Journal of Modern Applied Statistical Methods

Volume 9 | Issue 1 Article 3

5-1-2010

Inferences about the Population Mean: Empirical Likelihood versus Bootstrap-t

Rand R. Wilcox
University of Southern California, rwilcox@usc.edu

DOI: 10.22237/jmasm/1272686520

Journal of Modern Applied Statistical Methods
May 2010, Vol. 9, No. 1, 9-14

1538 – 9472/10/$95.00

Inferences about the Population Mean: Empirical Likelihood versus Bootstrap-t

Rand R. Wilcox
University of Southern California

The problem of making inferences about the population mean, μ, is considered. Known theoretical results suggest that a Bartlett corrected empirical likelihood method is preferable to two basic bootstrap techniques: a symmetric two-sided bootstrap-t and an equal-tailed bootstrap-t. However, simulations in this study indicate that, when the sample size is small, these two bootstrap methods are generally better in terms of Type I errors and probability coverage. As the sample size increases, situations are found where the Bartlett corrected empirical likelihood method performs better than the equal-tailed bootstrap-t, but the symmetric bootstrap-t gives the best results. None of the four methods considered are always satisfactory in terms of probability coverage or Type I errors, particularly when dealing with skewed distributions where the expected proportion of points flagged as outliers is somewhat high. If this proportion is 0.14, for example, all four methods can be unsatisfactory even with n=300, but if sampling from a symmetric distribution or a skewed distribution with relatively light tails the results suggest using a symmetric two-sided bootstrap-t method.

Key words: Level robust methods, Bartlett correction, bootstrap-t.

Introduction

One of the fundamental goals in statistics is making inferences about the population mean, μ; the classic and routinely used method to accomplish this is Student's t-test. However, when sampling from a skewed distribution, Student's t is known to be unsatisfactory in terms of Type I errors as well as probability coverage when computing a confidence interval (Rosenblum & van der Laan, 2009; Westfall & Young, 1993; Wilcox, 2005). With a relatively light-tailed distribution such as the lognormal, roughly meaning that the expected proportion of points declared outliers is relatively small, Student's t requires a sample size of about n = 200 in order to achieve reasonably accurate control over the probability of a Type I error. With a heavier-tailed distribution (a g-and-h distribution with g = h = 0.5), where the expected proportion of outliers is approximately 0.14 (based on the boxplot rule in Frigge, Hoaglin & Iglewicz, 1989), n > 300 is required.

Rand R. Wilcox is a Professor of Psychology. He is the author of seven textbooks on statistics, the most recent of which is *Basic Statistics: Understanding Conventional Methods and Modern Insights* (2009, New York, Oxford University Press). Email him at: rwilcox@usc.edu.

As a result, numerous alternative methods have been proposed. One general approach is to use nonparametric techniques, which include empirical likelihood methods (Owen, 2001) as well as bootstrap methods (Efron & Tibshirani, 1993). Asymptotic results suggest that a Bartlett corrected empirical likelihood approach is superior to using a bootstrap-t method (DiCiccio, Hall & Romano, 1991). However, with small to moderate sample sizes, it appears that little or nothing is known regarding how these two approaches compare. Moreover, simulation results on the empirical likelihood technique are limited to a rather narrow range of situations.

This study compared two basic variations of the bootstrap-t method to two variations of the empirical likelihood method. A minor result is that the simulations support extant results that the Bartlett corrected empirical likelihood method is preferable to the basic empirical likelihood technique. A practical issue, however, is whether a Bartlett corrected empirical likelihood method provides better control over the Type I error probability, versus a bootstrap-t method, when dealing with small to moderate sample sizes. Yet another issue is the extent to which a Bartlett corrected empirical likelihood method gives improved results when sampling from a heavy-tailed distribution, particularly when the distribution is also skewed.

With n = 20, none of the methods compared are satisfactory among all of the distributions considered; none of the methods are satisfactory when sampling from a skewed, heavy-tailed distribution with $n \leq 300$. With a small sample size, the simulations indicate that the bootstrap-t methods are generally better than the empirical likelihood methods. As the sample size gets large, situations are found where the Bartlett corrected empirical likelihood method performs better than the equal-tailed bootstrap-t, but all indications point to the symmetric bootstrap-t as best for general use.

Let $X_1,\ldots,X_n$ be a random sample from a distribution with mean μ. Note that Rosenblum and van der Laan (2009) described a method for computing a confidence interval for the mean. Their method is based on Hoeffding's inequality (Hoeffding, 1963), which guarantees probability coverage at least $1-\alpha$ if W can be specified such that with probability 1, $|X_i| \leq W$. For the special case $1-\alpha$ = .95, the resulting 0.95 confidence interval is $(\bar{X} - 2.72W/\sqrt{n}, \bar{X} + 2.72W/\sqrt{n})$.

A simple way of implementing this approach is to take W to be the maximum of the observed $|X_i|$ values, but a possible concern from a hypothesis testing point of view is that it is too conservative in terms of Type I errors. In the simulations herein, this approach was considered when sampling from various distributions, including a normal distribution, and based on 5,000 replications, the hypothesis $H_0: \mu = \mu_0$, where μ_0 is the true population mean, was never rejected with sample sizes n = 20 and n = 200. Consequently, this approach was eliminated from consideration.

Methods for Comparison:Descriptions

Equal-Tailed Bootstrap-t

The idea behind the bootstrap-t method is to use the observed data to approximate the distribution of

$$T = \frac{\bar{X} - \mu}{s/\sqrt{n}},$$

where $\bar{X}$ and s are the usual sample mean and sample standard deviation, respectively. The strategy begins by generating a bootstrap sample of size n; that is, randomly sample with replacement n values from $X_1,\ldots,X_n$ yielding $X_1^*,\ldots,X_n^*$. Let $\bar{X}^*$ and s^* be the mean and standard deviation based on this bootstrap sample, and let

$$T^* = \frac{\bar{X}^* - \bar{X}}{s^*/\sqrt{n}}. \qquad (1)$$

Repeat this process B times yielding $T_1^*,\ldots,T_B^*$ and let $T_{(1)}^* \leq \ldots \leq T_B^*$ be the B bootstrap T^* values written in ascending order. Let $\ell = \alpha B$, rounded to the nearest integer, and $u = B - \ell$, in which case an estimate of the $\alpha/2$ and 1-

$\alpha / 2$ quantiles of the distribution of T are $T^*_{(\ell+1)}$ and $T^*_{(u)}$, respectively. The resulting equal-tailed $1-\alpha$ confidence interval for μ is

$$(\bar{X} - T^*_{(u)} \frac{s}{\sqrt{n}}, \bar{X} - T^*_{(\ell+1)} \frac{s}{\sqrt{n}}) \qquad (2)$$

It might seem that $T^*_{(u)}$ should be used to compute the upper end of the confidence interval, not the lower end, but it can be shown that this not the case. Also, $T^*_{(\ell+1)}$ is negative, which helps explain why $T^*_{(\ell+1)} s / \sqrt{n}$ is subtracted from $\bar{X}$.

Symmetric Bootstrap-t

In contrast to the equal-tailed bootstrap-t is the symmetric confidence interval

$$\bar{X} \pm T^*_{(c)} \frac{s}{\sqrt{n}},$$

where $c = (1-\alpha)B$ rounded to the nearest integer and the absolute value of the right side of (1) is used to define T^*. This symmetric two-sided confidence interval enjoys some theoretical (asymptotic) advantages over the equal-tailed confidence interval (Hall, 1988a, 1988b), but it is known that - for small sample sizes - situations arise where an equal-tailed confidence interval is more satisfactory (Wilcox, 2005).

Empirical Likelihood

The empirical likelihood method can be used to construct a confidence interval for μ, but for simplicity it is described in terms of testing $H_0 : \mu = \mu_0$. Consider distributions F_p, $p = (p_1, \ldots, p_n)$, supported on the sample $X_1, \ldots, X_n$, where X_i is assigned mass p_i. For a specified value of μ, the empirical likelihood L(μ) is defined to be the maximum value of Πp_i over all such distributions that satisfy $\sum X_i p_i = \mu$. Because Πp_i attains its overall maximum when $p_i = 1 / n$, it follows that the empirical likelihood is maximized when $\mu = \bar{X}$. The empirical likelihood ratio for testing H_0 is $W = -2\log\{L(\mu_0) / L(\bar{X})\}$.

When the null hypothesis is true, W has approximately a Chi-squared distribution with 1 degree of freedom. In particular, H_0 will be rejected at the α level if $W \geq c$, where c is the 1-α quantile of a Chi-squared distribution with 1 degree of freedom.

Bartlett Corrected Empirical Likelihood

The Bartlett corrected empirical likelihood method is applied as follows. Let $\hat{\mu}_j = \sum (X_i - \bar{X})^j / n$ and

$$a = \frac{1}{2}\hat{\mu}_4 \hat{\mu}_2^{-2} - \frac{1}{3}\hat{\mu}_3^2 \mu_2^{-3};$$

the null hypothesis is rejected if $W(1 - an^{-1}) \geq c$.

Comments on Designing a Simulation Study

Presumably there are situations where sampling is from a relatively light-tailed, symmetric distribution and outliers are relatively rare, but in various situations it is known that the reverse is true. In a review of 440 large-sample psychological studies, Micceri (1989) reported that 97% (35 of 36 studies) "of those distributions exhibiting kurtosis beyond the double exponential (3.00) also showed extreme or exponential asymmetry" (p. 161). Moreover, 72% (36 of 50) of distributions that exhibited skewness greater than two also had tail weights that were heavier than the double exponential.

In a sexual attitude study by Pedersen, Miller, Putcha-Bhagavatula and Yang (2002), skewness and kurtosis, based on 105 participants, was estimated to be 15.9 and 256.3, respectively. In a related study based on 16,288 participants, the ten variables had estimated skewness that ranged between 52.1 and 115.5, and kurtosis that ranged between 3,290 and 13,357. Based on a boxplot, the proportion of points flagged as outliers ranged between 0.12 and 0.39. Consequently, there are some practical reasons for considering heavy-tailed distributions in simulation studies as well as

distributions that have a fairly high degree of skewness.

An important point is that extant simulation studies regarding empirical likelihood methods do not consider a very wide range of distributions. For example, DiCiccio, et al. (1991) considered a Student's t distribution with 5 degrees of freedom, which has a median proportion of outliers (over many studies) approximately equal to 0.03 based on the boxplot rule in Frigge, Hoaglin and Iglewicz (1989). In addition to a normal distribution, they also considered a Chi-squared distribution with 1 degree of freedom for which the median proportion of outliers is approximately 0.07. Their simulations reveal unsatisfactory control over the probability of a Type I error with n = 20, but with n = 40 the Bartlett corrected version was found to perform reasonably well. This study describes situations where it performs poorly with n = 300.

Results

Simulations were used to study the actual Type I error probability when testing $H_0 : \mu = \mu_0$. The distributions used were standard normal, Chi-squared with 1 degree of freedom, Student's t with 5 degrees of freedom, lognormal, contaminated normal, and three g-and-h distributions. For convenience these distributions are labeled distributions 1-8, respectively.

The family of contaminated (or mixed) normal distributions used is defined as follows. Let X be a standard normal random variable having the distribution $\Phi(x) = P(X \le x)$. Let ε be any constant, $0 \le \varepsilon \le 1$ and let K be any positive constant. The contaminated normal distribution is

$$H(x) = (1 - \varepsilon)\Phi(x) + \varepsilon\Phi(x \,/\, K).$$

Following Tukey (1960), K = 10 and $\varepsilon = .1$ are used resulting in a symmetric, heavy-tailed distribution, with the median proportion of points declared outliers approximately equal to 0.08. The first three distributions were chosen to illustrate how the bootstrap-t compares to the empirical likelihood methods for the same distributions used by DiCiccio, et al. (1991).

The g-and-h distributions (Hoaglin, 1985) arise as follows. If Z has a standard normal distribution, then

$$W = \frac{\exp(gZ) - 1}{g} \exp(hZ^2 \,/\, 2),$$

$g > 0$, has a g-and-h distribution where g and h are parameters that determine the first four moments. When $g = 0$,

$$W = Z \exp(hZ^2 \,/\, 2).$$

The three g-and-h distributions used were g = h = 0.2 and 0.5, and (g, h) = (0.2, 0). Table 1 shows the skewness (γ_1) and kurtosis (γ_2) for each of the g-and-h distributions considered. When g>0 and h>1/k, $E(W^k)$ is not defined and the corresponding entry in Table 1 is left blank. Additional properties of the g-and-h distribution are summarized by Hoaglin (1985).

Table 1: Some Properties of the g-and-h Distribution

g	h	γ_1	γ_2
0.2	0.0	0.61	3.68
0.2	0.2	2.81	155.98
0.5	0.5		

To add perspective, note that the median proportion of outliers generated, when dealing with g = h = 0.5, is approximately 0.11 when n = 100, based on the variation of the boxplot rule recommended by Frigge, Hoaglin & Iglewicz (1989). For g = h = 0.2 it is 0.05 and for (g, h) = (0.2, 0) it is 0.01. For a Chi-squared distribution with 1 degree of freedom, t_5, the lognormal and the contaminated normal, the median proportion of outliers is approximately 0.07, 0.03, 0.08 and 0.08, respectively. (These results are based on simulations with 5,000 replications.)

Table 2 shows the estimated Type I error probabilities. First consider n = 20, and note that

the Bartlett corrected empirical likelihood method always improves on the uncorrected approach. Both bootstrap methods have estimated Type I error probabilities less than the estimates using the empirical likelihood methods. Although the seriousness of a Type I error depends on the situation, Bradley (1978) has suggested that generally, at a minimum, the actual Type I error probability should be between 0.025 and 0.075. Based on this criterion, none of the methods are satisfactory. However, for skewed distributions for which the median proportion of outliers does not exceed 0.05, the symmetric bootstrap method gives satisfactory results.

The symmetric bootstrap method can be too conservative when sampling from a symmetric heavy-tailed distribution, but this might be judged to be less serious than having an actual Type I error greater than 0.075, as is the case when using the empirical likelihood methods. Note that with n = 20, the symmetric bootstrap method has a Type I error probability of 0.08 when sampling from a Chi- squared distribution with 1 degree of freedom. Increasing the sample size to n = 25, the estimate drops to 0.065, and for n = 30 it is 0.059.

For n = 50, the empirical likelihood methods compete better with the bootstrap-t methods, but the symmetric bootstrap-t performs well in situations where the empirical likelihood methods are unsatisfactory based on Bradley's criterion. Again, a criticism of the symmetric bootstrap-t is that for a symmetric heavy-tailed distribution (the contaminated normal), the Type I error probability drops below 0.025, but the other three methods have estimates greater than 0.12. Thus, for general use, the symmetric bootstrap-t seems best.

Additional simulations were conducted with n = 100 and it was found that the empirical likelihood methods continue to perform poorly when sampling from the heavy-tailed distributions considered here. With n = 200 they perform well when sampling from the contaminated normal but estimates exceed 0.15 when sampling from the g-and-h distribution with g = h = 0.5.

Table 2: Estimated Type I Error Probabilities

n	Distribution	Empirical Likelihood (EL)	Bartlett Corrected Empirical Likelihood (BCEL)	Bootstrap-t, Equal-Tailed (BEQ)	Bootstrap-t, Symmetric (BSYM)
20	1	0.074	0.064	0.058	0.045
	2	0.117	0.103	0.068	0.080
	3	0.075	0.059	0.067	0.036
	4	0.137	0.120	0.099	0.104
	5	0.169	0.138	0.116	0.010
	6	0.090	0.072	0.083	0.035
	7	0.094	0.080	0.083	0.047
	8	0.270	0.241	0.231	0.186
50	1	0.052	0.050	0.055	0.049
	2	0.074	0.069	0.055	0.059
	3	0.062	0.058	0.072	0.048
	4	0.068	0.062	0.058	0.054
	5	0.137	0.125	0.145	0.011
	6	0.061	0.057	0.073	0.037
	7	0.074	0.066	0.080	0.050
	8	0.215	0.203	0.207	0.194

Conclusion

In terms of controlling the probability of a Type I error, the most difficult situation seems to occur when sampling from an asymmetric distribution with heavy-tails. Even using n = 300 none of the methods considered are satisfactory. In particular, for the g-and-h distribution with g = h = 0.5, all four methods estimated Type I error probabilities exceeding 0.14. One of the main points is that - for symmetric distributions with heavy tails - the symmetric bootstrap-t avoids Type I errors well above the nominal level even with n = 20 (albeit with small sample sizes the actual level can drop below 0.025). By contrast, the Bartlett corrected empirical likelihood method has an actual level of approximately 0.09 with n = 100, and with n = 200 the level drops to 0.063. Consequently, it seems that the symmetric bootstrap-t is best for general use. Except for skewed heavy-tailed distributions, it performs reasonably well with n $\geq$ 50.

References

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, *31*, 144-152.

DiCiccio, T., Hall, P., & Romano, J. (1991). Empirical likelihood is Bartlett-correctable. *Annals of Statistics*, *19*, 1053-1061.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.

Frigge, M., Hoaglin, D. C., & Iglewicz, B. (1989). Some implementations of the boxplot. *American Statistician*, *43*, 50-54.

Hall, P. (1988a). On symmetric bootstrap confidence intervals. *Journal of the Royal Statistical Society*, *Series B*, *50*, 35-45.

Hall, P. (1988b). Theoretical comparison of bootstrap confidence intervals. *Annals of Statistics*, *16*, 927-953.

Hoaglin, D. C. (1985). Summarizing shape numerically: The g-and-h distribution. In D. Hoaglin, F. Mosteller, & J. Tukey (Eds.), *Exploring data tables trends and shapes*. New York: Wiley.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, *58*, 13–30.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, *105*, 156--166.

Owen, A. B. (2001). *Empirical likelihood*. New York: Chapman & Hall.

Pedersen, W. C., Miller, L. C., Putcha-Bhagavatula, A. D., & Yang, Y. (2002). Evolved sex differences in sexual strategies: The long and the short of it. *Psychological Science*, *13*, 157-161.

Rosenblum, M. A., & van der Laan, M. J. (2009). Confidence intervals for the population mean tailored to small sample sizes, with applications to survey sampling. *International Journal of Biostatistics*, *5*, Article 4.

Sutton, C. D. (1993). Computer-intensive methods for tests about the mean of an asymmetrical distribution. *Journal of the American Statistical Association*, *88*, 802-810.

Tukey, J. W. (1960). A survey of sampling from contaminated normal distributions. In I. Olkin, S. Ghurye, W. Hoeffding, W. Madow, & H. Mann (Eds.) *Contributions to probability and statistics*, 448-485. Stanford, CA: Stanford University Press.

Westfall, P. H., & Young, S. S. (1993). *Resampling based multiple testing*. New York: Wiley.

Wilcox, R. R. (2005). *Introduction to Robust Estimation and Hypothesis Testing* (*2nd Ed.*). San Diego, CA: Academic Press.