Cours 8 : 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 ce 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 éléction), 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 polution, proportion de vaches folles dans une région), en économie (proportion de ménages partant 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 de confiance $[0,71\ ; 0,81]$. Dans la suite de ce cours, 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 possèdant ce caractère est alors une variable aléatoire Binomiale, $Y = X_1 + X_2 + \ldots + X_n$ prenant les valeurs $0, 1, 2, \ldots, n$ pourvu que le choix de l'échantillon soit fait *au hasard* (c'est-à-dire comme un tirage au sort dans une urne avec remise) ; dans les sondages réels, on utilise plutôt des é*chantillons représentatifs* (méthode des quotas), ce que nous n'envisagerons pas ici. Nous considérons la suite des variables aléatoires (v.a.) $(X_i)_{i=1,2,..,n}$ comme une *suite de v.a. indépendantes et de même loi* (suite de variables iid). La loi commune des X_i est $\mathcal{B}(1,p)$ où p est la proportion que l'on cherche à déterminer et donc la loi de Y est $\mathcal{B}(n,p)$.

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 $k \in \{0, 1, \ldots, n\}$, $P\left(f = \frac{k}{n}\right) = \left(\begin{array}{c} n \\ k \end{array}\right) p^k (1-p)^{n-k}$. 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 son espérance vaut $\mathbb{E}(Y) = np$ et sa variance vaut $Var(Y) = np(1-p)$ (ou $\sigma(Y) = \sqrt{np(1-p)}$) et donc l'espérance et l'écart type de l'estimateur $\hat{p}$ sont donnés par

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

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 batons s'alignent approximativement sur une courbe en cloche appelée *gaussienne*. Le théorème de la limite centrale permet d'affirmer que 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'échantilonnage (ou loi de l'estimateur $\hat{p}$ de p) suit approximativement une loi

normale $\mathcal{N}\left(p, \frac{\sqrt{p(1-p)}}{\sqrt{n}}\right)$. Rappelons qu'une v.a. Z suit une loi normale $\mathcal{N}(m,\sigma)$ si on a $P(Z<x) = \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{x} e^{\frac{(t-m)^2}{2\sigma^2}}dt$ 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 par example :

$$P(m-1,64\sigma < Z < m+1,64\sigma) \simeq 0,9$$
$$P(m-1,96\sigma < Z < m+1,96\sigma) \simeq 0,95$$
$$P(m-2,58\sigma < Z < m+2,58\sigma) \simeq 0,99$$

En particulier, en majorant $1,96$ à 2, on peut affirmer que si x_0 est une valeur prise par une v.a. normale centrée (i.e. $m=0$) et réduite (i.e. $\sigma=1$), alors x_0 a moins de $\alpha = 5\%$ de risque d'être en dehors de l'intervalle $[-2\,;2]$. Dans le cas général $\mathcal{N}(m,\sigma)$ cet intervalle devient $[m-2\sigma, m+2\sigma]$.

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 $\mathcal{N}\left(p, \frac{\sqrt{p(1-p)}}{\sqrt{n}}\right)$ alors on peut *affirmer* qu'il y a $1-\alpha\%$ de chance que la vraie valeur (inconnue) de p appartienne à l'intervalle [1]

$$\left[f - q_{\frac{\alpha}{2}}\frac{\sqrt{p(1-p)}}{\sqrt{n}}\,;f + q_{\frac{\alpha}{2}}\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right] \tag{1}$$

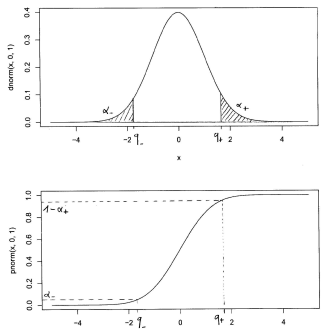
où $q_{\frac{\alpha}{2}}$ vaut par exemple $1,64$, $1,96$ et $2,58$ pour $\alpha = 10\%$, $\alpha = 5\%$ et $\alpha = 1\%$ respectivement. On appelle cet intervalle l'*intervalle de confiance symétrique* de l'estimation, *au seuil* α. [2]

Tout intervalle contenant l'intervalle de confiance au seuil α est aussi *un* l'intervalle de confiance au seuil α ; on peut simplifier un peu l'expression (1) en remarquant que $p\in[0,1]$, la quantité $p(1-p)$ est toujours au plus égale à $1/4$. Donc $\frac{\sqrt{p(1-p)}}{\sqrt{n}}$ peut être majoré par $0,5$, d'où l'intervalle de confiance $[f - q_{\frac{\alpha}{2}}\frac{1}{2\sqrt{n}}\,;f+q_{\frac{\alpha}{2}}\frac{1}{2\sqrt{n}}]$. Si $\alpha = 5\%$, comme $1,96<2$, on obtient l'intervalle de confiance $[f-\frac{1}{\sqrt{n}}\,;f+\frac{1}{\sqrt{n}}]$.

Plus généralement, on dit que l'intervalle $I=[q_-,q_+[$ est un *intervalle de confiance au seuil* α si

$$P(q_- \le Z < q_+) \ge 1-\alpha$$

Désignons par F_Z la fonction de répartition de Z, définie par $F_Z(q) = P(Z\le q)$. Si Z suit une loi $\mathcal{N}(\mu,\sigma)$ alors $p = F_Z(q) = \texttt{pnorm(q,}\mu\texttt{,}\sigma\texttt{)}$ et sa fonction réciproque est $q = F_Z^{-1}(p) = \texttt{qnorm(p,}\mu\texttt{,}\sigma\texttt{)}$. Si on découpe α en $\alpha = \alpha_- + \alpha_+$ (souvent on choisit $\alpha_- = \alpha_+ = \frac{\alpha}{2}$), il suffit de choisir $q_- \le F_Z^{-1}(\alpha_-)$ et $q_+ \ge F_Z^{-1}(1-\alpha_+)$. Si on choisit $\alpha_- = 0$ on obtient un intervalle illimité à gauche ($q_- = \mathrm{Inf}(Z)$ ou $-\infty$), et si on choisit $\alpha_+ = 0$ on obtient un intervalle illimité à droite ($q_+ = \mathrm{Sup}(Z)$ ou $+\infty$).



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 $[0,76 - \frac{1}{\sqrt{400}}\,;0,76+\frac{1}{\sqrt{400}}] = [0,76-\frac{1}{20}\,;0,76+\frac{1}{20}] = [0,71\,;0,81]$. D'où la réponse du statisticien.

1. ici choisi symétrique autour de f d'où la notation $q_{\frac{\alpha}{2}}$

2. Pour réduire la taille de cet intervalle, on peut soit accepter un risque d'erreur plus élevé (par exemple si le seuil est porté de 5% à 10%, le facteur $1,96$ peut être remplacé par $1,64$, mais on a alors une chance sur dix de se tromper au lieu de cinq chances sur cent), soit augmenter la taille de l'échantillon (c'est-à-dire n) mais bien souvent cela augmente alors le coût du sondage.