ALBERT-LUDWIGS-
UNIVERSITÄT FREIBURG

information
systems

Business Analytics

– SEMINAR SUMMER SEMESTER 2014 –

Resampling Methods: An Introduction to Cross Validation and Bootstrapping

– SEMINAR PAPER –

Submitted by:
Nicolas Pröllochs

Student-ID:

Advisor:
Prof. Dr. Dirk Neumann

Contents

Abstract

Resampling methods have become a major issue in modern statistics. These methods, in combination with modern computers, have revolutionized problem solving in science to a vast extend. Particularly in presence of sparse samples, these methods allow to infer to an unknown population by drawing simulated samples from the samples at hand, without making eventually unrealistic assumptions about the unknown population. In addition, resampling methods allow great accuracy in a wide field of applications even requiring few assumptions and providing large generalizability. Thus, this paper aims to give an introduction to the basic functionality of resampling methods. In this context, we introduce two wide-spread resampling approaches, namely, cross-validation and bootstrap. We provide additional guidance about the basic concepts of these methods in several variations and illustrate potential applications.

1 Introduction

Resampling methods are a indispensable tool in modern statistics [6]. Because of the fast growing computational power, it is possible to look at data in a previously inconceivable way. These methods are a set of statistical inference methods, based on drawing new samples from an initial sample.

Often, researchers have only at hand one sample and want to infer from the given sample to the distribution of an unknown population, while it is not possible to get additional samples from the population. The fundamental idea of resampling is to draw new samples from the existing sample to get a large amount of simulated samples. Assuming that all the information contained in the original sample is also contained in the distribution of the simulated samples, the motivation is now to analyze patterns in the simulated samples and to infer from the samples to the unknown distribution of the population. Thus, resampling is equivalent to generate completely new random samples from the population and allows us to make conclusions about the population strictly from the sample at hand, instead of making perhaps unrealistic assumptions about the population [7]. An advantage of resampling is the generality of these methods which allows the application in a wide range of statistics [6]. They provide greater accuracy than classical methods, even in absence of large amounts of data. Classical applications of these methods are model validation or bias reduction of an estimate.

Cross-validation is the most often used method for error estimation in econometrics [11]. Using this method, new samples are created by systematically removing objects from the given sample. There are several cross-validation methods available which can be used in a wide field of science, e. g. for accuracy estimation and overfitting reduction of a model. The bootstrap method is closely related to that method. Instead of leaving out objects from the sample, the bootstrap method creates new objects by resampling the observations of the original sample. This method allows the estimation of almost any statistic using relatively simple methods [10]. It can be applied for statistical learning and provides a measure of variability, even in cases where it is otherwise difficult to obtain [6].

This paper now provides an introduction to resampling methods. In this context, we present two common approaches, cross-validation and bootstrap, with the aim to give an idea about the

central functionality of these methods. In a first step, we introduce the cross-validation method and illustrate different types of implementations and possible applications of this approach (Section 2). Accordingly, we present the bootstrap method and provide an example applications using R code (Section 3). Section 4 closes with a summary of the main topics.

2 Cross-Validation

This chapter provides an introduction to the cross-validation approach. Thus, we introduce several types of cross-validation (Section 2.1) and give an overview of possible applications (Section 2.2).

Cross-validation is a statistical method of evaluating the performance of a learning algorithm [9]. In order to find out the best model for the available data, the cross-validation method allows e. g. to estimate the accuracy of a learned model and makes it possible to compare it with the performance of different models. The basic idea of cross-validation is the division of the data into a training part and a validation part. Cross-validation uses only the training segment for the learning of the model and uses the rest of the data to evaluate the trained model. Afterwards, it may repeat this procedure until each segment has served as validation part. The division of the available data allows to measure the performance on data which is not used for training of the model, i. e. the validation part, and thus simulates the process of measuring the performance on unseen data. Therefore, a great advantage of this approach is the reduction of overfitting which may occur [8], if we measure the performance of a learning algorithm on the same data from which it was learned.

2.1 Types of Cross-Validation

There are several validation and cross-validation variations available. In general, they distinguish on the way how to divide the data into learning part and validation part. The basic cross-validation approach is called k-fold cross-validation. In fact, all other cross-validation variations are special cases of k-fold cross-validation. In the following, we illustrate several validation and cross-validation variations.

Resubstitution Validation

In this simple approach, the model is learned and validated from all available data. Because of over-fitting, this approach performs well on the available dataset, but poorly on unknown data.

Hold-Out Validation

Hold-out validations splits the available data into two non-overlapping parts. One is used for training and the other for validation. This approach reduces overfitting but does not use all available data. Subsequently, hold-out validation is highly dependent on the way of splitting the data for training and validation.

K-Fold Cross-Validation

Using k-fold cross-validation, we divide the dataset into k folds of approximately equal size. We treat the first fold as validation set, and train the model with the remaining $k-1$ folds. Subsequently, we calculate the desired statistics for the held-out fold. This procedure is repeated

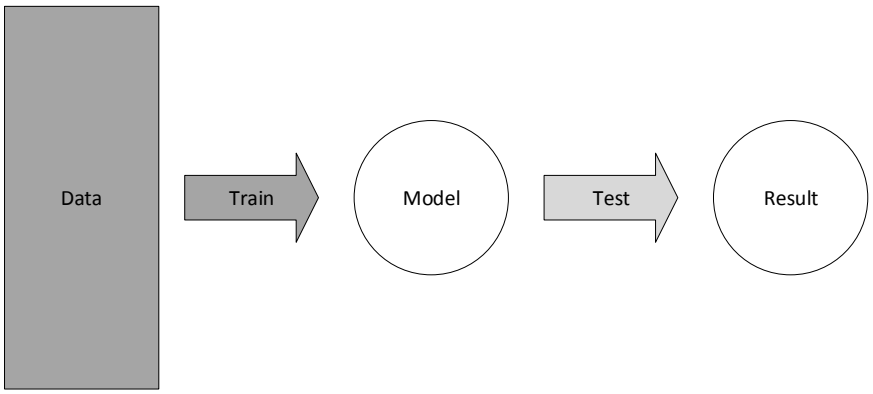


Figure 1: Schematic illustration of resubstitution validation.

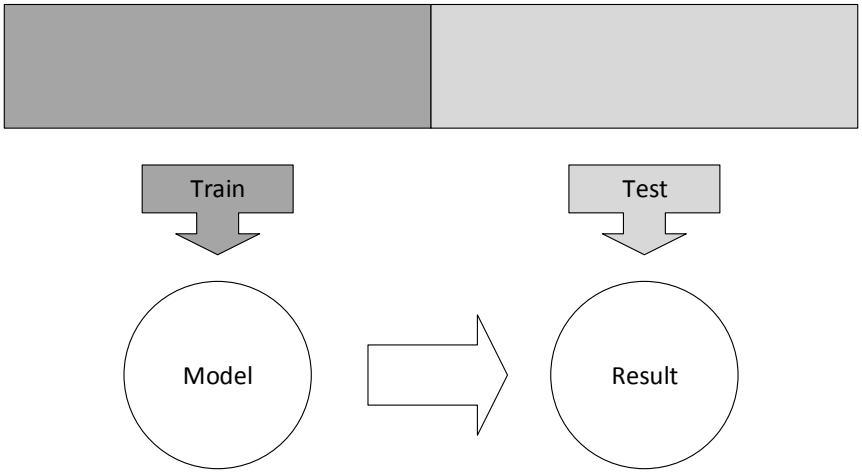


Figure 2: Schematic illustration of hold-out validation.

k times, while each time the next fold is treated as validation set. This leads us to k estimates of a statistic. Figure 3 illustrates this approach. Subsequently, we calculate the k-fold cross-validation estimate for the whole dataset by averaging the k estimates.

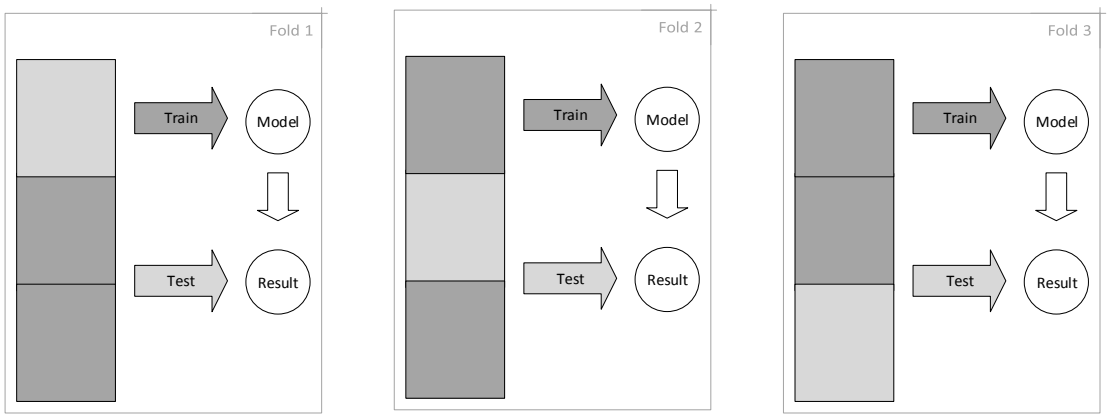


Figure 3: Schematic illustration of k-fold cross-validation.

Repeated K-Fold Cross-Validation

Repeated k-fold cross-validation extends the k-fold cross-validation approach with additional estimates. In this case, we run k-fold cross-validation multiple times and reshuffle the data before each round. This approach allows us to have a large number of performance estimates, but may have overlapping training and test data between each round and, thus, it may underestimate the performance variance.

Leave-One-Out Cross-Validation

Leave-One-Out cross-validation is a special case of k-fold cross-validation. In this approach, the number of folds k equals the number of instances in the data. The model is trained with all of the data except one single observation. Afterwards, the model is tested on that single observation (Figure 4). Leave-one-out cross-validation may have a high variance and can lead to unreliable results [9]. Nevertheless, it can be useful, if the available data is very rare.

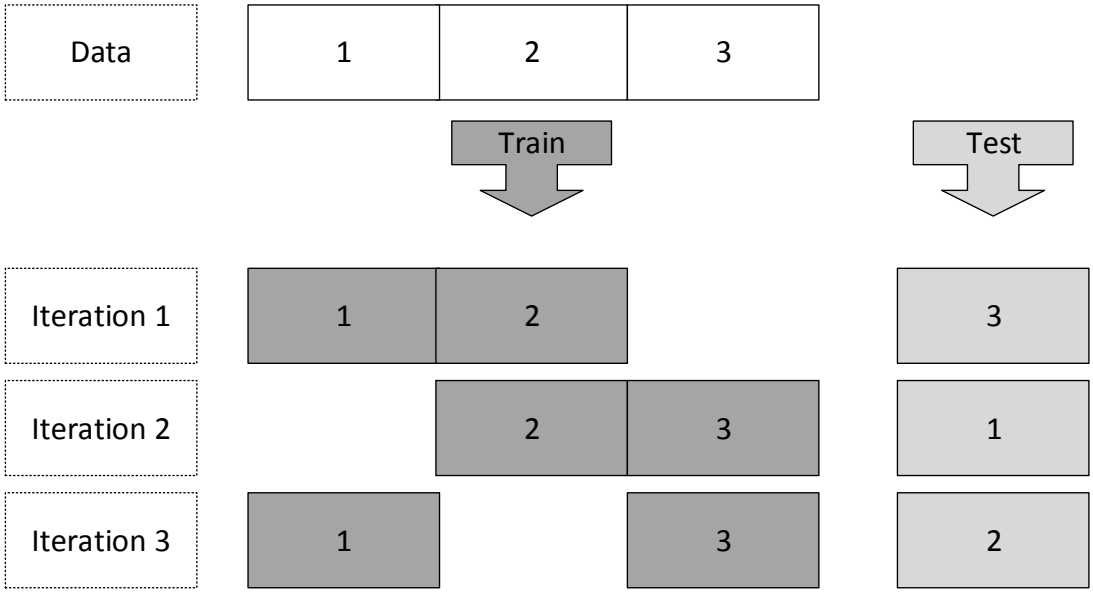


Figure 4: Schematic illustration of leave-one-out cross-validation.

2.2 Applications of Cross-Validation

Cross-validation is a frequently-used method for error estimation in econometrics [11]. Because of its universality, it is suitable for many applications in a wide field. Especially for performance estimation and model selection/tuning, cross-validation is an auxiliary method [9]. Thus, we illustrate possible application of cross-validation for several tasks in the following.

Performance Estimation

Cross-validation can be used for *performance estimation* of any measure, such as accuracy, precision, recall or F-score. Along with the objective to estimate the performance of a certain classifier on unseen data, this method reduces the problem of overfitting which may occur if the model is trained and validated with the same data. Especially, if there is only a limited amount of labeled data available, cross-validation (most common 10-fold cross-validation) is a suitable method to measure the performance of a model. Using 10-fold cross validation, the model gets

repeatedly trained with 90 % of the data and validated with a specific performance measure for the rest of the data. This method leads to reliable estimates, if the labeled data is sufficiently large and if the unlabeled data follows the same distribution as the labeled data [9].

Model Selection

If there are multiple models available, *model selection* describes the task of picking out the model that performs best on the dataset. The models and their performance can be compared using the cross-validation method. In case of comparing two classifiers, k-fold cross-validation leads to k pairs of performance values. Instead of comparing e. g. the average accuracies directly, it is possible to use a paired t-test [9] and set out a null hypothesis assumption that the algorithms perform equally, against a hypothesis using a two sample test.

Tuning

The optimal parameters of parametrized classifiers can be calculated using cross-validation and thus *tuned* to achieve the best result with a particular dataset. If a classifier contains only two classes $\{+,-\}$, the value of the parameters can be obtained by simply counting the positive or negative labeled instances and dividing this number by the total number of instances [9]. In cases, where the parameters do not have such intrinsic meaning the parameters can be obtained by trying out many values for the training part and picking out the one with the highest performance in the validation part.

3 Bootstrapping

This section defines and explains the bootstrap approach. First, we describe the method in general (Section 3.1), while we introduce an example application in a second step (Section 3.2). In this context, we provide exemplary R code for the calculation of t confidence intervals of a sample distribution using the bootstrap method.

3.1 The Bootstrap Idea

The bootstrap is a computer-based method for assigning measures of accuracy to sample estimates [4]. The method is derived from the jackknife method [3] and allows to generate an estimate for a sample distribution of almost any statistic.

The basic idea of bootstrapping is to infer from a given *sample* to a *population* by resampling the sample and treating the *sample* as the *population* from which it was drawn. Thus, the resamples represent the potential samples we would get if we are able to take many samples from the population. Thus, the distribution of a statistic, based on the resamples, namely the bootstrap distribution, represents the sampling distribution of the statistic [12]. This approach is an example for the plug-in principle [13] where we use a quantity based on the sample to approximate a similar quantity from the population.

The bootstrap method is recommended [1], if the theoretical distribution of a statistic of interest is complicated or unknown. In this case, bootstrapping provides an indirect method to assess the properties of the distribution of the underlying population and offers a measure of variability which is otherwise difficult to obtain [6]. In addition, bootstrapping allows to handle distortions, if the the sample size is insufficient and not fully representative [5]. In fact, the great

advantage of the resampling idea is that it often works even when theory fails. Assuming the existence of just one sample, the bootstrap method is still able to draw new samples from the existing single sample. These resamples can often serve as a good estimate for the case when it is possible to draw really new samples from the population [10]. Figure 5 illustrates that line of thought. Nevertheless, the bootstrap method can not increase the amount of information in the original data. Instead it estimates the variation of a statistic because of random sampling [5].

A great advantage of bootstrap is its simplicity. It can be easily applied to a wide range of statistical learning methods and provides a straightforward way for the calculation of complex parameters of a distribution, such as percentile points and correlation coefficient and leads to more accurate results than using sample variance and assumptions of normality [2].

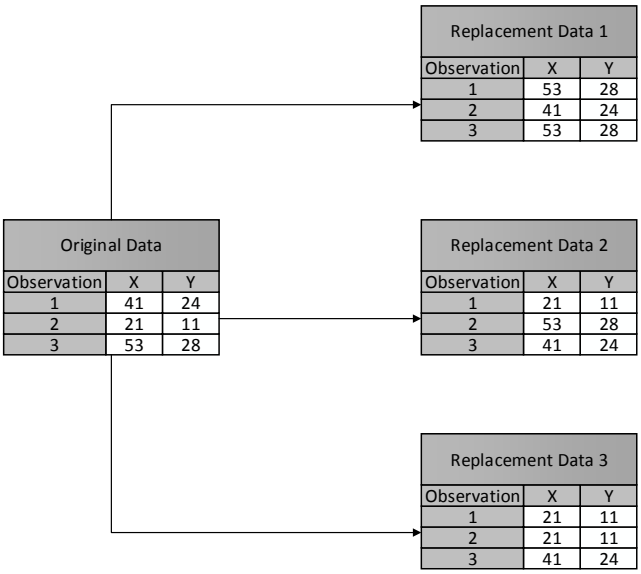


Figure 6: Procedure of drawing new bootstrap samples with replacement.

Bootstrapping generally follows the same procedure. At first, a given sample from a population is resampled a specified number of times, where each resample has the same size as the original sample. This resampling is done with replacement, which means that every observation in the sample can be drawn once, more than once, or not at all (Figure 6). In a second step, calculation of a specific statistic from each resample leads to a bootstrap distribution. This boostrap distribution provides information about the shape, center and spread of the original sample distribution. A flow diagram of the required steps for the bootstrap method is illustrated in Figure 7.

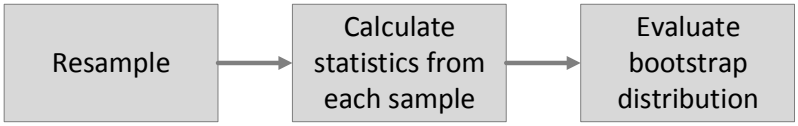
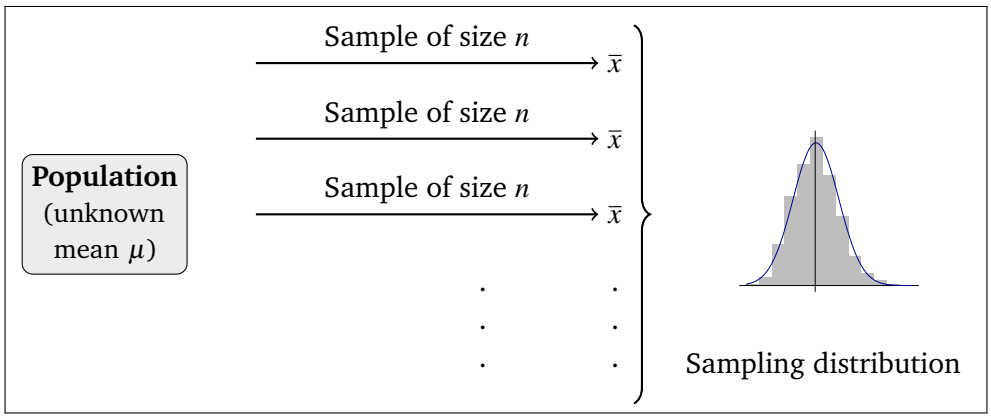
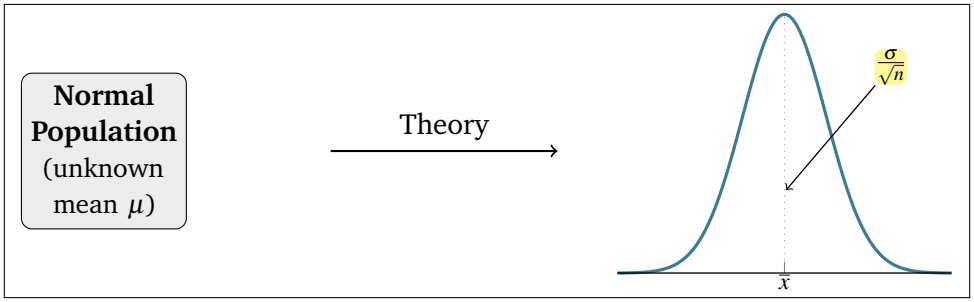


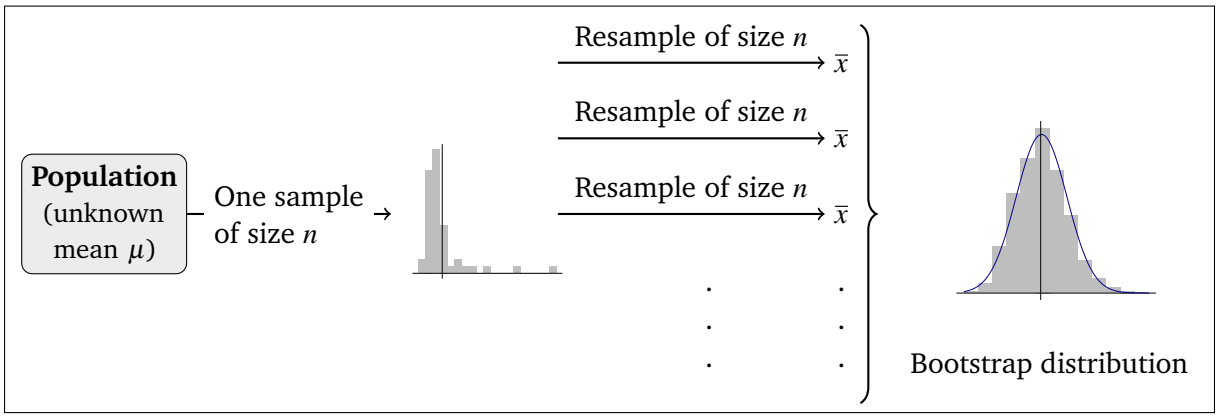
Figure 7: Illustration of the basic steps for the application of the bootstrap method.



(a) If a huge amount of samples are available, the values of $\bar{x}$ can be collected and evaluated in the sampling distribution.



(b) If the population follows a normal distribution, the central limit theorem indicates that the sampling distribution of $\bar{x}$ is also normal.



(c) If theory fails and only one sample is available, the sample simulates thee population and the bootstrap distribution of $\bar{x}$ simulates the sampling distribution.

Figure 5: The bootstrap idea, if the population is unknown and only one sample is available, application of the bootstrap method can simulate the process of drawing new samples from the population.

Bootstrap for the Standard Error of Mean

One possible application where the bootstrap method can be applied is the calculation of the standard error of mean of a sample distribution (nevertheless, this application is only useful if the central limit theorem is not practicable, e. g. if the the sample is too small). Assuming the central limit theorem, the estimated standard error of a mean $\bar{x}$, based on n independent data points $x_1,x_2,...,x_n$, $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, is given by

$$\frac{s}{\sqrt{n}}, \tag{1}$$

where s denotes the standard error of a given sample. We apply the bootstrap approach by generating a large number of independent bootstrap samples $x^{*1},x^{*2},...,x^{*B}$, each of size n from a given sample (Figure 8). Typical values for B in case of standard error estimation range from 50 to 200 [4].

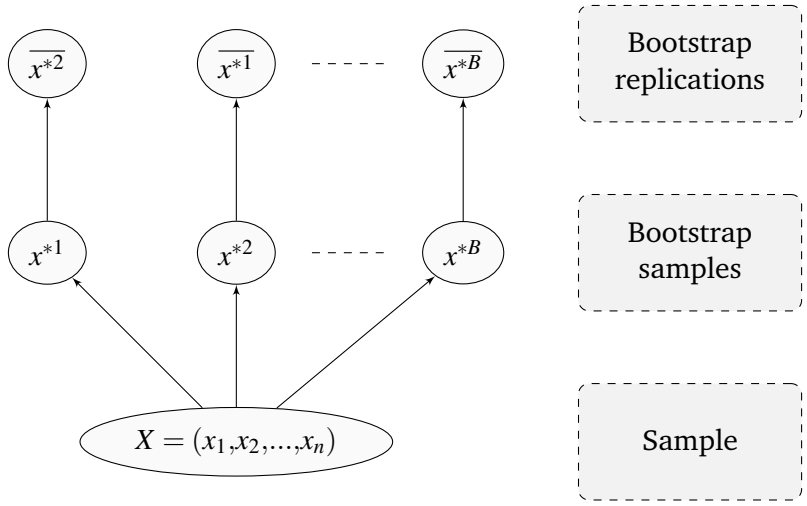


Figure 8: Schematic illustration of the bootstrap process for the calculation of the standard error of mean.

In the next step, we calculate the mean of each bootstrap sample $\overline{x^{*b}}$. At least, the bootstrap standard error based on B resamples, is calculated by

$$\hat{se}_{boot,\bar{x}} = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\bar{x}^{*b} - \frac{1}{B}\sum_{b=1}^{B}\bar{x}^{*b})^2}. \tag{2}$$

The bootstrap standard error $\hat{se}_{boot,\bar{x}}$ serves as an approximation for the standard error of $\bar{x}$ of the original sample.

In the following, we present a simple R Code example for this application. We use a sample consisting of 100 stock price observations and generate $B = 1000$ bootstrap samples. We find a standard error for the mean of the original sample, based on Equation (1), of 0.1062. Using our bootstrap samples, we find a value for the bootstrap standard error $\hat{se}_{boot,\bar{x}}$ of 0.1049 which agrees closely with the formula based estimate.

```
library(ISLR)
set.seed(1)
portfolio.mean=function(data,index){
  return (mean(data$X[index]))
}
boot(Portfolio,portfolio.mean,R=1000)
```

3.2 Example: Bootstrap for *t* Confidence Intervals

This section provides an example where the application of the bootstrapping method is highly useful. Our aim is to calculate the t confidence interval for a mean using a sample consisting of 50 sales prices of residential property in Seattle from 2002 [5]. This sample contains several outliers and its distribution is skewed to the right (Figure 9).

By drawing 1000 resamples from the original sample we get a bootstrap distribution of the sample means (Figure 10a) which is also skewed, but has got a small bias (the sample mean is approximately equal to the mean of the bootstrap distribution). Nevertheless, we can't assume normality in this case. Although, there are bootstrap methods which can handle distributions not based on normality [5], the easist way is to use a different statistic that is more resistant to skewness and outliers. In this context, we use the 25 % trimmed mean, which is the mean of the center observations in a dataset, leaving out the 25 % smallest and largest observations. By drawing again 1000 resamples and calculating the 25 % trimmed mean for each sample, we get a bootstrap distribution which is much closer to normality (Figure 10b). Using the bootstrap distribution and assuming approximately normal distribution and a small bias, we are able to calculate a t confidence interval by

$$\text{statistic} \pm t \cdot \text{SE}_{\text{boot,statistic}}. \tag{3}$$

Thus, the bootstrap t confidence interval for the trimmed mean is given by

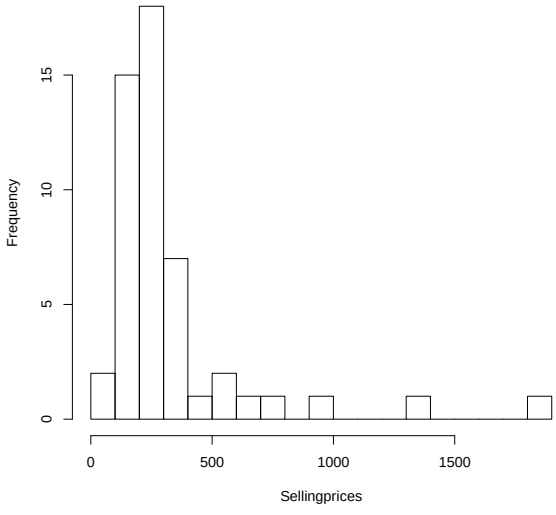
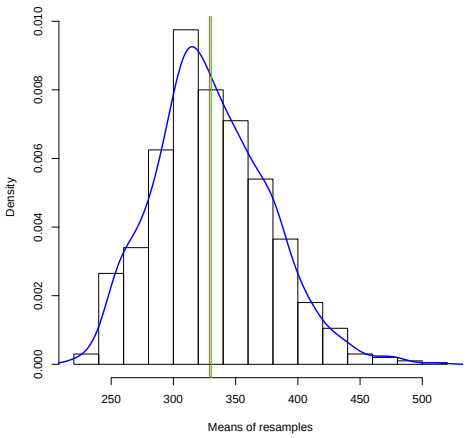
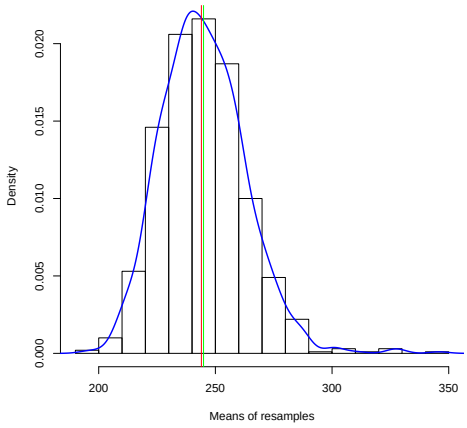


Figure 9: Histogram of 50 selling prices (in $1000) observing a skewed distribution with several outliers.



(a) The bootstrap distribution of the sample means of 1000 resamples (skewed).



(b) The bootstrap distribution of the 25 % trimmed means of 1000 resamples (close to normality).

Figure 10: Bootstrap distribution of 1000 resamples for sample means in comparison with bootstrap distribution of 25 % trimmed means. The blue curve in both plots displays a smoothed density curve for the corresponding bootstrap distribution.

$$\bar{x} \pm t \cdot \mathrm{SE}_{boot,\bar{x}_{25\%}}, \tag{4}$$

which leads to a 95 % confidence interval of (208.3, 278.5). Thus, we are 95 % confident, that the 25 % trimmed mean for the population of real estate sales in Seattle from 2002 is between $208,300 and $278,500.

```
sellingprices=c(142, 232, 132.5, 200, 362, 244.95, 335, 324.5, 222, 225,
                175, 50, 215, 260, 307, 210.95, 1370, 215.5, 179.8, 217,
                197.5, 146.5, 116.7, 449.9, 266, 265, 256, 684.5, 257, 570,
                149.4, 155, 244.9, 66.407, 166, 296, 148.5, 270, 252.95, 507,
                705, 1850, 290, 164.95, 375, 335, 987.5, 330, 149.95, 190)

trimmed_mean.fn=function(data,index){
  return (mean(data[index],trim=0.25))
}

set.seed(0)
selling.boot=boot(sellingprices, trimmed_mean.fn, R=1000)
boot.ci(selling.boot, conf=0.95, type="norm")
```

4 Summary

Resampling methods are a set of statistical inference methods based on drawing new samples from an initial sample. In absence of a sufficient amount of samples and under availability of sufficient computing power, these methods allow to infer to an unknown distribution of a population by simulating new samples from the given sample and thus simulating the process of generating new original samples from the population. Cross validation, which is often used for accuracy estimation of a model, creates new samples by systematically removing objects from a

given sample. The bootstrap method instead resamples the observations of a given sample, in order to draw new samples and to infer to the distribution of an unknown population. Because of its generalizability, this method allows the estimation of almost any statistic using relatively simple methods.

In this paper, we gave an introduction to two common resampling methods, namely cross-validation and bootstrap. In this context, we reviewed the motivation and basic functionality of these methods. Furthermore, we introduced the cross-validation method. In this context, we revealed several types of cross-validation and possible applications in research. In addition, we presented and illustrated the bootstrap method. For this purpose, we also introduced an example application for the calculation of t confidence intervals for a sample mean, providing the corresponding R code for additional illustration.

A References

[1] H. J. ADÈR, G. J. MELLENBERGH, and D. J. HAND. *Advising on Research Methods: A Consultant's Companion*. Huizen and Netherlands: Johannes van Kessel Pub., 2008. ISBN: 978-90-79418-01-5.

[2] T. J. DICICCIO and B. EFRON. *Bootstrap Confidence Intervals*. In: *Statistical Science* (1996), pp. 189–212.

[3] B. EFRON. *Bootstrap Methods: Another Look at the Jackknife*. In: *The Annals of Statistics*, Vol. 7, No. 1 (1979), pp. 1–26. ISSN: 0090-5364.

[4] B. EFRON and R. J. TIBSHIRANI. *An Introduction to the Bootstrap*. Vol. 57. CRC press, 1994.

[5] T. HESTERBERG et al. *Bootstrap Methods and Permutation Tests*. In: *Introduction to the Practice of Statistics*, Vol. 5 (2005), pp. 1–70.

[6] G. JAMES. *An Introduction to Statistical Learning: With Applications in R*. Vol. 103. Springer texts in statistics. New York and NY: Springer, 2013. ISBN: 1461471389.

[7] C. Z. MOONEY, R. D. DUVAL, and R. DUVAL. *Bootstrapping: A Nonparametric Approach to Statistical Inference*. Sage, 1993.

[8] A. Y. NG. *Preventing Overfitting of Cross-Validation Data*. In: *ICML*. Vol. 97. 1997, pp. 245–253.

[9] P. REFAEILZADEH, L. TANG, and H. LIU. *Cross-Validation*. In: *Encyclopedia of Database Systems*. Springer, 2009, pp. 532–538.

[10] H. VARIAN. *Bootstrap Tutorial*. In: *Mathematica Journal*, Vol. 9, No. 4 (2005), pp. 768–775.

[11] R. WEHRENS, H. PUTTER, and L. BUYDENS. *The Bootstrap: A Tutorial*. In: *Chemometrics and intelligent laboratory systems*, Vol. 54, No. 1 (2000), pp. 35–52.

[12] D. S. WILKS. *Statistical Methods in the Atmospheric Sciences*. Vol. 100. Academic press, 2011.

[13] D. WRIGHT. *Using Bootstrap Estimation and the Plug-in Principle for Clinical Psychology Data*. In: *Journal of Experimental Psychopathology*, Vol. 2, No. 2 (2011), pp. 252–270. ISSN: 20438087.

B List of Figures