Stochastic Modeling and Applications
Vol. 25 No. 2 (July-December, 2021)
ISSN: 0972-3641

Received: 19th June 2021 *Revised: 09th October 2021* *Selected: 23rd October 2021*

ANALYZING JOHNSON S_B DISTRIBUTION FOR DETECTION OF A PAIR OF UPPER OUTLIERS

TANUJA SRIWASTAVA[1], SHRUTI[2], SHISHIR KUMAR JHA[3] & SANDEEP MISHRA[4]

Abstract

Outlier can occur due to the chance error for any distribution. Test statistic for detection of outlier for Johnson S_B distribution is rarely available. An attempt has been made to develop a test statistic for the detection of a pair of upper outliers for a sample from a Johnson S_B distribution under the assumption that parameters are known. A simulation technique is carried out for obtaining the critical value of the test statistic. Some examples have also been constructed for highlighting the utility of the statistic.

Keywords: Johnson S_B distribution, Outliers, Slippage alternative, Critical region, Simulation technique.

1. Introduction

Johnson S_B distribution is a four-parameter distribution first described by N. L. Johnson in his classic paper Johnson (1949). In this paper, he considered transformations on random variables using method of translation that led to Normal distributions. This is popularly known as Johnson family of distributions with flexibleness of extensive coverage of distributional shapes.

Because of pliable character, S_B distribution is applied extensively e.g. model for human exposure data as described by Flynn (2004). Similarly, Mage (1980), Kotteguda (1987), Zang et.al. (2003) and Konduru et.al. (2013) used this distribution for air pollution, rainfall distribution, forestry study and for Infrared brightness study of the convective clouds respectively. Further, this distribution can be used in microarray data analysis as studied by Florence George (2007). Hafley and Schreuder (1977), Von Gadow (1983), Kiviste et al. (2003) and Parresol (2003) supported the use of this distribution to describe diameter distributions. Similarly, Hafley and Schreuder (1977) used S_B distribution for the description of height distributions.

Any observation which deviates from the rest of the data set in some sense is called an outlier. In the line of Sriwastava (2018), a test statistic is evolved for identification of an upper outlier pair assuming that all parameters are known. It

is done by considering a test statistic as given by Irwin (1925) using a sample from a Standard Normal distribution.

2. Johnson S_B Distribution and Test Statistics

Let $X_1, X_2, \cdots, X_n$ be the n observations from a Johnson S_B distribution. Let$X_{(1)}, X_{(2)}, \cdots, X_{(n)}$ be the order statistics of these observations.

Therefore, for Johnson S_B distribution

$$f(x) = \frac{\delta}{\sqrt{2\pi}} \frac{\lambda}{\{\lambda-(x-\xi)\}(x-\xi)} exp\left[-\frac{1}{2}\left\{\gamma + \delta \ln\left(\frac{x-\xi}{\lambda-(x-\xi)}\right)\right\}^2\right], \quad (2.1)$$

(2.1)

$$\xi \leq x \leq \xi + \lambda,\ \delta > 0,\ -\infty < \gamma < \infty,\ \lambda > 0,\ -\infty < \xi < \infty$$

where,ξ and λ are location and scale parameters respectively; δ and γ are shape parameters.

By making a transformation

$$z = \gamma + \delta \ln\left(\frac{x-\xi}{\lambda-(x-\xi)}\right) \quad , \quad (2.2)$$

the probability density function (2.1) transforms to a Standard Normal distribution

$$f(z) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}z^2\right], -\infty < z < \infty.$$

Let $z_1, z_2, \cdots, z_n$ ben observations of Normal distribution corresponding to the n observations $X_1, X_2, \cdots, X_n$. Further, let$z_{(1)}, z_{(2)}, z_{(3)} \cdots, z_{(n)}$is the order statistics of the n transformed observations. Note that, transformation (2.2) is order preserving.

Let H_0 (null hypothesis): there is no outlying observations in the sample. H_1 (alternative hypothesis): there is a pair of outlying observations on the right side. The test statistic for testing the presence of an upper outlier pair of a sample (from Normal distribution with σknown) is given by Irwin (1925) as

$$T = \frac{z_{(n-1)} - z_{(n-2)}}{\sigma}. \quad (2.3)$$

$T > T_\alpha$ is the α level critical region forT_α, where $T_\alpha$is the critical value of statistic T and its tabulated value is given in Barnett and Lewis (1994) at 5% and 1% level of significance.

When the transformed values of Johnson S_B distribution using transformation (2.2) was used in (2.3), it becomes,

$$T = z_{(n-1)} - z_{(n-2)}, \text{ as} \sigma = 1. \quad (2.4)$$

As the transformation (2.2) is order preserving, we have

$z_{(n-1)} = \gamma + \delta \ln\left(\frac{x_{(n-1)}-\xi}{\lambda-(x_{(n-1)}-\xi)}\right)$ and $z_{(n-2)} = \gamma + \delta \ln\left(\frac{x_{(n-2)}-\xi}{\lambda-(x_{(n-2)}-\xi)}\right)$.

Thus, $T = \gamma + \delta \ln\left(\frac{x_{(n-1)}-\xi}{\lambda-(x_{(n-1)}-\xi)}\right) - \gamma - \delta \ln\left(\frac{x_{(n-2)}-\xi}{\lambda-(x_{(n-2)}-\xi)}\right)$

$= \delta\left[\ln\left(\frac{x_{(n-1)}-\xi}{\lambda-(x_{(n-1)}-\xi)}\right) - \ln\left(\frac{x_{(n-2)}-\xi}{\lambda-(x_{(n-2)}-\xi)}\right)\right]$.

or, $T = \left[\ln\left\{\left(\frac{x_{(n-1)}-\xi}{x_{(n-2)}-\xi}\right)\left(\frac{\lambda-(x_{(n-2)}-\xi)}{\lambda-(x_{(n-1)}-\xi)}\right)\right\}\right]$.

Let, $x_{(n-1)} - \xi = y_{(n-1)}$, then T can be rewritten as

$T = \delta\left[\ln\left\{\frac{y_{(n-1)}}{y_{(n-2)}}\left(\frac{\lambda-y_{(n-2)}}{\lambda-y_{(n-1)}}\right)\right\}\right] = \left[\ln\left\{\frac{y_{(n-1)}}{y_{(n-2)}}\left(\frac{\lambda-y_{(n-2)}}{\lambda-y_{(n-1)}}\right)\right\}^{\delta}\right]$.

Then on taking exponential on both the sides, a statistic W which is suitable for variables from a Johnson S_B distribution can be obtained as

$W = e^T = \left\{\frac{y_{(n-1)}}{y_{(n-2)}}\left(\frac{\lambda-y_{(n-2)}}{\lambda-y_{(n-1)}}\right)\right\}^{\delta}$, $\{0 < T < \infty\}$.

or, $W = \left\{\left(\frac{x_{(n-1)}-\xi}{x_{(n-2)}-\xi}\right)\left(\frac{\lambda-(x_{(n-2)}-\xi)}{\lambda-(x_{(n-1)}-\xi)}\right)\right\}^{\delta}$, $1 < W < \infty$. (2.5)

Corresponding to α size critical region $T > T_\alpha$ of the test statistic T, $W > W_\alpha$ will be the α size critical region of the test statistic W.

The critical values W_α are tabulated in Table 2.1 at 5% and 1% levels of significance and for different values of n between 10 and 1000. As the critical values W_α were derived from the critical values T_α, they are independent of the parameters $(\xi, \lambda, \gamma, \delta)$. Thereby, they can be used for any Johnson S_B distribution.

Table 2.1
Critical Value of W_α for Different Sample Sizes

$W\alpha$					
N	α=0.05	α=0.01	n	α=0.05	α=0.01
10	2.6117	3.9749	80	1.8404	2.4596
20	2.2034	3.1268	100	1.786	2.3632
30	2.0751	2.8577	200	1.716	2.2479
40	1.9739	2.7183	500	1.616	2.0751
60	1.8776	2.5345	1000	1.5683	1.9542

It can be inferred from the table that the critical values of the test statistic W decreases with increase in sample size. This is because, higher the number of observations, closer will be the distance between them.

3. Examples

The utility of the statistic developed were verified with the following real-life data taken from Govt. of India census data 2011 and outlying observations were planted.

1028610, 1045547, 1062388, 1095722, 1112186, 1128521, 1160813, 1176742, 1192506, 1223581, 1238887, 1254019, 1283600, 1298041, 1312240, 1339741, 1352695, 1365302, 1388994, 1399838.

This is used in the following examples. An outlying observation was introduced by deviating each of the parameters as follows.

*Example.*3.1: For the purpose of introduction of an outlier, another sample with different location parameter$\xi + a\lambda$, where 0 <a < 1 with a as 0.5 of Johnson $S_B$distribution was generated (This is because the generated observations may lie beyond the range of the original variable). The largest observation of the first sample was replaced with the largest observation of the second sample. Similarly, the second largest observation of the first sample was replaced by the second largest sample of the second sample. Test statistic W was calculated, and it is equal to 3.30715, hence the null hypothesis is rejected at 5 percent level of significance. So, we conclude that the two replaced observations (largest and second largest) are the contaminant observations present in the sample.

*Example.*3.2: Again, for introduction of an outlier, a sample of Johnson S_B distribution was generated by shifting the scale parameter$b\lambda$, where ξ/λ<b< $1+\xi/\lambda$with b=1.5, (This is because the generated observations may lie outside the range of the original variable). The highest valued and the next highest valued observations of the first sample were replaced with these two observations of the second sample. Test statistic W was calculated and found to be equal to12.3665 and lying in the critical region at 5 percent level of significance. It shows that there are some outlying observations in the sample.

*Example.*3.3: As per previous examples, for introducing an outlying unit a sample from Johnson S_B distribution by shifting the shape parameter $c\gamma$, where $-(\xi + \lambda) < c < (\xi + \lambda)$ with $c = 0.75$ was generated (This is because beyond this range, it becomes infinity). The observation with highest value and the next largest observation of the sample one was replaced with these two observations of the sample two. Test statistic W was calculated and found to be equal to7.816926at 5% level of significance. On comparing with the critical value for a sample of size 20, the null hypothesis gets rejected. This shows that suspected observations are the outlying observations.

Example.3.4: After introduction of an outlying observation a new sample was generated with a shift in the shape parameter $d\delta$, d> 0; for this example, d was

taken as 0.5. The largest unit of the sample one was replaced with the largest unit of sample two. Similar exchange was done for the next largest value of the sample one with the corresponding unit of the second sample. Calculated value *W* was obtained as 2.635318 and the null hypothesis gets rejected at 5 percent significance level. This infers that the outlying observations are present in the sample.

Example.3.5: Now, for an outlying observation with a shift in all the four parameters, again for introducing an outlier, another sample of a Johnson S_B distribution having parameters$\xi + a\lambda$ (location parameter), $b\lambda$ (scale parameter) with shape parameters $c\gamma$ and $d\delta$, where $0 < a < 1, {}^{\xi}/_{\lambda} < b < 1 + {}^{\xi}/_{\lambda}, -(\xi + \lambda) < c < (\xi + \lambda)$, $d > 0$ with a=0.5,b=1.5, c=0.75 and d=0.5 was generated. The greatest valued unit of sample 1 was replaced with the greatest valued unit and the sample 2. The similar replacement was done for second greatest valued unit of the first sample with the second sample. Test statistic *W* was calculated and obtained as5.29684 and found to be lying in the critical region, which infers the presence of outlying observations in the sample.

4. Performance Study

For a slippage alternative of the location parameter, using R software a random sample of size n was generated from Johnson S_B distribution with location parameter ξ(=20), scale parameter λ(=10), with two shape parameters γ(=1) and δ(=2) (known). For introduction of a contaminant observation, a new sample was generated with a shift in the location parameter,$a\lambda$, where $0 < a < 1$. There by, there is no need of outlier detection tests for identification of contaminant observations. The largest and the second largest observations of the first sample were replaced with the corresponding observations of the second sample and hence, a new sample was formed.

Table 4.1

Probability of Identification of the Contaminant Observation with a Shift in Location Parameter

n \ a	0.2	0.3	0.4	0.5	0.6	0.8	1
10	0.9481	0.9984	1.0000	1.0000	0.9997	0.9995	0.9995
20	0.9883	0.9980	1.0000	1.0000	0.9995	0.9991	0.9980
30	0.9946	0.9970	1.0000	1.0000	0.9990	0.9987	0.9964
60	0.9991	0.9985	1.0000	1.0000	0.9991	0.9993	0.9710
100	0.9998	0.9999	1.0000	1.0000	0.9992	0.9994	0.9025
200	1.0000	1.0000	1.0000	1.0000	0.9995	0.9713	0.6154
500	1.0000	1.0000	1.0000	1.0000	0.9992	0.7498	0.0895
1000	1.0000	1.0000	1.0000	1.0000	0.9878	0.3000	0.0011

Then, the test statistic *W* was calculated at α level of significance and compared with the corresponding critical value. The probability of rejection (the ratio of total number of times the planted observations were detected to be outlying to the total number of repetitions) of the null hypothesis was calculated.

This study was carried out for different sample sizes and the simulation was done 10,000 times. The calculated value of the probability of identification of contaminant observation is shown in the Table 4.1.

The findings of Table 4.1 shows that the test statistic is reasonably performing well for different sample sizes up to *a*=0.5, beyond that it starts declining. This is due to the fact that larger values of the location parameter leading to a smaller value of the statistic *W*, as observed from (2.5).

Similarly, for slippage alternative in the scale parameter a sample was generated. After introduction of an outlying observation a new sample was generated with a shift in the scale parameter $b\lambda$, such that $\xi/\lambda < b < 1+\xi/\lambda$. This is because the generated observations may lie outside the range of the original variable, therefore, it becomes evident, and no outlier detection test would be needed for their identification. The largest and the second largest observations of sample 1 was replaced with the corresponding observations of the sample 2 and a new sample was created. Then, *W* (test statistic) was obtained at an appropriate significance level. Using simulation technique probability of rejection of H_0 was calculated. The Probability of rejectionwhich is the probability of identification of the outlying observation is then the ratio of number of outcomes to number of trials. This process was simulated 10,000 times for different sample sizes. The probability of identification of outlying observation is shown in Table 4.2.

Table 4.2

Probability of Identification of the Contaminant Observation with a Shift in Scale Parameter

n \ *b*	1.2	1.3	1.4	1.5	1.6	1.8	2
10	0.6083	0.7956	0.9088	0.9624	0.9874	0.9988	0.9999
20	0.7495	0.9235	0.9819	0.9964	0.9998	1.0000	0.9999
30	0.8148	0.9570	0.9947	0.9992	0.9999	1.0000	0.9999
60	0.9146	0.9916	0.9996	0.9998	1.0000	1.0000	0.9999
100	0.9598	0.9982	1.0000	0.9997	1.0000	1.0000	0.9999
200	0.9878	0.9999	1.0000	1.0000	1.0000	1.0000	0.9515
500	0.9992	1.0000	1.0000	1.0000	1.0000	1.0000	0.9390
1000	0.9997	1.0000	1.0000	1.0000	1.0000	1.0000	0.6557

As per Table 4.2, it can be inferred that the test statistic is performing well for different sample sizes up to *b*= 1.8; beyond that it starts declining slightly.

5. Comparison with other existing procedures

There exist several outlier detection methods, such as Box plot, Histogram, Scatter diagram and Normal probability plots. These methods graphically represent the outliers. In this paper, a mathematical approach was considered. It seems desirable to mention one criterion to compare the result. Here, Box plot method is considered. Considering the same data set as that was taken in earlier examples and using identify command of R software draw a box plot without introducing an outlier. This is shown in Figure 1 as given below:

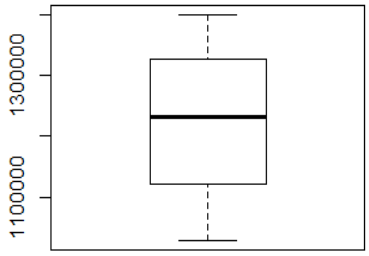

Fig.5.1. Box Plot

After introducing a pair of upper outliers *i.e.* the largest and the second largest observations with a shift in all the four parameters, the box plot comes out to be as follows.

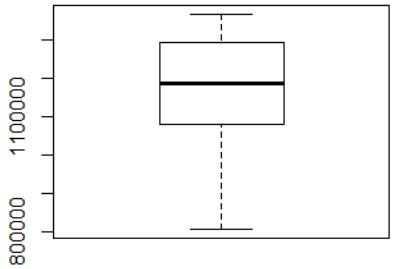

Fig.5.2. Box Plot after introducing outliers

Figure 2 shows that box plot did not identify any outlier. It indicates that the proposed method is more reliable than the earlier existing procedures.

6. Conclusion

For different sample sizes the proposed test statistic is performing well for population data as well as data coming from Johnson S_B distribution. When

parameters are not known then the parameters may be estimated and the test statistic is constructed. Then, outliers are detected in a sample from a Johnson S_B distribution. Further, its critical values may be analyzed using simulation technique.

7. References

Barnett V., Lewis T. (1994). *Outliers in Statistical Data. John Wiley*.

Flynn, M.R. (2004). *The 4-parameter lognormal (S_B) model of human exposure. Ann Occup Hyg:* 48: 617-22.

Gerald H. J. and Samuel S. S. (1967). *Statistical model in Engineering*, *John Wiley and sons*.

George, F. (2007). *Johnson's System of Distribution and Microarray Data Analysis*. *Graduate Theses and Dissertations*, *University of South Florida*.

Hafley, W.L., and M.T. Schreuder. (1977). *Statistical distributions for fitting diameter and height data in even-aged stands. Can. J. For. Res*. 7:481–487.

Irwin, J. O. (1925). *On a criterion for the rejection of outlying observations. Biometrika*, 17, 239-250. (31, 95, 249, 250).

Johnson, N.L. (1949). *Systems of frequency curves generated by methods of translation*. *Biometrica*, 58, 547.558.

Kiviste, A., A. Nilson, M. Hordo, AND M. Merenakk. (2003*). Diameter distribution models and height-diameter equations for Estonian forest. in Modelling Forest Systems. Amaro, A., D. Reed, and P. Soares (eds.). CABI Publishing, Wallingford, UK*.P. 169–179

Konduru, R.T., Kishtawal, C.M. and Shah, S. (2013). *A new perspective on the infrared brightness temprative distribution of the deep convective clouds*. *Springer*.

Kottegoda, N.T. (1987). *Fitting Johnson S_B curve by method of maximum likelihood to annual maximum daily rainfalls*. *Water Resour Res*:23: 728-732.

Mage, D.T. (1980). *An Explicit Solution for S_B Parameters Using Four Percentile Points*, *Technometrics*, 22, 247-251.

Paressol, B.R. (2003). *Recovering parameters of Johnson's SB distribution. US For. Ser. Res. Paper SRS*-31. 9 p.

Sriwastava, T. (2018). *An upper outlier detection procedure in a sample from a Johnson SB distribution with known parameters. International Journal of Statistics and Applied Mathematics,* 3(2), 194-198.

Von Gadow, K. (1983). *Fitting distributions in Pinus patula stands. Suid-Afrik. Bosboutydskrif*. 126:20–29.

Zhang L., Packard P.C., and Liu C. (2003). *A comparison of estimation methods for fitting Weibull and Johnson's S_B distributions to mixed spruce-fir stands in northeastern North America*. *Can J Forest Res*: 33: 1340–1347.

TANUJA SRIWASTAVA: Department of Statistics, Sri Venkateswara College, University of Delhi, Delhi
e-mail: tanuja@svc.ac.in

SHRUTI: School of Sciences, U. P. Rajarshi Tandon Open University, Prayagraj
e-mail: drshruti.uprtu@gmail.com

SHISHIR KUMAR JHA: Department of Statistics, Ramjas College, University of Delhi, Delhi
e-mail: skjha_statistics@ramjas.du.ac.in

SANDEEP MISHRA: Department of Statistics, Shaheed Rajguru College of Applied Sciences for Women, University of Delhi, Delhi
e-mail: sandeepstat24@gmail.com