The Estimation of Number of Bootstrap Replications for the Nonparametric Bootstrap

Bohdan Linda
Faculty of Economics and Administration, University of Pardubice

Abstract:
Non-parametric bootstrap can be implicated at small size of random sample from unknown distribution when statistical methods are used. The paper deals with the problem, how many non-parametric bootstrap replications have to be done at estimation of bias and which extent of errors might be caused.

Many of statistical methods are based on asymptotical theory, e.g. on assumption that we have great number of data for statistical examination. However in praxis we very often meet the problem of small number of measured data. In this case the asymptotic theories are either inapplicable or even less reliable. Application of bootstrap methods is one of possibilities how to relieve this shortage.

The principle of these methods is simulation of new samples on base of measured data – in other words data multiplication. The detailed description of these methods can be found for example in [3].

It is very important question at application of these methods how many such newly simulated datasets have to be created so as the conclusions of statistical investigation have required accuracy.

The above-mentioned problem is presented in this paper at population mean estimate. This mean is estimated by sample average. It is also our target is to determine the accuracy of bootstrap estimate of bias of average, e.g. of the statistics $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Let's $x_1, x_{2, \ldots,} x_n$ is some concrete realization of random sample from $N(\mu, \sigma)$ distribution and let's $\hat{F}$ is the empirical distribution function of this sample. The random sample from distribution $\hat{F}$ is marked $X_1^*, X_2^*, \ldots, X_n^*$. This way of resampling is called nonparametric bootstrap. The average of this sample is marked $\overline{X}^* = \frac{1}{n}\sum_{i=1}^{n} X_i^*$. For its mean and variance holds true: $E^*(\overline{X}^*) = E^*\left(\frac{1}{n}\sum_{i=1}^{n} X_i^*\right) =$

$$= \frac{1}{n}\sum_{i=1}^{n} E^* X_i^* = \frac{1}{n} n \cdot E^* X^* = \sum_{i=1}^{n} x_i \, p(x_i) = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x} \,, \qquad (i)$$

because all values in random sample have the same probability $\frac{1}{n}$.

Analogous to variance

$$D^*(\overline{X}^*) = D^*\left(\frac{1}{n}\sum_{i=1}^{n} X_i^*\right) =$$

$$= \frac{1}{n^2}\sum_{i=1}^{n} D^* X_i^* = \frac{1}{n^2} n \cdot D^* X^* = \frac{1}{n} E^*(X^* - E^* X^*)^2 =$$

$$= \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{x})^2 p(X_i) = \frac{1}{n^2}\sum_{i=1}^{n}(X_i - \overline{x})^2 = \frac{s^2}{n} \qquad (ii)$$

When we realize R nonparametric bootstrap replications of random sample from distribution $\hat{F}$, we indicate relevant averages $\overline{X}_i^*, \quad i = 1,2, \ldots ,n$.

The bootstrap estimate of bias of average $\overline{X}$ is the statistics B_R : $B_R = \frac{1}{R}\sum_{r=1}^{R}\overline{X}_r^* - \overline{x}$ [4].

In case in nonparametric bootstrap we are able to calculate mean and variance. While using terms (i) and (ii) we obtain for these variables terms:

$$E^*(B_R) = E^*\left(\frac{1}{R}\sum_{r=1}^{R}\overline{X}_r^* - \overline{x}\right) = E^*\left(\frac{1}{R}\sum_{r=1}^{R}\overline{X}_r^*\right) - E^*(\overline{x}) = \frac{1}{R}\sum_{r=1}^{R}E^*\overline{X}_r^* - \overline{x} = \frac{1}{R}R \cdot E^*\overline{X}^* - \overline{x} = \overline{x} - \overline{x} = 0$$

(iii)

$$D^*(B_R) = D^*\left(\frac{1}{R}\sum_{r=1}^{R}\overline{X}_r^* - \overline{x}\right) = D^*\left(\frac{1}{R}\sum_{r=1}^{R}\overline{X}_r^*\right) = \frac{1}{R^2}\sum_{r=1}^{R}D^*\overline{X}_r^* =$$

$$= \frac{1}{R^2}R \cdot D^*\overline{X}_r^* = \frac{1}{R^2} \cdot \frac{Rs^2}{n} = \frac{s^2}{nR} \qquad (iv)$$

The term (iv) can be used for standard error estimation at given size of random sample and number of bootstrap replications R.

To be able to compare the results of simulations with real values we will assume, that X_1, X_2, ..., X_n is random sample from $N(\mu,\sigma)$ distribution. At fulfilled presumption of normal distribution of variables X_i , i = 1, …, n, it is possible to calculate in the exact way the theoretical value of bias and variance of average . These values are 0 and $\frac{\sigma^2}{n}$ by turns.

We used two samples of 15 data, generated from normal distribution to verify the features of bootstrap bias estimate. The first sample was generated from N(100,10) distribution and the second one from N(0,1) distribution. The sample average $\overline{x}$ and sample variance s^2 are presented in table 1. Bootstrap estimates of these sample statistics were calculated on base of 10 000 replications of random sample. They are stated in table 1.

Tab.1

parameter	N(100,10) distribution	N(0,1) distribution
$\overline{x}$	101,267	0,1007
s^2	117,662	1,0614
$\overline{x}_R^*$	101,233	0,0977
s_R^{2*}	110,156	0,9908

Figures 1 and 2 show relation between bootstrap bias estimate and number of R simulated samples. Five repetitions each at 2000 replications were generated from N(100,10) distribution and the same number from N(0,1) distribution. Figures 1 and 2 suggest, if R >600 replications then the values of bias estimate differ in a minimal way.

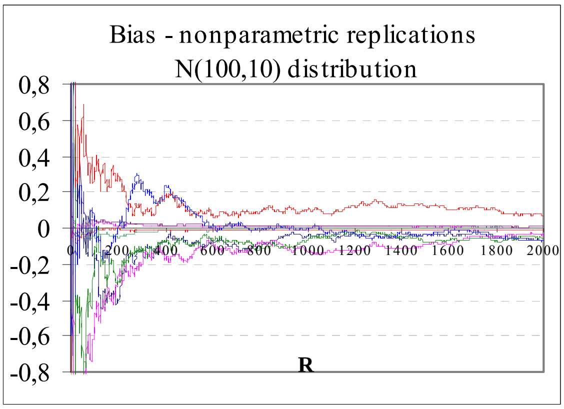

Fig.1

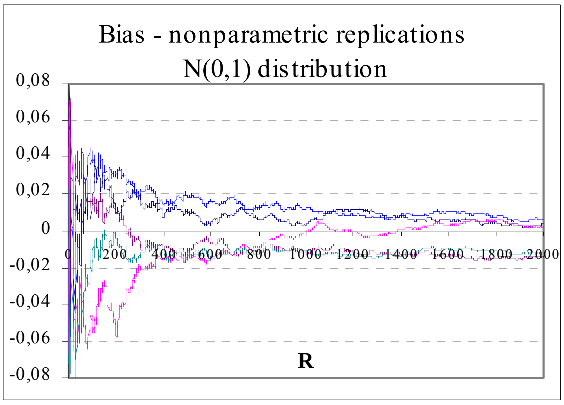

Fig.2

Bias estimate for each from Q repetitions at R bootstrap replications was calculated at first. Sample average $\overline{B}_{RQ} = \frac{1}{Q}\sum_{q=1}^{Q}\left|\frac{1}{R}\sum_{r=1}^{R}\overline{X}_r^* - \overline{x}\right|$ from absolute values of these bias estimates was calculated next. Variable $\overline{B}_{RQ}$ can be regarded as criterion of bias estimate error that is less sensitive to random deviations within one repetition. Convergence of this error is relatively fast. The error is approximately 0,07 for 600 replications from N(100,10) distribution and it didn't changed any more for higher number of bootstrap replications. Analogically stabilization of error started approximately for 900 replications at value 0,009 at samples generated from N(0,1) distribution. Convergence of this error is shown in the figures 3 and 4.

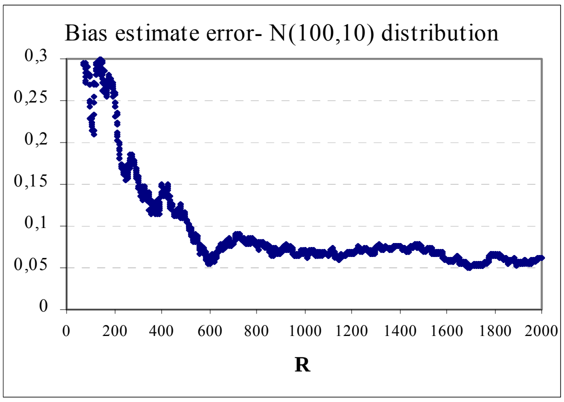

Fig.3

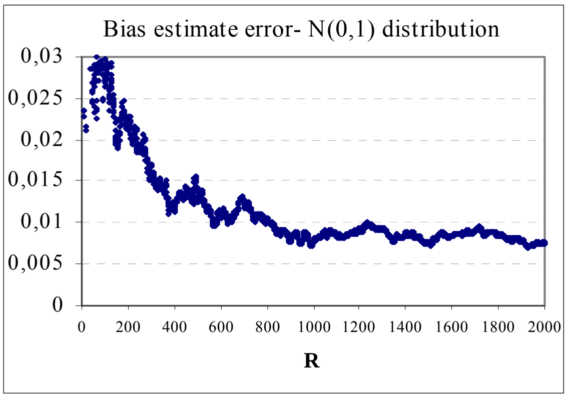

Fig.4

The value of bootstrap estimate of bias can be calculated in the theoretical way according to relation $D^*(B_R) = \frac{s^2}{nR}$. Comparison of these theoretical values with results of simulation is illustrated in the figures 5 and 6, where smooth curve express the theoretical value of $D^*(B_R)$ and the scatter diagram simulated values. These values are almost identical after 3000 simulations.

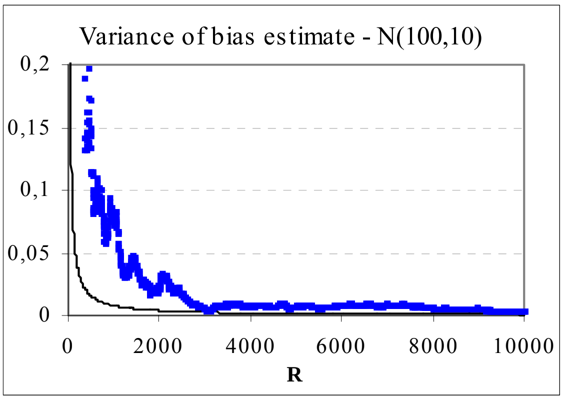

Fig.5

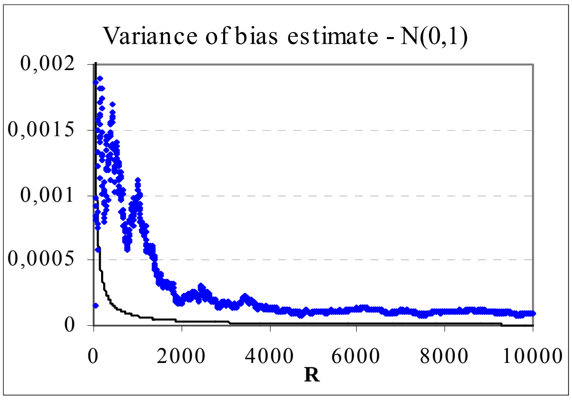

Fig.6

Absolute error of bias variance estimate is absolute value of difference between bias variance obtained from replications of original realization of random sample and theoretical value that was obtained by calculation and equals $\frac{s^2}{nR}$.

Figures 7 and 8 show the development of this error in dependence on number of replications. It is visible that this error doesn't change its value when more than 1000 bootstrap replications are made.

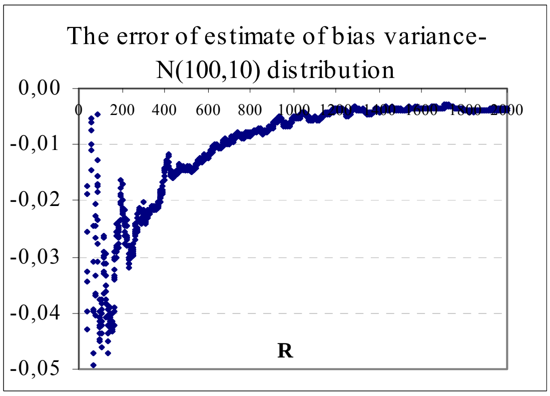

Fig.7

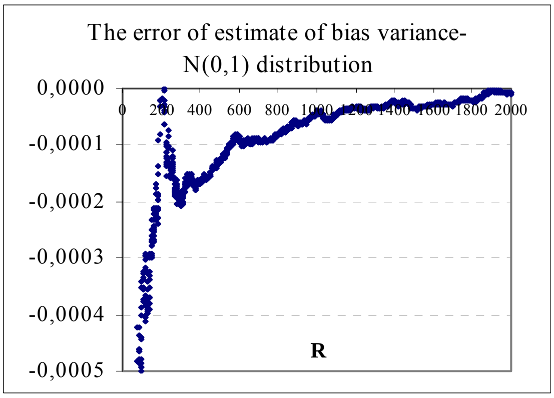

Fig.8

Finally it is possible to say that 600 bootstrap nonparametric replications were sufficient for bias estimate. The properties of bias estimate are not improved when more replications are made.

Literature:

[1.] Davison A.C., Hinkley D.V. Bootstrap Methods and their Application. Cambridge University Press, 1997.

[2.] Efron B. More efficient bootstrap computations. Journal of the American Statistical Association, 1990.

[3.] Efron B., Tibshirani R.J. An Introduction to the Bootstrap. Chapman & Hall, 1993

[4.] Hall P. On the bootstrap and confidence intervals. Annals of Statistics 14, s.1431-1452, 1986

Contacts address:
doc. RNDr. Bohdan Linda, CSc.
University of Pardubice
Faculty of Economics and Administration
Department of Mathematics
53210 Pardubice
e-mail: bohdan.linda@upce.cz
00420 466036020

Review:
prof. RNDr. Otakar Prachař, CSc.
University of Pardubice
Faculty of Economics and Administration
Department of Mathematics