Chapter 8
Bootstrapping

In the early days of statistics, calculations had to be done by hand or by unwieldy mainframe computers with very limited memory and computational power. Consequently, statisticians were permanently looking for ways to simplify calculations through approximation. Time consuming calculations, e.g. computation of quantiles and distribution functions, were performed once at high precision and the results were published in tables. Even today, tables for distributions, such as the standard normal or the t-distribution, are available in statistics textbooks. The generation of random numbers was also not a trivial task and thick books containing sequences of random numbers were published, e.g. for the purpose of survey sampling.

The rise in readily available computing power led to a corresponding increase in the range of techniques available to statisticians. In particular, in complex data analytic situations, where the derivation of asymptotic formulae is too cumbersome or where their validity or robustness may be questionable, these new resources could be exploited with a range of new statistical methods built upon sampling and resampling. Early papers by Quenouille (1956) and Tukey (1958) introduced the jackknife estimator, a predecessor to the nowadays more popular bootstrap and subsampling procedures. The bootstrap as we know it today was introduced in the seminal paper written by Efron (1979). The general idea is to use the actual sample to calculate the properties of statistic of interest. The idea is simple, but very powerful and extends readily to complex data analyses and will be explored throughout this chapter.

8.1 Nonparametric Bootstrap

8.1.1 Motivation

To introduce the basic idea of the nonparametric bootstrap, we begin with an independent and identically distributed sample of size n from a distribution function $F(.)$, i.e. we assume that $Y_1, \ldots, Y_n$ such that

$$Y_i \sim F(y) \quad i.i.d.. \tag{8.1.1}$$

Usually, $F(.)$ is unknown, at least up to some parameter θ. We will keep things as general as possible here and postulate a generic function $F(.)$ that is neither necessarily indexed by a parameter θ, nor defined by a particular distributional model, such as normality. Therefore, we address this approach as nonparametric bootstrap, i.e. we do not use any assumptions about $F(.)$. It proceeds as follows. First, we have our observed sample $y_i, i = 1, \ldots, n$, which we denote as $y = (y_1, \ldots, y_n)$. The bootstrap algorithm is given by

1. Calculate a statistic of interest, $t(y)$, e.g. the median.
2. Take n draws *with replacement* from the original data $y = (y_1, \ldots, y_n)$ to create a new dataset of the same size and call it $y^* = (y_1^*, \ldots, y_n^*)$. The sample y^* is called a bootstrap sample. Calculate the statistic of interest, $t(y^*)$. Store the result.
3. Repeat the previous step B times with indices $b = 1 \ldots, B$, such that y^{*b} is the b-th bootstrap sample from which $t(y^{*b})$ is calculated and stored.
4. Use the B bootstrap samples $t(y^{*1}), \ldots t(y^{*B})$ to obtain information about the statistical variation of $t(y)$.

The bootstrap process is sketched in Fig. 8.1. The collection of bootstrapped statistics $t(y^{*b})$, $b = 1, \ldots, B$ can be used to estimate properties of the statistic

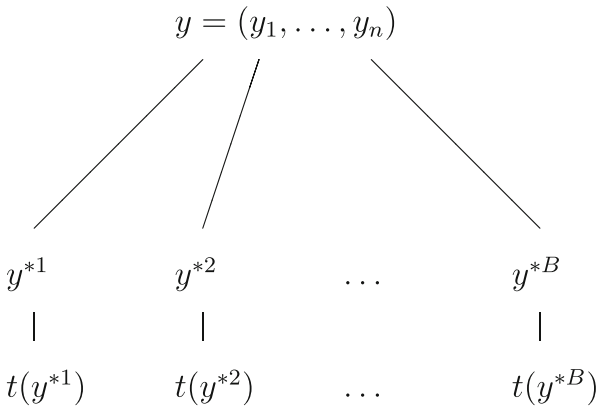

Fig. 8.1 Illustration of the bootstrap procedure

of interest, e.g. its variance. For instance, an estimate of the unknown variance $\mathrm{Var}_F(t(Y))$ is given by its bootstrap estimate

$$\mathrm{Var}_{\mathrm{Boot}}(t(Y)) \approx \frac{1}{B-1}\sum_{b=1}^{B}\left(t(y^{*b}) - \bar{t}_{\mathrm{Boot}}\right)^2 , \tag{8.1.2}$$

where

$$\bar{t}_{\mathrm{Boot}} = \frac{1}{B}\sum_{b=1}^{B} t(y^{*b}) .$$

Let us demonstrate the bootstrap idea with a simple example, shown in Fig. 8.2. In the top left are 50 simulated observations drawn from a standard normal

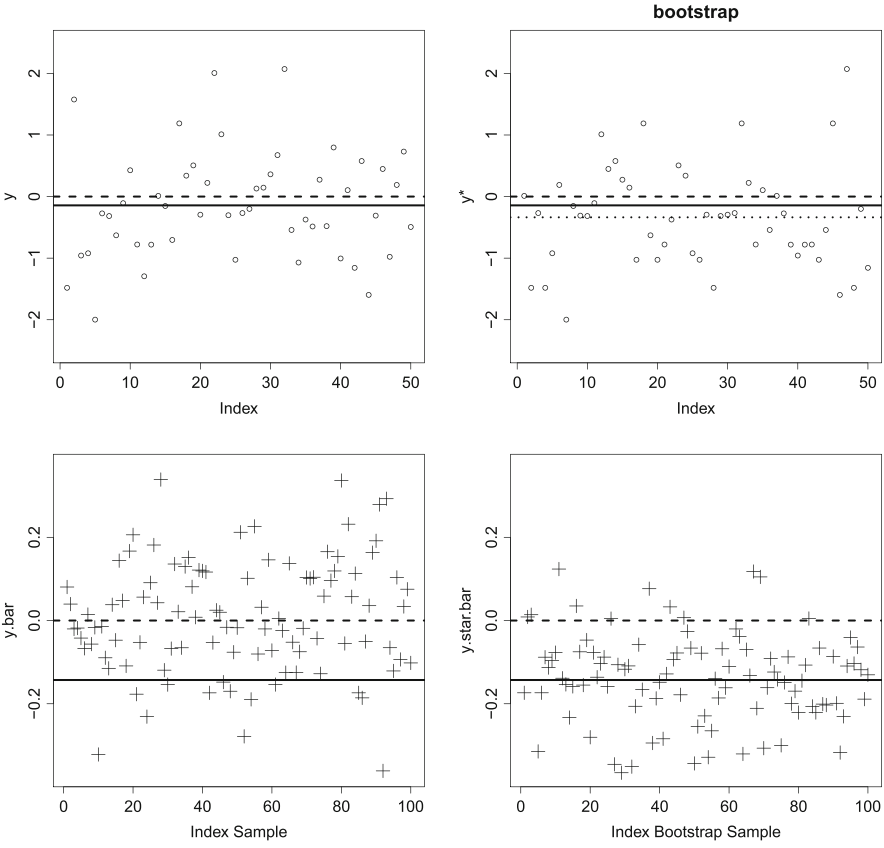

Fig. 8.2 Simulated data (top left) and bootstrapped data (top right). Simulated $\bar{y}$ (bottom left) and bootstrapped $\bar{y}^*$ (bottom right)

distribution, with the arithmetic mean of the 50 data-points represented by the solid line and the true mean by the dashed line. From the 50 observations $y_i, i = 1, \ldots, 50$, we draw with replacement a single bootstrap sample of 50 observations $y_i^*, i = 1, \ldots, 50$ shown in the top right. The arithmetic mean of this bootstrap sample $\bar{y}^*$ is represented by the dotted line. We repeat this step $B = 100$ times and calculate the arithmetic mean for each bootstrap sample, which we denote with $\bar{y}^{*b}$ for $b = 1, \ldots, B$. These values are shown in the bottom right plot, with the solid line representing the arithmetic mean of the bootstrapped means and the dashed line representing the true mean. Note that the mean of the original sample is slightly less than the true mean, and thus, the fact that the mean of the bootstrapped means is also less than the true mean is to be expected. The bootstrap aims to mimic the true variation, which we show in the bottom left plot. Here we simulated 100 times $Y_1, \ldots, Y_{50}$ from a $N(0, 1)$ distribution and calculated the arithmetic mean. Clearly, this is only possible if we know the true distribution, while the bootstrap makes use of only the available data and can therefore be applied without knowledge of the true distribution function. The distribution of the two sets of points appears similar, just with the mean shifted for the bootstrap samples.

It is important to note that we make two approximations above. Firstly, the bootstrap estimate $\mathrm{Var}_{\mathrm{Boot}}(t(Y))$ is only an approximation (we could also call it an estimate) of the true variance $\mathrm{Var}_F(t(Y))$ of the statistic of interest which clearly depends on the true but unknown distribution function $F(.)$. Secondly, we only use a limited number of bootstrap samples, i.e. B is finite, which itself only gives an approximation of the true or *ideal* bootstrap distribution. This is the distribution of the statistic of interest using all possible bootstrap samples.

8.1.2 Empirical Distribution Function and the Plug-In Principle

In this section, let us motivate the nonparametric bootstrap a little more formally, which will hopefully also help us to understand why bootstrapping works in general. Let $Y = (Y_1, \ldots, Y_n)$ with $Y_i \sim F(.)$ $i.i.d.$ from an unknown distribution function $F(.)$ and let $y = (y_1, \ldots, y_n)$ be the observed data. For simplicity, we also assume that all values of y are unique, i.e. there are no two indices i and j for which $y_i = y_j$. The empirical distribution function $\hat{F}_n(y)$ is given by

$$\hat{F}_n(y) = \frac{1}{n} \sum_{i=1}^{n} 1(y_i \leq y).$$

Let us start our explanation with the **plug-in principle**. To do this, we need to explain the concept that any statistic $t(Y)$ can be written as functional. This means that the behaviour of $t(Y)$ depends on the (unknown) distribution function $F()$. For example, assume we are interested in the variance of $t(Y)$. This variance depends

on the function $F()$. We denote this functional dependence in mathematical terms as a statistical functional which we motivate later. The concept of a functional is now combined with the plug-in principle as follows: Whenever the true distribution function F is involved in a statistical functional, it should be replaced by its empirical analogue $\hat{F}_n$. For the sake of simplicity, we avoid mathematical details and instead motivate the idea with an example.

Example 30 For example, let us consider the expectation of the random variable Y. The expectation of Y depends on the unknown distribution function. Altogether, this can be expressed as

$$T(F) = \mu = \int y dF(y).$$

The notation $\int y dF(y)$ refers, as previously mentioned, to an integral if $F(.)$ is both continuous and differentiable and a sum if $F(.)$ is discrete valued. Note that if $F(.)$ is continuous and differentiable we can rewrite the functional as

$$T(F) = \int y dF(y) = \int y \frac{dF(y)}{dy} dy = \int y f(y) dy$$

with $f(.)$ as the density function. If, in contrast, Y is discrete valued with $Y \in \{a_1 < a_2 < a_3, \ldots\}$, then $F(.)$ is a step function and the functional is given by

$$T(F) = \int y dF(y) = \sum_k a_k P(Y = a_k) = \sum_k a_k \{F(a_k) - F(a_{k-1})\},$$

where $F(a_0) = 0$.

Given that we do not know F, the plug-in principle states that we can replace F with its empirical approximation $\hat{F}_n$ to obtain $T(\hat{F}_n)$ as the statistic that estimates $T(F)$. For example, for the mean this gives

$$T(\hat{F}_n) = \int y \mathrm{d}\hat{F}_n(y) = \sum_{i=1}^{n} y_{(i)} \{\hat{F}_n(y_{(i)}) - \hat{F}_n(y_{(i-1)}))\} = \frac{1}{n} \sum_{i=1}^{n} y_i = \bar{y},$$

where $y_{(1)}, \ldots, y_{(n)}$ is the ordered sample. The plug-in principle is rather flexible and can be broadly applied to many other statistical quantities. It also allows us to calculate further properties of our estimates. If, for example, we want to calculate the variance of $\bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$, then

$$T(F) = \mathrm{Var}(\bar{Y}) = \int (\bar{y} - \mu)^2 dF(y) = \frac{1}{n} \sigma^2,$$

where $\sigma^2 = Var(Y_i), i = 1, \ldots, n$. Taking the sample $y_1, \ldots, y_n$ we can again apply the plug-in principle which gives

$$T(\hat{F}_n) = \int (\bar{y}^* - \bar{y})^2 d\hat{F}_n(y^*),$$

where $\bar{y}^* = (\bar{y}_1^*, \ldots, \bar{y}_n^*)$ is a sample drawn from $\hat{F}_n(.)$. A little bit of calculation shows that

$$T(\hat{F}_n) = \frac{1}{n}(\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2),$$

where the term in the inner brackets is the empirical variance of the sample. ▷

The plug-in principle is used when no further information about F is available, excluding the information contained in the sample y. As the explicit calculation of $T(\hat{F}_n)$ can be clumsy, the bootstrap approach is to draw a sample instead. If there are no repeated values in the original sample, it can be shown that the number of different bootstrap samples is $\binom{2n-1}{n}$. This number becomes quite large, even for small sample sizes. For example, $n = 15$ gives 77,558,760 possible unique samples. This explains why it is more practical, instead of using all samples, to randomly select a small number B, e.g. $B = 200$. Note that the samples are drawn from $\hat{F}_n()$ and need to be $i.i.d.$. This is achieved if we draw *with* replacement from our observed data $y_1, \ldots, y_n$. In other words, drawing n samples with replacement from $y_1, \ldots, y_n$ is equivalent to

$$y_i^* \sim \hat{F}_n(),\ i.i.d. \text{ for } i = 1, \ldots, n.$$

Example 31 Consider a toy example with the following ($n = 3$) sample: $y = (20, 25, 40)$. There are 10 possible bootstrap samples y^* when sampling with replacement:

(20, 20, 20); (25, 25, 25); (40, 40, 40); (20, 20, 25); (20, 20, 40);

(25, 25, 40); (20, 25, 25); (20, 40, 40); (25, 40, 40); (20, 25, 40).

It is important to note that the probabilities of the samples are not equal. While (20,20,20) occurs only once, (20,20,25) results from the three possible samples (20, 20, 25), (20, 25, 20), and (25, 20, 20) and is therefore more likely, but is listed only once above. For very small n, the multinomial distribution can be used to

calculate bootstrap estimates based on the ideal bootstrap distribution. Using the fact that we sample with replacement, we get

$$P\left(y^* = (20, 20, 20)\right) = \frac{3!}{3!0!0!}\left(\frac{1}{3}\right)^3\left(\frac{1}{3}\right)^0\left(\frac{1}{3}\right)^0 = \frac{1}{27}$$

$$P\left(y^* = (20, 20, 25)\right) = \frac{3!}{2!1!0!}\left(\frac{1}{3}\right)^2\left(\frac{1}{3}\right)^1\left(\frac{1}{3}\right)^0 = \frac{3}{27}$$

$$P\left(y^* = (20, 25, 40)\right) = \frac{3!}{1!1!1!}\left(\frac{1}{3}\right)^1\left(\frac{1}{3}\right)^1\left(\frac{1}{3}\right)^1 = \frac{6}{27}$$

with, for example, $P\,(y^* = (20, 20, 25)) = P\,(y^* = (25, 25, 40))$, etc. Let us focus on the median as the statistic of interest. The median of the original sample is 25. To get a bootstrap estimate of the variance of the median, we need to calculate the median of each of the above bootstrap samples, compute the arithmetic mean of all medians, and finally compute the variance using (8.1.2). Instead of enumerating all samples, we can use the multinomial probabilities to simplify the computation. In the end, this gives

$$\bar{t}_{\text{Boot}} = \frac{1}{27}(20 + 25 + 40) + \frac{3}{27}(20 + 20 + 25 + 25 + 40 + 40) + \frac{6}{27} \cdot 25 = 27.59259$$

$$\text{Var}_{\text{Boot}}(t\,(y)) = \frac{1}{27}\left((20 - \bar{t}_{\text{Boot}})^2 + \ldots + 6 \cdot (25 - \bar{t}_{\text{Boot}})^2\right) = 58.09328\ .$$

▷

To further motivate the plug-in principle, we consider two "worlds": the "real world" and the "bootstrap world". In the real world, the sample y is drawn from the unknown population distribution $F(.)$ and we calculate the statistic $t\,(y)$. In the bootstrap world, the bootstrap samples y^{*b} are drawn from the known empirical distribution $\hat{F}_n(.)$ and the statistic $t\,(y^{*b})$ is calculated B times. Therefore, we are trying to gain information about $F(.)$ and the properties of the statistic $t\,(.)$ through repeated sampling from the empirical distribution of the data $\hat{F}_n(.)$, which then gives an empirical distribution of our statistic. At first glance, it may appear unreasonable that re-use of the original sample y allows us to gain further information, which we were not able to extract from y alone. However, theoretical and empirical results show that the bootstrap works well with various kinds of data. This is a direct result of the plug-in principle. The bootstrap procedure as introduced above is also called the *nonparametric bootstrap*, as no parametric distributional assumptions are made. Later, we will also introduce the *parametric bootstrap* in Sect. 8.2, which streamlines the process when parametric assumptions are appropriate.

8.1.3 Bootstrap Estimate of a Standard Error

We have already introduced the procedure to estimate a variance using bootstrap samples and the estimation of a standard error follows the same principle. We now focus on a parameter of interest, ξ, which we estimate with $\hat{\xi} = t(y)$. The task is then to find a bootstrap approximation of the standard error of $\hat{\xi}$, which we calculate using B estimates $\hat{\xi}^{*b} = t(y^{*b})$ derived from the bootstrap samples. The bootstrap now works as follows

1. Generate B bootstrap samples $y^{*1}, \ldots, y^{*B}$.
2. Calculate $\hat{\xi}^{*b}$, $b = 1, \ldots, B$.
3. Estimate the standard error $\mathrm{se}_F(\hat{\xi}) = \sqrt{\mathrm{Var}_F(\hat{\xi})}$ with

$$\widehat{\mathrm{se}}_B = \left\{ \frac{1}{B-1} \sum_{b=1}^{B} \left[\hat{\xi}^{*b} - \bar{\hat{\xi}}^{*} \right]^2 \right\}^{\frac{1}{2}} \quad \text{with} \quad \bar{\hat{\xi}}^{*} = \frac{1}{B} \sum_{b=1}^{B} \hat{\xi}^{*b}.$$

Consequently, the bootstrap estimate for the standard error of an estimator $\hat{\xi}$ (with data from F) is the standard deviation of the bootstrap estimates $\hat{\xi}^{*b}$. These estimates are based on random samples y^{*b}, drawn with replacement from $\hat{F}_n$. It can be shown that

$$\lim_{B \to \infty} \widehat{\mathrm{se}}_B = \mathrm{se}_{\hat{F}_n}(\hat{\xi}^{*}), \tag{8.1.3}$$

where $\mathrm{se}_{\hat{F}_n}(\hat{\xi}^{*})$ is the estimate given all possible samples from the empirical distribution. The approximation $\widehat{\mathrm{se}}_B$ is often called the *nonparametric bootstrap estimate* of the standard error. The number of bootstrap samples B is, in general, governed by practical considerations. On the one hand, if the computation of the estimate $\hat{\xi}$ is complex and time consuming, this lends itself to a lower number of samples. On the other hand, the bootstrap variance estimate itself has its own variance, which is higher when B is small (due to the random process of drawing bootstrap samples). Therefore, B should be high enough to get a stable estimate of the variance. Usually, $B = 200$ is used in practice for variance estimates, but higher values of B may be useful, especially when one wants to estimate the bias of an estimator or construct confidence intervals.

Example 32 We consider a slightly more complex situation in which we have data that is possibly drawn from two different distributions and we are interested in the difference between the two distributions, e.g. in the difference of their mean values. To begin, let

$$\left.\begin{array}{ll} Y_1, \ldots, Y_n & \sim F \quad i.i.d. \\ Z_1, \ldots, Z_m & \sim G \quad i.i.d. \end{array}\right\} \text{ independent.}$$

Our target parameter is the standard error of the difference $\xi = \mathrm{E}(Y_i) - \mathrm{E}(Z_i) = \mu_Y - \mu_Z$, which we estimate with $\hat{\xi} = \bar{y} - \bar{z}$. For the b-th bootstrap sample

$$\begin{aligned} y^{*b} &= (y_1^{*b}, \ldots, y_n^{*b}) \text{ with replacement from } \hat{F}_n \\ z^{*b} &= (z_1^{*b}, \ldots, z_m^{*b}) \text{ with replacement from } \hat{G}_m . \end{aligned} \tag{8.1.4}$$

The bootstrap estimate is then:

$$\widehat{\mathrm{se}}_B = \left\{ \frac{1}{B-1} \sum_{b=1}^{B} \left[\hat{\xi}^{*b} - \bar{\hat{\xi}}^{*} \right]^2 \right\}^{\frac{1}{2}}$$

with

$$\hat{\xi}^{*b} = \bar{y}^{*b} - \bar{z}^{*b} = \frac{1}{n} \sum_{i=1}^{n} y_i^{*b} - \frac{1}{m} \sum_{i=1}^{m} z_i^{*b}$$

and

$$\bar{\hat{\xi}}^{*} = \frac{1}{B} \sum_{b=1}^{B} (\bar{y}^{*b} - \bar{z}^{*b}) = \frac{1}{B} \sum_{b=1}^{B} \hat{\xi}^{*b} .$$

▷

The variance of the mean difference can also be estimated by standard methods using the sampling variances. However, the variance of the ratio or other nonlinear functions of the two means cannot be easily calculated by standard methods, while the bootstrap estimate can be directly transformed to this case.

8.1.4 Bootstrap Estimate of a Bias

As demonstrated above, bootstrapping allows us to estimate the variance of a statistic $t(y)$ or more specifically an estimate $\hat{\xi}$. As we learned in Chap. 3, the Mean Squared Error (MSE) of an estimator is given by the variance plus the squared bias. This raises the question of whether it is possible to determine bias of an estimate with bootstrapping. The answer is yes, which we will now demonstrate. Let our samples again be $Y_1, \ldots, Y_n \sim F$ *i.i.d.* from an unknown distribution function F and ξ be some parameter of interest. We can express the parameter with the functional $\xi = T(F)$. The plug-in principle suggests that the corresponding estimate is $\hat{\xi} = T(\hat{F}_n)$. The bias of the estimator $\hat{\xi}$ is given by

$$\mathrm{bias}_F(\hat{\xi}, \xi) = \mathrm{E}_F(\hat{\xi}) - \xi = \mathrm{E}_F(\hat{\xi}) - T(F) . \tag{8.1.5}$$

It is clear that the bias cannot be estimated from a single sample y directly, as it depends upon the unknown parameter ξ. Instead, we want to calculate its bootstrap estimate by following the plug-in principle and substituting unknown components in (8.1.5) with their empirical counterparts. To do so, we substitute F with $\hat{F}_n$, $\hat{\xi}$ with $\hat{\xi}^*$ and finally ξ with $\hat{\xi} = T(\hat{F}_n)$. This gives the bootstrap estimate of the bias

$$\widehat{\text{bias}}_F(\hat{\xi}, \xi) = \text{bias}_{\hat{F}_n}(\hat{\xi}^*, \hat{\xi}) = \text{E}_{\hat{F}_n}[\hat{\xi}^*] - T(\hat{F}_n) \; .$$

In practice, the ideal bootstrap bias estimate is approximated by simulation. Consequently, let $y^{*1}, \ldots, y^{*B}$ be independent bootstrap samples, which give bootstrap estimates $\hat{\xi}^{*(b)}$, $b = 1, \ldots, B$. Taking

$$\bar{\hat{\xi}}^* = \frac{1}{B} \sum_{b=1}^{B} \hat{\xi}^{*(b)}$$

the bias can be estimated with

$$\widehat{\text{bias}}_B = \bar{\hat{\xi}}^* - \underbrace{T(\hat{F}_n)}_{\hat{\xi}} \; .$$

In principle, we now correct for the bias to construct the bias corrected estimate

$$\hat{\hat{\xi}} = \hat{\xi} - \widehat{\text{bias}}_B = \hat{\xi} - [\bar{\hat{\xi}}^* - \hat{\xi}] = 2\hat{\xi} - \bar{\hat{\xi}}^* \; .$$

It should be noted that bias correction in this form is not always recommendable, as correcting the bias typically leads to the bias corrected estimate having a higher variance. Nonetheless, bootstrapping does allow the quantification and estimation of the bias, which can often be helpful.

8.2 Parametric Bootstrap

The bootstrap approach discussed so far was built upon sampling from the empirical distribution function. Even though this approach is quite broadly applicable, it is not always appropriate, in particular, if data are not $i.i.d.$. We therefore propose the parametric bootstrap as an alternative approach to resampling. First of all, we assume that the unknown distribution function $F(.)$ depends on some parameter θ, which we denote with $F(y) = F(y; \theta)$. It is assumed that θ uniquely determines the distribution function and that it is estimated by some statistical method, e.g. Maximum Likelihood. Note that θ, the parameter of the distribution model and the parameter of interest ξ can be different. For instance, θ can be the parameter λ in a Poisson model, while the parameter of interest ξ is a quantile of the distribution.

As a special case we can take $\theta = \xi$. The parametric bootstrap procedure then uses samples taken from the distribution $F(.;\hat{\theta})$, where $\hat{\theta}$ is the estimate of the true parameter θ. An important consequence of this approach is that our bootstrap samples can be different from our original data, which is even guaranteed in the case of a continuous distribution. In this case, the parametric bootstrap samples can take any values. However, this comes at the cost of making stricter parametric assumptions about the distribution of Y. To demonstrate the parametric bootstrap, let us look at the subsequent two examples.

Example 33 We will first demonstrate the parametric bootstrap of the median of a Poisson distribution. Let $Y_1, \ldots, Y_n$ be an $i.i.d.$ sample from a Poisson distribution, $Po(\lambda)$, where $\lambda = \mathrm{E}(Y_i)$, $i = 1, \ldots, n$. We take the median as the parameter of interest ξ, for which we want to compute a bootstrap standard error estimate. The parametric bootstrap now proceeds as follows.

1. Compute the maximum likelihood estimate of λ which is $\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} y_i$.
2. For $b = 1, \ldots, B$:
   (a) Generate bootstrap samples $y^{*b} = (y_1^{*b}, \ldots, y_n^{*b})$ by generating random numbers
   $$y_i^{*b} \sim Po(\hat{\lambda}) \,.$$
   (b) Compute the median $\hat{\xi}^{*b}$ of sample y^{*b}.
3. Compute the standard error estimate
   $$\widehat{\mathrm{se}}_B = \left\{ \frac{1}{B-1} \sum_{b=1}^{B} \left[\hat{\xi}^{*b} - \bar{\hat{\xi}}^{*} \right]^2 \right\}^{\frac{1}{2}} \quad \text{with} \quad \bar{\hat{\xi}}^{*} = \frac{1}{B} \sum_{b=1}^{B} \hat{\xi}^{*b}$$
   exactly as for the nonparametric bootstrap.

▷

Example 34 (Bivariate Normal Distribution) With the next example we want to calculate the correlation coefficient for a bivariate normal distribution. Let $((Y_1, Z_1)', \ldots, (Y_n, Z_n)')$ with

$$\begin{pmatrix} Y_i \\ Z_i \end{pmatrix} \sim F_{Y,Z}(.) \quad i.i.d..$$

Assume, that $F_{Y,Z}(.)$ is a bivariate normal distribution with parameter vector $\theta = (\mu, \Sigma)$, where

$$\mu = \begin{pmatrix} \mathrm{E}(Y_i) \\ \mathrm{E}(Z_i) \end{pmatrix} \qquad \Sigma = \begin{pmatrix} \mathrm{Var}(Y_i) & \mathrm{Cov}(Y_i, Z_i) \\ \mathrm{Cov}(Y_i, Z_i) & \mathrm{Var}(Z_i) \end{pmatrix} .$$

Assume that the parameter of interest is the Pearson correlation coefficient ϱ between Y and Z and that we want to compute an estimate for the standard error of its empirical counterpart.

The Maximum Likelihood estimator $\hat{\theta}$ is

$$\hat{\mu} = \begin{pmatrix} \bar{y} \\ \bar{z} \end{pmatrix},$$

$$\hat{\Sigma} = \frac{1}{n} \begin{pmatrix} \sum_{i=1}^{n}(y_i - \bar{y})^2 & \sum_{i=1}^{n}(y_i - \bar{y})(z_i - \bar{z}) \\ \sum_{i=1}^{n}(y_i - \bar{y})(z_i - \bar{z}) & \sum_{i=1}^{n}(z_i - \bar{z})^2 \end{pmatrix}.$$

The distribution function $F(; \hat{\mu}, \hat{\Sigma}) = N(\hat{\mu}, \hat{\Sigma})$ is therefore used to generate the bivariate bootstrap samples. For each bootstrap sample the Pearson correlation coefficient is computed, i.e.

$$\hat{\varrho}^{*b} = \frac{\sum_{i=1}^{n}(y_i^{*b} - \bar{y}^{*b})(z_i^{*b} - \bar{z}^{*b})}{\sqrt{\sum_{i=1}^{n}(y_i^{*b} - \bar{y}^{*b})^2 \sum_{j=1}^{n}(z_j^{*b} - \bar{z}^{*b})}}$$

and the bootstrap estimated standard error is the usual standard deviation of the values $\hat{\varrho}^{*1}, \ldots, \hat{\varrho}^{*B}$. ▷

8.3 Bootstrap in Regression Models

Let us now look at the linear model discussed in Sect. 7.1 and assume that the data are given as pairs of a scalar response variable and a vector of covariates: $(y_i, \boldsymbol{x}_i)$, $i = 1, \ldots, n$ where $\boldsymbol{x}_i$ denotes a p-dimensional row vector whose first entry is 1 to simplify the calculation of the intercept. We assume the linear model

$$y_i = \boldsymbol{x}_i \boldsymbol{\beta} + \varepsilon_i,$$

where $\varepsilon_i \sim F(.)$ *i.i.d.* and $\mathrm{E}(\varepsilon_i) = 0$ and $Var(\varepsilon_i) = \sigma^2$ for $i = 1, \ldots, n$. Suppose that we want to determine the variance of the Ordinary Least Squares (OLS) estimator $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}$. We propose four approaches to bootstrapping in this scenario, each of which has its advantages and disadvantages.

Approach 1: Residual-Based Bootstrap

The first approach uses a nonparametric bootstrap of the residuals. The true distributional model depends on $\boldsymbol{\beta}$, the vector of regression coefficients, and the distribution function $F(.)$ of the residuals. We notate this with $F(.|\boldsymbol{\beta})$, which we call the "real world", in contrast to the "bootstrap world" that was sketched in Fig. 8.3. Bootstrapping now proceeds as follows.

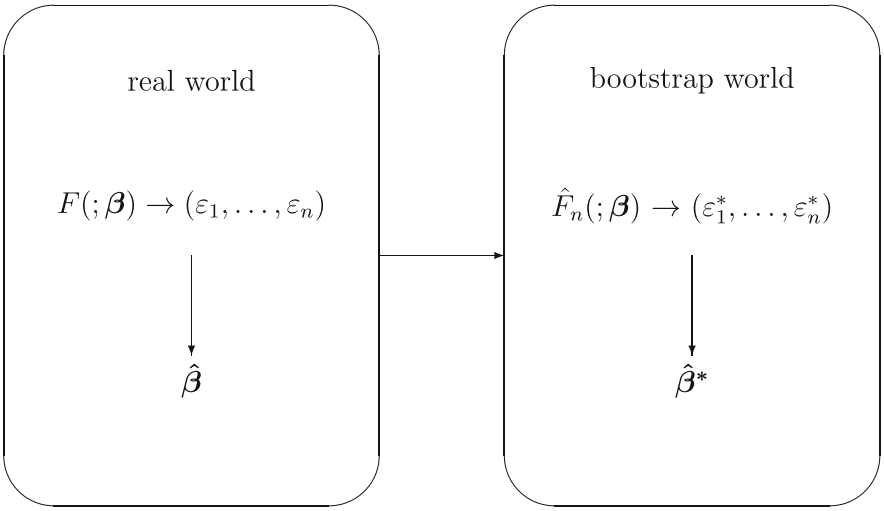

Fig. 8.3 "Real world" and "bootstrap world" for the residual bootstrap of a linear model

First we estimate $\hat{\boldsymbol{\beta}}$ using ordinary least squares (OLS) $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}$ where, as in Sect. 7.1, matrix $\boldsymbol{X}$ has rows $\boldsymbol{x}_i, i = 1, \ldots, n$ and $\boldsymbol{y} = (y_1, \ldots, y_n)^T$. This gives the fitted residuals $\hat{\boldsymbol{\varepsilon}} = (\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}})\boldsymbol{y} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$, giving $\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n$ and its empirical distribution function $\hat{F}_n(.)$. This empirical distribution function $\hat{F}_n$ is now used for bootstrapping by sampling with an equal probability for each residual $\hat{\varepsilon}_i, i = 1, \ldots, n$. A single bootstrap now takes place as follows:

1. Draw a sample $\boldsymbol{\varepsilon}^* = (\varepsilon_1^*, \ldots, \varepsilon_n^*)$ from $\hat{F}_n(.)$ with replacement.
2. Set new bootstrap response values as $y_i^* = \boldsymbol{x}_i\hat{\boldsymbol{\beta}} + \varepsilon_i^*$ for $i = 1, \ldots, n$, which can be written in matrix form as $\boldsymbol{y}^* = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{\varepsilon}^*$.
3. Calculate the bootstrap OLS estimate $\hat{\boldsymbol{\beta}}^* = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}^*$.

An important result for the residual bootstrap is that a simulation is in principle unnecessary, as the bootstrap variance estimate is equal to the standard OLS estimate of the variance of the error terms. This follows because

$$\mathrm{Var}_{\hat{F}_n}(\hat{\boldsymbol{\beta}}^*) = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\mathrm{Var}_{\hat{F}_n}(\boldsymbol{y}^*)\boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1} = \hat{\sigma}_F^2(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1},$$

where $\mathrm{Var}_{\hat{F}_n}(\boldsymbol{y}^*) = \mathrm{Var}_{\hat{F}_n}(\boldsymbol{\varepsilon}^*) = \hat{\sigma}_F^2\boldsymbol{I}$ with $\hat{\sigma}_F^2 = \hat{\boldsymbol{\varepsilon}}^{\mathrm{T}}\hat{\boldsymbol{\varepsilon}}/n$.

Approach 2: Model-Based Bootstrap

In the second approach, we use a parametric bootstrap, modelling our residuals with a normal distribution. Assuming $\varepsilon_i \sim N(0, \sigma^2)$ $i.i.d$, the strategy for a single draw is as follows:

1. Draw an $i.i.d.$ sample $\boldsymbol{\varepsilon}^* = (\varepsilon_1^*, \ldots, \varepsilon_n^*)$ from the normal distribution $N(0, \hat{\sigma}^2)$, where the variance $\hat{\sigma}^2$ is estimated with Maximum Likelihood.
2. Set new bootstrap response values as $y_i^* = \boldsymbol{x}_i \hat{\boldsymbol{\beta}} + \varepsilon_i^*$ for $i = 1, \ldots, n$, which can be written in matrix form as $\boldsymbol{y}^* = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{\varepsilon}^*$.
3. Calculate the OLS estimate with $\hat{\boldsymbol{\beta}}^* = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}^*$.

As in Approach 1, it can be shown that no simulations are necessary, as the bootstrap variance of $\hat{\boldsymbol{\beta}}^*$ converges for $n \to \infty$ to the Fisher information. In other words, both the residual and the model-based bootstrap provide the same variance estimate as the one calculated in Chap. 7 with Maximum Likelihood.

Approach 3: Pairwise Bootstrap
Thus far, we have treated the covariates $\boldsymbol{x}_i$ as a given. That is, the matrix $\boldsymbol{X}$ remained unchanged. Approach 3 now nonparametrically treats the pairs $(y_1, \boldsymbol{x}_1), \ldots, (y_n, \boldsymbol{x}_n)$ as a sample from which bootstrap samples can be directly drawn. Hence we proceed as follows:

1. Draw $(y_i^*, \boldsymbol{x}_i^*)$ with replacement from the original sample, $i = 1, \ldots, n$.
2. Calculate

$$\hat{\boldsymbol{\beta}}^* = (\boldsymbol{X}^{*T}\boldsymbol{X}^*)^{-1}\boldsymbol{X}^{*T}y^*,$$

where $\boldsymbol{X}^*$ is the bootstrapped design matrix with rows $\boldsymbol{x}_i^*$ and $\boldsymbol{y}^* = (y_1^*, \ldots, y_n^*)$.

This approach is less sensitive to the strict $i.i.d.$ assumption applied in the previous two approaches to ensure the exchangeability of the residuals. This means that the pairwise bootstrap can better cope with model violations, for instance, due to variance heterogeneity. This will be explored in an example a bit later.

Approach 4: Wild Bootstrap
Note that in pairwise bootstrapping we not only bootstrap observations y_i^* but also the covariates $\boldsymbol{x}_i^*$. This can be questionable if the covariates are not random, but fixed by design, for example, if they were taken at regular timepoints. In this case the bootstrap sample $(y_i^*, \boldsymbol{x}_i^*)$ has, in fact, a different empirical distribution to the observed values $\boldsymbol{x}_i$. To accommodate the restriction that the distribution of $\boldsymbol{x}_i$ should not change in the bootstrap, the wild bootstrap was proposed by Wu (1986). The idea is to bootstrap residuals $\hat{\varepsilon}_i^* = V_i^*\hat{\varepsilon}_i$, with $\hat{\varepsilon}_i$ as fitted residual and V_i^* being drawn $i.i.d.$ from the two point distribution

$$P(V_i^* = \frac{\sqrt{5}+1}{2}) = \frac{\sqrt{5}-1}{2\sqrt{5}}$$

$$P(V_i^* = -\frac{\sqrt{5}-1}{2}) = \frac{\sqrt{5}+1}{2\sqrt{5}}.$$

These numbers may look arbitrary but they are well chosen, as

$$E(V_i^*) = \frac{\sqrt{5}+1}{2}\frac{\sqrt{5}-1}{2\sqrt{5}} - \frac{\sqrt{5}-1}{2}\frac{\sqrt{5}+1}{2\sqrt{5}} = \frac{5-1}{4\sqrt{5}} - \frac{5-1}{4\sqrt{5}} = 0$$

$$\text{Var}(V_i^*) = \frac{5+2\sqrt{5}+1}{4}\frac{\sqrt{5}-1}{2\sqrt{5}} + \frac{5-2\sqrt{5}+1}{2\sqrt{5}}\frac{\sqrt{5}+1}{2\sqrt{5}} = 1.$$

It can also be shown that $E[(V_i^*)^3] = 1$, such that the first three moments of $\hat{\varepsilon}_i^*$ mimic the empirical estimates. The wild bootstrap proceeds now as follows:

1. Draw a sample $\boldsymbol{\varepsilon}^* = (\varepsilon_1^*, \ldots \varepsilon_n^*)$ using the two point distribution given above.
2. Set new bootstrap response values as $y_i^* = \boldsymbol{x}_i\hat{\boldsymbol{\beta}} + \varepsilon_i^*$ for $i = 1, \ldots, n$, which can be written in matrix form as $\boldsymbol{y^*} = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{\varepsilon^*}$.
3. Calculate the bootstrap OLS estimate $\hat{\boldsymbol{\beta}}\boldsymbol{^*} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y^*}$.

The wild bootstrap is simple to apply and proves to be quite flexible, as the following example shows.

Example 35 We demonstrate the different bootstrap strategies with the rent data from Chap. 7. The variable Y_i represents the rent of an apartment with a given floor space x_i. Figure 8.4 visualises how the different bootstrap approaches affect the determined relationship between these values. In the top left, we have plotted the data and the resulting least squares fit. Note that variance heterogeneity is clearly present, as larger apartments show more variability in their rents. If we apply a residual-based or a model-based bootstrap (middle row), we ignore this fact, as, e.g. for a model-based bootstrap we simulate "new" residuals from the variance homogeneous normal distribution $N(0, \hat{\sigma}_\varepsilon^2)$. In contrast to the original observations, the bootstrapped data now have a homogeneous variance. This is an unavoidable consequence of the bootstrap design and clearly demonstrates what follows when bootstrapped data do not mimic the original data. The least squares slope estimates of the $B = 200$ bootstrap samples are shown in grey, with the original fit shown with the solid line. Both approaches are not suitable to uncover variance heterogeneity. Note also that the bootstrapped rents y_i^* can also be negative, which clearly makes no sense. Hence the bootstrap can produce invalid samples that do not fulfil the positivity requirement. The pairwise bootstrap, seen in the bottom left, circumvents this problem as (y_i^*, x_i^*) always corresponds to an observed pair and hence y_i^* is positive. We see this approach more accurately reflects the variance heterogeneity of the original data. When interpreting the plot, note that points may represent multiple data entries, as we draw the data pairs (y_i, x_i) with replacement. Moreover, the distribution of the bootstrapped values of x_i^* does not match the empirical distribution of original x_i. Finally, we apply the wild bootstrap which is shown in the bottom right plot, where variance heterogeneity is also accounted for. We can still get negative values y_i^*, but with smaller probability than the model-based and residual-based bootstrap. A comparison of the bootstrapped slope parameters $\hat{\beta}_x$ for

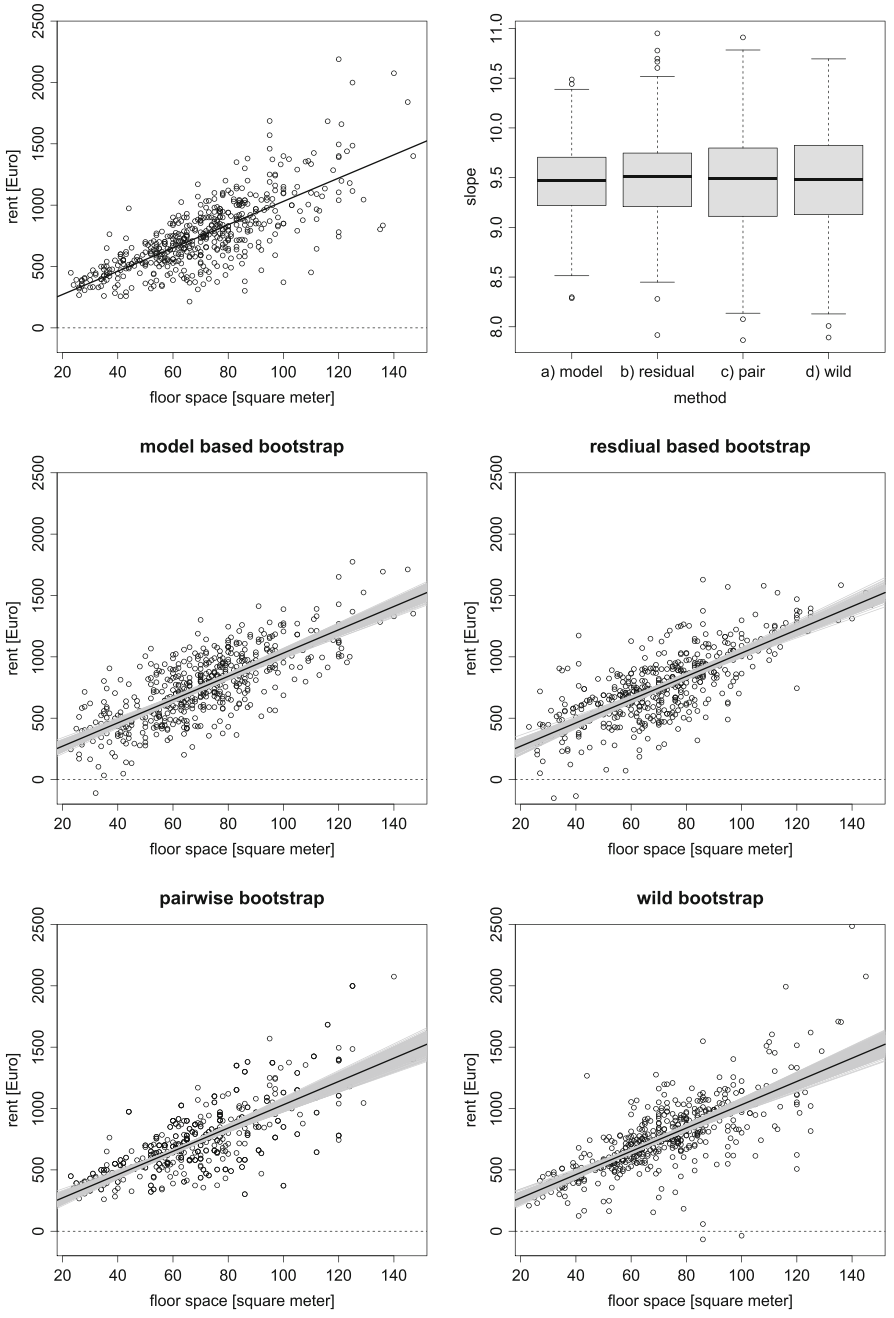

Fig. 8.4 Rental data and bootstrap samples. The top left plot shows the original data and least squares fit and the top right the distribution of bootstrapped slope estimates for each method. The remaining plots show, for each bootstrapping method, a single bootstrap sample, the original least squares slope estimate in black and the bootstrap slope estimates in grey

the four different bootstrap strategies $Y_i = \beta_0 + x_i\beta_x$ is shown in the top right plot. Clearly, incorrectly assuming variance homogeneity underestimates the variance of the slope estimate of $\hat{\beta}_x$, which is accounted for by the pairwise and wild bootstrap.

▷

8.4 Theory and Extension of Bootstrapping

8.4.1 Theory of the Bootstrap

The bootstrap aims to estimate the distribution of a statistic or some characteristic measures of this distribution, such as the expected value, median, variance or quantiles. In this section, we aim to elucidate some of the theory behind this process. Consistency of the bootstrap in estimating the asymptotic distribution of a statistic or estimator can be proven under certain conditions and in simple settings. We think that these proofs can be instructive and will explore some of them here. We will also give counterexamples where the bootstrap is inconsistent to demonstrate its limitations. Take note that, even though the use of bootstrapping is intended as a finite sample approximation of the distribution, asymptotic arguments can come into play. The following results are largely based on Horowitz (2001).

We first introduce some additional notation. Let $Y = (Y_1, \ldots, Y_n)$ be a random sample from an unknown distribution $F(.)$. With $\hat{F}_n$ we denote the empirical distribution $\hat{F}_n(y) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{Y_i \leq y\}}$ or some other (e.g. smooth, parametric) estimate of $F(.)$. Additionally, we take $T_n = \tilde{t}(Y_1, \ldots, Y_n)$ as some statistic of the data (with finite moments). Further, let $G_n(t, F) = P(T_n \leq t)$ be the *exact finite sample* distribution of T_n if the true distribution function of the sample Y is $F(.)$. Note that if $G_n(t, F)$ does not depend on $F(.)$, then T_n is called a pivotal statistic, see Definition 3.15. The bootstrap can be seen as a method for the estimation of $G_n(t, F)$, or some of its characteristic measures, by plugging in $\hat{F}_n$ into G_n, i.e. by using the approximation

$$G_n(t, \hat{F}_n) = P(t(Y^*) \leq t | \hat{F}_n),$$

where $Y^* = (Y_1^*, \ldots, Y_n^*)$ is a bootstrap sample drawn from $\hat{F}_n(.)$. Loosely speaking, the main idea is that for large n

$$G_n(\cdot, \hat{F}_n) \approx G_\infty(\cdot, \hat{F}_n) \approx G_\infty(\cdot, F) \approx G_n(\cdot, F) \ .$$

In the following, we give conditions for the consistency of a bootstrap estimator.

Definition 8.1 The bootstrap estimator $G_n(\cdot, \hat{F}_n)$ is consistent if

$$\lim_{n\to\infty} P_n \left\{ \sup_t |G_n(t, \hat{F}_n) - G_\infty(t, F)| > \epsilon \right\} = 0$$

for each $\epsilon > 0$. P_n is the joint probability distribution of the sample $(Y_1, \ldots, Y_n)$.

Clearly, $\sup_t |F(t) - G(t)|$ is only one possible measure to define the distance between distributions. Bickel and Freedman (1981) introduced a different metric to derive alternative conditions for consistency. They applied what is called the Wasserstein-Mallows metric (the special case of the Mallows metric for $p = 2$), which is defined as follows.

Definition 8.2 Set $p \geq 1$. Let $\mathcal{F}_p$ denote the set of all distribution functions F for which $\int_{-\infty}^{\infty} |t|^p dF(t) < \infty$. For $F, G \in \mathcal{F}_p$, the **Mallows metric** is defined as

$$\rho_p(F, G) = \inf_{\mathcal{T}_{XY}} \left\{ \mathrm{E}|X - Y|^p \right\}^{\frac{1}{p}} ,$$

where $\mathcal{T}_{XY}$ is set of all joint distributions of pairs of two random variables (X, Y) whose marginal distributions of X and Y are F and G, respectively. It implicitly assumes that the moments up to order p exist.

The idea is that it may be easier to show the convergence of $\rho_p(F, G)$ to 0 than that the equation in Definition 8.1 holds. For the special case $p = 2$, we get $\rho_2(F, G) = \inf_{\mathcal{T}_{XY}} \left\{ \mathrm{E}|X - Y|^2 \right\}^{\frac{1}{2}}$. The metric is useful because of the following property: $\rho_2(F, G) \to 0$ holds if and only if F converges in distribution to G and $\mathrm{E}_F(X^k) \to \mathrm{E}_G(X^k)$, $k = 1, 2$. Because the bootstrap is most often used to estimate the distribution function, mean and variance of a statistic, consistency in ρ_2 is well suited to this problem. The following theorem is fundamental to understanding the bootstrap and gives sufficient conditions for its consistency.

Theorem 1 (Beran and Ducharme 1991) *$G_n(\cdot, \hat{F}_n)$ is consistent, i.e. $G_n(t, \hat{F}_n) \to G_\infty(t, F)$, if, for any $\varepsilon > 0$ and F, the following three conditions hold:*

(i) $\lim_{n\to\infty} P_n(\rho_2(\hat{F}_n, F) > \varepsilon) = 0$,
(ii) *$G_\infty(t, F)$ is a continuous function of t,*
(iii) *for any t and any sequence $\{H_n\}$ such that* $\lim_{n\to\infty} \rho(H_n, F) = 0$:

$$G_n(t, H_n) \to G_\infty(t, F).$$

This theorem holds, for example, with the sample average $\bar{Y}$ and the $p = 2$ Mallows metric, which implies convergence of the corresponding sequences of first and second moments.

Example 36 Assume that Y_i has finite variance and define $T_n = \sqrt{n}(\bar{Y} - \mu)$ with $\mu = \mathrm{E}(Y_i), i = 1, \ldots, n$. Then

$$G_n(t, F) = P_n\left\{\sqrt{n}(\bar{Y} - \mu) \leq t\right\} .$$

Let $\hat{F}_n$ be the empirical distribution function of the data and $T_n^* = \sqrt{n}(\bar{Y}^* - \bar{Y})$ be the bootstrap analogue to T_n. Then

$$G_n(t, \hat{F}_n) = P_n^*\left\{\sqrt{n}(\bar{Y}^* - \bar{Y}) \leq t\right\} .$$

▷

We also want to show a counterexample to demonstrate where the bootstrap can fail. Let $Y = (Y_1, \ldots, Y_n)$ with $Y_i \sim \mathrm{Uniform}(0, \theta)$ *i.i.d.*. The sample maximum is the Maximum Likelihood estimator and is denoted by $\hat{\theta}_{\mathrm{ML}} = Y_{(n)} = \max\{Y_i\}$. The probability that $Y_{(n)}$ is not in the bootstrap sample is $\left(1 - \frac{1}{n}\right)^n$. The probability that $Y_{(n)}$ *is* in the bootstrap sample is therefore

$$1 - \left(1 - \frac{1}{n}\right)^n \to 1 - e^{-1} \approx 0.632 \quad \text{for } n \to \infty .$$

Thus, $P(\hat{\theta}^* = \hat{\theta}_{\mathrm{ML}}) \approx 0.632$ for $n \to \infty$ and the distribution of $\hat{\theta}^*$ has a point mass of 0.632, even asymptotically, on the Maximum Likelihood estimator. One can show that the distribution of $\hat{\theta}_{\mathrm{ML}}$ is not asymptotically normal and the bootstrap is not consistent. In fact it holds that

$$P(n(\theta - Y_{(n)}) \leq x) = 1 - \left(1 - \frac{y}{\theta n y}\right)^n \to 1 - e^{-\theta y} ,$$

i.e. the distribution of $n(\theta - Y_{(n)})$ converges to an exponential distribution with parameter θ, a continuous distribution with point mass zero everywhere. Therefore, the bootstrap distribution does not converge to the target distribution.

There are a number of other important exceptions. For example, the bootstrap is inconsistent for the maximum of a sample. The bootstrap is also inconsistent for the distribution of an estimator, when the true parameter is on the boundary of the parameter space. A simple example is the square $\bar{Y}^2$ of the sample average of *i.i.d.* variables with mean μ and variance σ^2 if $\mu = 0$. Further examples are given in Horowitz (2001).

8.4.2 Extensions of the Bootstrap

There are several extensions to the bootstrap for alternative data types, e.g. time series, panel data, longitudinal data and stochastic processes. These situations are characterised by the fact that the observations can be dependent upon each other. Recall that bootstrap essentially assumes $i.i.d.$ observations (univariate and multivariate) or, as in residual sampling in regression, $i.i.d.$ errors. This is not the case with dependent observations or heteroscedastic errors, which we saw could be managed with the wild and pairwise bootstrap.

A more general approach, however, is the Bayesian bootstrap proposed by Rubin (1981), which is able to directly address these more complex systems. It generates new data by weighting the original sample of size n and calculating weighted statistics. To begin with, we need a random weight for each sample, such that all weights add to 1 and $\mathrm{E}(G_i) = 1/n$, i.e. the expected proportion of a sample's inclusion in the nonparametric bootstrap. One simple approach is to draw $n-1$ random numbers $u_1, \ldots, u_{n-1}$ from a standard uniform distribution on $(0, 1)$ and order them as $u_{(0)}, u_{(1)}, \ldots, u_{(n-1)}, u_{(n)}$ with $u_{(0)} = 0$ and $u_{(n)} = 1$. Calculate now the n gaps $g_i = u_{(i)} - u_{(i-1)}, i = 1, \ldots, n$ leading to $g = (g_1, \ldots, g_n)$ as the vector of probabilities used for weighting the original sample.

8.4.3 Subsampling

In the bootstrap approaches above we sample single variables. For time series data, Politis et al. (1999) have proposed different possibilities for bootstrapping, including block bootstrapping, where instead of a single variable, entire blocks of observations are sampled from the observed data. This in turns maintains serial dependence during the bootstrapping process. Politis and Romano (1994) lists several ideas and bootstrap extensions for bootstrapping in the context of prediction.

Subsampling generally takes a sample of size m from the original sample of size $n > m$. This is also called the m-out-of-n bootstrap. In principle, we have two options: sampling with replacement (as in the usual bootstrap), called **replacement subsampling** or sampling without replacement, called **non-replacement subsampling**.

Sampling without replacement gives $\binom{n}{m}$ possible sub-samples and is conceptually different from the standard bootstrap, as the sub-sample of size $m < n$ is then theoretically a sample from the unknown F and not from $\hat{F}_n$. Both methods give consistent estimates under more general conditions than the standard bootstrap, as long as m and n tend to infinity and m/n tends to zero. It was shown by Politis and Romano (1994) and Politis et al. (1999) that non-replacement subsampling works under much weaker conditions than bootstrapping. The disadvantage is that one has to choose the sample size m, which is a tuning parameter that influences

the performance of subsampling, especially in real applications where samples are finite.

8.5 Bootstrapping the Prediction Error

8.5.1 Prediction Error

In predictive modelling, the most interesting target is the prediction error and its minimisation. Predictive models relate a response variable Y to a vector of explanatory variables x, also called predictor variables. We estimate a structural relationship or regression

$$\mathrm{E}(Y|x) = m(\mathrm{x}),$$

where $m(.)$ is the prediction model. This can be done with linear regression models, i.e. $m(x) = \beta_0 + x\beta_x$, but also with more complex models, such as regression trees or even neural networks. For a metric response variable Y, the interesting quantity is often the prediction error measured by the mean squared error of prediction (MSEP)

$$\mathrm{E}(Y - \hat{Y})^2 \,,$$

where $\hat{Y} = \hat{m}(x)$ is a prediction given by a fitted model (or any other prediction algorithm). Note that

$$E\{(Y - \hat{Y})^2\} = E\left\{(Y - m(x) + m(x) - \hat{Y})^2\right\} = \mathrm{Var}(Y) + \mathrm{MSE}(\hat{Y}).$$

If Y is categorical, then this is a classification problem and the prediction error is often measured as the probability of a false classification, i.e. $P(\hat{Y} \neq Y)$. In the following, we restrict ourselves to the estimation of the MSEP. The MSEP itself is a theoretical quantity, which therefore needs to be estimated. A naive estimator in the regression context would be

$$\widehat{\mathrm{MSEP}} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{8.5.1}$$

or

$$\widehat{\mathrm{MSEP}} = \frac{1}{n-p}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2,$$

where $\hat{y}_i = \hat{m}(x_i)$ and p is the number of predictor variables. This implies that we estimate the prediction with the same observations that were used to fit the

prediction model $m(x)$. Consider a simple linear regression model with $Y_i = \beta_0 + x_i\beta_x + \varepsilon_i$ and data $(y_i, x_i), i = 1, \ldots, n$. We fitted β_0 and β_x by minimising the squared error

$$(\hat{\beta}_0, \hat{\beta}_x) = \arg\min \sum_{i=1}^{n} (y_i - \beta_0 - x_i\beta_x)^2$$

which also minimises the MSEP in (8.5.1). However, because we used the same data for building the model and for assessing its predictive quality, this "in-sample" estimate is too optimistic, i.e. it underestimates the error. In an ideal situation, new independent data are available and the MSEP can be estimated with the following procedure:

- Use the **training data** $y_1, \ldots, y_n$ to build a model $\hat{m}(\cdot)$.
- Use the fitted model to predict new data y_j^0 with $\hat{y}_j^0$, for $j = 1, \ldots, m$. The new data pairs (y_j^0, x_j^0) are different than to the data used for fitting the model and are therefore usually called the **test data**.
- Estimate the prediction error with

$$\widehat{\text{MSEP}} = \frac{1}{m} \sum_{i=1}^{m} (y_i^0 - \hat{y}_i^0)^2.$$

In practice, new data are often not directly available or only available at a later date upon collection of more data. Cross validation is one approach to estimating the MSEP with only the current dataset.

8.5.2 Cross Validation Estimate of the Prediction Error

The most common strategy for cross validation is called k-fold cross validation and involves randomly splitting the data into K sections of roughly the same size. For each section $k = 1, \ldots, K$, a model is fitted with the data from $K - 1$ sections while prediction quality is assessed with data from the remaining section. In Fig. 8.5, we visualise the process for $K = 6$. Let $k(i)$ be the section which contains the i-th observation y_i. For example, in Fig. 8.5 we have $k(m) = 2$.

Furthermore, let $\hat{y}_i^{-k(i)}$ denote the prediction for y_i, calculated without the section $k(i)$, i.e. without the section which contains y_i. The estimation of the MSEP using cross validation is then given by

$$\widehat{\text{MSEP}} = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \hat{y}_i^{-k(i)} \right)^2 . \tag{8.5.2}$$

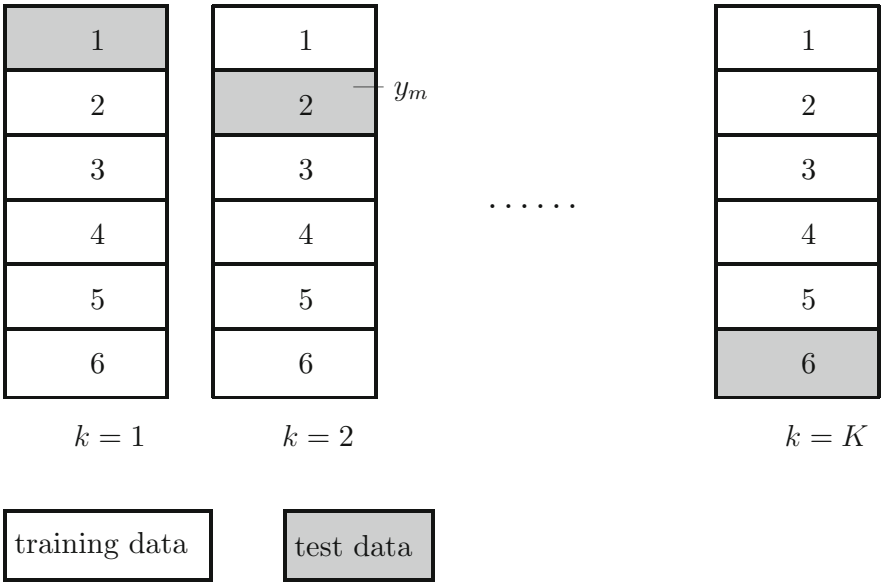

Fig. 8.5 Sketch of cross validation process

The major advantage of cross validation is, that y_i is independent of $\hat{y}^{-k(i)}$. Replacing the observations with random variables gives

$$\begin{aligned} E\left(\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i^{-k(i)})^2\right) &= \frac{1}{n}\sum_{i=1}^{n} E\left\{(Y_i - \hat{Y}_i^{-k(i)})^2\right\} \\ &= \frac{1}{n}\sum_{i=1}^{n} E\left\{(Y_i - m(x_i) + m(x_i) - \hat{Y}_i^{-k(i)})^2\right\} \\ &= \frac{1}{n}\sum_{i=1}^{n} E\left\{(Y_i - m(x_i))^2\right\} + \frac{1}{n}\sum_{i=1}^{n} E\{(m(x_i) - \hat{Y}_i^{-k(i)})^2\}. \end{aligned} \tag{8.5.3}$$

Clearly, as a result of independence this decomposes into the variance of Y (or the average variance in case of variance heterogeneity) and the MSE of the estimate $\hat{Y}_i^{-k(i)} = \hat{m}^{-k(i)}(x_i)$. When $K = n$, this is called leave-one-out cross validation, which has low variance but is less useful for large n and complex regression methods due to the high computational cost.

Note that our prediction error estimates are conditional on the configuration of the predictor variables, i.e. the estimates depend on the distribution of x. If this distribution differs between the training and test data, e.g. in future data, the true prediction error may be different and our MSEP estimate will not be

correct. Therefore, our estimates are only **conditional error estimates**. Alternative strategies can be used if a large amount of data is available. In this case, the data can be randomly split into two parts, simply using one part for training and the other for testing. One may even split the data randomly into three parts: training data, validation data and test data. This is recommended when the training data is used heavily for model fitting. The validation data then works as temporary test data until a final model is decided upon, which is then finally evaluated on the test data. Simple random sampling can also be replaced with more complex approaches, such as stratified sampling. The proportions in the splits can also vary. For example, with the simple train-test split approach, data can be split into two equal sized parts or sometimes a ratio of two thirds training data to one third test data is recommended. A survey of various cross validation strategies is given by Arlot and Celisse (2010).

8.5.3 Bootstrapping the Prediction Error

Let us use regression again as an example to describe a bootstrap procedure for estimating the prediction error. Given paired data $(y_i, x_i), i = 1, \ldots, n$, the goal is to predict a new observation Y for a given x, both drawn from a population distribution F. Let z denote the data $(y_i, x_i), i = 1, \ldots, n$. Note that the data z depend on the sample size n, which is, however, suppressed in the subsequent notation. The prediction is subsequently denoted as $\hat{m}(x)$, which clearly depends on the data $\boldsymbol{z}$ used for fitting. The prediction error for $\hat{m}(x)$ is then given by

$$\text{err}(\boldsymbol{z}, F) \equiv \text{E}_F \left\{ (Y - \hat{m}(x))^2 \right\}.$$

Note that the data z used to fit the prediction model is also drawn from $F(.)$. In other words, we are testing how our model, fitted with the data z, performs on the new datapoint (Y, x), which is also drawn randomly from F.

A naive approach to estimating this theoretical quantity is the **sample error**, which is defined as

$$\text{err}(z, \hat{F}_n) = E_{\hat{F}_n} \left\{ (Y - \hat{m}(x))^2 \right\} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{m}(x_i))^2,$$

where we replace the expectation with the average error over the data. We have already mentioned that, because it uses the same data for both fitting and evaluating the model, this quantity is too optimistic and hence gives a biased estimate of $\text{err}(\boldsymbol{z}, F)$. Note that attempting to bootstrap the prediction error with the existing model $\text{err}(\boldsymbol{z}, \hat{F}_n)$ is not a plug-in version of $\text{err}(\boldsymbol{z}, F)$, because the training data $\boldsymbol{z}$ is not drawn from $\hat{F}_n(.)$. In other words, we need to be more careful when constructing a plug-in version of $err(\boldsymbol{z}, F)$. To apply the plug-in principle rigorously we draw

the B bootstrap samples (y_i^{*b}, x_i^{*b}) with $i = 1, \ldots, n$ and $b = 1, \ldots, B$. Then, for any b,

$$\text{err}(\mathbf{z}^{*b}, \hat{F}_n) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{m}^{*b}(x_i))^2$$

results as a plug-in estimate of $\text{err}(\mathbf{z}, F)$, where y_i and x_i are from the original sample but $\hat{m}^{*b}(.)$ is the fitted predictor using the b-th bootstrapped data. Hence, the prediction model is estimated with bootstrapped data, while the prediction error is calculated from the original sample. The prediction error still depends on the data and we could now question how to calculate the expected prediction error, i.e. the average of $\text{err}(z, F)$. We will stop here, because this is getting a little bit clumsy and, while bootstrapping is easy and recommended for quantifying uncertainty, methods like cross validation are easier and more appropriate when calculating prediction error in practice.

8.6 Bootstrap Confidence Intervals and Hypothesis Testing

8.6.1 Bootstrap Confidence Intervals

Thus far, we have used bootstrapping to derive properties of estimates $\hat{\xi} = t(Y)$. However, the bootstrap distribution of the estimate that we calculate also allows us to directly obtain confidence intervals. Let, as above, $t(Y)$ be a statistic that estimates a parameter ξ. The confidence interval is defined as $[t_l(Y), t_r(Y)]$, such that (as in Definition 3.15)

$$P(\hat{\xi} \in [t_l(Y), t_r(Y)]) \geq 1 - \alpha.$$

Using the bootstrap principle, we replace the above expression with

$$P(\hat{\xi} \in [t_l(y^*), t_r(y^*)]) \geq 1 - \alpha,$$

where $y^* = (y_1^*, \ldots, y_n^*)$ is drawn from the empirical distribution $\hat{F}_n(.)$. In practice, one takes $t_l(.)$ and $t_r(.)$ to be the $\alpha/2$ and $1 - \alpha/2$ quantile of the bootstrap samples $\hat{\xi}^{*b}$. That is, we draw B bootstrap samples $y_1^{*b}, \ldots, y_n^{*b}$ with $b = 1, \ldots, B$ and derive the bootstrap estimates $\hat{\xi}^{*b}$. Note that B needs to be sufficiently large, as the $\alpha/2$ quantile of B bootstrap replicates relies on only $(\alpha B/2)$ observations. For $\alpha = 0.05$ and $B = 200$ this means that we have only 5 bootstrap samples below the 2.5% quantile, too few to provide reliable estimates of the confidence interval boundaries. Hence, if confidence intervals are derived from quantiles of bootstrap samples one should work with a relatively large B.

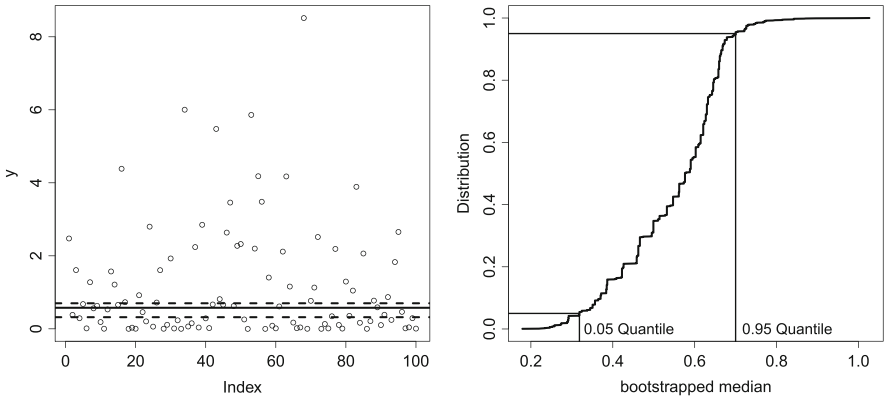

Fig. 8.6 Chi-squared distributed data (left) and the fitted median (solid line) with bootstrapped confidence intervals (dashed lines). The right-hand plot shows bootstrap distribution

Example 37 We simulate 100 data-points from a chi-squared distribution with one degree of freedom and consider the median as the parameter of interest. The data is shown on the left of Fig. 8.6, with the median delineated by the solid line and the 90% bootstrapped median values by the dashed lines. We can see a clear asymmetry of the confidence interval. On the right-hand side is the entire distribution of the bootstrapped median values. We see that in fact quite a lot of information is available and the entire bootstrap distribution can be used to derive properties of the estimate.

▷

8.6.2 Testing

Bootstrap hypothesis testing is similar to permutation testing, in that both are usually implemented using a Monte-Carlo procedure. However, for permutation testing one draws without replacement and for bootstrap with replacement. We will sketch both ideas in the coming section. In order to keep things simple, we limit ourselves here to the two-sample hypothesis. We assume

$$Y_i \sim F(.) \quad i.i.d., i = 1, \ldots, n$$

$$Z_j \sim G(.) \quad i.i.d., j = 1, \ldots, m$$

and question the hypothesis

$$H_0 : F = G.$$

The data are denoted as $\boldsymbol{y} = (y_1, \ldots, y_n)$ and $\boldsymbol{z} = (z_1, \ldots, z_m)$ and the combined sample is written as $\boldsymbol{x} = (\boldsymbol{y}, \boldsymbol{z})$. Note that both the individual samples and the combined sample are drawn from the same distribution under H_0.

The first step for testing is to choose a test statistic $t(\boldsymbol{x})$ which is sensitive to differences between F and G, that is, one that will also guarantee the power of the test. Once we have defined $t(\boldsymbol{x})$, we need to simulate the distribution of $t(\boldsymbol{x})$ under the null hypothesis. In principle, there are several possible test statistics for this situation, but for demonstration purposes let us just compare the arithmetic means. Let $t(\boldsymbol{x}) = \bar{y} - \bar{z}$ where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\bar{z} = \frac{1}{m}\sum_{j=1}^{m} z_j$. Then under $H_0 : F = G = F_0$ we can draw bootstrap samples of length $n + m$ from the combined sample $\boldsymbol{x}$. The procedure is then as follows:

1. Draw B bootstrap samples $\boldsymbol{x}^{*b}$ of size $n + m$ with replacement from $\boldsymbol{x}$.
2. Consider $\boldsymbol{x}^{*b}$ as $(\boldsymbol{y}^{*b}, \boldsymbol{z}^{*b})$ and use the first n observations to calculate $\bar{y}^{*b}$ and the remaining m observations to calculate $\bar{z}^{*b}$ and evaluate $t(\boldsymbol{x}^{*b}) = \bar{y}^{*b} - \bar{z}^{*b}$ for $b = 1, \ldots, B$.
3. The two-sided bootstrapped p-value of the test is then given by

$$p\text{-value}_{\text{boot}} = \frac{1}{B}\sum_{b=1}^{B} 1\{|t(x^{*b})| > |t(x)|\}.$$

If only a few bootstrap samples generate higher absolute values than the observed test statistic $|t(\boldsymbol{x})|$, then the bootstrapped p-value is low. In this case, $t(\boldsymbol{x})$ is rather extreme under the null hypothesis H_0, which therefore should be rejected. Note that the procedure is different from the one pursued in (8.1.4), where we generated a bootstrap estimate of the variance of $t(\boldsymbol{x})$ and the bootstrap samples were generated separately for each of the two groups.

A permutation test proceeds very similarly. Again, we pool the data in $\boldsymbol{x}$ but instead of drawing a sample from $\boldsymbol{x}$ with replacement, we draw a sample from $\boldsymbol{x}$ without replacement, i.e. we permute the order of the entries of $\boldsymbol{x}$. The remaining calculation of the p-value remains unchanged. Usually, using a studentised statistic increases the accuracy of testing and is therefore preferable. For example, we could use

$$t(\boldsymbol{x}) = \frac{\bar{y} - \bar{z}}{s}$$

with $s^2 = \frac{1}{n+m-2}(\sum_{i=1}^{n}(y_i - \bar{y})^2 + \sum_{j=1}^{m}(z_j - \bar{z})^2)$. This is the two-sample t statistic under the assumptions of equal variances in the two groups. One could also use

$$t(\boldsymbol{x}) = \frac{\bar{y} - \bar{z}}{s_y^2/n + s_z^2/m}$$

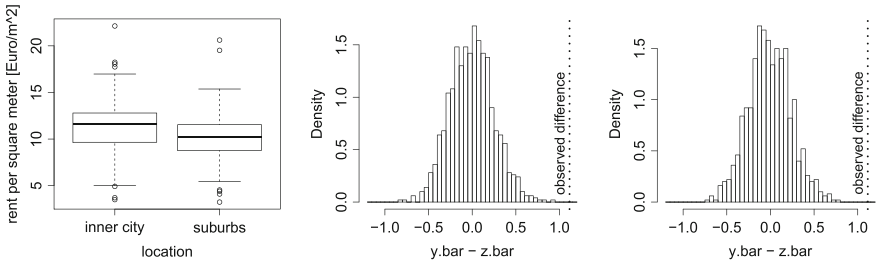

Fig. 8.7 Difference on rent per square metre for inner city apartments and apartments in the suburbs (left plot). Simulated differences using permutation (middle plot) and bootstrapping (right-hand plot)

with $s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$ and $s_z^2 = \frac{1}{m-1}\sum_{j=1}^{m}(z_j - \bar{z})^2$ and the bootstrap analogues for testing.

We have used a mean difference for testing H_0, but, of course, other statistics are also possible. This choice of statistic also affects the power to detect differences between the distributions. It is difficult to generally recommend a particular statistic as we can only approximate the distribution of $t(\boldsymbol{x})$ under H_0 but not under unspecified alternatives. The power is therefore dependent on $H_1 : F \neq G$ and cannot be calculated without further parametric assumptions on F and G. A comprehensive overview of bootstrap and permutation testing can be found in Good (2005). Let us round off this section with a few examples that hopefully demonstrate the ease and flexibility with which bootstrapping can be applied to testing.

Example 38 Let us once again take a look at the Munich rental data and question whether the location of the apartment has an influence on the rent. We therefore take the rent per square metre as response variable and make two location categories: *inner city* and *suburbs*. The rent per square metre is shown in Fig. 8.7. We denote the rent of apartments in the city centre with $\boldsymbol{y} = (y_1, \ldots, y_n)$ and the rent of apartments in the suburbs with $\boldsymbol{z} = (z_1, \ldots, z_m)$, where n=278 and m=222. We take the test statistic $t(\boldsymbol{x}) = t(\boldsymbol{y}, \boldsymbol{z}) = \bar{y} - \bar{z}$ and apply both the permutation and bootstrap tests. The resulting simulated differences are shown in the middle plot in Fig. 8.7 for the permutation test and on the right for the bootstrap test. Clearly, there is a significant difference and apartments in the city centre are significantly more expensive than those in the suburbs. ▷

Example 39 Let $Y_i, \ldots, Y_n \sim F(.)$ *i.i.d.* and let us test whether $F(.)$ is symmetric around a known c. Note that $F(.)$ is symmetric around c if and only if the distribution of Y and $c - Y$ is the same. In other words, we have the hypothesis

$$H_0 : F(y - c) = 1 - F(c - y)$$

for all $y \in \mathbb{R}$. A number of test statistics and approaches have been proposed for this problem and we refer to Zheng and Gastwirth (2010) for a recent summary. In

our bootstrap approach we take

$$t(Y) = \bar{Y} - Y_{med}$$

as the test statistic. That is, we take the difference between the mean estimate $\bar{Y}$ and the median estimate Y_{med}, where Y_{med} is the empirical median of $Y_1, \ldots, Y_n$. This leads to the observed value $t(y) = \bar{y} - y_{med}$, which, if close to zero, suggests a symmetric distribution and if large suggests asymmetry. To apply a bootstrap test, we need to bootstrap (simulate) from a symmetric distribution, that is, we need to simulate under H_0. Therefore, we need some way to force $\hat{F}_n$ to be symmetric. The trick is to complement the data with "symmetrised" values $y_{n+1}, .., y_{n+n}$ which are defined as

$$y_{n+i} = 2c - y_i = c + (c - y_i).$$

In other words, we mirror all observations around c. This gives an extended dataset $\tilde{\boldsymbol{y}} = (y_1, \ldots, y_n, y_{n+1}, \ldots, y_{2n})$, from which we now draw, with replacement, a bootstrap sample of size n. From the bootstrap sample y^{*b} we calculate $\bar{y}^{*b}$ and y^{*b}_{med}, from which we can derive the bootstrapped $t(y^{*b})$. Taking B bootstraps allows us to compare $t(y)$ with the empirical distribution function of $t(y^{*b})$. We apply the test to the apartment size, labelled as y, also from the Munich rental data. The question is whether this is symmetric around the value 69, which is the median. We take $c = 69$ as fixed. The left of Fig. 8.8 shows the empirical distribution (solid line) and the symmetrised distribution (dotted line). Clearly, there is some level of asymmetry that can be assessed with a test. The observed value of the test statistic is $t(\boldsymbol{y}) = \bar{y} - y_{med} = -1.83$. Bootstrapped values $t(\boldsymbol{y}^{*b})$ are shown on the right. We can see that the observed value lies far away from the bootstrapped values, with a

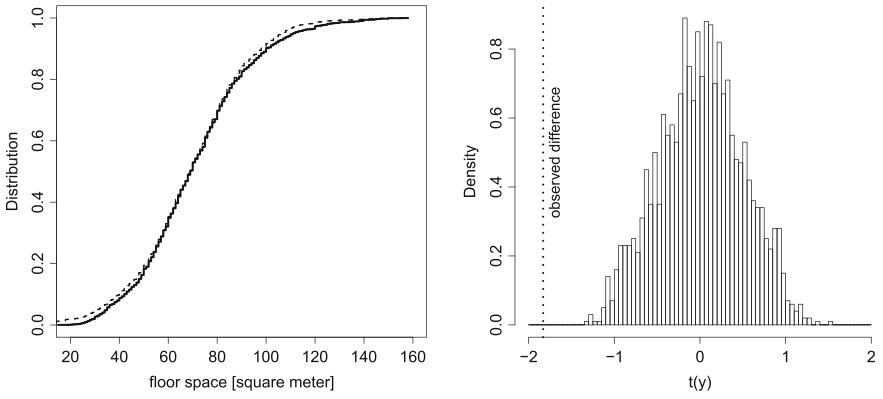

Fig. 8.8 Left: Empirical distribution of floor space and its symmetrised counterpart; Right: Bootstrapped differences between the mean and the median and the observed difference

bootstrapped p-value less than $1/B = 1/2000$, with our 2000 bootstrap replicates. This suggest clearly that the floor space is not symmetrically distributed.

▷

Example 40 In Sect. 6.5, we already discussed tests on independence. Let us take the correlation coefficient test and demonstrate its bootstrap version. Let Y_{i1} and Y_{i2} be drawn jointly from F:

$$(Y_{i1}, Y_{i2}) \sim F(y_1, y_2).$$

We want to test the hypothesis

$$H_0 : F(y_1, y_2) = F_1(y_1)F_2(y_2)$$

with $F_1(.)$ and $F_2(.)$ being the univariate marginal distributions of Y_{i1} and Y_{i2}, respectively. As a test statistic, we use the empirical correlation (even though other dependence statistics can also be used):

$$\hat{\rho}_{12} = \frac{\sum_{i=1}^{n}(Y_{i1} - \bar{Y}_1)(Y_{i2} - \bar{Y}_2)}{\sqrt{\left(\sum_{i=1}^{n}(Y_{i1} - \bar{Y}_1)^2\right)\left(\sum_{i=1}^{n}(Y_{i2} - \bar{Y}_2)^2\right)}}.$$

We are interested in the distribution of $\hat{\rho}_{12}$ in the case of H_0. This is easily bootstrapped by replacing $F_1(.)$ with its empirical counterpart $\hat{F}_{1n}(.)$ and likewise for $\hat{F}_2(.)$. In other words, we draw a univariate sample, with replacement, of the first variable $(y_{11}^*, \ldots, y_{n1}^*)$ and the second variable $(y_{12}^*, \ldots, y_{n2}^*)$. These two independent samples are put together to give the final bootstrap sample

$$(y_{i1}^*, y_{i2}^*) \text{ with } i = 1, \ldots, n.$$

From this sample we can calculate a bootstrap correlation coefficient $\hat{\rho}^*$, which is calculated under H_0. Repeating the bootstrap B times provides the reference distribution under H_0. ▷

8.7 Sampling from Data

The idea of sampling that we have thus far only applied to bootstrapping and other resampling methods can be seen from a different angle, which can also be useful in the age of big data. Assume that we need to analyse a massive database and the number of observations N is very large. While this is certainly a pleasant situation, more is not always better if the quality of the data is not sufficient, as we will discuss in more detail in Sect. 11.3. But for now let us focus on the computational effort for the data analysis, which increases with the size of available data. This

increase is particularly pronounced for some analytic models and procedures and often hardware limitations can mean that the entire dataset cannot be processed at once. In this case, it is both advisable and statistically sound to draw a sample from the data and run the analysis only with this sample. We define with n the sample size. As we have seen, the variance of estimates decreases with increasing sample size n, which also applies in this case. Hence, instead of analysing the entire database of N entries, we allow for a random error and draw n observations from the N data. If all observations get the same probability of inclusion, this is called **simple random sampling**. In fact, one usually draws without replacement and hence the central $i.i.d.$ assumption is violated. Indeed it can be shown (see Thompson 2002) that drawing without replacement reduces the variance compared to drawing with replacement. In other words, if we treat data which are drawn without replacement as if they were drawn with replacement, we overestimate the variance of statistics calculated from the sampled data. The difference is of order $\frac{n}{N}$, meaning that the variance without replacement is smaller by a factor $(1 - \frac{n}{N})$ compared to the variance with replacement. This also means that, if we pretend that data are drawn $i.i.d.$, we still develop valid confidence intervals or tests, as we would have effectively overestimated the variance of the quantity of interest by taking a sample with replacement. The variance of any estimate decreases with n, which we sketched in Fig. 2.2 already. Hence, the standard deviation decreases with $1/\sqrt{n}$ and with increasing sample size n the accuracy of our analysis increases. Let us informally define accuracy as the reciprocal of the width of a confidence interval calculated from the data. As the standard deviation decreases with order $1/\sqrt{n}$ (in the best case), accuracy then increases with order $\sqrt{n}$. If we calculate the estimates with more data, the computational complexity increases. In the best case, this is linear in n, but most algorithms have a higher complexity, e.g. $n \log(n)$ or n^2. In Fig. 8.9, we sketch both computational complexity and achieved accuracy and their dependence on the data size. We see that to simply analyse all of the data as a rule is counterproductive, as the computational burden increases sharply while the gain in accuracy is small. This suggests that in many cases it can be useful to simply sample data and run the analysis on sampled data. We emphasise that the message mirrored in Fig. 8.9 holds in general, that is not only for statistical models but for all data analytic algorithms. Sampling from massive databases therefore appears to be a recommendable strategy if the computational effort for analysing all of the data does not justify the gains in accuracy.

So far, we recommended using simple random sampling. This means each data entry (or individual) has the same probability of selection and inclusion in the data analysis of the sampled data. While this sounds sensible, it is not necessarily the most efficient way to draw a sample. In some situations, it is advisable to draw units with unequal probability, sometimes called oversampling of units. The resulting sampled dataset is clearly biased and subsequent data analysis needs to take this into account, which can be done through a weighted analysis. We already treated sampling weights and a corresponding weighted data analysis in regression already in Sect. 7.2. Here, we just want to mention and emphasise that the idea of drawing

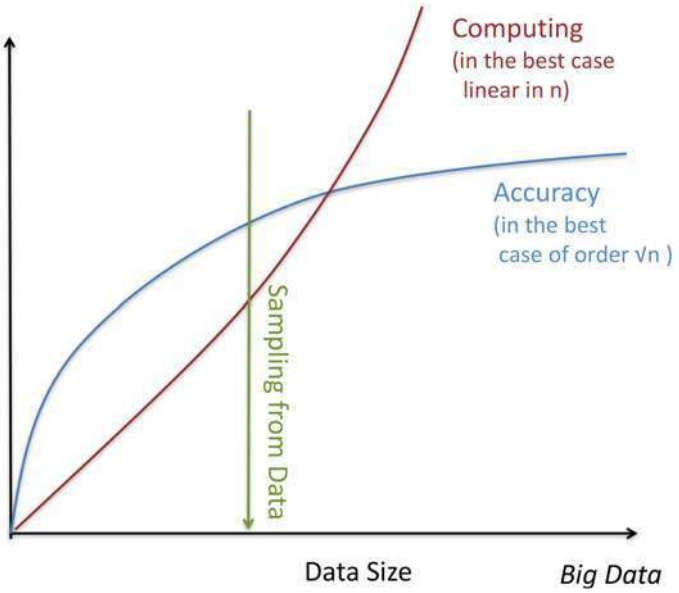

Fig. 8.9 Trade-off between computational effort and accuracy of data analysis for increasing sample size

a representative sample from a database does not require that the resulting sampled data are representative. We refer to Thompson (2002) for further details.

8.8 Exercises

Exercise 1 (Use R Statistical Software)
We consider n *i.i.d.* realisations $(y_1, \ldots, y_n)$ of a normal random variable $Y \sim N(\mu, \sigma^2)$ with unknown μ and $\sigma^2 > 0$.

1. Consider the Maximum Likelihood estimates

$$\hat{\mu} = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\mu})^2.$$

   Try to calculate *estimates* for the variances of $\hat{\mu}$ and $\hat{\sigma}^2$, $\widehat{Var}(\hat{\mu})$ and $\widehat{Var}(\hat{\sigma}^2)$ and the corresponding standard errors $\sqrt{\widehat{Var}(\hat{\mu})}$ and $\sqrt{\widehat{Var}(\hat{\sigma}^2)}$.
2. Write pseudo code for the estimation of these standard errors with a nonparametric bootstrap.

3. Simulate with varying sample sizes $n = 10, 50, 100, 500$ samples from a normal distribution with $\mu = 10$ and $\sigma^2 = 25$. Implement your pseudo code and test your function with $B = 500$ replications. Compare your results by using the function `boot` of the `R` package `boot`.
4. Compare your results additionally to the estimates in (1).

Exercise 2

The `spatial` data in the `R` package `bootstrap` contains the outcomes of two tests A and B on the spatial perception of 26 neurologically impaired children. We want to construct a 95% confidence interval for the correlation $\rho = Cor(A, B)$ between the two test results.

1. Construct a confidence interval using the formula $\hat{\rho} \pm 1.96\hat{se}$ where $\hat{\rho}$ is the usual Pearson correlation and $\hat{se}$ is a nonparametric bootstrap estimate of the standard error of $\hat{\rho}$.
2. Construct a confidence interval using the Fisher Z transformation

$$\theta = \frac{1}{2} \log \left(\frac{1+\rho}{1-\rho} \right) .$$

   Hint: Apply the inverse transformation on the endpoints of the interval to get an interval for ρ once you have an interval for θ.
3. Construct an interval using a parametric bootstrap assuming a bivariate normal distribution for the two outcomes.
4. Construct a nonparametric bootstrap interval for the Spearman rank correlation.
5. Compare the results of all the above methods.