Assistant Professor Iftikhar Hussain ADIL, PhD
Email:iftikhar.adil@s3h.nust.edu.pk
National University of Sciences and Technology, Islamabad
Professor Asad ZAMAN, PhD
Email: asadzaman@ilum.mit.edu
Director, Social Sciences, Al-Nafi Online Educational Platform
Member, Monetary Policy Committee, State Bank of Pakistan

OUTLIERS DETECTION IN SKEWED DISTRIBUTIONS: SPLIT SAMPLE SKEWNESS BASED BOXPLOT

Abstract. *Data analysis is the core heart of quantitative research. An important part of data analysis is outlier detection. Several outlier detection techniques have been developed in the past that include a popular technique named as Tukey's boxplot. Tukey's boxplot uses inter quartile range to detect outliers on both sides of median. This method works well in symmetric distributions while it constructs misleading fence in skewed distributions. Several amendments have been made in Tukey's method to detect outliers in skewed distributions. Latest amendment is suggested by Hubert and Vandervieren who incorporated the exponent of medcouple multiplied by different constants conditioned on the direction of skewness. This paper provides a novel approach to fix the problem by splitting the data in two halves from the median and constructing fence separately on both sides of the median. Splitting process enables detection process efficiently, making it robust. Mathematical calculation of probability of Type-I error and interval width proved that this method has superiority on Tukey and Hubert and Vandervieren method.*

Keywords: *Skewness, Outliers, Medcouple, Boxplot.*

JEL Classification: J13, C54

1. Introduction

Detection of outliers is very important in modeling, inference, and even data processing because outliers can lead to model misspecification, biased parameter estimation, and poor forecasting (Tsay, Pena and Pankratz, 2000). For univariate

DOI: 10.24818/18423264/54.3.20.17

distributions Tukey's (1977) boxplot is a very popular tool for detection of outliers. However, it uses symmetric criteria, and can fail on skewed data sets. Many naturally occurring data sets are skewed, and Tukey's boxplot fails to correctly detect outliers in such data sets.

To solve this problem, Hubert and Vandervieren (2008) created a modified boxplot (henceforth HVBP) which uses a robust measure of skewness. Their technique has certain limitations, as this paper will explore. This paper introduces an alternative technique based on splitting the sample around the median and then applying Tukey's technique separately on each half. This research will also show that newly introduced test improves on both the original Tukey technique, and the HVBP technique in terms of detecting outliers in skewed distributions.

It is clear that the Tukey boxplot will not place the fences for outliers correctly in a skewed distribution. Because it uses the same measure of spread IQR to make the upper and lower fence, the upper fence will be too short for the long tail, and the lower fence will be too long for the short tail. This means that it will miss outliers in the short tail, and find extra outliers in the long tail. For the sake of clarity, we illustrate this phenomenon with a simple example

2. Failure of Tukey-type symmetric techniques in skew distributions.

The boxplot of Tukey consists of a lower boundary and an upper boundary, defined as:

$$[L \quad U] = [Q_1 - g * IQR \qquad Q_3 + g * IQR] \ldots\ldots\ldots\ldots..(1)$$

Where L denotes lower bound or lower critical value and U represents upper bound or upper critical value, Q_i is the i-th quartile, and IQR is the difference of third and first quartile i.e. $Q_3 - Q_1$ is the interquartile range. L and U are called the lower and upper fences. In equation (1) value of constant g used by Tukey is 1.5. Data outside the fences is considered as outliers. The key issue to note here is that, in presence of outliers, distance from the central box to the lower fence and the upper fence is exactly the same, regardless of the level of skewness in the data. This causes problems in skew distributions as we illustrate with a simple example

Consider a standard normal random variable Z i.e. N (0, 1). The first and third quartile are Q1=-0.675 and Q3=+0.675 so the interquartile range is IQR=1.35. The lower and upper fences constructed by Tukey are LCV = -2.698, and UCV=+2.698. For a standard normal, P(-2.698<Z<+2.698)=99.3% so under the null hypothesis

DOI: 10.24818/18423264/54.3.20.17

$Z \sim N(0,1)$, the Tukey test will make a type I error in classifying an observation as an outlier only in 0.7% of the cases. However, the situation changes when we consider the skew log normal random variable X=exp(Z). In this case, Q1=exp(-0.675)=0.51 and Q3=exp(0.675)=1.96, so IQR=1.45. By using Tukey's boxplot LCV is -1.67. This lower fence is far below the smallest possible value of X which is 0. Similarly, UCV according to Tukey is 4.47. We have P(X<UCV)=P(Z<ln(4.47))=93%. Now the probability to a type I error on the upper side is 7%, while the probability of a Type-I error on the lower side is zero. By using the same tail (1.5 IQR) on both sides, the Tukey Boxplot makes the tail too long on the short side, and makes it too short on the long side of the skew distribution.

This paper proposes a very simple fix for this problem. First split the sample at the median. Then apply the Tukey technique to both sides of the distribution separately. Since the IQR is now estimated separately on both sides of the sample, we get a short tail on the short side, and a long tail on the long side, fixing the problem for skewed samples. We call this the SSSBB: Split-Sample Skewness Based Boxplot. We illustrate how this works in the simple example of the lognormal given above.

Applying this to the Lognormal distribution, we get Q_{1L}= value of standard normal at 12.5 percentile = -1.15 and Q3L = Value of standard normal at 37.5 percentile = -0.32, which leads to an IQR of 0.83, and an LCV =-1.15-1.5*0.83=-2.40 hence by symmetry, UCV = +2.4. Now P(-2.4<Z<+2.4) = 98.4% so for the symmetric distribution, this procedure has a higher probability of type I error 1.6% compared to Tukey's 0.7%. But this procedure now draws better fences for the Lognormal distribution. For lognormal distribution LCV is -0.299 which is closer to zero as compared to Tukey's LCV which is -1.67. On the right side of the lognormal distribution UCV's of Tukey and SSSBB are 4.47 and 5.84 respectively. So it may be observed that SSSBB's upper critical value is extended than Tukey. By taking log of both critical values and the finding the probability of Type-I error, it may be noted that probability of Type-I error for Tukey and SSSBB are 6.72 and 3.89 respectively on the upper side after taking log of UCV of lognormal distribution.

3. The Proposed Alternative based on Splitting the Sample

As discussed in the introduction, this problem was also noted by Huber and Vandervieren, who have proposed a different alternative, which we label HVBP.

In the paper, we will compare the three techniques on a number of different well known skewed distributions. Prior to undertaking a detailed comparison, we illustrate

DOI: 10.24818/18423264/54.3.20.17

how the three techniques perform on a simple small simulated data set from a log normal distribution.

Table 1: Comparison of Techniques in Example Data Set from N (0, 1)

N(0,1)	N(0,1)	Exp(N(0,1))	Exp(N(0,1))	Tukey Test		SSSBB	
-0.41	0.57	0.66	1.76	Q1	0.63	Q1L	0.43
-1.46	0.89	0.23	2.44	Q3	1.82	Q3L	0.68
1.04	0.59	2.83	1.80	IQR	1.19	IQRL	0.26
-0.39	-0.09	0.68	0.91	LCV	-1.15	Q1R	1.40
-0.86	-0.37	0.42	0.69	UCV	3.61	Q3R	2.68
0.64	-0.45	1.90	0.64	HVBP		IQRR	1.28
-1.45	0.38	0.23	1.47	LCV	0.19	LCV	0.04
-0.47	1.42	0.63	4.14	UCV	10.66	UCV	4.61
-0.32	1.05	0.72	2.87	Min	0.23		
-0.85	-0.34	0.43	0.71	Max	4.14	Skew	1.25

Applying the three techniques on a simple data set of twenty observations from a lognormal, we observe that probabilities of Type-I error on short tail by Tukey, HVBP, and SSSBB are 0, 5.1 and 0.07 percent respectively. Probabilities of Type-I error on long tail are 10, 1 and 6 percent by Tukey, HVBP and SSSBB respectively for this sample of twenty observations. So the total Type-I error probabilities are 10, 6.1 and 6.07 for Tukey, HVBP and SSSBB respectively. On the other hand, when we look in intervals made by the techniques, it may be observed that Tukey's, HVBP's, and SSSBB interval widths are 4.76, 10.47 and 4.57 respectively. It is clear from these statistics that probability of Type-I error for Tukey is more than all techniques under comparison while probabilities of Type-I error for HVBP and SSSBB are almost same but interval made by HVBP is more than 130 percent larger than SSSBB. This illustrates the strengths of our newly introduced technique in a simple special case. Now we turn to a more extensive comparison.

4. Hubert and Vandervieren's Medcouple Based Alternative

Noting the problems in detecting outliers in skewed distributions, Hubert and Vandervieren introduced a new technique for this purpose. Firstly, they noted that the classical measure of skewness is not robust and it is affected by outliers. They propose to use the Medcouple, introduced by Brys, Hubert, and Struyf (2003) as a robust alternative to classical skewness. Suppose $X_n = \{x_1, x_2, x_3, \dots\dots\dots\dots\dots..x_n\}$ is a random sample from the univariate distribution under consideration. For x_j> median denoted by med_k and x_i< median, define skewness via the kernel function$h(x_i, x_j)$ where

$$h(x_i, x_j) = \frac{(x_j - med_k) - (med_k - x_i)}{(x_j - x_i)} \dots\dots\dots\dots\dots (2)$$

Note that this measures the ratio of how much x_j exceeds the median relative to how much x_i is below the median, which is a natural measure of skewness. Next, the medcouple takes the median over all such pairs of observations to arrive at a measure of skewness for the sample:

$$MC = {}_{x_i \leq med_k \leq x_j}^{\;\;med} h(x_i, x_j) \ldots\ldots\ldots\ldots\ldots \qquad (3)$$

There are some complications in case of ties, which are ignored here, since these cannot arise in continuous distributions. This paper only deals with the continuous distributions.

After considering several alternatives, Mia Hubert and Ellen Vandervieren (2008) proposed the following skewness-adjusted boxplot; we will call this HVBP:

$$[L \quad U] = [Q_1 - \ 1.5 * IQR * e^{-3.5*MC} Q_3 + \ 1.5 * IQR * e^{4*MC}] \qquad \text{If MC >0}$$
$$[L \quad U] = [Q_1 - \ 1.5 * IQR * e^{-4*MC} \ Q_3 + \ 1.5 * IQR * e^{3.5*MC}] \qquad \text{If MC <0}$$

Where MC represents medcouple and IQR is the inter quartile range. In case of MC equals to zero implies that data is symmetric and by substituting zero value of MC leads to Tukey boxplot in both cases of the above HVBP equations. This modified boxplot computes the skewness and automatically makes the fence farther on the wide side and closer on the narrow side of the distribution.

Mia Hubert and Ellen Vandervieren (2008) used medcouple and proposed adjustment in the Tukey's technique as given in the previous section. But this modification can make the interval of critical values too long, especially on the skewed side. For example, we can apply this technique to the above mentioned hypothetical example. Medcouple for this data set is 0.33 which correctly shows that distribution is right skewed. By applying HVBP technique the critical values are calculated as [-729.8 1028.1]. The critical values calculated by HVBP lie beyond the extremes of the real data [-200 540]. This technique has erroneously extended critical values away from the data on both sides.

5. New Proposed Technique

This technique divides the data into two parts from the median, so that we have exactly 50% data on both lower and upper sides of the median. Treat these lower and upper sides as complete data sets and find the first quartile for the lower side Q_{1L}, third quartile for the lower side Q_{3L} and inter-quartile range for the lower side $IQR_{L.}$ Similarly, first quartile for upper side Q_{1R}, third quartile for the upper side Q_{3R} and inter-quartile range for the upper side IQR_R is also computed. Lower and upper critical values for detecting outliers in the skewed distributions are suggested by subtracting

1.5 times the inter quartile range of the lower side from the first quartile of the lower side of the median and adding 1.5 times the inter quartile range of the upper side with the third quartile of the right side of the median. On the basis of the above splitting into two parts from the median which is based on the skewness of the data, we call new technique as Split Sample Skewness Based Boxplot hereafter SSSBB. Mathematically for the complete data set

Q_{1L} = 12.5th percentile, Q_{1R} = 62.5th percentile,
Q_{3L} = 37.5th percentile, Q_{3R} = 87.5th percentile,
$IQR_L = Q_{3L} - Q_{1L}$=37.5th percentile - 12.5th percentile,
$IQR_R = Q_{3R} – Q_{1R}$ = 87.5th percentile - 62.5th percentile

Lower and upper boundaries are defined as

$$[L \quad U] = [Q_{1L} - g * IQR_L \qquad Q_{3R} + g * IQR_R]$$

Where L is the lower critical value and U is upper critical value of the data. An observation outside these boundaries $[L \quad U]$ would be labeled as an outlier. The value of g used in this technique is 1.5. By applying this technique on the same hypothetical data we calculated the boundaries as [-88.93 596.06]. It can be observed that SSSBB technique's boundaries are close to data on both sides. This technique successfully detects the left outlier at -200 in contrast to Tukey's and HVBP. Unlike Tukey, it also shows that 540 is not a right outlier.

6. Methodology

Every outlier detection technique makes a fence to discriminate between the usual observations and the outliers. A fence is the boundary constructed by the formula of specific technique to detect outliers. The observations inside the fence are treated as normal while outside the fence those observations are treated as outliers. The comparison of outlier detection techniques is based on the match between the Type-I error made by any technique and interval width simultaneously. If the distribution of the data is skewed the classical outlier detection techniques tend to treat symmetrically both sides of the data. Therefore, it leaves significant data on long tail side of the distribution and covers extra area on the shorter tail of the distribution. As a result, an unusual observation on the shorter tail of the distribution cannot be detected. In order to ensure the match between the distribution and the fence, this paper considers fence as 95% area of the distribution by allowing 5% probability of type I error; that is, the central 95% values are treated as normal, while the top and bottom 2.5% are treated as potential outliers. It is expected that all techniques will construct their fence over the true 95% boundary. If any of the techniques under comparison commits sum of Type-I error probabilities on both sides more than 5% may be treated as weak.

H_0: Sum of probabilities of Type-I error is less than or equal to 5 percent.
H_1: Sum of probabilities of Type-I error is more than 5 percent.

Now here are two options. First, if sum of Type-I error probabilities is more than 5% in any technique under comparison then it is not tolerable and vice versa. Secondly, interval width will be compared as technique having smaller interval will have greater precision. This study compares both criteria at the same time. At first step if all the techniques fulfill this criterion then the technique constructing smaller interval is better and powerful being smaller interval with having high precision.

The selected skewed distributions are chi square with 2, 5, 10, 15, 20 degrees of freedom. Beta with parameters (35, 2), (35, 3), (35, 4) and (35, 5) and lognormal distribution with parameters (0, 0.2), (0, 0.4), (0, 0.6), (0, 0.8), (0, 1) are taken for analysis. As this study compares HVBP technique in which medcouple is used. This study uses simulated value of medcouple for the above selected distributions with sample size of 300 and 10000 simulations. The simulated values are given as

Table 2: Medcouple

Beta	Parameter	Beta(35,1)	Beta(35,2)	Beta(35,3)	Beta(35,4)	Beta(35,5)	Beta(35,6)
	Medcouple	-0.32	-0.21	-0.16	-0.13	-0.11	-0.10
Chi Square	Degree of Freedom	Chi square (2)	Chi square (5)	Chi square (10)	Chi square (15)	Chi square (20)	Chi square (25)
	Medcouple	0.33	0.20	0.13	0.11	0.10	0.08
Lognormal	Parameters	Logn (0,.2)	Logn (0,.4)	Logn(0,.6)	Logn(0,.8)	Logn(0,1)	
	Medcouple	0.09	0.17	0.25	0.33	0.40	

As discussed in section 1 that Tukey technique constructs the fence around the standard normal having -2.698 and +2.698 as LCV and UCV respectively. For these boundaries of Tukey probability of Type-I error is 0.7% resulting 5.4 as an interval width. As medcouple for standard normal distribution is zero. This implies that HVBP reduces to Tukey technique in standard normal generating the same probability of Type-I error and interval width. For SSSBB, LCV and UCV in standard normal are -2.4 and +2.4 respectively resulting an interval width of 4.80. In case all the techniques under comparison have tolerable probability of Type-I error then a technique with smaller interval has an edge to have high precision. Comparison of Type-I error and interval width is done in the following tables and graphs given below.

6.1 Comparison of Techniques in Beta Distribution

In probability theory and statistics, the beta distribution is a family of continuous probability distributions defined on the interval [0, 1] parameterized by two positive shape parameters, denoted by α and β, that appear as exponents of the random variable and control the shape of the distribution. Parameter selected for this distribution are (35,1), (35,2), (35,3), (35,4), (35,5), (35,6) keeping in mind that its shape should be negatively skewed. The computed values of classical skewness and medcouple are reported in column 2 of the Table 3.

Table 3 reports the information of moment measure of skewness, medcouple, probability of Type-I error (short tail, long tail), LCV, UCV, and Interval width for the selected parameters in beta distribution by Tukey, HVBP and SSSBB techniques according to their formulae. Tukey's technique probability of Type-I error on short tail is zero while it is 4.5% on long tail. So the sum of probabilities of Type-I error is 4.5% which is less than tolerance level of 5%. In HVBP maximum probability of Type-I error is 0.08% while it is 0.18% on long tail. Total probability of Type-I error by HVBP is always less than 1% which meets the criteria of 5% tolerance level. For SSSBB, the highest probabilities of Type-I error are 0.10% and 2.25% on short and long tail respectively. This implies that all three techniques under comparison meet first criterion.

Table 3:Type-I Error Probability, Interval Width and Fences in Beta Distribution

Parameters	Moment measure of skewness	Type-I Error Probability (short tail)			Type-I Error Probability (long tail)			Type-I Error Probability (total)		
		Tukey	HVBP	SSSBB	Tukey	HVBP	SSSBB	Tukey	HVBP	SSSBB
(35, 1)	-1.84	0.0000	0.0000	0.0000	0.0450	0.0003	0.0225	0.0450	0.0003	0.0225
(35, 2)	-1.25	0.0000	0.0000	0.0000	0.0275	0.0009	0.0179	0.0275	0.0009	0.0179
(35, 3)	-0.98	0.0000	0.0000	0.0000	0.0205	0.0012	0.0157	0.0205	0.0012	0.0157
(35, 4)	-0.78	0.0000	0.0001	0.0000	0.0166	0.0015	0.0143	0.0166	0.0016	0.0144
(35, 5)	-0.71	0.0000	0.0005	0.0004	0.0140	0.0017	0.0134	0.0140	0.0021	0.0138
(35, 6)	-0.60	0.0000	0.0008	0.0010	0.0122	0.0018	0.0126	0.0122	0.0026	0.0136
Parameters	Medcouple SS = 300	Lower Critical Values			Upper Critical Values			Interval Width		
		Tukey	HVBP	SSSBB	Tukey	HVBP	SSSBB	Tukey	HVBP	SSSBB
(35, 1)	-0.32	0.92	0.79	0.90	1.04	1.01	1.01	0.12	0.21	0.11
(35, 2)	-0.21	0.86	0.77	0.85	1.04	1.01	1.01	0.19	0.24	0.17
(35, 3)	-0.16	0.81	0.74	0.80	1.04	1.00	1.01	0.23	0.26	0.20
(35, 4)	-0.13	0.77	0.71	0.77	1.03	0.99	1.00	0.25	0.28	0.23
(35, 5)	-0.11	0.74	0.68	0.74	1.02	0.98	0.98	0.28	0.30	0.24
(35, 6)	-0.10	0.71	0.66	0.71	1.01	0.97	0.97	0.29	0.31	0.26

For comparison of interval width, it can be observed from Table 3 that HVBP and SSSBB constructs largest and smallest interval respectively in all cases. Here we compare the shortest and highest percentage difference between Tukey and HVBP, SSSBB and HVBP, and Tukey and SSSBB. In β (35,6) which is less skewed, HVBP interval is (0.31-0.29)/0.29=6.9% larger than the interval made by Tukey's technique. And for β (35, 1) which is highly skewed, HVBP interval is 75% larger than Tukey's.

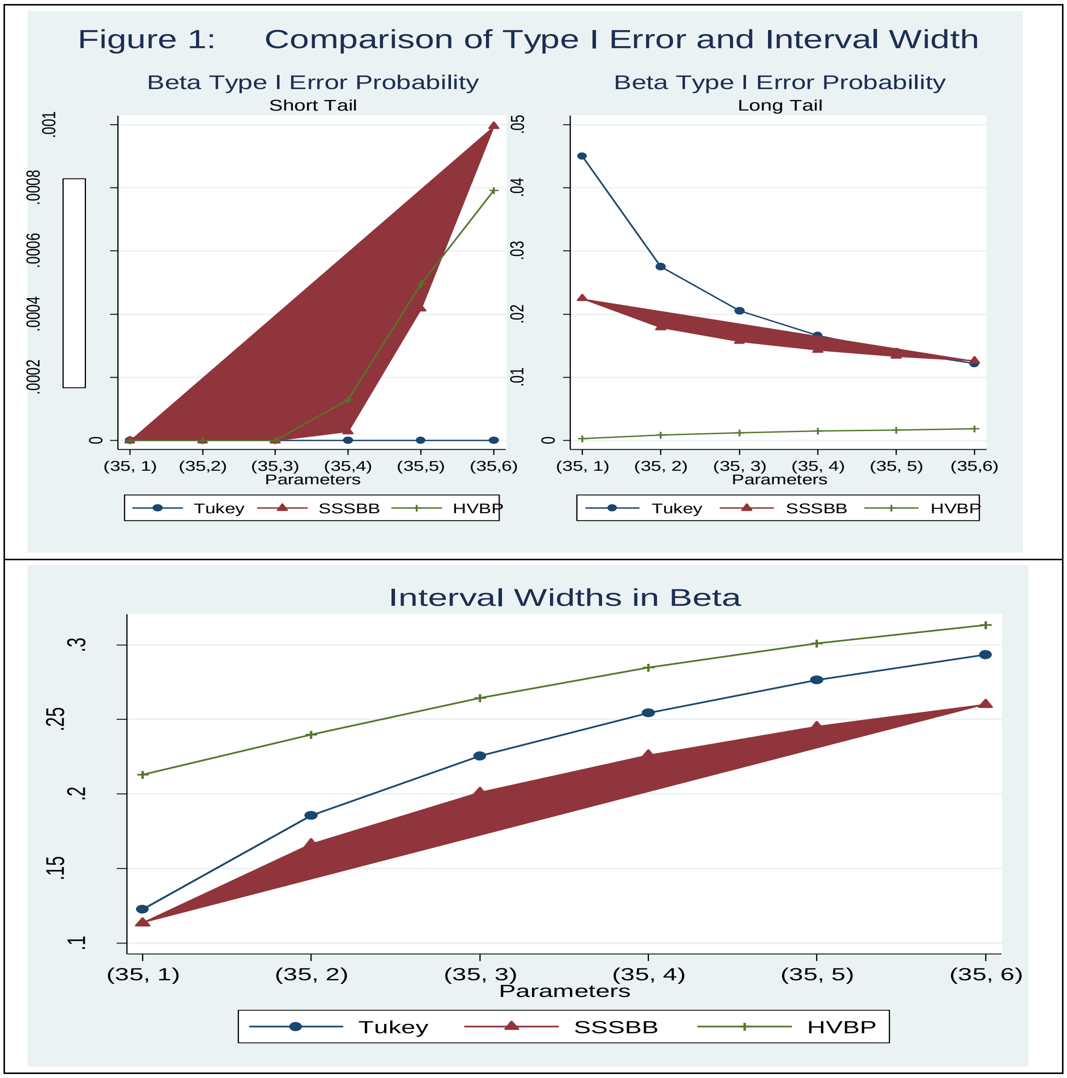


Figure 1: Comparison of Type I Error and Interval Width

In comparison of HVBP and SSSBB with respect to interval width in β (35, 6), HVBP interval is 19.2% larger. For β (35, 1), HVBP interval is 91% larger than SSSBB's

DOI: 10.24818/18423264/54.3.20.17

interval. Finally, for β (35, 6) Tukey's interval is 11.5% larger and in β (35, 1) it is 9.1% larger. Figure 1 shows clearly the interval widths and probability of type on both sides by each technique in beta distribution.

Table 4: Type-I Error Probability, Interval Width and Fences in χ^2 Distribution

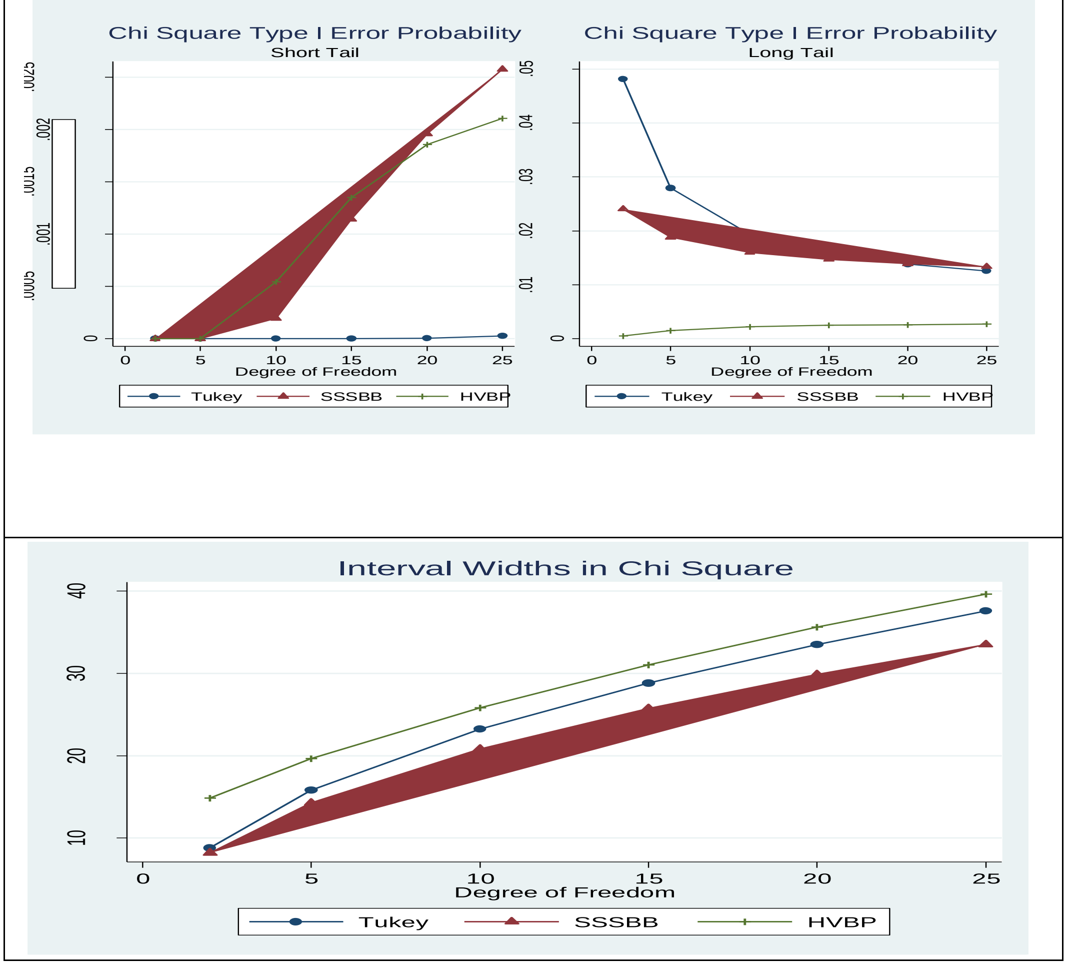
Degree of freedom	Moment measure of skewness	Type-I Error Probability (short tail)			Type-I Error Probability (long tail)			Type-I Error Probability (total)		
		Tukey	HVBP	SSSBB	Tukey	HVBP	SSSBB	Tukey	HVBP	SSSBB
25	0.57	0.0000	0.0021	0.0026	0.0125	0.0027	0.0134	0.0126	0.0049	0.0159
20	0.63	0.0000	0.0019	0.0020	0.0138	0.0026	0.0139	0.0138	0.0045	0.0159
15	0.73	0.0000	0.0014	0.0011	0.0158	0.0025	0.0147	0.0158	0.0038	0.0159
10	0.89	0.0000	0.0005	0.0002	0.0193	0.0022	0.0160	0.0193	0.0027	0.0162
5	1.26	0.0000	0.0000	0.0000	0.0280	0.0015	0.0188	0.0280	0.0015	0.0188
2	2	0.0000	0.0000	0.0000	0.0481	0.0005	0.0241	0.0481	0.0005	0.0241
Degree of freedom	Medcouple SS = 300	Lower Critical Values			Upper Critical Values			Interval Width		
		Tukey	HVBP	SSSBB	Tukey	HVBP	SSSBB	Tukey	HVBP	SSSBB
25	0.08	5.84	9.45	9.68	43.44	49.10	43.18	37.60	39.66	33.50
20	0.10	2.89	6.45	6.50	36.39	42.21	36.37	33.50	35.76	29.87
15	0.11	0.22	3.66	3.56	29.06	34.99	29.30	28.83	31.33	25.74
10	0.13	-1.98	1.29	1.03	21.27	27.47	21.83	23.25	26.18	20.80
5	0.20	-3.25	-0.31	-0.71	12.55	19.63	13.54	15.80	19.93	14.25
2	0.33	-2.72	-0.46	-0.74	6.07	15.12	7.45	8.79	15.59	8.20

6.2 Comparison of Techniques in Chi Square Distribution

Table 4 reports the information of moment measure of skewness, medcouple, probability of Type-I error (short tail, long tail), LCV, UCV and Interval width for the selected degree of freedom in chi square by Tukey, HVBP and SSSBB techniques according to their formulae. Looking at the Type-I error probability on short tail, Tukey's probability of Type-I error is zero and on long tail it is 1.25% in chi square (25). Same probability of Type-I error is 0 and 4.81% respectively in chi square (2). So total Type-I error probability is less than the level of tolerance which is 5 percent. Observing the short and long tailed probabilities for chi square (25) in HVBP is 0.21% and 0.27% respectively concluding the sum of probabilities is less the tolerance level of 5%. Similarly, for high skewed data in chi square (2) short and long tailed probabilities of Type-I error are 0 and 0.05% respectively resulting sum less than 5 percent. So HVBP fulfills criterion 1 for less and high skewed data. Considering SSSBB for least skewed chi square (25), on short and long tail probability of Type-I error is 0.26 and 1.34 whose sum is 1.59 percent which is less than 5 % level of tolerance. For high skewed data in chi square (2) probability of Type-I error is 0 and

2.41 percent respectively. So the total Type-I error probability is less than 5% and we conclude that all three techniques fulfill criterion of Type-I error probability less than 5%. Checking for the second criterion of interval width, Table 4 reports that the largest interval width by HVBP and smallest by SSSBB.

Figure 2: Comparison of Type I Error and Interval Width



If we check the interval width relatively, it may be observed that in chi square (25) interval width of HVBP is larger by 5.48% and 18.39% than Tukey and SSSBB respectively. For chi square (2) interval width of HVBP is 77.36% and 90.12% larger than Tukey and SSSBB respectively. Comparing Tukey and SSSBB interval width of

Tukey is 12.23% and 7.19% larger than SSSBB in chi (25) and chi square (2) respectively. Figure 2 reports the probability of Type-I error on short and long tail for techniques under comparison. This figure also clears that interval width of HVBP is always higher followed by Tukey technique.

Table 5. Type-I Error Probability, Interval Width, and Fences in Lognormal Distribution

Parameters	Moment measure of skewness	Type-I Error Probability (short tail)			Type-I Error Probability (long tail)			Type-I Error Probability (total)		
		Tukey	HVBP	SSSBB	Tukey	HVBP	SSSBB	Tukey	HVBP	SSSBB
(0, 0.2)	0.61	0.000	0.003	0.003	0.014	0.003	0.014	0.014	0.006	0.018
(0, 0.4)	1.32	0.000	0.002	0.000	0.029	0.003	0.021	0.029	0.005	0.021
(0, 0.6)	2.26	0.000	0.001	0.000	0.046	0.004	0.027	0.046	0.005	0.027
(0, 0.8)	3.69	0.000	0.000	0.000	0.062	0.005	0.033	0.062	0.005	0.033
(0, 1)	6.18	0.000	0.000	0.000	0.078	0.006	0.039	0.078	0.006	0.039
Parameters	Medcouple SS = 300	Lower Critical Values			Upper Critical Values			Interval Width		
		Tukey	HVBP	SSSBB	Tukey	HVBP	SSSBB	Tukey	HVBP	SSSBB
(0, 0.2)	0.09	0.47	0.57	0.58	1.55	1.72	1.55	1.08	1.14	0.97
(0, 0.4)	0.17	-0.06	0.32	0.26	2.13	2.95	2.26	2.18	2.63	2.00
(0, 0.6)	0.25	-0.58	0.15	0.01	2.75	4.92	3.17	3.33	4.77	3.15
(0, 0.8)	0.33	-1.12	0.04	-0.17	3.41	7.99	4.34	4.53	7.95	4.51
(0, 1)	0.40	-1.67	-0.04	-0.30	4.14	12.62	5.84	5.81	12.66	6.13

6.3 Comparison of Techniques in Lognormal Distribution

Table 5 reports the information of moment measure of skewness, medcouple, probability of Type-I error (short tail, long tail), LCV, UCV, and interval width for the selected parameters in lognormal distribution by Tukey, HVBP, and SSSBB techniques according to their formulae. First comparing Type-I error probability it may be observed that probability of Type-I error for all techniques. Tukey has 0% probability of Type-I on short tail of log normal distribution for all selected parameters. On long tail the minimum probability of Type-I error is 1.4% while maximum is 7.8%. Hence for highly skewed data Tukey technique failed to the meet first criterion of committing Type-I error less than five percent. The second technique HVBP has 0.3% maximum probability of Type-I error on short tail. On the long tail it also has a maximum of 0.6% probability of Type-I error so total probability of Type-I error is less than 5% and

HVBP passed this criterion. Considering SSSBB, its maximum probability of Type-I error is 0.3% and 1.4% on short tail and long tail respectively.

While comparing interval width we compare HVBP first with the remaining two as it has largest interval ever. It has 5.55% and 17.52% larger interval than Tukey and SSSBB respectively in lognormal (0, 0.2). In high skewed data like lognormal (0, 1) it has 118% and 107% larger interval than Tukey and SSSBB. Comparison of Tukey and SSSBB tells that in less skewed data, Tukey constructed a larger interval than SSSBB by 11% in lognormal (0, 0.2). Interestingly Tukey constructed a smaller interval than SSSBB in high skewed data. In lognormal (0, 1) interval width of Tukey is 5.5% smaller than SSSBB. The important point to note here is that for these specific parameters Tukey did not meet first criterion of probability of Type-I error less than five percent.

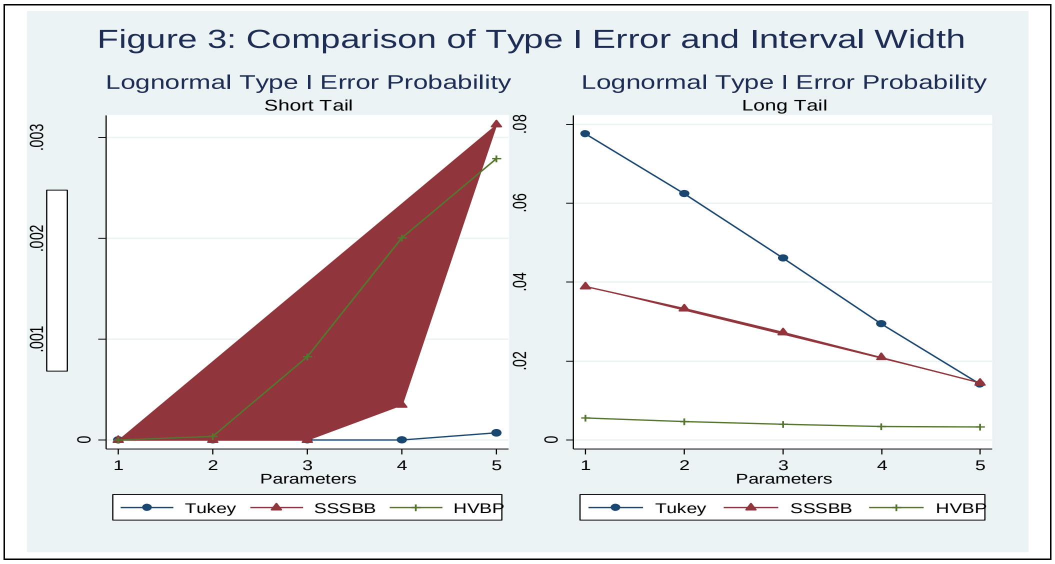
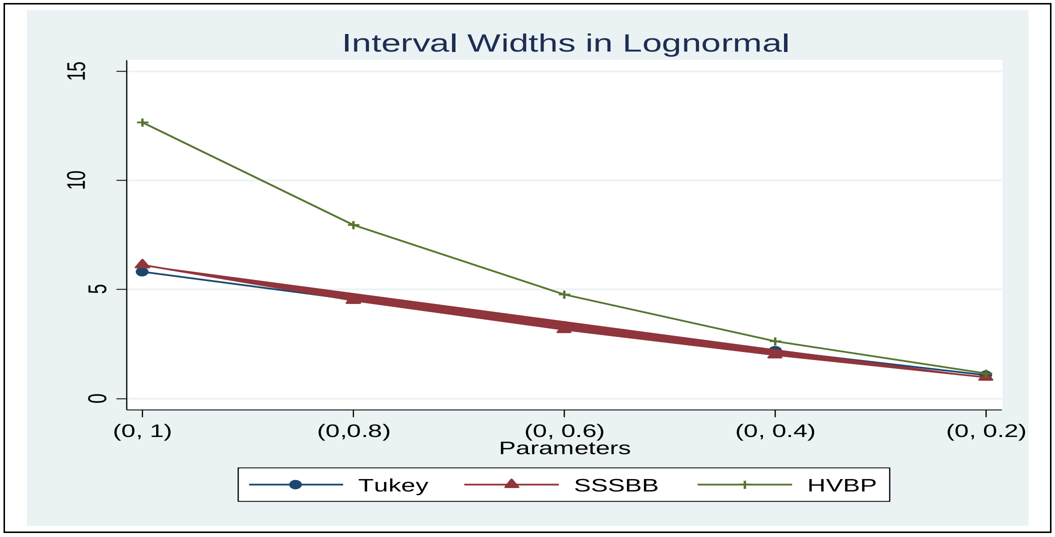


Figure 3: Comparison of Type I Error and Interval Width

DOI: 10.24818/18423264/54.3.20.17



7. Applications:

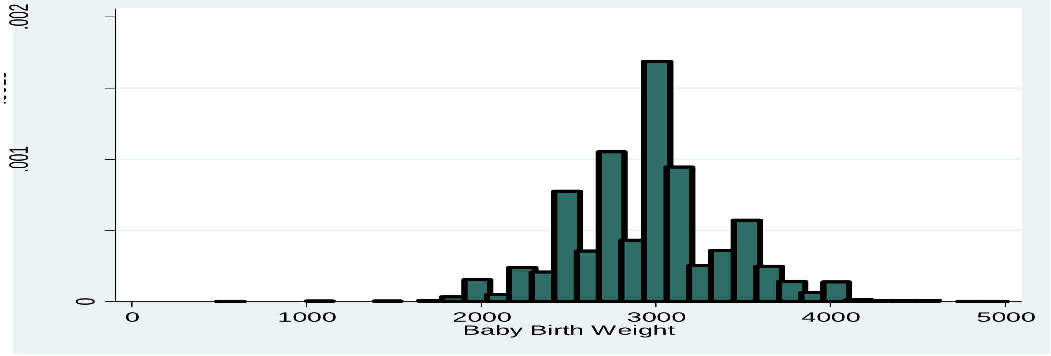
Data for baby birth weight has been taken from Agha Khan Hospital, Karachi. Here our assumption is that survivals of the babies depend upon their birth weight. So if a baby has joined this world with low birth weight he/she is more vulnerable to death as compared to a baby with higher birth weight. According to McIntire et al. (1999), infants born with low birth weight are more likely to die or succumb to morbidity. Vangen et.al (2002) proved that heavier is better. Babies with low birth weight, either due to short gestation period or because of fetal growth constraint, are at high risk for short- and long-term disabilities and death (Schieve et al, 2002). Checkup of very low birth weight children points toward increased deaths among all subpopulations.

There is a consensus over the role of socioeconomic conditions of the family and educational background, especially mother’s education, in the survival of the infant. Also medical facilities have improved to the extent that a baby with very low birth weight might survive by availing these facilities and a baby with relatively higher birth weight from low income family might not due to unavailability of medical facilities. But as it is mentioned above that the data have been taken from the similar income group (people approaching Agha khan hospital are well off and from educated families and can bear any cost in monetary terms for survival of their children). Agha Khan Hospital is one of the most efficient hospitals with latest facilities and equipment as compared with public sector hospitals. So it may be assumed that data belongs to a similar group with respect to

DOI: 10.24818/18423264/54.3.20.17

income and education and is comparable. Keeping remaining things constant, the probability of the survival increases as birth weight increases and vice versa.

Figure 4: Baby Birth Weight



Data consisting of 3613 observations of baby birth weight along with their follow up data till 4th week (28th day). Minimum weight is 500 grams and the highest weight is 5000 grams. Average weight is 2974 grams (nearly 3kg) and total deaths up to 4th week are nineteen. Mortality among the total population is just 0.5 percent. According to definition of low birth weight, an infant having weight less than 2500 grams is treated as low weight. This data itself proves the claim that low birth weight babies have more chances of mortality, as it can be observed that Tukey's technique has detected 26 observations as left outliers while proposed technique SSSBB has detected 16 observations on left side as outlier. By mining into data it was observed that there are five deaths in both cases (either in Tukey's or SSSBB). So it can be said that just 0.7% data (by Tukey's technique) and 0.4% data (by SSSBB) captures more than 25% of the deaths from the whole data set. This finding corroborates the claim that birth weight has a very close relation with mortality; secondly it shows the improvement of the proposed test on Tukey's as Tukey's technique detected the same number of deaths from 0.7% of the data while SSSBB from 0.4 percent. That is, newly introduced test is performing more efficiently than Tukey's does.

7.1 Comparison of Tukey's Technique, HVBP, and SSSBB in Baby Birth Weight Data

According to the assumption that birth weight has close relation with the survival, the babies with higher birth weights are more likely to survive than low birth weight babies. For this purpose, we should compare the left outliers (babies vulnerable for death due to low weight) for the mortality.

DOI: 10.24818/18423264/54.3.20.17

Left outliers detected by Tukey's technique are 26 while left outliers detected by SSSBB are 16. By analyzing the data with respect to left outliers we see that there are 5 deaths in both cases (in 26 left outliers by Tukey and 16 left outliers by SSSBB). So we can say that performance of Tukey's technique is 19.23% while the performance of SSSBB is 31.25%. Technically speaking Tukey technique is attempting type I error as it is detecting real observations as outliers. It can also be observed that deaths are also inliers so we compare the total number of deaths with respect to total number of outliers detected. Tukey's has detected 111 outliers in total while SSSBB has detected 29 outliers so the performance of Tukey's as a whole is 17% while performance of SSSBB is 66 percent.

Table 6. Outliers and IW for all Techniques in BBW* Data

Technique	Left OL	Right OL	Total OL	LCV	UCV	Interval Width
Tukey	26	85	111	1950	3950	2000
SSSBB	16	13	29	1900	4250	2350
HVBP	5	180	185	1621.10	3745.6	2124.5

*BBW: Baby Birth Weight

Table 7. Performance Comparison in BBW* Data

Technique	Left OL	Performance left outliers	Overall Performance
TUKEY	26	19.23%	17.12%
SSSBB	16	31.25%	65.52%
HVBP	5	40.00%	10.27%

*

BBW: Baby Birth Weight

In comparison of all techniques under consideration we see that HVBP is leading all the techniques under comparison by detecting just 5 left outliers and two deaths in these 5 outliers performing at 40% while SSSBB seems to chase it by 31% performance. Since deaths are also inliers so looking at total outlier's performance we see that HVBP have detected 180 right outliers and its performance falls drastically at 10.27% while SSSBB improves its performance from 31.25% to 65.52% by just detecting 13 outliers on the right side leading all the techniques.

8. Discussion and Conclusion

It may be observed that HVBP has the least probability of Type-I error in all selected distributions followed by SSSBB. Tukey technique has generally the least probability of Type-I error on short tail while it has more probability of Type-I error on long tail. It was also observed that in extreme skewed distribution Tukey technique constructs misclassified interval and it may also commit probability of Type-I error more than 5 percent. Split sample skewness based boxplot has less total probability of Type-I error than Tukey but it has more probability of Type-I error than HVBP. On the other hand, we observe that Tukey generally constructs medium interval except one case under consideration. Interval width of HVBP is higher in all cases of distributions under consideration. Even sometimes it is more than double than Tukey's and SSSBB's interval width. A large interval width means covering more area under the curve and reducing precision. So it may be inferred from above discussion that HVBP commits Type-I error less but constructs larger interval. Tukey technique sometimes commits Type-I error more than 5% and is condemnable with respect it. Also, its interval size is larger than SSSBB. On the other hand, we see that SSSBB constructs smaller interval and have Type-I error probability less than 5% in all cases. Figure 3 also shows the interval width of HVBP is ever larger followed by Tukey technique. Just for lognormal (0, 1) interval width of SSSBB is higher than Tukey's.

9. Recommendations

From the above discussion it is recommended that although Tukey's technique is bit easier than SSSBB but it fails to detect outlier in skewed distribution, hence for the safe side SSSBB should be used in skewed distributions. With certainty of distribution being symmetric, Tukey's technique should be used. In case of highly skewed distribution and with interest on the short tail, HVBP may be used having low probability of Type-I error without care of precision.

REFERENCES

[1] **Banerjee, S. & Iglewicz, B. (2007),** ***A Simple Univariate Outlier Identification Procedure Designed for Large Samples***. *Communications in Statistics- Simulation and Computation, 36*, 249-263;
[2] **Chen, Y., Miao, D. & Zhang, H. (2010),** ***Neighbourhood Outlier Detection***. *Expert Systems with Applications, 37* (12);
[3] **Efstathiou, C. E. (2006),** ***Estimation of Type I Error Probability from Experimental Dixon's "Q" Parameter on Testing for Outliers within Small Size Data Sets.****Talanta, 69,* 1068-1071;

DOI: 10.24818/18423264/54.3.20.17

[4] **G. Brys, M. H. (2004),** ***A Robust Measure of Skewness.*** *Journal of Computational and Graphical Statistics, 13* (4), 996-1017;

[5] **Hubert, M. & Veeken, S. V. (2007),** ***Outlier Detection for Skewed Data***. Katholieke Universiteit Leuven, DEPARTMENT OF MATHEMATICS. Technical Report;

[6] **Iglewicz, B. & Hoaglin, D. C. (1993),** ***How to Detect and Handle Outliers***. 16, Wisconsin: ASQC Quality Press;

[7] **Justel, A. & Pena, D. (1996),** ***Gibbs Sampling Will Fail in Outlier Problems with Strong Masking.****Journal of Computational and Graphical Statistics,, 5* (2), 176-189;

[8] **Kimber, A. C. (1990),** ***Exploratory Data Analysis for Possibly Censored Data from Skewed Distributions***. *Applied Statistics, 39* (1), 21-30;

[8] **Mansur, M. O. & Sap, M. N. (2005),** ***Outlier Detection Technique in Data Mining***. *Postgraduate Annual Research Seminar;*

[9] **McIntire, D. D., Bloom, S. L., Casey, B. M. & Leveno, K. J. (1999),** ***Birth Weight in Relation to Morbidity and Mortality Among Newborn Infants.****The New England Journal of Medicine, 340* (16), 1234-1238;

[10] **Schieve, L. A., Meikle, S. F., Ferre, C., Peterson, H. B., Jeng, G. & Wilcox, L. S. (2002),** ***Low and Very Low Birth Weight in Infants Conceived With Use of Assisted Reproductive Technology.*** *The New England Journal of Medicine, 346* (10), 731-737;

[11] **Tsay, R. S., Pena, D. & Pankratz, A. E. (2000),** ***Outliers in Multivariate Time Series.*** *Biometrika, 87* (4), 789-804;

[12] **Tukey, J. W. (1977),** ***Exploratory Data Analysis.*** *Addison-Wesely;*

[13] **Vangen, S., Stoltenberg, C., Skjaevern, R., Magnus, P., Harris, J. R. & Stray-Pedersen, B. (2002),** ***The Heaiver the Better: Birth Weight and Perinatal Mortality Different Ethnic Groups.****International Journal of Epidemiology, 31*, 654-660;

[14] **Zaman, A., Rousseeuw, P. J. & Orhan, M. (2001),** ***Econometric Applications of High-Breakdown Robust Regression Techniques.*** *Economics Letters, 71*, 1–8.

DOI: 10.24818/18423264/54.3.20.17