^2\right] = -V(\overline{X_n}) = \frac{\sigma^2}{n}
\end{aligned}$$

Ainsi, en réinjectant dans la somme plus haut, on obtient :

$$\mathbb{E}[S_n^2] = \sigma^2 - \tfrac{\sigma^2}{n} = \tfrac{n-1}{n}\sigma^2$$

□

Commentaires :

De la propriété précédente, on tire également

$$\mathbb{E}[S_n^2] \xrightarrow[n\to+\infty]{} \sigma^2$$

De plus, on peut montrer (exercice), que $V(S_n^2 - \sigma^2) \xrightarrow[n\to+\infty]{} 0$
Ainsi, si n est grand, la variable S_n^2 ne varie que très peu autour de σ^2, i.e.

$$\boxed{\text{si } n \text{ est grand,} \quad S_n^2 \simeq \sigma^2}$$

Remarque :

Du résultat précédent, on tire également une autre valeur approchée de σ^2 :

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

C'est ce qu'on appelle *la "variance corrigée"*.
Pour information, votre calculatrice vous propose généralement de calculer soit l'écart-type, soit l'écart-type corrigé (souvent notés resp. σ et $\widehat{\sigma}$). Remarquez que la variance corrigée est plus grande que la variance car $\frac{n}{n-1} > 1$, alors $\widehat{S}_n^2 = \frac{n}{n-1}S_n^2 > S_n^2$.

Nous avons vu dans cette partie comment approcher quelques paramètres élémentaires (probabilité, espérance, variance) si on est assuré d'avoir un échantillon de grande taille de notre variable. Nous allons maintenant aborder des résultats plus précis concernant non seulement certains paramètres, mais plus généralement la loi des "variables limites".

II Convergence en loi et variables à supports entiers

Dans cette partie, on va mettre en place des stratégies pour savoir dans quelle mesure deux variables X, Y "ont des lois approximativement identiques."

Dans le cadre des variables discrètes, on pourrait estimer que les lois sont approximativement identiques (on note $\mathcal{L}(X) \simeq \mathcal{L}(Y)$) si

$$Supp(X) = Supp(Y) \quad et \quad P(X=k) \simeq P(Y=k) \quad \forall k \in Supp(X) = Supp(Y)$$

Mais qu'en est-il par exemple, si X suit une loi uniforme sur $[\![1,6]\!]$ et Y suit elle aussi une loi uniforme, mais sur $\{1-10^{-4}, \dots, 6-10^{-4}\}$ (valeurs extrêmement proches). On a plutôt

$$P(X=k) = \frac{1}{6} = P(Y=k') \quad \forall k \in [\![1,6]\!] \text{ et } k' \simeq k$$

Là aussi on aurait envie de dire que $\mathcal{L}(X) \simeq \mathcal{L}(Y)$ non ?

Qu'en est-il de plus dans le cadre des variables à densité ? On peut se placer du point de vu de l'histogramme de la loi et penser que deux lois de variables sont proches ssi leurs histogrammes le sont, ce qui, en terme de densité, signifierait

$$f_X(x) \simeq f_Y(x) \quad \forall x \in \mathbb{R}$$

Pour conclure, le dénominateur commun à ces toutes ces approches consiste surtout à penser que les lois deux variables sont proches si

$$P(a < X \leq b) \simeq P(a < Y \leq b) \quad \forall a, b \in \mathbb{R}\ (a < b)$$

i.e. pour synthétiser :

$$\boxed{F_X(x) \simeq F_Y(x) \quad \forall x \in \mathbb{R}}$$

II-1 Convergence en loi : cas général

Généralement, on a besoin d'approcher une loi par une autre si les données de l'une sont plus compliquées ou longues à calculer que l'autre. Par exemple, si une variable est le résultat d'un grand nombre n de répétitions d'une même expérience. On aura envie de dire que "plus n est grand, plus la loi se rapproche d'une autre". Ceci se traduit par exemple par la notion de "convergence en loi" ci-dessous :

Définition

Soit $(Y_i)_{i\in\mathbb{N}}$ une suite de variables aléatoires de fonctions de répartition respectives F_{Y_i}, $\forall i \in \mathbb{N}$. On dit que (Y_n) *converge en loi vers* Y et on note $Y_n \xrightarrow{\mathcal{L}} Y$ si

$$\lim_{n\to+\infty} F_{Y_n}(u) = F_Y(u) \quad \forall u \in \mathbb{R} \text{ où } F_Y \text{ est continue en } u.$$

Ceci traduit en particulier que pour n grand, on a pour tout u :

$$P(Y_n \leq u) \simeq P(Y \leq u)$$

et donc que, pour tout $a, b \in \mathbb{R}$ $(a < b)$:

$$P(a < Y_n \leq b) \simeq P(a < Y \leq b)$$

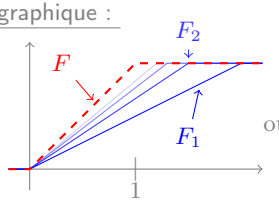
i.e. on peut approcher les probabilités concernant Y_n par celles concernant Y avec un n suffisamment grand.

■ Exemple 4 :

On considère (X_n) la suite de v.a. telle que $X_n \hookrightarrow \mathcal{U}_{[0;1+\frac{1}{n}]}$.

Alors $(X_n) \xrightarrow{\mathcal{L}} X \hookrightarrow \mathcal{U}_{[0;1]}$:

Version graphique :



où F_i est la fonction de répartition de X_i.

Version calculatoire : On a $F_n(x) = \begin{cases} 0 & \text{si } x < 0 \\ \frac{x}{1+\frac{1}{n}} & \text{si } 0 \leq x \leq 1+\frac{1}{n} \\ 1 & \text{si } x \geq 1+\frac{1}{n} \end{cases}$

Ainsi, à x fixé, on a :

⋆ Si $x < 0$,

$$F_n(x) = 0 \quad \text{et donc} \quad \lim_{n\to+\infty} F_n(x) = 0$$

⋆ Si $0 \leq x \leq 1$, alors $0 \leq x \leq 1+\frac{1}{n}$ $\forall n \in \mathbb{N}$ donc $F_n(x) = \frac{x}{1+\frac{1}{n}}$ et donc

$$\lim_{n\to+\infty} F_n(x) = x$$

⋆ Si $x > 1$, alors, pour n assez grand, on a $1+\frac{1}{n} < x$ et donc

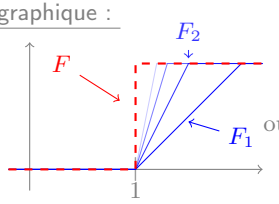
$$F_n(x) = 1 \quad \text{puis} \quad \lim_{n\to+\infty} F_n(x) = 1$$

`Problème des points de discontinuité de` F_Y `sur un exemple :`

■ Exemple 5 :

On considère (X_n) la suite de v.a. telle que X_n suit une loi uniforme sur $\left[1; 1+\frac{1}{n}\right]$. Alors (X_n) converge en loi vers la variable X constante égale à 1 p.s.

Version graphique :



où F_i est la fonction de répartition de X_i.

Version calculatoire : On a $F_n(x) = \begin{cases} 0 & \text{si } x < 1 \\ \frac{x-1}{\frac{1}{n}} & \text{si } 1 \leq x \leq 1+\frac{1}{n} \\ 1 & \text{si } x \geq 1+\frac{1}{n} \end{cases}$

Ainsi, à x fixé, on a :

⋆ Si $x < 1$,

$$F_n(x) = 0 \quad \text{et donc} \quad \lim_{n\to+\infty} F_n(x) = 0$$

⋆ Si $x > 1$, alors, pour n assez grand, on a $1 + \frac{1}{n} < x$ et donc

$$F_n(x) = 1 \quad \text{puis} \quad \lim_{n\to+\infty} F_n(x) = 1$$

⋆ Si $x = 1$, on remarque que

$$F_n(x) = 0 \quad \forall n \in \mathbb{N}^*$$

alors

$$\lim_{n\to+\infty} F_n(x) = \begin{cases} 0 & \text{si } x \leq 1 \\ 1 & \text{si } x > 1 \end{cases}$$

mais la fonction ainsi définie n'est pas une fonction de répartition. (Non continue à droite.)
Fort heureusement, ce n'est pas un problème, puisqu'on a a quand même

$$\lim_{n\to+\infty} F_n(x) = F(x) \quad \forall x \neq 1$$

Notation :
On notera également que pour n grand, si Y_n converge en loi vers Y, alors Y_n suit "à peu près" la loi de Y. On notera ce dernier point dans le cours $Y_n \overset{\sim}{\hookrightarrow} \mathcal{L}(Y)$.

II-2 Convergence en loi : cas des supports dans $\mathbb{N}$

Dans le cas des variables discrètes, plus précisément à valeurs entières, la problématique de la fonction de répartition peut être simplifiée :

II.2-a) En général

Proposition 7

Avec les notations de la définition, si $Supp(Y_i) \subset \mathbb{N}\ \forall i \in \mathbb{N}$ et $Supp(Y) \subset \mathbb{N}$, alors $Y_n \xrightarrow{\mathcal{L}} Y$ ssi

$$P(Y_n = k) \xrightarrow[n\to+\infty]{} P(Y = k) \quad \forall k \in \mathbb{N}$$

`Démonstration :`

- Supposons qu'il y ait convergence en loi.

Comme les supports de Y_n et Y sont entiers, on a :

$$P(Y_n = k) = F_{Y_n}\left(k + \frac{1}{2}\right) - F_{Y_n}\left(k - \frac{1}{2}\right)$$

Or, comme Y est à support entier, F_Y est continue en $\left(k + \frac{1}{2}\right)$ et $\left(k - \frac{1}{2}\right)$. Ainsi

$$P(Y_n = k) \xrightarrow[n\to+\infty]{} F_Y\left(k + \frac{1}{2}\right) - F_Y\left(k - \frac{1}{2}\right) = P(Y = k)$$

- Supposons que les différentes probabilités convergent.

Soit $x \in \mathbb{R} - \mathbb{N}$. Alors

$$\begin{aligned} F_{Y_n}(x) &= P(Y_n \leq x) = P\left(Y_n \leq \lfloor x \rfloor\right) \quad \text{car } Supp(Y_n) \subset \mathbb{N} \\ &= \sum_{k=0}^{\lfloor x \rfloor} P(Y_n = k) \xrightarrow[n\to+\infty]{} \sum_{k=0}^{\lfloor x \rfloor} P(Y = k) \\ &= F_Y(\lfloor x \rfloor) = F_Y(x) \quad \text{car } Supp(Y) \subset \mathbb{N} \end{aligned}$$

Ainsi, (Y_n) converge en loi vers Y. □

II.2-b) Approximation d'une loi binomiale par une loi de Poisson

Théorème 8

Si $(X_n)_{n\in\mathbb{N}^*}$ est une suite de variables aléatoires suivant respectivement une loi $\mathcal{B}\left(n, \frac{\lambda}{n}\right)$, avec $\lambda > 0$, alors elle converge en loi vers une variable suivant la loi de Poisson $\mathcal{P}(\lambda)$, i.e.

$$\lim_{n\to+\infty} P(X_n = k) = e^{-\lambda}\frac{\lambda^k}{k!} \qquad \forall k \in \mathbb{N}$$

`Démonstration :`
Soit $k \in \mathbb{N}$ fixé, $n \geqslant k$ et $\lambda \in]0; +\infty[$. On a

$$\begin{aligned} P(X_n = k) &= \binom{n}{k}\left(\frac{\lambda}{n}\right)^k\left(1 - \frac{\lambda}{n}\right)^{n-k} = \frac{n!}{k!(n-k)!}\left(\frac{\lambda}{n}\right)^k\left(1 - \frac{\lambda}{n}\right)^{n-k} \\ &= \frac{\lambda^k}{k!}\underbrace{\frac{n!}{(n-k)!}\frac{1}{n^k}}_{A_n}\underbrace{\left(1 - \frac{\lambda}{n}\right)^{n-k}}_{B_n} \end{aligned}$$

Or, $$A_n = \frac{n(n-1)\dots(n-k+1)}{n^k} = \underbrace{\frac{n}{n}\frac{n-1}{n}\dots\frac{n-k+1}{n}}_{k \text{ fois}} \xrightarrow[n\to+\infty]{} 1$$

et $$B_n = \left(1 - \frac{\lambda}{n}\right)^{n-k} = e^{(n-k)\ln\left(1-\frac{\lambda}{n}\right)}$$

Or, $$(n-k)\ln\left(1 - \frac{\lambda}{n}\right) \underset{n\to+\infty}{\sim} -(n-k)\frac{\lambda}{n} \underset{n\to+\infty}{\sim} -\lambda$$

D'où $B_n \xrightarrow[n\to+\infty]{} e^{-\lambda}$ puis la limite annoncée : $\lim_{n\to+\infty} P(X_n = k) = \frac{\lambda^k}{k!}e^{-\lambda}$ □

Commentaires :

Généralement les lois binomiales se présentent sous la forme $\mathcal{B}(n,p)$ et non $\mathcal{B}\left(n,\frac{\lambda}{n}\right)$. Dans les notations du théorème, si n est grand, on aurait donc $\frac{\lambda}{n}=p$ très petit. Le théorème dit donc que si n est grand, $\mathcal{B}(n,p)$ se rapproche de $\mathcal{P}(\lambda)$ avec $\lambda=np$. (D'où : On appelle quelquefois la loi de Poisson la "loi des événements rares".)

En pratique, on estime que, dès que $\boldsymbol{n\geqslant 30}$ et $\boldsymbol{p\leqslant 0,1}$, on peut approcher la loi binomiale $\mathcal{B}(n,p)$ par la loi de Poisson $\mathcal{P}(np)$.

■ Exemple 6 :

On considère une variable aléatoire X suivant une loi $\mathcal{B}(50;0,05)$. On souhaite calculer $P(X=3)$.

• Calcul exact :

$$P(X=3)=\binom{50}{3}(0,05)^3(0,95)^{47}=\frac{50\times 49\times 48}{3\times 2}(0,05)^3(0,95)^{47}\simeq 0,2199$$

• Calcul approché : On a

$$n=50\geqslant 30 \text{ et } p\leqslant 0,1.$$

On peut donc approcher $\mathcal{B}(50;p)$ par la loi $\mathcal{P}(np)=\mathcal{P}(2,5)$
D'où

$$P(X=3)\simeq e^{-2,5}\frac{2,5^3}{3\times 2}\simeq 0,214$$

Intérêt 1 : On élimine des calculs de factoriels et de puissances $(0,05)^3(0,95)^{47}$ pouvant donner lieu à des erreurs d'approximation : *(Par exemple, en disant "$(0,05)^3\simeq 0$, donc $P(X=3)\simeq 0$"...ce qui est donc faux ici.)*

Intérêt 2 : Par ordinateur, on gagne de la vitesse en éliminant un grand nombre de multiplications, (néanmoins au profit d'un calcul d'exponentiel)

? Exercice 1

Supposons donnée une urne contenant deux variétés de boules (rouges et vertes) en quantité totale N. On note V le nombre de boules vertes. On y fait un tirage d'un nombre $n\leqslant N$ de boules et on veut compter Y_N le nombre de boules vertes.

1. Montrer que

$$P(Y_N=k)=\frac{\binom{V}{k}\binom{N-V}{n-k}}{\binom{N}{n}}\quad\forall k\leqslant n$$

2. Montrer que si $N\to+\infty$, à proportion $p=\frac{V}{N}$ de boules vertes constante,

$$P(Y_N=k)\underset{N\to+\infty}{\sim}\binom{n}{k}p^k(1-p)^{n-k}\quad\forall k\leqslant n$$

3. En déduire que Y_N converge en loi vers une loi binomiale $\mathcal{B}(n,p)$.
Interprétation : un tirage sans remise se rapproche d'un tirage avec remise dans une très grande urne.

Solution

1. On se donne l'univers des possibilités équiprobables de tirages constitué des poignées de n éléments parmi les N de l'urne. Ainsi, il y a $\binom{N}{n}$ tirages possibles au total. L'événement $(Y_N=k)$ signifie qu'on a tiré une poignée avec k boules rouges : $\binom{k}{N}$ possibilités et en parallèle $N-k$ boules vertes : $\binom{N-V}{n-k}$, d'où la probabilité annoncée.

2. Si $N\to+\infty$, avec une propotion p de boules vertes constantes, on a nécessairement $V\to+\infty$ et dépasse nécessairement n au bout d'un certain temps. On peut donc supposer être dans ce cas de figure. Dans ce cas, d'après 1) :

$$\begin{aligned}P(Y_N=k) &= \frac{V!}{k!(V-k)!}\frac{(N-V)!}{(n-k)!(N-V-(n-k))!}\frac{n!(N-n)!}{N!}\\ &= \frac{n!}{k!(n-k)!}\frac{V!}{(V-k)!}\frac{(N-V)!}{(N-V-(n-k))!}\frac{(N-n)!}{N!}\text{(en réorganisant les termes)}\\ &= \binom{n}{k}\frac{\overbrace{V(V-1)\dots(V-k+1)}^{k\text{ termes}}\ \overbrace{(N-V)(N-V-1)\dots(N-V-(n-k)+1)}^{n-k\text{ termes}}}{\underbrace{N(N-1)\dots(N-n+1)}_{n\text{ termes}}}\end{aligned}$$

Le nombre de termes k, k et n étant constants, on obtient :

$$P(Y_N=k)\sim\binom{n}{k}\frac{V^k(N-V)^{n-k}}{N^n}=\binom{n}{k}\frac{V^k(N-V)^{n-k}}{N^kN^{n-k}}=\binom{n}{k}\left(\underbrace{\frac{V}{N}}_{p}\right)^k\left(\underbrace{\frac{N-V}{N}}_{1-p}\right)^{n-k}$$

3. Comme le terme $\binom{n}{k}p^k(1-p)^{n-k}$ est constant par rapport à N, on a bien

$$\lim_{\substack{N\to+\infty\\ Np\in\mathbb{N}}}P(Y_N=k)=\binom{n}{k}p^k(1-p)^{n-k}\quad\forall k\leqslant n$$

ce qui est la convergence en loi annoncée.

Commentaires :

Que se passe-t-il dans les autres cas ? Et bien il existe des résultats sur les variables de type $\overline{X_n}$ extrêment utiles. Ils s'intègrent dans un théorème s'intitulant "théorème de la limite centrée" ou encore "th. central limite" que l'on retrouve dans notre programme sous deux formes que l'on découvrira ci-dessous.

III Théorème central limite : deux formes

III-1 Si on connait σ^2

III.1-a) Le théorème ; première forme

Rappelons la définition suivante :

Définition

Soit Y une variable aléatoire réelle admettant une variance non nulle. Alors, lorsque l'on note

$$\mu = E(Y), \quad \sigma^2 = \mathbb{V}(Y) \quad \text{et} \quad Y^* = \frac{Y-\mu}{\sigma}$$

on appelle Y^* la *variable centrée réduite* associée à Y. (car $E(Y)=0$ et $\mathbb{V}(Y)=1$).

Commentaires :

Le théorème ci-dessous va établir de quelle manière on pourra supposer que la loi de $\overline{X_n}^*$ peut être remplacée par la loi $\mathcal{N}(0;1)$, quelquesoit la variable X de départ !! Ce résultat est donc extrêment puissant...

Théorème 9 *central limite (ou de la limite centrée) ; première forme*

Soit $(X_1,\dots,X_n)$ un n-échantillon d'une variable aléatoire X de variance non nulle. On note

$$\mu = E(X), \quad \sigma^2 = V(X), \quad \overline{X_n} = \frac{X_1+\ldots+X_n}{n} \quad et \quad \overline{X_n}^* = \frac{\overline{X_n}-\mu}{\sigma/\sqrt{n}}$$

on obtient que

$$\overline{X_n}^* \text{ converge en loi vers } Y \hookrightarrow \mathcal{N}(0,1)$$

i.e. pour tout $a,b\in\bar{\mathbb{R}}$ avec $a<b$,

$$\lim_{n\to+\infty} P(a<\overline{X_n}^*\leqslant b) = \frac{1}{\sqrt{2\pi}}\int_a^b e^{-t^2/2}\,dt = \phi(b)-\phi(a)$$

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

`Démonstration :`

admise □

Remarque :

On rappelle que $\overline{X_n}^* = \dfrac{\overline{X_n}-\mu}{\sigma/\sqrt{n}}$ est la loi centrée réduite de $\overline{X_n}$ car

$$\mathbb{E}[\overline{X_n}] = \frac{1}{n}\sum_{k=1}^n \mathbb{E}[X_i] = \frac{1}{n}\sum_{k=1}^n \mu = \frac{n}{n}\mu = \mu$$

$$\mathbb{V}[\overline{X_n}] \underbrace{=}_{\text{indépendance des } X_i} \frac{1}{n^2}\sum_{k=0}^n \mathbb{V}[X_i] = \frac{1}{n^2}\sum_{k=1}^n \sigma^2 = \frac{n}{n^2}\sigma^2 = \frac{1}{n}\sigma^2$$

i.e. :

$$\forall n\in\mathbb{N}, \quad \boxed{\mathbb{E}[\overline{X_n}^*]=0, \quad et \quad V(\overline{X_n}^*)=1}$$

Il est donc cohérent que la loi de la variable $\overline{X_n}^*$ s'approche d'une loi d'espérance 0 et de variance 1

Remarque :

Le TCL ayant comme conséquence le fait que pour n grand,

$$\mathcal{L}(\overline{X_n}^*) \simeq \mathcal{N}(0,1),$$

par les propriétés des lois normales, on pourrait dire également que

$$\mathcal{L}(\overline{X_n}) \simeq \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

ou encore $\mathcal{L}\left(\sum_{k=1}^n X_k\right) \simeq \mathcal{N}\left(n\mu, n\sigma^2\right)$

mais **ATTENTION**, on ne dit certainement pas que chacun des X_n suit approximativement une loi normale. ça peut en effet être complètement faux...

■ Exemple 7 :

On choisit 500 fois au hasard un nombre compris entre 0 et 1. Quelle est la probabilité approximative que la somme de ces nombres soit comprise entre 240 (strictement) et 260 ?

On introduit les variables aléatoires X_i correspondant au nombre obtenu au $i^{\text{ème}}$ choix. Alors, $(X_1,\dots,X_{500})$ est un 500-échantillon de $X\hookrightarrow\mathcal{U}([0;1])$
En posant

$$T_n = X_1+\ldots+X_{500},$$

la question revient à chercher $P(240<T_n\leqslant 260)$.

Résolution exacte ?
Ne connaissant pas la loi de T_n, on pourrait peut être la calculer, mais ceci signifie qu'il faudrait appliquer la formule de convolution 499 fois. Qui se lance ... ?

Résolution approchée : (à l'aide du TCL) Posons

$$\mu = E(X_1) = 1/2 \quad \text{et} \quad \sigma^2 = \frac{1}{12}$$

on a alors, $\dfrac{T_n/n - \mu}{\sigma/\sqrt{n}} \rightsquigarrow \mathcal{N}(0,1)$ d'où

$$\begin{aligned} P(240 < T_n \leqslant 260) &= P\left(\frac{240/n-\mu}{\sigma/\sqrt{n}} < \frac{T_n/n-\mu}{\sigma/\sqrt{n}} \leqslant \frac{260/n-\mu}{\sigma/\sqrt{n}}\right) \\ &= P\Big(\underbrace{\frac{240/500-0,5}{1/\sqrt{500\times 12}}}_{\alpha\simeq -1,55} < \frac{T_n/n-\mu}{\sigma/\sqrt{n}} \leqslant \underbrace{\frac{260/500-0,5}{1/\sqrt{500\times 12}}}_{\beta\simeq 1,55}\Big) \end{aligned}$$

D'après le TCL, on peut donc estimer que

$$P(240 < T_n \leqslant 260) \simeq \phi(1,55) - \phi(-1,55) = 2\phi(1,55) - 1 \simeq 0.879$$

III.1-b) Appl.1 : approximation d'une loi binomiale par une loi normale

Commentaires :

Voyons le cas particulier de X qui suit une loi de Bernoulli de paramètre p, où on sait alors que $T_n = X_1 + \ldots + X_n$ suit une loi binomiale $\mathcal{B}(n,p)$

Théorème 10 *de Moivre-Laplace*

Soit $(T_n)_{n\in\mathbb{N}^*}$ une suite de variables aléatoires qui suivent respectivement une loi $\mathcal{B}(n,p)$, où $p \in]0;1[$. Alors,

$$T_n^* \xrightarrow{\mathcal{L}} Y \hookrightarrow \mathcal{N}(0,1)$$

i.e., pour tous les $a, b \in \bar{\mathbb{R}}$ où $a < b$, on a

$$P\Big(a < \underbrace{\frac{T_n - np}{\sqrt{np(1-p)}}}_{T_n^*} \leqslant b\Big) \xrightarrow[n\to+\infty]{} \phi(b) - \phi(a) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{t^2}{2}}\, dt$$

Démonstration :

C'est tout simplement la traduction du théorème central limite à une suite de variables de Bernoulli. En effet, on pose $(X_1, \ldots, X_n)$ un n-échantiloon d'une variable $X \hookrightarrow B(p)$, de manière à ce que $T_n = X_1 + \ldots + X_n$. Alors, d'après le TCL, on

$$P(a < X_n^* \leqslant b) \xrightarrow[n\to+\infty]{} \Phi(b) - \Phi(a)$$

Or $X_n^* = \dfrac{T_n/n - p}{\sigma/\sqrt{n}} = \dfrac{T_n - np}{\sqrt{n}\,\sigma} = \dfrac{T_n - np}{\sqrt{np(1-p)}}$ □

Commentaires :

En traduisant ce théorème, ceci signifie que la loi de T_n^* peut être approchée par une loi $\mathcal{N}(0;1)$ si n est "grand". Or,

$$T_n = \sigma\sqrt{n}\cdot T_n^* + n\mu, \quad \text{avec } \sigma^2 = p(1-p)$$

Donc la loi de T_n (i.e. $\mathcal{B}(n,p)$ peut être approchée par la loi $\mathcal{N}(\underbrace{np}_{\mathbb{E}[T_n]}\ ;\ \underbrace{n\sigma^2}_{\mathbb{V}[T_n]})$

Dans la pratique, on estime que l'on peut approcher la loi de T_n^* par la loi $\mathcal{N}(0,1)$ ou $\mathcal{B}(n,p)$ par la loi $\mathcal{N}\big(np, np(1-p)\big)$ dès que $\boldsymbol{n \geqslant 30}$, $\boldsymbol{np \geqslant 5}$ et $\boldsymbol{n(1-p) \geqslant 5}$.

■ Exemple 8 :

On lance un dé équilibré 100 fois. On souhaite approximer la probabilité pour que le nombre de 3 soit compris entre 20 et 30 (au sens large).

On note T_n la variable aléatoire donnant le nombre de 3 dans la série de lancés. Alors

$$T_n \rightsquigarrow \mathcal{B}(n,p), \text{ où } p = 1/6.$$

On a $n \geqslant 30$, $np = 100/6$ $(\simeq 17) \geqslant 5$ et $n(1-p) = 100\times 5/6$ $(\simeq 83) \geqslant 5$.

D'après le théorème de Moivre-Laplace (ou d'après le TCL), la loi de T_n peut être approchée par $\mathcal{N}(np, np(1-p))$. Or :

$$P(20 \leqslant T_n \leqslant 30) = P(19 < T_n \leqslant 30)$$

Si votre calculatrice le permet, vous pouvez donc dès à présent faire ce calcul. Sinon, on passe à la loi normake centrée réduite en constatant que $T_n^* = \dfrac{T_n - np}{\sqrt{np(1-p)}}$ peut être considérée comme suivant par la loi normale centrée réduite $\mathcal{N}(0;1)$. D'où

$$\begin{aligned} P(20 \leqslant T_n \leqslant 30) &= P\left(\frac{19-np}{\sqrt{np(1-p)}} < T_n^* \leqslant \frac{30-np}{\sqrt{np(1-p)}}\right) \\ &= P\left(\underbrace{\frac{19-100/6}{\sqrt{100\times 5/36}}}_{\alpha\simeq 0,626} < T_n^* \leqslant \underbrace{\frac{30-100/6}{\sqrt{100\times 5/36}}}_{\beta\simeq 3,578}\right) \\ &\left(\simeq P(\alpha < Z \leqslant \beta) \quad \text{où } Z \rightsquigarrow \mathcal{N}(0;1)\right) \end{aligned}$$

Ainsi, d'après les tables de la loi normale (ou avec votre calculatrice)

$$P(\alpha < T_n^* \leqslant \beta) \simeq \phi(\beta) - \phi(\alpha) \simeq \phi(3,578) - \phi(0,626) \simeq 0,9998 - 0,7357 \simeq 0,2651$$

On peut également formuler ceci légèrement différemment à l'aide cette fois ci des proportions :

Corollaire *(thm de Moivre-Laplace, 2ème version)*

Soit $(F_n)_{n\in\mathbb{N}^*}$ La fréquence d'obtention d'un événement A de probabilité p, dans une suite de n répétitions identiques et indépendantes d'une même expérience. alors, quand

$$n \geqslant 30,\ np \geqslant 5 \text{ et } n(1-p) \geqslant 5,$$

on estime que

$$\mathcal{L}\left(\frac{F_n - p}{\sqrt{\frac{p(1-p)}{n}}}\right) \simeq \mathcal{N}(0,1), \quad i.e. \quad \mathcal{L}(F_n) \simeq \mathcal{N}\left(p, \frac{p(1-p)}{n}\right)$$

■ Exemple 9 :

On lance une pièce équilibrée $n = 100$ fois. On cherche à savoir quelle sera la probabilité d'avoir au moins $1/4$ des tirages qui soient des 1.

On note F_n la proportion de 1 dans la série de lancés. Alors, d'après le théorème de Moivre Laplace, comme on lance n fois de manière identique et indépendante, que la probabilité d'obtenir 1 à chaque lancer est $p = \frac{1}{6}$, que pour finir

$$n \geqslant 30, \quad np \simeq 17 \geq 5, \quad n(1-p) \simeq 84 \geq 5$$

On en déduit qu'on peut approcher la loi de F_n par $\mathcal{N}\left(p, \frac{p(1-p)}{n}\right) = \mathcal{N}\left(\frac{1}{6}, \frac{5/36}{100}\right)$. D'où, directement à la calculatrice (ou en passant à la loi normale centrée réduite)

$$P(0.25 \leqslant F_n) = 1 - P(F_n < 0.25) \simeq 0.0127$$

Très faible ! Mieux vaut ne pas compter dessus...

Correction de continuité

Lorsque l'on approche une loi discrète par une loi continue, on a un problème d'approximation du type suivant : Si X suit une loi discrète (par exemple entière), on a par exemple

$$P(10 \leqslant X \leqslant 15) = P(9,9 < X \leqslant 15,3) = \dots$$

Évidemment, l'approximation s'en trouve donc légèrement modifiée. On souhaite alors équilibrer l'erreur obtenue "de chaque coté de X". La solution la moins douloureuse en général consiste donc à introduire une *correction de continuité*, c'est-à-dire, si n et m sont deux entiers, on écrira

$$\boxed{P(n \leqslant X \leqslant m) = P(n - 0,5 < X \leqslant m + 0,5)}$$

Cependant, cette manipulation n'est pas exigible au concours. Dans le cas ci-dessus, toute valeur issue de $P(10 \leqslant X \leqslant 15) = P(9,9 < X \leqslant 15,3) = P(9 < X < 16) = \dots$ sera acceptée.

■ Exemple 10 :

On reprend l'exemple des 100 lancers de dés en utilisant une correction de continuité.

$$\begin{aligned} P(20 \leqslant T_n \leqslant 30) &= P(19,5 < T_n \leqslant 30,5) = P\left(\frac{19,5 - np}{\sqrt{np(1-p)}} < T_n^* \leqslant \frac{30,5 - np}{\sqrt{np(1-p)}}\right) \\ &= P\left(\underbrace{\frac{19,5 - 100/6}{\sqrt{100 \times 5/36}}}_{\alpha \simeq 0,760} < T_n^* \leqslant \underbrace{\frac{30,5 - 100/6}{\sqrt{100 \times 5/36}}}_{\beta \simeq 3,712}\right) \\ \left(&\simeq P(\alpha < Z \leqslant \beta) \quad \text{où } Z \rightsquigarrow \mathcal{N}(0;1) \right) \end{aligned}$$

Ainsi, d'après les tables de la loi normale,

$$P(\alpha < Y_n^* \leqslant \beta) \simeq \phi(\beta) - \phi(\alpha) \simeq \phi(3,712) - \phi(0,760) \simeq 0,9999 - 0,7764 \simeq 0,2235$$

À titre d'information :
- le calcul exact sur la loi binomiale effectué par ordinateur donne environ 0,2195.
- Sans correction de continuité :

$$P(20 \leqslant T_n \leqslant 30) \simeq 0.185 \quad ; \quad P(19 < T_n \leqslant 30) = 0.265 \quad ; \quad P(20 \leqslant T_n < 31) = 0.185$$

III.1-c) Appl 2 : approximation d'une loi de Poisson par une loi normale

Théorème 11

Soit $(T_n)_{n\in\mathbb{N}^*}$ une suite de variables aléatoires qui suivent respectivement une loi $\mathcal{P}(np)$, où $p > 0$. Alors,

$$T_n^* \xrightarrow{\mathcal{L}} Y \hookrightarrow \mathcal{N}(0,1)$$

i.e., pour tous les $a, b \in \bar{\mathbb{R}}$ où $a < b$, on a

$$P(a < T_n^* \leqslant b) \xrightarrow[n\to+\infty]{} \phi(b) - \phi(a).$$

Démonstration :

Encore une fois, ce n'est que l'application du TCL à la suite $(T_n)_{n\in\mathbb{N}^*}$.
En effet, on peut écrire que pour tout $n \in \mathbb{N}$ $T_n = X_1 + \dots + X_n$, où $(X_n)_{n\in\mathbb{N}^*}$ est une suite de variables aléatoires mutuellement indépendantes, de même loi de Poisson $\mathcal{P}(\lambda)$ (et donc de variance $\sigma^2 = \lambda$ non nulle.) □

- En pratique, on estime que l'on peut approcher la loi de T_n^* par $\mathcal{N}(0;1)$ si $\boldsymbol{np \geqslant 18}$.
- Pour $\lambda \geqslant 18$, ceci signifie également que la loi $\mathcal{P}(\lambda)$ peut être approchée par $\mathcal{N}(\lambda, \lambda)$.

■ Exemple 11 :

On pose $n = 40$ et $\lambda = 20$. On suppose que X suit une loi

$$\mathcal{P}(20).$$

On cherche $P(X \leqslant 17)$.

Dans notre calculatrice ou Python, on peut obtenir que

$$P(X \leqslant 17) \simeq 0.2970$$

L'approximation par loi normale donne :

$$\begin{aligned} P(X \leqslant 17) &= P(X \leqslant 17,5) \qquad \text{(correction de continuité)} \\ &= P\left(\frac{X-np}{\sqrt{np}} \leqslant \frac{17,5-np}{\sqrt{np}}\right) = P\left(\frac{X-np}{\sqrt{np}} \leqslant \underbrace{\frac{17,5-20}{\sqrt{20}}}_{\alpha \simeq -0.5590}\right) \\ &\simeq 0.2881 \end{aligned}$$

III.1-d) Les approximations en bref

Condition	On peut approcher	par
$n \geqslant 30$ et $p \leqslant 0,1$	$\mathcal{B}(n,p)$	$\mathcal{P}(np)$
$n \geqslant 30$, $np \geqslant 5$ et $n(1-p) \geqslant 5$	$\mathcal{B}(n,p)$	$\mathcal{N}\big(np, np(1-p)\big)$
$\lambda \geqslant 18$	$\mathcal{P}(\lambda)$	$\mathcal{N}(\lambda,\lambda)$.

Commentaires :

Astuce pour se souvenir des paramètres à appliquer dans les approximations :

Les lois ci-dessus par lesquelles on approche admettent toutes des paramètres que l'on peut retrouver grâce à leur espérance et variance. (Par exemple, pour $\mathcal{P}(np)$, son espérance est np.)
Pour retrouver ces paramètres, il suffit d'aller les chercher dans la variable de départ. Dans l'ordre ci-dessus :

$$\underbrace{\mathcal{B}(n,p)}_{\text{espérance } np} \longrightarrow \underbrace{\mathcal{P}(np)}_{\text{paramètre = espérance } np}$$

$$\underbrace{\mathcal{B}(n,p)}_{\text{espérance } np,\ \text{variance } np(1-p)} \longrightarrow \underbrace{\mathcal{N}\big(np, np(1-p)\big)}_{\text{espérance } np,\ \text{variance } np(1-p)}$$

$$\underbrace{\mathcal{P}(\lambda)}_{\text{espérance = variance } =\lambda} \longrightarrow \underbrace{N(\lambda,\lambda)}_{\text{espérance = variance}=\lambda}$$

Remarque :

Si la situation le permet, il est également possible d'enchaîner les approximations. Par exemple, *(cas "extrême")* : On pioche n boules dans une très grande urne avec remise et on compte X le nombre de "bonnes" boules.
si $n \geqslant 30$ et $p \leqslant 0,1$:

$$\mathcal{B}(n,p) \text{ peut être approchée par } \mathcal{P}(np).$$

Rien ne s'oppose à ce qu'on ait également $np \geqslant 18$. Ainsi,

$$\mathcal{P}(np) \text{ peut être approchée par } \mathcal{N}(np, np).$$

Au final, on peut donc supposer que

$$X \text{ suit une loi } \mathcal{N}(np, np)$$

On ne s'attendra néanmoins pas à ce que les résultats soient très précis. Par exemple, si on a également $n(1-p) \geq 5$, alors il vaut mieux passer directement de la loi binomiale à la loi normale, car dans ce cas, on tombe sur $\mathcal{N}(np, np(1-p))$, dont la variance est plus petite que $\mathcal{N}(np, np)$.

III-2 Si on ne connait pas σ^2 : Deuxième forme du TCL

Commentaires :

La première version du TCL utilise l'espérance μ et la variance σ de la variable X. Or, dans de multiples cas, contrairement aux exemples de la partie précédente où on connait σ, on ne dispose pas nécessairement de toutes les données théoriques lorsqu'on étudie un caractère sur un échantillon de population. D'après la première partie, on sait néanmoins que l'on peut approcher l'espérance et la variance à l'aide (respectivement) de $\overline{X_n}$ et S_n^2 (notation de la partie I.)

On va alors remarquer que le TCL est encore valable en remplaçant simplement σ par S_n :

Théorème 12 *TCL (deuxième forme)*

Soit $(X_1, \dots, X_n)$ un n-échantillon d'une variable aléatoire X. On note

$$\mu = E(X), \quad \overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \text{et} \quad S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X_n})^2 = \overline{X_n^2} - \overline{X_n}^2$$

on obtient que

$$\frac{\overline{X_n} - \mu}{S_n/\sqrt{n}} \text{ converge en loi vers } Y \hookrightarrow \mathcal{N}(0,1)$$

i.e., pour tous $a, b \in \bar{\mathbb{R}}$ avec $a < b$,

$$\lim_{n\to+\infty} P\left(a < \frac{\overline{X_n} - \mu}{S_n/\sqrt{n}} \leqslant b\right) = \frac{1}{\sqrt{2\pi}}\int_a^b e^{-t^2/2}\,dt = \phi(b) - \phi(a)$$

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

Remarque :

Par rapport à la première version, on a remplacé σ par S_n. Ceci revient à remplacer $\overline{X_n}^*$ par $\dfrac{\overline{X_n}-\mu}{S_n/\sqrt{n}}$.

Remarquez que $\overline{X_n}^*$ était la "centrée réduite" de $\overline{X_n}$, ce qui n'est plus le cas de $\dfrac{\overline{X_n}-\mu}{S_n/\sqrt{n}}$. Le TCL est toutefois encore valable mais néanmoins, il y a fort à parier que les approximations soient moins bonnes !

Exercice 2

On a planté deux types de graines A et B pour lesquels on note respectivement $\mu_A, \sigma_A, \mu_B, \sigma_B$ leur moyennes et écart-type théoriques (inconnus) au jour 100 de leur croissance. Après mesure sur un échantillon de 50 graines chacunes, notant m_A, s_A, m_B, s_B leur moyennes et écart-type empirique au jour 100 de leur croissance, on trouve

$$m_A = 15,8, \quad s_A = 2, \quad m_B = 17, \quad s_B = 3$$

1. Déterminer, pour la variété A, une valeur a telle que

$$P\left(m_A - a\frac{s_A}{\sqrt{n}} < \mu_A < m_A + a\frac{s_A}{\sqrt{n}}\right) \simeq 0.95$$

2. En déduire un intervalle I_A tel que

$$P(\mu_A \in I_A) = 0.95$$

3. Trouver de même un intervalle I_B correspondant à

$$P(\mu_B \in I_B) = 0.95$$

4. Peut-on répondre à la question "est-ce que $\mu_A = \mu_B$?" avec une probabilté raisonnable 0.95 d'avoir raison ?

Solution

On observe que

$$P\left(m_A - a\frac{s_A}{\sqrt{n}} < \mu_A < m_A + a\frac{s_A}{\sqrt{n}}\right) = P\left(-a \leq \frac{m_A - \mu_A}{s_A/\sqrt{n}} \leq a\right)$$

D'après la deuxième forme du TCL, on sait que

$$P\left(-a \leq \frac{m_A - \mu_A}{s_A/\sqrt{n}} \leq a\right) \simeq \Phi(a) - \Phi(-a) = 2\Phi(a) - 1$$

Ainsi,

$$P\left(m_A - a\frac{s_A}{\sqrt{n}} < \mu_A < m_A + a\frac{s_A}{\sqrt{n}}\right) = 0,95 \quad \Leftrightarrow \quad 2\Phi(a) - 1 \simeq 0.95 \quad \Leftrightarrow \quad a \simeq 1,96$$

On trouve alors l'intervalle

$$I_A \simeq [15,25; 16,35]$$

Les mêmes calculs avec les valeurs de B donnent

$$I_B \simeq [16.17; 17.83]$$

Ce qui signifie que

$$\mu_A \in I_A, \quad \mu_B \in I_B, \quad I_A \cap I_B \neq \emptyset$$

On pourrait très bien avoir $\mu_A = \mu_B$ mais on n'est est pas certain ! **on ne peut donc pas conclure...**

Corollaire *Application à la loi binomiale*

Si $T_n \hookrightarrow \mathcal{B}(n,p)$. Alors, pour n grand, si on note r la proportion empirique de succès, on peut dire estimer que

$$\frac{T_n - np}{\sqrt{nr(1-r)}} \hookrightarrow \mathcal{N}(0,1)$$

Démonstration :

Comme déjà utilisé précédemment, T_n peut s'écrire $T_n = X_1 + \ldots + X_n$ où $X_i \hookrightarrow \mathcal{B}(p)$ qui rend 1 si un succès arrive au rang i.
Ainsi, on sait d'après le TCL (deuxième forme), comme $\mathbb{E}[X_i] = p$, on peut estimer que

$$\frac{\overline{X_n} - p}{S_n/\sqrt{n}} \hookrightarrow \mathcal{N}(0,1)$$

En multipliant par n le numétateur et dénominateur :

$$\frac{T_n - np}{S_n\sqrt{n}} \hookrightarrow \mathcal{N}(0,1)$$

Par définition dans le TCL de la variance empirique :

$$S_n^2 = \overline{X_n^2} - \overline{X_n}^2$$

Or

$$\overline{X_n}^2 = \left(\frac{T_n}{n}\right)^2 = p'^2 \quad \text{et} \quad X_i^2 = X_i$$

d'où

$$\overline{X_n^2} = \overline{X_n} = \frac{T_n}{n} = p'.$$

Ainsi

$$S_n^2 = p' - p'^2 = p'(1-p')$$

ce qui donne la conclusion annoncée. □

Commentaires :

Afin de retenir cette formule, on peut remarquer que la variance correspond exactement à la variance empirique.

Dans la pratique, on estime qu'on peut approcher la loi $\mathcal{B}(n,p)$ par la loi $\mathcal{N}\big(np, nr(1-r)\big)$ dès que $\boldsymbol{n \geqslant 30}$, $\boldsymbol{nr \geqslant 10}$, $\boldsymbol{n(1-r) \geqslant 10}$,
(Conditions plus contraignantes que pour une approximation grâce au TCL1)

Remarque :

Comme dans la première version, (thm de Moivre Laplace) on peut traduire ceci en version "calcul de fréquence" en posant $F_n = \frac{T_n}{n}$. qui représente la fréquence d'obtention d'un événement A dans une série de n répétitions indépendantes d'une même expérience. Ainsi, on pourrait estimer que dans les conditions pré-citées, on aurait

$$\mathcal{L}\left(\frac{F_n - p}{\sqrt{\frac{r(1-r)}{n}}}\right) \simeq \mathcal{N}(0,1), \quad i.e. \quad \mathcal{L}(F_n) \simeq \mathcal{N}\left(p, \frac{r(1-r)}{n}\right)$$

ce qui est particulièrement utile quand on veut estimer des probabilités !

Exercice 3

Un institut de sondage a été contacté afin de tenter de prévoir le gagnant avant la fin de la totalité du dépouillement, le soir du second tour des élections présidentielles. Les bureaux de vote ferment normalement à 19h, mais certains ont une dérogation préfectorale afin d'ouvrir jusqu'à 20h. À cette heure là, l'institut dispose donc déjà d'une grande partie des données déjà dépouillées. Les journaux ont donc pris l'habitude d'annoncer un potentiel gagnant dès l'ouverture du journal de 20h.
Pour ce faire, on note p la proportion finale de français ayant voté pour le candidat A, n le nombre de votes déjà connus et T_n le nombre de français ayant voté pour A sur les n connus.

1. Soit r la proportion empirique de votes pour A qu'on estime $> 10\%$. Déterminer x tel que

$$P\left(\left|\frac{T_n - np}{\sqrt{nr(1-r)}}\right| \leq x\right) \simeq 0,99$$

2. À 19h10, on effectue d'abord une première estimation après $n = 2\,000$ bulletins déjà dépouillés. On obtient 1 019 bulletins en faveur de A. Déterminer la proportion empirique de succès. A-t-on envie de déclarer la candidat A gagnant ?

3. Déterminer maintenant un intervalle dans lequel se trouve p avec une probabilité de 99%. Peut-on donc véritablement affirmer que A sera gagnant ?

4. Vers 19h30, on dispose déjà de $n = 200\,000$ résultats. On obtient maintenant 102 234 bulletins en faveur de A. Quelle est la margeur d'erreur à 99% en annonçant que le candidat A a une probabilité r de l'emporter ? Peut-on enfin raisonnablement déclarer A gagnant ?

5. Est-ce encore le cas avec une probabilité de $0,999$?

Solution

1. D'après le TCL, comme n est très grand ($n \geq 30$, $nr \geq 10$, $n(1-r) \geqslant 10$), on peut estimer que la variable dans la probabilité suit une loi normale $\mathcal{N}(0,1)$. Ainsi, on obtiendra

$$x \simeq 2.5758$$

2. On a ici $r = \frac{T_n}{n} = \frac{1\,019}{2\,000} = 0.5095 > 0.5$. On pourrait penser que A sera gagnant.

3. D'après le calcul de la question 1, après en isolant p au centre de l'inégalité, on obtient que

$$P\left(r - x\frac{\sqrt{r(1-r)}}{\sqrt{n}} \leq p \leq r + x\frac{\sqrt{r(1-r)}}{\sqrt{n}}\right) \simeq 0,99$$

Ce qui, avec les chiffres annoncés ici donne l'intervalle approximatif demandé

$$[0.4983; 0.5207]$$

Ainsi, on peut très bien trouver en réalité $p < 0,5$. On ne peut pas affirmer que p sera gagnant.

4. La proportion empirique de succès est maintenant

$$r = 0.51117$$

avec une marge d'erreur de

$$x\frac{\sqrt{r(1-r)}}{\sqrt{n}} \simeq 0,00112$$

Ainsi, on est certain à 99% que le candidat A a une probabilité de remporter les élections supérieure à $r - 0,00112 \simeq 0,51 > 0.5$. On peut maintenant envisager de le déclarer gagnant.

5. On obtient cette fois-ci $x \simeq 3.29053$. L'intervalle devient donc $[0.5100, 0.5123]$. C'est donc encore le cas !

IV Introduction aux tests de conformité

Commentaires :

Le TCL a de multiples autres utilités issues de l'approximation par une loi normale. Il permet par exemple

- *de prévoir dans quels intervalles on va trouver les données empiriques ;*
- *ou au contraire, d'extrapoler des intervalles dans lesquelles se situent les données théoriques à partir d'échantillons (ex. exercice de la partie précédente).*

Il existe d'énormes variétés de procédures que l'on peut mettre en place pour tester un grand nombre de choses, mais à votre programme se trouve explicitement les "tests de conformité", qui sont liés au premier point.

IV-1 Principe d'un test de conformité sur un exemple

Globalement, un test de conformité consiste à comparer un échantillon avec la population théorique de référence et à tenter d'émettre un avis sur le fait que l'échantillon corresponde ou non à ladite population théorique.

La procédure est la suivante :

- On pose deux hypothèses opposées (classiquement nommées H_0 et H_1) afin de tester laquelle est la bonne (nous préciserons ceci plus tard) ;
- On effectue certains calculs avec les données dont on dispose ;
- On tente d'apporter une conclusion à la problématique.

Exemple 12 :

On considère deux variétés de fraises différentes, conditionnées sur deux palettes distinctes pour leur vente. Le responsable de l'étiquetage a peur d'avoir confondu les deux palettes. Il sait néanmoins que les deux variétés ont des poids moyens significativement différents et que sur une des deux palettes, il y a des gariguettes dont il connait le poids moyen par fruit : 16g.
Sa seule manière d'indentifier les palettes est de faire un prélèvement d'échantillon (suffisamment conséquent).

On pose alors les hypothèses :

H_0 : "La palette n° 1 est celle des gariguettes" *(hypothèse précise)*

H_1 : "La palette n° 1 n'est pas celle des gariguettes" *(hypothèse contraire)*

Sur la palette testée, il trouve un poids moyen empirique de 16,5g.
Est-ce ou non la palette de gariguette ? La suite de cette partie consiste à trouver des outils donnant des éléments de réponse à cette question.

Commentaires :

Nous verrons dans cette partie que le travail de test de conformité se rapproche d'un raisonnement par l'absurde. Les calculs nous permettant (avec nos connaissances actuelles en Bcpst) de donner un élément de réponse sont issus du TCL (1ère ou 2ème forme selon besoin).
On rappelle ainsi que, si n est grand

$$\overline{X_n}^* \xrightarrow{\mathcal{L}} Z \hookrightarrow \mathcal{N}(0,1)$$

et de manière plus large

$$\frac{\overline{X_n}-\mu}{S_n/\sqrt{n}} \xrightarrow{\mathcal{L}} Z \hookrightarrow \mathcal{N}(0,1)$$

Notre but est donc de ramener chaque problématique à un problème de moyenne afin d'utilise le TCL pour établir des résultats.

? Exercice 4

Revenons à notre exemple des 2 palettes de fraises. Supposons que les poids des gariguettes ait un écart type de $\sigma = 2g$ et que le responsable de l'étiquetage ait prélevé aléatoirement $n = 100$ fraises sur la palette n° 1. On note X_i le poids de la fraise i et on pose l'hypothèse

H_0 : "La palette n° 1 est celle des gariguettes"

1. Sous l'hypothèse H_0, quelle est la valeur de a qui correspond à la probabilité $P\left(\left|\frac{\overline{X_n}-\mu}{\frac{\sigma}{\sqrt{n}}}\right| \leq a\right) \simeq 0.95$
En déduire un intervalle I pour lequel $P\left(\overline{X_n} \in I\right) \simeq 0.95$.

2. La valeur moyenne empirique effectivement observée sur l'échantillon prélevée est $m_A = 16,5$. Permet-elle de savoir quelle est la palette de gariguette ?

Solution

1. Pour la probabilité 0.95, D'après d'après le TCL, avec $2\Phi(a) - 1 \simeq 0.95$, la calculatrice nous donne

$$\Phi(a) = \frac{1+0,95}{2}, \quad \text{d'où } a \simeq 1,96.$$

Ainsi, en développant : l'intervalle I est le suivant :

$$I \simeq \left[\mu - a\frac{\sigma}{\sqrt{n}}; \mu + a\frac{\sigma}{\sqrt{n}}\right] \simeq [15.61; 16.39]$$

2. Comme $16,5 \notin I$, on peut en déduire avec une probabilité d'au moins 95% d'avoir raison, que l'on peut rejeter l'hypothèse H_0. Ce n'est pas la palette de gariguette. Par élimination, la palette de gariguettes est donc la deuxième.

Explications sur l'exercice et l'exemple :
Posons $\mu = 16$ le poids théorique des gariguettes et μ_2 le poids théorique correspondant à la palette étudiée. Nous souhaitons en fait conclure ici sur les hypothèses

H_0 : "c'est la palette de gariguette", H_1 : "ce n'est pas la palette de gariguette"

que l'on remplace ici par

$$H_0 : "\mu = \mu_2",\ H_1 : "\mu \neq \mu_2"$$

Dans le cadre de l'exercice précédent, on a fait un **un raisonnement par l'absurde**, en supposant $\mathcal{H}_0$ vraie. Sous cette hypothèse, on a trouvé :

"C'est la palette de gariguette" $\Rightarrow$ "95% des valeurs possibles de $\overline{X_n}$ sont dans I"

Or, m_A est une valeur (celle effectivement obtenue) de la variable $\overline{X_n}$ et on a trouvé

$$m_A = 16,5 \notin I.$$

Comme on trouve finalement une valeur m_A de $\overline{X_n}$ qui est très peu probable (moins de 5% de chance de tomber en dehors de I), on estime avoir une **contradiction**. Ainsi, l'hypothèse initiale $\mathcal{H}_0$ est fausse. Ainsi, on en déduit que **ce n'est pas la palette escomptée**, H_1 est vraie.

Commentaires :

Qu'en est-il si en revanche la valeur empirique de $\overline{X_n}$ est dans l'intervalle I ? Que peut-on en déduire ?

■ Exemple 13 :

Reprenons l'exercice précédent sur les fraises avec $m_A = 16,3$. Peut-on conclure sur le fait que ce soit ou non la palette de gariguettes ?

On avait $I = [15.61; 16.39]$ Cette fois-ci, on a $16,3 \in I$. **On ne peut donc pas exclure que ce soit la palette de gariguette** mais peut-on affirmer en conséquence que c'est bien la palette de gariguettes ?
En fait, non. Pour comprendre ceci, relisez les explications faites un peu plus haut et souvenez-vous que dans un raisonnement par l'absurde, on ne peut conclure si on n'aboutit à une contradiction !

On ne peut donc en réalité rien décider.

PEUT-ON TROUVER UNE SOLUTION POUR DÉCIDER SI H_0 EST VRAIE ?

Avec les techniques précédentes, afin d'accepter l'hypothèse H_0 : "C'est la palette de gariguette", il nous faudrait donc **rejeter** H_1 : "Ce n'est pas la palette de gariguette". Or, le problème est que nous n'avons aucun cadre précis de calcul en partant de H_1 ! Impossible donc de fournir une formule de comparaison du type précédent...

En revanche, si nous avions les informations théoriques de l'autre variété de fraise, ce serait éventuellement possible (à méditer !)

Remarque :

Retour sur la signification de la probabilité $\alpha = 0.95$ du $''P(\overline{X_n} \in I) = 0.95\ ''$:

On comprend qu'on a

- 95% de chance que cet événement se produise
- 95% des valeurs de $\overline{X_n}$ sont dans I
- 5% des valeurs de $\overline{X_n}$ ne sont pas dans I.

Autrement dit, on a (au maximum) 5% de chance de se tromper quand on rejette H_0. Si on veut réduire les chances de se tromper, il faut donc augmenter le $''\alpha''$. On pourra d'ailleurs se demander pourquoi on ne pourra jamais prendre $\alpha = 1$ en obtenant un intervalle I raisonnable !...

? Exercice 5

Reprendre l'exercice précédent (p. 15) avec $P\left(\overline{X_n} \in I\right) \simeq 0.99$ *(On veut conclure avec moins de risque de se tromper que pour l'exercice précédent où c'était* 0.95*)*. Peut-on conclure sur le fait que ce soit ou non la palette de gariguettes si $m_A = 16,5$?

Solution

Les calculs nous amènent à

$$a \simeq 2.576, \quad I = [15.48; 16.51]$$

Cette fois-ci, on a $16,5 \in I$. **On ne peut donc pas exclure que ce soit la palette de gariguette** mais on ne peut pas non plus affirmer que c'est bien la palette de gariguettes.

IV-2 Résumé de vocabulaire et principe général

`On peut donc maintenant résumer le principe général d'un test de conformité en détail :`

Définition

On appelle *population de référence* la population générale pour laquelle on dispose des données théoriques.

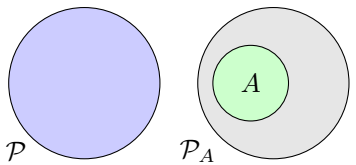
Notation :

On note ici $\mathcal{P}$ une population de référence et A un échantillon d'une population $\mathcal{P}_A$ dont on ne sait pas si elle correspond ou non à la population de référence. *(ex : deux variétés de fraises peut être différentes)*
On souhaite savoir si $\mathcal{P} = \mathcal{P}_A$ en effectuant un test sur un même caractère (*ex : le poids des fraises)* pour lequel on note X et X_A les variables aléatoires mesurant ledit caractère sur un individu respectivement de $\mathcal{P}$ et $\mathcal{P}_A$.
On estime évidemment que chaque individu est indépendant concernant le caractère étudié. On note

$$\mu = \mathbb{E}[X],\ \mu_A = \mathbb{E}[X_A] \text{ et } \overline{X_n} \text{ la moyenne empirique qui sera observée sur } A.$$

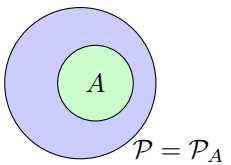
Résumé des 2 situations possibles :

A n'est pas dans $\mathcal{P}$ ($\mathcal{P} \neq \mathcal{P}_A$) :



$\mu \neq \mu_A$

A est dans $\mathcal{P}$ ($\mathcal{P} = \mathcal{P}_A$)



$\mu = \mu_A$

Définition

On pose

$$H_0 \ : \ "\mu = \mu_A".$$

Cette hypothèse est appelée *l'hypothèse nulle*. On pose H_1 l'hypothèse contraire. Elle est appelée *hypothèse alternative*.

Remarque :

H_0 traduit notre hypothèse de travail $\mathcal{P} = \mathcal{P}_A$, et c'est la seule hypothèse sous laquelle on peut effectuer nos calculs *(ex : l'échantillon est dans la palette de garigguettes)*

Il existe d'autres types d'hypothèses nulles (par exemple sur des variances $\sigma = \sigma_A$ ou autre) mais ce sera toujours une hypothèse précise avec laquelle on peut procéder à des calculs.

`Raisonnement :`

Pour une probabilité donnée α (grande), on cherche un intervalle I adapté à nos hypothèses tel que

$$P(\overline{X_n} \in I) = \alpha.$$

Si finalement la moyenne effective vaut $\overline{X_n} = m_A \notin I$, on rejette H_0, d'où la définition suivante :

Définition

On appelle $\mathbb{R} - I$ la *zone de rejet*.

Commentaires :

Pour les tests de conformité, il faut donc globalement comprendre le fait suivant : quand la probabilité $P\left(\overline{X_n} \in I\right)$ est grande,

- *une situation d'exclusion ($m_A \notin I$) permet de rejeter l'hypothèse.*

  En effet, si l'échantillon était conforme, on devrait être dans I. On ne l'est pas, donc l'échantillon est non conforme.

- *une situation d'inclusion ($m_A \in I$) ne permet pas de conclure.*

- *Plus on veut être certain d'avoir raison, (α augmente) moins on prend de décisions.*

On va maintenant voir qu'il existe classiquement deux types d'hypothèses alternatives :

IV-3 Tests bilatéraux

Ce sont les tests les plus courants. Pour un échantillon A et une population de référence $\mathcal{P}$ de moyenne théorique resp. μ_A et μ, ils consistent à opposer deux situations de type

$$H_0 \ : \ \mu_A = \mu \qquad ; \qquad H_1 \ : \ \mu_A \neq \mu$$

Pour faire les calculs, on va donc tenter de trouver des intervalles symétriques par-rapport à μ de type $I = [\mu - ...; \mu + ...]$ (comme dans l'exemple de la partie précédente !)
En effet, il n'y a pas de raison de choisir une zone de rejet différente suivant qu'on soit plus grand ou plus petit que μ.

IV.3-a) Tests de conformité d'une moyenne connaissant σ^2

Propriété 13

Soit $(X_1, \ldots, X_n)$ est un n-échantillon d'une variable aléatoire X d'espérance μ et d'écart-type σ. En notant :

$$\overline{X_n} = \frac{X_1 + \ldots + X_n}{n}$$

Alors, pour tout $a \geqslant 0$, on a

$$P\left(\left|\frac{\overline{X_n} - \mu}{\sigma/\sqrt{n}}\right| < a\right) = P\left(\mu - a\frac{\sigma}{\sqrt{n}} < \overline{X_n} < \mu + a\frac{\sigma}{\sqrt{n}}\right) \xrightarrow[n\to+\infty]{} 2\Phi(a) - 1$$

`Démonstration :`

C'est une conséquence de la première forme du théorème central limite. En effet, si $a > 0$, Posons

$$\begin{aligned} P_n &= P\left(\left|\frac{\overline{X_n} - \mu}{\sigma/\sqrt{n}}\right| < a\right) \\ &= P\left(-a < \frac{\overline{X_n} - \mu}{\sigma/\sqrt{n}} < a\right) \\ &= P\left(-a\frac{\sigma}{\sqrt{n}} < \overline{X_n} - \mu < a\frac{\sigma}{\sqrt{n}}\right) \\ &= P\left(\mu - a\frac{\sigma}{\sqrt{n}} < \overline{X_n} < \mu + a\frac{\sigma}{\sqrt{n}}\right) \end{aligned}$$

Or, d'après le théorème central limite, on sait que

$$P_n = P\left(-a < \frac{\overline{X_n} - \mu}{\sigma/\sqrt{n}} < a\right) \xrightarrow[n\to+\infty]{} \phi(a) - \phi(-a) = 2\phi(u) - 1$$

□

Cette propriété signifie que si n est grand, **quelle que soit** la variable X, on peut supposer que

$$P\left(\mu - a\frac{\sigma}{\sqrt{n}} < \overline{X_n} < \mu + a\frac{\sigma}{\sqrt{n}}\right) \simeq 2\Phi(a) - 1$$

Mais comment utiliser ceci ? ? Voyons ça dans l'exemple ci-dessous.

■ Exemple 14 :

Reprenons le cadre de l'exemple des palettes de gariguettes et observons ce qui a été fait :

On a noté μ_2 la valeur moyenne théorique du poids d'une fraise sur la palette étudiée, ainsi que l'hypothèse nulle

$$H_0 : \mu_2 = 16$$

et comme hypothèse alternative

$$H_1 : \mu_2 \neq 16$$

La propriété précédente nous dit alors que

$$I = \left[16 - a\frac{\sigma}{\sqrt{n}}; 16 + a\frac{\sigma}{\sqrt{n}}\right]$$

où a vérifie

$$2\Phi(a) - 1 = 0,95$$

et où $\sigma = 2$, $n = 100$. Il suffit donc de finir le calcul. On trouvait ici $a \simeq 1,96$ et

$$I = [15,61; 16,39]$$

⚠ Remarque :

Il n'est normalement pas nécessaire d'apprendre cette formule par coeur, étant donné que l'exercice vous amène généralement à la retrouver.

`Cas particulier classique d'une proportion :`

Comme nous l'avons vu dans ce chapitre, on rappelle qu'une probabilité p d'un événement B (donc la "proportion théorique") est approchée par la proportion empirique de réalisation dans un n échantillon d'une même variable X (avec n grand) :

$$\overline{X_n} = \frac{X_1 + \ldots + X_n}{n} = \frac{\textit{nombre de fois où } B \textit{ se produit}}{n}$$

avec

$$X = \begin{cases} 1 & \textit{si } B \textit{ se réalise} \\ 0 & \textit{sinon} \end{cases}$$

et où on rappelle que

$$p = \mathbb{E}[X] = \mu, \quad \sigma = \sqrt{p(1-p)}$$

La propriété permettant de faire le test est donc la suivante :

Propriété 14

Soit p la probabilité d'apparition d'un événement E dans une population de référence, et $\overline{X_n}$ la proportion empirique d'apparition sur un échantillon de la population. Alors

$$P\left(p - a\frac{\sqrt{p(1-p)}}{\sqrt{n}} < \overline{X_n} < p + a\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right) \xrightarrow[n\to+\infty]{} 2\Phi(a) - 1$$

En pratique, on peut dire, comme dans le théorème de Moivre-Laplace, que dès que $n \geqslant 30$, $np \geqslant 5$, $n(1-p) \geqslant 5$, on a

$$P\left(p - a\frac{\sqrt{p(1-p)}}{\sqrt{n}} < \overline{X_n} < p + a\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right) \simeq 2\Phi(a) - 1$$

■ Exemple 15 :

Un individu souhaite savoir si son voisin a traité les pommes de son verger avec un pesticide. Il sait par expérience qu'en moyenne, les vergers non traités aux alentours fournissent un taux de 10% de fruits véreux.
Le verger étant immense, il "emprunte" 50 pommes, les coupe en 2, et constate une proportion de pommes véreuses effective de 2%. Trouver un intervalle I symétrique par-rapport à p tel que $P(\overline{X_n} \in I) = 0.95$. Peut-on en conclure si le verger est traité avec une probabilité de 0.95 ?

On pose l'hypothèse

$$H_0 : \text{"le verger n'est pas traité"}$$

(la seule hypothèse sur laquelle on peut faire des calculs avec les données que nous avons.) Sous cette hypothèse, on a

$$P\left(p - a\frac{\sqrt{p(1-p)}}{\sqrt{n}} < \overline{X_n} < p + a\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right) \simeq 2\Phi(a) - 1$$

Or, ayant $n \geqslant 30$, $np = 50 \times 0,1 = 5 \geqslant 5$, $n(1-p) = 45 \geqslant 5$, on a

$$2\Phi(a) - 1 \simeq 0.95 \quad \Leftrightarrow \quad a \simeq 1.96$$

ainsi, un intervalle I correspondant à la demande est

$$I \simeq \left[p - a\frac{\sqrt{p(1-p)}}{\sqrt{n}}; p + a\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right] \simeq \left[0.1 - a\frac{\frac{3}{10}}{\sqrt{50}}; 0.1 + a\frac{\frac{3}{10}}{\sqrt{50}}\right] \simeq [0.017; 0.183]$$

Ici,

$$0.02 \in I$$

L'hypothèse ne peut être rejetée et ainsi :

on ne peut pas exclure le fait que le verger ne soit pas traité.

■ Exemple 16 :

On considère à nouveau la population de fraises gariguettes, dont le poids moyen par fruit est de $\mu = 16g$ et dont la palette a été confondue avec cette d'une autre variété. Le nouveau responsable de l'étiquetage ne connait malheureusement pas la variance théorique liée à cette variété de fruits. Comment faire pour identifier les palettes ? *(essayons de répondre à cette question ci-dessous)*

Le TCL seconde forme donne lieu au résultat suivant :

Propriété 15

Soit $(X_1, \dots, X_n)$ est un n-échantillon d'une variable aléatoire X d'espérance μ et d'écart-type σ. En notant :

$$\overline{X_n} = \frac{X_1 + \dots + X_n}{n} \qquad \text{et} \qquad S_n^2 = \frac{1}{n}\sum_{k=1}^{n}(X_k - \overline{X_n})^2$$

Alors, pour tout $a \geqslant 0$, on a

$$P\left(\left|\frac{\overline{X_n} - \mu}{S_n/\sqrt{n}}\right| < a\right) = P\left(\mu - a\frac{S_n}{\sqrt{n}} < \overline{X_n} < \mu + a\frac{S_n}{\sqrt{n}}\right) \xrightarrow[n\to+\infty]{} 2\Phi(a) - 1$$

`Démonstration :`

La démonstration est en tout point identique à celle faite à la page 17 concernant le premier test de conformité, mais en remplaçant σ par S_n. □

? Exercice 6

Revenons à notre exemple de gariguettes. Prenant exemple sur son prédecesseur, le nouveau responsable de l'étiquetage a prélevé aléatoirement $n = 100$ fraises. Il trouve un poids moyen $\overline{X_n} = 16,5$ et une variance empirique $S_n^2 = 9$.

1. Quelle est la valeur de a qui correspond à la probabilité $P\left(\overline{X_n} \in \left[16 - a\frac{S_n}{\sqrt{n}}; 16 + a\frac{S_n}{\sqrt{n}}\right]\right) = 0.95$

2. En déduire un intervalle I pour lequel $P\left(\overline{X_n} \in I\right) \simeq 0.95$.

3. La valeur moyenne observée $\overline{X_n} = 16,5$ permet-elle de savoir si la palette considérée est la palette de gariguette ?

4. Le responsable prélève maintenant un échantillon sur l'autre palette et trouve cette fois-ci un poids moyen $\overline{X_n} = 17.2$ et une variance empirique $S_n^2 = 4$. Peut-il conclure ?

Solution

1. Pour la probabilité 0.95, D'après la calculatrice, on a encore

$$a \simeq 1,96.$$

2. d'après le TCL, en se plaçant dans l'hypothèse nulle $\mathcal{H}_0$: "la palette est celle de gariguettes", l'intervalle I est le suivant :

$$I \simeq \left[\mu - a\frac{S_n}{\sqrt{n}}; \mu + a\frac{S_n}{\sqrt{n}}\right] = \left[16 - a\frac{\sqrt{9}}{\sqrt{n}}; 16 + a\frac{\sqrt{9}}{\sqrt{n}}\right] \simeq [15.41; 16.59]$$

3. Comme $16,5 \in I$, on ne peut rien conclure.

4. On se place dans l'hypothèse nulle $\mathcal{H}_0$: "la deuxième palette est la palette de gariguette". On obtient alors, avec les mêmes calculs, un intervalle

$$I_2 \simeq \left[\mu - a\frac{\sqrt{4}}{\sqrt{n}}; \mu + a\frac{\sqrt{4}}{\sqrt{n}}\right] = [15,6; 16.4]$$

Cette fois-ci, $17.2 \notin I_2$. On exclu donc l'hypothèse nulle au profit de l'hypothèse alternative. On peut conclure que **ce n'est pas la palette de gariguettes.** Par élimination, la première était donc la bonne !

IV-4 Tests unilatéraux

Ce sont les tests que l'on peut mettre en place si on est déjà certain que l'une des valeurs est supérieure (ou inférieure à l'autre)
Pour un échantillon A et une population de référence $\mathcal{P}$ de moyenne théorique resp. μ_A et μ, ils consistent à opposer deux situations de type

$$H_0 \ : \ \mu_A = \mu \qquad ; \qquad H_1 \ : \ \mu_A > \mu$$

ou

$$H_0 \ : \ \mu_A = \mu \qquad ; \qquad H_1 \ : \ \mu_A < \mu$$

Pour faire nos calculs, on va donc tenter de trouver des intervalles bornés d'un seul côté par rapport à μ, de manière cohérente avec le problème : i.e. de type :

- *pour le premier cas : $I =]-\infty; \mu + \beta]$ (on donne un "maximum" afin de rejeter ce qui est trop loin au dessus)*
- *pour le deuxième cas : $I = [\mu - \beta; +\infty[$ (on donne un "minimum" afin de rejeter ce qui est trop loin en dessous).*

On a malgré tout toujours la possibilité de faire un test bilatéral comme dans la partie précédente, mais nous verrons sur un exemple quel est l'intérêt de faire un test unilatéral.

IV.4-a) On sait déjà que $\mu_A \geqslant \mu$; donc cas de H_1 : "$\mu_A > \mu$"

On cherche à rejeter des intervalles de type $]\mu+\beta, +\infty[$, et donc avoir des intervalles de type

$$\overline{X_n} \in]-\infty; \mu+\beta].$$

Cas de la variance théorique connue, toujours directement par TCL :

Propriété 16

Soit $(X_1, \ldots, X_n)$ est un n-échantillon d'une variable aléatoire X d'espérance μ et d'écart-type σ. En notant : $\overline{X_n} = \frac{X_1+\ldots+X_n}{n}$, alors, pour tout $b \geqslant 0$, on a

$$P\left(\overline{X_n} \leq \mu + b\frac{\sigma}{\sqrt{n}}\right) = P\left(\frac{\overline{X_n}-\mu}{\sigma/\sqrt{n}} \leq b\right) \xrightarrow[n\to+\infty]{} \Phi(b)$$

? Exercice 7

Un individu souhaite savoir si son verger est contaminé par les pesticides de son voisin, qu'il soupçonne d'avoir trop largement diffusés par les airs. Il sait que la probabilité pour une pomme d'être véreuse après traitement par le pesticide en question est de $p = 1\%$.
On note p_A la proportion théorique de pommes véreuses dans son verger. On pose

$$H_0 : \text{"le verger est traité"} \ : "p = p_A"$$

et

$$H_1 : \text{"le verger n'est pas traité"} \ : "p < p_A"$$

1. Soit $\overline{X_n}$ donnant la moyenne empirique de pommes véreuses dans son verger sur n pommes cueillies. Montrer que sous l'hypothèse H_0, avec n suffisamment grand, on a

$$P\left(\overline{X_n} < p + b\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right) \simeq \Phi(b)$$

2. En déduire un intervalle tel que $P(\overline{X_n} \in I) = 0.95$.

3. Sur les 500 pommes dans son verger, son taux effectif de pommes véreuses prélevées est de 3.5%. Peut-on en conclure si son verger a été contaminé avec une probabilité de 0.95 ?

Solution

1. On a $\overline{X_n} = \frac{T_n}{n}$, où T_n est le nombre de pommes véreuses trouvées sur les n cueillies (et indépendantes), d'où $T_n \hookrightarrow \mathcal{B}(n,p)$.

Ainsi, sous l'hypothèse H_0, σ est connu et vaut

$$\sigma = \sqrt{p(1-p)}$$

D'après le théorème de Moivre Laplace, on sait de plus que, si $n \geq 30,\ np \geq 5, n(1-p) \geq 5$, on a

$$P\left(\overline{X_n} \leq \mu + b\frac{\sigma}{\sqrt{n}}\right) \simeq \Phi(b)$$

ce qui correspond à ce qui est demandé.

2. On a bien $n \geq 30,\ np \geq 5, n(1-p) \geq 5$. Ainsi

$$\Phi(b) \simeq 0.95 \quad \Leftrightarrow \quad b \simeq 1.645$$

ainsi, un intervalle I correspondant à la demande est

$$I \simeq \left]-\infty; p + b\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right] \simeq]-\infty; 0.0173]$$

3. Ici,

$$0.035 \notin I$$

L'hypothèse H_0 est rejetée.

On en déduit que le verger n'est pas contaminé.

Cas de la variance théorique inconnue :

Propriété 17

Soit $(X_1, \ldots, X_n)$ un n-échantillon d'une variable aléatoire X d'espérance μ et d'écart-type σ. En notant

$$\overline{X_n} = \frac{X_1+\ldots+X_n}{n} \qquad \text{et} \qquad S_n^2 = \frac{1}{n}\sum_{k=1}^{n}(X_k - \overline{X_n})^2$$

Alors, pour tout $b \geqslant 0$, on a

$$P\left(\overline{X_n} \leq \mu + b\frac{S_n}{\sqrt{n}}\right) = P\left(\frac{\overline{X_n}-\mu}{S_n/\sqrt{n}} \leq b\right) \xrightarrow[n\to+\infty]{} \Phi(b)$$

? Exercice 8

Un ingénieur agronome souhaite savoir si le nouvel engrais qu'il envisage de produire augmente oui ou non le nombre moyen de tomate par pied. L'ancien engrais donnait un nombre moyen de $\mu = 14,38$ tomates. On note μ_A le nombre moyen de tomates avec le nouvel engrais. Il effectue le test sur un échantillon de $n = 300$ pieds. On pose

$$\mathcal{H}_0 : \text{'l'engrais n'a pas d'effet'}, \quad \mathcal{H}_1 : \text{'l'engrais a un effet positif'}$$

i.e.

$$\mathcal{H}_0 : \mu = \mu_A, \quad \mathcal{H}_1 : \mu_A > \mu$$

4. Soit $\overline{X_n}$ le nombre moyen de tomate par plant ainsi que S_n l'écart type observé. Montrer que sous l'hypothèse H_0, on a

$$P\left(\overline{X_n} \leq \mu + b\frac{S_n}{\sqrt{n}}\right) \simeq \Phi(b)$$

5. Les données effectives obtenues donnent un nombre moyen de $15,2$ avec un écart-type de $S = \sqrt{3,5}$. En déduire un intervalle tel que $P(\overline{X_n} \in I) = 0.99$ et conclure.

Solution

1. Le calcul et les arguments sont exactement les même que dans la propriété.

2. On obtient $b \simeq 2.326$ et

$$I \simeq]-\infty, 14.63]$$

Ainsi, $15,2 \notin I$ et on peut en effet conclure à une amélioration de l'engrais par rapport à l'ancienne version.

Commentaires généraux sur la méthode unilatérale :

Remarque :

Avec le test unilatéral, par ex. dans l'exercice sur les pesticides précédent, on trouve

$$I \simeq]-\infty; 0.033]\,.$$

Avec un test bilatéral, i.e. H_1 : "$p \neq p_A$" *(que vous n'hésiterez pas à faire pour vous entrainer)*, les résultats auraient été :

$$I' \simeq [-0.017; 0.037].$$

Dans ce cas, on aurait eu

$$0.035 \in I'$$

et on aurait conclut sur le fait qu'on ne peut pas conclure !

Commentaires :

En bref, le test unilatéral est plus "puissant" que le test bilatéral. En effet, comme on sait que nécessairement $\mu_A \geq \mu$, il a en réalité une zone de rejet plus "grande" (pour les cas supérieurs) :

$$> 0.033 \text{ pour le test unilatéral, et } > 0,037 \text{ pour le test bilatéral}$$

IV.4-b) Cas de H_1 : "$\mu_A < \mu$"

On inverse la procédure par-rapport à la partie précédente pour avoir des intervalles de type $\overline{X_n} \in [\mu - \beta; +\infty[$.

Cas de la variance théorique connue, toujours directement par TCL :

Propriété 18

Soit $(X_1, \ldots, X_n)$ est un n-échantillon d'une variable aléatoire X d'espérance μ et d'écart-type σ. En notant :

$$\overline{X_n} = \frac{X_1 + \ldots + X_n}{n}$$

Alors, pour tout $b \geqslant 0$, on a

$$P\left(\mu - b\frac{\sigma}{\sqrt{n}} \leq \overline{X_n}\right) = P\left(b \leq \frac{\overline{X_n} - \mu}{\sigma/\sqrt{n}}\right) \xrightarrow[n\to+\infty]{} 1 - \Phi(b)$$

Démonstration :

On utilise toujours le TCL, mais en passant à l'événement contraire :

$$P\left(b \leq \frac{\overline{X_n} - \mu}{\sigma/\sqrt{n}}\right) = 1 - P\left(\frac{\overline{X_n} - \mu}{\sigma/\sqrt{n}} < b\right) = 1 - \Phi(b)$$

□

Exercice 9

Un producteur des pesticides a mis point un autre produit qu'il sait être au moins aussi efficace que la version précédente, dont on rappelle qu'elle donnait seulement $p = 1\%$ de pommes véreuses.
Il veut maintenant savoir si l'écart est significatif et si la nouvelle version est en effet plus efficace que l'autre. Il effectue donc son test dans le verger de votre voisin...
On note p_A la proportion théorique de pommes véreuses dans ledit jardin. On pose

$$H_0 : \text{"le pesticide n'est pas plus efficace"} : "p = p_A"$$

et

$$H_1 : \text{"le pesticide est plus efficace"} : "p > p_A"$$

1. Soit $\overline{X_n}$ donnant la moyenne empirique de pommes véreuses dans le verger sur n pommes cueillies. Montrer que sous l'hypothèse H_0, on a

$$P\left(p - b\frac{\sqrt{p(1-p)}}{\sqrt{n}} \leqslant \overline{X_n}\right) \simeq \Phi(b)$$

2. En déduire un intervalle tel que $P(\overline{X_n} \in I) = 0.95$.

3. Il fait un test sur 500 pommes. Son taux effectif de pommes véreuses prélevées est de 0.5%. Peut-on en conclure si son pesticide est plus efficace ?

Solution

1. On a $\overline{X_n} = \frac{T_n}{n}$, où T_n est le nombre de pommes véreuses trouvées sur les n cueillies (et indépendantes), d'où $T_n \hookrightarrow \mathcal{B}(n,p)$ où $p = 0.1$.
Ainsi, sous l'hypothèse H_0, σ est connu et vaut

$$\sigma = \sqrt{p(1-p)}$$

D'après le TCL, on sait de plus que, comme $n \geq 30$, $np \geq 5$, $n(1-p) \geq 5$,

$$P\left(\mu - b\frac{\sigma}{\sqrt{n}} \leq \overline{X_n} < +\infty\right) \simeq 1 - \Phi(-b) = \Phi(b)$$

ce qui correspond à ce qui est demandé.

2. On a

$$\Phi(b) \simeq 0.95 \quad \Leftrightarrow \quad b \simeq 1.645$$

ainsi, un intervalle I correspondant à la demande est

$$I \simeq \left[p - b\frac{\sqrt{p(1-p)}}{\sqrt{n}}; +\infty\right[\simeq [0.0027; +\infty[$$

3. Ici,

$$0.005 \in I$$

Ce n'est pas significatif, on ne peut pas conclure.

Cas de la variance théorique inconnue :

Propriété 19

Soit $(X_1, \ldots, X_n)$ est un n-échantillon d'une variable aléatoire X d'espérance μ et d'écart-type σ. En notant :

$$\overline{X_n} = \frac{X_1 + \ldots + X_n}{n} \qquad \text{et} \qquad S_n^2 = \frac{1}{n}\sum_{k=1}^{n}(X_k - \overline{X_n})^2$$

Alors, pour tout $b \geqslant 0$, on a

$$P\left(\mu - b\frac{S_n}{\sqrt{n}} \leq \overline{X_n}\right) = P\left(b \leq \frac{\overline{X_n} - \mu}{S_n/\sqrt{n}}\right) \xrightarrow[n\to+\infty]{} 1 - \Phi(b)$$

? Exercice 10

Un laboratoire souhaite savoir si le nouveau médicament qu'il envisage de distribuer contre les migraines améliore oui ou non le temps total des crises. L'ancien médicament donne une vitesse de "guérison" moyenne de $\mu = 3,25h$. On note μ_A la durée moyenne (pour l'instant inconnue) avec le nouveau médicament. Il effectue le test sur un échantillon de $n = 200$ personnes atteintes de ladite maladie. On pose

$$\mathcal{H}_0 : \text{'le médicament n'a pas d'effet'}, \quad \mathcal{H}_1 : \text{'le médicament a un effet positif'}$$

i.e.

$$\mathcal{H}_0 : \mu = \mu_A, \quad \mathcal{H}_1 : \mu_A < \mu$$

1. Soit $\overline{X_n}$ la vitesse moyenne de guérison de n patients ainsi que S_n l'écart type observé. Montrer que sous l'hypothèse H_0, on a

$$P\left(\overline{X_n} \geq \mu - b\frac{S_n}{\sqrt{n}}\right) \simeq \Phi(b)$$

2. Les données effectives obtenues sont une durée moyenne de crise de $2,8h$ avec un écart-type de $S = \sqrt{1,6}$. En déduire un intervalle tel que $P(\overline{X_n} \in I) = 0.99$ et conclure.

Solution

1. D'après le TCL, le calcul et les arguments sont exactement les même que dans le cas où σ est connue, mais en remplaçant σ par S_n. On peut donc reprendre la même démonstation que dans l'exercice précédent.

2. On obtient $b \simeq 2.326$ et

$$I \simeq [3.042 : +\infty[$$

Ainsi, $2,8 \notin I$ et on peut en effet conclure à une amélioration du médicament par-rapport à l'ancien.