Biometrika (1990), 77, 1, *pp.* 179-85
Printed in Great Britain

Bootstrap bias

BY G. A. YOUNG AND H. E. DANIELS

Statistical Laboratory, University of Cambridge, Cambridge CB2 1SB, U.K.

SUMMARY

Use of the bootstrap in estimation of the sampling distribution of a pivot is considered for two simple situations. It is shown by a simulation study that in each case the bootstrap is noticeably biased for small sample sizes. The techniques of computer algebra are used to obtain an exact assessment of the bias in both problems and an approximate theoretical analysis based on saddlepoint approximation is presented.

Some key words: Bias; Bootstrap; Computer algebra; Pivot; Saddlepoint approximation; Tail probability.

1. INTRODUCTION

The bootstrap resampling method of statistical estimation (Efron, 1979) operates as follows. It is required to estimate the distribution of the random variable $T(X_1, \ldots, X_m; F)$, possibly depending on the distribution F, for $X_1, \ldots, X_m$ as a random sample from F. The distribution F itself is unknown, but a random sample $x_1, \ldots, x_n$, with empirical distribution function F_n, is available from F. The bootstrap method approximates the sampling distribution of $T(X_1, \ldots, X_m; F)$ under F by that of $T(Y_1, \ldots, Y_m; F_n)$ under F_n, where $Y_1, \ldots, Y_m$ denotes a random sample of size m from F_n.

The bootstrap procedure therefore involves resampling from a distribution F_n which is of finite support, whereas F may not be. It might be expected that in certain circumstances, such as the estimation of small tail probabilities, this would lead to noticeable bias.

In this paper we study bias in bootstrap estimation of tail probabilities of the form $P(a) = \mathrm{pr}\{T(X_1, \ldots, X_m; F) > a \mid F\}$ for two situations. In the case of an underlying normal distribution, estimation for the pivot $T(X_1, \ldots, X_m; F) = \bar{X}_m - \mu$ is investigated, and in the case of an underlying exponential distribution for the ratio pivot $T(X_1, \ldots, X_m; F) = \bar{X}_m/\mu$, where $\bar{X}_m = m^{-1}\Sigma X_i$ and $\mu = E_F(X)$. In each case $P(a)$ is estimated by

$$\tilde{P}(a) = \mathrm{pr}\{T(Y_1, \ldots, Y_m; F_n) > a \mid F_n\}.$$

Further, in each case T is exactly pivotal under sampling from the underlying family of distributions, so an immediate assessment of the extent of the bias that derives from the finite support of F_n may be obtained by studying estimation for the specific cases of the standard normal distribution and the standard exponential distribution.

2. SIMULATION APPROACH

Suppose the observed dataset consists of n observations $x_1, \ldots, x_n$, with mean $\bar{x}_n = n^{-1}\Sigma x_i$. Let $Y_1, \ldots, Y_m$ denote a bootstrap sample of m observations drawn independently from the empirical distribution F_n, and let $\bar{Y}_m = m^{-1}\Sigma Y_j$.

In the normal case, the tail probability $P(a)=\mathrm{pr}\,(\bar{X}_m-\mu>a\,|\,F)$ is estimated by $\tilde{P}(a)=\mathrm{pr}\,(\bar{Y}_m-\bar{x}_n>a\,|\,F_n)$. In principle, $\tilde{P}(a)$ should be constructed by considering all n^m possible bootstrap samples. In practice, $\tilde{P}(a)$ is itself estimated by drawing a large number of bootstrap samples from F_n. Throughout the simulation, bootstrap estimators were constructed by drawing 50 000 bootstrap samples from each F_n. This sampling procedure was then repeated over different F_n to estimate $E\{\tilde{P}(a)\}$. In all, 1000 replications were performed for each of the combinations of sample size considered. All simulations were performed using NAG generation routines, on a Hewlett-Packard 9000/330 workstation. It took some days to obtain a complete set of results, though use of efficient bootstrap methods, such as importance sampling, might reduce this time considerably.

Table 1 gives sample results of the normal simulation, for two combinations of data and bootstrap sample sizes: $(m, n)=(5, 10)$ and $(20, 20)$. The comparison of immediate interest is that between the true probability $P(a)$ and the simulation estimate of $E\{\tilde{P}(a)\}$. An estimated standard error of the estimate of $E\{\tilde{P}(a)\}$ is also given in the table. The results are striking and indicate that, in each case, for small values of a, $\tilde{P}(a)$ is biased downwards as an estimator of $P(a)$. For larger values of a it is seen that $\tilde{P}(a)$ is biased upwards. The bias of the bootstrap estimator is least when $P(a)$ is about 0·01 in each case. These results are, perhaps, rather counter-intuitive, as we might expect the bootstrap on average to underestimate the mass in the extreme tails of the distribution. Breakdown of the results shows, however, that in the tail most bootstrap estimators $\tilde{P}(a)$ do indeed underestimate $P(a)$: the simulation shows that the upward bias results from occasional extreme overestimation of $P(a)$.

Table 2 gives sample results of the exponential simulation, for the same two combinations of sample sizes. Results here have been split into results for bootstrap estimation of left tail probability $P_L(a)=\mathrm{pr}\,(\bar{X}_m/\mu<a\,|\,F)$, where the bootstrap estimator is $\tilde{P}_L(a)=\mathrm{pr}\,(\bar{Y}_m/\bar{x}_n<a\,|\,F_n)$, and results for estimation of right tail probability $P_R(a)=1-P_L(a)$,

Table 1. *Simulated and theoretical expectations, normal distribution*

m, n	a	$P(a)$	$E\{\tilde{P}(a)\}$ (sim.)	Est. st. error	$E\{\tilde{P}(a)\}$ (exact)	$E\{\tilde{P}(a)\}$ (SP)
5, 10	0·1	0·41153	0·40066	0·00103	0·40138	0·40103
	0·3	0·25117	0·22807	0·00186	0·22901	0·22789
	0·5	0·13178	0·11246	0·00174	0·11220	0·11098
	0·7	0·05876	0·04928	0·00120	0·04851	0·04797
	0·9	0·02209	0·01973	0·00070	0·01909	0·01934
	1·1	0·00695	0·00735	0·00036	0·00702	0·00755
	1·3	0·00183	0·00266	0·00018	0·00247	0·00290
	1·5	0·00040	0·00094	0·00009	0·00085	0·00110
	1·7	0·00007	0·00033	0·00004	0·00029	0·00041
20, 20	0·1	0·32736	0·31809	0·00094	0·31753	0·31780
	0·2	0·18555	0·17403	0·00124	0·17331	0·17352
	0·3	0·08986	0·08226	0·00103	0·08174	0·08171
	0·4	0·03682	0·03406	0·00065	0·03381	0·03357
	0·5	0·01267	0·01254	0·00034	0·01249	0·01210
	0·6	0·00365	0·00417	0·00016	0·00420	0·00297
	0·7	0·00087	0·00129	0·00007	0·00131	0·00059
	0·8	0·00017	0·00037	0·00002	0·00038	0·00016
	0·9	0·00003	0·00010	0·00001	0·00011	0·00005

SP, saddlepoint

Table 2. *Simulated and theoretical expectations, exponential distribution*

(*a*) *Left tail probability*

m, *n*	*a*	$P_L(a)$	$E\{\tilde{P}_L(a)\}$ (sim.)	Est. st. error	$E\{\tilde{P}_L(a)\}$ (exact)
5, 10	0·1	0·00017	0·00092	0·00011	0·00085
	0·3	0·01858	0·02111	0·00090	0·02136
	0·5	0·10882	0·09349	0·00203	0·09599
	0·7	0·27456	0·23406	0·00258	0·23836
	0·9	0·46790	0·42528	0·00160	0·42905
20, 20	0·3	0·00001	0·00006	0·00001	0·00005
	0·5	0·00345	0·00529	0·00025	0·00489
	0·7	0·07650	0·07079	0·00124	0·06894
	0·9	0·34908	0·32827	0·00125	0·32620

(*b*) *Right tail probability*

m, *n*	*a*	$P_R(a)$	$E\{\tilde{P}_R(a)\}$ (sim)	Est. st. error	$E\{\tilde{P}_R(a)\}$ (exact)
5, 10	1·0	0·44049	0·47171	0·00063	0·47021
	1·3	0·22367	0·20832	0·00171	0·21091
	1·6	0·09963	0·06976	0·00134	0·07167
	1·9	0·04026	0·02281	0·00077	0·02418
	2·2	0·01510	0·00762	0·00044	0·00828
	2·5	0·00535	0·00243	0·00021	0·00276
	2·8	0·00181	0·00084	0·00011	0·00096
20, 20	1·0	0·47026	0·48005	0·00030	0·48016
	1·3	0·09682	0·08278	0·00116	0·08102
	1·6	0·00934	0·00775	0·00033	0·00740
	1·9	0·00051	0·00070	0·00006	0·00067
	2·2	0·00002	0·00007	0·00001	0·00007

where the bootstrap estimator is $\tilde{P}_R(a)=1-\tilde{P}_L(a)$. Comparison of the true population probabilities with the simulation estimates of the expected bootstrap estimate again indicates distinct and systematic bias.

3. Analytic approach

An analytic examination of $E\{\tilde{P}(a)\}$ can be performed as follows. Suppose that the moment generating function $M(\tau)$ of the underlying distribution exists for real τ in an open interval containing the origin, so that all moments exist.

The moment generating function of the empirical distribution is

$$M(\tau|x_1,\ldots,x_n)=(e^{\tau x_1}+\ldots+e^{\tau x_n})/n,$$

and the conditional moment generating function for $m\tilde{Z}=m(\bar{Y}_m-\bar{x}_n)$ is

$$E(e^{m\tau\tilde{Z}}|x_1,\ldots,x_n)=\exp\left\{-\frac{m}{n}(x_1+\ldots+x_n)\tau\right\}M^m(\tau|x_1,\ldots,x_n). \qquad (3\cdot1)$$

Davison & Hinkley (1988) use this moment generating function (3·1) to obtain a saddlepoint approximation to $\tilde{P}(a)$ for a given dataset. But as we are interested in the bias of $\tilde{P}(a)$, we have first to average the moment generating function over all possible datasets before approximating. The simplest way of evaluating this expectation is to observe that

$M^m(\tau | x_1, \ldots, x_n)$ is the coefficient of $\lambda^m/m!$ in

$$\prod_{j=1}^{n} \left(1 + \frac{\lambda}{n} e^{\tau x_j} + \frac{\lambda^2}{n^2 2!} e^{2\tau x_j} + \ldots + \frac{\lambda^m}{n^m m!} e^{m\tau x_j}\right),$$

so that the unconditional moment generating function $E(e^{m\tau \tilde{Z}})$ is the coefficient of $\lambda^m/m!$ in

$$E_{X_1 \ldots X_n}\left\{e^{-mn^{-1}\tau(X_1 + \ldots + X_n)} \prod_{j=1}^{n} \left(\sum_{r=0}^{m} \frac{\lambda^r}{n^r r!} e^{r\tau X_j}\right)\right\} = E_{X_1 \ldots X_n} \prod_{j=1}^{n} \sum_{r=0}^{m} \frac{\lambda^r}{n^r r!} e^{(r-m/n)\tau X_j}$$

$$= R^n(\lambda, \tau) \qquad (3 \cdot 2)$$

where

$$R(\lambda, \tau) = \sum_{r=0}^{m} \frac{\lambda^r}{n^r r!} M\left\{\left(r - \frac{m}{n}\right)\tau\right\}.$$

Then

$$E(e^{m\tau \tilde{Z}}) = \frac{m!}{2\pi i} \int_C R^n(\lambda, \tau) \frac{d\lambda}{\lambda^{m+1}}, \qquad (3 \cdot 3)$$

C being a contour in the λ plane enclosing the origin.

The true tail probability is

$$P(a) = \mathrm{pr}\,(\bar{X}_m - \mu > a) = \frac{1}{2\pi i} \int_{c-i\infty}^{c+i\infty} M^m(\tau)\, e^{-m(a+\mu)\tau} \frac{d\tau}{\tau},$$

where $c > 0$. Correspondingly, $E\{\tilde{P}(a)\}$ is given by the formula

$$E\{\tilde{P}(a)\} = \frac{1}{2\pi i} \int_{c-i\infty}^{c+i\infty} E(e^{m\tau \tilde{Z}})\, e^{-ma\tau} \frac{d\tau}{\tau}. \qquad (3 \cdot 4)$$

Explicit determination, by differentiation, of the coefficient of $\lambda^m/m!$ in (3·2) shows that in the case $m = 5$, $n = 10$, for instance,

$$E(e^{m\tau \tilde{Z}}) = \frac{1}{10\,000} \exp(45\tau^2/4) + \frac{9}{2000} \exp(29\tau^2/4) + \frac{9}{1000} \exp(21\tau^2/4)$$
$$+ \frac{9}{125} \exp(17\tau^2/4) + \frac{27}{250} \exp(13\tau^2/4)$$
$$+ \frac{63}{125} \exp(9\tau^2/4) + \frac{189}{625} \exp(5\tau^2/4). \qquad (3 \cdot 5)$$

From (3·5) it is easily seen that the inversion (3·4) yields

$$E\{\tilde{P}(a)\} = \frac{1}{10\,000}\{1 - \Phi(5a/\surd 22 \cdot 5)\} + \frac{9}{2000}\{1 - \Phi(5a/\surd 14 \cdot 5)\} + \frac{9}{1000}\{1 - \Phi(5a/\surd 10 \cdot 5)\}$$
$$+ \frac{9}{125}\{1 - \Phi(5a/\surd 8 \cdot 5)\} + \frac{27}{250}\{1 - \Phi(5a/\surd 6 \cdot 5)\}$$
$$+ \frac{63}{125}\{1 - \Phi(5a/\surd 4 \cdot 5)\} + \frac{189}{625}\{1 - \Phi(5a/\surd 2 \cdot 5)\}. \qquad (3 \cdot 6)$$

The manipulations required to derive (3·5) and evaluate (3·6) are most easily performed using a computer-algebraic manipulation package. The REDUCE package enabled evaluation of the corresponding expressions for the larger sample sizes considered, where there are many more terms involved, and hence calculation of the exact values of $E\{\tilde{P}(a)\}$ shown in Table 1.

To obtain an easily computed numerical approximation to $E\{\tilde{P}(a)\}$, observe that by (3·3) and (3·4)

$$E\{\tilde{P}(a)\}=\frac{m!}{(2\pi i)^2}\int\int e^{n\Omega(\theta,\tau)}\,d\theta\frac{d\tau}{\tau}, \qquad (3\cdot 7)$$

where

$$\lambda=e^{\theta}, \quad Q(\theta,\tau)=R(e^{\theta},\tau), \quad \Omega(\theta,\tau)=\log Q(\theta,\tau)-m\theta/n-ma\tau/n,$$

and the contours in the θ and τ planes are vertical lines passing to the right of the origin.

The method used by Skovgaard (1987) to extend the approximation due to Lugannani & Rice (1980) is directly applicable to (3·7), and yields an approximation to $E\{\tilde{P}(a)\}$ under appropriate conditions, satisfied in the case when F is $N(0, 1)$ and $M(\tau)=\exp(\frac{1}{2}\tau^2)$. Write $\dot{\Omega}$, Ω' for $\partial\Omega/\partial\theta$ and $\partial\Omega/\partial\tau$, and so on. The saddlepoint $(\hat{\theta}, \hat{\tau})$ of the exponent in (3·7) satisfies $\dot{\Omega}(\hat{\theta}, \hat{\tau})=0, \Omega'(\hat{\theta}, \hat{\tau})=0$, and is easily found using packaged iterative root finding routines, such as those in the NAG subroutine library. Suppose also that $\dot{\Omega}(\hat{\theta}_0, 0)=0$, and define $\hat{v}=[2\{\Omega(\hat{\theta}_0, 0)-\Omega(\hat{\theta}, \hat{\tau})\}]^{\frac{1}{2}}\operatorname{sgn}\hat{\tau}$.

Then the approximation is

$$E\{\tilde{P}(a)\}\sim\frac{m^{m+\frac{1}{2}}e^{-m}\,e^{n\Omega(\hat{\theta}_0,0)}}{n^{\frac{1}{2}}\{\ddot{\Omega}(\hat{\theta}_0,0)\}^{\frac{1}{2}}}\left[1-\Phi(\hat{v}n^{\frac{1}{2}})-\phi(\hat{v}n^{\frac{1}{2}})\left\{\frac{1}{\hat{v}n^{\frac{1}{2}}}-\frac{\ddot{\Omega}(\hat{\theta}_0,0)^{\frac{1}{2}}}{\hat{\tau}n^{\frac{1}{2}}\hat{\Delta}^{\frac{1}{2}}}\right\}\right], \qquad (3\cdot 8)$$

where $\hat{\Delta}=\ddot{\Omega}(\hat{\theta}, \hat{\tau})\Omega''(\hat{\theta}, \hat{\tau})-\{\dot{\Omega}'(\hat{\theta}, \hat{\tau})\}^2$, and Φ, ϕ denote the standard normal distribution function and density function respectively. The relative error of the approximation is the larger of $O(n^{-1})$ and $O(m^{-3/2})$.

The computed saddlepoint approximation (3·8) to $E\{\tilde{P}(a)\}$ is shown in Table 1. Overall, the approximation is seen to give adequate agreement with the simulation estimates, even for small probability levels, though it underestimates $E\{\tilde{P}(a)\}$ considerably in the case $m=n=20$, for tail probabilities less than about 0·01. A probable cause is the anomalous behaviour of $E(e^{m\tau\bar{Z}})$ when τ is large. The curve of $K(\tau)=\log E(e^{m\tau\bar{Z}})$ follows $10\tau^2$, the population cumulant generating function of $\bar{X}_m-\mu$, reasonably up to $\tau=0\cdot 55$, then rises dramatically, as would happen if $E(e^{m\tau\bar{Z}})$ were dominated by a mixture of two distinct exponential functions in this region. Inaccuracy of (3·8) is, presumably, related to this dramatic transition in the functional form of $K(\tau)$. The deviation also, of course, explains the poor approximation of the bootstrap in the tail: noting that the saddlepoint for approximation to $P(a)$ is at $\tau=a$, we see that $\tau=0\cdot 55$ can be related roughly to $a=0\cdot 55$, where the bias of the bootstrap becomes most noticeable.

For the case of estimating

$$P_R(a)=\operatorname{pr}(\bar{X}_m/\mu>a\,|\,F)$$

by $\tilde{P}_R(a)=\operatorname{pr}(\bar{Y}_m-a\bar{x}_n>0\,|\,F_n)$, it is found, on writing $\bar{Z}=\bar{Y}_m-a\bar{x}_n$, that

$$E\{\tilde{P}_R(a)\}=\frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty}E(e^{m\tau\bar{Z}})\frac{d\tau}{\tau},$$

for $c > 0$. The unconditional moment generating function $E(e^{m\tau\bar{Z}})$ is now the coefficient of $\lambda^m/m!$ in $R^n(\lambda, \tau, a)$, where

$$R(\lambda, \tau, a) = \sum_{r=0}^{m} \frac{\lambda^r}{n^r r!} M\left\{\left(r - a\frac{m}{n}\right)\tau\right\}.$$

In the case of an underlying standard exponential distribution, when $M(\tau) = (1-\tau)^{-1}$ and for $m = 5$, $n = 10$, $a = 1{\cdot}5$, for instance, differentiation of $R^n(\lambda, \tau, a)$ shows that

$$\begin{aligned} E(e^{m\tau\bar{Z}}) = \{&-1048576(680053005\tau^9 + 2537736860\tau^8 - 87649472\tau^7 - 7720774912\tau^6 \\ &- 1978825216\tau^5 + 8815052800\tau^4 + 223232000\tau^3 \\ &- 3973120000\tau^2 + 1515520000\tau - 163840000)\} \\ &\times \{625(17\tau-4)(13\tau-4)(9\tau-4)(5\tau-4)^2(3\tau+4)^9(\tau-4)^5\}^{-1}. \end{aligned} \qquad (3{\cdot}9)$$

Given $E(e^{m\tau\bar{Z}})$, the numerical value of $E\{\tilde{P}_R(a)\}$ may be computed, using the residue theorem, as the sum of the residues of $g(\tau) = E(e^{m\tau\bar{Z}})/\tau$ at positive poles, the contour of integration being completed by a large semicircle in the right halfplane in the usual way. Recall that if τ_0 is a pole of g of order k, then the residue of g at τ_0 is $h^{(k-1)}(\tau_0)/(k-1)!$, where $h(\tau) = (\tau - \tau_0)^k g(\tau)$.

In the case (3·9), for example, computation of $E\{\tilde{P}_R(a)\}$ involves calculation of the residues of g at the poles $\tau_0 = \frac{4}{17}, \frac{4}{13}, \frac{4}{9}, \frac{4}{5}$ and 4, of orders 1, 1, 1, 2 and 5 respectively. These calculations require at most fourth order differentiation, and are simply performed using REDUCE. Exact values for $E\{\tilde{P}_L(a)\}$ and $E\{\tilde{P}_R(a)\}$ computed in this way are shown in Table 2.

For the saddlepoint approximation, corresponding to (3·7) we have

$$E\{\tilde{P}_R(a)\} = \frac{m!}{(2\pi i)^2} \iint e^{n\Omega(\theta, \tau, a)}\, d\theta \frac{d\tau}{\tau},$$

where $\Omega(\theta, \tau, a) = \log Q(\theta, \tau, a) - m\theta/n$ and

$$Q(\theta, \tau, a) = \sum_{r=0}^{m} \frac{e^{r\theta}}{n^r r!} M\left\{\left(r - a\frac{m}{n}\right)\tau\right\}.$$

This differs from (3·7) crucially in that there are $m+1$ poles at $\tau_r = (r - am/n)^{-1}$ which appear to interfere with the region containing the saddlepoint, and cause the approximation analogous to (3·8) to underestimate the correct values as estimated by the simulations. We hope to pursue this matter in a subsequent paper.

4. DISCUSSION

We have studied two simple bootstrap estimation problems for small sample sizes and noted that in each case the bootstrap is noticeably and systematically biased. Hartigan (1986) has shown previously that the error in approximating the sampling distribution of $\bar{X}_n - \mu$ by that of $\bar{Y}_n - \bar{x}_n$ is $O(n^{-\frac{1}{2}})$. Table 1 shows that this bias is appreciable even for $n = 20$ in the case of a normal population. Simulation shows also that the finite support of F_n leads to marked bias in bootstrap estimation of the distribution of $\bar{X}_m/\mu$ in the case of an exponential underlying distribution.

The problems studied yield to an exact theoretical analysis, which is entirely straightforward in the normal case, if computer algebra is employed. In the exponential case such analysis is feasible, if computationally expensive. In principle, of course, such computer-algebraic techniques may be extended beyond study of the bias of the bootstrap estimator.

An approximate theoretical analysis is made feasible by the representation of the moment generating function $E(e^{m\tau\bar{Z}})$ as a complex integral: see (3·3). In the normal case the saddlepoint method developed works well, except in the extreme tail for $m=20$, $n=20$, where the approximation fails for reasons associated with extreme behaviour of the generating function. In the exponential case, however, the method yields difficulties which represent a challenge to the saddlepoint approximation technique.

An interesting alternative saddlepoint approach has recently been proposed by Suojin Wang of the University of Texas, for the case of the difference pivot $\bar{X}_m - \mu$. Starting with the saddlepoint approximation to the density of the mean of a bootstrap sample from a single data sample having mean $\bar{x}_n$, he expands the formula in powers of $\bar{x}_n - \mu$, before averaging over $\bar{x}_n$. He considers only the case $m=n$ and obtains results which, when computed for the standard normal case, are similar to those obtained via (3·8). We find that the problems of inaccuracy in the extreme tail, which we encountered for the case $m=20$, $n=20$, occur for his method also.

References

Davison, A. C. & Hinkley, D. V. (1988). Saddlepoint approximations in resampling methods. *Biometrika* **75**, 417-31.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Ann. Statist.* **7**, 1-26.

Hartigan, J. A. (1986). Discussion of paper by B. Efron and R. Tibshirani. *Statist. Sci.* **1**, 75-7.

Lugannani, R. & Rice, S. O. (1980). Saddlepoint approximation for the distribution of the sum of independent random variables. *Adv. Appl. Prob.* **12**, 475-90.

Skovgaard, I. M. (1987). Saddlepoint expansions for conditional distributions. *J. Appl. Prob.* **24**, 875-87.

[*Received February* 1989. *Revised July* 1989]