Available online at www.scholarsresearchlibrary.com

Scholars Research Library

Annals of Biological Research, 2014, 5 (12):16-20
(http://scholarsresearchlibrary.com/archive.html)

Annals of Biological Research
Scholars Research Library
ISSN 0976-1233
CODEN (USA): ABRNBW

Analysis strategy for comparison of skewed outcomes from biological data: A recent development

Devika Shanmugasundaram[1], L. Jeyaseelan[1*], Sebastian George[2] and Geevar Zachariah[3]

[1]Department of Biostatistics, Christian Medical College, Vellore, India
[2]Department of Statistics, St. Thomas College, Palai, Kerala, India
[3]Chief Cardiologist, Mother Hospital, Thrissur, Kerala, India

---

ABSTRACT

When faced with the problem of comparing positively skewed outcome values, data transformations such as log and square root etc, are often used. However, this approach suffers with the difficulty in interpretability, lack of accuracy etc. That is, while the back transformation of mean is possible, but not for the standard deviation. This paper presents the analysis of comparing positively skewed outcome data by using generalized pivotal and log transformation approach for lognormally distributed data. Simulation experiment was conducted to examine the characteristics of generalized pivotal approach for small sample sizes and for large standard deviations. For the analysis of positively skewed biological data between two groups generalized p value and confidence interval approach for lognormal distribution is considered to be efficient as this provides direct statistical inference such as estimates, 95% CI and its p values.

Key words: Positively skewed distribution, normal distribution, lognormal distribution, log transformation, generalized p value, generalized confidence intervals.

INTRODUCTION

In many biomedical studies, researchers are interested in estimating the difference of two sample means. One of the ways to test the above hypothesis is by doing an Independent sample t-test. However, two sample t-test approach is appropriate only if the observations are normally distributed. Many biological variables such as triglyceride levels, skinfold thickness, serum bilirubin levels etc., which are encountered in medical research are positively skewed. Data transformations are frequently used effectively in normalizing the data. Bland and Altman suggested that logarithmic transformation is frequently used for skewed outcomes as this gives nearly normal distribution [1]. Basically, the analysis is performed on the transformed scale, which can then be back transformed to the original scale. However, this will not lead to a reasonable estimate on the original scale as back transformation results in geometric mean of the original data rather than the arithmetic mean [2].

In vaccine and immunogenicity studies, the antibody titre values are log transformed and the results are summarized in terms of geometric mean titre or geometric mean ratio [3]. As such, the antilog of arithmetic means computed on log scale (geometric mean) is readily interpretable, but there is no straightforward interpretation available for the antilog of the standard deviation of the logged values [4]. Consider the example given in [5], the mean (SD) of triglyceride values of original data was 0.51 (0.22) mmol/l. The mean (SD) of the log transformed data was -0.33 (0.17). The back transformation of the mean on the log scale leads to 0.47 mmol/l which is geometric mean but the standard deviation on the log scale cannot be back transformed. Also the confidence interval for the mean in the original data cannot be regained back from the confidence interval for the mean of logged data [6]. To avoid these

circumstances, researchers make inference in terms of the transformed scale itself; however, the interest is in the means of the original data. Manning and Manning and Mullahy discussed the issues of transforming skewed outcomes, such as interpretability, lack of accuracy, and inefficiency [7, 8]

In this paper we disseminate the strategy and method of handling positively skewed biological outcomes and testing the significance of mean difference of positively skewed outcome - triglycerides among males and females based on generalized pivotal approach [9]. We also examine the characteristics of this approach for small sample sizes using a simulation study.

MATERIALS AND METHODS

Data:
The data used in this paper is a cross sectional data collected by Cardiological Society of India, Kerala (CSI Kerala, CRP study). The main objectives of the study were to find the prevalence and the risk factors of CAD among men and women aged 20 to 79 years in urban and rural Kerala [10]. For illustrative purpose, we have randomly chosen 200 triglyceride values from this data which are usually positively skewed and compared them among males and females.

Generalized P value and Confidence Interval:
The generalized p value and confidence interval was introduced for testing the hypothesis which involves the presence of nuisance parameters [11, 12]. For example, for testing the difference in the means of two exponential distributions or testing the difference in the means of two lognormal or in inverse Gaussian distributions, the means of these distributions involves the nuisance parameters. The details about the generalized p value and confidence interval approach for lognormal distribution can be found in [9, 13]. Details of R-software codes for calculation of difference of these means assuming lognormal distribution is given in Appendix 1.

Simulation study:
We conducted a simulation experiment to study the coverage probability of generalized confidence interval approach for small to large sample sizes. We arbitrarily chose various parameter values of the lognormal distribution and created 5000 data sets, each with different sample of sizes; n1 = n2 = (5,20,50,100) from lognormal distribution by using different parameter values of mean μ_1 = 5 and μ_2 = 7 (in log scale) and standard deviations σ_1 = (0.5,0.6,0.8) and σ_2 = (0.8,0.7,0.9) (in log scale). We used Monte Carlo method to obtain the 95% coverage probability of generalized confidence interval approach. We conducted the experiment by using R statistical software [14].

RESULTS

In CSI Kerala CRP study, triglyceride dataset consists of 200 observations, of which 100 (50%) were males and remaining were females. For original data, the mean (SD) triglyceride level of males was 153.93 (99.83) mg/dL while for females it was 110.01 (52.40) mg/dL. Median (IQR) triglyceride for males and females were 136.5 (88.75) and 98.0 (60.25) mg/dL respectively. By rule of thumb the standard deviation being more than half of the mean indicates that the values were not normally distributed [15]. Secondly, since the triglyceride values were positive random variables we checked for the assumption of normality after log transformation. The histogram of the original triglyceride values separately for males and females are presented in Figure 1.

Table 1: Comparison of log transformation and Generalized Pivotal approach for the difference of the mean triglyceride values

	Male (n = 100)		Female (n = 100)				
	Mean	**SD**	**Mean**	**SD**	**Difference**	**P value**	**95% CI**
Triglyceride:							
Original Data	153.93	99.83	110.01	52.40			
Log Transformed data	4.89	0.52	4.61	0.42	0.28	<0.0001	(0.15, 0.42)
Back Transformed (Taking Exponential)	132.95	-	100.48	-	1.32		(1.16, 1.52)
Generalized Pivotal approach	152.45	85.14	109.56	48.30	42.89	0.0000	(24.55, 63.70)

We carried out Shapiro-Wilk normality test for testing the normality of each of the log transformed data. The test yielded the p-value 0.1582 and 0.1164 for the male and female respectively. This indicates that the log transformed data follows normal distribution. We also performed Quantile-Quantile (Q-Q) plot (not presented here) of log transformed data to check for the normality assumption. Both the methods do not provide any evidence against normality. If the log data follows normal distribution it implies that the original data follows lognormal distribution. The statistical inference based on both these methods (log transformation & generalized pivotal approach) are presented in Table 1.

Figure 1: Histograms showing the distribution of triglyceride values among males and females

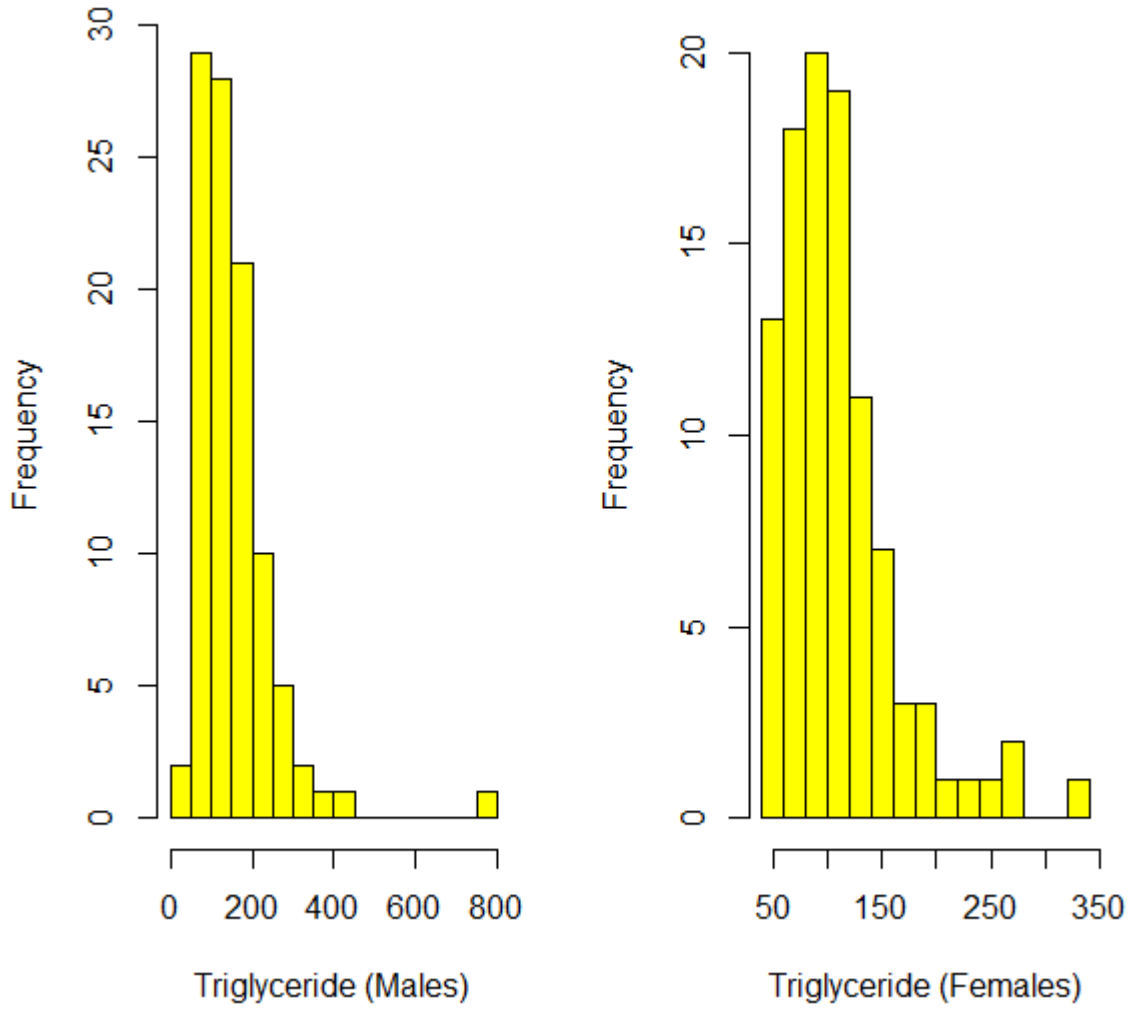

Table 2: Results from simulation study for the difference of two lognormal means - Coverage Probability of 95% C.I

n_1	n_2	μ_1	μ_2	σ_1	σ_2	Cov. Prob of 95% CI
5	5	5	7	0.5	0.8	94.88
				0.6	0.7	95.60
				0.8	0.9	95.76
20	20	5	7	0.5	0.8	94.86
				0.6	0.7	94.86
				0.8	0.9	94.74
50	50	5	7	0.5	0.8	95.26
				0.6	0.7	95.26
				0.8	0.9	95.16
100	100	5	7	0.5	0.8	94.82
				0.6	0.7	94.68
				0.8	0.9	94.74

The mean of the log transformed data for males is 4.89 and the standard deviation is 0.52 and for females it is 4.61 and 0.42 respectively. If we take means on the transformed scale and back transform by taking exponentiation that results in 132.95 mg/dL for males and 100.48 mg/dL for females. However, the back transformation of their standard deviations on the log scale does not make any sensible interpretation. Hence the variability of the estimates may not be obtained by back transformation. In addition to that, the difference in the means of triglycerides among males and females of the back transformed values (132.95-100.48=32.47 mg/dL) is not equal to the back transformation of the difference in the log scale (1.32 mg/dL). Based on lognormal distribution of the original data the mean (SD) of triglyceride values for males is 152.45 (85.14) mg/dL and for females is 109.56 (48.30) mg/dL. The difference in the means of the triglyceride values among males and females is 42.89 (Generalized 95% CI: 24.55, 63.70) mg/dL which is statistically significant (Generalized p value <0.001). As the 95% generalized CI does

not include the null value of 0, it provides evidence that the triglyceride values are different among males and females.

The empirical coverage probability of the simulation study of generalized confidence interval for the difference of two lognormal means obtained for various parameter values are given in Table 2. The results of the simulation study clearly shows that the estimated coverage probabilities of the generalized confidence intervals are almost near to the nominal level of 95%. This is even true for the studies with small sample sizes and for large standard deviations. Thus, simulation implies that generalized pivotal approach can be applied for positively skewed outcomes and even for small no. of observations.

DISCUSSION AND CONCLUSION

The lognormal distribution for positively skewed outcome is currently used in many situations like occupational exposure and pollution data [13], in the application of breath analysis [16], to study the mean carbon monooxide levels in the air [17], and also for analyzing the medical costs data for patients with type I diabetics and patients being treated for diabetics ketoacidosis (DKA) [18]. The main advantage of this approach is the statistical inference with regards to the means of the original data [16]. In this paper we tried to compare the means of biological outcome which is positively skewed and therefore requires special analysis strategy and data transformation. We applied generalized pivotal quantity of lognormal means to calculate generalized p value and its confidence interval.

There are many alternative approaches available in the literature to study the distribution of lognormal means [18]. For testing the mean of a lognormal distribution power function of four testing procedures were compared based on student-t, Edgeworth expansion, generalized p-value and permutation test [19]. Generalized p-value approach was used to study the effect of silver nitrate seeding by comparing the amount of rainfall between seeded and unseeded clouds [20]. The performance of generalized p value approach was compared against many methods using simulation study [17]. Though the generalized p value and confidence interval method is computationally intensive, it works better even for small sample sizes. This simulation study indicated that the coverage probability is near to the nominal level of 95% for both small and the large sample sizes. From the data analysis of comparison of two lognormal means of triglyceride values, there is a significant difference in the values between males and females. Thus generalized pivotal approach can be a useful approach in the comparison of lognormal means in many biological data as this does not need any transformation. Moreover in this approach the statistical inference from the testing procedure is straight forward.

REFERENCES

[1] J. M. Bland and D. G. Altman, *BMJ*, **1996**, vol. 312, no. 7039, p. 1153.
[2] R. M. Nixon and S. G. Thompson, *Stat. Med.*, **2004**, vol. 23, no. 8, pp. 1311–1331.
[3] G. W. Ph.D and J. Shostak, *Common Statistical Methods for Clinical Research with SAS Examples, Third Edition*, 3 edition. Cary, NC: SAS Publishing, **2010**.
[4] P. Armitage, G. Berry, and J. N. S. Matthews, *Statistical Methods in Medical Research*, 4 edition. Malden, MA: Wiley-Blackwell, **2001**.
[5] J. M. Bland and D. G. Altman, *BMJ*, **1996**, vol. 312, no. 7038, p. 1079.
[6] D. S. Moore, G. P. McCabe, and B. Craig, *Introduction to the Practice of Statistics*, 6th edition edition. New York: W. H. Freeman, **2008**.
[7] W. G. Manning, *J. Health Econ.*, **1998**, vol. 17, no. 3, pp. 283–295.
[8] W. G. Manning and J. Mullahy, *J. Health Econ.*, **2001**, vol. 20, no. 4, pp. 461–494.
[9] K. Krishnamoorthy and T. Mathew, *J. Stat. Plan. Inference*, **2003**, vol. 115, no. 1, pp. 103–121.
[10] G. Zachariah, S. Harikrishnan, M. N. Krishnan, P. P. Mohanan, G. Sanjay, K. Venugopal, and K. R. Thankappan, *Indian Heart J.*, **2013**, vol. 65, no. 3, pp. 243–249.
[11] K.-W. Tsui and S. Weerahandi, *J. Am. Stat. Assoc.*, **1989**, vol. 84, no. 406, p. 602.
[12] S. Weerahandi, *J. Am. Stat. Assoc.*, **1993**, vol. 88, no. 423, p. 899.
[13] K. Krishnamoorthy, T. Mathew, and G. Ramachandran, *J. Occup. Environ. Hyg.*, **2006**, vol. 3, no. 11, pp. 642–650.
[14] R Core Team, *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, **2014**.
[15] D. G. Altman, *Practical Statistics for Medical Research*, 1 edition. Boca Raton, Fla: Chapman and Hall/CRC, **1990**.
[16] K. Cimermanová, *Meas. Sci. Rev.*, **2007**, vol. 7, no. 4, pp. 31–36.
[17] Ulf Olsson, *J. Stat. Educ.*, **2005**, vol. 13, no. 1.
[18] Y.-H. Chen and X.-H. Zhou, *Stat. Med.*, **2006**, vol. 25, no. 23, pp. 4099–4113.
[19] K. B. Dulal, K. Kush, B. Runa, and J. R. Domenic, *J. Environ. Stat.*, **2013**, vol. 5, no. 1, pp. 1–21.

[20] K. Abdollahnezhad, M. Babanezhad, and A. A. Jafari, *J. Stat. Econom. Methods*, **2012**, vol. 1, no. 2, pp. 125–131.