ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 152, 2000

II. STATISTICAL INFERENCE

*Janusz Wywial**

ESTIMATION OF MODE ON THE BASIS OF A TRUNCATED SAMPLE

Abstract. The problem of estimation of the mode of a continuous distribution function is considered. The estimation of the mode based on estimators of the density function is well known, see e.g. Härdle (1991) and Parzen (1962). New parameters of the continuous distribution function will be defined: the quasi-mode and mean median. They are parameters of the appropriately truncated random variable. Next, the estimators of the mode, such as the sample quasi-mode or sample mean-median, are determined. These statistics are usually biased estimators of the mode. Well known "jackknife" procedure is adapted to estimate their mean square error. The accuracy of the mode estimation is studied on the basis of computer simulation.

1. THE BASIC DEFINITIONS AND NOTATION

The mode, the median and the mean of a continuous distribution function are denoted by d, m and μ, respectively. The third central moment is denoted by η_3. Let A_+ be the set of such continuous and one-modal density functions that the inequalities $d \leqslant m \leqslant \mu$ and $\eta_3 \geqslant 0$ are fulfilled. The set A_+ can be treated as the class of right skewed continuous distribution functions. Similarly, let A_- be such a class of continuous one-modal density functions that $d \geqslant m \geqslant \mu$ and $\eta_3 \leqslant 0$. Hence, the set A_- will be called the class of the left skewed distribution functions.

Let us define the following moments of the truncated distribution:

$$\mu(a_t,\ b_t) = \frac{1}{F(b_t) - F(a_t)} \int_{a_t}^{b_t} xf(x)dx \tag{1}$$

where: $t = 0,\ 1,\ 2,\ \ldots,\ a_0 = a,\ b_0 = b,\ F(b) = 1,\ F(a) = 0$ and

* Academy of Economics, Chair of Econometrics, Katowice.

$$F(u) = \int_{-\infty}^{u} f(x)dx.$$

Moreover, let us assume:

$$\mu(a_{t+1},\ b_{t+1}) = \frac{1}{F(\mu(a_t,\ b_t) - F(a_t)} \int_{a_t}^{\mu(a_t,\ b_t)} xf(x)dx, \qquad \text{if } \eta_3(a_t,\ b_t) > 0 \quad (2)$$

where: $a_{t+1} = a_t,\ b_{t+1} = \mu(a_t,\ b_t)$ and

$$\eta_3(a_t,\ b_t) = \frac{1}{F(b_t) - F(a_t)} \int_{a_t}^{b_t} (x - \mu(a_t,\ b_t))^3 dx \quad (3)$$

$$\mu(a_{t+1},\ b_{t+1}) = \frac{1}{F(\mu(b_t) - F(\mu(a_t,\ b_t)} \int_{\mu(a_t,\ b_t)}^{b_t} xf(x)dx, \quad \text{if } \eta_3(a_t,\ b_t) < 0 \quad (4)$$

where: $a_{t+1} = \mu(a_t, b_t)$ and $b_{t+1} = b_t$.

Definition 1. Let $f(x)$ be the density function in the interval $I = \langle a, b \rangle$ and let a_t and b_t be such truncation points that $a_t,\ b_t \in I$ and

$$g_1(a_t,\ b_t) = \mu(a_t, b_t), \quad \text{if} \quad \eta_3(a_t, b_t) = 0,\ t = 0,\ 1,\ 2,\ \ldots \quad (5)$$

The parameter $g_1 = g_1(a_t, b_t)$ will be called the quasi-mode of the continuous distribution function.

Hence, the quasi-mode is the mean value of a distribution function appropriately truncated in such a way that its third central moment takes the value zero or the left and right truncated points are equal to each other.

The median $m(a_t, b_t)$ of the truncated distribution in the points $a_t < b_t$ is determined in the following way:

$$\frac{1}{F(b_{t-1}) - F(a_{t-1})} \int_{a_{t-1}}^{m(a_t, b_t)} f(x)dx\, f(x)dx = F_t(m(a_t,\ b_t)) = \frac{1}{2} \quad (6)$$

Let the expected value of the truncated distribution be defined as follows:

$$\mu(a_{t+1}, b_{t+1}) = \frac{1}{F(m(a_t, b_t)) - F(a_t)} \int_{a_t}^{m(a_t, b_t)} xf(x)dx, \quad \text{if } m(a_t, b_t) < \mu(a_t, b_t) \quad (7)$$

where: $a_{t+1} = a_t,\ b_{t+1} = m(a_t, b_t)$,

$$\mu(a_{t+1}, b_{t+1}) = \frac{1}{F(\mu(b_t) - F(m(a_t, b_t))} \int_{m(a_t, b_t)}^{b_t} xf(x)dx, \quad \text{if } m(a_t, b_t) < \mu(a_t, b_t) \quad (8)$$

where: $a_{t+1} = m(a_t, b_t)$ and $b_{t+1} = b_t$.

Definition 2. Let $f(x)$ be the density function in the interval $I = \langle a, b\rangle$ and let a_t and b_t be such truncation points that $a_t, b_t \in I$ and

$$g_t(a_t, b_t) = \mu(a_t, b_t), \quad \text{if} \quad \mu(a_t, b_t) = m(a_t, b_t), \quad t = 0, 1, 2, \ldots \tag{9}$$

The parameter $g_2 = g_2(a_t, b_t)$ will be called the mean-median of the distribution function.

The estimators of the parameters $d(u)$ and $m(u)$ are going to be defined in the next paragraphs. They can be used to estimate the mode of continuous distribution functions.

2. ESTIMATION OF THE MODE BY MEANS OF THE AVERAGE FROM A TRUNCATED SAMPLE

Let $X_{(1)} \leqslant X_{(2)} \leqslant \ldots \leqslant X_{(n)}$ be the sequence of the order statistics. The third central moment from the truncated simple sample is defined by the expression:

$$C_3(a, b) = \frac{1}{k} \sum_{i=a}^{i} (X_{(i)} - \overline{X}(a, b))^3, \quad 1 < k = b - a + 1 \tag{10}$$

where:

$$\overline{X}(a, b) = \frac{1}{k} \sum_{i=a}^{b} X_{(i)}.$$

Let us define the following statistics:

$$C_3(a_{t+1}, b_{t+1}) =$$

$$= \begin{cases} C_3(a_t, b_t - h), & \text{if } C_3(a_t, b_t) > 0 \text{ and } X_{(b_t - h)} \leqslant X(a_t, b_t) < X_{b_t - h + 1)}, \\ C_3(a_t, b_t), & \text{if } C_3(a_t, b_t) = 0, \\ C_3(a_t + h, b_t), & \text{if } C_3(a_t, b_t) < 0 \text{ and } X_{(b_t + h - 1)} \leqslant X(a_t, b_t) < X_{(a_t - h)}. \end{cases} \tag{11}$$

Then, the sample quasi-mode G_1 is defined as follows:

$$G_t = \overline{X}(a_t, b_t), \text{ if } C_3(a_t, b_t) = 0 \neq C_3(a_{t-1}, b_{t-1}) \tag{12}$$

The statistic G_1 will be usually a biased estimator of the mode d. Its bias can be reduced by means of the well-known jackknife method. Let $G_1^{(i)}$, $i = 1, \ldots, n$, be the quasi-mode from the sample without the i-th observation. The pseudovalues are determined by the expression:

$$Z_i = nG_1 - (n-1)G_1^{(i)}.$$

Hence, the jackknife type estimator of the mode d is as follows:

$$G_{1J} = \overline{Z} = \frac{1}{n}\sum_{i=1}^{n} Z_i \tag{13}$$

The estimators of the mean-square error are defined by the following expressions:

$$S^2(G_1) = \frac{1}{n(n-1)}\sum_{i=1}^{n}(Z_i - G_1)^2 \tag{14}$$

$$S^2(G_{1J}) = \frac{1}{n(n-1)}\sum_{i=1}^{n}(Z_i - G_{1J})^2 \tag{15}$$

3. ESTIMATION OF THE MODE BY MEANS OF THE MEAN-MEDIAN FROM THE TRUNCATED SAMPLE

Let us define the median from a truncated sample in the following way:

$$M(a_t, b_t) = X_{(e)}, \quad \text{where} \quad e = \left[\frac{b_t - a_t + 1}{2}\right] + 1 \tag{16}$$

$$M(a_{t+1}, b_{t+1}) =$$

$$= \begin{cases} M(a_t, b_t - h), & \text{if } M(a_t, b_t) < X(a_t, b_t) \text{ and } X_{(b_t-h)} \leqslant M(a_t, b_t) < X_{(b_t-h+1)}, \\ M(a_t, b_t), & \text{if } M(a_t, b_t) = X(a_t, b_t), \\ M(a_t + h, b_t), & \text{if } M(a_t, b_t) > X(a_t, b_t) \text{ and } X_{(a_t+h)} \geqslant M(a_t, b_t) > X_{(a_t+h-1)}. \end{cases} \tag{17}$$

The following estimators of the mean-median or the mode of a skewed distribution can be defined. The sample mean-median G_2 of the skewed distribution are determined by the expression:

$$G_2(a_t, b_t) = \overline{X}(a_t, b_t), \text{ if } M(a_t, b_t) = X(a_t, b_t), \quad t = 0, 1, 2, \ldots \tag{18}$$

Then, the sample median $M(a_t, b_t)$ is the median of the truncated sample determined by the following sequence of the order statistics: $X_{(a_t)}, \ldots, X_{(b_t)}$ and the sample mean-median $G_2(a_t, b_t)$ is the average of these statistics.

Usually, both statistics G_1 and G_2 are biased estimators of the mode d of the continuous distribution function. Just like in the case of the statistics G_1, we can try to reduce this bias by means of the jackknife method.

4. ESTIMATION OF A MODE OF A MULTIDIMENSIONAL DISTRIBUTION

Let $f(x_1, \ldots, x_r)$ be a density function of an r-dimensional random variable. The central moments of the order 3 of the r-dimensional variable are denoted by:

$$\eta_{uw}(X_i, X_j) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x_i - E(X_i))^u (x_j - E(X_j))^v f(x_1, \ldots, x_r) dx_1, \ldots, dx_r \tag{19}$$

and the sample moments:

$$C_{uv}(X_i, X_j) = \frac{1}{n} \sum_{t=1}^{n} (X_{it} - \overline{X}_i)^u (X_{jt} - \overline{X}_j)^v, \quad \overline{X}_i = \frac{1}{n} \sum_{t=1}^{n} X_{it} \tag{20}$$

where $u + v = 3$ and $u, v = 0, 1, 2, 3$. Let us introduce the following vectors: $\Theta = [\theta_1, \ldots, \theta_w] = [|\eta_{30}, X_1, X_2)| \; |\eta_{03}(X_1, X_2)| \; \ldots \; |\eta_{12}(X_{r-1}, X_r)| \; |\eta_{21}(X_r, X_{r-1})|]$ and $\mathbf{L} = [l_1, \ldots, l_w] = [|C_{30}(X_1, X_2)| \; |C_{03}(X_2, X_1)| \ldots C_{12}(X_{r-1}, X_r)| \; |C_{21}(X_r, X_{r-1})|]$, where $w = r^2$.

It is well known that if a distribution function is symmetric, then all central moments of the order 3 of the marginal one or two dimensional distributions are equal to zero. Hence: $\Theta = \mathbf{0}$.

In order to simplify our analysis, let us consider a two-dimensional random variable. The data observed in the simple sample will be denoted by $\{(x_i, y_i)\}$, $i = 1, \ldots, n$. From the geometrical point of view (x_i, y_i) are the coordinates of a point $A_i (i = 1, \ldots, n)$ in a two dimensional plane. Let $P_1, P_2, \ldots, P_h$, $h < n$, be points selected from the set $A = \{A_i\}$ in such a way that they are apexes of a convex polygon and the points of the set A are inside this polygon. The set $P = \{P_1, P_2, \ldots, P_h\}$ determines the polygon P. Hence the edge of the polygon P can be treated as a convex envelope of the points $\{A_i\}$.

Let us construct the polygons: $P = P^{(0)} \supseteq P^{(1)} \supseteq P^{(2)} \ldots \supseteq P^{(1)}$. The polygon $P^{(t)}$ is obtained trough rejecting one apex $P_q \in P^{(t-1)}$, $t = 1, 2, \ldots$ The central moments of the data creating the coordinates of points belonging to the polygon $P^{(t)}$ are as follows:

$$C_{uv}(X_i, X_j | B(t, z)) = \frac{1}{n - t} \sum_{\{e:\, A_e \in \beta(t, z_t)\}} (X_{ie} - \overline{X}_i^{(t)})^u (X_{je} - \overline{X}_j^{(t)})^v \tag{21}$$

where:

$$\overline{X}_i^{(t)} = \frac{1}{n-t} \sum_{(e:\, A_e \in B(t,\, z))} X_{ie},$$

$$B(t, z) = P^{(t-1)} - P_z \quad \text{and} \quad P_z \in P^{(t)}, \quad t = 1, 2, \ldots \text{ and } B(0, z) = P.$$

Moreover:

$$L(t, z) = [l_1(t, z) \ldots l_w(t, z)] = [|C_{30}(X_1, X_2|B(t, z))| \ldots |C_{21}(X_{r-1}, X_r|B(t, z))|]$$

The polygon $P^{(t)}$ is determined through dropping the point P_q from the polygon $P^{(t-1)}$ in such a way that

$$l(t, q) = \min_{P_z \in P^{(t-1)}} \max_{j=1,\ldots,w} \{l_j(t, z)\} \tag{22}$$

This procedure leads to truncation of the two dimensional sample. This algorithm is stopped if $l(t, q) = 0$ or $t = n - 2$ and $\mathbf{G} = (G_i, G_j) = (\overline{X}_i^{(t)}, \overline{X}_j^{(t)})$ are estimators of dominants (d_i, d_j) of two-dimensional random variables (X_i, X_j). Hence:

$$\mathbf{G} = (\overline{X}_i^{(t)}, \overline{X}_j^{(t)}) \quad \text{if} \quad l(t, q) = 0 \quad \text{and} \quad l(t-1, q) > 0 \tag{23}$$

The presented method of estimation of the dominants of a two dimensional variable can be easily generalized on a case of distribution of more than two variables. W y w i a ł (1998) considered a one-dimensional case of this method including simulation study of estimation precision.

5. SIMULATION ANALYSIS

In order to study the basic properties of the introduced estimators of the mode, a simulation analysis is developed. The estimation of the mode of the following triangular distribution is considered[1]:

[1] The basic properties of the triangular distribution can be found in H e l l w i g (1995).

$$f(x) = \begin{cases} 0 & \text{for} \quad x < 0 \vee x > b > 2a, \\ \dfrac{2(x+b-2a)}{b^2-2a^2} & \text{for} \quad 0 < x < a, \\ \dfrac{2x}{b^2-2a^2} + \dfrac{2b}{b^2-2a^2} & \text{for} \quad a < x < b. \end{cases} \tag{24}$$

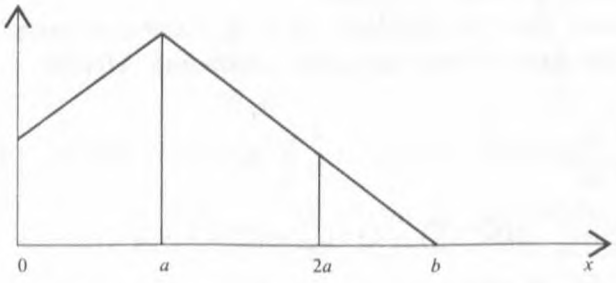

Fig. 1. The density function of the triangular distribution

The distribution function is as follows:

$$F(x) = \begin{cases} 0 & \text{for} \quad x < 0, \\ \dfrac{x^2}{b^2-2a^2} + \dfrac{2(b-2a)}{b^2-2a^2} & \text{for} \quad 0 < x < a, \\ 1 - \dfrac{(x-b)^2}{b^2-2a^2} & \text{for} \quad a < x < b, \\ 0 & \text{for} \quad x > b. \end{cases} \tag{25}$$

The function inverse to the distribution function is as follows:

$$F^{-1}(y) = \begin{cases} 2 - b + \sqrt{(b-2)^2 + (b^2-2)y} & \text{for} \quad 0 \leqslant y \leqslant y_0, \\ b - \sqrt{(b^2-2)(1-y)} & \text{for} \quad y_0 < y \leqslant 1, \end{cases} \tag{26}$$

where:

$$y_0 = F(a) = \frac{(2b-3a)a}{b^2-2a^2}.$$

The moments are as follows:

$$E(X) = \frac{b^3-2a^3}{3(b^2-2a^2)}, \quad E(X^2) = \frac{b^4-2a^4}{6(b^2-2a^2)}, \quad E(X^3) = \frac{b^5-2a^5}{10(b^2-2a^2)}.$$

In the simulation experiment we assumed that $a = 1$ and $b = 5$. Hence, the mode $d = 1.0$, $E(X) = 1.7826$, $D^2(X) = 1.3368$, $\eta_3(X) = 0.7646$ and the skewness coefficient $\beta_1 = 0.4947$.

The pseudo-random values of the triangular distribution are generated. Values of the estimators are determined on the basis of 2000 samples of a fixed size. The simulation experiment was developed on the basis of the well known SPSS statistical package.

Let us assume that the expected value E(.) and the variance $D^2(.)$ are estimated on the basis of the computer simulation. Hence:

$$\mathrm{E}(G_k) = \frac{1}{N}\sum_{\{xi\}} g_k(\{x_i\}); \quad \mathrm{D}^2(G_k) = \frac{1}{N}\sum_{\{xi\}}(g_k(\{x_i\}) - E(G_k))^2, \quad k = 1, 2,$$

$$\mathrm{MSE}(G_k) = \mathrm{D}^2(G_k) + [\mathrm{E}(G_k) - d]^2,$$

$$e(G_k) = (E(S(G_{kf}))/\sqrt{\mathrm{MSE}(G_k)} - 1, \quad b(G_k) = 100\%(\mathrm{E}(G_k) - d)/\mathrm{MSE}(G_k),$$

where: $k = 1, 2$, $g_k(x_i)$ is the value of the estimator determined on the basis of the simple sample $\{x_i\}$ of size n and the number of such samples is denoted by N. The mean square error of the estimator G_k is denoted by $\mathrm{MSE}(G_k)$. The results of the simulation study of the quasi-mode distribution are shown in the Tab. 1 and 2.

Table 1

The simulation results for the distribution of the estimator G_1 in the case of the left side truncated triangular distribution for the parameters $a = 1$ and $b = 5$

n	$E(G_1)$	$D^2(G_1)$	$\sqrt{\mathrm{MSE}(G_1)}$	$b(G_1)\%$	$e(G_{1f})\%$
10	1.591	1.146	1.289	21.0	113.3
20	1.373	0.983	1.051	12.6	246.4
30	1.296	0.857	0.907	10.7	218.6
50	1.235	0.673	0.713	10.9	295.1

Source: the author's own elaboration.

Table 2

The simulation results for the distribution of the estimator G_2 in the case of the left side truncated triangular distribution for the parameters $a = 1$ and $b = 5$

n	$E(G_2)$	$D^2(G_2)$	$\sqrt{\text{MSE}(G_2)}$	$b(G_2)\%$	$e(G_{2f})\%$
10	1.950	0.943	1.339	50.4	−12.9
20	1.669	0.928	1.144	34.2	20.5
30	1.516	0.826	0.974	28.1	54.6
50	1.134	0.655	0.669	4.0	126.3
100	1.147	0.516	0.537	7.5	2.778

Source: the author's own elaboration.

The analysis of the Tab. 1 and 2 lead to the following conclusions: The bias of both estimators are rather large. The bias of the estimator G_1 is larger then the bias of the G_2 only for the size of the sample $n = 50$. Similarly, the relative efficiency $b(G_1)$ is larger than $b(G_2)$ for the standard deviations, and the mean square errors of both estimators decrease when the sample size increases. The relative biases $e(G_f)$ and $e(G_2)$ increase when the sample size increases. Especially, the bias of the estimator $S(G_1)$ is too large. In conclusion, neither statistic $S(G_1)$ nor $S(G_2)$ are useful as estimators of the mean square errors $\sqrt{\text{MSE}(G_1)}$ and $\sqrt{\text{MSE}(G_2)}$, respectively. Generally however, the estimator G_2 is slightly better than G_1

REFERENCES

Härdle W. (1991), *Smoothing Techniques*, Springer Verlag, New York–Berlin–Heidelberg–London–Paris–Tokyo–Hong Kong–Barcelona.

Hellwig Z. (1995), *Elementy rachunku prawdopodobieństwa i statystyki matematycznej*, PWN, Warszawa.

Parzen E. (1962), *On Estimation of a Probability Density and Mode*, "Annals of Mathematical Statistics", **35**, 1005–1076.

Wywiłł J. (1998), *Estimation of Distribution Function Mode on the Basis of Sample Moment or Sample Median*, Submitted to Badania Operacyjne i Decyzje.