Statistiques

1 Intervalle de fluctuation

Si la variable aléatoire X_n suit une loi binomiale $\mathscr{B}(n, p)$ et si l'on se trouve dans les conditions de l'approximation normale de la loi binomiale ($n \geqslant 30$, $np \geqslant 5$ et $n(1-p) \geqslant 5$), on définit alors l'**intervalle de fluctuation asymptotique au seuil de 95 %** par :

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

Cette intervalle peut éventuellement être simplifié par :

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

2 Prise de décision

Soit f_{obs} la fréquence d'un caractère observée d'un échantillon de taille n 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 $n(1-p) \geqslant 5$.

Hypothèse :
La proportion du caractère étudié dans la population est p.

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.

3 Estimation - Intervalle de confiance

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 échantilon de taille n. On calcule alors la fréquence f_{obs} des individus de cet échantillon ayant ce caractère.

On observe la fréquence f_{obs} sur un échantillon de taille n. On appelle **intervalle de confiance de 95%** l'intervalle :

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

Si l'on souhaitre encadrer p dans un intervalle de longueur a, on doit avoir : $n \geqslant \dfrac{4}{a^2}$

Explications

les formules sur l'intervalle de fluctuation ASYMPTOTIQUE d'un échantillon de taille n ***au seuil de 95%***

Rappel : *Formules donnant l'espérance et l'écart type d'une variable aléatoire* X

$$E(X) = \sum x_i \times P(X = x_i) \text{ ou } E(X) = \int_{-\infty}^{+\infty} t \times f(t)\, dt \quad \text{et} \quad \sigma_X = \sqrt{VAR(X)}$$

Avec $$Var(X) = E\left(X^2\right) - \left[E(X)\right]^2 = \left(\sum x_i^2 \times P(X = x_i)\right) - \left[E(X)\right]^2$$

ou avec $$Var(X) = E\left(X^2\right) - \left[E(X)\right]^2 = \int_{-\infty}^{+\infty} t^2 \times f(t)\, dt - \left[E(X)\right]^2$$

1) Si la variable X suit la loi binomiale $\mathcal{B}(n, p)$ alors $\begin{cases} E(X) = np \\ VAR(X) = np(1-p) \\ \sigma_X = \sqrt{np(1-p)} \end{cases}$

2) La variable $Y = \dfrac{X - E(X)}{\sigma_X}$ est une variable aléatoire centrée réduite $\begin{cases} E(X) = 0 \\ VAR(X) = 1 \\ \sigma_X = \sqrt{1} = 1 \end{cases}$

3) Si la variable X suit la loi binomiale $\mathcal{B}(n, p)$ alors

la v.a. $Y = \dfrac{X - np}{\sqrt{np(1-p)}}$ « converge » vers la loi Normale centrée réduite $\mathcal{N}(0, 1)$

4) La fréquence observée **dans un échantillon de taille n** peut être modélisée par la v.a. $F_n = \dfrac{X}{n}$

et $Y = \dfrac{n\left(\dfrac{X}{n} - p\right)}{\sqrt{np(1-p)}} = \dfrac{\dfrac{X}{n} - p}{\sqrt{\dfrac{p(1-p)}{n}}} = \dfrac{F_n - p}{\sqrt{\dfrac{p(1-p)}{n}}}$ « converge » vers la loi Normale centrée réduite $\mathcal{N}(0,1)$

Comme $P\left(\mathcal{N}(0,1) \in [-1{,}96\ ;\ 1{,}96]\right) \approx 0{,}95 \Rightarrow P\left(\dfrac{F_n - p}{\sqrt{\dfrac{p(1-p)}{n}}} \in [-1{,}96\ ;\ 1{,}96]\right) \approx 0{,}95$

5) On a $-1{,}96 \le \dfrac{F_n - p}{\sqrt{\dfrac{p(1-p)}{n}}} \le 1{,}96 \Leftrightarrow p - 1{,}96\sqrt{\dfrac{p(1-p)}{n}} \le F_n \le p + 1{,}96\sqrt{\dfrac{p(1-p)}{n}}$

6) On peut démontrer que pour tout $p \in]0\,;1[$ on a $p(1-p) \le \dfrac{1}{4}$ ***(voir annexe en fin de page 3)***

on a donc pour tout $p \in]0\,;1[$: $F_n \le p + 1{,}96\sqrt{\dfrac{p(1-p)}{n}} \Rightarrow$

$$F_n \le p + 1{,}96\sqrt{\frac{1}{4n}} = p + \frac{1{,}96}{2}\sqrt{\frac{1}{n}} \le p + \sqrt{\frac{1}{n}} \Rightarrow \boxed{I_n \in \left[p - \sqrt{\frac{1}{n}}\,;\, p + \sqrt{\frac{1}{n}}\right]}$$

Exemples

b) Prise de décision à partir d'un intervalle de fluctuation

PROPRIÉTÉ

Étant donné une population dans laquelle **on suppose que la proportion d'un certain caractère est** p. Si on prélève, avec remise, un échantillon de taille n dans cette population et si la **fréquence réelle observée** f du caractère dans cet échantillon **est comprise dans l'intervalle de fluctuation** alors on dit qu'**on accepte au seuil de 95% l'hypothèse que la proportion réelle du caractère dans la population est bien** p (dans le cas contraire, on dit qu'on rejette l'hypothèse).

► *Exemple :* Un candidat pense que 52% des électeurs lui sont favorables. On prélève avec remise un échantillon de 500 électeurs : 47% des électeurs interrogés de cet échantillon se déclarent favorable au candidat en question. L'intervalle de fluctuation de l'échantillon associé à la proportion de 52% est $[0{,}476\,;\,0{,}564]$ car $0{,}52-1{,}96\sqrt{\dfrac{0{,}52\times 0{,}48}{500}}\approx 0{,}476$ et $0{,}52+1{,}96\sqrt{\dfrac{0{,}52\times 0{,}48}{500}}\approx 0{,}564$.
0,47 étant en dehors de l'intervalle de fluctuation, on peut rejeter au seuil de 95% l'hypothèse du candidat selon laquelle 52% des électeurs lui sont favorables.

c) Estimation par un intervalle de confiance

PROPRIÉTÉ

On cherche à connaitre une estimation de la proportion p inconnue d'un certain caractère au sein d'une population. Pour cela, on prélève avec remise un échantillon de taille n au sein de la population et on note f la proportion observée du caractère au sein de l'échantillon. Il y alors 95% de chance (dans certaines conditions) que la proportion p du caractère au sein de la population totale soit comprise dans l'intervalle :

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

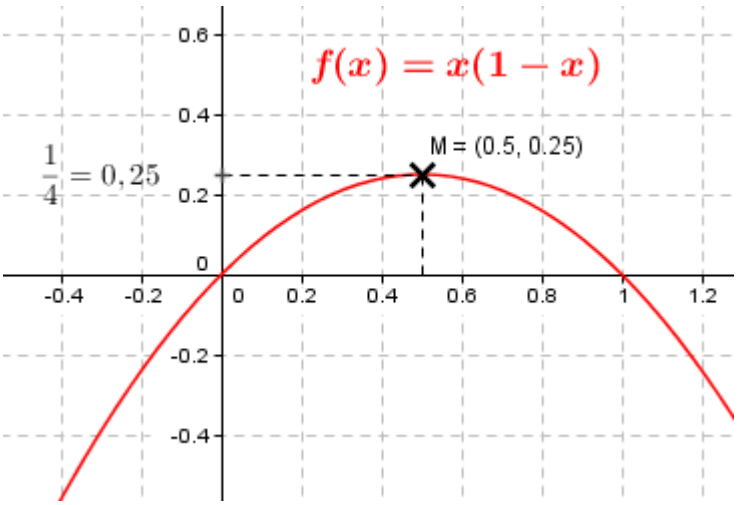
Cet intervalle est appelé **intervalle de confiance à 95%** associé à la proportion f.

► *Exemple :* Un sondage réalisé sur un échantillon de 1000 personnes attribue à un candidat un score de 18%. L'intervalle de confiance à 95% associé à cette proportion observée de 18% dans l'échantillon est [14,8% ; 21,2%] car $0{,}18-\dfrac{1}{\sqrt{1000}}\approx 0{,}148$ et $0{,}18+\dfrac{1}{\sqrt{1000}}\approx 0{,}212$.

Annexe

Montrons que pour tout $p\in\left]0\,;1\right[$ on a $p(1-p)\le\dfrac{1}{4}$

Une étude de la fonction $f(x)=x(1-x)$ sur $[\,0\,;1\,]$ montre que cette fonction admet un maximum pour $x=\dfrac{1}{2}$ égal à $f\left(\dfrac{1}{2}\right)=\dfrac{1}{2}\left(1-\dfrac{1}{2}\right)=\dfrac{1}{4}$



Savoir-faire 1 — Prendre une décision à partir d'un intervalle de fluctuation

ÉNONCÉ Pour leur mariage, Cécile et Pascal ont commandé à un traiteur des chouquettes au sucre et à la crème pâtissière. Le nombre de convives étant de 200, ils ont commandé 600 chouquettes de chaque sorte. Le matin du mariage, la mère de Cécile lui dit : « *Avec tes cousins, nous avons goûté quelques chouquettes. Vous auriez dû en commander plus à la crème pâtissière !* ». Après discussion avec ses cousins, Cécile apprend qu'ils ont goûté 40 chouquettes et seulement 15 étaient à la crème pâtissière.
La commande a-t-elle été respectée par le traiteur ?

SOLUTION

- On s'intéresse à la proportion de chouquettes à la crème pâtissière dans la commande réceptionnée le matin du mariage.
- Cécile et Pascal ont commandé 600 chouquettes de chaque sorte : la proportion de chouquettes à la crème pâtissière est alors **supposée être égale à** $p = 0{,}5$. Comme les cousins de Cécile ont goûté 40 chouquettes sur les 1 200 chouquettes commandées*, la taille de l'échantillon est $n = 40$. D'après eux, 15 d'entre elles étaient à la crème pâtissière :

$$f = \frac{15}{40} = \frac{3}{8} = 0{,}375$$

- Les paramètres n et p sont tels que :
  $n = 40 \geqslant 30$; $n \times p = 20 \geqslant 5$; $n \times (1-p) = 20 \geqslant 5$.

Les trois conditions sont bien respectées.
Par suite, l'intervalle de fluctuation asymptotique au seuil 0,95 est bien défini et il est donné par :

$$\left[p - 1{,}96 \times \frac{\sqrt{p(1-p)}}{\sqrt{n}} ; p + 1{,}96 \times \frac{\sqrt{p(1-p)}}{\sqrt{n}}\right] = \left[0{,}5 - \frac{0{,}98}{\sqrt{40}} ; 0{,}5 + \frac{0{,}98}{\sqrt{40}}\right] \approx \left[0{,}345 ; 0{,}655\right]$$

- La fréquence observée $f = 0{,}375$ appartient à l'intervalle de fluctuation asymptotique au seuil 0,95. Selon cet échantillon, « l'hypothèse faite sur p est acceptable » mais on ne connaît pas le risque de se tromper dans cette affirmation. La commande semble donc être respectée par le traiteur.

MÉTHODE

- On précise la population et le caractère étudié dans cette population.
- On identifie la proportion du caractère étudié dans la population (p), la taille de l'échantillon (n) et la fréquence observée du caractère étudié dans cet échantillon (f).
- Si les conditions sont vérifiées sur les paramètres n et p ($n \geqslant 30$, $n \times p \geqslant 5$ et $n \times (1-p) \geqslant 5$), on détermine l'intervalle de fluctuation asymptotique au seuil 0,95. Sinon on détermine l'intervalle étudié en classe de 2de ou éventuellement l'intervalle déterminé à l'aide de la loi binomiale et étudié en classe de 1re.
- On applique la règle de décision et on conclut.

* La taille de l'échantillon ($n = 40$) n'excède pas 10 % de la taille de la population (120) : l'échantillon considéré peut alors être assimilé à un échantillon aléatoire non exhaustif.

Savoir-faire 2 — Estimer une proportion inconnue par un intervalle de confiance

ÉNONCÉ Une semaine avant le second tour de l'élection municipale dans la commune de Buis-en-Provence, un sondage est effectué sur 1 024 personnes choisies au hasard parmi les 42 821 inscrites sur la liste électorale. 532 personnes interrogées ont déclaré qu'elles voteront pour le maire sortant Mr Ladent. En supposant que les votes seront conformes aux intentions de vote, le maire sortant a-t-il raison de penser qu'il sera élu au second tour ?

SOLUTION

- On s'intéresse à la proportion des électeurs qui voteront lors de cette élection municipale pour le maire sortant Mr Ladent. Cette proportion, généralement notée p, est **inconnue**.
- 1 024 personnes ont été interrogées : la taille de l'échantillon est $n = 1\,024$. Comme 532 voteront pour le maire sortant, la fréquence observée dans cet échantillon est $f = \frac{532}{1024} \approx 0{,}519\,5$.
- Les conditions étant vérifiées sur les paramètres n et f (la proportion inconnue p étant approchée par la fréquence observée f), on peut utiliser l'intervalle de confiance au niveau de confiance 0,95 qui est donné par :

$$\left[f - \frac{1}{\sqrt{n}} ; f + \frac{1}{\sqrt{n}}\right] = \left[\frac{532}{1\,024} - \frac{1}{\sqrt{1\,024}} ; \frac{532}{1\,024} + \frac{1}{\sqrt{1\,024}}\right] \approx \left[0{,}488\,3 ; 0{,}550\,8\right]$$

- Pour être réélu, la proportion inconnue p doit être strictement supérieure à 0,5. Or selon ce sondage, le maire sortant peut recueillir plus de 50 % des voix comme moins de 50 % des voix. Il est donc envisageable que le maire sortant ne soit pas réélu.

MÉTHODE

- On précise la population et le caractère étudié dans cette population.
- On identifie la taille de l'échantillon (n) et la fréquence observée (f) du caractère étudié dans cet échantillon.
- Si les conditions sont vérifiées sur les paramètres n et f ($n \geqslant 30$, $n \times f \geqslant 5$ et $n \times (1-f) \geqslant 5$), on détermine l'intervalle de confiance au niveau de confiance 0,95.
- On conclut.

Savoir-faire 3 — Déterminer une taille suffisante d'un échantillon pour estimer une proportion inconnue

ÉNONCÉ Une banque désire savoir si son site est bien adapté aux besoins de plus de 5 millions de clients. Elle commande alors à un institut de sondage une enquête afin d'estimer la proportion de ses clients satisfaits par les fonctions proposées par ce site. Elle impose un niveau de confiance de 0,95 avec une amplitude d'au plus quatre centièmes. Combien de personnes au minimum doit interroger l'institut de sondage ?

SOLUTION

MÉTHODE

- On précise la population et le caractère étudié dans cette population.
- On écrit l'intervalle de confiance au niveau de confiance 0,95 puis on précise l'amplitude de cet intervalle.
- On écrit la condition imposée sur l'amplitude puis on conclut.

- La population est constituée des clients de cette banque, soit plus de 5 millions de personnes. Le caractère étudié est la satisfaction du client aux fonctionnalités du site bancaire par rapport à ses besoins.
- La proportion des clients de cette banque satisfaits par les fonctionnalités offertes par ce site est inconnue et elle est notée p. **Une estimation** de cette proportion p peut être obtenue à l'aide de l'intervalle de confiance au niveau de confiance 0,95 (imposée par la banque) qui est défini par $\left[f-\frac{1}{\sqrt{n}}\,;\,f+\frac{1}{\sqrt{n}}\right]$ où f est la fréquence observée de clients de cette banque satisfaits par les fonctionnalités offertes par le site **sur un échantillon**. Cet intervalle permet d'en obtenir une estimation au niveau de confiance 0,95 avec une amplitude de $\left(f+\frac{1}{\sqrt{n}}\right)-\left(f-\frac{1}{\sqrt{n}}\right)=\frac{2}{\sqrt{n}}$.
- L'amplitude imposée par la banque implique que $\frac{2}{\sqrt{n}} \leqslant 0{,}04$. Par conséquent, $n \geqslant \frac{4}{0{,}04^2} = 2500$. L'institut de sondage doit interroger au minimum 2 500 personnes pour répondre aux exigences de la banque.

VOCABULAIRE

Dans ce chapitre, on s'intéresse à un caractère dans une population donnée dont la proportion est notée p. Cette **proportion** sera dans quelques cas **connue** (échantillonnage), dans certains cas **supposée connue** (prise de décision) et dans d'autres cas **inconnue** (estimation).
Pour des raisons généralement économiques, on étudie le caractère, non pas sur la population entière, mais sur des échantillons de taille n extraits de cette population. Pour ce faire, on peut **prélever au hasard** des individus de cette population un par un avec remise. On parle d'**échantillons aléatoires non exhaustifs**. Dans des situations telles qu'un sondage, un tel prélèvement est impensable : on pourrait interroger la même personne plusieurs fois. On prélève alors **successivement et sans remise** n individus de cette population. Si la taille de cet échantillon n'excède pas 10 % de la taille de la population entière, ce prélèvement ne modifie pas sensiblement la proportion du caractère dans la population. L'échantillon ainsi construit est assimilé à un échantillon aléatoire non exhaustif.
Tels sont les échantillons considérés dans ce chapitre.

COMMENT REDIGER

Sur 140 personnes interrogées lors du sondage, 99 se déclarent satisfaites.
Estimer par un intervalle de confiance au seuil 95% la proportion de personnes satisfaites parmi les utilisateurs de la crème.

Dans un échantillon de 140 personnes, 99 sont satisfaites. La fréquence observée est donc $f = \frac{99}{140}$.

On a $n = 144$, $f = \frac{99}{140} \approx 0{,}707$ alors :

$$\begin{cases} \checkmark & n = 140 \geq 30 \\ \checkmark & nf = 99 \geq 5 \\ \checkmark & n(1-f) = 41 \geq 5 \end{cases}$$

Un intervalle de confiance au seuil de 95% est alors :

$$I_n = \left[f-\frac{1}{\sqrt{n}}\,;\,f+\frac{1}{\sqrt{n}}\right] = \left[\frac{99}{140}-\frac{1}{\sqrt{140}}\,;\,\frac{99}{140}+\frac{1}{\sqrt{140}}\right]$$

soit puisque les borne sont :

- $\frac{99}{140}-\frac{1}{\sqrt{140}} \approx 0{,}622\,627$. On arrondit la borne inférieure par défaut au millième soit **0, 622**.
- $\frac{99}{140}+\frac{1}{\sqrt{140}} \approx 0{,}791\,658$. On arrondit la borne supérieure par excès au millième soit **0, 792**.

$$\boxed{I \approx [0{,}622\,;\ 0{,}792]}$$