Shanghai Archives of Psychiatry, 2015, Vol. 27, No. 5

•Biostatistics in psychiatry (29)•

The effect of simple imputation on inferences about population means when data are missing in biomedical research due to detection limits

Hongyue WANG[1,*], Guangqing CHEN[1], Xiang LU[1], Hui ZHANG[2], Changyong FENG[1,3]

Summary: The sample geometric mean has been widely used in biomedical and psychosocial research to estimate and compare population geometric means. However, due to the detection limit of measurement instruments, the actual value of the measurement is not always observable. A common practice to deal with this problem is to replace missing values by small positive constants and make inferences based on the imputed data. However, no work has been carried out to study the effect of this naïve imputation method on inference. In this report, we show that this simple imputation method may dramatically change the reported outcomes of a study and, thus, make the results uninterpretable, even if the detection limit is very small.

Keywords: sample geometric mean; population geometric mean; two-sample test

[*Shanghai Arch Psychiatry.* 2015; **27**(5): 319-325. doi: http://dx.doi.org/10.11919/j.issn.1002-0829.215121]

1. Introduction

Detection limit is a long standing problem in experimental sciences. It refers to the limited ability of an instrument in measuring an outcome of interest in a certain range (typically small values close to 0). Many instruments cannot return meaningful measurements if signals fall below a certain threshold value. This problem is especially prevalent in biomedical sciences as signals are sometimes too weak to be detected in the presence of ambient noise. Although detection limits are often due to limitations of physical devices, the problem may also arise in psychosocial research with assessments based on instruments (questionnaires). For example, in alcohol and substance use research, alcohol or drug use may not be detected in a subject if the blood level is not sufficiently high. Also, if answers for all or most subjects to an item in a questionnaire fall below (or above) a certain score in the potential range of scores, the lack of variability in the outcome may prevent any useful analysis of the data.

Detection limit presents problems for statistical analysis since no data (or very little data) is observed in part of the potential range of the variable. It is not possible to gauge the variability of the outcome below the detection limit, but this information is needed to conduct standard statistical inference on the data in the whole range (for example, to estimate the geometric mean of the population from which the sample is drawn). A commonly used ad-hoc method to deal with this problem is to impute data below the detection limit and then apply standard statistical methods.[1] This practice is especially prevalent in biomedical research. Geometric means are the most popular method of imputing values below the detection limit because data are often log-transformed to reduce skewness before being analyzed (even though the log-transformation may not actually reduce the skewness[2]). The arithmetic mean of the log-transformed outcome is the logarithm of the sample geometric mean.

Although imputation seems natural and intuitive, it has significant implications for statistical inference and, thus, on the reported results of research.[3,4] In this report we discuss the pitfalls of using this common method of imputation in research and in clinical practice.

[1] Departments of Biostatistics and Computational Biology, University of Rochester, Rochester, NY

[2] Department of Biostatistics, St. Jude Children's Research Hospital, Memphis, TN

[3] Department of Anesthesiology, University of Rochester, Rochester, NY

*correspondence: hongyue_wang@urmc.rochester.edu

A full-text Chinese translation of this article will be available at http://dx.doi.org/10.11919/j.issn.1002-0829.215121 on February 26, 2016.

2. Geometric mean of a non-negative random variable

The so-called 'geometric mean' used in biomedical and psychosocial research is actually the sample geometric mean, that is, the geometric mean of a sample of observations from an underlying distribution.

For example, let a_i, i=1,...,n, be a sequence of non-negative numbers; then the geometric mean of the sequence is $\sqrt[n]{a_1 a_2 \ldots a_n}$. This is what is commonly referred to as the 'geometric mean', but it is important to keep in mind that the *sample* and *population* means are completely different concepts. The former is computable based on the sample, while the latter is an unknown quantity, or a parameter in the nomenclature of statistics. The geometric mean described above is actually a *sample* geometric mean because it is a computable quantity. Unlike the arithmetic, the population geometric mean had never been clearly defined in the literature until the recent work of Feng and colleagues[2,3] who presented a formal definition of this elusive quantity. The population geometric mean of a non-negative variable X defined if either X has non-zero probability at zero (that is, X may equal 0) or X is positive with $|\log X|$ having a finite mean value, that is, $E|\log X|<\infty$. Subsequently the definition was further broadened to only require that $E|\log X|$ exists (which includes $E|\log X|<\infty$ as a special case) and the properties of the population geometric mean were elaborated.[4] This work lays a conceptual foundation to interpret the sample geometric mean (as an estimate of the underlying population geometric mean) and clarifies some ambiguities in using the geometric mean in biomedical research. A brief summary of this work is shown in Box 1.

The geometric mean has a very unusual property. We know that for a positive random variable, arithmetic mean, if it exists, is always positive. However, for some positive random variables, geometric means can be zero. This fact is counterintuitive as the sample geometric mean obtained from a positive random variable is always positive. This unusual property can have significant implications for inference about the population geometric mean when the data in the sample is left-censored due to a detection limit.

Another issue is the relationship between the geometric mean and the arithmetic mean for a positive random variable. In biomedical research, data is often right-skewed with most values close to the lower limit. A popular approach for dealing with this is to log-transform the data, analyze the transformed data, and then transform the result back to the original scale. For a non-negative random variable X, in general there is no connection between the geometric mean GM_X and the arithmetic mean $E(X)$, even if they both exist. For example, for two log-normally distributed random variables, if they have the same log-mean values but different log-variances, then their geometric means are equal but their arithmetic means are not equal. This means that we cannot test the hypothesis that they have the same (arithmetic) mean values by testing that the log-transformed data have the same (arithmetic) mean values. This fact is not well appreciated in biomedical research.[2]

BOX 1. Relationship of population and sample geometric means

- *Suppose $F(x)$ is the probability distribution function of a non-negative random variable X. If $F(0) > 0$, then the geometric mean of X, denoted by GM_X, is defined as 0. If X is positive and the expected value of $|\log X|$ exists, then we define GM_X as $\exp(E \log X)$. If $F(0)\neq 0$ and the expected value of $|\log X|$ does not exist, then it is not possible to define the geometric mean of X. Hence, we cannot define the geometric mean for all non-negative random variables, just like the mean cannot be defined for all random variables. However, if GM_X is well defined, it is the population geometric mean, which can be used to interpret the sample geometric mean of X.*
- *Suppose the geometric mean of X exists, and, X_i, i=1,...,n is a random sample from the distribution of X. Let $\overline{GM_n} = \sqrt[n]{X_1 \ldots X_n}$ be the sample geometric mean. Then the sample geometric mean is strongly consistent[3,4] and, thus, is a consistent estimate of the population geometric mean.*

3. Geometric mean with detection limit

In this section, we discuss the effect of the naïve imputation method on the geometric mean in the presence of a detection limit. Let X be a positive random variable and δ be the lower detection limit; X is unobservable (missing) if $X < \delta$. A common approach in biomedical research is to define a modified version of X by:

$$x^* = \begin{cases} x & \text{if } x \geq \delta, \\ \eta & \text{if } x < \delta, \end{cases}$$

where η is some positive constant. Usually, $\eta=\delta/2$, or a small positive constant less than δ.[1,5-15] After the imputation, inference about the population geometric mean of the original data proceeds by treating the imputed data as if they were observed.

To discuss potential effects of this naive imputation on inference about the population geometric mean, we assume that X is a positive random variable, which is the case in most real-study applications.

a) Case 1: $GM_X>0$. The geometric mean of X^* can be greater than, less than, or equal to GM_X depending on the distribution of X below the detection limit. If the detection limit δ is small enough, then with relatively large sample sizes inferences based on the imputed data (such as confidence intervals, the two-sample t-test, and the paired t-test) yield valid results for the original data. However, if δ is large, imputation may yield invalid results.

b) Case 2: $GM_X=0$. With the imputation, the geometric mean of X^* depends on how the imputed value η is selected and is always greater than η. This means that the estimated geometric mean based on the imputed data may be very far away from the theoretical geometric mean of zero. Another effect is that the imputation brings some arbitrariness into the statistical inference.

Thus whether $GM_X>0$ or $GM_X=0$, imputation has significant implications for inference about the population geometric mean. If $GM_X>0$, inference using common statistical methods is reasonably robust if the detection limit is small; but if $GM_X=0$, any analysis of the geometric mean based on the imputed data is invalid and the result is uninterpretable. Unfortunately, the detection limit makes it impossible to determine whether $GM_X>0$ or $GM_X=0$.

4. Simulation results

As described above, when the sample geometric mean of a positive random variable is 0, the geometric mean of the modified observation (which imputes values below the detection limit δ) may be very different from 0 and, thus, inferences based on the modified sample may be misleading.

Suppose Y has a standard log-normal distribution with its probability distribution function Φ, and U is independent of Y and uniformly distributed on (0,1). Let C_0 be a positive constant. The random variable X is defined as:

$$X = \begin{cases} Y & \text{if } Y \geq C_0, \\ C_0 \exp(1 - U^{-1}) & \text{if } Y < C_0. \end{cases} \quad (1)$$

The distribution function of X is:

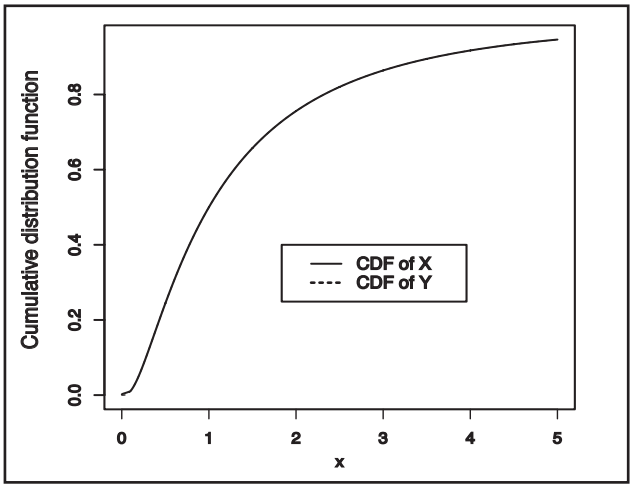
$$F_X(x) = \Pr\{X \leq x\} = \begin{cases} \dfrac{\Phi(C_0)}{1 + \log C_0 - \log x} & \text{if } 0 < x < C_0, \\ \Phi(x) & \text{if } x \geq C_0. \end{cases}$$

In the simulation study, C_0 is set at 0.277602. The data $X_1,.....,X_n$ is generated from the distribution of X defined in equation (1) above.

4.1 Properties of geometric mean

Figure 1 shows the cumulative distribution function of X and Y when C_0=0.1. Since $\Phi(0.1)=0.01$, it is nearly impossible to distinguish between the two distribution function curves in the figure. However, their geometric means are very different. It is easy to prove that $GM_Y=1$ and $GM_X=0$, no matter how small C_0 is (see Example 2 in Feng and colleagues[4] for a proof).

Figure 1. Cumulative distribution functions of X and Y in formula (1) with c_0=0.1



Since X is positive, the sample geometric mean is always positive. Although the sample geometric mean is a consistent estimator of GM_X (which is 0 in this case), it may be quite a large number. Table 1 is a sample of n=100 observations from the distribution of Z=100X, where X is defined in equation (1). The sample geometric mean is $\overline{GMn}$=85.72. However, the population geometric mean is actually GM_Z=0. It is very difficult to imagine that the data in Table 1 is from a distribution with a geometric mean of 0. This strange property of the geometric mean makes it difficult to test whether or not a sequence of positive numbers is a sample from a distribution with a geometric mean equal to 0.

In the simulation study, we set the detection limit at δ=0.277602 such that there was a 10% probability that X_i is below the detection limit, that is, Pr{X<0.277602}=0.1. No data is observed below this detection limit, so if the value of $\delta/2$ is imputed for all cases in which X_i falls below δ=0.277602, then the modified observations are

$$X_i^* = \begin{cases} X_i & \text{if } X_i \geq \delta = 0.277602, \\ 0.133801 & \text{if } X_i < \delta = 0.277602. \end{cases} \quad (2)$$

Let $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $\bar{X}_n^* = \frac{1}{n}\sum_{i=1}^{n} X_i^*$ be the sample means, and let $\overline{GM_n} = \sqrt[n]{X_1 \ldots X_n}$ and $\overline{GM}_n^* = \sqrt[n]{X_1^* \ldots X_n^*}$ be the sample geometric means of (X_1, ..., X_n) and (X_1^*, ..., X_n^*) respectively.

Table 2 shows the means and standard deviations of $\bar{X}_n$, $\bar{X}_n^*$, $\overline{GM_n}$, and $\overline{GM}_n^*$ for samples of different sizes after 100,000 Monte Carlo replications. In each replicate a random sample X_1, ..., X_n is generated, and $\bar{X}_n$, $\bar{X}_n^*$, $\overline{GM_n}$ and $\overline{GM}_n^*$ are calculated. For each n, the mean and standard deviation of $\overline{GM_n}$ is the sample mean and sample standard deviation based on 100,000 Monte Carlo replicates.

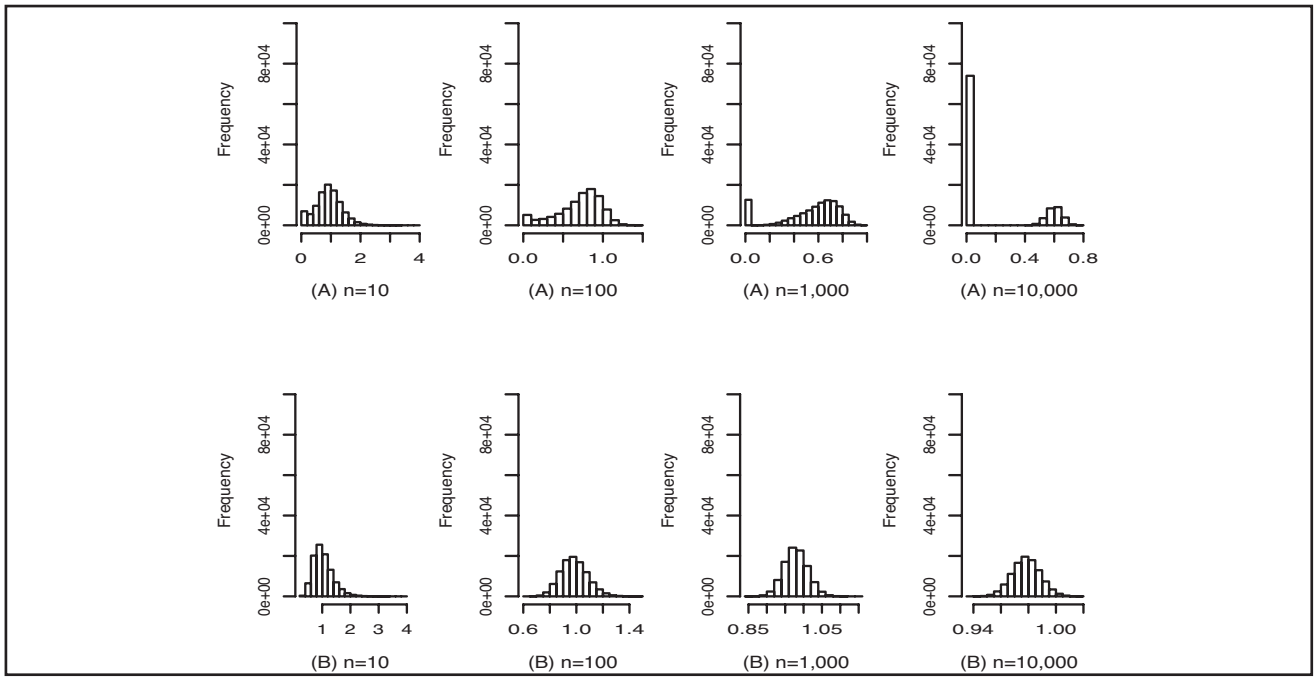
Table 1: A random sample from a distribution with geometric mean 0 (sample size n=100)

166.95	70.67	75.68	2.61	264.39	55.30	129.93	87.55	172.43	59.95
211.11	127.63	71.91	362.70	73.12	293.65	292.67	369.40	139.59	304.00
155.42	16.80	109.80	18.34	190.47	29.37	53.43	62.93	137.79	44.72
152.19	84.66	172.02	45.94	437.89	110.13	53.51	152.44	75.92	60.48
151.47	513.60	34.72	69.70	492.94	42.03	4.48	82.01	445.03	35.22
2.67	41.08	205.55	73.19	713.21	182.35	43.62	67.32	37.21	65.01
108.44	747.98	15.69	59.55	122.46	475.55	0.95	261.28	96.82	168.29
44.53	191.05	74.81	143.88	194.59	26.63	90.69	141.91	25.92	251.09
55.08	154.57	53.82	66.33	53.58	17.57	115.23	6.69	49.44	303.29
118.96	48.13	39.11	690.46	170.17	217.58	62.74	79.84	26.43	106.79

Table 2: Means and standard deviations of sample means and sample geometric means

sample size	$\overline{X}_n$		$\overline{X}_n^*$		$\overline{GM}_n$		$\overline{GM}_n^*$	
	mean	sd	mean	sd	mean	sd	mean	sd
10	1.6459	0.6936	1.6485	0.6930	0.9169	0.4460	1.0326	0.3445
50	1.6406	0.3068	1.6433	0.3065	0.7573	0.3079	0.9887	0.1450
100	1.6415	0.2167	1.6442	0.2165	0.7027	0.2797	0.9834	0.1017
500	1.6418	0.0969	1.6445	0.0968	0.5944	0.2389	0.9795	0.0453
1,000	1.6413	0.0688	1.6440	0.0688	0.5399	0.2464	0.9789	0.0321
5,000	1.6411	0.0306	1.6438	0.0306	0.3083	0.3056	0.9783	0.0143
10,000	1.6413	0.0217	1.6440	0.0216	0.1571	0.2651	0.9783	0.0101

Figure 2. Histogram of sample geometric means from the distributions of X (part A) and X* (part B) in formula (2) for different sample sizes



The same interpretation applies to the other columns in the table. There are two main findings shown in the table:

a) Since the detection limit is relatively small, the difference between the means $\overline{X}_n$ and $\overline{X}_n^*$ is very small. They are very close to $E(X_1)$ and $E(X_1^*)$ even for small sample sizes.

b) The geometric means behave very differently. $\overline{GM}_n$ converges to 0, while $\overline{GM}_n^*$ converges to a constant far away from 0. The sample geometric means $\overline{GM}_n$ and $\overline{GM}_n^*$ also change substantially as the sample size increases.

The panels in Figure 2 show the histograms of $\overline{GM}_n$ (the 'A' series) and $\overline{GM}_n^*$ (the 'B' series) after 100,000 Monte Carlo replications. Although the distribution of $\overline{GM}_n^*$ is skewed for relatively small sample sizes (n=10), the skewness almost disappears for relatively large sample sizes. However, the distribution of $\overline{GM}_n$ is skewed for all sample sizes, particularly for large sample sizes. Since GM_X=0, most of the sample geometric means clustered around 0 when n=10,000.

4.2 Hypothesis testing using geometric means

Let X_{11}, ..., $X_{1,n1}$ and X_{21}, ..., $X_{2,n2}$ be be two independent samples. Suppose we want to test the hypothesis:

$$H_0 : GM_{X11}=GM_{X21}.$$

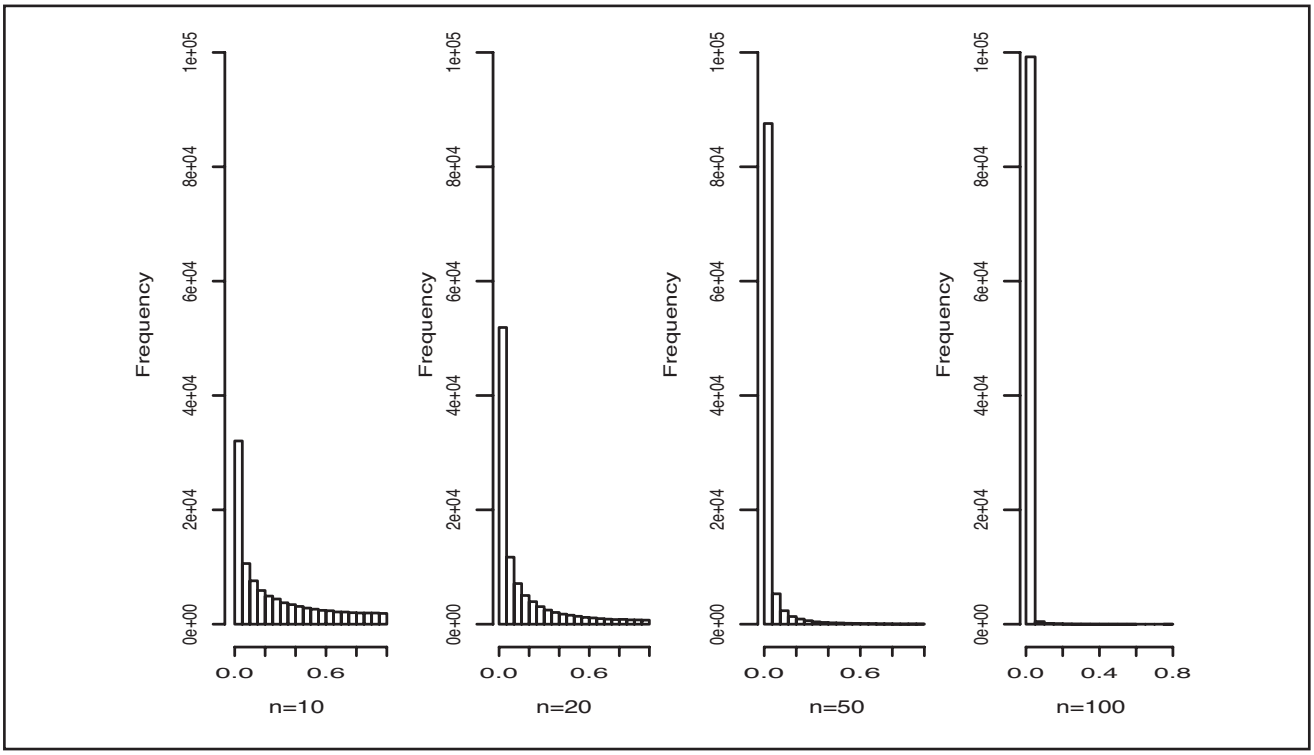
Due to detection limit, only the modified data can be used. The test statistic used in biomedical research is of the form

$$T^* = \frac{n_1^{-1}\sum_{j=1}^{n_1} \log X_{1j}^* - n_2^{-1}\sum_{j=1}^{n_2} \log X_{2j}^*}{\sqrt{S_1^{*2}/n_1 + S_2^{*2}/n_2}}, \qquad (3)$$

where S_k^{*2} is the sample variance of $\log X_{k1}^*, \ldots, \log X_{k,nk}^*$ (k=1,2).

In the simulation studies, X_{11} has the same distribution as defined in equation (2) with c_0=0.277602 and X_{21} is defined as $2X_{11}$. Figure 3 shows the histograms of p-values of the test statistic T^* for different sample sizes. In our example both samples have the same geometric means, so the null hypothesis is true and the distribution of the p-values of the statistic test T^* should be close to the uniform distribution, at least for large sample sizes. However, the histograms shown in Figure 3 clearly indicate otherwise. Thus results of testing the null hypothesis when using the modified data are difficult to interpret and can be quite misleading.

Figure 3. Histograms of p-values of the test statistic in formula (3) for different sample sizes



5. Discussion

In this paper we consider the effect of the most common method of data imputation used in biomedical research for results that are below a detection limit. Despite its popularity, this method of using imputed values to compute a sample geometric mean which is used to estimate the population geometric mean (needed in many common statistical analyses) has not been adequately reviewed in the statistical literature. We use simulation studies to show that that the sample geometric mean is a very unstable statistic, so even small modifications introduced by this common imputation method can have a major effect on the estimated true (population) geometric mean and, thus, on statistical inference.[4] The sample geometric mean based on data that includes imputed values can be quite different from the true geometric mean, so the conclusions of hypothesis testing based on the use of modified data can be misleading.

All these problems stem from a very special property of the geometric mean: a positive random variable may have a geometric mean of 0. However, given a random sample from the distribution of a positive random variable, there is no method to determine whether or not the population geometric mean is 0, a problem that is compounded by the detection limit issue that requires the use of imputed values when computing the sample geometric mean. Any computed estimate of the geometric mean under the detection limit is uninterpretable.

Another issue with detection limit is measurement error. In this paper we assume that there is no measurement error from the device or instrument. The effect of potential measurement error on the detection limit requires further investigation.

Acknowledgements

This study was supported by a pilot grant (PI: Feng) from the Clinical and Translational Sciences Institute at the University of Rochester Medical Center.

Conflict of interest statement

The authors report no conflict of interest related to this manuscript.

生物医学研究中因检测范围所限致数据缺失时简单填补法对人口几何均数推断的影响

Wang HY, Chen GQ, Lu X, Zhang H, Feng CY

概述：在生物医学和社会心理学研究中采用样本几何均值估计、比较人口几何均值的方法十分普遍。然而，由于测量工具的检测局限，有时无法观察到测量的实际值。处理这个问题的一种常见做法是用较小的正值常数来替代缺失值，然后在这些填补数据基础上进行统计推断。然而，这种简单的填补方法对推论的影响还没有研究过。我们在本文中阐明了这种简单的填补方法可能会大幅度地改变一项研究所报告的结果，因此即使检测限非常小，也会使结果难以解释。

关键词：样本几何均值；人口几何均值；双样本检测

本文全文中文版从 2016 年 2 月 26 日起在 http://dx.doi.org/10.11919/j.issn.1002-0829.215121 可供免费阅览下载

References

1. Green CA, Scarselli E, Sande CJ, Thompson AJ, de Lara CM, Taylor KS et al. Chimpanzee adenovirus- and MVA-vectored respiratory syncytial virus vaccine is safe and immunogenic in adults. *Sci Transl Med.* 2015; **7**(300): 300ra126. doi: http://dx.doi.org/10.1126/scitranslmed.aac5745
2. Feng C, Wang H, Lu N, Tu XM. Log-transformation: Application and interpretation in biomedical research. *Stat Med.* 2013; **32**(21): 230-239. doi: http://dx.doi.org/10.1002/sim.5840
3. Feng C, Wang H, Tu XM. Geometric mean of nonnegative random variable. *Communication in Statistics – Theory and Methods.* 2013; **42**(15): 2714-2717. doi: http://dx.doi.org/10.1080/03610926.2011.615637
4. Feng C, Wang H, Zhang Y, Han Y, Liang Y, Tu XM. Generalized definition of the geometric mean of a nonnegative random variable. *Communications in Statistics – Theory and Methods.* 2015; (In press)
5. Chen H, Matsuoka Y, Chen Q, Cox NJ, Murphy BR, Subbarao K, et al. Generation and characterization of a cold-adapted influenza A H9N2 reassortant as a live pandemic influenza virus vaccine candidate. *Vaccine.* 2003; **21**(27-30): 4430-4436
6. Chiesa C, Signore F, Assumma M, Buffone E, Tramontozzi P, Osborn J F, et al. Serial measurements of c-reactive protein and interleukin-6 in the immediate postnatal period: reference intervals and analysis of maternal and perinatal confounders. *Clin Chem*. 2001; **47**(6): 1016-1022
7. Donnenberg AD. Statistics of Immunological Testing. Edited by O'Gorman MRG and Donnenberg AD. *Handbook of Human Immunology (Second Edition).* CRC Press. 2008; p. 29-62
8. Ishida Y, Suzuki K, Taki K, Niwa, T, Kurotsuchi S, Ando H, et al. Significant association between Helicobacter pylori infection and serumC-reactive protein. *Int J Med Sci.* 2008; **5**(4): 224-229
9. Miranda-Novales G, Arriaga-Pizano L, Herrera-Castillo C, Pastelin-Palacios R, Valero-Pacheco N, Pérez-Toledo M, et al. Antibody responses to influenza viruses in paediatric patients and their contacts at the onset of the 2009 pandemic in Mexico. *J Infect Dev Ctries.* 2015; **9**: 259-266. doi: http://dx.doi.org/10.3855/jidc.5052
10. Pantazi H, Papapetrou PD. Calcitonin levels are similar in goitrous euthyroid patients with or without thyroid antibodies, as well as in hypothyroid patients. *Eur J Endocrinol.* 1998; **138**(5): 530-535
11. Petersen KM, Bulkow LR, McMahon BJ, Zanis C, Getty M, et al. Duration of hepatitis B immunity in low risk children receiving hepatitis B vaccinations from birth. *Pediatr Infect Dis J.* 2004; **23**(7): 650-655.
12. Slack MH, Schapira D, Thwaites RJ, Burrage M, Southern J, Goldblatt D, et al. Responses to a fourth dose of Haemophilus influenzae type B conjugate vaccine in early life. *Arch Dis Child Fetal and Neonatal Ed*. 2004; **89**(3): F269-F271

13. Sternthal MJ, Enlow MB, Cohen S, Canner MJ, Staudenmayer J, Tsang K, et al. Maternal interpersonal trauma and cord blood IgE levels in an inner-city cohort: A life-course perspective. *J Allergy Clin Immunol.* 2009; **124**(5): 954-960. doi: http://dx.doi.org/10.1016/j.jaci.2009.07.030
14. Whitney JB, Hill AL, Sanisetty S, Penaloza-MacMaster P, Liu J, Shetty M, et al. Rapid seeding of the viral reservoir prior to SIV viraemia in rhesus monkeys. *Nature.* 2014; **512**(7512): 74-77. doi: http://dx.doi.org/10.1038/nature13594
15. Yunker MB, Belickab LL, Harvey HR, Macdonald RW. Tracing the inputs and fate of marine and terrigenous organic matter in arctic ocean sediments: a multivariate analysis of lipid biomarkers. *Deep-Sea Research II: Topical Studies in Oceanography.* 2005; **52**(24-26): 3478-3508. doi: http://dx.doi.org/10.1016/j.dsr2.2005.09.008

(received, 2015-10-20; accepted, 2015-10-23)

Hongyue Wang obtained her Bachelors of Science in Scientific English from the University of Science and Technology of China in 1995, and a PhD in Statistics from the University of Rochester in 2007. She is a Research Assistant Professor in the Department of Biostatistics and Computational Biology at the University of Rochester Medical Center. Her research interests include longitudinal data analysis, missing data, survival data analysis, and design and analysis of clinical trials. She has extensive collaboration with investigators in infectious diseases, nephrology, neonatology, cardiology, neurodevelopmental and behavioral science, radiation oncology, pediatric surgery, and dentistry. She has published 65 papers about statistical methodology and other topics in peer-reviewed journals.