Journal of Statistical Theory and Applications
Volume 11, Number 2, 2012, pp. 183-195 ISSN 1538-7887

Bootstrap Procedures for Testing Homogeneity Hypotheses

Bimal Sinha[1], Arvind Shah[2], Dihua Xu[1], Jianxin Lin[2] and Junyong Park[1]

[1]Department of Mathematics and Statistics, University of Maryland,
Baltimore County, 1000 Hilltop Circle, Baltimore, MD 21250, USA

[2]Clinical Biostatistics, RY34-A312, P.O. Box. 2000,
Merck Research Laboratories, Rahway, New Jersey 070565, USA

Abstract

Before pooling data on effect sizes (a generic term for parameters of interest in the context of meta-analysis) from different studies, it is important to test for homogeneity of the effect sizes. A well known test for homogeneity is based on Cochran's chisquare statistic. Our recent investigation showed that when the effect size of interest is a pairwise correlation, Cochran's homogeneity test is inaccurate; it has a highly inflated type I error probability, and hence cannot be recommended for practical use. However, we also noted that the homogeneity test for the correlations, performed after Fisher's variance stabilizing z transformation, is quite accurate. In general, such a transformation is not known for every parameter of interest. A natural approach to try is then to use the bootstrap. We propose to investigate the accuracy of the bootstrap for testing the homogeneity hypothesis of *two* natural problems for which the Cochran's test is known to be inaccurate.

Keywords: Bootstrap, Cochran's test, effect size, homogeneity hypothesis, Type I error.

AMS 2000 Subject Classifications: 62F03, 62G09

1 Introduction

Before pooling data on effect sizes (a generic term for parameters of interest in the context of meta-analysis) from different studies, it is important to test for homogeneity of the effect sizes. A well known test for homogeneity is based on Cochran's chi-square statistic. Our recent investigation (Mathew et al. (2010)) showed that when the effect size of interest is a pairwise correlation, Cochran's homogeneity test is inaccurate; it has a highly inflated type I error probability, and hence cannot be recommended for practical use. However, we also noted that the homogeneity test for the correlations, performed after Fisher's variance stabilizing z transformation, is quite accurate. In general, such a transformation is not known for every parameter of interest. A natural approach to try is then to use the bootstrap. We propose to investigate the accuracy of the bootstrap for testing the homogeneity hypothesis of *two* natural problems for which the Cochran's test is known to be inaccurate.

The organization of the paper is as follows. In Section 2 we discuss the test of homogeneity of bivariate correlations and in Section 3 we consider the problem of testing the equality of univariate normal means under heterogeneous variances. We demonstrate that in both the cases tests based on bootstrap accurately maintain Type I error rates. Section 4 contains our conclusion.

2 Test for the equality of several bivariate correlations

Assume that there are k bivariate normal populations with ρ_i as the correlation coefficient from the ith population, and consider the problem of testing $H_o : \rho_1 = \cdots = \rho_k$ versus H_1: not all correlations are equal. A popular test in this context is the familiar Cochran's chi-square test based on

$$T_C = \sum_{i=1}^{k} (r_i - \bar{r}^*)^2/\sigma_i^2 \tag{1}$$

where $\bar{r}^* = \frac{\sum_{i=1}^{k} r_i/\sigma_i^2}{\sum_{i=1}^{k} 1/\sigma_i^2}$, $\sigma_i^2 = (1 - r_i^2)^2/(n_i - 1)$, and the test rejects H_0 at level α when T_C exceeds $\chi^2_{\alpha,k-1}$. Here r_i is the sample correlation coefficient based on n_i paired samples from the ith population, $i = 1, \cdots, k$. It has been recently demonstrated that this widely used test has a highly inflated Type I error probability based on the cut-off point $\chi^2_{\alpha,k-1}$ and also that

the test based on Fisher's z-transformation of r does maintain the nominal Type I error rate quite well (see Mathew et al. (2010) for details).

Below we carry out the bootstrap procedure to determine the bootstrapped cut-off point T_S of T_C and show that the test which rejects H_0 when $T_C > T_S$ clearly maintains the nominal Type I error rate.

Bootstrap Procedure

1. Given k pairs of (n_i, r_i), compute $\bar{r}^{(0)} = \frac{\sum_{i=1}^{k} r_i/\sigma_i^2}{\sum_{i=1}^{k} 1/\sigma_i^2}, \sigma_i^2 = (1-r_i^2)^2/(n_i-1)$ and Cochran's chi-square test statistic T_C defined in equation (1).

2. Generate r_i^* from the exact sampling distribution of Pearson's correlation with true parameters n_i and $\bar{r}^{(0)}$, $i = 1, \cdots, k$.

3. Compute Cochran's chi-square test statistics T_{C1}^* based on $\{(r_i^*, n_i), i = 1, \cdots, k\}$.

4. Repeat steps 2 and 3 B times to get $T_{C1}^*, T_{C2}^*, \ldots, T_{CB}^*$.

5. Find the 95% percentile of $\{T_{C1}^*, T_{C2}^*, \ldots, T_{CB}^*\}$, namely, T_S.

6. Reject H_0 if $T_C > T_S$. We call this the bootstrap test, and denote it by T_{boot}.

Type I error rate. To study the properties of the tests, we performed Monte Carlo simulations under the null hypothesis of homogeneity of population correlations. When comparing T_C, T_Z (Fisher's Transformation), and T_{boot} (Bootstrap), we are mainly interested in Type I error rates, given the nominal Type I error of $\alpha = 0.05$.

The simulation study was carried out using the statistical software R (2010) (Version 2.11.0, 2010-04-22). In step 2 above, we need to generate bootstrap samples r_i^* for $\rho = \bar{r}^{(0)}$. The density of sample correlation r for sample size n and the population correlation ρ is given by (see Rao (1973) for details)

$$\frac{2^{n-3}}{\pi(n-3)}(1-\rho^2)^{(n-1)/2}(1-r^2)^{(n-4)/2}\sum_{s=0}^{\infty}\Gamma^2\left(\frac{n+s-1}{2}\right)\frac{(2\rho r)^s}{s!}.$$

To generate $r_i^{(*)}$, we used the function "rPearson" in the package "SuppDists" in R. Each estimated Type I error rate is based on 10,000 simulation runs. The real data set used in Table 1 refers *only* to the sample sizes used in the computation of the correlations between two of

the three cholesterol related variables Low-Density Lipoprotein cholesterol (LDL), Non-High Density Lipoprotein cholesterol (NHDL) and Apo lipoprotein B (APOB) obtained from $k = 29$ studies. The sample sizes of these studies are

$$875, 664, 210, 700, 745, 619, 1528, 2832, 1790, 2854, 1167, 179, 551,$$
$$1069, 1021, 465, 296, 237, 545, 531, 625, 100, 406, 362, 170, 269, 422, 627, 593$$

The studies were meant to investigate the efficacy of Ezetimibe in combination with statins in patients with hypercholesterolemia. We refer to Mathew et al. (2010) for details. The common null correlation ρ being unknown, we have taken values of ρ from 0.1 to 0.9. *Conclusion.* We have considered several scenarios in our simulation study with small and large values of k and N, N being the common paired sample size from the k populations. The results are presented in Tables 1 to 4. It is clear that the test based on Fisher's z nicely maintains Type I error rate, which is to be expected under the normality assumption. However, the performance of T_{boot} is also remarkable. Naturally, as already reported in Mathew et al. (2010), T_C does not maintain Type I error rate at an acceptable level.

3 Test for the equality of several normal means

Suppose we have k independent univariate normal populations where the ith population follows $N(\mu_i, \sigma_i^2)$, $i = 1, \cdots, k$, and we are interested in testing the homogeneity of normal means

$$H_0 : \ \mu_1 = \cdots = \mu_k \quad \text{versus} \quad H_1 : \ \mu_i \neq \mu_j, \ \text{for some } i \neq j. \tag{2}$$

Let $\bar{Y}_i$ denote the sample mean and S_i^2 the sample variance from n_i observations in the ith population. Then, we have

$$\bar{Y}_i \sim N\left(\mu_i, \frac{\sigma_i^2}{n_i}\right) \quad \text{or} \quad \bar{Y}_i \overset{H_0}{\sim} N\left(\mu, \frac{\sigma_i^2}{n_i}\right) \tag{3}$$

and

$$\frac{(n_i - 1)\ S_i^2}{\sigma_i^2} \sim \chi^2_{n_i - 1}. \tag{4}$$

Note that $\bar{Y}_i$ and S_i^2 are stochastically independent. We discuss below two popular tests and study their Type I error rates.

Likelihood ratio test. The LRT was discussed in details in Tongsai et al. (2010). Briefly, under H_0, the ML estimators are solutions of the estimating equations

$$\tilde{\sigma}^2_{i(ML)} = \frac{(n_i - 1)S_i^2}{n_i} + (\bar{Y}_i - \hat{\mu}_{ML})^2 \ , \ i = 1, \ldots, k, \tag{5}$$

and

$$\tilde{\mu}_{ML} = \frac{\sum_{i=1}^k n_i \ \bar{Y}_i/\tilde{\sigma}^2_{i(ML)}}{\sum_{i=1}^k n_i/\tilde{\sigma}^2_{i(ML)}}. \tag{6}$$

On the other hand, the unrestricted MLE's of $\mu_1, \ldots, \mu_k$ and $\sigma_1^2, \ldots, \sigma_k^2$ are given by

$$\hat{\mu}_{i(ML)} = \bar{Y}_i \quad \text{and} \quad \hat{\sigma}^2_{i(ML)} = \frac{(n_i - 1)}{n_i} S_i^2. \tag{7}$$

Consequently, the LRT statistic is simply

$$\lambda = \prod_{i=1}^k \left(\frac{\hat{\sigma}^2_{i(ML)}}{\tilde{\sigma}^2_{i(ML)}} \right)^{n_i/2}. \tag{8}$$

The following large sample test procedure is often proposed.

$$\text{Reject } H_0 \text{ at level } \alpha \text{ if } -2 \log \lambda \geq \chi^2_{k-1;1-\alpha},$$

where $\chi^2_{k-1;1-\alpha}$ denotes the upper $(1-\alpha)$-quantile of the χ^2-distribution with $k-1$ degrees of freedom.

Cochran's test. The familiar homogeneity test suggested by Cochran (1937) in this context is based on

$$Q_C = \sum_{i=1}^k w_i \ \left(\bar{Y}_i - \sum_{j=1}^k h_j \bar{Y}_j \right)^2, \tag{9}$$

where $w_i = n_i/S_i^2$, $h_i = w_i / \sum_{i=j}^k w_j$. Under H_0, the Cochran's statistic is distributed approximately as a χ^2-variable with $k-1$ *df*. The test rejects H_0 at level α if $Q_C > \chi^2_{k-1;1-\alpha}$. Note that the null distributions for Cochran's test and LRT are identical. Due to the obvious computational difficulties associated with the LRT, Cochran's test is often used as the standard test for testing homogeneity in meta-analysis.

Bootstrap Procedure. As in Problem 1, we carried out the bootstrap procedure for this problem to determine the bootstrapped cut-off points and demonstrate that the *same* two test statistics

which reject H_0 based on bootstrapped cut-off points do maintain the nominal type I error rate. Here are the steps we followed.

1. Given k pairs of $(\bar{Y}_i, S_i^2, n_i)$, compute $\tilde{\mu}_{ML}$, and $\tilde{\sigma}^2_{i(ML)}$ defined in equations (5) and (6), respectively. Also, compute Cochran's chisquare test statistic Q_C defined in equation (9) and the LRT $-2\log\lambda$, where λ is defined in equation (8).

2. Generate $(\bar{Y}_i^*, S_i^{*2})$ from equation (3) and equation (4), respectively, with true parameters $(\tilde{\mu}_{ML}\ ,\tilde{\sigma}^2_{i(ML)},\ n_i)$, $i = 1, \cdots, k$.

3. Compute Cochran's chisquare test statistic Q_C^* and $-2\log\lambda^*$ based on $\{(\bar{Y}_i^*, S_i^{*2}, n_i), i = 1, \cdots, k\}$.

4. Repeat steps 2 and 3 B times to get $Q^*_{C1}, Q^*_{C2}, \ldots, Q^*_{CB}$, and $-2\log\lambda_1^*, -2\log\lambda_2^*, \ldots, -2\log\lambda_B^*$.

5. Find the 95% percentile of $Q^*_{C1}, \ldots, Q^*_{CB}$ and the 95% percentile of $2\log\lambda_1^*, \ldots, -2\log\lambda_B^*$, namely, Q_S and $-2\log\lambda_S$, respectively.

6. Reject H_0 if $Q_C > Q_S$. We call this the bootstrap test based on Cochran's test, and denote it by $Q_{C,boot}$. Similarly, reject H_0 if $-2\log\lambda > -2\log\lambda_S$. We call this the bootstrap test based on LRT, and denote it by LRT_{boot}.

Type I error rate. To study the properties of the above tests, namely, Q_C (Cochran's test), LRT (Likelihood Ratio Test), $Q_{C,boot}$ (Bootstrap test based on Cochran's test statistic), and LRT_{boot} (Bootstrap test based on likelihood ratio), we performed Monte Carlo simulations under the null hypothesis of homogeneity of means. Tongsai et al. (2010) carried out an extensive Monte Carlo simulation to investigate the performance of Cochran's test and likelihood ratio test of homogeneity in regard to Type I error rates under various scenarios of sample sizes and variances. We remark that the various cases considered in Tongsai et al. (2010) are essentially those used in Hartung, Argaç, and Makambi (2002), and these are reproduced below in Table 5 for $k = 9$ populations. For $k = 3$ populations, we take the patterns of the columns $i = 1, 2, 3$ from Table 5, and for $k = 6$, the patterns of the columns $i = 1, \ldots, 6$ from Table 5 are used.

The simulation study was carried out using the statistical software R (2010) (Version 2.11.0, 2010-04-22). We are mainly interested in Type I error rates, given the nominal Type I error of

$\alpha = 0.05$. Each estimated Type I error rate is based on 10,000 simulation runs; for bootstrap method, we select 300 independent bootstrap samples. Our simulation results are shown in Table (6) through Table (8).

It is obvious from the above tables that both the test procedures Q_C and LRT fail to maintain the nominal type I error rate in all the cases considered. The last two columns in the above tables represent bootstrapped attained significance levels of Q_C and LRT, denoted as $Q_{C,boot}$ and LRT_{boot}, and the conclusions are obvious.

Real datasets. In the context of inference about a common mean, two well known examples given below repeatedly appear in the literature. We analyze them further from the point of view of maintenance of Type I error rate of two standard tests.

Example 1. Meier (1953) (reanalyzed in Jordan and Krishnamoorthy, 1996) considered four experiments about the percentage of albumin in plasma protein in human subjects, and reported the following statistics.

Example 2. Eberhardt, Reeve, and Spiegelman (1989) estimated mean Selenium in nonfat milk powder by combining the results of four methods. Details of their statistics are given below.

For each of the above data sets, we computed Q_C, LRT as well as their bootstrapped versions $Q_{C,boot}$ and LRT_{boot}. The values of test statistics and related P-values appear above. Although in all the cases, the P values are high, leading to acceptance of a common mean, the bootstrapped P values (based on $10,000$ simulation runs) turn out to be higher than the original P-values in both the data sets.

4 Conclusion

In conclusion and summary, in this paper we have demonstrated that bootstrap procedures better control type I error rate than commonly used procedures for testing homogeneity of bivariate correlations and normal means. We believe that such conclusions would hold in other similar situations.

Acknowledgement

Our sincere thanks are due to a reviewer for some helpful comments. This collaborative research was funded through a research contract from Merck & Co.

References

[1] Cochran, W.G. (1937). Problems arising in the analysis of a series of similar experiments. Journal of the Royal Statistical Society (Suppl.), 4, 102–118.

[2] Eberhardt, K.R., Reeve, C.P., and Spiegelman, C.H. (1989). A minimax approach to combining means, with practical examples. Chemometrics and Intelligent Laboratory Systems, 5, 129–148.

[3] Hartung, J., Argaç, D., and Makambi, K.H. (2002). Small sample properties of tests on homogeneity in one-way anova and meta-analysis. Statistical Papers, 43, 197–235.

[4] Hartung, J.,Knapp, G., and Sinha, B.K. (2008). Statistical Meta-Analysis with Applications. Wiley, New York.

[5] Jordan, S.M. and Krishnamoorthy, K. (1996). Exact confidence intervals for the common mean of several normal populations. Biometrics, 52, 77-86.

[6] Meier, P. (1953). Variance of a weighted mean. Biometrics, 9, 59–73.

[7] Rao, C.R. (1973). Linear statistical inference and its applications, second edition. John Wiley & Sons.

[8] R Development Core Team (2009). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0, URL http://www.R-project.org.

[9] Mathew, T., Sinha, B.K., Shah, A., Jianxin, L., Sharma, G. and Xu, D. (2010). Meta-Analysis and Cumulative Meta-Analysis for Correlations: Methodology and Applications in a Study to Assess the Efficacy of Ezetimibe Coadministered with Statin in Patients with Hypercholesterolemia. Technical Report, Department of Mathematics and Statistics, University of Maryland Baltimore County.

[10] Tongsai, S., Knapp, G., Sinha, B.K. and Wattanachayakul, S. (2010). On Testing Equality of Several Normal Means. Journal of Statistical Theory and Applications, 9, 459.

[11] Welch, B.L. (1951). On the comparison of several mean values: an alternative approach. Biometrika, 38, 330–336.

Appendix

Table 1: attained significance level in real data set when $k = 29$

ρ	T_C	T_Z	T_{boot}
0.1	6.1	5.3	5.6
0.2	5.7	4.8	5.1
0.3	6.3	5.3	5.6
0.4	6.0	5.0	5.3
0.5	6.2	5.3	5.5
0.6	6.1	5.2	5.5
0.7	5.8	4.8	5.1
0.8	6.1	5.1	5.5
0.9	6.4	5.3	5.7

Table 2: attained significance level when $N = 20$, $k = 32$

ρ	T_C	T_Z	T_{boot}
0.1	41.1	5.5	5.6
0.2	38.6	5.4	6.0
0.3	36.0	5.7	6.8
0.4	30.9	5.1	7.0
0.5	26.3	5.0	7.7
0.6	21.6	5.1	7.9
0.7	16.5	5.0	7.5
0.8	11.1	4.7	6.8
0.9	7.8	4.3	6.1

Table 3: attained significance level when $N = 20$, $k = 8$

ρ	T_C	T_Z	T_{boot}
0.1	17.3	4.9	5.2
0.2	17.5	5.4	5.8
0.3	16.0	5.0	6.3
0.4	14.8	5.2	7.1
0.5	12.5	4.9	7.2
0.6	9.7	5.0	7.0
0.7	7.7	5.1	7.5
0.8	5.3	4.8	7.0
0.9	3.3	4.8	6.3

Table 4: attained significance level when $N = 100$, $k = 32$

ρ	T_C	T_Z	T_{boot}
0.1	9.3	5.1	5.3
0.2	9.1	5.2	5.4
0.3	8.9	5.2	5.4
0.4	8.7	4.7	5.5
0.5	8.3	5.0	5.5
0.6	8.0	5.5	5.5
0.7	6.8	4.9	5.4
0.8	6.5	4.8	5.5
0.9	5.0	4.4	4.7

Table 5: Sample designs for k=9 populations

		k=9								
Pattern	i	1	2	3	4	5	6	7	8	9
1	n_i	5	5	5	5	5	5	5	5	5
	σ_i^2	2	6	10	2	6	10	2	6	10
2	n_i	10	10	10	10	10	10	10	10	10
	σ_i^2	2	6	10	2	6	10	2	6	10
3	n_i	5	10	15	5	10	15	5	10	15
	σ_i^2	2	6	10	2	6	10	2	6	10
4	n_i	10	20	30	10	20	30	10	20	30
	σ_i^2	2	6	10	2	6	10	2	6	10
5	n_i	5	10	15	5	10	15	5	10	15
	σ_i^2	10	6	2	10	6	2	10	6	2
6	n_i	10	20	30	10	20	30	10	20	30
	σ_i^2	10	6	2	10	6	2	10	6	2

Table 6: Attained significance level (in %) of homogeneity tests given a nominal level of $\alpha = 0.05$ for k=3 populations

Pattern	Q_C	LRT	$Q_{C,boot}$	LRT_{boot}
1	12.9	10.2	4.8	5.2
2	8.4	7.2	5.2	5.3
3	8.6	7.7	4.9	5.1
4	7.0	6.6	5.6	5.6
5	12.2	9.5	5.7	5.6
6	7.3	6.2	4.9	5.0

Table 7: Attained significance level (in %) of homogeneity tests given a nominal level of $\alpha = 0.05$ for k=6 populations

Pattern	Q_C	LRT	$Q_{C,boot}$	LRT_{boot}
1	22.1	14.4	4.7	5.2
2	11.9	8.9	5.5	5.6
3	14.2	10.5	5.6	5.8
4	8.7	7.3	5.3	5.4
5	16.3	10.9	5.3	5.3
6	9.3	7.0	5.0	4.9

Table 8: Attained significance level (in %) of homogeneity tests given a nominal level of $\alpha = 0.05$ for k=9 populations

Pattern	Q_C	LRT	$Q_{C,boot}$	LRT_{boot}
1	30.4	18.6	5.1	5.3
2	14.5	9.6	5.4	5.4
3	17.8	11.8	5.3	5.3
4	9.6	7.6	4.9	4.8
5	21.2	12.6	5.4	5.3
6	11.1	8.2	5.2	5.2

Table 9: Percentage of albumin in plasma protein

Experiment	n_i	Mean	Variance
A	12	62.3	12.986
B	15	60.3	7.840
C	7	59.5	33.433
D	16	61.5	18.513

Table 10: Test results in Example 1

Methods	Test Statistics	P value
Q_C	3.19	0.36
LRT	3.32	0.34
$Q_{C,boot}$	3.19	0.42
LRT_{boot}	3.32	0.41

Table 11: Selenium in nonfat milk powder

Methods	n_i	Mean	Variance
Atomic absorption spectrometry	8	105.0	85.711
Neutron activation:			
1). Instrumental	12	109.75	20.748
2). Radiochemical	14	109.5	2.729
Isotope dilution mass spectrometry	8	113.25	33.640

Table 12: Test results in Example 2

Methods	Test Statistics	P Value
Q_C	5.21	0.16
LRT	5.04	0.17
$Q_{C,boot}$	5.21	0.23
LRT_{boot}	5.04	0.23