Eötvös Loránd University

Faculty of Science

An overview of the Bootstrap

Nemes Balázs Amadé

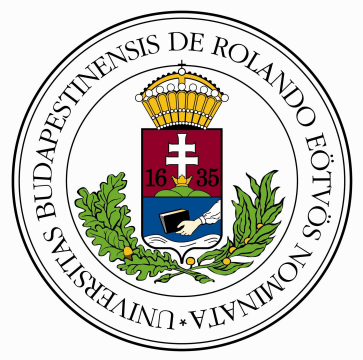
Mathematics BSc, Pure Mathematics specialization

Bachelor Thesis

Supervisor:

Pröhle Tamás

Department of Probability Theory and Statistics



Budapest, 2019

Contents

Chapter 1

Introduction

The bootstrap is a computer based method of assigning measures of accuracy to estimators, falling into the general category of resampling methods. The spiritual successor of the jackknife, another resampling method mainly used for variance and bias estimation, the bootstrap proves much more versatile and accurate. This is achieved at the cost of computing power, a commodity that is becoming ever cheaper with the exponential growth of computing power.

The first section discusses the nonparametric bootstrap, the most conceptually simple setting for the bootstrap. In the following section, the basic idea of bootstrapping is generalized, making it applicable in more complicated situations, such as linear regression and autoregressive models for time series. We shall then discuss the jackknife and it's geometric connection to the bootstrap. Finally, we shall take a look at confidence interval construction using the bootstrap, a topic that has sparked much theoretical discussion.

All chapters of this thesis, with the exception of chapter 5, are reconstructions of material found in selected chapters from Bradley Efron and Robert J. Tibshirani's 1994 monograph, An Introduction to the Bootstrap [1], according to my understanding of them, with some alternative proofs and datasets. Chapter 5 is mainly based on Efron's 1987 paper [2].

Chapter 2

Non-parametric bootstrap

2.1 Plug-in estimates

We start with one of the most basic problems of statistical inference: a probability distribution F and a random sample of values $\mathbf{x} = (x_1, x_2, ..., x_n)$ drawn from F, based on which we wish to estimate some parameter $\theta = t(F)$ of F. By random sample we mean the usual:

Definition 2.1.1. $\mathbf{x} = (x_1, x_2, ..., x_n)$ is a random sample drawn from F, if $x_1, x_2, ..., x_n$ are independent and identically distributed random variables with $x_i \sim F$.

Let's say we estimate θ with a statistic based on $\mathbf{x}$, $\hat{\theta} = s(\mathbf{x})$. How accurate of an estimate is $\hat{\theta}$? The bootstrap was initially introduced in 1979 by Bradley Efron [3] as a computer based method of estimating the standard error of $\hat{\theta}$. By standard error we mean the standard deviation of a summary statistic, in this case $\hat{\theta}$. Standard error is of course not the only quantifier of accuracy, we could try to estimate $\hat{\theta}$'s bias, or construct a confidence interval. Indeed, bootstrapping can be used to do all of the above, but for now we shall only discuss standard error.

Before we dive into how the bootstrap works, it is first necessary to become familiar with some concepts. In the inferential problem described above, we can define the the following.

Definition 2.1.2. The empirical distribution $\hat{F}$ is the discrete probability distribution putting $1/n$ probability on all values $x_i, i = 1, ..., n$ of the observed sample.

The empirical distribution $\hat{F}$ is a simple estimate of the original distribution F based on $\mathbf{x}$, without making any parametric assumptions, eg.: that x_i has a normal distribution

with unknown mean parameter μ and variance of 1, or with more compact notation $x_i \sim N(\mu, 1)$. A standard result of statistics, the Glivenko-Cantelli theorem tells us that the cumulative distribution functions (CDFs) $G_n(t)$ of the empirical distributions $\hat{F}_n$ based on a sample of size n from F, converge uniformly with probability one to the CDF of F as $n \to \infty$. This reassures that $\hat{F}$ is a good choice, in the asymptotic sense at least. For small sample sizes, however, $\hat{F}$ can sometimes be a quite misleading estimate of F, so one must be careful when using non-parametric inference in small sample size situations.

The empiric distribution gives rise to the natural idea of estimating the parameter of interest $\theta = t(F)$ by applying $t(\cdot)$ to our estimated distribution $\hat{F}$, thus giving a so called *plug-in estimate* $\hat{\theta} = t(\hat{F})$ of θ. This method of estimation is sometimes called the *plug-in principle*. Some examples of plug-in estimates are:

- The plug in estimate of an expectation parameter $\theta = E_F(x)$ is known as the *sample mean* and is equal to
$$\hat{\theta} = E_{\hat{F}}(x) = \frac{1}{n}\sum_{i=1}^{n} x_i := \overline{x}. \tag{2.1}$$
- The plug in estimate of a variance parameter $\theta = E_F((x - E_F(x))^2)$, called the *sample variance* is
$$\hat{\theta} = E_{\hat{F}}((x - E_{\hat{F}}(x))^2) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2. \tag{2.2}$$

Note that often a corrected version of the sample variance is used

$$\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 \tag{2.3}$$

as this makes the estimate unbiased for the real variance of the distribution F.

2.2 Bootstrap estimate of standard error

Returning to our problem of estimation: the quantity we are interested in is the standard error of $\hat{\theta} = s(\mathbf{x})$ under distribution F, denoted by $se_F(\hat{\theta})$. As we do not wish to make any parametric assumptions of F, the plug-in principle will serve us well. Using the plug in principle, we can estimate this quantity with $se_{\hat{F}}(\hat{\theta}^*)$, the *ideal bootstrap estimate* of standard error. This warrants a definition.

Definition 2.2.1. The ideal bootstrap estimate of the standard error of $\hat{\theta}$ is $se_{\hat{F}}(\hat{\theta}^*)$

The $\hat{\theta}^*$ notation serves to differentiate between $\hat{\theta} = s(\mathbf{x})$, which is based on a $\mathbf{x}$ sample drawn from F, and $\hat{\theta}^* = s(\mathbf{x}^*)$, where the $\mathbf{x}^*$ *bootstrap sample* is drawn from the estimated distribution $\hat{F}$. Now that we have defined the ideal estimate, the question is, how do we calculate it? First we take a look at the mean as an estimator, for which there is an explicit formula, then the general case.

The mean estimator $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, usually used to estimate the mean of a distribution, has a quite simple formula for it's variance:

$$var\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{\sum_{i=1}^{n} var(x_i)}{n^2} = \frac{var(x)}{n}. \tag{2.4}$$

Therefore, the ideal bootstrap estimate of the mean's standard error $\sigma_F/\sqrt{n}$, is simply $\sigma_{\hat{F}}/\sqrt{n}$. Having already calculated the plug in estimate of a variance parameter in (2.2), we have

$$se_{\hat{F}}(\overline{x}^*) = \sigma_{\hat{F}}/\sqrt{n} = \left(\sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{n^2}\right)^{\frac{1}{2}}. \tag{2.5}$$

For other, more complicated statistics such simple expressions for the ideal bootstrap estimate may be hard figure out, or may not even exist. We can, however, approximate it using bootstrap sampling. As the distribution $\hat{F}$, unlike F, is readily available to us, we can draw as many bootstrap samples $\mathbf{x}^*$ as we wish. Drawing samples from $\hat{F}$ is exactly the same as drawing with replacement from $\mathbf{x}$. Applying s$(\cdot)$ to these bootstrap samples gives us *bootstrap replications* of the original $\hat{\theta}$ statistic, and using these we can calculate their (corrected) sample standard deviation, giving us an approximation to $se_{\hat{F}}(\hat{\theta}^*)$.

The algorithmic description of the non-parametric bootstrap procedure for estimating standard error is the following.

1. Draw B independent bootstrap samples $\mathbf{x}^{*1}, \mathbf{x}^{*2}, \ldots, \mathbf{x}^{*B}$, each consisting of n values drawn with replacement from $\mathbf{x}$.

2. Calculate the bootstrap replication based on each sample

$$\hat{\theta}^*(b) = s(\mathbf{x}^{*b}) \qquad b = 1, 2, \ldots, B$$

3. Estimate the standard error $se_F(\hat{\theta})$ by the corrected sample standard deviation of

the B replications.

$$\hat{se}_B = \left(\frac{1}{B-1} \sum_{b=1}^{B} \left(\hat{\theta}^*(b) - \hat{\theta}^*(\cdot) \right)^2 \right)^{\frac{1}{2}}$$

where $\hat{\theta}^*(\cdot) = \sum_{b=1}^{B} s(\mathbf{x}^{*b})/B$

This gives us, finally, a new definition.

Definition 2.2.2. The bootstrap estimate of $se_F(\hat{\theta})$ based on B bootstrap replications is

$$\hat{se}_B = \left(\frac{1}{B-1} \sum_{b=1}^{B} \left(\hat{\theta}^*(b) - \hat{\theta}^*(\cdot) \right)^2 \right)^{\frac{1}{2}}$$

In essence, we solved the problem of approximating $se_{\hat{F}}(\hat{\theta}^*)$ by applying the plug-in principle again, in this instance to the sample of size B consisting of our bootstrap replications. While the original $F \longrightarrow \hat{F}$ approximation's accuracy is limited by the size n of the original $\mathbf{x}$ sample, this approximation can be made arbitrarily accurate by increasing the number B of bootstrap replications used, as shown by the Glivenko-Cantelli theorem. The recommended number B of bootstrap samples for high accuracy is about 200, though fewer may be used when accuracy is less important, or in cases when the calculational complexity of the function $s(\cdot)$ makes the cost of creating many bootstrap replications prohibitive.

We end this section by providing an example use of the bootstrap, using an increasing number of replications. Consider 40 normal variables with mean 5 and variance 4, the realizations of which are pictured in Figure (2.1), with the mean statistic as an estimator of the expected value of the population. In this case, as shown in (2.5) we can explicitly calculate the ideal bootstrap estimate of standard error. Table 2.1 shows the bootstrap estimate of standard error using an increasing number of replications. We can see that the values converge, as we would expect, to the ideal bootstrap estimate of standard error. 200 replications give a solid estimate, while 2000 replications serve as a very close approximation.

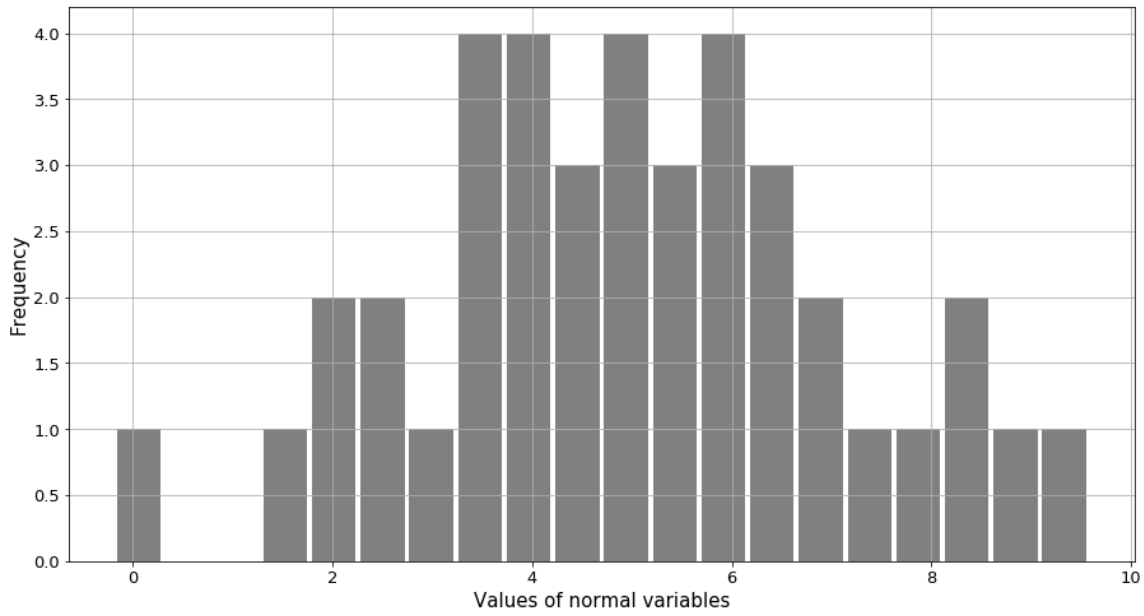


Figure 2.1: *Histogram of 40 normal variables with a mean of* 5 *and variance of* 4

No. of replications	$B = 50$	$B = 100$	$B = 200$	$B = 2000$	$\sigma_{\hat{F}}/\sqrt{n}$
$\hat{se}_B$	0.2948	0.3132	0.3352	0.3176	0.3237

Table 2.1: *Bootstrap estimate of standard error, using ever larger number of bootstrap samples*

2.3 Bootstrap estimate of bias

In this section we will remain in the nonparametric one sample situation as before, but will now be discussing *bias* instead of standard error.

Definition 2.3.1. The bias of $\hat{\theta} = s(\mathbf{x})$ as an estimate of θ is defined to be

$$bias_F = bias_F(\hat{\theta}, \theta) = E_F[s(\mathbf{x})] - \theta. \tag{2.6}$$

Large bias is an undesirable property of an estimate, so methods of estimating an estimator's bias can prove quite useful when assessing the accuracy of said estimator. Unbiased estimates are nonetheless not always desirable over other estimates; we would for example prefer a slightly biased estimate with small standard error over an unbiased estimate with large standard error.

We can use the same logic as before to define an *ideal bootstrap estimate of bias*.

Definition 2.3.2. The ideal bootstrap estimate of $bias_F$ is $bias_{\hat{F}} = E_{\hat{F}}[s(\mathbf{x}^*)] - t(\hat{F})$

Here we substituted in $\hat{F}$ for F twice according to the plug-in principle, in $t(F)$ and $E_F[s(\mathbf{x})]$. Now, again according to the logic of the previous section, we approximate $E_{\hat{F}}[s(\mathbf{x}^*)]$ by the sample mean of B bootstrap replications.

$$E_{\hat{F}}[s(\mathbf{x}^*)] \approx \hat{\theta}^*(\cdot) = \sum_{b=1}^{B} \hat{\theta}^*(b)/B = \sum_{b=1}^{B} s(\mathbf{x}^{*b})/B$$

Definition 2.3.3. The bootstrap estimate of bias based on B replications is

$$\widehat{bias}_B = \hat{\theta}^*(\cdot) - t(\hat{F}).$$

The accuracy of the formula above can be improved upon when the original estimate is a plug-in estimate $\hat{\theta}^* = t(\hat{F})$. We will describe how this more efficient method works, as the resampling notation required for it will play an important role in Chapter 3, but we will not prove the improvements made. For more details on the method see Chapter 23 of Efron 1994 [1].

We need to define the concept of resampling vectors for our new method. Let P_j^* denote the the proportion of data values in a bootstrap sample $\mathbf{x}^* = (x_1^*, x_2^*, \ldots, x_n^*)$ that equal the j-th original data value.

$$P_j^* = \#\{x_i^* = x_j\}/n \qquad j = 1, 2, \ldots, n$$

The resampling vector

$$\mathbf{P}^* = (P_1^*, P_2^*, \ldots, P_n^*)$$

consists of these components, the sum of which equals one. We can now represent a bootstrap sample by it's corresponding resampling vector, though the order of the values x_j will be lost. Note that $n\mathbf{P}^*$ has a multinomial distribution with n draws and equal class probabilities.

$$\mathbf{P}^* \sim \frac{1}{n} Mult(n, \mathbf{P}^0).$$

Let us define $\mathbf{P}^0 = (\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n})$, which can be thought of as the resampling vector corresponding to a bootstrap sample identical to the original sample, except for the possible reordering of the values. A resampling vector $\mathbf{P}^*$ also implies a corresponding probability measure F^*, namely the one putting probability P_j on the value x_j for $j = 1, 2, \ldots, n$,

which is independent of the ordering in the bootstrap sample. This gives us an alternate way of thinking of our original plug-in estimate

$$\hat{\theta} = t(\hat{F}) = t(F^0) := T(\mathbf{P}^0)$$

and of the bootstrap replications

$$\hat{\theta}^*(b) = t(F^{*b}) = T(\mathbf{P}^{*b}).$$

Thus we can rewrite the bootstrap bias estimate as

$$\widehat{bias}_B = \hat{\theta}^*(\cdot) - T(\mathbf{P}^0).$$

Defining $\overline{P}^* := \sum_{b=1}^{B} \mathbf{P}^{*b}/B$, the *better bootstrap bias estimate*, denoted by $\overline{bias}_B$, is

$$\overline{bias}_B = \hat{\theta}^*(\cdot) - T(\overline{P}^*).$$

The benefit of switching $T(\mathbf{P}^0)$ to $T(\overline{P}^*)$ has to do with Monte Carlo sampling adjustments. In essence, the bootstrap estimates of bias and variance can be viewed as Monte Carlo estimates, that is, estimates gained by performing Monte Carlo sampling. Therefore, sampling adjustments commonly used for Monte Carlo estimates can be used to improve the efficiency of bootstrap calculations. Sampling adjustments can also be used to improve bootstrap sampling for the confidence intervals discussed in Chapter 5. For more details, see Chapter 23 of Efron 1994 [1].

Chapter 3

The Jackknife

3.1 The Jackknife estimate of standard error and variance

The jackknife, the predecessor of the bootstrap, is a resampling technique used for variance and bias estimation. It was first proposed by Maurice Quenouille in the mid 1950-s, and refined in 1956. John Tukey expanded on the method in 1956 and coined the name, since like an actual jackknife, it's versatile tool that gives a quick solution for most problems, though each individual problem may be more efficiently solved by a more specific approach.

In the usual one sample, non-parametric situation, given a data set $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, the ith *jackknife sample* $\mathbf{x}_{(i)}$, is defined to be simply $\mathbf{x}$ with the ith data value removed

$$\mathbf{x}_i = (x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n). \tag{3.1}$$

The ith *jackknife replication* $\hat{\theta}_{(i)}$ of the statistic $\hat{\theta} = s(\mathbf{x})$ is $s(\cdot)$ evaluated for $\mathbf{x}_{(i)}$. The *jackknife estimate of bias* is defined by

$$\widehat{bias}_{jack} = (n-1)(\hat{\theta}_{(\cdot)} - \hat{\theta}) \tag{3.2}$$

where

$$\hat{\theta}_{(\cdot)} = \sum_{i=1}^{n} \hat{\theta}_{(i)}/n. \tag{3.3}$$

Note that this definition implicitly assumes that the statistic $s(\cdot)$ can also be applied to

the jackknife data set.

The jackknife estimate of bias works only for plug-in statistics $\hat{\theta} = t(\hat{F})$, and only in cases when the statistic $T(\cdot)$ in resampling form $T(\mathbf{P}^*)$, is twice differentiable. We will show in the following section that $\widehat{bias}_{jack}$ is a quadratic approximation to the ideal bootstrap estimate of bias $bias_{\hat{F}}$, which explains why the formula breaks down for cases when $T(\mathbf{P}^*)$ is not twice differentiable. Though a less accurate approximation to $bias_{\hat{F}}$ than $\widehat{bias}_B$ or $\overline{bias}_B$, it has the advantage of being far less computationally intensive, only requiring n recomputations of the function $t(\cdot)$ as opposed to the B required by bootstrap estimates (where B must be at least 200 even for $\overline{bias}_B$).

There is also a jackknife estimate of standard error, denoted here by $\widehat{bias}_{jack}$, which was developed by John Tukey in the late 1950's. It is only applicable when working with a plug-in estimate and serves as a quickly computable, less accurate alternative to se_B. We will show the jackknife estimate of standard error in the following to be a linear approximation to the ideal bootstrap estimate of standard error, meaning the formula is reliable only when $T(\cdot)$ is differentiable.

Definition 3.1.1. The jackknife estimate of standard error is defined to be

$$\hat{se}_{jack} = \left(\frac{n-1}{n} \sum_{i=1}^{n} (\hat{\theta}_{(i)} - \hat{\theta}_{(\cdot)})^2 \right)^{1/2}.$$

3.2 Geometric relation between the bootstrap and jackknife

In this section we are still in the one sample, non parametric situation. We assume that the paramater $\theta = t(F)$ is estimated by a plug in estimate $\hat{\theta} = t(\hat{F})$. Returning to the resampling vector notation of Section 2.3, recall that we defined $\mathbf{P}^0 = (\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n})^T$. The observed value of the statistic is then $T(\mathbf{P}^0)$. The bootstrap replications of $T(\mathbf{P}^0)$ can be thought of as being realizations of $T(\mathbf{P}^*)$, where $\mathbf{P}^*$ has distribution

$$\mathbf{P}^* \sim \frac{1}{n} Mult(n, \mathbf{P}^0). \tag{3.4}$$

For future reference, the mean and covariance of this distribution is

$$\mathbf{P}^* \sim \left(\mathbf{P}^0, \left[\frac{\mathbf{I}}{n^2} - \frac{\mathbf{P}^0 \mathbf{P}^{0^T}}{n} \right] \right) \tag{3.5}$$

where $\mathbf{I}$ is the $n \times n$ identity matrix. Because resampling vectors $\mathbf{P}^* = (P_1^*, P_2^*, \ldots, P_n^*)$ have the property $\sum_{i=1}^n P_i = 1$ and $\forall i : 0 \leq P_i \leq 1$, they take values on the n-dimensional simplex, denoted by S_n.

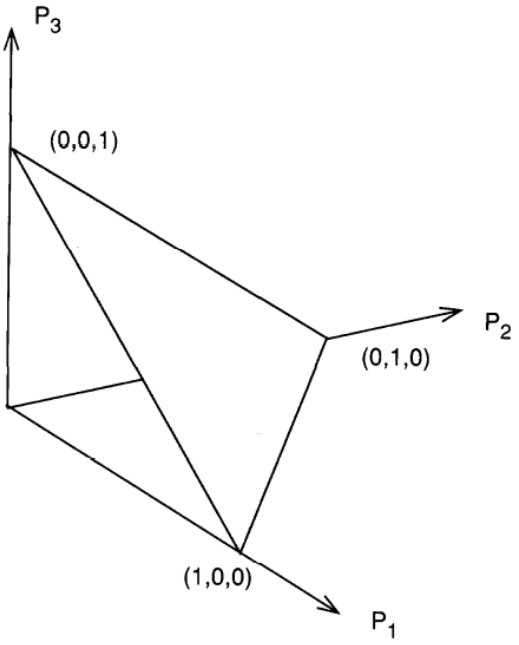


Figure 3.1: *Simplex in three dimensional space, aka.* S_3

The images in this section, taken from An Introduction to the Bootstrap [1] will represent the $n = 3$ case, to help give a geometric image for the statements made. In this section we will use $\mathbf{P}^*$ to denote resampling vectors, while $\mathbf{S}$ indicates a generic point of the simplex S_n. Figure 3.1 shows the simplex for $n = 3$, while Figure 3.2 shows it laid flat, with points of interest marked. We assume that the statistic $T(\cdot)$ can be evaluated on any point on the simplex, meaning that $t(\cdot)$ can be evaluated for any distribution putting arbitrary positive probability weights, summing up to one, on the observed values of the original sample.

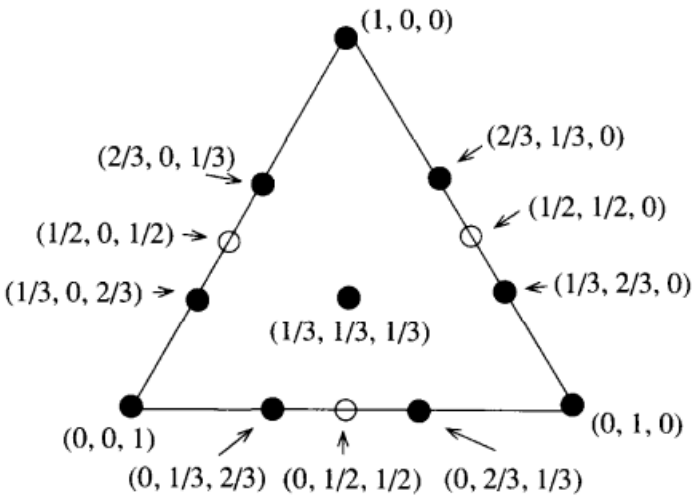


Figure 3.2: *S_3, laid flat. The solid points show possible values of the resampling vector $\boldsymbol{P}^*$, while the open circles show the jackknife points*

Recall that the jackknife values of the statistic are defined to be

$$\hat{\theta}_{(i)} = T(\mathbf{P}_{(i)}) \tag{3.6}$$

where

$$\mathbf{P}_{(i)} = \left(\frac{1}{n-1}, \frac{1}{n-1}, \ldots, 0, \frac{1}{n-1}, \ldots, \frac{1}{n-1}\right)^T, \tag{3.7}$$

the 0 being in coordinate i. These points are shown on the three dimensional simplex in Figure 3.2. The statistic $T(\mathbf{S})$ can be thought of as a surface over S_n, as shown in Figure 3.3. Every point of the simplex is a vector of probabilities $\mathbf{S}$ summing up to 1, and the value above it is $t(\cdot)$ applied to the corresponding probability distribution $F_{\mathbf{S}}$.

The ideal bootstrap estimate of variance can be expressed, using the resampling notation, as

$$var_* T(\mathbf{P}^*),$$

where var_* indicates variance under the distribution (3.5). The ideal bootstrap estimate of variance will in this chapter serve as the gold standard, and we will show how the jackknife and other estimates of variance serve as approximations to it.

Definition 3.2.1. A *linear statistic* T($\mathbf{S}$) is defined to be a statistic of the form

$$T(\mathbf{S}) = c_0 + (\mathbf{S} - \mathbf{P}^0)^T \mathbf{U}$$

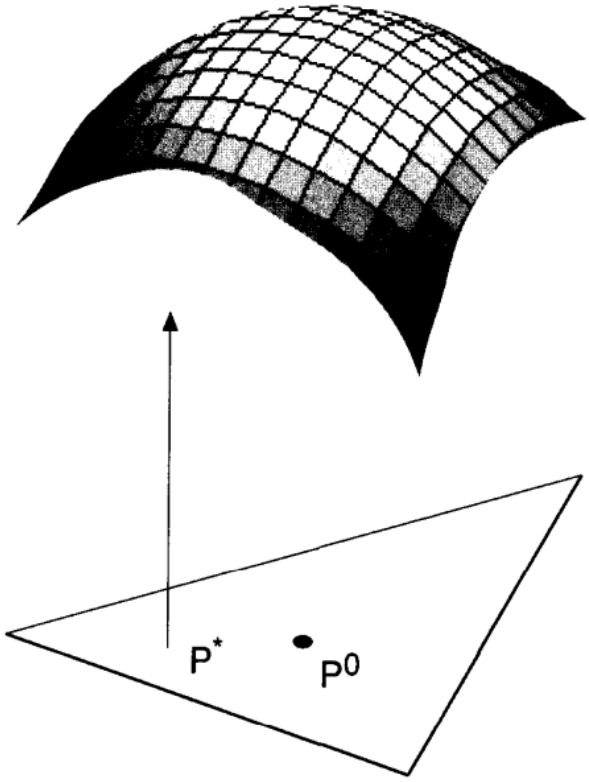


Figure 3.3: $T(\boldsymbol{S})$ *viewed as a surface above the simplex*

where c_0 is a constant and $\mathbf{U} = (U_1, U_2, \ldots, U_n)$ is a vector satisfying $\sum_{i=1}^{n} U_i = 0$.

A linear statistic, as a surface, defines a hyperplane over the simplex S_n. A simple example of a linear statistic is the sample mean $\overline{\mathbf{x}}^* = \sum_1^n P_i^* x_i$, satisfying the definition with $c_0 = \overline{\mathbf{x}}$ and $U_i = x_i - \overline{\mathbf{x}}$. Now, with definitions and notations in hand, we can present the following result, which states that the jackknife estimate of variance for $T(\mathbf{P}^*)$ is (almost) equal to the bootstrap estimate of variance for a statistic that is a linear approximate of $T(\mathbf{P}^*)$.

3.2.1. Theorem. *Let T^{LIN} be the unique hyperplane passing through the jackknife points $(\boldsymbol{P}_{(i)}, T(\boldsymbol{P}_{(i)}))$ for $i = 1, 2, \ldots, n$. Then*

$$var_* T^{LIN}(\boldsymbol{P}^*) = \frac{n-1}{n} \widehat{var}_{jack} \hat{\theta}, \tag{3.8}$$

where $\widehat{var}_{jack}\hat{\theta}$ *is the jackknife estimate of variance for* $\hat{\theta}$*:*

$$var_{jack}\hat{\theta} = \frac{n-1}{n}\sum_{1}^{n}(\hat{\theta}_{(i)} - \hat{\theta}_{(\cdot)})^2 \tag{3.9}$$

and $\hat{\theta}_{(\cdot)} = \sum_1^n \hat{\theta}_{(i)}/n$.

Proof. The definition of T^{LIN} implies a set of n linear equations

$$\hat{\theta}_{(i)} = T^{LIN}(\mathbf{P}_{(i)}) = c_0 + (\mathbf{P}_{(i)} - \mathbf{P}^0)^T\mathbf{U} = c_0 + \mathbf{P}_{(i)}\mathbf{U}$$

The last equality holding because of $\sum_1^n U_i = 0$. Written explicitly:

$$\begin{aligned}
\hat{\theta}_{(1)} &= c_0 + 0 + \frac{U_2}{n-1} + \cdots + \frac{U_n}{n-1} = c_0 - \frac{U_1}{n-1} \\
\hat{\theta}_{(2)} &= c_0 + \frac{U_1}{n-1} + 0 + \cdots + \frac{U_n}{n-1} = c_0 - \frac{U_2}{n-1} \\
&\vdots \\
\hat{\theta}_{(n)} &= c_0 + \frac{U_1}{n-1} + \frac{U_2}{n-1} + \cdots + 0 = c_0 - \frac{U_n}{n-1}
\end{aligned}$$

By summing the n equations and dividing by n we get

$$c_0 = \sum_{i=1}^{n}\frac{\hat{\theta}_{(i)}}{n} = \hat{\theta}_{(\cdot)}.$$

which in turn implies

$$U_i = (n-1)(\hat{\theta}_{(\cdot)} - \hat{\theta}_{(i)}) \quad i = 1, 2, \ldots, n.$$

Finally, using (3.5) and $\sum_{i=1}^n U_i = 0$,

$$\begin{aligned}
var_* T^{LIN}(\mathbf{P}^*) &= \mathbf{U}^T\mathbf{\Sigma}_{\mathbf{P}^*}\mathbf{U} = \mathbf{U}^T\left(\frac{\mathbf{I}}{n} - \frac{\mathbf{P}^0\mathbf{P}^{0^T}}{n}\right)\mathbf{U} \\
&= \frac{1}{n^2}\mathbf{U}^T\mathbf{U} = \frac{n-1}{n}\left(\frac{n-1}{n}\sum_{1}^{n}(\hat{\theta}_{(i)} - \hat{\theta}_{(\cdot)})^2\right)
\end{aligned}$$

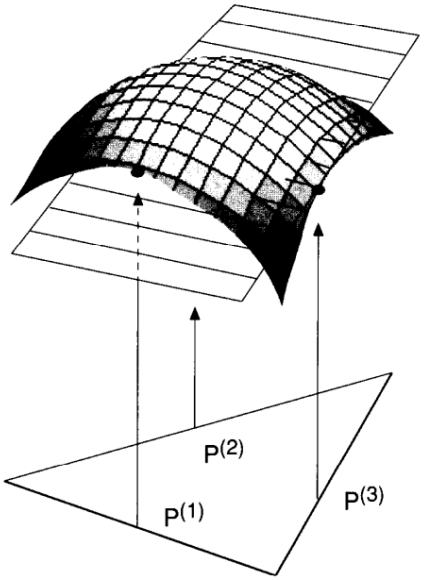


Figure 3.4: *The jackknife plane approximation to the surface $T(\boldsymbol{S}^*)$, resulting in the jackknife estimate of variance.*

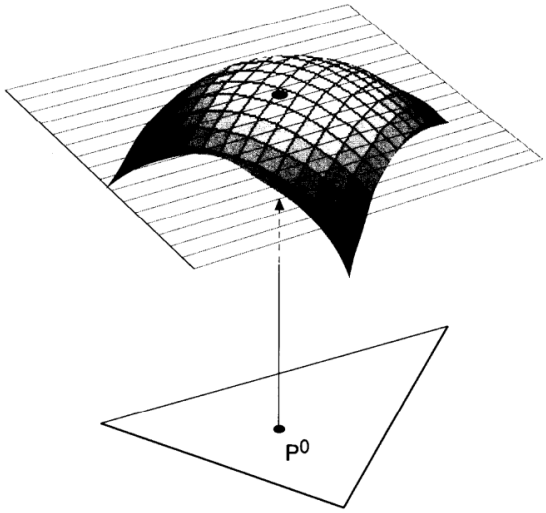


Figure 3.5: *The tangent plane approximation to the surface $T(\boldsymbol{S}^*)$, giving the infinitesimal jackknife estimate of variance.*

□

The "jackknife plane" T^{LIN} is visualized in Figure 3.4. The theorem proved above gives us information about how good of an approximation the jackknife estimate of variance (or standard error) is to the ideal bootstrap estimate: it all depends on how well T^{LIN} serves as an approximate of the surface $T(\mathbf{S})$.

The previous calculations also tell us what the variance of a general linear statistic is under (3.5) is, namely

$$\frac{1}{n^2}\sum_{1}^{n} U_i^2. \tag{3.10}$$

While jackknife estimate of variance uses the hyperplane passing through the jackknife points to approximate $T(\mathbf{S})$, another natural idea would be to approximate using the tangent plane at $T(\mathbf{P}^0)$, as shown in Figure 3.5. This plane has the form

$$T^{TAN}(\mathbf{S}) = T(\mathbf{P}^0) + (\mathbf{S} - \mathbf{P}^0)^T\mathbf{U} \tag{3.11}$$

where $\mathbf{U}$ is defined to be

$$U_i = \lim_{\epsilon \to 0} \frac{T((1-\epsilon)\mathbf{P}^0 + \epsilon(\mathbf{e}_i - \mathbf{P}^0)) - T(\mathbf{P}^0)}{\epsilon}, \quad i = 1, 2, \ldots, n \tag{3.12}$$

and $\mathbf{e}_i = (0, 0, \cdots, 0, 1, 0, \cdots, 0)^T$ is the ith coordinate vector. This tangent approximating plane leads to the *infinitesimal jackknife* estimate of variance

$$var^{IJ}\hat{\theta} = \frac{1}{n^2} \sum_1^n U_i^2. \tag{3.13}$$

The U_i, known as *empirical influence values*, can be thought of as the rate of change in the value of $T(\mathbf{P}^0)$ when shifting an infinitesimal amount of probability onto the ith data value. The empirical influence values will make an appearance again in section 5.5.

The relationship between jackknife and bootstrap estimate of bias can also be expressed using this geometric framework. The approximating statistic will not be linear in this case, as the bootstrap estimates of bias for any linear statistic is

$$E_*[T(\mathbf{P}^*)] - T(\mathbf{P}^0) = c_0 + (E_*[\mathbf{P}^*] - P^0)\mathbf{U} - c_0 = 0. \tag{3.14}$$

We shall therefore consider so called *quadratic statistics*.

Definition 3.2.2. A quadratic statistic is defined to a statistic of the form

$$T^{QUAD}(\mathbf{S}) = c_0 + (\mathbf{S} - \mathbf{P}^0)^T\mathbf{U} + \frac{1}{2}(\mathbf{S} - \mathbf{P}^0)^T\mathbf{V}(\mathbf{S} - \mathbf{P}^0), \tag{3.15}$$

where $\sum_1^n U_i = 0$ and V is an $n \times n$ symmetric matrix satisfying $\forall i, j \ \sum_i V_{ij} = \sum_j V_{ij} = 0$

3.2.2. Theorem. *Let $T^{QUAD}(\boldsymbol{S})$ be a quadratic statistic passing through the center point $(\boldsymbol{P}^0, T(\boldsymbol{P}^0)$ and the jackknife points $(\boldsymbol{P}_i, T(\boldsymbol{P}_{(i)}))$ for $i = 1, 2, \ldots, n$. Then*

$$E_*(T^{QUAD}(\boldsymbol{P}^*) - \hat{\theta}) = \frac{n-1}{n}\widehat{bias}_{jack}(\hat{\theta}). \tag{3.16}$$

Here $\widehat{bias}_{jack}$ is the jackknife estimate of bias for $\hat{\theta}$:

$$\widehat{bias}_{jack} = (n-1)(\hat{\theta}_{(\cdot)} - \hat{\theta}) \tag{3.17}$$

That is, the jacknife estimate of bias for estimate $\hat{\theta} = t(\hat{F})$ is $n/(n-1)$ times the bootstrap estimate of bias for the quadratic approximation T^{QUAD}.

Proof. Because of the points T^{QUAD} must pass through, we have the equations:

$$\begin{aligned} c_0 &= T^{QUAD}(\mathbf{P}^0) = T(\mathbf{P}^0) \\ \hat{\theta}_{(i)} &= c_0 + (\mathbf{P}_{(i)} - \mathbf{P}^0)^T\mathbf{U} + \frac{1}{2}(\mathbf{P}_{(i)} - \mathbf{P}^0)^T\mathbf{V}(\mathbf{P}_{(i)} - \mathbf{P}^0) \end{aligned} \tag{3.18}$$

for $i = 1, 2, \ldots, n$. Using the previous equation, we can express the jackknife estimate of bias as:

$$\begin{aligned} (n-1)(\hat{\theta}_{(\cdot)} - \hat{\theta}) &= (n-1)\sum_1^n \left(\frac{\hat{\theta}_{(i)} - \hat{\theta}}{n}\right) \\ &= \frac{(n-1)}{n}\sum_1^n (\mathbf{P}_{(i)} - \mathbf{P}^0)^T\mathbf{U} + \frac{(n-1)}{2n}\sum_1^n (\mathbf{P}_{(i)} - \mathbf{P}^0)^T\mathbf{V}(\mathbf{P}_{(i)} - \mathbf{P}^0). \end{aligned} \tag{3.19}$$

Considering that $\sum_1^n U_i = 0$ and $\sum_i V_{ij} = \sum_j V_{ij} = 0$, the expression simplifies to

$$\begin{aligned} \frac{(n-1)}{n}\sum_{i=1}^n \mathbf{P}_{(i)}^T\mathbf{U} + \frac{(n-1)}{2n}\sum_{i=1}^n \mathbf{P}_{(i)}^T\mathbf{V}\mathbf{P}_{(i)} &= \frac{(n-1)}{2n}\sum_{i=1}^n\sum_{j=1}^n\sum_{k=1}^n \mathbf{V}_{jk}\mathbf{P}_{(i)_j}\mathbf{P}_{(i)_k} \\ &= \frac{1}{2n}\sum_{i=1}^n\sum_{j=1}^n -\mathbf{V}_{ji}\mathbf{P}_{(i)_j} = \frac{1}{2n(n-1)}\sum_{i=1}^n \mathbf{V}_{ii} \end{aligned} \tag{3.20}$$

Now to expand on the left side of equation (3.16), we must first prove a statement.

3.2.3. Statement. *For a symmetric matrix* $\boldsymbol{A}$*, and a random vector* $\boldsymbol{X}$ *with mean* $\boldsymbol{\mu}$ *and covariance matrix* $\boldsymbol{\Sigma}$

$$E(\mathbf{X}^T\mathbf{A}\mathbf{X}) = \boldsymbol{\mu}^T\mathbf{A}\boldsymbol{\mu} + tr(\boldsymbol{\Sigma}\mathbf{A}) \tag{3.21}$$

Proof.

$$\begin{aligned} E(\mathbf{X}^T\mathbf{A}\mathbf{X}) &= \sum_{i=1}^n\sum_{j=1}^n A_{ij}E(X_iX_j) = \sum_{i=1}^n\sum_{j=1}^n A_{ij}E(X_i)E(X_j) + \sum_{i=1}^n\sum_{j=1}^n A_{ij}\Sigma_{ij} \\ &= \sum_{i=1}^n\sum_{j=1}^n A_{ij}E(X_i)E(X_j) + \sum_{i=1}^n\sum_{j=1}^n \Sigma_{ij}A_{ji} = \boldsymbol{\mu}^T\mathbf{A}\boldsymbol{\mu} + tr(\boldsymbol{\Sigma}\mathbf{A}) \end{aligned} \tag{3.22}$$

□

Using the previous statement, we can expand the left side of equation (3.16)

$$E_*(T^{QUAD}(\mathbf{P}^*)) - T^{QUAD}(\mathbf{P}^0) = c_0 + \frac{1}{2}E_*(\mathbf{P}^{*T}\mathbf{V}\mathbf{P}^*) - c_0$$
$$= \frac{1}{2}tr\left(\left(\frac{\mathbf{I}}{n^2} - \frac{\mathbf{P}^0\mathbf{P}^{0^T}}{n}\right)\mathbf{V}\right) = \frac{1}{2n^2}\sum_{i=1}^{n} V_{ii} \tag{3.23}$$

Finally, comparing equation (3.20) and (3.23) yields

$$E_*(T^{QUAD}(\mathbf{P}^*)) - T^{QUAD}(\mathbf{P}^0) = \frac{n-1}{n}(n-1)(\hat{\theta}_{(\cdot)} - \hat{\theta}). \tag{3.24}$$

□

Similarly as previously, another natural choice for an approximating quadratic statistic is a two term Taylor series around $\mathbf{P}^0$, which will have the form 3.15. Note that the derivatives in the Taylor series expansion is of the empirical influence value variety. The bootstrap estimate of bias for this quadradtic statistic is also of the form $\sum_{i=1}^{n} V_{ii}/2n^2$, and is called the *infinitesimal jackknife estimate of bias* for T.

Chapter 4

Bootstrap for general data structures

4.1 The general bootstrap scheme

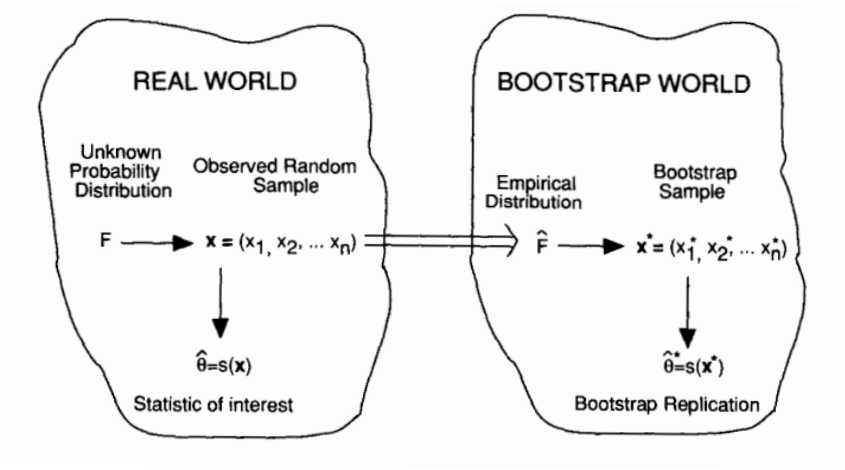


Figure 4.1: *Nonparametric bootstrapping scheme*

So far we have restricted ourselves to the one sample, non parametric setup for bootstrapping. The basic idea behind bootstrapping can, however, be extended to the general case where inference has to be made based on a sample. To show how this can be done, we will review the logic behind the nonparametric bootstrap scheme. Figure (4.1), taken from Efron, Tibshirani [1] summarizes the idea concisely. Given a statistic $s(\mathbf{x})$ of a sample from unknown distribution F, we estimated it's sampling distribution by constructing an estimate of F based on the sample, namely the empirical distribution, denoted here by $\hat{F}$. We then take as many samples as we deem appropriate from this estimated distribution,

and calculate the corresponding bootstrap replications, the distribution of which give us and estimate of the sampling distribution of $s(\mathbf{x})$.

In the general case, shown in Figure (4.2) from Efron, Tibshirani [1], the process is conceptually very similar. The difference is in generality: the data $\mathbf{x}$, given by a probability mechanism $P \rightarrow$, could be something more complicated than a single sample vector; it could be a set of multiple samples, or any data structure in general. The mapping $\Rightarrow$ is conceptually the key step: deciding how we construct the estimated probability mechanism $\hat{P}$ from $\mathbf{x}$. There is no general rule, as in the specific nonparametric case; the statistician must make a decision based on the characteristics of the data. The sampling from $\hat{P}$ is then done according to the same mechanism $\rightarrow$ that yielded our original sample, and the bootstrap replications $\hat{\theta} = s(\mathbf{x})$ are calculated based on these bootstrap samples. The sampling of bootstrap samples and the calculation of bootstrap replications are conceptually simple, but computationally they can be quite cumbersome, especially if the sampling process $\rightarrow$ is complicated or the calculation of $s(\cdot)$ is costly.

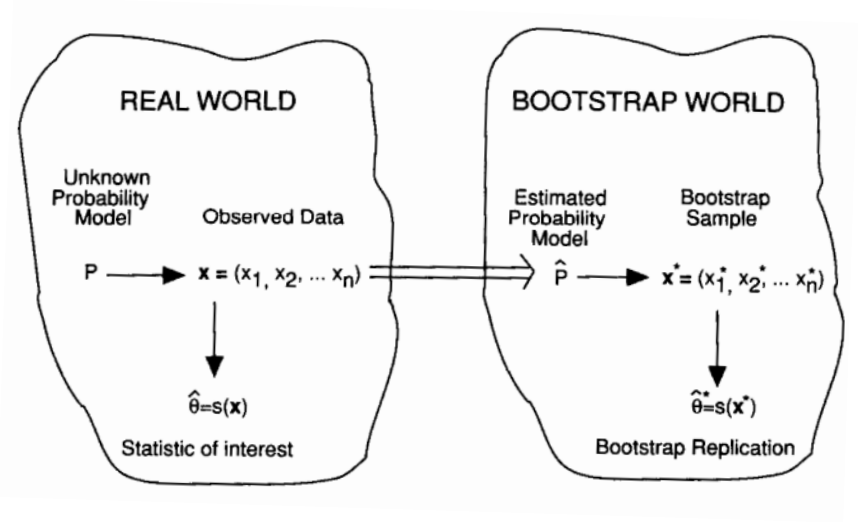


Figure 4.2: *Bootstrapping scheme for general data structures*

As an example, the bootstrap scheme implemented for one sample problems when we are willing and have reason to make parametric assumptions of the unknown distribution, is the so-called *parametric bootstrap*. Here we assume that the sample is from a parametric family of densities $\mathbf{x} \sim f_\theta(\mathbf{x})$. We have an estimate $\hat{\theta}$ of θ, and draw bootstrap samples from the distribution $f_{\hat{\theta}}$.

In the remainder of this chapter we will demonstrate a few distinct ways of applying the bootstrap scheme to time series data.

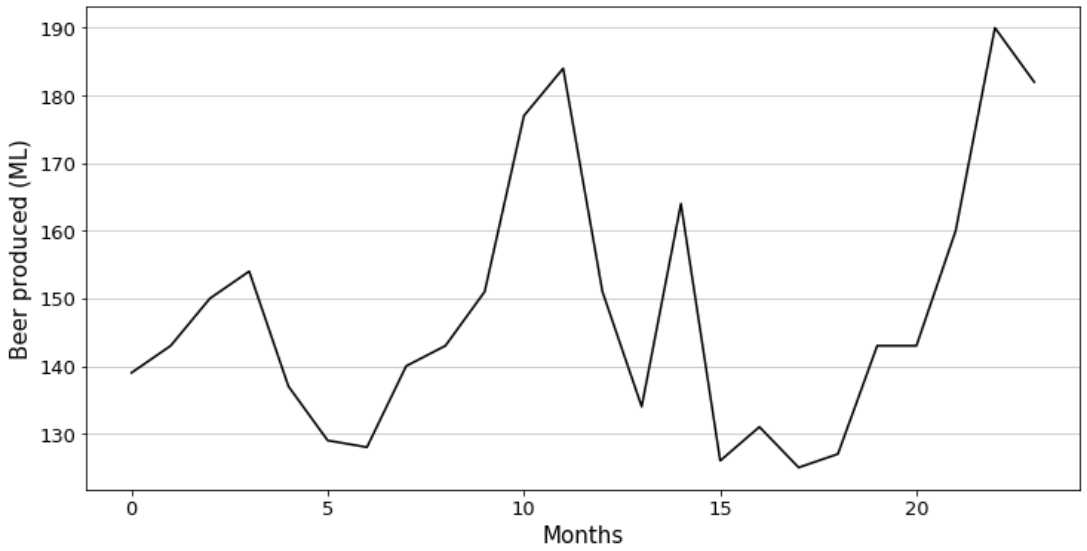


Figure 4.3: *Australian monthly beer production in megalitres, including ale and stout and excluding beverages with alcohol percentage less than 1.15, January 1993 through December 1994 (Australian Bureau of Statistics)*

4.2 Bootstrap scheme for first order autoregression

Figure (4.3) shows Australian monthly beer production in megalitres, including ale and stout and excluding beverages with alcohol percentage less than 1.15, from January 1993 to December 1994, courtesy of the Australian Bureau of Statistics. This dataset is an example of *time series data*: a data set indexed in temporal order. As is the case with many real world time series data, we assume the values obtained are not simply a random sample from some distribution; we assume the contiguous values are related to each other in some manner, that is that the time series data is *short term dependent*.

There are many statistical approaches used to analyze this type of data; we shall begin here with a simple model, a *first order autoregressive scheme*. Let t index the observation times, $t \in \{1, 2, 3, \ldots, 24\}$, and y_t denote the measured quantity at time t. We assume that the time series is *stationary*, that is μ, the expectation of y_t is the same for all times t; this is a reasonable assumption upon viewing the data. Define the centered measurements as

$$z_t = y_t - \mu. \tag{4.1}$$

These centered measurements all have expectation 0. In the first order autoregressive scheme we also assume that z_t is a linear combination of the previous value z_{t-1} and an

independent disturbance term ϵ_t,

$$z_t = \beta z_{t-1} + \epsilon_t \quad \text{for} \quad t = 2, 3, \ldots, 24 \tag{4.2}$$

where β is an unknown parameter, a real number between -1 and 1. The disturbances ϵ_t are assumed to be a sample from an unknown distribution F with expectation equaling 0. Now, if we believe that this model applies to our beer production data, the question is how do we estimate the unknown parameter β based on our data? A simple approach is based the method of least squares. Let us first estimate μ by the sample average $\overline{y}$. We can now calculate the approximately centered measurements

$$\hat{z}_t = y_t - \overline{y}. \tag{4.3}$$

Suppose that b is a guess at the true value of β. The residual square error for this guess is defined to be

$$\mathrm{RSE}(b) = \sum_{t=2}^{24} (z_t - b z_{t-1})^2. \tag{4.4}$$

The estimate given by the method of least squares is the b which minimizes $\mathrm{RSE}(b)$

$$\hat{\beta} = \arg\min_b \mathrm{RSE}(b). \tag{4.5}$$

The beer dataset gives us least squares estimate

$$\hat{\beta} = 0.623. \tag{4.6}$$

We can now implement the general bootstrap scheme to get an idea of how accurate the estimate $\hat{\beta}$ is. The probability mechanism P has three elements, β, μ and F, for short $P = (\beta, \mu, F)$. The data $\mathbf{x}$ consists of the pairs (t, y_t), and the mechanism $P \rightarrow$ that yielded it is described by equations (4.1), (4.2). $\hat{\beta}$ is the statistic of interest, the construction of which is given by (4.5).

Now comes the more difficult part: how do execute $\mathbf{x} \Rightarrow \hat{P}$, that is, how do we estimate the probability mechanism $P = (\beta, \mu, F)$ based on the observed data. The least squares technique already gave us an estimate of β, and we estimated μ with $\overline{y}$ while calculating it. We need now only to estimate F in some manner. The problem is that we do not

directly have access to the disturbance terms

$$F \to (\epsilon_1, \epsilon_2, \ldots, \epsilon_{24}). \tag{4.7}$$

We could, if we knew the value of β, calculate these epsilons using $\epsilon_t = z_t - \beta z_{t-1}$. Not knowing the value of β, we can instead use our estimate $\hat{\beta}$ to calculate the *approximate disturbances*

$$\hat{\epsilon}_t = z_t - \hat{\beta} z_{t-1} \quad \text{for} \quad t = 2, 3, \ldots, 24. \tag{4.8}$$

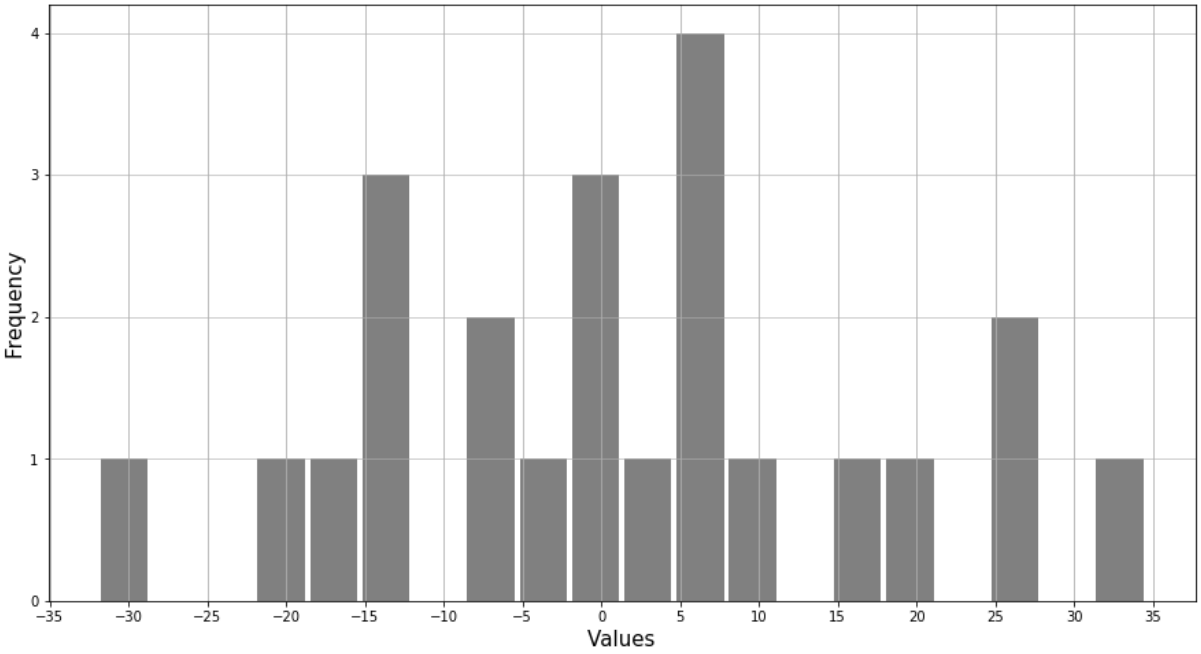


Figure 4.4: *Histogram of approximate disturbances*

Figure (4.4) shows a histogram representing the approximate disturbances. We can now estimate F simply with $\hat{F}$, the empirical distribution of the approximate disturbances.

Having constructed $\hat{P} = (\hat{\beta}, \hat{\mu}, \hat{F})$, now carry out a bootstrap accuracy analysis of $\hat{\beta}$. A bootstrap sample $\mathbf{x}^*$ from $\hat{P}$ is a time series dataset created recursively. We start, for all bootstrap samples, with the initial fixed value $\hat{z}_1^* = y_1 - \overline{y}$. The further data points are generated as follows:

$$z_i^* = \hat{\beta} z_{i-1}^* + \epsilon_i^* \quad for \quad i = 2, \ldots, 24, \tag{4.9}$$

where ϵ_i^* is drawn from $\hat{F}$. With bootstrap sample in hand, we can create a bootstrap replication $\hat{\beta}^*$ for the original least squares esimate $\hat{\beta}$ by calculating the least squares

estimate of β for the bootstrap sample. We created 200 bootstrap replications of β, their distribution is pictured in Figure (4.5). These replications gave us bootstrap estimate of standard error $\hat{se}_{200} = 0.172$.

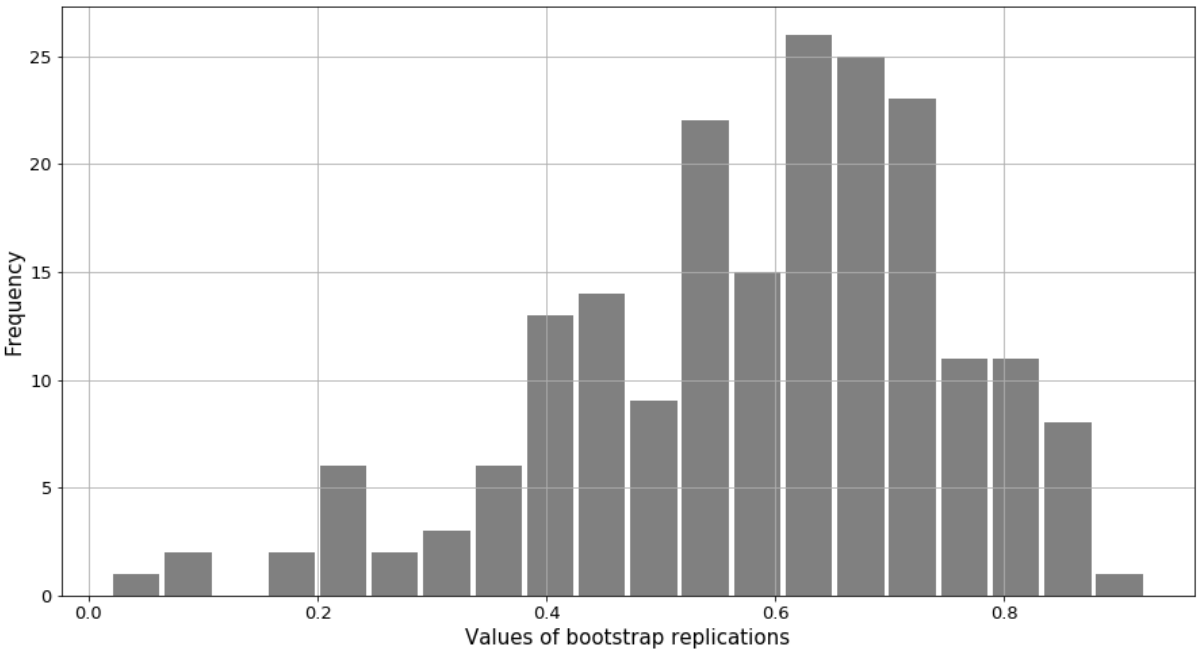


Figure 4.5: *Histogram of boostrap replications of* $\hat{\beta}$

4.3 Bootstrap scheme for second order autoregression

A more refined model than the first order autoregressive scheme is the second order autoregressive scheme, where the dependence of z_t on previous values is extended to also include z_{t-2}. More explicitly, the second order model assumes that

$$z_t = \beta_1 z_{t-1} + \beta_2 z_{t-2} + \epsilon_t \quad for \quad t = 3, 4, \ldots, 24, \tag{4.10}$$

where the ϵ_t-s are independent disturbances as before. Let $\boldsymbol{\beta}$ denote the two dimensional parameter vector $\boldsymbol{\beta} = (\beta_1, \beta_2)^T$. We will use the least squares approach to estimate the parameters, as in the first order case. Let $\mathbf{z}$ denote the vector $(z_3, z_4, \ldots, z_{24})^T$, and $\mathbf{Z}$ be a matrix with two columns, the first being $(z_2, z_3, \ldots, z_{23})$, and the second

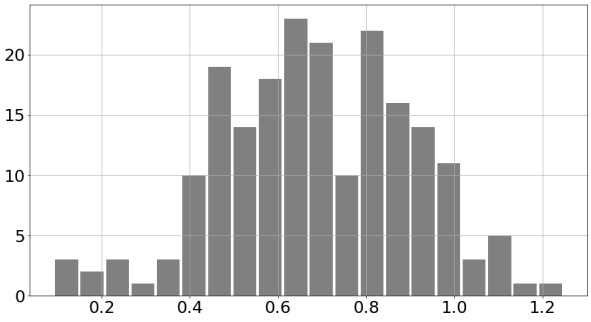


Figure 4.6: $\hat{\beta}_1^*$

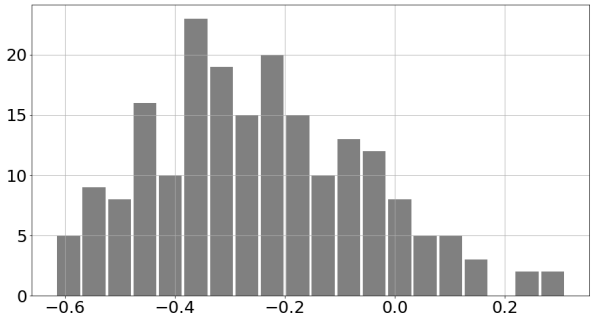


Figure 4.7: $\hat{\beta}_2^*$

$(z_1, z_2, \ldots, z_{22})$. The second order model can then be expressed compactly as

$$\mathbf{z} = \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{4.11}$$

where $\boldsymbol{\epsilon} = (\epsilon_3, \epsilon_4, \ldots, \epsilon_{24})$ is the vector of independent disturbances. Per definition $\hat{\boldsymbol{\beta}}$ is a least squares estimate of $\boldsymbol{\beta}$ if and only if $\|\mathbf{z} - \mathbf{Z}\mathbf{b}\|$ is minimized by $\mathbf{b} = \hat{\boldsymbol{\beta}}$. This is achieved iff. $\mathbf{Z}\hat{\boldsymbol{\beta}}$ is an orthogonal projection of $\mathbf{z}$ onto $\mathrm{Im}(\mathbf{Z})$. This, and some basic linear algebra gives us the following series of equivalences:

$$\hat{\boldsymbol{\beta}} \text{ is a least squares estimate} \Leftrightarrow \mathbf{Z}\hat{\boldsymbol{\beta}} \text{ is an orthogonal projection of } \mathbf{z} \text{ onto } \mathrm{Im}(\mathbf{Z}) \Leftrightarrow$$
$$\mathbf{z} - \mathbf{Z}\hat{\boldsymbol{\beta}} \perp \mathrm{Im}(\mathbf{Z}) \Leftrightarrow \mathbf{z} - \mathbf{Z}\hat{\boldsymbol{\beta}} \in \mathrm{Ker}(\mathbf{Z}^T) \Leftrightarrow \mathbf{Z}^T\mathbf{z} - \mathbf{Z}^T\mathbf{Z}\hat{\boldsymbol{\beta}} = 0 \Leftrightarrow \mathbf{Z}^T\mathbf{Z}\hat{\boldsymbol{\beta}} = \mathbf{Z}^T\mathbf{z}.$$

This implies that if $\mathbf{Z}^T\mathbf{Z}$ is invertable, we can express $\hat{\boldsymbol{\beta}}$ uniquely: $\hat{\boldsymbol{\beta}} = (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{z}$. When not invertable, one can obtain a solution by using the Moore-Penrose pseudoinverse of $\mathbf{Z}^T\mathbf{Z}$. In the case of our dataset, $\mathbf{Z}^T\mathbf{Z}$ happens to be invertable, and unique least squares estimate is $\hat{\boldsymbol{\beta}} = (0.714, -0.206)$. We implemented the bootstrap scheme similarly as before by generating bootstrap samples recursively, to estimate the standard error of $\hat{\boldsymbol{\beta}}$, using 200 replications. The histograms of the first and second coordinates of these replications are pictured in Figures (4.6), (4.7).

The replications gave us bootstrap estimates of standard error $\hat{se}_{200}(\hat{\beta}_1^*) = 0.219$ and $\hat{se}_{200}(\hat{\beta}_2^*) = 0.192$. Notice that according to this estimate of standard error $\hat{\beta}_2$ is almost within one standard error of 0; therefore we cannot confidently state that the extra parameter of the second order model captures any further information in the beer sales dataset as compared to the first order model.

4.4 Moving blocks bootstrap

A different, interesting yet simple method of bootstrapping time series is the *moving blocks bootstrap*. Note that the assumptions we made previously, that the time series is stationary and short term dependent is crucial to this approach as well. Instead of fitting a model to the data and then resampling the residuals, this method uses an approach more similar to the one-sample bootstrap. The method is illustrated in Figure (4.8) from Efron, Tibshirani [1]. We create a set of all blocks of contiguous datapoints with length

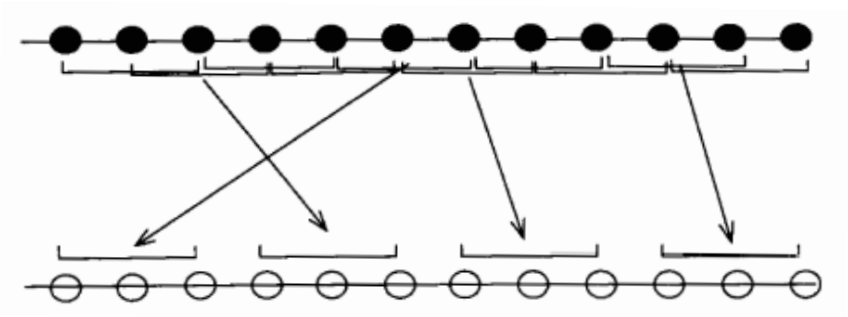


Figure 4.8: *Diagram representing the moving blocks bootstrap. The black circles are the original time sereis data points. The bootstrap sample (white circles) is constructed by selecting a block size (here 3) and sampling with replacement from all contiguous blocks of this size.*

k from the original time series data, then sample with replacement from this set l times, where $n \approx k \cdot l$. We then construct bootstrap replications by pasting these l blocks next to each other. This method refines the crude approach of sampling from the data points with replacement by preserving the relationship between contiguous datapoints in the k sized blocks from the dataset. The trick is to select k to be big enough so that data points more than k far in the original sample can be considered almost independent. There is no standard methodology of selecting the block size, but many compelling ideas can be found in the literature. An advantage of the moving blocks bootstrap is that it is mostly model independent; as long as we have reason to believe that the time series is stationary, and short term dependent, the method can be used validly.

We executed the moving blocks bootstrap on the beer production dataset, with a block length of 4 and 200 replications. We chose the statistic of interest to be the least squares estimate of the parameter in the first order autoregressive scheme. This produced the histogram shown in Figure (4.9), and a bootstrap estimate of standard error $\hat{se}_{200}(\hat{\beta}^*) = 0.182$. Interestingly, the moving blocks bootstrap construction shifted the replications to

have mean 0.391, as opposed to the autoregression based resampling methods, which produced replication distributions with a mean close to the value of the statistic being replicated. This is not surprising, as in autoregressive bootstrapping we used the statistic, $\hat{\beta}$ or $\hat{\boldsymbol{\beta}}$ as the parameter for the construction of the bootstrap samples; in the moving blocks bootstrap no such considerations were made.

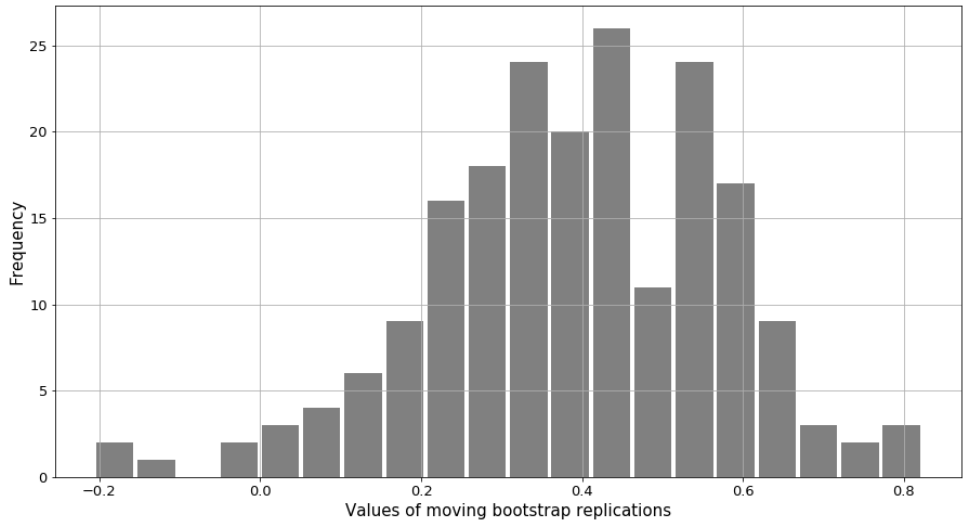
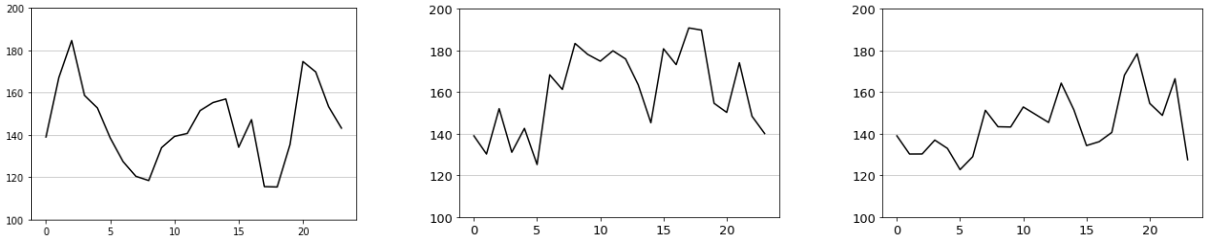
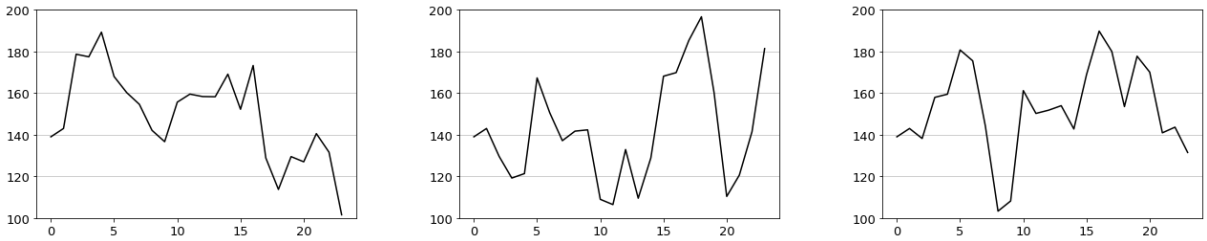


Figure 4.9: *Histogram of bootstrap replications of $\hat{\beta}^*$, using moving blocks bootstrap*

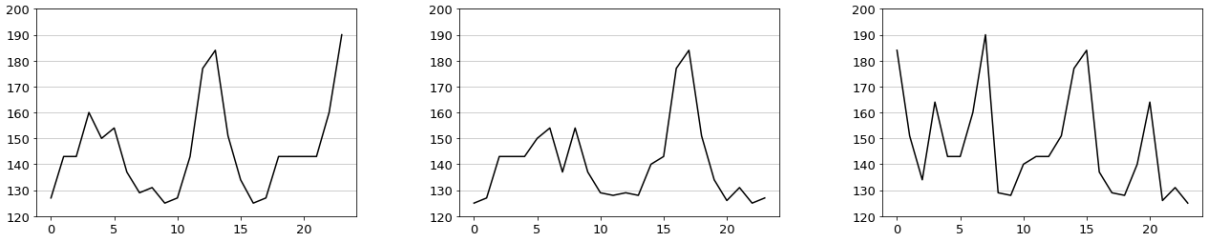
In conclusion, which of these model choices seem the most appropriate? This is hard to answer definitively without conducting further analysis, or using a larger sample. A rough picture can be formed by viewing the line plots of the bootstrap samples. Figure (4.10) shows three-three samples corresponding to each method mentioned in this chapter, as well as the original dataset, and bootstrap samples obtained by non-paramatricly sampling single values with replacement from the beer dataset. As we can see, all of the reviewed methods produce datasets of similar shape to the beer dataset, compared to the simple nonparametric method. Nonparametric bootstrap sampling produces datasets with much higher volitility and no discernable relationship between contiguous datapoints, as we would expect.



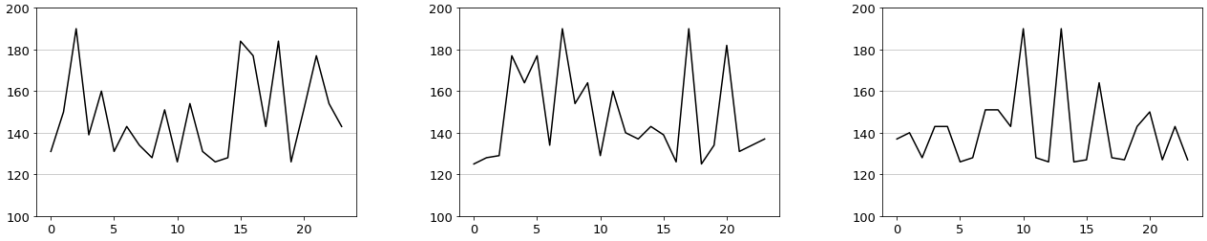
(a) *Bootstrap samples obtained using first order autoregressive construction*



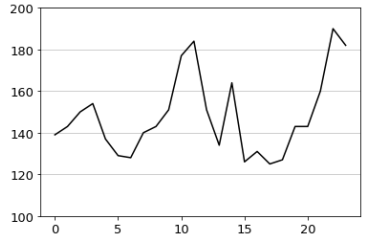
(b) *Bootstrap samples obtained using second order autoregressive construction*



(c) *Bootstrap samples obtained using moving blocks construction*



(d) *Bootstrap samples obtained by sampling with replacement from data values*



(e) *Original times series data, in the same scale*

Figure 4.10: *Comparison of bootstrap samples obtained via the methods reviewed in this chapter. The line plots of these samples are seperated into groups of three. For reference, there is also a group of samples obtained using simple sampling with replacement (d) as well as the original beer production dataset (e).*

Chapter 5

Bootstrap confidence intervals

5.1 Standard confidence interval

Confidence intervals are a useful concept within the context of parameter estimation. While an estimate gives only a single value to estimate the parameter, a confidence interval gives more information about the possible values of the parameter. An α level or $\alpha \cdot 100\%$ confidence interval is an interval constructed based on a sample in such a way that the parameter lies inside it with probability α. To clarify, when a confidence interval is constructed using the *realized values* of our sample, chance has no more part to play; the parameter is either inside the confidence interval or not. Probability comes into play earlier, when we draw the sample from the underlying distribution.

Exact confidence intervals, where the probability is exactly α, can be constructed in quite a few specific problems, but often one must make do with approximate confidence intervals. In this case, even while operating under the assumed probability model of the sample, the probability of the parameter of the parameter lying in the approximate confidence intervals is approximately, but not exactly α.

Before getting into confidence intervals constructed using bootstrapping techniques, we will first give a brief review of the 90% *standard confidence interval*

$$\hat{\theta} \pm 1.656 \cdot \hat{se}. \tag{5.1}$$

Suppose that we are in the one-sample situation $F \to \mathbf{x}$, attempting to estimate a parameter θ of distribution F using the estimator $\hat{\theta}$. Let $\hat{se}$ be an estimate of $\hat{\theta}$'s standard error, possibly obtain using bootstrapping. The standard interval operates on the as-

sumption that as $n \to \infty$ the distribution of $\hat{\theta}$ converges to $N(\theta, \hat{se})$, denoted shortly by $\hat{\theta} \dot{\sim} N(\theta, \hat{se}^2)$. This can be stated equivalently as

$$\frac{\hat{\theta} - \theta}{\hat{se}} \dot{\sim} N(0, 1). \tag{5.2}$$

This assumtion is a strong one, though many common estimators do in fact have a normal limiting distribution similar to $N(\theta, \hat{se})$ owing to the central limit theorem and the asymptotic distribution of maximum likelihood estimators. Let z^α indicate the α-th quantile of the $N(0, 1)$ distribution. If we take (5.2) to hold exactly, then

$$\mathrm{P}(z^\alpha \leq \frac{\hat{\theta} - \theta}{\hat{se}} \leq z^{(1-\alpha)}) = 1 - 2\alpha. \tag{5.3}$$

This can be reorganized as

$$\mathrm{P}(\theta \in [\hat{\theta} - z^{(1-\alpha)} \cdot \hat{se}, \hat{\theta} - z^\alpha \cdot \hat{se}]) = 1 - 2\alpha. \tag{5.4}$$

$z^\alpha = -z^{(1-\alpha)}$ as the standard normal distribution is symmetric around 0; therefore, if we take (5.2) to hold exactly, the confidence interval

$$\hat{\theta} \pm z^\alpha \cdot \hat{se} \tag{5.5}$$

has coverage probability $1 - 2\alpha$. Using the above formula with $\alpha = 0.5$ yields the 90% standard interval, as $z^\alpha = 1.645$. Note that a more accurate name would be an approximate standard confidence interval, as the interval is exact only if we take the asymptotic assumtion (5.2) to hold exactly for small sample sizes as well.

5.2 Bootstrap percentile interval

The the standard interval was constructed based on the percentiles of the normal distribution. This idea can be used in tandem with bootstrap resampling to produce the so called *bootstrap percentile interval*. Instead of the normal distribution, the percentile interval is constructed using the distribution of bootstrap replications. Let $\hat{\theta}^{*(\alpha)}$ denote the α quantile of $\hat{\theta}^*$'s distribution. The $1-2\alpha$ level ideal percentile interval is then defined

to be

$$[\hat{\theta}^{*(\alpha)}, \hat{\theta}^{*(1-\alpha)}]. \tag{5.6}$$

We can approximate this ideal percentile interval by first drawing B bootstrap samples $\mathbf{x}^{*1}, \mathbf{x}^{*2}, \ldots, \mathbf{x}^{*B}$, and generating B bootstrap replications $\hat{\theta}^*(b) = s(\mathbf{x}^{*b})$. We then approximate $\hat{\theta}^{*(\alpha)}$ using the $B \cdot \alpha$-th value in the ordered list of these replications, which we will mark as $\hat{\theta}_B^{*(\alpha)}$. When $B \cdot \alpha$ is not an integer, we round it to an integer using some convention. The approximate $1 - 2\alpha$ percentile interval is

$$[\hat{\theta}_B^{*(\alpha)}, \hat{\theta}_B^{*(1-\alpha)}]. \tag{5.7}$$

As a rule of thumb, the recommended number of bootstrap replications for the construction of confidence intervals is aboout 2000. It is natural that bootstrap confidence intervals require more samples than standard error estimates, as confidence intervals requires an accurate estimation of the tails of a distributions, which cannot be achieved using as few samples as 200.

For most common estimators, thanks to the central limit theorem and properties of ML estimates, as $n \to \infty$ the bootstrap histogram will become normal shaped, often quite similar to the PDF of $N(\theta, \hat{se})$. For smaller sample sizes, however, the histogram may look non-normal, and therefore the standard and percentile interval will differ. In these cases the percentile interval can have an advantage, which we will become clear shortly.

In some statistical problems, one can find a perfectly normalizing and variance stabilizing transformation, that is a monotone increasing mapping $m(\cdot)$, for which $\hat{\phi} = m(\hat{\theta})$ has distribution

$$\hat{\phi} \sim N(\phi, c^2), \tag{5.8}$$

where $\phi = m(\theta)$, and c is some non-zero constant. Some examples of this are $m(\cdot) = log(\cdot)$ for a log-normal distribution, or the Fischer-transformation for the sample correlation coefficient of two random variables. When there is in fact such a transformation, the percentile interval has the following good property.

5.2.1. Statement. *Suppose the transformation $\hat{\phi} = m(\hat{\theta})$ perfectly normalizes the distribution of $\hat{\theta}$:*

$$\hat{\phi} \sim N(\phi, c^2). \tag{5.9}$$

The ideal percentile percentile interval based on $\hat{\theta}$ is $[m^{-1}(\hat{\phi} - z^{(1-\alpha)}c), m^{-1}(\hat{\phi} - z^{(\alpha)}c)]$

Proof. The proof is quite straightforward; because the mapping $m(\cdot)$ is monotone increasing, it preserves the standard ordering on $\mathbb{R}$, which means the percentiles simply map to each other through $m(\cdot)$. □

This result implies that when there exists some normalizing transformation, the bootstrap percentile interval will produce an accurate interval by, in the background, calculating the percentiles of the normalized distribution, and transforming it back to our original scale. This shows the advantage of the percentile interval over the standard interval. The standard interval only gives correct results when the distribution of $\hat{\theta}$ is normal, while the percentile interval will produce correct results for any distribution where a perfectly normalizing transformation exists, without requiring us to actually know the transformation.

5.3 Bias corrected and accelerated bootstrap confidence interval

We saw in the previous section that percentile interval can be viewed as an improvement over of the standard interval, in the sense that in order to be correct, the standard interval requires the distribution of $\hat{\theta}$ to be $N(\theta, \hat{se}^2)$, while the percentile interval more laxly also allows for the distribution of $\hat{\psi} = m(\hat{\theta})$ to be $N(\theta, c^2)$ with constant c through the transformation $m(\cdot)$. The bias corrected and accelerated (abbreviated as BC_a) bootstrap confidence interval is in an improvement of the percentile interval.

The BC_a interval improves upon the percentile interval in two respects. Firstly, it allows the transformation $m(\cdot)$ to give $\hat{\psi} = m(\hat{\theta})$ a normal distribution which is biased for ψ. The predecessor of the BC_a interval, the simpler *bias corrected* (BC) interval, gave only this allowance. More explicitly the BC interval assumes that

$$(\hat{\psi} - \psi)/c \sim N(-z_0, 1) \tag{5.10}$$

where c is a non-zero variance constant, and z_0 is a bias constant.

The BC_a confidence interval pushes even further; it allows for the variance of the normalized distribution to change as a linear function of ψ. The assumption of the BC_a interval is that

$$(\hat{\psi} - \psi)/c \sim N(-z_0\sigma_\psi, \sigma_\psi^2), \qquad \sigma_\psi = 1 + a\psi, \tag{5.11}$$

where c is some non-zero constant, z_0 is a bias constant, and a is an acceleration constant.

The shed some light on why the acceleration constant is so called, note that for any fixed parameter value of ψ_0

$$\sigma_\phi = \sigma_{\phi_0} + a(\phi - \phi_0). \tag{5.12}$$

The value of a gives the rate at which the standard deviation changes while the parameter ϕ shifts from any fixed value.

These improvements made by the BC_a over the standard and percentile interval result in better asymptotic coverage properties, which are stated and proven for the one-parameter and multiparameter case in Efron (1987) and Efron (1996), respectively. The BC_a interval approach is not the only way to improve the coverage properties of the percentile interval. A commonly used and conceptually simpler, but more computationally intensive approach is *confidence point calibration*. For a description of this method see Chapter 18 of Efron, Tibshirani [1].

Returning to the BC_a: now that we have introduced these extra factors to the normalized distribution, the question is, how do we construct a confidence interval? We will first discuss the simple one parameter situation for calculating the BC_a interval, and the more general multiparameter situation later.

Beginning with the one parameter situation, let's assume that our sample $\mathbf{x}$ is from a probability distribution f_θ parameterized by a one dimensional real parameter θ. The statistic $\hat{\theta}$ estimates θ. We use parametric bootstrap resampling to produce bootstrap replications, as discussed in Chapter 3, that is

$$\hat{\theta}^* \sim f_{\hat{\theta}}. \tag{5.13}$$

We will denote the cumulative distribution function of the parametric bootstrap distribution as $\hat{G}(s)$

$$\hat{G}(s) = \int_{-\infty}^{s} f_{\hat{\theta}}\, d\lambda = \mathrm{P}_{\hat{\theta}}(\hat{\theta}^* < s). \tag{5.14}$$

Now let's assume that there is a monotone increasing transformation m and constants z_0 and a so that

$$\hat{\psi} = m(\hat{\theta}), \quad \psi = m(\theta) \tag{5.15}$$

satisfy

$$\hat{\psi} = \psi + \sigma_\psi(Z - z_0), \quad Z \sim N(0,1) \tag{5.16}$$

where $\sigma_\psi = 1 + a\psi$. This is simply a reformulation of (5.11), with the exception that $c = 1$.

We can assume that $c = 1$ without any loss of generality, as if we have $\sigma_\theta = c(1 + A\psi)$, so that $\sigma_0 = c$ ($c \neq 1$), then the transformation $\hat{\psi}' := \hat{\psi}/c$ gives $\hat{\psi}' = \psi' + \sigma'_{\psi'}(Z - z_0)$, where $\sigma'_{\psi'} = 1 + a\psi'$ and $a = Ac$. We also assume that $\psi > -1/a$ if $a > 0$ so that $\sigma_\psi > 0$, and likewise $\psi < -1/a$ if $a < 0$; this assumption is not very limiting for most real use cases, for more details see section 4 of Efron 1987 [2]. Now, under these assumptions, we can state a result.

5.3.1. Lemma. *Under the preceding assumptions, an exact confidence interval of level* $1 - 2\alpha$ *for* θ *is*

$$\theta \in [\hat{G}^{-1}(\Phi(z[\alpha])), \hat{G}^{-1}(\Phi(z[1-\alpha]))], \tag{5.17}$$

where Φ *is the cumulative distribution function of the standard normal distribution, and*

$$z[\alpha] = z_0 + \frac{(z_0 + z^{(\alpha)})}{1 - a(z_0 + z^{(\alpha)})}. \tag{5.18}$$

Proof. Since

$$\hat{\psi} = \psi + (1 + a\psi)(Z - z_0) \tag{5.19}$$

we have

$$1 + a\hat{\psi} = 1 + a[\psi + (1 + a\psi)(Z - z_0)] = (1 + a\psi)(1 + a(Z - z_0)) \tag{5.20}$$

Taking the logarithm of this equation gives us the nice form

$$\hat{\xi} = \xi + W \tag{5.21}$$

where $\hat{\xi} = log(1 + a\hat{\psi})$, $\xi = log(1 + a\psi)$, and $W = log(1 + a(Z - z_0))$. Note that here we disregarded the fact that the arguments of the logarithms may have been negative. The lemma still holds for this case; section 4 and 8 of Efron 1982 [4] gives a more careful discussion where negativity is also considered.

The form of (5.21) gives a simple way of constructing a $1 - 2\alpha$ an interval for ξ:

$$\xi \in [\hat{\xi} - w^{(1-\alpha)}, \hat{\xi} - w^{(\alpha)}] \tag{5.22}$$

where $w^{(\alpha)}$ is the $100 \cdot \alpha$ percentile point of W. It is convenient to introduce a new notation here. Let $\theta[\alpha]$ indicate the endpoint of an α level one sided confidence interval for parameter θ that is unbounded on the left side. (5.22) then says that $\xi[\alpha] = \hat{\xi} - w^{(1-\alpha)}$,

and $\xi[1-\alpha] = \hat{\xi} - w^{(\alpha)}$. Resuming the proof, we can transform the interval back using the transformation $t(x) = (e^x - 1)/a$. This transformation gives us $\hat{\psi} = (e^{\hat{\xi}} - 1)/a$, $\psi = (e^{\xi} - 1)/a$, and $Z - z_0 = (e^W - 1)/a$.

$$\begin{aligned}\phi[\alpha] = \frac{(e^{\xi[\alpha]} - 1)}{a} &= \frac{(e^{\hat{\xi}}/e^{w^{(1-\alpha)}} - 1)}{a} = \frac{e^{\hat{\xi}} - e^{w^{(1-\alpha)}}}{e^{w^{(1-\alpha)}} a} \\ &= \frac{1 + a\hat{\phi} - (1 + a(z^{(1-\alpha)} - z_0))}{a(1 + a(z^{(1-\alpha)} - z_0))} = \frac{\hat{\phi} + (z^{(\alpha)} + z_0)}{1 - a(z^{(\alpha)} + z_0)} \\ &= \hat{\phi} + \frac{\hat{\phi} + z^{(\alpha)} + z_0 - \hat{\phi}(1 - a(z^{(\alpha)} + z_0))}{1 - a(z^{(\alpha)} + z_0)} = \hat{\phi} + \sigma_{\hat{\phi}} \frac{z_0 + z^{(\alpha)}}{1 - a(z_0 + z^{(\alpha)})}, \quad (5.23)\end{aligned}$$

which, in summary, leaves us with

$$\phi[\alpha] = \hat{\phi} + \sigma_{\hat{\phi}} \frac{z_0 + z^{(\alpha)}}{1 - a(z_0 + z^{(\alpha)})}. \quad (5.24)$$

The CDF of $\hat{\phi}$ according to (5.16) is

$$H(s) = \Phi\left(\frac{s - \phi}{\sigma_\phi} + z_0\right), \quad (5.25)$$

which implies that the CDF of bootstrap replication $\hat{\psi}^*$ is

$$\hat{H}(s) = \Phi\left(\frac{s - \hat{\phi}}{\sigma_{\hat{\phi}}} + z_0\right). \quad (5.26)$$

The inverse of $\hat{H}(s)$ is therefore $\hat{H}^{-1}(\alpha) = \hat{\phi} + \sigma_{\hat{\psi}}(\Phi^{-1}(\alpha) - z_0)$. Now, if we take a look at our previous result (5.24) and the definition of $z[\alpha]$ (5.18), we can see that $\hat{H}^{-1}(\Phi(z[\alpha])) = \phi[\alpha]$. This is similar to the form of the lemma we want to prove, all we have to do now is transform back to the scale of θ.

Thankfully, the BC_a interval transforms in the natural way, that is given the monotone increasing transformation m, $\hat{\phi} = m(\hat{\theta})$, $\phi = m(\theta)$, then $\phi[\alpha] = m(\theta[\alpha])$. This follow from the fact that $\hat{H}(m(s)) = \hat{G}(s)$, or equivalently $\hat{H}^{-1}(s) = m(\hat{G}^{-1}(s))$:

$$\hat{H}(m(s)) = \mathrm{P}(\hat{\phi} < m(s)) = \mathrm{P}(m(\hat{\theta}) < m(s)) = \mathrm{P}(\hat{\theta} < s) = \hat{G}(s). \quad (5.27)$$

Thus, we have finally proven the lemma; the transformations $\hat{\theta} \to \hat{\phi} \to \hat{\xi}$ give a translation form where we can easily construct the interval, and transforming back this interval gives

us the BC_a interval points stated in the lemmma. □

The confidence interval points in Lemma 5.3.1 are defined to be the ideal $1-2\alpha$ level BC_a confidence interval points.

How do we approximate these ideal points? The bootstrap distribution's cumulative distribution function can be easily approximated by the empirical CDF of B bootstrap replications; the two constants prove more difficult, as we do not know the normalizing transformation they are associated with. It turns out, however, that they too can be approximated well using the bootstrap replication distribution. The approximation of z_0 is rather straightforward:

$$z_0 \doteq \Phi^{-1}(\hat{G}(\hat{\theta})). \tag{5.28}$$

To prove (5.28) note that because of (5.15), (5.16) and our assumption directly after it that $\sigma_\theta > 0$, the following equalities hold for every value of θ

$$\mathrm{P}_\theta(\hat{\theta} < \theta) = \mathrm{P}_\phi(\hat{\phi} < \phi) = \mathrm{P}(Z < z_0) = \Phi(z_0). \tag{5.29}$$

Substituting $\theta = \hat{\theta}$ into this equation gives us $\hat{G}(\hat{\theta}) = P_{\hat{\theta}}(\hat{\theta}^*) = \Phi(z_0)$, which is equivalent to (5.28).

Approximating the acceleration constant a is more complicated, so will omit a detailed discussion here. Efron 1987 [2] connects the role of a to transformation theory discussed in Efron 1982 [4] and shows that a good approximation of a is

$$a \doteq \frac{1}{6} SKEW_{\theta=\hat{\theta}}(\dot{l}_\theta). \tag{5.30}$$

Here $SKEW_{\theta=\hat{\theta}}(X)$ denotes the skewness random variable X, $\mu_3(X)/\mu_2(X)^{3/2}$ at parameter point $\theta = \hat{\theta}$, and $\dot{l}_\theta$ is the score function of the family of densities f_θ,

$$\dot{l}_\theta = \frac{\delta}{\delta\theta} log f_\theta. \tag{5.31}$$

We now have methods of approximating a and z_0 without knowing transformation m, based on the bootstrap distribution $\hat{G}$; therefore we can now compute approximate bootstrap BC_a intervals using Lemma 5.3.1 with the approximate values of a and z_0, and approximating the distribution $\hat{G}$ using the empirical CDF of B bootstrap replications (the recommended number of replications B typically being around 2000).

5.4 Parametric BC_a for multiparameter families

So far we have discussed only the simple case, when our data $\mathbf{x}$ is from a one parameter family. We will now discuss the case of multiparameter families, which will somewhat strangely lead us to be able to apply the BC_a interval to non-parametric situations as well.

Let's assume that our sample $\mathbf{x}$ comes from a parametric family $\mathcal{F}$ of density functions $f_{\boldsymbol{\eta}}$, where $\boldsymbol{\eta}$ is a vector of parameters, based on which we are trying to construct a confidence interval for the real valued parameter $\theta = t(\boldsymbol{\eta})$. In this section we also assume that $\hat{\theta}$ is a maximum likelihood estimator (MLE); from $\mathbf{x}$ we calculate $\hat{\boldsymbol{\eta}}$, the MLE of $\boldsymbol{\eta}$, and $\hat{\theta} = t(\hat{\boldsymbol{\eta}})$, the MLE of θ. Parametric bootstrap sampling is done by

$$f_{\hat{\boldsymbol{\eta}}} \to \mathbf{x}^*. \tag{5.32}$$

We calculate the MLE of $\boldsymbol{\eta}$ based on bootstrap sample $\mathbf{x}^*$, which gives us a bootstrap replication $\hat{\boldsymbol{\eta}}^*$, which in turn gives us the bootstrap replication of θ, $\hat{\theta}^* = t(\hat{\boldsymbol{\eta}}^*)$. The distribution of $\hat{\theta}^*$ is the parametric bootstrap distribution of $\hat{\theta}$, which gives us the bootstrap CDF

$$\hat{G}(s) = P_{\hat{\boldsymbol{\eta}}}(\hat{\theta}^* < s). \tag{5.33}$$

The bias-correction constant z_0 is equal to $\Phi^{-1}(\hat{G}(\hat{\theta}))$ as previously, and can be approximated in the same way. The accelerating constant a is more complicated, as our previous approximation for one parameter families $a \doteq \frac{1}{6}SKEW_{\theta=\hat{\theta}}(\dot{l}_\theta)$ cannot be used outright in the multiparameter case. We can, however, still make use of this approximation, by replacing the multiparameter family $\mathcal{F}$ with a special one parameter family $\hat{\mathcal{F}}$ contained within $\mathcal{F}$: the *least favorable* one parameter family.

Let $\dot{\mathbf{l}}_{\boldsymbol{\eta}}$ denote the vector with ith coordinate $\delta/\delta\eta_i \log f_{\boldsymbol{\eta}}(\mathbf{x})$, then by definition of the MLE $\dot{\mathbf{l}}_{\hat{\boldsymbol{\eta}}}(\mathbf{x}) = 0$, and let $\ddot{\mathbf{I}}_{\hat{\boldsymbol{\eta}}}(\mathbf{x})$ be the $k \times k$ matrix with ijth coordinate of $\delta^2/(\delta\eta_i\delta\eta_j) \log f_{\boldsymbol{\eta}}(\mathbf{x})|_{\boldsymbol{\eta}=\hat{\boldsymbol{\eta}}}$. Furthermore, let $\hat{\boldsymbol{\Delta}}$ be the gradient vector of $\theta = t(\boldsymbol{\eta})$ evaluated at $\hat{\boldsymbol{\eta}}$, that is $\hat{\Delta}_i = (\delta/\delta\eta_i)t(\boldsymbol{\eta})|_{\boldsymbol{\eta}=\hat{\boldsymbol{\eta}}}$. The *least favorable direction* at $\boldsymbol{\eta} = \hat{\boldsymbol{\eta}}$ is defined to be

$$\hat{\boldsymbol{\mu}} := (-\ddot{\mathbf{I}}_{\hat{\boldsymbol{\eta}}})^{-1}\hat{\boldsymbol{\Delta}}. \tag{5.34}$$

The least favorable family $\hat{\mathcal{F}}$ is the one-parameter subfamily of $\mathcal{F}$ passing through $\hat{\boldsymbol{\eta}}$ in

the direction $\hat{\boldsymbol{\mu}}$,

$$\hat{\mathcal{F}} = \{f_\lambda := f_{\hat{\boldsymbol{\eta}}+\lambda\hat{\boldsymbol{\eta}}}\}, \tag{5.35}$$

We will give some intuition about the meaning behind the name of the least favorable family. Let's say that we want to estimate $\theta(\lambda) := t(\hat{\boldsymbol{\eta}}+\lambda\hat{\boldsymbol{\mu}})$ based on observation $f_\lambda \to \mathbf{y}$. The following statement is then true.

5.4.1. Statement. *The Fischer information bound for an unbiased estimate of λ in the one parameter least favorable family, evaluated at $\lambda = 0$, is $\hat{\boldsymbol{\Delta}}^T(-\ddot{\boldsymbol{I}}_{\hat{\boldsymbol{\eta}}})^{-1}\hat{\boldsymbol{\Delta}}$, which is the same as the corresponding bound for estimating $\theta = t(\boldsymbol{\eta})$, at $\boldsymbol{\eta} = \hat{\boldsymbol{\eta}}$, in the multiparameter family $\mathcal{F}$. Furthermore, any other one-dimensional subfamily passing though $\hat{\boldsymbol{\eta}}$ has Fischer information at least as great as this.*

For brevity's sake we will omit the proof of this statement, as it is relatively straightforward calculation.

Statement (5.4.1) says that in the Fischer information lower bound sense, the least favorable family is the one-dimensional family passing though $\hat{\boldsymbol{\eta}}$ for which inference for $\lambda = 0$ is the most difficult, as difficult as inference for $\hat{\theta} = t(\hat{\boldsymbol{\eta}})$ in the multiparameter family $\mathcal{F}$. It makes sense heuristically, to select this one dimensional family $\hat{\mathcal{F}}$ to represent the whole multiparameter family $\mathcal{F}$ at the parameter point $\hat{\boldsymbol{\eta}}$ when calculating our approximation of the acceleration constant a.

The approximation of a based on family $\hat{\mathcal{F}}$ is

$$a \doteq \frac{1}{6}\text{SKEW}_{\lambda=0}\left(\frac{\delta}{\delta\lambda}\log f_{\hat{\boldsymbol{\eta}}+\lambda\hat{\boldsymbol{\mu}}}\right) \tag{5.36}$$

This formula simplifies nicely in the exponential family case, because of the natural relation between cumulants and the log-partition function of exponential families. We will look at a specific exponential family formula, when the densities of family $\mathcal{F}$ are of the form

$$f_{\boldsymbol{\eta}}(\mathbf{y}) = e^{n[\boldsymbol{\eta}^T\mathbf{y}-\psi(\boldsymbol{\eta})]}f_0(\mathbf{y}). \tag{5.37}$$

We use this form for our calculations because when discussing the non-parametric BC_a we will deal with an exponential family of this form.

5.4.2. Lemma. *For an exponential family of form (5.37), formula (5.36) gives us*

$$a \doteq \frac{1}{6\sqrt{n}}\frac{\hat{\psi}^{(3)}(0)}{(\hat{\psi}^{(2)}(0))^{3/2}} \tag{5.38}$$

where

$$\hat{\psi}^{(j)}(0) = \left.\frac{\delta^j \psi(\hat{\boldsymbol{\eta}} + \lambda\hat{\boldsymbol{\mu}})}{\delta\lambda^j}\right|_{\lambda=0} \tag{5.39}$$

.

Proof. To use formula (5.36) we calculate

$$\left.\frac{\delta \log f_{\hat{\boldsymbol{\eta}}+\lambda\hat{\boldsymbol{\mu}}}(\mathbf{y}^*)}{\delta\lambda}\right|_{\lambda=0} = n\hat{\boldsymbol{\mu}}^T(\mathbf{y}^* - \hat{\psi}(\hat{\boldsymbol{\eta}})^T). \tag{5.40}$$

Therefore, $\mathrm{SKEW}_{\lambda=0}\left(\frac{\delta}{\delta\lambda}\log f_{\hat{\boldsymbol{\eta}}+\lambda\hat{\boldsymbol{\mu}}}(\mathbf{y}^*)\right)$ equal the skewness of $\hat{\boldsymbol{\mu}}^T\mathbf{y}^*$ where $\mathbf{y}^* \sim f_{\hat{\boldsymbol{\eta}}}$. To calculate the skewness of $\hat{\boldsymbol{\mu}}^T\mathbf{y}$, we will express it's moment generating function, and take the logarithm to obtain the cumulant generating function.

$$\begin{aligned} M_{\hat{\boldsymbol{\mu}}^T\mathbf{y}^*}(t) = E(e^{t\hat{\boldsymbol{\mu}}^T\mathbf{y}^*}) &= \int_{\mathbf{y}} \left(\exp(t\hat{\boldsymbol{\mu}}^T\mathbf{y})\, \exp(n[\hat{\boldsymbol{\eta}}^T\mathbf{y} - \psi(\hat{\boldsymbol{\eta}})])\, f_0(\mathbf{y})\right) \\ &= \int_{\mathbf{y}} \left(\exp(n[(\hat{\boldsymbol{\eta}} + \tfrac{t}{n}\hat{\boldsymbol{\mu}})^T\mathbf{y} - \psi(\hat{\boldsymbol{\eta}})])f_0(\mathbf{y})\right) = \exp(n[\psi(\hat{\boldsymbol{\eta}} + \tfrac{t}{n}\hat{\boldsymbol{\mu}}) - \psi(\hat{\boldsymbol{\eta}})]). \end{aligned} \tag{5.41}$$

The cumulant generating function of $\hat{\boldsymbol{\mu}}^T\mathbf{y}^*$ is then $K(t) = n[\psi(\hat{\boldsymbol{\eta}} + \frac{t}{n}\hat{\boldsymbol{\mu}}) - \psi(\hat{\boldsymbol{\eta}})]$; therefore, the kth cumulant is equal to

$$\mu_k = \left.\frac{\delta^k}{\delta t^k}K(t)\right|_{t=0} = n\ \psi^{(k)}(\hat{\boldsymbol{\eta}})\left(\frac{\hat{\boldsymbol{\mu}}}{n}\right)^k. \tag{5.42}$$

The skewness of $\hat{\boldsymbol{\mu}}^T\mathbf{y}^*$ is defined to be

$$\frac{\mu_3}{(\mu_2)^{3/2}} = \frac{n\ \psi^{(3)}(\hat{\boldsymbol{\eta}})\left(\frac{\hat{\boldsymbol{\mu}}}{n}\right)^3}{\left(n\ \psi^{(2)}(\hat{\boldsymbol{\eta}})\left(\frac{\hat{\boldsymbol{\mu}}}{n}\right)^2\right)^{3/2}} = \frac{\hat{\psi}^{(3)}(0)}{\sqrt{n}(\hat{\psi}^{(2)}(0))^{3/2}} \tag{5.43}$$

which proves the statement. □

5.5 Non-parametric BC_a

In this section we will apply the results relating to multiparameter families to construct a non-parametric BC_a confidence interval. We are in the non-parametric situation where we have a sample $\mathbf{y} = (x_1, x_2, \ldots, x_n)$ obtained for probability distribution F. $\theta = t(F)$

is a real-valued parameter for which we wish construct a confidence interval based on the plug-in estimate $\hat{\theta} = t(\hat{F})$, where $\hat{F}$ is the empirical distribution.

We can resample from $\hat{F}$ to produce bootstrap replications $\hat{\theta}^*$, allowing us to estimate the bootstrap CDF of $\hat{\theta}^*$,

$$\hat{G}(s) = P_{\hat{F}}(\hat{\theta}^* < s). \tag{5.44}$$

The bias correction constant z_0 equals $\Phi^{-1}(\hat{G}(\hat{\theta}))$, as in (5.28).

To gain an estimate of the acceleration constant, we will consider the multiparameter family of distributions supported on the observed sample values,

$$F(\mathbf{w}) : \text{probability mass of } w_i \text{ on } x_i, \tag{5.45}$$

where $\mathbf{w}$ is a vector in the simplex S_n. We consider the sample $\mathbf{y}$ fixed, the parameter being $\mathbf{w} \in S_n$. This choice of this family is in line with the non-parametric approach to inference discussed in previous chapters. The parameter θ can be formulated as $\theta(\mathbf{w}) := t(F(\mathbf{w}))$. We will denote the central point of the simplex as

$$\mathbf{w}^0 := (1/n, 1/n, \ldots, 1/n), \tag{5.46}$$

which corresponds to $F(\mathbf{w}^0) = \hat{F}$, and the plug-in estimate $\theta(\mathbf{w}^0) = \hat{\theta} = t(\hat{F})$.

The underlying assumption behind the non-parametric BC_a is then that we drew a hypothetical sample $x_1^*, x_2^*, \ldots, x_n^*$ from $F(\mathbf{w})$, which happened to equal our actual sample $x_1, x_2, \ldots, x_n$. The vector of proportions $P_i = \#\{x_j^* = x_i\}/n$ is a sufficient statistic for $\mathbf{w}$, and has distribution

$$\mathbf{P} = (P_1, P_2, \ldots, P_n) \;\sim\; \text{mult}_n(n, \mathbf{w})/n, \quad \mathbf{w} \in S_n. \tag{5.47}$$

As $\mathbf{P}$ is sufficient, we we can use it instead of the hypothetical sample for the construction of a confidence interval. Our observed sample in terms of $\mathbf{P}$ was then $\mathbf{P} = \mathbf{w}^0$.

The distributions (5.47) have PMF

$$f_{\mathbf{w}}(\mathbf{P}) = \frac{n!}{\prod_{i=1}^{n}(nP_i)!} \prod_{i=1}^{n} w_i^{nP_i}, \tag{5.48}$$

and thus form an n-parameter exponential family of form (5.37) with $\mathbf{y} = \mathbf{P}$, $\eta_i =$

$\log(nw_i)$, $\psi(\eta) = \log(\sum_1^n \exp(\eta_i)/n)$, and $f_0(\mathbf{P}) = n!/(\prod_1^n (nP_i)!\ n^n)$.

This means that we can use Lemma (5.4.2) to calculate an approximation to the acceleration constant.

5.5.1. Statement. *The approximation of the acceleration constant a for the non-parametric BC_a confidence interval is*

$$a \doteq \frac{1}{6}\left[\sum_{i=1}^{n} U_i^3 \Big/ \left(\sum_{i=1}^{n} U_i^2\right)^{3/2}\right] \tag{5.49}$$

where U_i is the ith empirical influence value defined in equation (3.12).

Proof. The proof consists thinking through these of four steps.

1. The MLE of $\boldsymbol{\eta}$ is $\hat{\boldsymbol{\eta}} = 0$.

2. The least favorable direction passing through $\hat{\boldsymbol{\eta}}$ is $\hat{\boldsymbol{\mu}} = \mathbf{U}$.

3. The least favorable family passing through $\hat{\boldsymbol{\eta}} = 0$ is

$$\mathbf{P}^* \sim \text{mult}(n, \mathbf{w}^\lambda)/n, \quad w_i^\lambda = \exp(\lambda U_i) \Big/ \sum_{j=1}^{n} \exp(\lambda U_j) \tag{5.50}$$

4. Finally, the statement follows from Lemma (5.4.2) by differentiating $\hat{\psi}(\lambda) = \log(\sum_1^n \exp(\lambda U_i)/n)$ and using the fact that $\sum_1^n U_i = 0$.

Steps 1., 3., 4. are straightforward, only step 2. warrants further explanation. As a consequence of it's definition, $\mathbf{U}$ is essentially the gradient of $\theta(\mathbf{w})$ at $\mathbf{w}^0$; therefore, it is proportional to $\hat{\boldsymbol{\Delta}}$ in equation (5.34). Furthermore, $-\ddot{\mathbf{I}}_{\hat{\boldsymbol{\eta}}} = \mathbf{1}\mathbf{1}^T/n$, which has pseudo inverse $\mathbf{I}$. Thus $\hat{\boldsymbol{\mu}}$ is proportional to $\mathbf{U}$, which in effect means $\hat{\boldsymbol{\mu}} = \mathbf{U}$ as multiplying the direction by a scalar results in the same least favorable family and approximation of a. □

Having derived the approximations for z_0 and a, the construction of the BC_a happens in the same manner as described in equations (5.17) and (5.18), using these approximations.

Bibliography

[1] Efron, B. and Tibshirani, R. *An Introduction to the Bootstrap.* Chapman and Hall, New York, 1993.

[2] Efron, B. *Better bootstrap confidence intervals.* Journal of the American Statistical Association, Vol. 82, No. 397, 171-185, 1987.

[3] Efron, B. *Bootstrap methods: another look at the jackknife.* The Annals of Statistics, Vol. 7, 1-26, 1979.

[4] Efron, B. *Transformation theory: how normal is a family of distributions?* The Annals of Statistics, Vol. 10., No. 2. 323-339, 1982.

[5] Tibshirani, R. J. and Efron, B. *Bootstrap confidence intervals.* Statistical Science, Vol. 11, No. 3, 189-228, 1996.

NYILATKOZAT

Név:

ELTE Természettudományi Kar, szak:

NEPTUN azonosító:

Szakdolgozat címe:

A **szakdolgozat** szerzőjeként fegyelmi felelősségem tudatában kijelentem, hogy a dolgozatom önálló szellemi alkotásom, abban a hivatkozások és idézések standard szabályait következetesen alkalmaztam, mások által írt részeket a megfelelő idézés nélkül nem használtam fel.

Budapest, 20

_____

a hallgató aláírása