BIOMETRICS 47, 757–762
June 1991

THE CONSULTANT'S FORUM

Two Guidelines for Bootstrap Hypothesis Testing

Peter Hall and Susan R. Wilson

Centre for Mathematics and Its Applications,
Australian National University, G.P.O. Box 4,
Canberra, A.C.T. 2601, Australia

SUMMARY

Two guidelines for nonparametric bootstrap hypothesis testing are highlighted. The first recommends that resampling be done in a way that reflects the null hypothesis, even when the true hypothesis is distant from the null. The second guideline argues that bootstrap hypothesis tests should employ methods that are already recognized as having good features in the closely related problem of confidence interval construction. Violation of the first guideline can seriously reduce the power of a test. Sometimes this reduction is spectacular, since it is most serious when the null hypothesis is grossly in error. The second guideline is of some importance when the conclusion of a test is equivocal. It has no direct bearing on power, but improves the level accuracy of a test.

1. Introduction

The nonparametric bootstrap is a particularly versatile tool for data analysis. Its good performance in many important statistical problems has been established by theoretical analysis, by simulation study, and by application to real data. DiCiccio and Romano (1988) and Hinkley (1988) give recent surveys with discussions. However, use of nonparametric bootstrap methods in hypothesis testing problems is not nearly so well understood as many other applications.

We have found it useful to identify two fundamental guidelines which, we argue, should guide the implementation of a bootstrap hypothesis test. The first guideline says that care should be taken to ensure that even if the data might be drawn from a population that fails to satisfy H_0, resampling is done in a way that reflects H_0. The second guideline argues that bootstrap hypothesis testing should use methods that are already recognized as having good features in the closely related problem of confidence interval construction. This often amounts to pivoting, which usually means correcting for scale.

In a sense, both of these guidelines have been mentioned before. For example, Young (1986), Beran (1988), Hinkley (1988, 1989), and Hall and Fisher (unpublished manuscript) have suggested that resampling should reflect the null hypothesis; Efron (1987), DiCiccio and Romano (1988), and Hall (1988) have discussed the drawbacks of nonpivotal methods such as percentile. Silverman (1981) has suggested using bootstrap methods to test for multimodality of a probability density. However, these relatively technical contributions are not easily accessible to biometricians, and so their theoretical recommendations can go unheeded in practice. Indeed, the present note was partly motivated by an article of Wahrendorf, Becher, and Brown (1987), wherein both guidelines are ignored. That paper is proving influential in the biometric literature; see, for example, Cole and McDonald (1989).

The first guideline can have a profound effect on power. If it is not adhered to, then the resulting bootstrap test may have very low power, even against alternatives that are a

Key words: Bootstrap; Goodness of fit; Hypothesis testing; Level accuracy; Power.

considerable distance from the null. The second guideline does not have a direct bearing on power, but does influence coverage accuracy. Therefore it will usually affect the conclusion of a test only when the result is rather equivocal.

The references cited above do not adequately address the problem that if the null and alternative hypotheses are markedly dissimilar then it can be quite difficult, in practice, to ensure that resampling is conducted as though H_0 were correct. For example, this will be the case if one is testing for the appropriateness of two quite different models. Such a situation would usually mean considering an alternative approach to the testing problem, and two of our examples highlight this difficulty.

Section 2 will briefly discuss the first and second guidelines. We shall focus on the simple parameter testing problem, so as to make our account as succinct as possible. Section 3 will analyse three examples that illustrate our main points.

2. Two Guidelines

Assume for the sake of simplicity that the hypothesis under test is H_0: $\theta = \theta_0$, to be tested against the two-sided alternative H_1: $\theta \neq \theta_0$. We treat the nonparametric bootstrap, which involves resampling from the sample. The examples in Section 3 will treat cases that are more complex than simple hypotheses about parameter values, and show that the guidelines continue to be relevant in those circumstances.

Let $\hat{\theta}$, a function of the sample $X_1, \ldots, X_n$, denote an estimator of the unknown quantity θ, and write $\hat{\theta}^*$ for the value of $\hat{\theta}$ computed for a resample $X_1^*, \ldots, X_n^*$ drawn from the sample with replacement. A test of H_0 against H_1 would usually be based on the difference $\hat{\theta} - \theta_0$, whose distribution under H_0 we would wish to estimate. Most importantly, we must estimate this distribution under H_0, even when the sample is drawn from a population that fails to satisfy H_0. This leads to our first guideline, which here has the following form.

First Guideline: Resample $\hat{\theta}^* - \hat{\theta}$, not $\hat{\theta}^* - \theta_0$.

To appreciate why this is important, observe that the test will involve rejecting H_0 if $|\hat{\theta} - \theta_0|$ is "too large." If θ_0 is a long way from the true value of θ (i.e., if H_0 is grossly in error) then the difference $|\hat{\theta} - \theta_0|$ will never look very much too big compared to the nonparametric bootstrap distribution of $|\hat{\theta} - \theta_0|$. A more meaningful comparison is with the bootstrap distribution of $|\hat{\theta}^* - \hat{\theta}|$. In fact, if the true value of θ is θ_1 then the power of the bootstrap test increases to 1 as $|\theta_1 - \theta_0|$ increases, provided the test is based on resampling $|\hat{\theta}^* - \hat{\theta}|$, but the power decreases to at most the significance level (as $|\theta_1 - \theta_0|$ increases) if the test is based on resampling $|\hat{\theta} - \theta_0|$.

Thus, the first guideline of bootstrap hypothesis testing has the effect of increasing power. The second guideline, which we introduce next, reduces error in the level of significance. (Level error is defined as the difference between the actual significance level of a bootstrap test, and the nominal level such as 5%.) The second guideline is essentially the hypothesis testing version of an idea that is currently receiving attention in the literature on confidence intervals. To explain the second guideline, let $\hat{\sigma}$ denote an estimate of the scale of $\hat{\theta}$, with the property that the distribution of $(\hat{\theta} - \theta_0)/\hat{\sigma}$ is virtually free of unknowns when H_0 is true. For example, it might be approximately normal $N(0, 1)$. Let $\hat{\sigma}^*$ denote the value of $\hat{\sigma}$ computed for the resample rather than the sample. Then the bootstrap distribution of $(\hat{\theta}^* - \hat{\theta})/\hat{\sigma}^*$ estimates the distribution of $(\hat{\theta} - \theta_0)/\hat{\sigma}$ under the null hypothesis. This leads us to the second guideline.

Second Guideline: Base the test on the bootstrap distribution of $(\hat{\theta}^* - \hat{\theta})/\hat{\sigma}^*$, not on the bootstrap distribution of $(\hat{\theta}^* - \hat{\theta})/\hat{\sigma}$ or of $\hat{\theta}^* - \hat{\theta}$. That is, for a test at the 5% level first compute a number $\hat{t}$ such that

$$\Pr{}^*(|\hat{\theta}^* - \hat{\theta}|/\hat{\sigma}^* > \hat{t}) = .05 \tag{2.1}$$

(where Pr* denotes probability measure under the bootstrap distribution), and then reject H_0 in favour of H_1 if $|\hat{\theta} - \theta_0|/\hat{\sigma} > \hat{t}$. Note that we have incorporated the first guideline here, by working with $(\hat{\theta}^* - \hat{\theta})/\hat{\sigma}^*$ rather than $(\hat{\theta}^* - \theta_0)/\hat{\sigma}^*$.

The rationale behind the second guideline is that the bootstrap distribution of $T^* = (\hat{\theta}^* - \hat{\theta})/\hat{\sigma}^*$ is a better approximation to the distribution of $T^* = (\hat{\theta} - \theta_0)/\hat{\sigma}$ under H_0, than the bootstrap distribution of $S^* = \hat{\theta}^* - \hat{\theta}$ is to the distribution of $S = \hat{\theta} - \theta_0$ under H_0. Intuitively, the reason is that the asymptotic distributions of S and S^* depend on the unknown scale, and there are significant differences between the appropriate scale factors in the cases of S and S^*; see Hall (1988), for example.

The device of dividing by $\hat{\sigma}^*$ is known as "bootstrap pivoting," because it produces a statistic $(\hat{\theta}^* - \hat{\theta})/\hat{\sigma}^*$ that is asymptotically pivotal—i.e., whose asymptotic distribution does not depend on any unknowns. This technique may be used whenever a good variance estimate $\hat{\sigma}^2$ is available. (In this context, "good" means that $\hat{\sigma}^2$ should have low variance.) However, there do exist circumstances where σ^2 cannot be estimated well—examples include cases where θ is a ratio of two or more unknowns, such as a ratio of two means or a correlation coefficient. Estimation of σ^2 can also be difficult when the number of nuisance parameters is unduly large, and there it might be appropriate to disregard the second guideline, or to use a technique such as accelerated bias correction suggested by Efron (1987).

The arguments above should be modified in obvious ways when the parametric bootstrap is in use, or when one is testing a composite null hypothesis against a one-sided alternative hypothesis.

3. Three Examples

The first example discusses part of data set D of Cox and Snell (1981), which is discussed further there. To help the uninitiated reader we give extra detail about the bootstrap algorithm for this example. In Table 1 we present 20 readings of temperature. Suppose we are interested in testing hypotheses concerning the mean temperature, θ. Then for these data, $\hat{\theta} = 454.6$ and its estimated standard error is $\hat{\sigma} = 4.034$.

To illustrate our first guideline, consider testing the hypothesis H_0: $\theta = \theta_0 = 440$, against the two-sided alternative H_1: $\theta \neq \theta_0$. We drew $B = 499$ bootstrap resamples and found the value of $\hat{t}$ such that $\Pr{}^*(|\hat{\theta}^* - \theta_0| > \hat{t}) = \alpha$; for $\alpha = 0.05$, $\hat{t}$ was 21.0, and for $\alpha = .10$ it was 19.7. [Specifically, $\hat{t}$ was the 24th largest of the 499 values of $|\hat{\theta}^* - \theta_0|$ when $\alpha = .05$, and the 49th largest when $\alpha = .10$. If $\hat{t}$ were taken to be the νth largest then the corresponding value ofα would be $(\nu + 1)/(B + 1)$.] Since $\hat{\theta} - \theta_0 = 14.6$, this approach, which ignores both the first and second guidelines, would find our null hypothesis acceptable.

Applying the first guideline, but not the second, we determined the value of $\hat{t}$ such that

$$\Pr{}^*(|\hat{\theta}^* - \hat{\theta}| > \hat{t}) = \alpha; \tag{3.1}$$

Table 1
20 *readings of temperature, in* °C (*from Cox and Snell,* 1981).

431	450	431	453	481	449	441	476	460	482
472	465	421	452	451	430	458	446	466	476

for $\alpha = .05$, $\hat{t}$ was 7.4, and for $\alpha = .01$ it was 9.3. (On this occasion $\hat{t}$ was the 24th largest value of $|\hat{\theta}^* - \hat{\theta}|$ when $\alpha = .05$, and the 49th largest value when $\alpha = .10$.) This analysis leads us to reject the null hypothesis, in marked contrast to the conclusion that we reached when we ignored the first guideline. The loss of power described in the previous section increases as the null hypothesis becomes more extreme. (For example, consider the above argument with $\theta_0 = 0$!)

If we use the second as well as the first guideline in the analysis in the previous paragraph, we come to the same conclusion: the null hypothesis is not acceptable. [Here we compute $\hat{t}$ such that $\Pr^*(|\hat{\theta}^* - \hat{\theta}|/\hat{\sigma}^* > \hat{t}) = \alpha$, where $\hat{\sigma}^*$ is the standard deviation of a resample. When $\alpha = .05$, $\hat{t}$ is the 24th largest of the 499 values of $|\hat{\theta}^* - \hat{\theta}|/\hat{\sigma}^*$.] This concurrence is hardly surprising, since we know from the argument in Section 2 that the second guideline will have an effect only when the differences between H_0 and H_1 are somewhat equivocal.

We used balanced bootstrap resampling (Davison, Hinkley, and Schechtman, 1986) in the Monte Carlo experiment that gave these figures. That is, each set of $B = 499$ resamples used each of the sample observations equally often. This improved the efficiency of the Monte Carlo algorithm. The NAG subroutine G05EHF was employed to permute the vector containing the n observations, copied B times.

For nonstatistical readers, a reviewer suggested we draw to their attention a more general but analogous situation, namely that of a linear model. Here, resampling would usually be from a set of centred residuals, rather than directly from the sample.

Our next two examples illustrate the major role that the first guideline plays in determining the power of a bootstrap test. If it should happen that the null hypothesis is false in a rather striking manner, but no attempt is made to resample in a way that emulates what would happen if the null hypothesis were true, then it may well be the case that a bootstrap test will fail to reject the null hypothesis. There do exist circumstances where it is particularly difficult to apply the first guideline to ensure respectable power, as will be described, and an alternative approach should be considered.

The second example is one of those used by Wahrendorf et al. (1987). The data are given in Table 2. Let λ_{jk} be the death rate in each age group j ($j = 1, \ldots, 4$) for nonsmokers ($k = 1$) and smokers ($k = 2$). The multiplicative model takes the form $\lambda_{jk} = \exp(\alpha_j + \beta z)$, where the effect of smoking z ($= k - 1$) is considered as a covariate. For the additive model this representation is $\lambda_{jk} = \alpha_j + \beta z$. The hypothesis under consideration is that the multiplicative model and additive model fit the data equally well.

The denominators of the death rates, the person-years, are considered as fixed constants, whereas the numerators, the numbers of deaths, are considered to be Poisson random variables. The likelihood ratio goodness-of-fit statistic, with 3 degrees of freedom, for the multiplicative model has value $M = 13.05$ and for the additive model is $A = 3.37$.

Initially, resampling of the original data was done in a parametric fashion from the saturated

Table 2
Death rates from coronary heart disease among British male doctors (from Breslow, 1985)

	Person-years		Coronary deaths	
Age	Nonsmokers	Smokers	Nonsmokers	Smokers
35–54[a]	29,463	95,655	14	136
55–64	5,710	28,612	28	206
65–74	2,585	12,663	28	186
75–84	1,462	5,317	31	102

[a] The original two age categories have been combined for comparability with previous bootstrap results.

model, as described by Wahrendorf et al. (1987). Our program was based on the NAG subroutine G05ECF. To each bootstrap sample the multiplicative and additive models were fitted, yielding estimates M^* and A^*. These authors based their evaluation of the hypothesis that the two models fit equally well on the bootstrap distribution of $|\hat{\theta}^*|$, where $\hat{\theta} = M - A$. As one would expect from the argument presented above, the mean of this bootstrap distribution is very close to the data value, $\hat{\theta} = 9.68$. So, ignoring the first guideline completely, the resulting statistic has virtually no power.

It is not straightforward to apply the first guideline properly here, because the null and alternative hypotheses are entirely disjoint. The naive application—namely, considering the statistic $|\hat{\theta}^* - \hat{\theta}|$, improves the power, but it is still low. For these data, the estimated 10% and 5% points of this statistic's bootstrap distribution are 12.2 and 15.1.

One appropriate evaluation of the two models is to simulate *separately* under fitted multiplicative and additive models. Using the bootstrap to estimate (separately) the significance levels in the case of each statistic, M and A, gives values $\alpha = .005$ for M and $\alpha = .15$ for A. This indicates that the additive model is satisfactory while the multiplicative model is not. Breslow (1985) reached this conclusion using an alternative approach, namely embedding the two models in a parametric family that contains both as special cases.

It is worth pointing out that the likelihood ratio statistic is internally pivoted or ''Studentized'' —its asymptotic distribution under the null hypothesis does not depend on any unknown parameters. Hence the second guideline is automatically applied in this case. [Recall from Section 2 that the second guideline removes the effect of unknowns, such as the σ in formula (2.2a), from null distributions.]

Our third and final example also illustrates the lack of power that results from disregarding the first guideline. The data are given in Table 3 (from Freeman, 1987) and show LI values (a measure of cell activity) for 27 cancer patients of whom 9 went into remission. Consider a logistic model for the probability, p, of remission, $\ln\{p/(1-p)\} = \alpha + \beta x$, where x is the LI value. The value of the likelihood ratio statistic under the null hypothesis H_0: $\beta = 0$ is $\hat{L} = 8.3$, on 1 degree of freedom. Nonparametric balanced bootstrap resampling, simulating the distribution of $\hat{L}^* - \hat{L}$, yielded estimates 9.6 and 13.1 of the 10% and 5% points, respectively.

This analysis provides very little evidence against H_0. However, note that our resampling scheme has violated the first guideline. Under H_0 the distribution of $\hat{L}$ is approximately central chi-squared, but should H_0 be false the distribution would be noncentral chi-squared, with higher variance than central chi-squared. Therefore, if H_0 is false then the distribution with which we have compared $\hat{L}$ has substantially higher variability than the distribution that $\hat{L}$ would enjoy under H_0. (Centring $\hat{L}^*$ at $\hat{L}$ does not remove this difficulty.) In consequence, we would once again expect the bootstrap test to have low power.

In fact, a more sophisticated analysis using results from Davison (1988) and Reid (1988) indicates that inclusion of LI in the linear part of the model is *highly* statistically significant (beyond the 1% level, and confirmed by simulation). This result delineates the low power of the naive bootstrap test. Substantive considerations indicate that a threshold model is quite reasonable. There, cells having an LI value above a threshold might respond to treatment, and hence lead to remission, while those below would not. Further analysis and simulation indicate that this model provides an even better explanation for the data than the linear logistic model. Exploration

Table 3
Cancer remission (counts) and LI *values*

LI value	.4	.5	.6	.7	.8	.9	1.0	1.1	1.2	1.3	1.4	1.6	1.7	1.9
Remission	0	0	0	0	0	1	2	1	0	1	1	0	1	2
No. patients	2	2	3	3	3	1	3	2	1	1	1	1	1	3

of the data in this direction was certainly not encouraged, or even suggested, by our original bootstrap test.

Acknowledgements

We thank Ms. Yvonne Pittelkow for computational assistance, and the three reviewers for their helpful comments.

Résumé

On apporte un nouvel éclairage, par deux guides pour effectuer des test d'hypothèses nonparamétriques. Le premier recommande de faire le rééchantillonnage de façon à refléter l'hypothèse nulle, même si l'hypothèse vraie en est éloignée. Le second présente l'argumentation suivante: les tests d'hypothèse cyrano doivent utiliser des méthodes dont on sait qu'elles ont un bon comportement pour le problème proche de la construction d'intervalles de confiance. La violation du premier guide peut réduire de façon drastique la puissance du test. Quelquefois, cette réduction est spectaculaire, puisqu'elle est plus importante quand l'hypothèse nulle est grossièrement inexacte. Le second a une importance notable quand la conclusion du test est ambigüe. Il n'a pas d'effet direct sur la puissance, mais il améliore la précision du niveau du test.

References

Beran, R. (1988). Prepivoting test statistics: A bootstrap view of asymptotic refinements. *Journal of the American Statistical Association* **83**, 682–697.

Breslow, N. (1985). Cohort analysis in epidemiology. In *A Celebration of Statistics*, A. C. Atkinson and S. E. Fienberg (eds), 109–143. New York: Springer-Verlag.

Cole, M. J. and McDonald, J. W. (1989). Bootstrap goodness-of-link testing in generalized linear models. In *Statistical Modelling*, A. Decarli, B. J. Francis, R. Gilchrist, and G. U. H. Seber (eds), 84–94. Lecture Notes in Statistics **57**. New York: Springer-Verlag.

Cox, D. R. and Snell, E. J. (1981). *Applied Statistics: Principles and Examples*. London: Chapman and Hall.

Davison, A. C. (1988). Approximate conditional inference in generalised linear models. *Journal of the Royal Statistical Society*, *Series B* **50**, 445–461.

Davison, A. C., Hinkley, D. V., and Schechtman, E. (1986). Efficient bootstrap simulation. *Biometrika* **74**, 555–566.

Diciccio, T. J. and Romano, J. P. (1988). A review of bootstrap confidence intervals (with Discussion). *Journal of the Royal Statistical Society*, *Series B* **50**, 338–354.

Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association* **82**, 171–200.

Freeman, D. H. (1987). *Applied Categorical Data Analysis*. New York: Marcel Dekker.

Hall, P. (1988). Theoretical comparison of bootstrap confidence intervals (with Discussion). *Annals of Statistics* **16**, 927–985.

Hinkley, D. V. (1988). Bootstrap methods (with Discussion). *Journal of the Royal Statistical Society*, *Series B* **50**, 321–337.

Hinkley, D. V. (1989). Bootstrap significance tests. In *Proceedings of the 47th Session of the International Statistical Institute*, Paris, 29 August–6 September 1989, **3**, 65–74.

Reid, N. (1988). Saddlepoint methods and statistical inference (with Discussion). *Statistical Science* **3**, 213–238.

Silverman, B. W. (1981). Using kernel density estimates to investigate multimodality. *Journal of the Royal Statistical Society*, *Series B* **43**, 97–99.

Wahrendorf, J., Becher, H., and Brown, C. C. (1987). Bootstrap comparison of non-nested linear models: Applications in survival analysis and epidemiology. *Applied Statistics* **36**, 72–81.

Young, A. (1986). Conditional data-based simulations: Some examples from geometrical statistics. *International Statistical Review* **54**, 1–13.

Received August 1989; *revised February* 1990; *accepted April* 1990.