Resampling Tests of Statistical Hypotheses

A. Young, Cambridge

1. Introduction

A direct approach to statistical tests of significance is possible using bootstrap techniques to simulate the null distribution of the test statistic of interest. The approach is outlined by Hinkley(1988) and by Young(1986). In this paper issues involved in the construction of such tests are considered in the context of testing the mean of a univariate population. The general method is summarized in Section 2, while questions relating to choice of reference distribution and test statistic are considered and illustrated empirically in Section 3. Section 4 discusses the importance of appropriate conditioning in resampling tests of significance.

The approach has much in common with the Monte Carlo test of Barnard(1963), though the methods considered here contrast with more conventional simulation tests through the interventionalist role played by the sample data in specification of the resampling mechanism.

2. Resampling Tests

Suppose a test statistic T is proposed for testing a statistical hypothesis H_0 about an underlying population F, with large values of T being evidence against H_0. Let $\underline{x} = (x_1, ..., x_n)$ denote the observed sample data and let $T(\underline{x}) = t_{obs}$. The resampling method is to be used to estimate the sampling distribution of T under the constraint imposed by H_0. Implementation of the test consists of choice of a set $\mathcal{F}$ of distributions satisfying H_0 and consideration of the sampling distribution of T under $\tilde{F}$, where $\tilde{F}$ is that member of the set $\mathcal{F}$ closest to, or most consistent with, the observed data. Formally, denoting by F_n the empirical distribution of the observed sample and supposing δ is a distance measure between distributions, the significance probability is computed by considering the distribution of T under sampling from $\tilde{F}$, where $\tilde{F}$ minimises $\delta(F_n, F)$ for $F \in \mathcal{F}$. The bootstrap test significance probability corresponding to t_{obs} is then

$$p = Pr\{T \geq t_{obs} \mid \tilde{F}\}.$$

Compstat 1988

In general p will be estimated by simulating B 'bootstrap data samples' from $\tilde{F}$ to produce B values $T_1, ..., T_B$ of the test statistic. The estimate of p is then

$$\hat{p} = B^{-1} \sharp \{T_i \geq t_{obs}\}.$$

The simulation size B should be chosen large enough so that $\hat{p}$ is an accurate estimate of the true bootstrap significance probability p when p is close to the nominal level of the test (c.f. Marriott, 1979). For a test of nominal size 0.05, for example, 500 to 1000 simulations would be advisable.

3. Tests on the Mean

Let $\underline{x}$ represent an i.i.d. sample from a univariate distribution F with mean μ and suppose it is required to test $H_0 : \mu = \mu_0$, for specified μ_0.
One possibilty for choice of the bootstrap reference distribution $\tilde{F}$ is to embed F_n in a class of distributions whose support is $x_1, ..., x_n$: see Efron(1981) and Owen(1988). The distribution $\tilde{F}$ will attach probabilities $w_1, ..., w_n$ to $x_1, ..., x_n$, where, since $\tilde{F} \in \mathcal{F}$, $\sum_{i=1}^n w_i x_i = \mu_0$, and the w_i are chosen, say, to

(1) minimize the Kullback-Leibler distance $\delta(\tilde{F}, F_n) = \sum_{i=1}^n w_i log(nw_i)$, or,
(2) maximize $\prod_{i=1}^n w_i$, or equivalently minimize $\delta(\tilde{F}, F_n) = -\sum_{i=1}^n n^{-1} log(nw_i)$, or
(3) minimize $\delta(\tilde{F}, F_n) = max_i \mid w_i - \frac{1}{n} \mid$.

In the first case the optimal weights w_i are given by $w_i = e^{\lambda x_i} / \sum_{i=1}^n e^{\lambda x_i}, i = 1, ..., n$, where λ is uniquely defined by $\sum_{i=1}^n x_i e^{\lambda x_i} / \sum_{i=1}^n e^{\lambda x_i} = \mu_0$: see Efron(1981). The distribution $\tilde{F}$ is then the 'exponentially tilted' version of F_n.
The optimal weights w_i in the second case are given by $w_i = n^{-1}\{1 + \lambda(x_i - \mu_0)\}^{-1}$, where λ is the unique solution of $\sum_{i=1}^n \{1+\lambda(x_i-\mu_0)\}^{-1}(x_i-\mu_0) = 0$: see Owen(1988). The distribution $\tilde{F}$ is then the non-parametric constrained maximum likelihood estimator of F.
The optimal weights in the third case may be obtained as the solution of a simple linear programming problem. Numerically, such a construction for $\tilde{F}$ is less attractive than the one- dimensional optimization required in the other two cases. In the simulation study below only the first two methods, referred to as 'TILT' and 'MLE' respectively, are considered.

A simpler method for construction of $\tilde{F}$ allows use of distributions of modified support. We might consider, for example, taking $\tilde{F}$ as a 'shifted' version of F_n, with the hypothesised mean μ_0. Then $\tilde{F}$ attaches equal weight n^{-1} to each of the points

$x_i - \bar{x} + \mu_0, i = 1, ..., n$,where $\bar{x} = n^{-1} \sum_{i=1}^{n} x_i$. This method is refered to as 'SHIFT' below.

By analogy with the usual parametric situation, T might be taken as the one-sample t-statistic. Performance of tests based on this statistic and bootstrap data distributions $\tilde{F}$ constructed by the above methods are considered for two underlying distributions, the standard normal distribution and the exponential distribution of variance 1 centered at mean 0, and for sample size 20. For both distributions a two-term Edgeworth expansion was found to give an adequate approximation to the distribution of T under $\tilde{F}$, and therefore the bootstrap significance probabilities could be estimated without simulation. For each F, 20000 replications were used to compare performances of tests of nominal size 0.05. Results are summarized in Tables 1 and 2.

$\mu_0 =$	0.0	-0.2	-0.4
TILT	0.053	0.219	0.536
MLE	0.053	0.221	0.538
SHIFT	0.048	0.204	0.513

Table 1 : Simulated powers, normal distribution.

$\mu_0 =$	0.0	-0.2	-0.4
TILT	0.038	0.230	0.686
MLE	0.039	0.234	0.673
SHIFT	0.032	0.206	0.664

Table 2 : Simulated powers, exponential distribution.

The results show that there is little to choose between the TILT and MLE procedures, but that the simpler SHIFT procedure is perhaps less powerful. In the exponential case, all three resampling tests have observed sizes substantially lower than the nominal size.

The success of a resampling test will depend crucially on the choice of test statistic T. A general procedure for choice of statistic may be based on the notion of empirical likelihood, described by Owen(1988). For the mean testing problem define the *empirical likelihood ratio* as

$$R_0 = sup \prod_{i=1}^{n} nw_i,$$

where the supremum is taken over sets of probability weights satisfying $\sum_{i=1}^{n} w_i x_i = \mu_0$. The required numerical procedure is outlined above. Theorem 1 of Owen(1988) shows that, on H_0, $T_{LR} = -2logR_0$ is asymptotically distributed as χ_1^2. A test of H_0 may

be based on T_{LR}, which is an empirical version of the usual likelihood ratio statistic. Table 3 shows simulation based estimates of the null expectation and variance of T_{LR}, for the two distributions considered previously, and for a range of sample sizes n. Each figure is based on 20000 simulations.

n	Normal		Exponential	
	$E(T_{LR})$	$Var(T_{LR})$	$E(T_{LR})$	$Var(T_{LR})$
15	1.2449	4.3483	1.6319	15.0809
20	1.1546	3.2746	1.4250	9.9239
50	1.0352	2.1778	1.1271	2.7587
100	1.0057	1.9966	1.0536	2.2567
200	1.0041	1.9847	1.0286	2.0873

Table 3 : Null Distribution of T_{LR}.

The table indicates that the chi-squared approximation will be inadequate, without Bartlett correction, for sample sizes less than about 100. For small sample sizes, the resampling method can be used to simulate the distribution of the empirical likelihood ratio statistic under suitable $\tilde{F}$. It would be most convenient to construct $\tilde{F}$ by the maximum likelihood procedure described above.

Performance of empirical likelihood ratio tests of nominal size 0.05 was studied for the two underlying distributions, using 1000 simulations from $\tilde{F}$ to estimate the bootstrap significance probability p, for each of 500 replications. In the normal case, the estimated powers were $0.052, 0.164$ and 0.396 for $\mu_0 = 0.0, -0.2, -0.4$ respectively. In the exponential case the corresponding figures were $0.068, 0.102$ and 0.378. Though our figures here are based on a smaller number simulations than used previously, and are therefore less accurate estimates of true power, they would indicate a less powerful procedure than that based on the t- statistic.

4. Conditional Inference

The principle of conditioning on ancillary statistics is an important and well-developed component of statistical inference. It is clear that appropriate conditioning is also a necessary part of the implementation of resampling tests. The idea (Young, 1986) should be to condition on those features of the data under test which reflect the propensity for extreme values of the test statistic to occur for artificial reasons. Such features may or may not be quantifiable as ancillary statistics.

Here, as discussed by Hinkley and Schechtman(1987), it is important to distinguish between an experimental ancillary, which determines a performable sub-experiment, and

a mathematical ancillary, which is a function of the random responses. In the parametric context, a statistic S would be said to be ancillary, and should be conditioned on, if its distribution does not depend on the parameter of interest, and therefore it contains no information on that parameter, but stratification of data samples by the value of S indicates the test statistic to have different distributions across strata.

In the mean testing problem of Section 3, it is possible that the data might represent a known mixture from two distributions with the same mean. Table 3 would indicate, for example, that the null distribution of the empirical likelihood ratio statistic might depend on the mixture proportion. This proportion is then an experimental ancillary. Conditioning could be achieved by construction of separate bootstrap data distributions from the two sub-samples.

Let $x_{(1)} < x_{(2)} < \ldots < x_{(n)}$ denote the order statistics of the observed sample. The configuration statistic $C = (x_{(2)} - x_{(1)}, \ldots, x_{(n)} - x_{(1)})$ has distribution not depending on μ. Let $A = n^{-1} \sum_{i=2}^{n} \{x_{(i)} - x_{(1)}\}^2$. Simulation shows that, for sample size $n = 20$, the empirical likelihood ratio statistic T_{LR} has null distribution which depends on A. For example, by simulating 20000 datasets from the standard normal distribution, it was found that:

$$E(T_{LR} \mid H_0, A \in [2.5, 3.0]) = 1.4578,$$

$$E(T_{LR} \mid H_0, A \in [5.5, 6.0]) = 0.9532.$$

The bootstrap test should therefore be conditioned on the observed value of A. Only samples from $\tilde{F}$ with A close to its observed value, say within 0.5 of its observed value, should be used in estimation of the significance probability p. If unconditional simulation from $\tilde{F}$ is to be used, this will mean throwing away a large proportion of bootstrap data samples. Empirically, this proportion is seen to be about 0.7 in the normal case.

A conditional bootstrap inference can be obtained using the notion, expressed by Kendall and Kendall(1980), of a resampling distribution with modified support and a simulation mechanism specifically designed to provide conditioning on A. The procedure here might be to assess significance of t_{obs} by simulating the distribution of T_{LR} under random displacements of F_n. Such displacements might be taken as normal: in order to simulate datasets representing H_0, but with the same configuration C and value of A as the data under test, the displacements should have mean $\mu_0 - \bar{x}$ and variance σ^2/n, where σ^2 is the variance of F.

This conditional procedure was applied to the two underlying distributions F considered previously, and for sample size 20 as before. In the normal case the conditional simulation of nominal size 0.05 gave powers 0.040, 0.160 and 0.442 against $\mu_0 = 0.0, -0.2$ and -0.4 respectively. The corresponding figures in the exponential case were 0.052,

0.202 and 0.536. Each figure here is based on 500 replications from F, 1000 simulations being performed for each replication.

5. References

Barnard, G. (1963). Contribution to discussion of Bartlett's paper. J.Roy.Statist.Soc. **B**, **25**, 294.

Efron, B. (1981). Nonparametric standard errors and confidence intervals. Can. J.Statist., **9**, 139-172.

Hinkley, D.V. (1988). Bootstrap methods. J.Roy.Statist.Soc. **A**, **151** (To appear).

Hinkley, D.V. and Schechtman, E. (1987). Conditional bootstrap methods in the mean-shift model. Biometrika, **74**, 85-93.

Kendall, D.G. and Kendall, W.S. (1980). Alignments in two-dimensional random sets of points. Adv.Appl.Probab., **12**, 380-424.

Marriott, F.H.C. (1979). Barnard's Monte Carlo test : how many simulations? Appl. Statist., **28**, 75-77.

Owen, A.B. (1988). Empirical likelihood ratio confidence intervals for a single functional. Biometrika, **75** (To appear).

Young, A. (1986). Conditioned data-based simulations : some examples from geometrical statistics. Int.Statist.Rev., **54**, 1-13.