Chap. n°

ECHANTILLONNAGE ET ESTIMATION

I Deux situations à distinguer

Le cadre commun :
On s'intéresse à la fréquence p d'apparition d'un caractère dans une population.
Sur un échantillon de taille n individus, on relève la fréquence observée f_{obs} de ce caractère

Echantillonnage et prise de décision	Estimation
• On connaît la proportion $\boldsymbol{p}$ (fréquence) du caractère dans la population ou • On fait une hypothèse sur la valeur de la proportion $\boldsymbol{p}$ du caractère dans la population.	On n'a aucune idée de la proportion $\boldsymbol{p}$ (fréquence) du caractère dans la population On cherche à l'estimer
→ utilisation d'un intervalle de fluctuation	→ utilisation d'un intervalle de confiance

Exemples :

- On trouve une pièce de monnaie : est-elle équilibrée ?
  On réalise une série de lancers (un « échantillon » de lancers) et on prend la décision de valider ou pas l'hypothèse « la pièce est équilibrée » en fonction de la fréquence d'apparition de pile observée
  → **situation d'échantillonnage et de prise de décision**
- On fait un sondage sur un échantillon de population : 25 % se disent favorables à un projet
  Ce pourcentage est-il représentatif de l'opinion de toute la population ?
  → **situation d'estimation**

II Intervalle de fluctuation et prise de décision

1. Intervalle de fluctuation asymptotique au seuil de 95 %

Soit la $\boldsymbol{p}$ la proportion (ou fréquence) d'un caractère sur une population.
On s'intéresse à un échantillon de taille $\boldsymbol{n}$ de cette population.
On relève la fréquence observée f_{obs} de ce caractère sur cet «échantillon

On appelle intervalle de fluctuation asymptotique au seuil de 95 % l'intervalle :

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

Dans les conditions où $n \geq 30$, $np \geq 5$ et $n(1-p) \geq 5$ on démontre que $P(f_{obs} \in I_n) = 0{,}95$

2. Le protocole de prise de décision

On veut par exemple contrôler l'affirmation d'un fournisseur qui prétend que 2% seulement des pièces fabriquées dans son usine ne sont pas conformes.

- On formule l'hypothèse « la proportion du caractère dans la population est égale à p »
- On vérifie que les conditions d'utilisation de l'intervalle de fluctuation sont réunies
- On détermine l'intervalle I_n de fluctuation asymptotique au seuil de 95 %
- On calcule la fréquence observée du caractère sur l'échantillon
- On conclut : - si $f_{obs} \in I_n$ on accepte l'hypothèse
  - si $f_{obs} \notin I_n$ on rejette l'hypothèse, avec un risque de la rejeter à tort égal à 5%

III Estimation d'une proportion inconnue

Le problème :

Considérons une expérience aléatoire à deux issues possibles (succès ou échec) dont on ne connaît pas la probabilité $\boldsymbol{p}$ de succès.

On désire estimer au mieux $\boldsymbol{p}$ à partir de la réalisation de n expériences indépendantes.
En notant X_n le nombre de succès parmi ces n expériences, il est naturel de proposer comme estimateur de $\boldsymbol{p}$ la proportion de succès $F_n = \dfrac{X_n}{n}$

→ A quel point peut-on se fier à cette estimation ?
Peut-on évaluer une marge d'erreur ?

Théorème

Soit X_n une variable aléatoire suivant la loi binomiale $B\,(n\,;p)$ où p est la proportion inconnue d'apparition d'un caractère.
Soit $F_n = \dfrac{X_n}{n}$ la fréquence associée à la variable aléatoire X_n.
Alors, pour n suffisamment grand, l'intervalle $\left[F_n - \dfrac{1}{\sqrt{n}}\,;\,F_n + \dfrac{1}{\sqrt{n}}\right]$ contient la proportion p avec une probabilité au moins égale à 0,95.

Définition

On observe une réalisation f_{obs} de la variable aléatoire F_n en effectuant n expériences indépendantes.
L'intervalle $\left[f_{obs} - \dfrac{1}{\sqrt{n}}\,;\,f_{obs} + \dfrac{1}{\sqrt{n}}\right]$ est appelé intervalle de confiance de $\boldsymbol{p}$ au niveau de confiance 0,95.

- Chaque tirage d'un échantillon de taille n fournit un intervalle de confiance généralement différent.
- Si on faisait 100 échantillons issus de la même population, environ 95% des intervalles de confiance obtenus contiendraient p.
- En pratique, ayant observé une réalisation f_{obs} de la variable aléatoire $F_n = \dfrac{X_n}{n}$, l'intervalle $\left[f_{obs} - \dfrac{1}{\sqrt{n}}\,;\,f_{obs} + \dfrac{1}{\sqrt{n}}\right]$ n'est plus aléatoire : il contient ou non $\boldsymbol{p}$ mais il est faux de dire que la probabilité que $\boldsymbol{p}$ soit dans cet intervalle est supérieure ou égale à 0,95.

Remarque :

p étant inconnu, il n'est pas possible de vérifier si les conditions énoncées sur n et *p* en introduction de chapitre sont vérifiées.

Cependant, il faudra les vérifier sur la fréquence observée f_{obs}:
$n \geq 30$, $n \times f_{obs} \geq 5$ et $n \times (1 - f_{obs}) \geq 5$

Exemple :

On dispose d'une urne contenant un grand nombre de boules blanches et noires. La proportion de boules blanches contenues dans l'urne n'est pas connue.
On réalise un tirage de 100 boules et on obtient 54 boules blanches.
La fréquence observée est donc $f = 0{,}54$.
L'intervalle de confiance de la proportion de boule blanche dans l'urne au niveau de confiance 95% est
$\left[0{,}54 - \dfrac{1}{\sqrt{100}}\,;\,0{,}54 + \dfrac{1}{\sqrt{100}}\right] = [0{,}44\,;\,0{,}64]$.