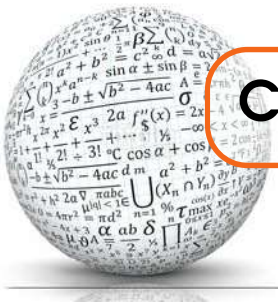

Chapitre 8 : Echantillonage et estimation

Savoirs faire à acquérir

- Connaître et maîtriser le vocabulaire du chapitre.
- Construire un intervalle de fluctuation en respectant les conditions.
- Prendre une décision.
- Construire un intervalle de confiance en respectant les conditions.
- Déterminer la taille d'un échantillon en connaissant la proportion effective.

1 Échantillon, simulation, fluctuation (rappels)

Définition 1 Une expérience ou situation est dite **aléatoire** si on ne peut pas en prévoir le résultat à l'avance.

Exemple 1

- Lancer un dé.
- Un sondage d'opinion avant une élection.
- etc.

■

Définition 2 Un **échantillon** de taille n est constitué des résultats de n répétitions indépendantes de la même expérience.

Exemple 2

- On lance une pièce 50 fois et on regarde si on obtient pile.
- On tire 20 fois une carte d'un jeu de 32 cartes, on regarde si cette carte est un coeur et on la replace dans le paquet.
- On interroge 1000 personnes et on leur demande si elles voteront.

■

Définition 3 Deux échantillons de même taille issus de la même expérience aléatoire ne sont généralement pas identiques.

On appelle **fluctuation d'échantillonnage**, les variations des fréquences des valeurs relevées.

Notations :

- n est le nombre d'éléments de l'échantillon. C'est **l'effectif** ou **la taille de l'échantillon**. On dit que l'échantillon est de taille n.
- f est la **fréquence** du caractère observé dans l'échantillon.
- p est la **proportion effective** du caractère observé dans la population.

R Plus la taille de l'échantillon est grand, plus les fréquences se rapprochent de p.

2 Intervalle de fluctuation (rappels)

Définition 4 L'intervalle de fluctuation au seuil de 95%, relatif aux échantillons de taille n, est l'intervalle centré autour de p qui contient la fréquence observée f dans un échantillon de taille n avec une probabilité égale à 0,95.

Propriété 1 — Vue en seconde. Soit p la proportion effective d'un caractère d'une population comprise entre 0,2 et 0,8 et f la fréquence du caractère observé dans un échantillon de taille n supérieur ou égale à 25. f appartient à l'intervalle $\left[p-\frac{1}{\sqrt{n}};p+\frac{1}{\sqrt{n}}\right]$ avec une probabilité d'environ 0,95.

Propriété 2 — Vue en première (avec loi binomiale). L'intervalle de fluctuation à 95 % d'une fréquence correspondant à la réalisation, sur un échantillon de taille n, d'une variable aléatoire X de loi binomiale $\mathscr{B}(n;p)$, est l'intervalle $\left[\frac{a}{n};\frac{b}{n}\right]$ avec :

1. a le plus petit entier tel que $P(X \leqslant a) > 2,5$
2. b le plus petit entier tel que $P(X \leqslant b) \geqslant 97,5$

3 Intervalle de fluctuation asymptotique

On considère X_n une variable aléatoire suivant une loi binomiale $\mathscr{B}(n;p)$, ainsi que la variable aléatoire $F_n = \frac{X_n}{n}$ représentant la fréquence de succès d'un schéma de Bernouilli de paramètres n et p.

On admet que pour un n suffisamment grand, X_n suit approximativement la loi normale de paramètres $\mu = np$ et $\sigma = \sqrt{np(1-p)}$.

On sait que depuis le chapitre 7 que $P(\mu - 1,96\sigma \leqslant X_n \leqslant \mu + 1,96\sigma) \approx 0,95$. Ce qui est équivalent à

$$P(X_n \in [\mu - 1,96\sigma;\mu + 1,96\sigma]) \approx 0,95$$

$$\Leftrightarrow P(X_n \in [np - 1,96\sqrt{np(1-p)};np + 1,96\sqrt{np(1-p)}]) \approx 0,95$$

$$\Leftrightarrow P(np - 1,96\sqrt{np(1-p)} \leqslant X_n \leqslant np + 1,96\sqrt{np(1-p)}) \approx 0,95$$

On divise par n :

$$\Leftrightarrow P\left(\frac{np}{n} - 1,96\frac{\sqrt{np(1-p)}}{n} \leqslant \frac{X_n}{n} \leqslant \frac{np}{n} + 1,96\frac{\sqrt{np(1-p)}}{n}\right) \approx 0,95$$

$$\Leftrightarrow P\left(\frac{np}{n} - 1,96\frac{\sqrt{n}\sqrt{p(1-p)}}{\sqrt{n}^2} \leqslant \frac{X_n}{n} \leqslant \frac{np}{n} + 1,96\frac{\sqrt{n}\sqrt{p(1-p)}}{\sqrt{n}^2}\right) \approx 0,95$$

$$\Leftrightarrow P\left(p - 1,96\frac{\sqrt{p(1-p)}}{\sqrt{n}} \leqslant F_n \leqslant p + 1,96\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right) \approx 0,95$$

$$\Leftrightarrow P\left(F_n \in \left[p - 1,96\frac{\sqrt{p(1-p)}}{\sqrt{n}};p + 1,96\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right]\right) \approx 0,95$$

Propriété 3 L'intervalle de fluctuation asymptotique au seuil de 95 % de la variable aléatoire F_n, telle que $n \geqslant 30$, $np \geqslant 5$ et $n(1-p) \geqslant 5$ est l'intervalle :

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

R Lien entre l'intervalle de fluctuation en seconde et l'intervalle de fluctuation asymptotique.
Pré-requis : On considère $f(p) = p(1-p) = p - p^2$ définie sur $[0;1]$.
f admet un maximum de $\frac{1}{4}$ atteint pour $p = \frac{1}{2}$.
Par conséquent : $p(1-p) \leqslant \frac{1}{4}$ et donc $\sqrt{p(1-p)} \leqslant \frac{1}{2}$.

$$\begin{aligned}
\sqrt{p(1-p)} &\leqslant \frac{1}{2} \\
\frac{\sqrt{p(1-p)}}{\sqrt{n}} &\leqslant \frac{1}{2} \times \frac{1}{\sqrt{n}} \\
1{,}96 \times \frac{\sqrt{p(1-p)}}{\sqrt{n}} &\leqslant 1{,}96 \times \frac{1}{2} \times \frac{1}{\sqrt{n}} \leqslant 2 \times \frac{1}{2} \times \frac{1}{\sqrt{n}} \\
1{,}96 \times \frac{\sqrt{p(1-p)}}{\sqrt{n}} &\leqslant 1{,}96 \times \frac{1}{2} \times \frac{1}{\sqrt{n}} \leqslant \frac{1}{\sqrt{n}}
\end{aligned}$$

Par conséquent : $1{,}96 \times \frac{\sqrt{p(1-p)}}{\sqrt{n}} \leqslant \frac{1}{\sqrt{n}}$.
Les intervalles $I = \left[p - 1{,}96\frac{\sqrt{p(1-p)}}{\sqrt{n}}; p + 1{,}96\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right]$ et $J = \left[p - \frac{1}{\sqrt{n}}; p + \frac{1}{\sqrt{n}}\right]$ ont le même centre p et l'amplitude de J est plus petite que l'amplitude de I car $1{,}96 \times \frac{\sqrt{p(1-p)}}{\sqrt{n}} \leqslant \frac{1}{\sqrt{n}}$.
Explications : Avec ces informations on peut conclure que $I \subset J$. L'intervalle de fluctuation vue en seconde est un tout petit plus grand que l'intervalle de fluctuation asymptotique.

4 Prise de décision

La proportion p est supposée. On considère que l'échantillon est représentatif de la population.

(On émet donc une conjecture sur la proportion p).

La question : Peut-on, à partir de l'observation d'un échantillon de fréquence f, valider la conjecture faite sur p ?

Réponse :

- Si $f \in \left[p - 1{,}96\frac{\sqrt{p(1-p)}}{\sqrt{n}}; p + 1{,}96\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right]$
  On n'a pas de raison de rejeter la proportion supposée/conjecturée p.
- Si $f \notin \left[p - 1{,}96\frac{\sqrt{p(1-p)}}{\sqrt{n}}; p + 1{,}96\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right]$
  On rejette la proportion supposée/conjecturée p avec un risque de 5% d'erreur.

Activité :

Dans une réserve indienne d'Aamjiwnaag, située au Canada, à proximité d'industries chimiques, il est né entre 1999 et 2003, 132 enfants dont 46 garçons. Est-ce normal ?

On fait l'hypothèse que la proportion p suivante : "Le sexe d'un enfant qui naît dans une cette réserve est un garçon avec une probabilité de 0,5"

Étape 1 : On vérifie les conditions.

- $n = 132\,;\, n \geqslant 30$.
- $np = 132 \times 0,5 = 66\,;\, np \geqslant 5$.
- $n(1-p) = 132 \times (1-0,5) = 132 \times 0,5 = 66\,;\, n(1-p) \geqslant 5$.

Les conditions sont vérifiées, on peut construire l'intervalle de fluctuation.

Étape 2 : On détermine l'intervalle de fluctuation.

$$\begin{aligned} I &= \left[p - 1,96\frac{\sqrt{p(1-p)}}{\sqrt{n}}; p + 1,96\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right] \\ &= \left[0,5 - 1,96\frac{\sqrt{0,5(1-0,5)}}{\sqrt{132}}; 0,5 + 1,96\frac{\sqrt{0,5(1-0,5)}}{\sqrt{132}}\right] \\ &\approx [0,415; 0,585] \end{aligned}$$

Étape 3 : Prise de décision.

$f = \dfrac{46}{132} \approx 0,348$. $f \notin I$, on rejette la proportion supposée avec un risque d'erreur 5%.

. . . . . .

La proportion p est connue.

La question : Peut-on, à partir de l'observation d'un échantillon de fréquence f, valider la véracité du sondage nous permettant d'obtenir la fréquence f ?

Réponse :

- Si $f \in \left[p - 1,96\dfrac{\sqrt{p(1-p)}}{\sqrt{n}}; p + 1,96\dfrac{\sqrt{p(1-p)}}{\sqrt{n}}\right]$
  On n'a pas de raison de rejeter la fréquence observée f.
- Si $f \notin \left[p - 1,96\dfrac{\sqrt{p(1-p)}}{\sqrt{n}}; p + 1,96\dfrac{\sqrt{p(1-p)}}{\sqrt{n}}\right]$
  On considère que l'échantillon observé ne représente pas la population, avec un risque de 5 % d'erreur.

Activité :

Le 4 mai 2007 soit deux jours avant le second tour des élections présidentielles, on publie le sondage suivant réalisé auprès de 992 personnes.

S. Royal 45% d'intention de vote.

N. Sarkozy 55% d'intention de vote.

Le 6 mai 2007 N.Sarkozy a gagné les élections avec 53,06% des voix et S.Royal 46,94 %. Le sondage était-il représentatif de l'élection ?

Étape 1 : On vérifie les conditions.

- $n = 992\,;\, n \geqslant 30$.
- $np = 992 \times 0,5306 = 526,3552\,;\, np \geqslant 5$.
- $n(1-p) = 992 \times (1-0,5306) = 992 \times 0,4694 = 465,648\,;\, n(1-p) \geqslant 5$.

Les conditions sont vérifiées, on peut construire l'intervalle de fluctuation.

Étape 2 : On détermine l'intervalle de fluctuation.

$$\begin{aligned} I &= \left[p - 1,96\frac{\sqrt{p(1-p)}}{\sqrt{n}}\,;p + 1,96\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right] \\ &= \left[0,5306 - 1,96\frac{\sqrt{0,5306(1-0,5306)}}{\sqrt{992}}\,;0,5306 + 1,96\frac{\sqrt{0,5306(1-0,5306)}}{\sqrt{922}}\right] \\ &\approx [0,4995;0,5617] \end{aligned}$$

Étape 3 : Prise de décision.

$f = 0,55$. $f \in I$, on admet que l'échantillon est représentatif de la population avec un risque d'erreur 5%.

. . . . . .

5 Intervalle de confiance (Estimation)

Nous allons maintenant travailler avec la fréquence d'un échantillon, afin d'estimer la proportion d'un caractère d'une population. C'est ce procédé que l'on utilise pour établir des sondages.

Propriété 4 Soit F_n, la variable aléatoire fréquence, qui à tout échantillon de taille n. Alors l'intervalle $\left[F_n - \frac{1}{\sqrt{n}}\,;F_n + \frac{1}{\sqrt{n}}\right]$, pour n suffisamment grand contient la proportion p avec une probabilité au moins égale à 0,95.

Définition 5 Soit f la fréquence observée d'un caractère dans un échantillon de taille n d'une population dans lequel ce caractère est représenté par une proportion p.

On appelle intervalle de confiance de la proportion p au seuil de 95%, l'intervalle : $\left[f - \frac{1}{\sqrt{n}}\,;f + \frac{1}{\sqrt{n}}\right]$, tel que : $n \geqslant 30$, $nf \geqslant 5$ et $n(1-f) \geqslant 5$.

Activité :

Le 4 mai 2007 soit deux jours avant le second tour des élections présidentielles, on publie le sondage suivant réalisé auprès de 992 personnes.

S. Royal 45% d'intention de vote.

N. Sarkozy 55% d'intention de vote.

Interpréter ce sondage.

Étape 1 : On vérifie les conditions.

- $n = 992\,;\, n \geqslant 30$.
- $nf = 992 \times 0,55 = 545,6\,;\, nf \geqslant 5$.
- $n(1-f) = 992 \times (1-0,55) = 992 \times 0,45 = 446,4\,;\, n(1-f) \geqslant 5$.

Les conditions sont vérifiées, on peut construire l'intervalle de confiance.

Étape 2 : On détermine l'intervalle de confiance.

$$\begin{aligned} I &= \left[f - \frac{1}{\sqrt{n}}; f + \frac{1}{\sqrt{n}}\right] \\ &= \left[0,55 - \frac{1}{\sqrt{992}}; 0,55 + \frac{1}{\sqrt{922}}\right]] \\ &= [0,51825; 0,58175] \end{aligned}$$

On estime que la proportion d'individus qui voteront pour Nicolas Sarkosy se situe entre 51,825% et 58,175%.

R L'intervalle $\left[f - 1,96\frac{\sqrt{f(1-f)}}{\sqrt{n}}; f + 1,96\frac{\sqrt{f(1-f)}}{\sqrt{n}}\right]$ est aussi un intervalle de confiance au seuil de 95%. (plus précis)

Définition 6 On appelle l'amplitude d'un intervalle $[a;b]$ la valeur $b-a$.

Propriété 5 L'amplitude de notre intervalle de confiance est de $\frac{2}{\sqrt{n}}$.

Démonstration.

$$\begin{aligned} \left(f + \frac{1}{\sqrt{n}}\right) - \left(f - \frac{1}{\sqrt{n}}\right) &= f + \frac{1}{\sqrt{n}} - f + \frac{1}{\sqrt{n}} \\ &= \frac{1}{\sqrt{n}} + \frac{1}{\sqrt{n}} \\ &= \frac{2}{\sqrt{n}} \end{aligned}$$

■ **Exemple 3** Un institut de sondage souhaite réaliser un sondage pour mesurer le niveau de satisfaction de la clientèle d'une société.

L'institut souhaite estimer la proportion des clients satisfaits à l'aide d'un intervalle de confiance au seuil de 95 % avec une amplitude de deux centièmes.

Combien de personnes l'institut de sondage doit-elle interroger ?

$$\begin{aligned} \frac{2}{\sqrt{n}} &= 0,02 \\ \frac{1}{\sqrt{n}} &= \frac{0,02}{2} \\ \frac{1}{\sqrt{n}} &= 0,01 \\ \sqrt{n} &= \frac{1}{0,01} \\ \sqrt{n} &= 100 \\ \sqrt{n}^2 &= 100^2 \\ n &= 10000 \end{aligned}$$

Pour que l'amplitude de l'intervalle de confiance soit égale à 0,02, il faut interroger 10 000 individus.

■

6 QCM Test

Entraine-toi pour le devoir ou pour simplement réviser avec le QCM en ligne en scannant le QR-code ou en cliquant tout simplement dessus ;)