REVUE DE STATISTIQUE APPLIQUÉE

A. VESSEREAU

Une propriété peu connue : l'intervalle de confiance de la médiane

Revue de statistique appliquée, tome 35, n° 1 (1987), p. 5-7

<http://www.numdam.org/item?id=RSA_1987__35_1_5_0>

© Société française de statistique, 1987, tous droits réservés.

L'accès aux archives de la revue « Revue de statistique appliquée » (http://www.sfds.asso.fr/publicat/rsa.htm) implique l'accord avec les conditions générales d'utilisation (http://www.numdam.org/conditions). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques
http://www.numdam.org/

Rev. Statistique Appliquée, 1987, **XXXV** (1), 5-8

UNE PROPRIÉTÉ PEU CONNUE : L'INTERVALLE DE CONFIANCE DE LA MÉDIANE

A. VESSEREAU

Peu connue certes, mais pas inconnue : elle est proposée comme exercice dans les premières éditions de "M. G. KENDALL : The Advanced Theory of Statistics", elle est l'objet de tables dans "OWEN : Handbook of statistical Tables", et elle est sans doute mentionnée par d'autres auteurs.

Cette propriété fait partie des "Distribution Free Methods", que la terminologie anglo-saxonne distingue des "Non Parametric Methods". Distinction quelque peu subtile : les auteurs français adoptent généralement le terme global de « Méthodes non Paramétriques ». Dans le cas actuel la médiane est-elle un « paramètre » d'une loi de probabilité ? Certainement pas, si l'on entend par paramètre(s) un ou plusieurs nombres qui déterminent entièrement la loi.

La seule restriction signalée par KENDALL et OWEN est que la fonction de répartition $F(x)$ de la variable aléatoire X doit être continue : la médiane est la valeur $x = M$ définie par $F(M) = 0{,}5$. On a tenté d'élargir la définition à une variable discrète : ce serait la valeur M de la variable satisfaisant à la double inégalité $F(X \leqslant M) \leqslant 0{,}5$, $F(X \geqslant M) \geqslant 0{,}5$. Dans la représentation graphique de la fonction de répartition (courbe en escalier), M serait l'abscisse du point où la droite d'ordonnée 0,5 rencontre la courbe : la rencontre a lieu généralement sur une marche de l'escalier, exceptionnellement sur un palier, auquel cas la médiane serait indéterminée (loi binomiale $p = 0{,}5$, n pair par exemple). Cette définition peut conduire à des situations tout à fait paradoxales : pour la loi binomiale ($n = 20$, $p = 0{,}02$) par exemple, on aurait :

$$M = 0, \quad \Pr[X < M] = 0 \quad \Pr[X = M] = 0{,}668 \quad \Pr[X > M] = 0{,}332$$

Cette remarque de simple bon sens suffirait, s'il en était besoin, pour montrer qu'il est vain de chercher à définir un intervalle de confiance de la médiane (plus généralement de tout fractile) pour une variable autre que continue.

*

Les n valeurs x constituant l'échantillon issu de la loi étant rangées de façon non décroissante :

$$x_{(1)} \leqslant x_{(2)} \leqslant \ldots \leqslant x_{(i)} \leqslant \ldots \leqslant x_{(n)}$$

la probabilité que la valeur de rang k tombe dans l'intervalle (x, x + dx), donc que (k − 1) lui soient inférieures et (n − k) supérieures est :

$$\Pr[x_{(k)} < X < x_{(k)} + dx_{(k)}] = \frac{[F(x)]^{k-1}\,[1 - F(x)]^{n-k}\,dF(x)}{\int_0^1 [F(x)]^{k-1}\,[1 - F(x)]^{n-k}\,dF(x)}$$

La probabilité que cette $k^{ième}$ valeur soit au plus égale à une valeur x donnée est :

$$\Pr[x_{(k)} \leqslant x] = \frac{\int_0^{F(x)} F^{k-1}\,(1 - F)^{n-k}\,dF}{\int_0^1 F^{k-1}\,(1 - F)^{n-k}\,dF} \tag{1}$$

$$= I_{F(x)}(k, n - k + 1) \tag{2}$$

où I désigne la fonction bêta incomplète de paramètres (k, n − k + 1).

On a d'autre part entre la fonction bêta incomplète et la fonction cumulative de la loi binomiale (n, p) la relation :

$$I_p(k, n - k + 1) = 1 - \sum_{j=0}^{j=k-1} \binom{n}{j} p^j\,(1 - p)^{n-j} \tag{3}$$

Si, dans les relations (1) et (2), x = M est la médiane, F (x) = 0,5

$$\Pr[x_{(k)} \leqslant M] = I_{0,5}(k, n - k + 1) = 1 - \left(\frac{1}{2}\right)^n \sum_{j=0}^{j=k-1} \binom{n}{j} \tag{4}$$

Somme des k premiers termes de la loi binomiale $\left(n, p = \frac{1}{2}\right)$

La borne inférieure M_i de l'intervalle de confiance unilatéral « à droite » de M, au niveau de confiance (1 − α), correspond donc au plus grand rang k_i tel que :

$$\left(\frac{1}{2}\right)^n \sum_{j=0}^{j=k_i-1} \binom{n}{j} \leqslant \alpha \tag{5}$$

expression qui définit $k_i - 1$, donc k_i, d'où l'on déduit $M_i = x_{(k_i)}$.

Le rang k_i peut s'obtenir, pour toutes valeurs de n et α, au moyen des probabilités cumulées de la loi binomiale (n, p = 1/2) dont il existe des tables jusqu'à n = 100, ou, pour les valeurs pas trop élevées de n, à l'aide d'une calculatrice de bureau. On obtient en même temps la valeur du risque réel $\alpha' \leqslant \alpha$. Pour les mêmes conditions (n, α), et par raison de symétrie évidente, la borne supérieure M_s de l'intervalle de confiance unilatéral « à gauche » correspond au rang $k_s = n + 1 - k_i$. Enfin pour un intervalle bilatéral, on obtient k_i (d'où M_i) en remplaçant α par α/2 dans la relation (5), puis k_s (d'où M_s) par $k_s = n + 1 - k_i$.

Pour n ⩾ 50, on a une bonne approximation de la loi binomiale (n, p = 1/2) par la loi normale de paramètres μ = 1/2, $\sigma = \sqrt{n}/2$. Compte tenu d'une correction de continuité, k_i et (ou) k_s sont les valeurs arrondies à l'entier

immédiatement inférieur (supérieur) de $1/2\,[n + 1 \pm u\sqrt{n}]$, où u représente, suivant le cas, le fractile d'ordre $1 - \alpha/2$ ou $1 - \alpha$ de la variable normale réduite.

Exemple

n = 14. Aux niveaux de confiance 0,95 — 0,99 — 0,999 — les rangs k_i, k_s et les risques réels α' déterminés au moyen d'une table de la loi binomiale sont les suivants :

Niveau de confiance	Intervalles unilatéraux	Intervalle bilatéral
0,95 ($\alpha = 0{,}05$)	$k_i = 4$ ($k_s = 11$) $\alpha' = 0{,}029$	$k_i = 3$, $k_s = 12$ $\alpha' = 0{,}013$
0,99 ($\alpha = 0{,}01$)	$k_i = 3$ ($k_s = 12$) $\alpha' = 0{,}006$	$k_i = 2$, $k_s = 13$ $\alpha' = 0{,}0018$
0,999 ($\alpha = 0{,}001$)	$k_i = 2$ ($k_s = 13$) $\alpha' = 0{,}0009$	$k_i = 1$, $k_s = 14$ $\alpha' = 0{,}00012$ (les bornes coincident avec les valeurs extrêmes observées)

Généralisation

La méthode se généralise de façon évidente à la détermination de l'intervalle de confiance d'un fractile d'ordre quelconque (relation (2) où F (x) correspondra au fractile considéré). Pour les 1er et 3ème quantiles par exemple, les intervalles de confiance sont déterminés par les probabilités cumulées des lois binomiales [n, p = 1/4] et [n, p = 3/4]; les rangs correspondant à k_i et (ou) k_s sont donnés dans "OWEN : Handbook of Statistical Tables" pour $\alpha = 0{,}05$ et $\alpha = 0{,}01$.