

Option Statistique C4:

Estimation de la fréquence d'un caractère dans une population à partir d'un échantillon

On étudie dans ce cours une population donnée d'effectif N et un caractère que chaque individu de la population possède ou non. On désigne par p la fréquence de ce caractère, c'est-à-dire le nombre d'individus ayant ce caractère divisé par l'effectif total. Pour déterminer une telle proportion, on peut soit examiner l'ensemble de la population et dénombrer exactement les individus ayant le caractère (méthode exhaustive), soit estimer cette proportion à partir d'un échantillon extrait de la population (méthode de *sondage*). Nous allons voir que dans le dernier cas, le résultat n'est pas (et ne devrait jamais être) une valeur mais plutôt un intervalle, dit *intervalle de confiance*.

Le chapitre du cours de statistique qui traite de cette question s'appelle l'*estimation d'un paramètre* (ici une fréquence) par intervalle de confiance. Il appartient à la statistique *inférentielle* ou *inductive*.

L'estimation est utilisée dans beaucoup de domaines: en politique (sondages avant élection), en marketing (sondages auprès de consommateurs), en médecine (proportion d'individus porteurs d'une maladie, de malades guéris par un médicament), dans les sciences de l'environnement (proportion de plantes ou d'animaux victimes d'une pollution, proportion de vaches folles dans une région), en économie (proportion de ménages pratant en vacances), en sociologie ...

Exemple: Cet exemple est tiré du livre *Itinéraires en statistiques et probabilités*, H. Carnec, R. Seroux, J-M Dagoury et M. Thomas, Editions Ellipses, 2000.

Pour déterminer la proportion de ménages d'une ville donnée possédant au moins un téléviseur, on prélève *au hasard* un échantillon de 400 ménages et on constate que 304 d'entre eux ont un téléviseur, soit une proportion de $f = \frac{304}{400} = 0,76$. Le statisticien répondra alors que la proportion p exacte appartient, *au seuil de 5%*, à l'intervalle $[0,71 ; 0,81]$.

Nous allons expliquer ce que signifie ce seuil et comment est calculé cet intervalle.

1 Distribution d'échantillonnage

On modélise le caractère considéré par une variable aléatoire X de Bernouilli qui à chaque individu associe la valeur 1 s'il possède le caractère et la valeur 0 s'il ne le possède pas. Le nombre d'individus d'un échantillon de taille n possédant ce caractère est alors une variable aléatoire Binomiale, $Y = X_1 + X_2 + \dots + X_n$ prenant les valeurs $0, 1, 2, \dots, n$ pourvu que le choix de l'échantillon soit fait *au hasard* (comme un tirage au sort dans une urne, avec remise); dans les sondages réels, on utilise plutôt des *échantillons représentatifs* (méthode des quotas), ce que nous ne faisons pas ici. Nous considérons la suite des v.a. $(X_i)_{i=1,2,\dots,n}$ comme une *suite de v.a. indépendantes et de même loi*.

Si on considère différents échantillons de taille n ($n < N$) issus de la population, et qu'on associe à chacun d'eux la fréquence f du caractère pour l'échantillon, on obtient un ensemble de valeurs pour f que l'on peut voir comme des valeurs prises par la v.a. Y/n . Comme Y suit une loi Binomiale, on sait que pour tout $k \in \{0, 1, \dots, n\}$,

$$P\left(f = \frac{k}{n}\right) = \binom{n}{k} p^k (1-p)^{n-k}.$$

La v.a. prenant ces valeurs, avec ces probabilités s'appelle un *estimateur de la fréquence* p et elle est souvent notée \hat{p} . Comme Y suit une loi Binomiale $\mathcal{B}(n, p)$, on se souvient que $\mathbb{E}(Y) = np$ et $\text{Var}(Y) = np(1-p)$ (ou $\sigma(Y) = \sqrt{np(1-p)}$) et donc l'espérance et l'écart type de l'estimateur \hat{p} sont donnés par

$$\mathbb{E}(\hat{p}) = \mathbb{E}\left(\frac{Y}{n}\right) = \frac{np}{n} = p, \text{ et } \sigma(\hat{p}) = \sigma\left(\frac{Y}{n}\right) = \frac{\sqrt{np(1-p)}}{n} = \frac{\sqrt{p(1-p)}}{\sqrt{n}}.$$

2 Rappels sur la loi normale

Si l'on trace l'histogramme d'une loi binomiale $\mathcal{B}(n, p)$ lorsque n est suffisamment grand ($n > 30$), on s'aperçoit que les sommets des btons s'alignent approximativement sur une courbe en cloche appelée *gaussienne*. Le théorème de la limite centrale permet d'affirmer que :

Proposition 1 Si l'on prélève un échantillon aléatoire de taille n ($n > 30$) dans une population dans laquelle la fréquence d'un caractère donné est p alors la distribution d'échantillonnage suit approximativement une loi normale $\mathcal{N}(p, \frac{\sqrt{p(1-p)}}{\sqrt{n}})$.

Rappelons qu'une v.a. X suit une loi normale $\mathcal{N}(m, \sigma)$ si on a $P(X < x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{(t-m)^2}{2\sigma^2}} dt$ pour tout $x \in \mathbb{R}$. L'écart type σ est d'autant plus petit que les observations sont groupées autour de l'espérance m . On a :

$$\begin{aligned} P(m - 1,64\sigma < X < m + 1,64\sigma) &\simeq 0,9 \\ P(m - 1,96\sigma < X < m + 1,96\sigma) &\simeq 0,95 \\ P(m - 2,58\sigma < X < m + 2,58\sigma) &\simeq 0,99 \end{aligned}$$

En particulier, en assimilant 1,96 à 2, on peut affirmer que si x_0 est une valeur prise par une v.a. normale centrée (i.e. pour laquelle $m = 0$) et réduite (i.e. pour laquelle $\sigma = 1$), alors x_0 a 95% de chance d'appartenir à l'intervalle $[-2 ; 2]$ et seulement 5% de chance d'être à l'extérieur.

3 Intervalle de confiance

L'estimation d'une fréquence p à partir d'un échantillon fournit une valeur f (dite estimation ponctuelle) mais cette valeur n'est pas égale en général à la valeur exacte de p . Cependant si on considère que la valeur trouvée est *une* valeur de l'estimateur \hat{p} qui est une v.a. dont la loi (distribution d'échantillonnage) peut être assimilée à une loi normale $\mathcal{N}(p, \frac{\sqrt{p(1-p)}}{\sqrt{n}})$ alors on peut *affirmer* qu'il y a 95% de chance que la vraie valeur (inconnue) de p appartienne à l'intervalle

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

On appelle cet intervalle l'*intervalle de confiance* de l'estimation. On peut simplifier un peu son expression en remarquant que comme $p \in [0, 1]$, la quantité $p(1-p)$ est toujours au plus égale à 1/4. Donc $1,96 \frac{\sqrt{p(1-p)}}{\sqrt{n}}$ peut être majoré par $\frac{1}{\sqrt{n}}$, d'où l'intervalle de confiance $\left[f - \frac{1}{\sqrt{n}} ; f + \frac{1}{\sqrt{n}} \right]$.

Revenant à l'exemple indiqué dans l'introduction, l'échantillon de taille 400 a fourni une fréquence de 0,76. La vraie valeur appartient donc, avec un risque d'erreur n'excédant pas 5%, à l'intervalle $\left[0,76 - \frac{1}{\sqrt{400}} ; 0,76 + \frac{1}{\sqrt{400}} \right] = \left[0,76 - \frac{1}{20} ; 0,76 + \frac{1}{20} \right] = [0,71 ; 0,81]$.

4 Une mise en garde

L'exemple suivant a été indiqué par J-P Kahane. Il révèle que lorsqu'on donne l'estimation ponctuelle obtenue à partir d'un sondage en arrondissant au dixième le plus proche et non l'intervalle de confiance, il arrive qu'on ait bien plus de chance de se tromper que de donner la valeur correcte !

Avant une élection, on sonde 900 personnes issues d'une population dans laquelle $\frac{1}{5}$ des individus votent A et $\frac{4}{5}$ votent B. On décide de donner le résultat du sondage en arrondissant au dixième le plus proche : ainsi on annoncera 20% si la fréquence trouvée est comprise entre 19,5% et 20,5% et 21% si elle est comprise entre 20,5% et 21,5%. Quelle est la probabilité que le résultat annoncé soit exact ?

Cette probabilité égale à $P(0,195 < \hat{p} < 0,205)$ peut se calculer puisqu'on connaît la loi de \hat{p} . Ici $p = 0,2$ et $n = 900$ (et donc $\sqrt{n} = 30$). On assimile la loi de \hat{p} à $\mathcal{N}(0,2 ; \sigma)$ avec $\sigma = \frac{\sqrt{p(1-p)}}{\sqrt{n}} = \frac{\sqrt{0,16}}{30} \simeq \frac{4}{3}10^{-2}$. On en déduit que la probabilité cherchée s'écrit encore $P(0,195 < \hat{p} < 0,205) = P(p - \frac{1}{2}10^{-2} < \hat{p} < p + \frac{1}{2}10^{-2}) = P(p - \frac{3}{8}\sigma < \hat{p} < p + \frac{3}{8}\sigma) \leq \frac{1}{3}$. En conséquence, le score obtenu n'est juste qu'au plus une fois sur trois !