The Basic Idea of Bootstrap Methods

Jana Kubanová

Abstract

Two methods of bootstrap simulation – parametric and nonparametric – are described in this article. Parametric simulation assumes that distribution of random variable \( X \) is known. Nonparametric simulation doesn’t require this assumption. Concrete examples demonstrate both ways of simulation.

Key words: Parametric and nonparametric bootstrap simulations, resampling, bias, variance estimates.

1. Introduction

The case when the frequency of random sample is very small is very often in research in economics, sciences and technical sciences. Then it is very difficult and untrustworthy to express any conclusions about standard errors, hypotheses testing or confidence intervals. We can use the bootstrap method based on resampled data in this case.

The substantial and frequent problem of statistical data analysis is to determine theoretical properties of some statistics \( \hat{\Theta} \). In simple cases it is possible to use classical statistical methods. But when it is difficult to determine theoretical properties of \( \hat{\Theta} \) estimate by exact way then it is possible to use the bootstrap method.

The bootstrap method was elaborated by Bradley Efron in 1977. This method makes use advantage of the high-speed power and number-crunching power of computers. The principle is to resample new samples from the original data set with the same rate. This approach involves repeating the original data analysis procedure with many replicated datasets. Very important advantage of this method is that it allows to construct artificial data sets without making any assumptions about bell shaped curves. Problems that can be solved with the help of bootstrap method can be divided into two groups that are later called parametric and nonparametric bootstrap (Efron, Tibshirani, 1993).

2. Theoretical assumptions

We assume that \( X = (X_1, X_2, \ldots, X_n) \) is random sample from distribution \( F(x, \psi) \), belonging to some family of distribution functions \( \mathcal{F} \) differentiated with parameter \( \psi \).

Frequent object in view is to determine the properties if statistics \( \hat{\Theta} = \hat{\Theta}(X_1, \ldots, X_n) \). When the data are obtained from distribution \( F \), we indicate the distribution function of statistics \( \hat{\Theta} \) \( G_n(x, F) = P(\hat{\Theta} < x) \). Indication \( G_n(x, F) \) expresses that distribution function \( G \) of statistics \( \hat{\Theta} \) is determined from \( n \) randomly selected values of random variable \( X \) with distribution function \( F \). Generally function \( G_n(x, F) \) depends on parameters of distribution \( F \). In case of pivot statistics \( \hat{\Theta} \) the distribution function isn’t dependent on these parameters.

Asymptotic theory is one of common classical methods of distribution function \( G_n(x, F) \) estimate. The asymptotic approximation enables to substitute the unknown distribution function \( G_n \) with known function \( G_{\psi} \). The estimates in econometric applications aren’t in principle pivotal. Distribution of them usually depends on one or more unknown parameters. Bootstrap method provides one alternative approximation of distribution of statistics \( \hat{\Theta}(X_1, \ldots, X_n) \). Unknown distribution function \( F \) is replaced by known estimate that is indicated \( \hat{F} \).

1 CSc., University of Pardubice, Faculty of Economics and Administration, Department of Mathematics, Czech Republic.
When bootstrap methods are applied, two estimates of distribution function are differentiated and similarly two bootstrap simulation techniques are used. Following paragraph describes how the distribution function was obtained, we use special signification for it.

2.1. The parametric estimate of distribution function $F$

Let’s suppose that for distribution function $F$ of random variable $X$ holds true $F(x) = F(x, \psi)$ for some unknown parameter $\psi$, that is consistently estimated by statistics $\hat{\psi}$. If the function $F(x, \psi)$ is the continuous function of parameter $\psi$, then $F(x, \hat{\psi}) \to F(x, \psi)$ for $n \to \infty$. The distribution function $F = F(x, \hat{\psi})$ obtained in this way is signified $F_{\hat{\psi}}$ and called the parametric estimate of distribution function $F$.

2.2. The nonparametric estimate of distribution function $F$

When the distribution function $F$ is unknown we estimate this distribution with empirical distribution function $\hat{F}$ that is indicated $F_n$. It holds:

$$F_n = \frac{1}{n} \sum_{i=1}^{n} I(x - x_i),$$

where function $I$ is an indicator, it is $I(x - x_i) = \begin{cases} 0 & x - x_i < 0 \\ 1 & x - x_i \geq 0 \end{cases}$.

Function $F_n$ is called nonparametric estimate of distribution function $F$.

No matter how the estimate $\hat{F}$ of distribution function $F$ was obtained, the bootstrap estimate of distribution function $G_n(x, \hat{F})$ is the function $G_n(x, \hat{F})$. But we often fail to find the function $G_n(x, \hat{F})$ by analytical way. Bootstrap method enables to perform the approximation of distribution function $F$ with function $\hat{F}$ and consequently to estimate the distribution $G_n(x, F)$ of the statistics $\hat{\Theta}$ with function $G_n(x, \hat{F})$. As the functions $\hat{F}$ and $F$ aren’t identical then the functions $G_n(x, \hat{F})$ and $G_n(x, F)$ are different except for the case when $\hat{\Theta}$ is the pivotal statistics. That is why the bootstrap estimate $G_n(x, \hat{F})$ is only approximation of distribution function $G_n(x, F)$ of the statistics $\hat{\Theta}$.

3. Parametric bootstrap simulations

3.1. Principle of parametric simulations

Let’s assume that $X$ is a random variable with known distribution function $F$, $X = (X_1, X_2, \ldots, X_n)$ random sample from this distribution with distribution function $F$ and $\psi$ is some parameter of distribution of random variable $X$. $\hat{\psi}$ is the estimate of parameter $\psi$ that was obtained from random sample $X_1, X_2, \ldots, X_n$. Parametric estimate of distribution function $F$ is its parametric estimate $\hat{F} = F_{\hat{\psi}}$. New random samples are obtained when values of random variable with $F_{\hat{\psi}}$ distribution are generated. This way of new samples generation is called parametric simulation. The model based on function $F_{\hat{\psi}}$ is called parametric model.

Technique of parametric simulations:

- Simulate random sample $X' = (X_1', X_2', \ldots, X_n')$ of rate $n$ from distribution $F_{\hat{\psi}}$.
- Calculate statistics $\hat{\Theta} = \hat{\Theta}(X_1', X_2', \ldots, X_n')$.
- Repeat first and second items and use results for next calculations.

Indication $'$ is used to express the reality that relevant variable relates to the model $F_{\hat{\psi}}$.

For example $X'$ means that distribution of this random variable is equivalent with model $F_{\hat{\psi}}$. 
3.2. Moments estimates

When the statistic of interest \( \hat{\Theta} \) is calculated from a simulated dataset, we denote it \( \hat{\Theta}^* \). From \( R \) repetitions of the data simulation we obtain \( \hat{\Theta}_1^*, \ldots, \hat{\Theta}_R^* \) statistics. Properties of \( \hat{\Theta} \) are then estimated from \( \hat{\Theta}_1^*, \ldots, \hat{\Theta}_R^* \).

The moment estimates will be realized in following way:

The estimator of bias of \( \hat{\Theta} \)

\[
b(\hat{\Theta}) = E(\hat{\Theta} | F) - \Theta,
\]

is statistics

\[
B = b(F_{\hat{\theta}}) = E(\hat{\Theta} | F_{\hat{\theta}}) - \hat{\Theta}^* - \hat{\Theta} - t
\]

(Peracchi, 2000)

Distribution function \( F \) is estimated with function \( \hat{F}_{\hat{\theta}} \), and then estimate of bias is

\[
B_{\hat{\theta}} = \frac{1}{R} \sum_{r=1}^{R} \hat{\Theta}_r^* - t = \hat{\Theta}^* - t
\]

(Peracchi, 2000)

where \( t \) is the concrete value of statistics \( \hat{\Theta} \), calculated from original dataset. \( \hat{\Theta} - t \) is the simulation analogy of \( \hat{\Theta} - \Theta \).

Variance \( \sigma^2 = D(\hat{\Theta} | F) \) of random variable \( \hat{\Theta} \) is estimated with statistics

\[
S_{\hat{\theta}}^2 = \frac{1}{R-1} \sum_{r=1}^{R} (\hat{\Theta}_r^* - \hat{\Theta}^*)^2
\]

(Peracchi, 2000)

Similar estimators can be derived for other moments. These empirical approximations are justified by the law of large numbers.

3.3. Example – simulation from normal distribution

Let’s assume that \( x_1, x_2, \ldots, x_n \) is some realization of random sample from \( N(\mu, \sigma) \) distribution. When parameter \( \mu \) is estimated with sample average \( \bar{x} \) and parameter \( \sigma \) – with sample standard deviation \( s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2} \), then the parametric bootstrap replication of sample \( x_1, x_2, \ldots, x_n \) is random sample \( x_1^*, x_2^*, \ldots, x_n^* \) that comes from \( N(\bar{x}, s) \) distribution. We obtain \( R \) realizations of bootstrap samples after \( R \) bootstrap replications.

Let’s indicate bootstrap average \( \bar{X}^* = \frac{1}{n} \sum_{i=1}^{n} X_i^* \), where \( X_1^*, X_2^*, \ldots, X_n^* \) is random sample from \( N(\bar{x}, s) \) distribution. Random variable \( \bar{X}^* \) is normally distributed with mean \( \bar{x} \) and standard deviation \( s/\sqrt{n} \). Bootstrap estimate of bias of sample average \( \bar{X}^* \) is the statistics

\[
B_{\bar{X}} = \frac{1}{R} \sum_{r=1}^{R} \bar{X}_r^* - \bar{x}
\]

(Davison, Hinkley, 1997).
Random variable $\frac{1}{R} \sum_{r=1}^{R} X^*_r$ is approximately normally distributed with mean $\bar{x}$ and standard deviation $\sqrt{\frac{s^2}{Rn}}$. Then the bootstrap bias $\frac{1}{R} \sum_{r=1}^{R} X^*_r - \bar{x}$ is approximately normally distributed with mean 0 and standard deviation $\sqrt{\frac{s^2}{Rn}}$ as well.

We are interested in bootstrap bias distribution and mean of bias $E^*(B_R)$ distribution, also $B_R - E^*(B_R)$ in the following step. This bootstrap bias is obtained from $R$ bootstrap replications. But $E^*(B_R) = 0$ in our case. Therefore $B_R - E^*(B_R) = \frac{1}{R} \sum_{r=1}^{R} X^*_r - \bar{x}$.

Random variable $B_R$ is approximately normally distributed with mean 0 and standard deviation $\sqrt{\frac{s^2}{Rn}}$. Then random variable $B_R - E^*(B_R)$ is also approximately normally distributed with mean 0 and standard deviation $\sqrt{\frac{s^2}{Rn}}$. This random variable exactly expresses the estimation of an error that is made after finite number of bootstrap replications. After normal transformation $\frac{B_R - E^*(B_R)}{s} \sqrt{Rn}$ we obtain normally distributed random variable with mean 0 and standard deviation 1. Interval estimate of variable $E^*(B_R)$ is then $D_R = \frac{s}{\sqrt{Rn}}$ $1 < \frac{s}{\sqrt{Rn}}$.

Where $z_\alpha$ is quantile of $N(0,1)$ distribution; $z_\alpha = \Phi^{-1}\left(\frac{\alpha}{2}\right)$. We can estimate by means of this relation extent of the error that we can cause at given number $R$ of bootstrap replications, at given extent of $n$ of original random sample and at chosen value of $\alpha$.

To be able to compare the results of simulations with real values we assumed, that $X_1, X_2, \ldots, X_n$ is random sample from $N(\mu, \sigma)$ distribution. At fulfilled presumption of normal distribution of variables $X_i, i = 1, \ldots, n$, 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.

Table 1 shows the method of parametric simulations at the example of data from $N(100,10)$ distribution. The concrete values of random sample realization are 109, 80, 97, 115, 113, 83, 89, 110, 98, 114, 95, 100, 105, 112, 99. These values are introduced in the first column of the table. Following nine columns show simulated values of random sample, rate of each of these samples is identical to rate of original random sample. These simulated samples come again from normal distribution with estimated parameter $\bar{x} = 101.267$ and $s = 10.847$. These simulated samples enable us to estimate parameters and results can be used for next calculations.

To compare results of parametric simulations, we used two normally distributed random samples: the first one with mean 100, standard deviation 10 and size 15 and the second one with mean 0, standard deviation 1 and size 15 to illustrate solved problem. 10 000 bootstrap replications were made for each of above samples and basic statistical characteristics were computed.

These characteristics are presented in Table 2. It is visible that results obtained after 10 000 bootstrap parametric replications are “closer” to the real parameters $\mu$ and $\sigma$ of original normal distribution than results obtained on the base of random sample from this normal distribution.
Table 1

<table>
<thead>
<tr>
<th>Original</th>
<th>1 simulation</th>
<th>2 simulation</th>
<th>3 simulation</th>
<th>4 simulation</th>
<th>5 simulation</th>
<th>6 simulation</th>
<th>7 simulation</th>
<th>8 simulation</th>
<th>9 simulation</th>
</tr>
</thead>
<tbody>
<tr>
<td>109</td>
<td>116,552</td>
<td>111,248</td>
<td>101,449</td>
<td>92,454</td>
<td>102,736</td>
<td>84,239</td>
<td>120,159</td>
<td>104,762</td>
<td>91,297</td>
</tr>
<tr>
<td>80</td>
<td>107,456</td>
<td>93,269</td>
<td>112,719</td>
<td>95,816</td>
<td>118,465</td>
<td>86,591</td>
<td>115,768</td>
<td>96,795</td>
<td>84,646</td>
</tr>
<tr>
<td>97</td>
<td>97,026</td>
<td>90,843</td>
<td>101,826</td>
<td>100,452</td>
<td>66,183</td>
<td>93,034</td>
<td>93,477</td>
<td>96,961</td>
<td>98,727</td>
</tr>
<tr>
<td>115</td>
<td>95,409</td>
<td>99,438</td>
<td>99,603</td>
<td>98,760</td>
<td>63,181</td>
<td>100,595</td>
<td>80,231</td>
<td>100,052</td>
<td>110,125</td>
</tr>
<tr>
<td>113</td>
<td>97,782</td>
<td>103,790</td>
<td>95,423</td>
<td>98,941</td>
<td>108,465</td>
<td>100,431</td>
<td>101,741</td>
<td>117,758</td>
<td>77,064</td>
</tr>
<tr>
<td>83</td>
<td>96,492</td>
<td>107,183</td>
<td>115,801</td>
<td>94,384</td>
<td>111,088</td>
<td>122,378</td>
<td>98,795</td>
<td>116,095</td>
<td>84,694</td>
</tr>
<tr>
<td>89</td>
<td>89,235</td>
<td>98,433</td>
<td>108,453</td>
<td>90,943</td>
<td>90,843</td>
<td>112,738</td>
<td>104,762</td>
<td>100,862</td>
<td>92,745</td>
</tr>
<tr>
<td>110</td>
<td>112,429</td>
<td>114,113</td>
<td>117,733</td>
<td>90,493</td>
<td>92,132</td>
<td>102,115</td>
<td>101,000</td>
<td>105,073</td>
<td>103,552</td>
</tr>
<tr>
<td>98</td>
<td>99,070</td>
<td>101,942</td>
<td>103,745</td>
<td>88,700</td>
<td>103,963</td>
<td>92,489</td>
<td>104,074</td>
<td>77,667</td>
<td>76,697</td>
</tr>
<tr>
<td>114</td>
<td>104,879</td>
<td>104,497</td>
<td>92,209</td>
<td>99,492</td>
<td>91,969</td>
<td>99,681</td>
<td>107,494</td>
<td>119,469</td>
<td>102,898</td>
</tr>
<tr>
<td>95</td>
<td>111,479</td>
<td>85,621</td>
<td>103,537</td>
<td>99,726</td>
<td>89,840</td>
<td>103,731</td>
<td>88,210</td>
<td>107,638</td>
<td>100,003</td>
</tr>
<tr>
<td>100</td>
<td>120,649</td>
<td>97,936</td>
<td>93,525</td>
<td>102,964</td>
<td>102,539</td>
<td>109,414</td>
<td>101,412</td>
<td>89,768</td>
<td>103,873</td>
</tr>
<tr>
<td>105</td>
<td>76,794</td>
<td>97,471</td>
<td>111,277</td>
<td>108,909</td>
<td>111,856</td>
<td>88,535</td>
<td>91,839</td>
<td>103,767</td>
<td>111,609</td>
</tr>
<tr>
<td>112</td>
<td>111,465</td>
<td>95,953</td>
<td>115,436</td>
<td>77,358</td>
<td>94,297</td>
<td>100,911</td>
<td>101,214</td>
<td>92,962</td>
<td>107,961</td>
</tr>
<tr>
<td>99</td>
<td>126,222</td>
<td>84,854</td>
<td>105,237</td>
<td>93,610</td>
<td>93,622</td>
<td>97,758</td>
<td>80,513</td>
<td>102,767</td>
<td>121,041</td>
</tr>
</tbody>
</table>

Table 2

<table>
<thead>
<tr>
<th>Parameter</th>
<th>N(100, 10) distribution</th>
<th>N(0, 1) distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\bar{x}$</td>
<td>101,267</td>
<td>0,1007</td>
</tr>
<tr>
<td>$s^2$</td>
<td>117,662</td>
<td>1,0614</td>
</tr>
<tr>
<td>$\bar{x}_R$</td>
<td>101,238</td>
<td>0,0982</td>
</tr>
<tr>
<td>$s_R^2$</td>
<td>109,947</td>
<td>0,9971</td>
</tr>
</tbody>
</table>

Figures 1 and 2 show the results of several simulations, changes of bias with increasing number of bootstrap replications at sampling from $N(101,267; 10,847)$ and $N(0,100; 1,030)$ distributions. The problem is demonstrated at 5 repetitions of 2000 replications. Empirical biases were calculated for each value of $R$. We can note how the variability decreases as the simulation size increases and how the simulated values converge to the exact value. To answer the question, how many bootstrap replications are needed, Figures 1 and 2 suggest that $R = 400$ bootstrap replications could be adequate. Values of bias don’t markedly change for larger values of $R$. 

Fig. 1. Sampling from $N(101,267; 10,847)$

Fig. 2. Sampling from $N(0,100; 1,030)$
4. Nonparametric bootstrap simulations

4.1. Principle of nonparametric simulations

In case of nonparametric simulation we assume, that \( X = (X_1, X_2, \ldots, X_n) \) is random sample from distribution with unknown distribution function \( F \). Empirical distribution function \( F_n \) is used for estimate of unknown distribution function \( F \). Application of this distribution function is analogical to parametric model.

Realization of further samples are obtained from digital data (originally measured) \( x_1, x_2, \ldots, x_n \) that way that we apply random sample with replacement of rate \( n \). That is why random variables \( X_i^* \) have really distribution \( F_n \). We use indication \( X_i^* \) for simulated variable \( X_i \), \( n \)-tuple \( X_1^*, \ldots, X_n^* \) is random sample from \( F_n \) distribution. New concrete samples rise from original data by original data resampling (random reordering and confusion). This technique is called nonparametric simulation (nonparametric bootstrap).

4.2. Moments estimates

Following consideration can be used in connection with theoretical calculation:

The estimate of parameter \( \theta \) is random variable \( \hat{\theta} \). We obtain the concrete value \( t \) of random variable \( \hat{\theta} \) from concrete realization of random sample. Both value \( t \) of statistics \( \hat{\theta} \) and empirical distribution function depend on values of random sample \( x_1, x_2, \ldots, x_n \). The value \( t \) can be considered as function of empirical distribution function \( F_n \). This relation can be expressed \( t = g(F_n) \), where \( g \) is the relevant function. Relation \( t = g(F_n) \) expresses the way how to determine the value of \( t \) on base of empirical distribution function \( F_n \). The elementary examples of such functions are terms for mean and variance calculation, that are generally defined in the following way:

\[
EX = g(F) = \int_{-\infty}^{\infty} x dF(x) \quad \text{and} \quad DX = g(F) = \int_{-\infty}^{\infty} (x - EX)^2 dF(x).
\]

If we substitute the distribution function \( F \) in relation \( EX = g(F) = \int_{-\infty}^{\infty} x dF(x) \) with empirical distribution function \( F_n \), we obtain estimate for mean

\[
g(F_n) = \int_{-\infty}^{\infty} x dF_n(x) = \int_{-\infty}^{\infty} x d\left( \frac{1}{n} \sum_{i=1}^{n} I(x - x_i) \right) = \frac{1}{n} \sum_{i=1}^{n} x_i = \bar{x}
\]

(because generally holds true \( \int g(y) dI(y-x)g(x) \) for each continuous function \( g \)).

The similar relation can be found for variance estimate.

If we substitute in term \( DX = g(F) = \int_{-\infty}^{\infty} (x - EX)^2 dF(x) \) distribution function \( F \) with empirical distribution function \( F_n \), we obtain:

\[
g(F_n) = \int_{-\infty}^{\infty} (x - EX)^2 dF_n(x) = \int_{-\infty}^{\infty} (x - EX)^2 d\left( \frac{1}{n} \sum_{i=1}^{n} I(x - x_i) \right) = \frac{1}{n} \sum_{i=1}^{n} \int_{-\infty}^{\infty} (x - EX)^2 dI(x - x_i)
\]

\[
= \frac{1}{n} \sum_{i=1}^{n} x_i^2 - (EX)^2 = \frac{1}{n} \sum_{i=1}^{n} x_i^2 - \bar{x}^2 = s^2
\]
4.3. Example

To verify theoretical results, we assume that $x_1, x_2, \ldots, x_n$ is some concrete realization of a random sample from $N(\mu, \sigma)$ distribution and $\hat{F}$ is the empirical distribution function of this sample. The random sample from distribution $\hat{F}$ is marked $X^*_1, X^*_2, \ldots, X^*_n$. The average of this sample is marked $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X^*_i$. For its mean holds true:

$$E^*(\bar{X}) = E\left(\frac{1}{n} \sum_{i=1}^{n} X^*_i\right) = \frac{1}{n} \sum_{i=1}^{n} E^* X^*_i = \frac{1}{n} n \cdot E^* X^* = \frac{1}{n} \sum_{i=1}^{n} x_i = \bar{x}, \quad (1)$$

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

Analogous to variance

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

$$= \frac{1}{n} \sum_{i=1}^{n} (X^*_i - \bar{x})^2 \cdot p(X^*_i) = \frac{1}{n^2} \sum_{i=1}^{n} (X^*_i - \bar{x})^2 = \frac{s^2}{n}. \quad (2)$$

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

The bootstrap estimate of bias of average $\bar{X}$ is the statistics $B_r = \frac{1}{R} \sum_{r=1}^{R} \bar{X}^*_r - \bar{x}$ (Davidson Hinkley, 1997).

In case of nonparametric bootstrap we are able to calculate mean and variance. While using terms (1) and (2) we obtain for these variables terms:

$$E^*(B_r) = E\left(\frac{1}{R} \sum_{r=1}^{R} \bar{X}^*_r - \bar{x}\right) = E\left(\frac{1}{R} \sum_{r=1}^{R} \bar{X}^*_r\right) - E^*(\bar{x}) = \frac{1}{R} \sum_{r=1}^{R} E^* \bar{X}^*_r - \bar{x} =$$

$$= \frac{1}{R} R \cdot E^* \bar{X}^*_r - \bar{x} = \bar{x} - \bar{x} = 0. \quad (3)$$

$$D(B_r) = D\left(\frac{1}{R} \sum_{r=1}^{R} \bar{X}^*_r - \bar{x}\right) = D\left(\frac{1}{R} \sum_{r=1}^{R} \bar{X}^*_r\right) = \frac{1}{R^2} \sum_{r=1}^{R} D^* \bar{X}^*_r = \frac{1}{R^2} R \cdot D^* \bar{X}^*_r = \frac{1}{R^2} \cdot \frac{R s^2}{n} = \frac{s^2}{nR}. \quad (4)$$

The term 4 can be used for standard error estimation at given size of random sample and number of bootstrap replications $R$.

The concrete example of nonparametric simulations is shown in Table 3. The original realization of random sample is in the first column of the table. The same probability $\frac{1}{n}$ of each value is assumed. It is 0.0667 in our concrete case. These values are stated in the second column of the table. Next nine columns show simulated samples with rate $n = 15$ as well. Seeing that sampling with replacement is realized, individual values can repeat in the random sample or they can’t be in this sample at all. On the base of samples obtained in a described way we can estimate parameters or express further statistical conclusions.
To compare results of nonparametric simulations and to verify the features of bootstrap bias estimate, we used two samples of 15 data, where the original samples were generated from normal distribution. The first original sample was generated from $N(100, 10)$ distribution and the second one from $N(0, 1)$ distribution. The sample average $\bar{x}$ and sample variance $s^2$ are presented in Table 4. We continued in calculation as though we have never known the distribution of original dataset. Bootstrap estimates of these sample statistics were calculated on base of 10 000 replications of random sample and they are stated in Table 4 as well. We can see similarly to parametric bootstrap that estimated statistics from 10 000 bootstrap replication are “closer” to values of original distribution than values estimated from original sample.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>$N(100, 10)$ distribution</th>
<th>$N(0, 1)$ distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\bar{x}$</td>
<td>101.267</td>
<td>0.1007</td>
</tr>
<tr>
<td>$s^2$</td>
<td>117.662</td>
<td>1.0614</td>
</tr>
<tr>
<td>$\bar{x}_R$</td>
<td>101.233</td>
<td>0.0977</td>
</tr>
<tr>
<td>$s_R^2$</td>
<td>110.156</td>
<td>0.9908</td>
</tr>
</tbody>
</table>

Figures 3 and 4 show relation between nonparametric bootstrap bias estimate and number of $R$ simulated samples. Five repetitions each at 2000 replications [original samples were from $N(100,10)$ distribution and from $N(0,1)$ distribution respectively] were generated. Figures 3 and 4 suggest, if $R > 600$ replications then the values of bias estimate differ in a minimal way.
5. Conclusion

The basic differences of described approaches are as follows:

At parametric bootstrap we know probability model and thereby exact distribution of some important statistics. This knowledge can be used at confidence intervals construction, hypothesis testing and in further statistical analyses.

At nonparametric bootstrap any assumptions about distribution model aren’t determined. It is possible to realize a lot of simulations that help to determine the properties of random variable \( \hat{\theta} \) and even to estimate its distribution.

The difference in parameters estimates is the following:

Mean estimate in case of normal distribution – the difference between parametric and nonparametric approach is 0.005 in both cases. It can be considered as insignificant.

Variance estimate in case of normal distribution – the difference between parametric and nonparametric approach is 0.2090 at \( N(100, 10) \) distribution and 0.0063 at \( N(0,1) \) distribution.

It is possible to state, that estimates of parameters obtained after 10 000 bootstrap replications were in all cases “closer” to values of original parameters than estimates obtained only from 15 values of original sample.

Number of bootstrap replications necessary for bias estimate:

It was found out in presented examples that it is necessary to make 400 parametric bootstrap replications and 600 nonparametric replications to obtain relevant results. The values of estimated bias didn’t improve, when more replications were made.

References