Chapter 14
Monte Carlo Methods

Abstract The term *Monte Carlo* refers to the use of random variables to evaluate quantities such as integrals or parameters of fit functions that are typically too complex to evaluate via other analytic methods. This chapter presents elementary Monte Carlo methods that are of common use in data analysis and statistics, in particular the bootstrap and jackknife methods to estimate parameters of fit functions.

14.1 What is a Monte Carlo Analysis?

The term *Monte Carlo* derives from the name of a locality in the Principality of Monaco known for its resorts and casinos. In statistics and data analysis Monte Carlo is an umbrella word that means the use of computer-aided numerical methods to solve a specific problem, typically with the aid of random numbers.

Traditional Monte Carlo methods include numerical integration of functions that can be graphed but that don't have a simple analytic solution and simulation of random variables using random samples from a uniform distribution. Another problem that benefits by the use of random numbers is the estimation of uncertainties in the best-fit parameters of analytical models used to fit data. There are cases when an analytical solution for the error in the parameters is not available. In many of those cases, the bootstrap or the jackknife methods can be used to obtain reliable estimates for those uncertainties.

Among many other applications, Monte Carlo Markov chains stand out as a class of Monte Carlo methods that is now commonplace across many fields of research. The theory of Markov chains (Chap. 15) dates to the early twentieth century, yet only over the past 20 years or so it has found widespread use as Monte Carlo Markov chains (Chap. 16) because of the computational power necessary to implement the method.

© Springer Science+Busines Media New York 2017

M. Bonamente, *Statistics and Analysis of Scientific Data*, Graduate Texts in Physics, DOI 10.1007/978-1-4939-6572-4_14

14.2 Traditional Monte Carlo Integration

A common numerical task is the evaluation of the integral of a function $f(x)$ for which analytic solution is either unavailable or too complicated to calculate exactly,

$$I = \int_A f(x)dx. \tag{14.1}$$

We want to derive a method to approximate this integral by randomly drawing N samples from the support A. For simplicity, we assume that the domain of the variable $f(x)$ is a subset of real numbers between a and b. We start by drawing samples from a uniform distribution between these two values,

$$g(x) = \begin{cases} \dfrac{1}{b-a} & \text{if } a \leq x \leq b \\ 0 & \text{otherwise.} \end{cases} \tag{14.2}$$

Recall that for a random variable X with continuous distribution $f(x)$, the expectation (or mean value) is defined as

$$E[X] = \int_{-\infty}^{\infty} xg(x)dx \tag{14.3}$$

(2.6); we have also shown that the mean can be approximated as

$$E[X] \simeq \frac{1}{N}\sum_{i=1}^{N} x_i$$

where x_i are independent measurements of that variable. The expectation of the function $f(x)$ of a random variable is

$$E[f(x)] = \int_{-\infty}^{\infty} f(x)g(x)dx,$$

and it can be estimated using the Law of Large Numbers (Sect. 4.5):

$$E[f(x)] \simeq \frac{1}{N}\sum_{i=1}^{N} f(x_i). \tag{14.4}$$

These equations can be used to approximate the integral in (14.1) as a simple sum:

$$I = (b-a)\int_a^b f(x)g(x)dx = (b-a)E[f(x)] \simeq (b-a)\frac{1}{N}\sum_{i=1}^{N} f(x_i). \tag{14.5}$$

Equation (14.5) can be implemented by drawing N random uniform samples x_i from the support, then calculating $f(x_i)$, and evaluating the sum. This is the basic *Monte Carlo integration* method, and it can be easily implemented by using a random number generator available in most programming languages.

The method can be generalized to more than one dimension; if the support $A \subset \mathbb{R}^n$ has volume V, then the integration of an n-dimensional function $f(x)$ is given by the following sum:

$$I = \frac{V}{N} \sum_{i=1}^{N} f(x_i) \tag{14.6}$$

It is clear that the precision in the evaluation of the integral depends on the number of samples drawn. The error made by this method of integration can be estimated using the following interpretation of (14.6): the quantity $Vf(x)$ is the random variable of interest, and I is the expected value. Therefore, the variance of the random variable is given by the usual expression,

$$\sigma_I^2 = \frac{V^2}{N} \sum_{i=1}^{N} (f(x_i) - \bar{f})^2. \tag{14.7}$$

This means that the relative error in the calculation of the integral is

$$\frac{\sigma_I}{I} = \frac{1}{\sqrt{N}} \frac{\sqrt{\sum_{i=1}^{N} (f(x_i) - \bar{f})^2}}{\sum_{i=1}^{N} f(x_i)} \propto \frac{1}{\sqrt{N}}; \tag{14.8}$$

as expected, the relative error decreases like the square root of N, same as for a Poisson variable. Equation (14.8) can be used to determine how many samples are needed to estimate an integral with a given precision.

14.3 Dart Monte Carlo Integration and Function Evaluation

Another method to integrate a function, or to perform related mathematical operations, can be shown by way of an example. Assume that we want to measure the area of a circle of radius R. One can draw a random sample of N values in the (x, y) plane, as shown in Fig. 14.1, and count all the points that fall within the circle, $N(R)$. The area of the circle, or any other figure with known analytic function, is accordingly estimated as

$$A = \frac{N(R)}{N} \times V \tag{14.9}$$

in which V is the volume sampled by the two random variables. In the case of a circle of radius $R = 1$ we have $V = 4$, and since the known area is $A = \pi R^2$, this method provides an approximation to the number π.

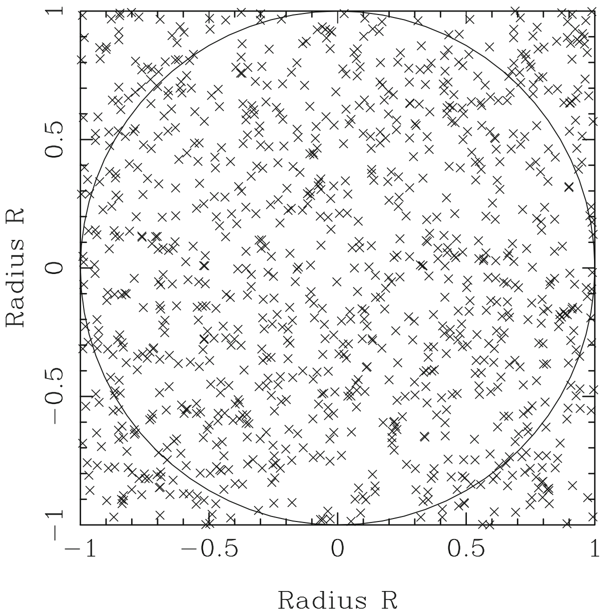

Fig. 14.1 Monte Carlo method to perform a calculation of the area of a *circle* (also a simulation of the number π), with $N = 1000$ iterations

Notice that (14.9) is equivalent to (14.6), in which the sum $\sum f(x_i)$ becomes $N(R)$, where $f(x_i) = 1$ indicates that a given random data point x_i falls within the boundaries of the figure of interest.

Example 14.1 (Simulation of the Number π) Figure 14.1 shows a Monte Carlo simulation of the number π, using 1000 random numbers drawn in a box of linear size 2, encompassing a circle of radius $R = 1$. The simulation has a number $N(R) = 772$ of points within the unit circle, resulting in an estimate of the area of the circle as $\pi R^2 = 0.772 \times 4 = 3.088$. Compared with the exact result of $\pi = 3.14159$, the simulation has an error of 1.7 %. According to (14.8), a 1000 iteration simulation has an expected relative error of order 3.1 %, therefore the specific simulation reported in Fig. 14.1 is consisted with the expected error, and more numbers must be drawn to improve the precision. ◇

14.4 Simulation of Random Variables

A method for the simulation of a random variable was discussed in Sect. 4.8. Since the generation of random samples from a uniform random variable was involved, this method also falls under the category of Monte Carlo simulations.

The method is based on (4.42):

$$X = F^{-1}(U),$$

in which F^{-1} represents the inverse of the cumulative distribution of the target variable X, and U represents a uniform random variable between 0 and 1. In Sect. 4.8 we provided the examples on how to use (4.42) to simulate an exponential distribution, which has a simple analytic function for its cumulative distribution.

The Gaussian distribution is perhaps the most common variable in many statistical applications, and its generation cannot be accomplished by (4.42), since the cumulative distribution is a special function and $F(x)$ does not have a close form. A method to overcome this limitation was discussed in Sect. 4.8.2, and it consists of using two uniform random variables U and V to simulate two standard Gaussians X and Y of zero mean and unit variance via (4.45),

$$\begin{cases} X = \sqrt{-2\ln(1-U)} \cdot \cos(2\pi V) \\ Y = \sqrt{-2\ln(1-U)} \cdot \sin(2\pi V). \end{cases} \quad (14.10)$$

A Gaussian X' of mean μ and variance σ^2 is related to the standard Gaussian X by the transformation

$$X = \frac{X' - \mu}{\sigma},$$

and therefore it can be simulated via

$$X' = \left(\sqrt{-2\ln(1-U)} \cdot \cos(2\pi V)\right)\sigma + \mu. \quad (14.11)$$

Figure 14.2 shows a simulation of a Gaussian distribution function using (14.11). Precision can be improved with increasing number of samples.

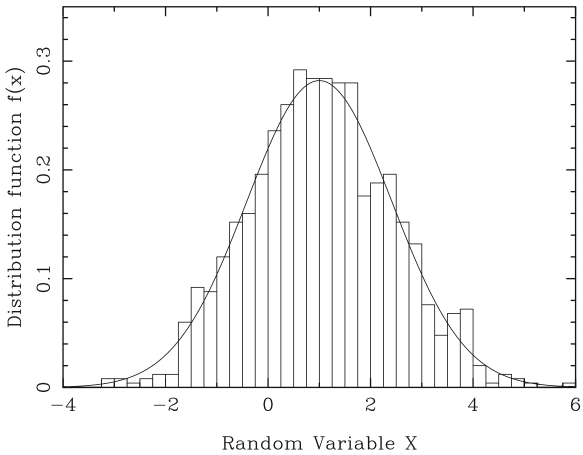

Fig. 14.2 Monte Carlo simulation of the probability distribution function of a Gaussian of $\mu = 1$ and $\sigma^2 = 2$ using 1000 samples according to (14.11)

14.5 Monte Carlo Estimates of Errors for Two-Variable Datasets

The two methods presented in this section, the bootstrap and the jackknife, are among the most common techniques to estimate best-fit parameters and their uncertainties in the fit to two-variable datasets. We have seen in previous chapters that the best-fit parameters and their uncertainties can be estimated analytically, for example, in the case of a linear regression with known errors in the dependent variable. In those cases, the exact analytical solution is typically the most straightforward to implement. When the analytic solution to a maximum likelihood fit is unavailable, then χ^2 minimization followed by the $\chi^2_{min} + \Delta\chi^2$ criterion can also be used to measure best-fit values and uncertainties in the parameters. Finally, Markov chain Monte Carlo methods to be presented in Chap. 16 can also be used in virtually any case for which the likelihood can be calculated.

The two methods presented in this section have a long history of use in statistical data analysis, and had been in use since well before the Markov chain Monte Carlo methods became of wide use. The bootstrap and jackknife methods are typically easier to implement than a Monte Carlo Markov chain. In particular, the bootstrap uses a large number of repetitions of the dataset, and therefore is computer intensive; the older jackknife method instead uses just a small number of additional random datasets, and requires less computing resources.

14.5.1 The Bootstrap Method

Consider a dataset Z composed of N measurements of either a random variable or, more generally, a pair of variables. The bootstrap method consists of generating as large a number of random, "synthetic" datasets based on the original set. Each set is then used to determine the distribution of the random variable (e.g., for the one-dimensional case) or of the best-fit parameters for the $y(x)$ model (for the two–dimensional case). The method has the following steps:

1. Draw at random N datapoints from the original set Z, with replacement, to form a synthetic dataset Z_i. The new dataset has therefore the same dimension as the original set, but a few of the original points may be repeated, and a few missing.
2. For each dataset Z_i, calculate the parameter(s) of interest a_i. For example, the parameters can be calculated using a χ^2 minimization technique.
3. Repeat this process as many times as possible, say N_{boot} times.
4. At the end of the process, the parameters a_n, $n = 1, \ldots, N_{boot}$, approximate the posterior distribution of the parameter of interest. These values can therefore be used to construct the sample distribution function for the parameters, and therefore obtain the best-fit value and confidence intervals.

Notice that one advantage of the bootstrap method is that it can be used even in cases in which the errors on the datapoints are not available, which is a very common occurrence. In this situation, the direct maximum likelihood method applied to the original set Z alone would not provide uncertainties in the best-fit parameters, as explained in Sect. 8.5. Since at each iteration the best-fit parameters alone must be evaluated, a dataset without errors in the dependent variable can still be fit to find the best-fit parameters, and the bootstrap method will provide an estimate of the uncertainties. This is one of the main reasons why the bootstrap method is so common.

Example 14.2 (Bootstrap Analysis of Hubble's Data) We perform a bootstrap analysis on the data from Hubble's experiment of page 157. The dataset Z consists of the ten measurements of the magnitude m and logarithm of the velocity $\log v$, as shown in Fig. 8.2. We generate 10,000 random synthetic datasets of ten measurements each, for which typically a few of the original datapoints are repeated. Given that error bars on the dependent variable $\log v$ were not given, we assume that the uncertainties have a common value for all measurement (and therefore the value of the error is irrelevant for the determination of the best-fit parameters). For each dataset Z_i we perform a linear regression to obtain the best-fit values of the parameters a_i and b_i.

The sample distributions of the parameters are shown in Fig. 14.3; from them, we can take the median of the distribution as the "best-fit" value for the parameter, and the 68 % confidence interval as the central range of each parameter that contains 68 % of the parameter occurrences. It is clear that both distributions are somewhat asymmetric; the situation does not improve with a larger number of bootstrap samples, since there is only a finite number of synthetic datasets that

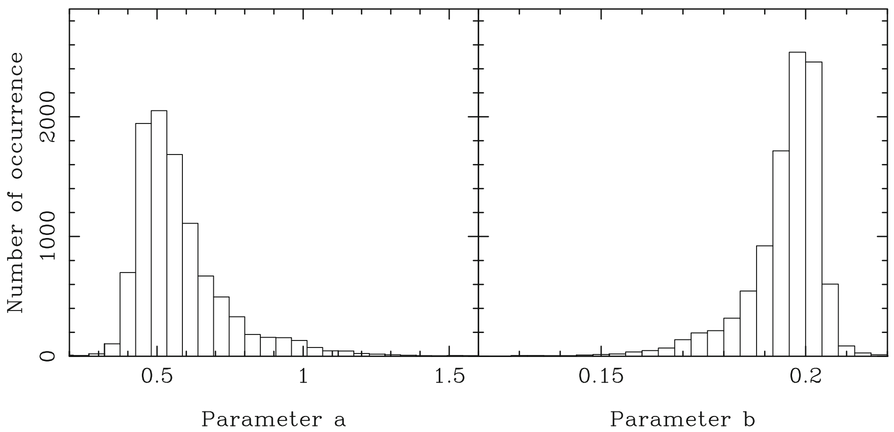

Fig. 14.3 Monte Carlo bootstrap method applied to the data from Hubble's experiment. (*Left*) Sample distribution of parameter a, with a median of $a = 0.54$ and a 68 % central range of 0.45–0.70. (*Right*) Distribution of b, with median $b = 0.197$ and a central range of 0.188–0.202. The best-fit values of the original dataset Z were found to be $a = 0.55$ and $b = 0.197$ (see page 159)

can be generated at random, with replacement, from the original dataset (see Problem 14.1). ◇

A key feature of the bootstrap method is that it is an unbiased estimator for the model parameters. We can easily prove this general property in the special case of a one-dimensional dataset, with the goal of estimating the sample mean and variance of the random variable X from N independent measurements. It is clear that we would normally not use the bootstrap method in this situation, since (2.8) and (5.4) provide the exact solution to the problem. The following proof is used to show that the bootstrap method provides unbiased estimates for the mean and variance of a random variable.

Proof The sample average calculated for a given bootstrap dataset Z_i is given by

$$\bar{x}_i = \frac{1}{N}\sum_{j=1}^{N} x_j n_{ji} \tag{14.12}$$

where n_{ji} is the number of occurrence of datapoint x_j in the synthetic set Z_i. If $n_{ji} = 0$ it means that x_j was not selected for the set, $n_{ji} = 1$ it means that there is just one occurrence of x_j (as in the original set), and so on. The number $n_{ji} \leq N$, and it is a random variable that is distributed like a binomial with $p = 1/N$, since the drawing for each bootstrap set is done at random, and with replacement. Therefore, we find that

$$\begin{cases} E[n_{ji}] = Np = 1 \\ Var(n_{ij}) \equiv \sigma_i^2 = Np(1-p) = \dfrac{N-1}{N} \end{cases} \tag{14.13}$$

where the expectation is calculated for a given dataset Z, drawing a large number of bootstrap sample based on that specific set. It follows that $\bar{x}_i$ is an unbiased estimator of the sample mean,

$$E[\bar{x}_i] = \frac{1}{N}\sum_{j=1}^{N} x_j E[n_{ji}] = \bar{x}. \tag{14.14}$$

The expectation operator used in the equation above relates to the way in which a specific synthetic dataset can be drawn, i.e., indicates an "average" over a specific dataset. The operation of expectation should also be repeated to average over all possible datasets Z consisting of N measurements of the random variable X, and that operation will also result in an expectation that is equal to the parent mean of X,

$$E[\bar{x}] = \mu. \tag{14.15}$$

Although we used the same symbol for the expectation of (14.14) and (14.15), the two operations are therefore different in nature.

The proof that the variance of the sample mean of dataset Z_i is an unbiased estimator of the parent variance σ^2/N is complicated by the fact that the random variables n_{ij} are not independent. In fact, they are related by

$$\sum_{i=1}^{N} n_{ij} = N, \tag{14.16}$$

and this enforces a negative correlation between the variables that vanishes only in the limit of very large N. It can be shown that the covariance of the n_{ij}'s (say, the covariance between n_{ij} and n_{ki}, were $i \neq k$, and i labels the dataset) is given by

$$\sigma^2_{jk} = -\frac{1}{N}. \tag{14.17}$$

The proof of (14.17) is left as an exercise, and it is based on the use of (14.16), and (4.3) (see Problem 14.2).

The variance of $\bar{x}_i$ can be calculated using (4.3), since $\bar{x}_i$ is a linear combination of N random variables n_{ij}:

$$Var(\bar{x}_i) = Var\left(\frac{1}{N}\sum_{j=1}^{N} x_j n_{ji}\right) =$$

$$\frac{1}{N^2}\left(\sum_{j=1}^{N} x_j^2 \sigma_i^2 + 2\sum_{j=1}^{N}\sum_{k=j+1}^{N} x_j x_k \sigma^2_{jk}\right) =$$

$$\frac{1}{N^2}\left(\frac{N-1}{N}\sum_{j=1}^{N} x_j^2 - \frac{2}{N}\sum_{j=1}^{N}\sum_{k=j+1}^{N} x_j x_k\right)$$

in which we have used the results of (14.13) and (14.17). Next, we need to calculate the expectation of this variance, in the sense of varying the dataset Z itself:

$$E[Var(\bar{x}_i)] = \frac{N-1}{N^3}E[\sum_{j=1}^{N} x_j^2] - \frac{2}{N^3}\left(\frac{1}{2}\sum_{j\neq k} E[x_j x_k]\right) \tag{14.18}$$

The last sum in the equation above is over all pairs (j, k); the factor 1/2 takes into account the double-counting of terms such as $x_j x_k$ and $x_k x_j$, and the sum contains a total of $N(N-1)$ identical terms. Since the measurements x_i, x_j are independent and identically distributed, $E[x_i x_k] = E[x_j]^2$, it follows that

$$E[Var(\bar{x}_i)] = \frac{N-1}{N^2}\left(E[x_i^2] - E[x_i]^2\right) = \frac{N-1}{N^2}\sigma^2 = \frac{N-1}{N}\sigma_\mu^2$$

where σ^2 is the variance of the random variable X, and $\sigma_\mu^2 = \sigma^2/N$ the variance of the sample mean. The equation states that $E[Var(\bar{x}_i)] = E[s^2]$, where s^2 is the sample variance of X. We showed in Sect. 5.1.2 that the sample variance is an unbiased estimator of the variance of the mean, provided it is multiplied by the known factor $N/(N-1)$. In practice, when calculating the variance from the N bootstrap samples, we should use the factor $1/(N-1)$ instead of $1/N$, as is normally done according to (5.6). □

14.5.2 The Jackknife Method

The jackknife method is an older Monte Carlo method that makes use of just N resampled datasets to estimate best-fit parameters and their uncertainties. As in the bootstrap method, we consider a dataset Z of N independent measurements either of a random variable X or of a pair of random variables. The method consists of the following steps:

1. Generate a resampled dataset Z_j by deleting the jth element from the dataset. This resampled dataset has therefore dimension $N-1$.
2. Each dataset Z_j is used to estimate the parameters of interest. For example, apply the linear regression method to dataset Z_j and find the best-fit values of the linear model, a_j and b_j.
3. The parameters of interest are also calculated from the full-dimensional dataset Z, as one normally would. The best-fit parameters are called $\hat{a}$.
4. For each dataset Z_j, define the *pseudo-values* $a_j^\star$ as

$$a_j^\star = N\hat{a} - (N-1)a_j \tag{14.19}$$

5. The jackknife estimate of each parameter of interest and its uncertainty are given by the following equations:

$$\begin{cases} a^\star = \dfrac{1}{N}\displaystyle\sum_{j=1}^{n} a_j^\star \\ \sigma_{a^\star}^2 = \dfrac{1}{N(N-1)}\displaystyle\sum_{j=1}^{N}(a_j^\star - a^\star)^2. \end{cases} \tag{14.20}$$

To prove that (14.20) provide an accurate estimate for the parameters and their errors, we apply them to the simple case of the estimate of the mean from a sample of

N measurements. In this case we want to show that the expectation of the jackknife estimate of the mean $a^{\star}$ is equal to the parent mean μ, and that the expectation of its variance $\sigma^2_{a^{\star}}$ is equal to σ^2/N.

Proof For a sample of N measurements of a random variable x, the sample mean and its variance are given by

$$\begin{cases} \bar{x} = \dfrac{1}{N}\sum_{j=1}^{N} x_i \\ \dfrac{s^2}{N} = \dfrac{1}{N(N-1)}\sum_{j=1}^{N}(x_i - \bar{x})^2. \end{cases} \tag{14.21}$$

The proof consists of showing that $a^{\star}_j = x_j$, so that $a^{\star}$ is the sample mean and $\sigma^2_{a^{\star}}$ is the sample variance. The result follows from:

$$a_j = \frac{1}{N-1}\sum_{i \neq j} x_i, \; \hat{a} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\Rightarrow a^{\star}_j = N\frac{1}{N}\sum_{i=1}^{N} x_i - \frac{N-1}{N-1}\sum_{i \neq j} x_i = x_j.$$

Notice that the factor of $1/(N-1)$ was used in the calculation of the sample variance, according to (5.6). □

Example 14.3 In the case of the Hubble experiment of page 157, we can use the jackknife method to estimate the best-fit parameters of the fit to a linear model of m versus $\log v$. According to (14.20), we find that $a^{\star} = 0.52$, $\sigma_{a^{\star}} = 0.13$, and $b^{\star} = 0.199$, $\sigma_{b^{\star}} = 0.008$. These estimates are in very good agreement with the results of the bootstrap method, and those of the direct fit to the original dataset for which, however, we could not provide uncertainties in the fit parameters. ◇

Summary of Key Concepts for this Chapter

- □ *Monte Carlo method*: Any numerical method that makes use of random variables to perform calculations that are too complex to be performed analytically, such as Monte Carlo integration and "dart" methods.
- □ *Bootstrap method*: A common method to estimate model parameters that uses a large number of synthetic datasets obtained by re-sampling of the original data.
- □ *Jackknife method*: A simple method to estimate model parameters that uses just N re-sampled datasets.

Problems

14.1 Calculate how many synthetic bootstrap datasets can be generated at random from a dataset Z with N unique datapoints. Notice that the order in which the datapoints appear in the dataset is irrelevant.

14.2 For a bootstrap dataset Z_j constructed from a set Z of N independent measurements of a variable X, show that the covariance between the number of occurrence n_{ji} and n_{jk} is given by (14.17),

$$\sigma_{ik}^2 = -\frac{1}{N}.$$

14.3 Perform a numerical simulation of the number π, and determine how many samples are sufficient to achieve a precision of 0.1 %. The first six significant digits of the number are $\pi = 3.14159$.

14.4 Perform a bootstrap simulation on the Hubble data presented in Fig. 14.3, and find the 68 % central confidence ranges on the parameters a and b.

14.5 Using the data of Problem 8.2, run a bootstrap simulation with $N = 1000$ iterations for the fit to a linear model. After completion of the simulation, plot the sample probability distribution function of the parameters a and b, and find the median and 68 % confidence intervals on the fit parameters. Describe the possible reason why the distribution of the fit parameters are not symmetric.

14.6 Use the data of Problem 8.2, but assuming that the errors in the dependent variable y are unknown. Run a bootstrap simulation with $N = 1000$ iterations, and determine the median and 68 % confidence intervals on the parameters a and b to the fit to a linear model.

14.7 Using the data of Problem 8.2, assuming that the errors in the dependent variable y are unknown, estimate the values of a and b to the fit to a linear model using a jackknife method.

14.8 Given two uniform random variables U_1 and U_2 between $-R$ and $+R$, as often available in common programming software, provide an analytic expression to simulate a Gaussian variable of mean μ and variance σ^2.