Canad. Math. Bull. Vol. **14** (2), 1971

ESTIMATEURS DE LA MOYENNE POUR LES DISTRIBUTIONS SYMETRIQUES A DOMAINE BORNE

PAR
J. P. DION

1. **Introduction**. L'emploi de $\bar{X}$ comme estimateur de la moyenne est très fréquent en statistiques pour plusieurs bonnes raisons. Ainsi pour une densité régulière de la classe exponentielle, il est l'estimateur sans biais pour la moyenne, à variance minimale. Mais il perd cette importante propriété pour la classe des densités tronquées étudiées ici: cet article démontre que lorsque la taille de l'échantillon est suffisamment grande, il faut préférer $\Delta = [(\min_{i=1,\ldots n} X_i + \max_{i=1,\ldots n} X_i)/2]$ à $\bar{X}$.

Soit θ fixé, réel, et considérons la famille des densités $f(x, \theta)$ symétriques par rapport à θ et à domaine borné: $[\theta - c, \theta + c]$, c réel positif. Supposons aussi que $f(x, \theta)$ soit non décroissante pour x inférieur ou égal à θ. Nous trouvons alors une borne supérieure B à la variance de Δ et déterminons la taille d'échantillon $N(c)$ suffisante pour que la variance de Δ soit inférieure à la variance de $\bar{X}$. Il ressort que si la troncation est sévère ou la taille grande, Δ constitue un meilleur estimateur sans biais pour la moyenne que $\bar{X}$.

Enfin suit une discussion de cette borne B et une comparaison avec des valeurs "expérimentales" $\mathcal{N}(c)$ obtenues par simulation sur ordinateur, dans le cas de la densité normale tronquée.

2. **Résultats et preuves**. Soit donc $\{X_1, X_2, \ldots, X_n\}$ un échantillon d'une population de distribution F et $X_n^{(1)} < \cdots < X_n^{(n)}$ les statistiques d'ordre correspondantes. La distribution F possède une densité f telle que:

(i) $f(x) = f(-x)$, $\forall x$ réel,
(ii) $f(x) = 0$ sii $|x| > c$, $c \in R^+$, et
(iii) $xf'(x) \leq 0$, $\forall x$ réel.

THÉORÈME. *Pour toute distribution F avec les propriétés* (i) *à* (iii), *on a*

$$\operatorname{Var} \Delta \leq \frac{[1 + (\frac{1}{4})(f(0) - f(c))/f(0)]}{2(n+1)(n+2)f^2(c)} \leq 5/(8(n+1)(n+2)f^2(c)),$$

où

$$\Delta = [(X_n^{(1)} + X_n^{(n)})/2]$$

COROLLAIRE. *Soit un tel F, c son point de troncation et* $\sigma^2 = \operatorname{Var}(X)$. *Alors* $\operatorname{Var} \Delta < \operatorname{Var} \bar{X}$ *pour tout échantillon de taille* $n > N(c) = 5/(8\sigma^2 f^2(c))$.

Preuve du théorème. A cause de la symétrie, $E(X_n^{(1)}) = -E(X_n^{(n)})$ et $E(X_n^{(1)^2}) = E(X_n^{(n)^2})$. Faisant le changement de variables: $M_n = (c - X_n^{(n)}) \cdot n$ et $m_n = (c + X_n^{(1)}) \cdot n$, on trouve $\operatorname{Var} \Delta = [(E(M_n^2) - E(m_n \cdot M_n))/2n^2]$.

Nous allons montrer que

(a) $$E(M_n^2) \leq 2n^2/[(n+1)(n+2)f^2(c)]$$

et

(b) $$E(M_n \cdot m_n) \geq E(M_n^2)/2 - [(f(0)-f(c))n^2/(4f^2(c)\cdot f(0)\cdot(n+1)(n+2))]$$

(A) $E(M_n^2) \equiv E(V^2) = 2\int_0^{2cn} v[F(c-v/n)]^n\,dv$, après une intégration par partie. D'où

$$\begin{aligned} E(V^2) &\leq (2/f(c))\int_0^{2cn} v[F(c-v/n)]^n f(c-v/n)\,dv \\ &= (2n/((n+1)f(c)))\int_0^{2cn} [F(c-v/n)]^{n+1}\,dv \\ &\leq (2n/((n+1)f^2(c)))\int_0^{2cn} [F(c-v/n)]^{n+1} f(c-v/n)\,dv \\ &= 2n^2/((n+1)(n+2)f^2(c)). \end{aligned}$$

(B) $E(M_n \cdot m_n) \equiv E(U\cdot V) = \int_0^{2cn} vf(c-v/n)\int_0^{2cn-v}[F(c-v/n)-F(u/n-c)]^{n-1}\,du\,dv$ après une intégration par partie. On peut écrire $E(U\cdot V) = I_1 + I_2$ où

$$I_1 = \int_0^{cn} vf(c-v/n)\int_0^{2cn-v}[F(c-v/n)-F(u/m-c)]^{n-1}\,du\,dv$$

et

$$I_2 = \int_{cn}^{2cn} vf(c-v/n)\int_0^{2cn-v}[F(c-v/n)-F(u/n-c)]^{n-1}\,du\,dv.$$

Comme $v \in (cn, 2cn)$ implique que $f(c-v/n) > f(u/n-c)$,

$$I_2 \geq \int_{cn}^{2cn} v\int_0^{2cn-v}[F(c-v/n)-F(u/n-c)]^{n-1} f\left(\frac{u}{n}-c\right)\,du\,dv,$$

i.e. $I_2 \geq \int_{cn}^{2cn} v[F(c-v/n)]^n\,dv$.

D'autre part,

$$I_1 = \int_0^{cn} vf(c-v/n)\left\{\int_0^v + \int_v^{2cn-v}\right\}[F(c-v/n)-F(u/n-c)]^{n-1}\,du\,dv,$$

i.e. $I_1 = I_{11} + I_{12}$. Et $v \in (0, cn) \Rightarrow f(c-v/n) > f(u/n-c)$, $\forall u < v$, de sorte que

$$\begin{aligned} I_{11} &\geq \int_0^{cn}\int_0^v [F(c-v/n)-F(u/n-c)]^{n-1} f(u/n-c)\,du\,dv \\ &= \int_0^{cn}[F(c-v/n)]^n\,dv - 2^n\int_0^{cn}[F(c-v/n)-\tfrac{1}{2}]^n\,dv. \end{aligned}$$

Et

$$\begin{aligned} I_{12} &\geq \frac{2^n}{f(0)}\int_0^{cn} vf(c-v/n)[F(c-v/n)-\tfrac{1}{2}]^n\,dv \\ &\geq (n\cdot 2^n/((n+1)f(0)))\int_0^{cn}[F(c-v/n)-\tfrac{1}{2}]^{n+1}\,dv. \end{aligned}$$

https://doi.org/10.4153/CMB-1971-044-1 Published online by Cambridge University Press

Et en regroupant

$$E(M_n \cdot m_n) \geq E(M_n^2)/2 - \frac{n \cdot 2^n(1-f(c)/f(0))}{(n+1)f(c)} \int_0^{cn} [F(c-v/n) - \tfrac{1}{2}]^{n+1}\, dv,$$

i.e.

$$E(M_n \cdot m_n) \geq (E(M_n^2)/2) - n^2(1-f(c)/f(0))/(4(n+1)(n+2)f^2(c)).$$

Q.E.D.

3. **Discussion.** Ainsi une condition suffisante pour que Δ soit un meilleur estimateur que $\bar{X}$ est que la taille de l'échantillon soit supérieure à $N(c) = 5/(8\sigma^2 f^2(c))$.

Quant à la borne $B = \left[1 + (\tfrac{1}{4})\left(\frac{f(0)-f(c)}{f(0)}\right)\right]/(2(n+1)(n+2)f^2(c))$ elle est d'autant plus proche de Var Δ que $f(0)$ l'est de $f(c)$. Aussi pour le cas uniforme la borne est atteinte, i.e. Var $\Delta \equiv B$. D'autre part elle est une bonne approximation de la variance dans le cas de grands échantillons, puisque $\lim_{n\to\infty} n^2 \operatorname{Var} \Delta = 1/(2f^2(c))$ et $\lim_{n\to\infty} n^2 B \approx 5/(8f^2(c))$.

Enfin dans un cas particulier: la distribution normale tronquée, une simulation sur ordinateur a permis de calculer pour différentes valeurs de n l'estimateur $S^2(\Delta)$ de la variance de Δ. Nous avons ainsi déterminé "expérimentalement" la taille $\mathcal{N}(c)$ suffisante pour que $S^2(\Delta)$ soit inférieure à Var $\bar{X}$. La densité normale tronquée est définie par:

$$f(x) = 0 \quad \text{si} \quad |x| > c$$

$$f(x) = [e^{-x^2/2}/k_c\sqrt{2\pi}] \quad \text{si} \quad |x| \leq c, \text{ où}$$

$$k_c = \int_{-c}^{c} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt.$$

Le tableau suivant donne pour cette densité les valeurs respectives de $\mathcal{N}(c)$ et $N(c)$, en fonction de quelques valeurs de c.

Densité normale tronquée

c	1.0	1.5	2.0	2.5
$\mathcal{N}(c)$	6	18	130	1600
$N(c)$	18	53	272	2315

Les résultats obtenus par simulation sont nettement meilleurs et ils justifient une préférence pour l'estimateur Δ pour des tailles même inférieures à $N(c)$. En général on peut utiliser la borne B comme approximation de Var Δ mais pour chaque cas particulier (sauf le cas uniforme) une simulation donnerait une meilleure approximation de la variance.

REMERCIEMENTS. Je voudrais remercier le professeur C. Kraft du département de mathématiques de l'Université de Montréal pour m'avoir suggéré le problème. Monsieur N. Buckle, du département d'informatique, a fait la programmation pour la simulation, je l'en revercie vivement. Je remercie également le département de mathématiques pour l'aide financière reçue durant ce travail.

BIBLIOGRAPHIE

1. Hajek et Sidak, *Theory of rank tests*, Academic Press, New York, 1967.

2. B. Gnedenko, *Sur la distribution limite du terme maximum d'une série aléatoire*, Ann. of Math. **44** (3), 1943.

3. M. Fisz, *Probability theory and mathematical statistics*, Wiley, New York. 1963.

UNIVERSITÉ DE MONTRÉAL,
MONTRÉAL, QUÉBEC