Theory of Stochastic Processes
Vol. 14 (30), no. 1, 2008, pp. 1–6

UDC 519.21
ROMAN I. ANDRUSHKIW, DMITRY D. KLYUSHIN, AND YURIY I. PETUNIN

A NEW TEST FOR UNIMODALITY

A distribution function (d.f.) of a random variable is unimodal if there exists a number such that d.f. is convex left from this number and is concave right from this number. This number is called a mode of d.f. Since one may have more than one mode, a mode is not necessarily unique. The purpose of this paper is to construct nonparametric tests for the unimodality of d.f. based on a sample obtained from the general population of values of the random variable by simple sampling. The tests proposed are significance tests such that the unimodality of d.f. can be guaranteed with some probability (confidence level).

1. Introduction

Testing the unimodality of a distribution function is a widely investigated issue. The most popular tests include the DIP test proposed by J.A.Hartigan and the kernel density estimation test proposed by B.W.Silverman [1–4]. However, all of these tests are computationally quite complex and asymptotic. That is why it is useful to develop elementary tests which are based on simple computational procedures and are non-asymptotic.

According to A.Ya.Khinchin, a distribution function $F(u)$ of a random variable x is unimodal if there exists a number M such that d.f. $F(u)$ is convex in $(-\infty, M)$ and concave in (M, ∞). The number M is said to be a mode of d.f. $F(u)$. A mode can not be unique, since d.f. $F(u)$ can have several modes. Also, d.f. $F(u)$ can have break at M and be continuous in $(-\infty, M)$ and (M, ∞).

The purpose of the paper is to construct nonparametric tests for the unimodality of d.f. $F(u)$ based on a sample $x_1, x_2, \ldots, x_n$ obtained by the simple sampling from the general population of values of a random variable x. The tests proposed are significance tests, so the unimodality of d.f. $F(u)$ can be guaranteed with some probability α (confidence level), where $\beta = 1 - \alpha$ is the significance level of a test. To formulate the tests, we introduce new estimations of the probability density (d.p.) and the distribution function based on a sample $x_1, x_2, \ldots, x_n$.

2. Unified histogram and modified empirical d.f.

Let $x_1, x_2, \ldots, x_n$. be a sample obtained from a general population $F(u)$ by the simple sampling which has d.p. $F(u)$. Since these functions are unknown, we call them hypothetical. To estimate d.p., we use the relation

$$p\left(x_{n+1} \in \left(x_{(i)}, x_{(i+1)}\right)\right) = \frac{1}{n+1}, \tag{1}$$

where $x_{(i)}$ is the order statistics ($i = 1, 2, ..., n$). Using estimation (1), we can define the estimation $h_n(u)$ for a hypothetical d.p.

$$h_n(u) = \begin{cases} \frac{1}{(n-1)\left(x_{(i+1)} - x_{(i)}\right)}, & \text{if } u \in (x_i, x_{i+1}), \\ 0, & \text{otherwise.} \end{cases}$$

2000 *AMS Mathematics Subject Classification.* Primary 62G05.
Key words and phrases. Unimodality, distribution function, significance test.

In such a case, the probability that the value of a random variable $\tilde{x}$ with d.p. (2) belongs to $\left[x_{(i)}, x_{(i+1)}\right)$ is equal to

$$p\left(\tilde{x} \in \left(x_{(i)}, x_{(i+1)}\right)\right) = \frac{1}{n+1},$$

where x_i are considered as constants. For large n, this probability is close to probability (1), so we refer to the value $h_n(u)$) as a unified histogram constructed on $x_1, x_2, \ldots, x_n$.. This histogram has some advantage over all other histograms, because it is unambiguously defined by the sample $x_1, x_2, \ldots, x_n$.. Also, the integral

$$\tilde{F}_n^*(u) = \int_{x_{(1)}}^{u} h_n(v)\, dv = \frac{u + (i-1)\, x_{(i+1)} - i x_{(i)}}{(n-1)\left(x_{(i+1)} - x_{(i)}\right)} \tag{2}$$

is a linear spline $x_{(i)} \le v < x_{(i+1)}$ which is a more precise estimation of the hypothetical d.f $F(u)$ than that of a piecewise empirical d.f.

$$F_n^*(u) = \frac{i}{n},\ if\ x_{(i)} \le v < x_{(i+1)}.$$

We refer to the function $F_n^*(u)$ as e.d.f. and to $\tilde{F}_n^*(u)$ defined by (2) as a modified e.d.f. (m.e.d.f.). Its advantages over the conventional e.d.f. are obvious: 1) when d.f. $F(u)$ is continuous, linear splines are more precise approximations than the piecewise e.d.f $F_n^*(u)$, and 2) $\tilde{F}_n^*(u)$ is continuous, so it is possible to estimate quantiles of any order and to construct an inverse d.f. (Quetelet curve), whereas it is impossible to do by using the piecewise e.d.f. $F_n^*(u)$. However, at large n, e.d.f. $F_n^*(u)$ and $\tilde{F}_n^*(u)$ are close. Let us prove that

$$\left|F_n^*(u) - \tilde{F}_n^*(u)\right| \le \frac{1}{n}. \tag{3}$$

Indeed, for all $u \in [x_i, x_{i+1})$, the following relation holds:

$$\left|F_n^*(u) - \tilde{F}_n^*(u)\right| = \left|\frac{u + (i-1)\, x_{(i+1)} - i x_{(i)}}{n(n-1)\left(x_{(i+1)} - x_{(i)}\right)} - \frac{i}{n}\right| = \left|\frac{nu - (n-i)\, x_{(i+1)} + i x_{(i)}}{n(n-1)\left(x_{(i+1)} - x_{(i)}\right)}\right|.$$

Granting that

$$\tilde{F}_n^*(u) = \frac{u - x_{(i+1)} + i\left(x_{(i+1)} - x_{(i)}\right)}{(n-1)\left(x_{(i+1)} - x_{(i)}\right)} = \frac{1}{n-1}\left[\frac{u - x_{(i+1)}}{x_{(i+1)} - x_{(i+1)}} + 1\right],$$

we have

$$\frac{i-1}{n-1} \le \tilde{F}_n^*(u) \le \frac{i}{n-1}$$

and

$$\frac{i-1}{n-1} - F_n^*(u) \le \tilde{F}_n^*(u) - F_n^*(u) \le \frac{i}{n-1} - F_n^*(u),$$

$$\frac{n(i-1) - (n-1)\, i}{(n-1)\, i} \le \tilde{F}_n^*(u) - F_n^*(u) \le \frac{ni - (n-1)\, i}{n(n-1)},$$

$$\frac{i-n}{n(n-1)} \le \tilde{F}_n^*(u) - F_n^*(u) \le \frac{1}{n};$$

hence,

$$-\frac{1}{n} \le \tilde{F}_n^*(u) - F_n^*(u) \le \frac{1}{n},$$

i.e.

$$\left|\tilde{F}_n^*(u) - F_n^*(u)\right| \le \frac{1}{n}. \tag{4}$$

Estimation (4) implies that m.e.d.f has similar asymptotic properties as conventional e.d.f., i.e. it is consistent, asymptotically unbiased, etc.

3. Confidence limits for hypothetical d.f.

Let us define the lower and upper bounds of a hypothetical d.f.$F(u)$ by means of an empirical d.f., under the assumption that $F(u)$ is continuous and strictly increasing. This problem was solved in [5–10]. Hence, given a significance level β^* (e.g., $\beta^* = 0.05$), we can define ε so that

$$p\left(\Delta = \max_{x_{(1)} \le u \le x_{(n)}} |F(u) - F_n(u)| > \varepsilon\right) = \beta^*.$$

It follows that, for a given β^*, we can find ε according to statistical tables [11] and construct a strip Π_{β^*}, whose bounds are stepwise linear: $y = F_n^*(u)+\varepsilon$ and $y = F_n^*(u)-\varepsilon$. The strip Π_{β^*} completely covers the true d.f. $y = F(u)$ with the confidence probability $\alpha^* = 1 - \beta^*$. Hereinafter, we refer to the strip Π_{β^*} as the confidence strip for d.f. with significance level β^* constructed for the empirical d.f.

4. Test for unimodality based on e.d.f.

Let $x_1, x_2, \ldots, x_n$ be a sample obtained from a general population G by the simple sampling with continuous and strictly monotone d.f. $F(u)$. Using this sample, we construct the empirical d.f. $F_n^*(u)$ and the strip Π_{β^*}. Denote, by $\varphi(u)$, the upper bound of Π_{β^*} described by the equation $y = F_n^*(u) + \varepsilon$, and let $\psi(u)$ be the lower bound of Π_{β^*} described by the equation $y = F_n^*(u) - \varepsilon$. Then

$$p\left(\varphi(u) \le F(u) \le \psi(u)\right) = \alpha^* = 1 - \beta^*.$$

Definition 1. *Let $y=\varphi(u)$ be an arbitrary function defined on $[a, b]$. Then the set*

$$G_U = \{(u, y): y \ge \varphi(u), a \le u \le b\}$$

is an epigraph of $\varphi(u)$, and the set $G_L = \{(u, y) : y \le \varphi(u), a \le u \le b\}$ is a subgraph of $\varphi(u)$.

Definition 2. *The lower bound of a convex hull of the epigraph of a function $\varphi(u)$ is a convex minorant of $\varphi(u)$,*

$$\varphi_{\inf}(u) = \inf\left\{v : (u, v) \in \operatorname*{conv}_{a \le u \le b} G_U\right\},$$

where conv G_U *is the convex hull of G_U. Analogously, the upper bound of a convex hull of the subgraph of a function $\varphi(u)$ is a concave majorant of $\varphi(u)$:*

$$\psi_{\sup}(u) = \sup\left\{v : (u, v) \in \operatorname*{conv}_{a \le u \le b} G_L\right\}.$$

Theorem 1. *Let $\varphi_{\inf}(u)$ and $\psi_{\sup}(u)$ be the convex minorant and concave majorant of $\varphi(u)$ and $\psi(u)$, respectively, and*

$$c = \sup\left\{u : \varphi_{\inf}(u) \le \psi(u),\ x_{(1)} \le u \le x_{(n)}\right\},$$

$$d = \inf\left\{u : \psi_{\sup}(u) \ge \varphi(u),\ x_{(1)} \le u \le x_{(n)}\right\}.$$

Then, the hypothetical distribution $F(u)$ is unimodal iff

1) $\varphi_{\inf}(u) \ge \psi(u)$ or $\psi_{\sup}(u) \le \varphi(u)$ $\forall u \in \left[x_{(1)}, x_{(n)}\right]$;

or

2) $c \ge d$.

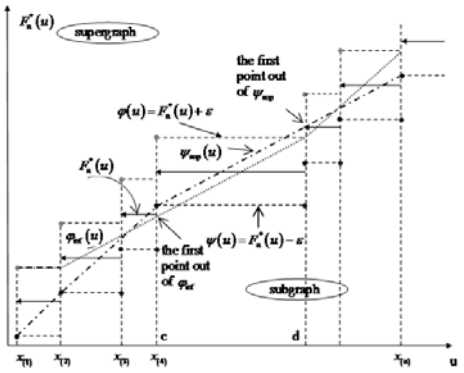

Fig. 1. If $c < d$, the unimodality is absent.

Moreover, the significance level of this criterion is β^*.

Proof. Necessity. Suppose that the hypothetical d.f. $F(u)$ is unimodal, and M is its mode. If $M \le x_{(1)}$ or $M \ge x_{(n)}$, then $F(u)$ on $\left[x_{(1)}, x_{(n)}\right]$ can be convex or concave. Then condition 1) holds.

If $x_{(1)} \ge M \le x_{(n)}$, then $F(u)$ is convex on $\left[x_{(1)}, M\right]$ and concave on $\left[M, x_{(n)}\right]$. In such a case, it follows from Definition 2 (see Fig. 1) that $\varphi_{\inf}(u) \ge F(u)$ on $\left[x_{(1)}, M\right]$. Also, $F(u) \ge \psi_{\inf}(u) \forall u \in \left[x_{(1)}, M\right]$, so $\varphi_{\inf}(u) \ge \psi(u)$, and $d \ge M$.

On the other hand, $F(u) \ge \psi_{\sup}(u)$, as far as $F(u)$ is concave on $\left[M, x_{(n)}\right]$. Definition 2 implies that $F(u) \ge \psi_{\sup}(u)$ on $\left[M, x_{(n)}\right]$. Also, $\varphi(u) \ge F(u) \forall u \in \left[M, x_{(n)}\right]$. Thus, $\varphi(u) \ge \psi_{sup}(u) \forall u \in \left[M, x_{(n)}\right]$ and $d \le M$. Consequently, $c \ge d$, and condition 2) holds.

Sufficiency. Note that $\varphi(u)$ and $\psi(u)$ are increasing. If condition 1) holds, then

$$\psi(u) \le \varphi_{sup}(u) \le \varphi(u) \ \forall u \in \left[x_{(1)}, x_{(n)}\right]$$

or

$$\psi(u) \le \psi_{inf}(u) \le \varphi(u) \ \forall u \in \left[x_{(1)}, x_{(n)}\right].$$

Thus, $\varphi_{sup}(u)$ (or $\psi_{inf}(u)$) lies in the strip Π_β. Therefore, $\varphi_{sup}(u)$ (or $\psi_{inf}(u)$) can be used as an estimation of the hypothetical d.f. $F(u)$ of a general population G. Since $F(u) = \varphi_{inf}(u)$ or $F(u) = \psi_{inf}(u)$, the hypothetical d.f. increases, is convex or concave on $\left[x_{(1)}, x_{(n)}\right]$, and is unimodal. The significance level of this test is β^*.

Now, we suppose that condition 2) holds, i.e. $c \ge d$. Put $\hat{F}(u) = \varphi_{inf}(u)$, if $u \in \left[x_{(1)}, c\right]$, and $\hat{F}(u) = \psi_{sup}(u)$, if $u \in \left(c, x_{(n)}\right]$. It is easy to see that $\hat{F}(u)$ lies in Π_β, because $c \ge d$. Also, $\hat{F}(u)$ is convex on $\left[x_{(1)}, c\right]$ and concave on $\left(c, x_{(n)}\right]$. Let us prove that $\hat{F}(u) \ge \hat{F}(c+0) = lim_{u \to c, u > c} = \gamma$. Indeed, if $\gamma < \hat{F}(c)$, then $\gamma \notin \Pi_\beta$. Therefore, the abscissa of the first exit point d, where $\varphi_{inf}(u)$ exceeds the bounds Π_β while moving from $x_{(n)}$ to $x_{(1)}$, is greater than c. This contradicts condition 2. Thus, $\hat{F}(u)$ increases, is convex on $\left[x_{(1)}, c\right]$, and concave on $\left(c, x_{(n)}\right]$. But $\hat{F}(u)$ can have a breakpoint in c. To exclude this breakpoint, we take $\varepsilon > 0$ sufficiently small so that the segment with the ends $\left(c - \varepsilon, \hat{F}(c - \varepsilon)\right)$ completely lies in Π_β. Then the function

$$\hat{F}_\varepsilon(u) = \begin{cases} \hat{F}(u), & \text{if } u \in \left[x_{(1)}, c - \varepsilon\right], \\ \hat{u}\frac{\gamma - \hat{F}(c-\varepsilon)}{\varepsilon} + \gamma - c\frac{\gamma - \hat{F}(c-\varepsilon)}{\varepsilon}, & \text{if } u \in (c - \varepsilon, c), \\ \hat{F}(u), & \text{if } u \in \left[c, x_{(n)}\right], \end{cases}$$

increases, is continuous, convex on $\left[x_{(1)}, c\right]$, concave on $\left[c, x_{(n)}\right]$, and its graph lies in Π_β. Thus, d.f. $\hat{F}_\varepsilon(u)$ is unimodal, and we can consider it as an estimation of the hypothetical d.f. of a general population G. The significance level of this test is β. Theorem 1 is proved.

Remark 1. Theorem 1 has the following geometric sense: let c be the abscissa of the first exit point, where the convex minorant $\varphi_{inf}(u)$ exceeds the upper bound of Π_β while moving from the maximal order statistics to the minimal one, and let d be the abscissa of the first exit point, where the convex minorant $\psi_{sup}(u)$ exceeds the upper bound of Π_β while moving from the minimal order statistics to the maximal one. Then the hypothetical d.f. $F(u)$ is unimodal iff the exit points c and d lie outside $\left[x_{(1)}, x_{(n)}\right]$ or $c \geq d$.

5. Test for unimodality based on m.e.d.f.

The confidence strip $\hat{\Pi}_\beta$ for a hypothetical d.f. can be constructed on m.e.d.f. $\tilde{F}_n^*(u)$ in the following way: let the significance level β^* be given, let ε be the width of Π_β, and let

$$p\left(\Delta = \max_{x_{(1)} \leq u \leq x_{(n)}} |F(u) - F_n^*(u)| > \varepsilon\right) = \beta^*.$$

We put $\tilde{\varphi}(u) = \tilde{F}_n^*(u) + \varepsilon + \frac{1}{n}$ and $\tilde{\psi}(u) = \tilde{F}_n^*(u) - \varepsilon - \frac{1}{n}$. It is easy to see that $\tilde{\Pi}_{\beta^*}$ with lower bound $\tilde{\psi}(u)$ and upper bound $\tilde{\varphi}(u)$ has the significance level not exceeding β^*. Indeed, by the virtue of (4),

$$\left|F(u) - \tilde{F}_n^*(u)\right| = \left|F(u) - F_n^*(u) + F_n^*(u) - \tilde{F}_n^*(u)\right| \leq |F(u) - F_n^*(u)| + \frac{1}{n}$$

Therefore,

$$\tilde{\Delta} = \max_{x_{(1)} \leq u \leq x_{(n)}} |F(u) - F_n^*(u)| < \Delta + \frac{1}{n}$$

Hence,

$$p\left(\tilde{\Delta} - \frac{1}{n} \geq \varepsilon\right) \leq p\left(\Delta \geq \varepsilon\right) = \beta^*,$$
$$p\left(\tilde{\Delta} \geq \varepsilon + \frac{1}{n}\right) \leq p\left(\tilde{\Delta} \geq \tilde{\varepsilon}\right) = \beta^*.$$

Thus, the significance level of $\tilde{\Pi}_{\beta^*}$ does no exceed β^*. Since we increase the validity of the test by selecting β^* as a significance level, we can use the m.e.d.f $\tilde{F}_n^*(u)$ to construct $\hat{\Pi}_\beta$ without decrease in the significance level. However, doing this, we increase the width of $\tilde{\Pi}_{\beta^*}$ by $\frac{1}{n}$ relative to Π_{β^*}. For moderate samples ($30 \leq n \leq 200$), this increment varies from 7 to 13

Now, we can formulate the test for the unimodality of a hypothetical d.f. based on m.e.d.f.

Theorem 2. *Let $\tilde{\varphi}_{\inf}(u)$ and $\tilde{\psi}_{\sup}(u)$ be the convex minorant and concave majorant of $\tilde{\varphi}(u)$ and $\tilde{\psi}(u)$, respectively, and let*

$$c = \sup\left\{u : \tilde{\varphi}_{\inf}(u) \leq \tilde{\psi}(u),\ x_{(1)} \leq u \leq x_{(n)}\right\},$$

$$d = \inf\left\{u : \tilde{\psi}_{\sup}(u) \geq \tilde{\varphi}(u),\ x_{(1)} \leq u \leq x_{(n)}\right\}.$$

Then, the hypothetical distribution $F(u)$ is unimodal iff

1) $\tilde{\varphi}_{\inf}(u) \geq \tilde{\psi}(u)$ or $\tilde{\psi}_{\sup}(u) \leq \tilde{\varphi}(u)$ $\forall u \in \left[x_{(1)}, x_{(n)}\right]$;

or

2) $c \geq d$.

Moreover, the significance level of this criterion is β^*.

The proof of Theorem 2 is similar to that of Theorem 1.

6. Conclusion

It is shown in [12] that if the distribution function of a general population is unimodal, then the confidence interval $(m(x) - 3\sigma(x), m(x) + 3\sigma(x))$, where $m(x)$ is the mathematical expectation of G and $\sigma(x)$ is the standard deviation of G, has the significance level which does not exceed 0.05. That is why, this nonparametric test for unimodality can be used to construct the confidence interval for the bulk of the general population G.

Bibliography

1. B.W. Silverman, *Using kernel density estimates to investigate multimodality*, J. of the Royal Statistical Society B **43** (1981), 97-99.
2. J.A. Hartigan, *Computation of the dip statistics to test for unimodality*, Applied Statistics **34** (1985), 320-325.
3. J.A. Hartigan, *The span test of multimodality*, Classification and Related Methods of Data Analysis, (H. H. Bock, ed.), North-Holland, Amsterdam, 1988, pp. 229-236.
4. J.A. Hartigan, S. Mohanty, *The RUNT test for multimodality*, Applied Statistics **9** (1992), 63-70.
5. A.N. Kolmogoroff, *Determinatione empirica di una legge di distributione*, Giornale Instit. Ital. Attuari **4** (1933), 83-91.
6. N.V. Smirnov, *Sur les ecarts de la courbe de distribution empiric*, Mat. Sb. **6** (1939), 3-26.
7. A. Wald, J. Wolfowitz, *Confidence limits for continuous distribution functions*, Ann. Math. Statist. **10** (1939), 199-326.
8. W. Feller, *On the Kolmogorov–Smirnov limit theorems for empirical distributions*, Ann. Math. Statist. **19** (1948), 177-189.
9. F.J. Massey, *A note on the estimation of a distribution function by confidence limits*, Ann. Math. Statist. **21** (1950), 125-128.
10. Z.W. Birnbaum, F.H. Tingey, *One-sided confidence contours for distribution functions*, Ann. Math. Statist. **22** (1951), 592-596.
11. B.L. Van der Waerden, *Mathematische Statistik*, Springer, Berlin, 1957.
12. D.F. Vysochanskij, Yu.I. Petunin, *Justification of the 3-σ rule for unimodal distribution*, Theor. Probability Math. Stat. **21** (1980), 25-36.

Department of Mathematical Sciences and Center for Applied Mathematics and Statistics, New Jersey Institute of Technology, Newark, NJ 07102, USA

Taras Shevchenko Kyiv National University, Department of Cybernetics, 64, Volodymyrska Str., Kyiv 01033, Ukraine
E-mail: vm214@dcp.kiev.ua