Statistiques et estimation

Table des matières

1 Intervalle de fluctuation

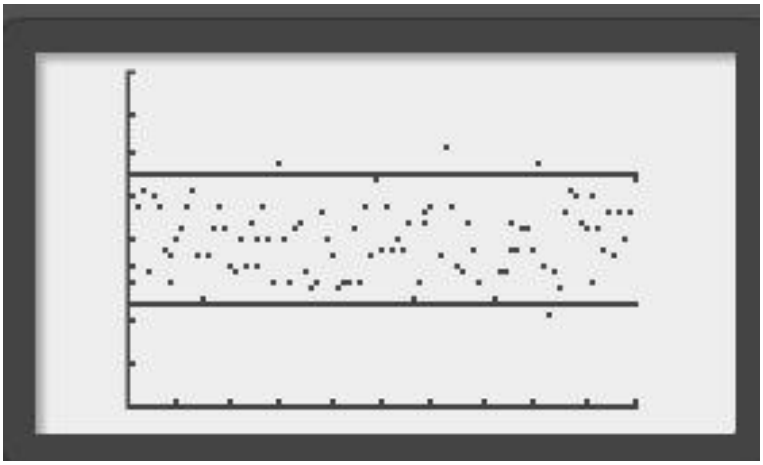
1.1 Simulation

On lance 120 fois un dé à jouer bien équilibré. On appelle N la variable aléatoire qui associe le nombre de fois que le dé affiche la face 6. On voudrait savoir la probabilité que la variable aléatoire N soit comprise dans l'intervalle [12 ;28].
On écrit le programme ci contre. Ce programme effectue 100 fois ces 120 lancers. On affiche à chaque expérience I le point (I, N) ainsi que les droites d'équations $y = 12$ et $y = 28$. A la fin de ces 100 expériences, on affiche le nombre de points M qui se situe dans l'intervalle [12 ;28].

On trouve alors : $M = 96$. On peut alors dire qu'à 96 %, le nombre d'apparitions de la face 6 se situe dans l'intervalle [12 ;28]. On nomme alors cet l'intervalle, **intervalle de fluctuation de N au seuil de 96 %**.

Variables
A, B, I, J, M, N, X
Initialisation
Effacer dessin
$0 \to M$
$12 \to A$
$28 \to B$
Tracer $y = A$
Tracer $y = B$
Traitement
Pour I de 1 à 100
  $0 \to N$
    Pour J de 1 à 120
    randInt(1,6) $\to X$
    Si $X = 6$
      $N + 1 \to N$
    FinSi
  FinPour
  Afficher le point $(I; N)$
  Si $N \geqslant A$ et $N \leqslant B$
    $M + 1 \to M$
  Fin Si
FinPour
Sortie
Afficher M

Sur une calculatrice TI 82 plus, on obtient le graphe suivant :



1.2 Définition

Définition 1 : X est une variable aléatoire qui suit une loi binomiale $\mathscr{B}(n, p)$.

α est un réel tel que $0 < \alpha < 1$ et a et b sont deux réels.

On dit que $[a; b]$ est un **intervalle de fluctuation de X** au seuil de $1-\alpha$, si, et seulement si :

$$P(a \leqslant X \leqslant b) \geqslant 1 - \alpha$$

1.3 Intervalle de fluctuation asymptotique

Théorème 1 : Si la variable aléatoire X_n suit une loi binomiale $\mathscr{B}(n,p)$ alors pour tout réel α de]0 ;1[, on a :

$$\lim_{n\to+\infty} P\left(\frac{X_n}{n} \in I_n\right) = 1-\alpha \quad \text{où} \quad I_n = \left[p - u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}} \ ; \ p + u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}}\right]$$

u_α étant le nombre tel que $P(-u_\alpha \leqslant Z_n \leqslant u_\alpha) = 1-\alpha$ lorsque Z_n suit une loi normale centrée réduite.

On appelle **variable fréquence**, la variable aléatoire $F_n = \dfrac{X_n}{n}$ qui à tout échantillon de taille n associe la fréquence f obtenue.

Remarque : Le mot asymptotique vient du passage à la limite de l'intervalle I_n, la loi binomiale $\mathscr{B}(n,p)$ peut alors être assimilé à la loi normale $\mathscr{N}(np, np(1-p))$.

ROC

Démonstration : On pose $Z_n = \dfrac{X_n - np}{\sqrt{np(1-p)}}$. On pourra utiliser cet intervalle de fluctuation dans les conditions de l'approximation normale de la loi binomiale ($n \geqslant 30$, $np \geqslant 5$ et $n(p-1) \geqslant 5$)

D'après le théorème Moivre-Laplace : $\lim\limits_{n\to+\infty} P(-u_\alpha \leqslant Z_n \leqslant u_\alpha) = \displaystyle\int_{-u_\alpha}^{u_\alpha} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\,dt$ c'est à dire que Z_n suit une loi normale centrée réduite.

On sait d'après les propriétés de la loi normale centrée réduite que pour tout α de]0 ;1[, il existe un unique réel strictement positif u_α tel que :

$$P(-u_\alpha \leqslant Z_n \leqslant u_\alpha) = 1-\alpha$$

De plus :

$$-u_\alpha \leqslant Z_n \leqslant u_\alpha$$

$$-u_\alpha \cdot \sqrt{np(1-p)} \leqslant X_n - np \leqslant u_\alpha \cdot \sqrt{np(1-p)}$$

$$np - u_\alpha \cdot \sqrt{np(1-p)} \leqslant X_n \leqslant np + u_\alpha \cdot \sqrt{np(1-p)}$$

$$p - u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}} \leqslant \frac{X_n}{n} \leqslant p + u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}}$$

Donc $\lim\limits_{n\to+\infty} P\left(\dfrac{X_n}{n} \in I_n\right) = 1-\alpha$

Propriété : Il faut connaître l'intervalle I_n de **fluctuation au seuil de 95%** correspondant à $\alpha = 0,05$ et qui donne (voir chapitre précédent p 12) $u_\alpha = 1,96$

$$I_n = \left[p - 1,96 \frac{\sqrt{p(1-p)}}{\sqrt{n}} \ ; \ p + 1,96 \frac{\sqrt{p(1-p)}}{\sqrt{n}}\right]$$

Exemple : Si on reprend l'exemple sur les 120 lancers de dé à jouer avec N comme variable aléatoire. L'intervalle de fluctuation asymptotique au seuil de 95 % (dans les conditions de l'approximation normale) est alors :

$$p - 1,96\frac{\sqrt{p(1-p)}}{\sqrt{n}} = \frac{1}{6} - 1,96\frac{\sqrt{\frac{1}{6}\times\frac{5}{6}}}{\sqrt{120}} \simeq 0,100$$

$$p + 1,96\frac{\sqrt{p(1-p)}}{\sqrt{n}} = \frac{1}{6} + 1,96\frac{\sqrt{\frac{1}{6}\times\frac{5}{6}}}{\sqrt{120}} \simeq 0,233$$

Donc $\mathrm{I}_n = [0,100; 0,233]$ qui correspond à la variable aléatoire fréquence $\dfrac{N}{120}$.

Si on revient à la variable N, l'intervalle de fluctuation est alors :
$[120 \times 0,100\ ;\ 120 \times 0,233] = [12\ ;\ 28]$, ce qui confirme notre expérience (on avait trouvé 96 %).

Remarque : Cet intervalle peut être simplifié par l'intervalle

$$\mathrm{J}_n = \left[p - \frac{1}{\sqrt{n}}\ ;\ p + \frac{1}{\sqrt{n}}\right]$$

En effet la fonction $x \mapsto x(1-x) = x - x^2$ est une fonction du second degré qui s'annule en 0 et 1, elle admet donc un maximum (coefficient négatif devant x^2) en 0,5. On a alors $f(0,5) = 0,25$. Elle est positive entre 0 et 1. On a alors :

$$0 \leqslant p(1-p) \leqslant 0,25 \quad \Leftrightarrow \quad 0 \leqslant \sqrt{p(1-p)} \leqslant \sqrt{0,25} = 0,5$$

On en déduit alors que : $0 \leqslant 1,96\sqrt{p(1-p)} \leqslant 1$

On a alors $0 \leqslant 1,96\dfrac{\sqrt{p(1-p)}}{\sqrt{n}} \leqslant \dfrac{1}{\sqrt{n}}$

On a ainsi $\mathrm{I}_n \subset \mathrm{J}_n$. On a alors dans la plupart des cas $P(F_n \in \mathrm{J}_n) \geqslant 0,95$

1.4 Prise de décision

Propriété 1 : Soit f_{obs} la fréquence observée d'un caractère sur un échantillon de taille n issu d'une population donnée. On suppose que les conditions de l'approximation normale de la loi binomiale sont remplies : $n \geqslant 30$, $np \geqslant 5$ et $n(1-p) \geqslant 5$.

Test d'hypothèse : On fait une conjecture sur la valeur de la proportion p du caractère étudié dans la population toute entière.

Soit I_n l'intervalle de fluctuation asymptotique au seuil de 95 %.
- Si $f_{\text{obs}} \in I_n$; on ne peut rejeter l'hypothèse faite sur p.
- Si $f_{\text{obs}} \notin I_n$; on rejete l'hypothèse faite sur p.

Exemple : Pour créer ses propres colliers, on peut acheter un kit contenant des perles de cinq couleurs différentes (marrons, jaunes, rouges, vertes et bleues), dans des proportions affichées sur le paquet.

Ainsi les perles marron et les perles jaunes sont annoncées comme représentant chacunes 20 % de l'ensemble des perles tandis que les perles rouges sont annoncées à 10 %.
On veut vérifier cette information. Pour cela, on choisit d'observer un échantillon aléatoire de perles et de construire un intervalle de fluctuation asymptotique au seuil de 95 % pour la proportion de perles marron.
On constitue donc un échantillon, que l'on considère aléatoire, de 690 perles. On a dénombré 140 perles marron.
La **prise de décision** est la suivante : si la proportion de perles marron dans l'échantillon n'appartient pas à intervalle de fluctuation, on rejette l'hypothèse selon laquelle les perles marron représentent 20 % des perles

a) Déterminer l'intervalle de fluctuation asymptotique I au seuil de 95 % pour la proportion de perles marron.
b) Calculer la proportion de perles marron dans l'échantillon. Que peut-on en conclure ?
c) Dans le même échantillon, il y avait 152 perles jaunes et 125 perles rouges. Que peut-on conclure de ces résultats ?

✏ ✏

a) En ce qui concerne les perles marron, on a : $n = 690$ et $p = 0,2$, donc :

$$n \geqslant 30 \quad np = 138 \geqslant 5 \quad \text{et} \quad n(1-p) = 552 \geqslant 5$$

Nous sommes bien dans les hypothèses du théorème de Moivre-Laplace.

On calcule ensuite

$$p - 1,96\frac{\sqrt{p(1-p)}}{\sqrt{n}} = 0,2 - 1,96\frac{\sqrt{0,2 \times 0,8}}{\sqrt{690}} \simeq 0,1702$$
$$p + 1,96\frac{\sqrt{p(1-p)}}{\sqrt{n}} = 0,2 + 1,96\frac{\sqrt{0,2 \times 0,8}}{\sqrt{690}} \simeq 0,2298$$

On a donc : $\mathrm{I} = [0,17; 0,23]$

b) On calcule la fréquence $f_m = \dfrac{140}{690} \simeq 0,203$

Comme $f_m \in \mathrm{I}$, **on ne peut pas rejeter** l'hypothèse selon laquelle les perles marron représentent 20 % des perles.

c) On calcule la fréquence des perles jaunes : $f_j = \dfrac{152}{690} \simeq 0,220$

Comme $f_j \in \mathrm{I}$, **on ne peut pas rejeter** l'hypothèse selon laquelle les perles jaunes représentent 20 % des perles.

Pour les perles rouges, il faut calculer un nouvel intervalle de fluctuation :

$$p - 1,96\frac{\sqrt{p(1-p)}}{\sqrt{n}} = 0,1 - 1,96\frac{\sqrt{0,1 \times 0,9}}{\sqrt{690}} \simeq 0,0776$$
$$p + 1,96\frac{\sqrt{p(1-p)}}{\sqrt{n}} = 0,1 + 1,96\frac{\sqrt{0,1 \times 0,9}}{\sqrt{690}} \simeq 0,1224$$

On a donc : $I' = [0,07; 0,13]$ (on prend l'intervalle par excès)

On calcule la fréquence des perles rouges : $f_r = \dfrac{125}{690} \simeq 0,18$

Comme $f_r \notin I'$, **on doit rejeter** l'hypothèse selon laquelle les perles rouges représentent 10 % des perles.

2 Estimation

2.1 Présentation du problème

Pour des raisons de coût et de faisabilité, on ne peut étudier un certain caractère sur l'ensemble d'une population. La proportion p de ce caractère est donc inconnue.

On cherche alors à estimer p à partir d'un échantillon de taille n. On calcule alors la fréquence f_{obs} des individus de cet échantillon ayant ce caractère.

Estimation : On estime la proportion p par un intervalle de confiance déterminé à partir de la fréquence f_{obs} et de la taille n de l'échantillon.

Remarque : La fréquence f_{obs} calculée varie d'un échantillon à l'autre du fait de la fluctuation d'échantillonnage. Il est donc nécessaire d'apprécier l'incertitude en donnant une estimation par un intervalle.

2.2 Intervalle de confiance

On suppose les trois conditions d'approximations remplies :

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

Théorème 2 : Soit F_n la variable aléatoire qui à chacun des échantillons de taille n associe la fréquence du caractère dans cet échantillon.

La proportion inconnue p est telle que :

$$P\left(F_n - \frac{1}{\sqrt{n}} \leqslant p \leqslant F_n + \frac{1}{\sqrt{n}}\right) \geqslant 0,95$$

ROC

Démonstration : On a vu que l'intervalle de fluctuation au seuil de 95% peut être simplifié par : $\left[p - \dfrac{1}{\sqrt{n}}; p + \dfrac{1}{\sqrt{n}}\right]$

On a donc :

$$\begin{aligned}
p - \frac{1}{\sqrt{n}} \leqslant F_n \leqslant p + \frac{1}{\sqrt{n}} \\
-\frac{1}{\sqrt{n}} \leqslant F_n - p \leqslant \frac{1}{\sqrt{n}} \\
-F_n - \frac{1}{\sqrt{n}} \leqslant -p \leqslant -F_n + \frac{1}{\sqrt{n}} \\
F_n - \frac{1}{\sqrt{n}} \leqslant p \leqslant F_n + \frac{1}{\sqrt{n}}
\end{aligned}$$

Ainsi : $P\left(F_n - \dfrac{1}{\sqrt{n}} \leqslant p \leqslant F_n + \dfrac{1}{\sqrt{n}}\right) \geqslant 0,95$

Définition 2 : On observe la fréquence f_{obs} sur un échantillon de taille n et p désigne la proportion inconnue d'apparition du caractère dans la population entière. On appelle **intervalle de confiance** de p au niveau asymptotique de 95% l'intervalle :

$$\left[f_{obs} - \frac{1}{\sqrt{n}} \,;\, f_{obs} + \frac{1}{\sqrt{n}}\right]$$

Cet intervalle de confiance a pour amplitude $\dfrac{2}{\sqrt{n}}$. Ainsi si l'on souhaite encadrer p dans un intervalle de longueur a, on doit avoir :

$$\frac{2}{\sqrt{n}} \leqslant a \quad \Leftrightarrow \quad n \geqslant \frac{4}{a^2}$$

Exemple : Un sondage pour l'élection présidentielle du 21 avril 2002

Voici les résultats d'un sondage IPSOS réalisé avant l'élection présidentielle de 2002 pour Le Figaro et Europe 1, les 17 et 18 avril 2002 auprès de 989 personnes, constituant un échantillon national représentatif de la population française âgée de 18 ans et plus et inscrite sur les listes électorales.
On suppose cet échantillon constitué de manière aléatoire (même si en pratique cela n'est pas le cas). Les intentions de vote au premier tour pour les principaux candidats sont les suivantes :

Jacques Chirac : 20 % Lionel Jospin : 18 % Jean-Marie Le Pen : 14 %.

Les médias se préparent pour un second tour entre Jacques Chirac et Lionel Jospin.

a) Déterminer pour chaque candidat, l'intervalle de confiance au niveau de confiance de 0,95 de la proportion inconnue d'électeurs ayant l'intention de voter pour lui.

b) Le 21 avril, les résultats du premier tour des élections sont les suivantes : Jacques Chirac : 19,88 %, Lionel Jospin : 16,18 %, Jean-Marie Le Pen : 16,86 %. Les pourcentages de voix recueillies par chaque candidat sont-ils bien dans les intervalles de confiance précédents ?

c) Pouvait-on, au vu de ce sondage, écarter avec un niveau de confiance de 0,95, l'un de ces trois candidats second tour ?

a) Nous sommes dans les trois hypothèses d'approximation :
$989 \geqslant 30$, $138 \leqslant np \leqslant 198$ et $791 \leqslant n(1-p) \leqslant 851$
On calcule : $\dfrac{1}{\sqrt{n}} = \dfrac{1}{\sqrt{989}} \simeq 0,032$
On obtient alors les intervalles de confiance à 0,95 suivant :

- Pour J. Chirac $I_1 = [0,168 \,;\, 0,232]$
- Pour L. Jospin $I_2 = [0,148 \,;\, 0,212]$

- Pour J.M. Le Pen $I_2 = [0,108\ ;\ 0,172]$

b) Les résultats sont bien dans les intervalles de confiance.

c) Les trois intervalles de confiance ont une intersection non vide :
$I_1 \cap I_2 \cap I_3 = [0,168\ ;\ 0,172]$
Il n'était donc pas possible de donner le classement final des trois candidats. Tous les classements étaient possibles.