Kuzmin V.M., Lapach S.M., Paltsun S.V. The new approach to bimodal testing of small samples // International Conference "Modern Stochastics: Theory and Applications II", September 7-11, 2010. Kyiv, Ukraine. Abstracs –K.: 2010, P.89.

Full version

Описывается проверка гипотезы об унимодальности распределения для малых выборок с использованием бипараболической интервальной функции

Кл. слова

Малые выборки, тестирование бимодальности, бипараболическая интервальная функция

bimodal testing, small samples, biparabolic interval function

Kuzmin V.M., Lapach S.M., Paltsun S.V.

The new approach to bimodal testing of small samples

Problem

Absence of reliable criteria, which permit to define whether it is unimodal for small sample.

Objective

To develop the uni- or bimodal checkout methodology of small samples with calculation of modes.

The determination of random sample bimodality (if it take place) is very complicated, especially for small samples. It requires development of non-standard approaches and methods for the task to be solved. This work offers the approach, based on **biparabolic interval function** (BPIF). The interval function was offered for small samples (mode calculation) statistical analysis [2]. **The interval function is constructed as follows. Two new statistics D_i and R_i, where $D_i=z_{i+1} - z_i$ successive differences, while $R_i = (z_i+z_{i+1})/2$ (i=1,n) half-sums of neighbour random values**, are calculated for random values sample, ranged in increasing order (z_i). Take $y_i \equiv D_i$ and $x_i \equiv R_i$ and obtain the interval function for sample under examination. The interval function is the dependency of all neighbour random value intervals on their location in the sample space. The biparabolic interval function is the combination of two parabolic curves, which have a point in common at section x=g, constructed by least squares method.

The equation of biparabolic interval function is folloing

$$y = a + bx + cx^2 + d(x-g)_+ + e(x-g)_+^2 \quad (1)$$

Using a least squares method to equation (1) we get system of equations

$$\sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i + c\sum_{i=1}^{n} x_i^2 + d\sum_{i=1}^{n} (x_i - g)_+ + e\sum_{i=1}^{n} (x_i - g)_+^2$$

$$\sum_{i=1}^{n} y_i x_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 + c\sum_{i=1}^{n} x_i^3 + d\sum_{i=1}^{n} x_i (x_i - g)_+ + e\sum_{i=1}^{n} x_i (x_i - g)_+^2$$

$$\sum_{i=1}^{n} y_i x_i^2 = a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i^3 + c\sum_{i=1}^{n} x_i^4 + d\sum_{i=1}^{n} x_i^2 (x_i - g)_+ + e\sum_{i=1}^{n} x_i (x_i - g)_+^3 \qquad (2)$$

$$\sum_{i=1}^{n} y_i (x_i - g)_i = a\sum_{i=1}^{n} (x_i - g)_+ + b\sum_{i=1}^{n} x_i (x_i - g)_+ + c\sum_{i=1}^{n} x_i^2 (x_i - g)_+ + d\sum_{i=1}^{n} (x_i - g)_+^2 + e\sum_{i=1}^{n} (x_i - g)_+^3$$

$$\sum_{i=1}^{n} y_i (x_i - g)_i^2 = a\sum_{i=1}^{n} (x_i - g)_+^2 + b\sum_{i=1}^{n} x_i (x_i - g)_+^2 + c\sum_{i=1}^{n} x_i^2 (x_i - g)_i^2 + d\sum_{i=1}^{n} (x_i - g)_+^3 + e\sum_{i=1}^{n} (x_i - g)_+^4$$

Algorithm

In order to solve the problem, whether the sample under examination is uni- **or bimodal, the following algorithm is offered.**

1. The interval function is constructed.
2. The biparabolic regression is constructed on interval function.
3. The observed parabolic curve is constructed on interval function.
4. The response values are found for both regression models at switch point of polygonal.
5. The confidence intervals are found at this switch point section for both regressions.
6. The comparison of confidence intervals is made. If they disjoint, the sample is considered bimodal, if not – unimodal.
7. The modes values are determined.

Let’s consider the suggested procedure by the way of example. The input data for the example were taken from [1]. These data (Table 1) represent the weight of monkeys. Should be mentioned these data have a point, which looks like outlier, but from the viewpoint of the biologist is not such – it is the weight of the leader-monkey. The calculations were made either according to the standard macros from “Data Analysis” of Excel, or with the help of macros, developed in [3].

Table 1. Input data (weight, kg)

Female animal	10	10	10,1	10,2	10,8	11	11,1	11,3	11,3	11,4
	11,8	12	12	12,1	12,3	13	13,4	13,5	14,5	15,6
Male animal	17,7	18	18	19	19	20	21	22	30	

As is clear from fig. 1, these data are referred to different main entities (t_{exp}= 6,534 > t_{cr} =2,306 as a check on hypothesis on the averages). The check is performed on this population of data, as their bimodality is obvious.

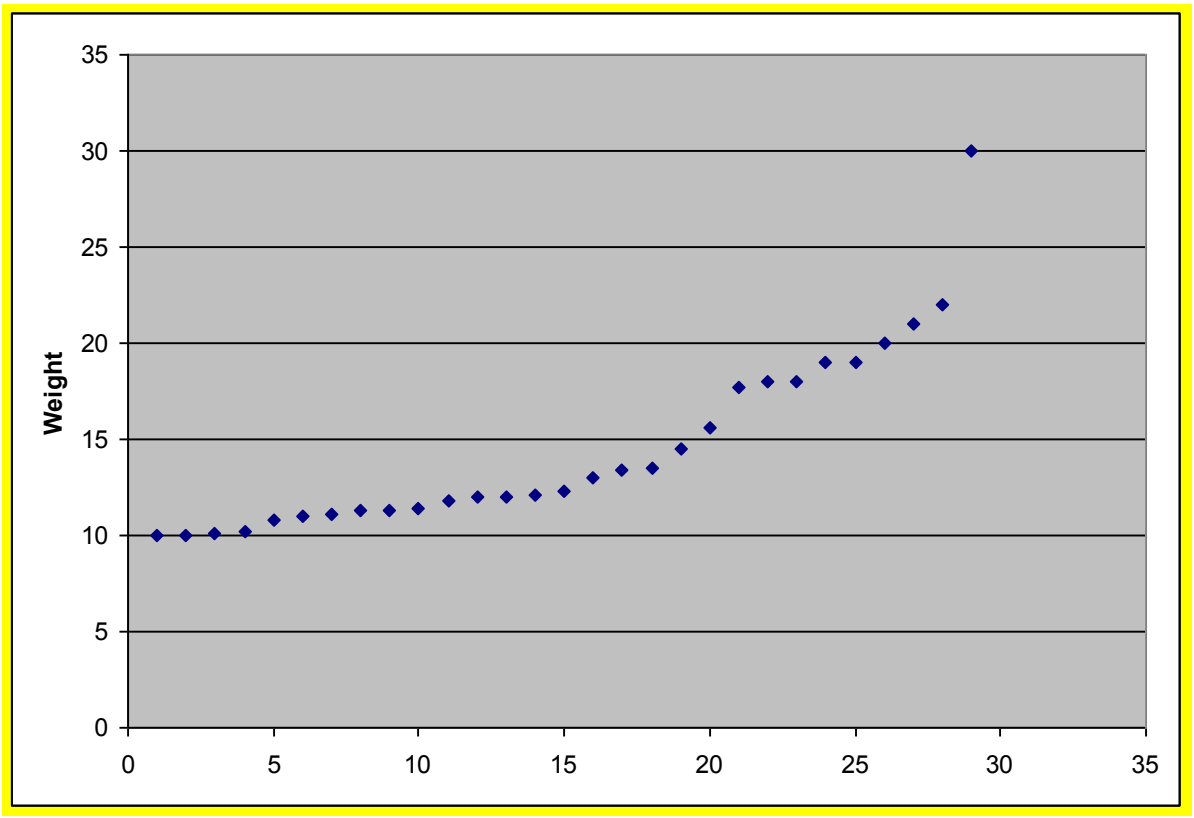

Fig.1. Simply ordered values of animals weight

The interval function is constructed (Table 2).

Table 2. Data for interval function

Ordered series	Difference between neighbour elements (y)	Half-sum of neighbour values (x)
10		
10	0	10
10,1	0,1	10,05
10,2	0,1	10,15
10,8	0,6	10,5
11	0,2	10,9
11,1	0,1	11,05
11,3	0,2	11,2
11,3	0	11,3

11,4	0,1	11,35
11,8	0,4	11,6
12	0,2	11,9
12	0	12
12,1	0,1	12,05
12,3	0,2	12,2
13	0,7	12,65
13,4	0,4	13,2
13,5	0,1	13,45
14,5	1	14
15,6	1,1	15,05
17,7	2,1	16,65
18	0,3	17,85
18	0	18
19	1	18,5
19	0	19
20	1	19,5
21	1	20,5
22	1	21,5
30	8	26

In case of bimodality the data from Table 2 are represented by piecewise continuous function

$$y = a + bx + cx^2 + d\frac{(x-g)+|x-g|}{2} + e\left(\frac{(x-g)+|x-g|}{2}\right)^2 \quad (3)$$

The coefficients of the model equation are defined by the least squares method. The break point is in such a way that provides minimal residual dispersion [3]. In this case the break point is 16,33816, and model equation is as follows

$$y = 7{,}164643 - 1{,}29038x + 0{,}059248x^2 - 1{,}64623\frac{(x-16{,}33816)+|x-16{,}33816|}{2} + 0{,}109853\left(\frac{(x-16{,}33816)+|x-16{,}33816|}{2}\right)^2 \quad (4)$$

Multiple correlation coefficient R=0,978; statistically significant F_R=126,34 > F_{cr}=2,796; residual dispersion 0,116993. Geometric representation is given in Fig. 2.

For unimodal function the approximation by the parabola order should be adequate.

$$\hat{y} = a + bx + cx^2 \quad (5)$$

In this case $\hat{y} = 9{,}066185 - 1{,}31007x + 0{,}046861x^2$ (6)

Multiple correlation coefficient R=0,873; statistically significant F_R=39,94 > F_{cr}=3,385; residual dispersion 0,589349.

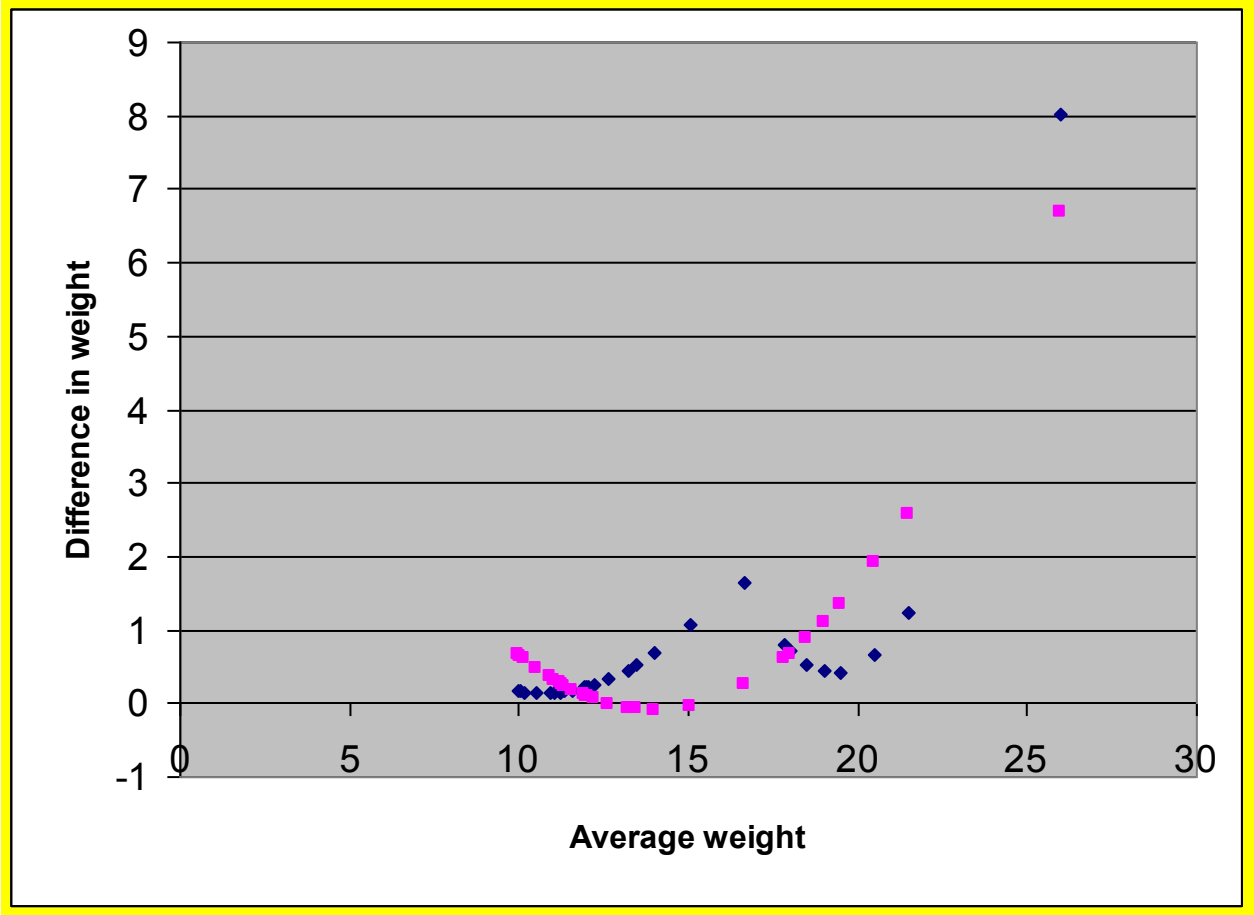

Fig. 2. Model-made points for interval functions

Define the confidence intervals for both equations at pick up point piecewise continuous model. The confidence intervals are calculated according to the formula

$$\hat{Y}_k \pm t(\nu;1-\alpha/2)s_\nu\sqrt{X_k'(X'X)^{-1}X_k}$$ ([4]).

The results of calculations are summarized in Table 3.

Table 3. Confidence intervals

Figures	Parabola	Piecewise continuous
Root-mean-square error	0,767691	0,342042361
Student's t-test	2,059539	2,068657599
Half-width of interval	0,485028	0,562794899
Response	0,17109	1,897567064
Left class boundary	-0,31394	1,334772165
Right class boundary	0,656118	2,460361963

As the confidence intervals disjoint, we accept the hypothesis about bimodality of this sample.

To find mode for each of subsamples we derive from equation (1) and set it to zero.

As the equation has modulus signs, it falls into two parts. Namely:

If x > g $y' = b + 2(c+e)x + d - 2eg$,

$$x_1 = \frac{2eg - b - d}{2(c+e)} \cdot \qquad (7)$$

If x < g $y' = b + 2cx$, $x_2 = -\frac{b}{2c}$. (8)

Therefore x_2=19, 30; x_1=10,98.

Decrease the weight of all male animals by 4 kg. And check this population of data, for which the bimodality is not self-evident (Fig. 3).

These data are referred to different main entities (t_{exp}= 3,326 > t_{cr} =2,306 as a check on hypothesis on the averages), but during sorting in magnitude they are mixed and are not visually differentiated.

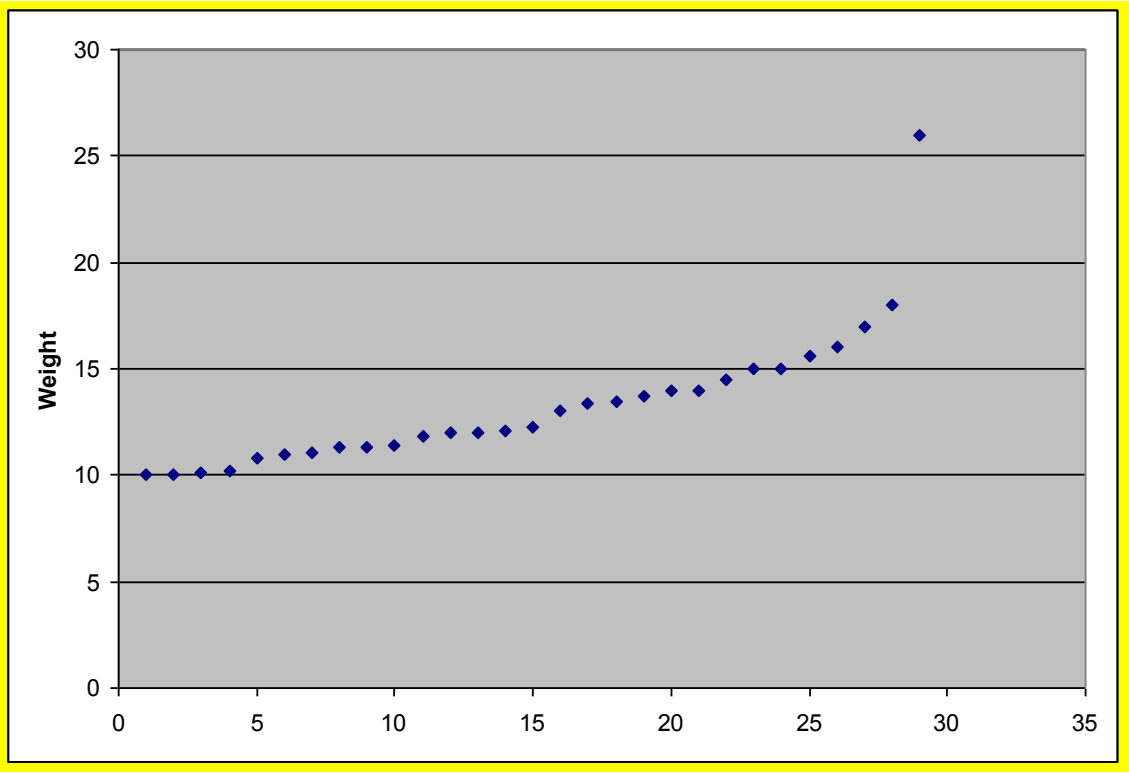

Fig. 3 Corrected ordered series

In this case the break point is 16,23329, and model equation is as follows

$$y = 8{,}070027 - 1{,}44461x + 0{,}065749x^2 - 1{,}68442\frac{(x-16{,}23329)+|x-16{,}23329|}{2} + 0{,}096869\left(\frac{(x-16{,}23329)+|x-16{,}23329|}{2}\right)^2$$

(9)

Multiple correlation coefficient R=0,926; statistically significant F_R=34,37 > F_{cr}=2,796; residual dispersion 0,118127.

For hypothesis about unimodal differentiation, the model of parabola is as follows

$$y = 3{,}5035531 - 0{,}52281x + 0{,}020066x^2 \quad (10)$$

Multiple correlation coefficient R=0,739; statistically significant F_R=15,06 > F_{cr}=3,385; residual dispersion 0,343884.

The results of confidence intervals calculation are summarized in Table 4.

Table 4. Confidence intervals

Figure	Parabola	Piecewise continuous
Root-mean-square error	0,586416	0,343696
Student's t-test	2,068658	2,068658
Half-width of interval	0,398563	0,617731
Response	0,460947	1,945327
Left class boundary	0,062384	1,327596
Right class boundary	0,85951	2,563057

As the confidence intervals disjoint, we accept the hypothesis about bimodality of samples.

Conclusion

The offered methodology enables to determine whether distribution of sample is uni- or bimodal and to calculate the modes. It is very importantly for small samples.

References

1. Lakin G.F. Biometry – 4th edition, revised and enlarged - M.; Higher School 1990. - 352 p.: il.
2. Kuzmin V.N. The interval function as a tool for statistical analysis of small samples Materials of XI international scientific conference named after academician Kravchuk –K.: 2006, p. 718
3. Kuzmin V.M., Lapach S.M. Application of polygonal regression in economic researches // Economics and management –2004, –№3. –p. 79–84.
4. Norman Draper, Harry Smith. Applied regression analysis, third edition. –M.: PH Williams. 2007. –912 p.