Statistiques et estimation

Intervalle de fluctuation

X variable aléatoire qui suit la loi binomiale $\mathscr{B}(n\,;\,p)$

- On dit que l'intervalle $[a\,;\,b]$ est un intervalle de fluctuation de X au seuil de $1-\alpha$ ssi :

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

  Le seuil le plus utilisé est 95 % pour $\alpha = 0,05$.
  Parfois, on utilise le seuil de 99 % pour $\alpha = 0,01$

- Dans le cas : $n \geqslant 30$, $np \geqslant 5$, $n(1-p) \geqslant 5$.
  L'intervalle de fluctuation asymptotique I_n de la variable fréquence $F_n = \dfrac{X}{n}$ au seuil de 95 % est égal à :

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

Exemple

On lance 120 fois un dé non pipé, l'intervalle de fluctuation asymptotique au seuil de 95 % de la fréquence d'apparition de la face 6 est avec $n = 120$ et $p = \dfrac{1}{6}$

- $p - 1,96\dfrac{\sqrt{p(1-p)}}{\sqrt{n}} = \dfrac{1}{6} - 1,96\dfrac{\sqrt{\frac{1}{6}\times\frac{5}{6}}}{\sqrt{120}} \approx 0,100$
- $p + 1,96\dfrac{\sqrt{p(1-p)}}{\sqrt{n}} = \dfrac{1}{6} + 1,96\dfrac{\sqrt{\frac{1}{6}\times\frac{5}{6}}}{\sqrt{120}} \approx 0,233$
- $\mathrm{I}_{120} = [0,100\,;\,0,233]$

Valeurs à retenir

X suit une loi normale centrée réduite alors :

- $p(-1,96 \leqslant X \leqslant 1,96) = 0,95$
- $p(-2,58 \leqslant X \leqslant 2,58) = 0,99$

Proportion inconnue

On cherche à estimer p, la proportion d'un caractère d'une population, à partir d'un échantillon de taille n. On ne connaît pas *a priori* cette proportion.

On suppose que : $n \geqslant 30$, $np \geqslant 5$ et $n(1-p) \geqslant 5$.

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 :

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

Intervalle de confiance

Soit la fréquence f_{obs} sur un échantillon de taille n et soit p la proportion inconnue d'apparition du caractère dans la population entière. On appelle intervalle de confiance $\mathrm{I_c}$ de p au niveau asymptotique de 95 % l'intervalle :

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

Si l'on veut encadrer p dans un intervalle de longueur a :

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

Exemple

Un sondage donne les résultats suivants :

> 52,9 % des électeurs* voteraient pour le candidat A.
> *estimation après redressement, fondée sur un sondage d'un échantillon représentatif de 1 200 personnes.

$f_{\text{obs}} = 0,529$ et $n = 1200$.

Conditions : $n \geqslant 30$, $nf_{\text{obs}} \geqslant 5$ et $n(1-f_{\text{obs}}) \geqslant 5$.

$$\mathrm{I_c} = \left[0,529 - \frac{1}{\sqrt{1200}}\,;\, 0,529 + \frac{1}{\sqrt{1200}}\right] = [0,500\,;\,0,558]$$

Le candidat peut croire en sa victoire (au seuil de 95 %)

Prise de décision

Soit f_{obs} la fréquence observée d'un caractère sur un échantillon de taille n issu d'une population donnée.
On suppose que : $n \geqslant 30$, $np \geqslant 5$ 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.

I_n 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 rejette l'hypothèse faite sur p.

Exemple

On lance 120 fois un dé. On obtient 30 fois la face 6. Le dé est-il bien équilibré.

- $f_{\text{obs}} = \dfrac{30}{120} = 0,25$
- $\mathrm{I}_{120} = [0,100\,;\,0,233]$

On constate que $f_{\text{obs}} \notin \mathrm{I}_{120}$, on doit donc rejeter l'hypothèse que le dé est bien équilibrée au seuil de 95 % (soit au risque de 5 %).
Le dé est considéré comme pipé.

Un intervalle de fluctuation pour quoi faire ?

Contexte du Bac

On utilise l'intervalle I_n, avec $n \geqslant 30$ pour prendre une décision, typiquement suspecter une fraude, mettre en évidence une publicité mensongère :

- Étape 1 : On s'assure que l'on est dans les conditions d'application de l'approximation normale : $n \geqslant 30,\ np \geqslant 5$ et $n(1-p) \geqslant 5$.
- Étape 2 : On détermine les bornes de l'intervalle de I_n en respectant les consignes d'arrondi. On arrondit en effet la borne inférieure par défaut et la borne supérieure par excès.
- Étape 3 : On estime f_{obs}, la fréquence observée sur le terrain, i.e. sur l'échantillon de taille n prélevé.
  Si $f_{\text{obs}} \notin \mathrm{I}_n$, on rejette l'hypothèse de p au seuil de 95 % i.e. au risque de 5 % de commettre une erreur.
  Si $f_{\text{obs}} \in \mathrm{I}_n$, on valide l'hypothèse sur la valeur de p au seuil de 95 %.

Exemple

Pour des raisons pratiques, un astronome responsable du club souhaiterait installer un site d'observation sur les hauteurs d'une petite ville de 2 500 habitants. Mais la pollution lumineuse due à l'éclairage public nuit à la qualité des observations. Pour tenter de convaincre la mairie de couper l'éclairage nocturne pendant les nuits d'observation, l'astronome réalise un sondage aléatoire auprès de 100 habitants et obtient 54 avis favorables à la coupure de l'éclairage nocturne.
L'astronome fait l'hypothèse que 50 % de la population du village est favorable à la coupure de l'éclairage nocturne. Le résultat de ce sondage l'amène-t-il à changer d'avis ?

- Étape 1 : $n = 100 \geqslant 30,\ p = 0,5$ donc $np = 50 \geqslant 5$ et $n(1-p) = 50 \geqslant 5$.
  Les conditions de l'approximation normale sont vérifiées.
- Étape 2 : $\mathrm{I}_{100} = \left[0,5 - 1,96\dfrac{\sqrt{0,5^2}}{\sqrt{100}}\ ;\ 0,5 + 1,96\dfrac{\sqrt{0,5^2}}{\sqrt{100}}\right] = [0,402\ ;\ 0,598]$
- Étape 3 : $f_{\text{obs}} = \dfrac{54}{100} = 0,54$, donc $f_{\text{obs}} \in \mathrm{I}_{100}$

  Ce résultat conforte donc l'astronome dans son hypothèse.

La taille de l'échantillon dans un intervalle de confiance

⚠ La taille n de l'échantillon est déterminant car plus n est grand, plus les bornes de l'intervalle sont resserrées et l'estimation « fine ». Cependant recueillir des données sur un grand échantillon coûte cher et demande du temps.

Lorsque les bornes sont trop « lâche » l'information fournie est peu exploitable. On fait donc un compromis pour la taille de l'échantillon : suffisamment grand pour obtenir une information crédible et pas trop grand pour être facile à réaliser.

Exemple

Un institut effectue un sondage pour connaître, dans une population donnée, la proportion de personnes qui sont favorables à un projet d'aménagement du territoire. Pour cela, on interroge un échantillon aléatoire de n personnes, $n \geqslant 50$, de cette population, et l'on pose une question à chaque personne.
Parmi ces personnes, 29 % sont favorables au projet d'aménagement.
Déterminer la valeur minimale de l'entier n pour que l'intervalle de confiance, au niveau de confiance de 95 %, ait une amplitude inférieure ou égale à 0,04.

- On vérifie les conditions d'application de l'intervalle de confiance :
  $$\begin{cases} n \geqslant 50 \geqslant 30 \\ nf_{\text{obs}} \geqslant 0,29 \times 50 \Rightarrow nf_{\text{obs}} \geqslant 14,5 \geqslant 5 \\ n(1-f_{\text{obs}}) \geqslant 0,71 \times 50 \Rightarrow nf_{\text{obs}} \geqslant 35,5 \geqslant 5 \end{cases}$$
- $a \leqslant 0,04 \Leftrightarrow \dfrac{2}{\sqrt{n}} \leqslant 0,04 \Leftrightarrow \sqrt{n} \geqslant \dfrac{2}{0,04} \Leftrightarrow n \geqslant \left(\dfrac{2}{0,04}\right)^2 \Leftrightarrow n \geqslant 2\,500$

Le nombre minimum de personnes à interroger est 2 500