# 17

# Basic statistical models

In this chapter we introduce a common statistical model. It corresponds to the situation where the elements of the dataset are repeated measurements of the same quantity and where different measurements do not influence each other. Next, we discuss the probability distribution of the random variables that model the measurements and illustrate how *sample statistics* can help to select a suitable statistical model. Finally, we discuss the *simple linear regression model* that corresponds to the situation where the elements of the dataset are paired measurements.

## 17.1 Random samples and statistical models

In Chapter 1 we briefly discussed Michelson's experiment conducted between June 5 and July 2 in 1879, in which 100 measurements were obtained on the speed of light. The values are given in Table 17.1 and represent the speed of light in air in km/sec minus 299 000. The variation among the 100 values suggests that measuring the speed of light is subject to random influences. As we have seen before, we describe random phenomena by means of a probability model, i.e., we interpret the outcome of an experiment as a realization of some random variable. Hence the first measurement is modeled by a random variable $X_1$ and the value 850 is interpreted as the realization of $X_1$. Similarly, the second measurement is modeled by a random variable $X_2$ and the value 740 is interpreted as the realization of $X_2$. Since both measurements are obtained under the same experimental conditions, it is justified to assume that the probability distributions of $X_1$ and $X_2$ are the same. More generally, the 100 measurements are modeled by random variables

$$X_1, X_2, \ldots, X_{100}$$

with the same probability distribution, and the values in Table 17.1 are interpreted as realizations of $X_1, X_2, \ldots, X_{100}$. Moreover, because we believe that

**Table 17.1.** Michelson data on the speed of light.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 850 | 740 | 900 | 1070 | 930 | 850 | 950 | 980 | 980 | 880 |
| 1000 | 980 | 930 | 650 | 760 | 810 | 1000 | 1000 | 960 | 960 |
| 960 | 940 | 960 | 940 | 880 | 800 | 850 | 880 | 900 | 840 |
| 830 | 790 | 810 | 880 | 880 | 830 | 800 | 790 | 760 | 800 |
| 880 | 880 | 880 | 860 | 720 | 720 | 620 | 860 | 970 | 950 |
| 880 | 910 | 850 | 870 | 840 | 840 | 850 | 840 | 840 | 840 |
| 890 | 810 | 810 | 820 | 800 | 770 | 760 | 740 | 750 | 760 |
| 910 | 920 | 890 | 860 | 880 | 720 | 840 | 850 | 850 | 780 |
| 890 | 840 | 780 | 810 | 760 | 810 | 790 | 810 | 820 | 850 |
| 870 | 870 | 810 | 740 | 810 | 940 | 950 | 800 | 810 | 870 |

*Source:* E.N. Dorsey. The velocity of light. *Transactions of the American Philosophical Society.* 34(1):1-110, 1944; Table 22 on pages 60-61.

Michelson took great care not to have the measurements influence each other, the random variables $X_1, X_2, \ldots, X_{100}$ are assumed to be *mutually independent* (see also Remark 3.1 about physical and stochastic independence). Such a collection of random variables is called a random sample or briefly, sample.

RANDOM SAMPLE. A *random sample* is a collection of random variables $X_1, X_2, \ldots, X_n$, that have the same probability distribution and are mutually independent.

If $F$ is the distribution function of each random variable $X_i$ in a random sample, we speak of *a random sample from* $F$. Similarly, we speak of a random sample from a density $f$, a random sample from an $N(\mu, \sigma^2)$ distribution, etc.

QUICK EXERCISE 17.1 Suppose we have a random sample $X_1, X_2$ from a distribution with variance 1. Compute the variance of $X_1 + X_2$.

Properties that are inherent to the random phenomenon under study may provide additional knowledge about the distribution of the sample. Recall the software data discussed in Chapter 15. The data are observed lengths in CPU seconds between successive failures that occur during the execution of a certain real-time command. Typically, in a situation like this, in a small time interval, either 0 or 1 failure occurs. Moreover, failures occur with small probability and in disjoint time intervals failures occur independent of each other. In addition, let us assume that the rate at which the failures occur is constant over time. According to Chapter 12, this justifies the choice of a Poisson process to model the series of failures. From the properties of the Poisson process we know that the interfailure times are independent and have the same exponential distribution. Hence we model the software data as the realization of a random sample from an exponential distribution.

In some cases we may not be able to specify the type of distribution. Take, for instance, the Old Faithful data consisting of observed durations of eruptions of the Old Faithful geyser. Due to lack of specific geological knowledge about the subsurface and the mechanism that governs the eruptions, we prefer not to assume a particular type of distribution. However, we *do* model the durations as the realization of a random sample from a continuous distribution on $(0, \infty)$.

In each of the three examples the dataset was obtained from repeated measurements performed under the same experimental conditions. The basic statistical model for such a dataset is to consider the measurements as a random sample and to interpret the dataset as the realization of the random sample. Knowledge about the phenomenon under study and the nature of the experiment may lead to partial specification of the probability distribution of each $X_i$ in the sample. This should be included in the model.

STATISTICAL MODEL FOR REPEATED MEASUREMENTS. A dataset consisting of values $x_1, x_2, \ldots, x_n$ of repeated measurements of the same quantity is modeled as the realization of a random sample $X_1, X_2, \ldots, X_n$. The model may include a partial specification of the probability distribution of each $X_i$.

The probability distribution of each $X_i$ is called the *model distribution*. Usually it refers to a collection of distributions: in the Old Faithful example to the collection of all continuous distributions on $(0, \infty)$, in the software example to the collection of all exponential distributions. In the latter case the parameter of the exponential distribution is called the *model parameter*. The unique distribution from which the sample actually originates is assumed to be one particular member of this collection and is called the *"true" distribution*. Similarly, in the software example, the parameter corresponding to the "true" exponential distribution is called the *"true" parameter*. The word *true* is put between quotation marks because it does not refer to something in the real world, but only to a distribution (or parameter) in the statistical model, which is merely an approximation of the real situation.

QUICK EXERCISE 17.2 We obtain a dataset of ten elements by tossing a coin ten times and recording the result of each toss. What is an appropriate statistical model and corresponding model distribution for this dataset?

Of course there are situations where the assumption of *independence* or *identical distributions* is unrealistic. In that case a different statistical model would be more appropriate. However, we will restrict ourselves mainly to the case where the dataset can be modeled as the realization of a random sample.

Once we have formulated a statistical model for our dataset, we can use the dataset to infer knowledge about the model distribution. Important questions about the corresponding model distribution are

- *which feature of the model distribution* represents the quantity of interest and *how do we use our dataset* to determine a value for this?
- *which model distribution* fits a particular dataset best?

These questions can be diverse, and answering them may be difficult. For instance, the Old Faithful data are modeled as a realization of a random sample from a continuous distribution. Suppose we are interested in a complete characterization of the "true" distribution, such as the distribution function $F$ or the probability density $f$. Since there are no further specifications about the type of distribution, our problem would be to estimate the *complete curve* of $F$ or $f$ on the basis of our dataset.

On the other hand, the software data are modeled as the realization of a random sample from an exponential distribution. In that case $F$ and $f$ are completely characterized by a single parameter $\lambda$:

$$F(x) = 1 - \mathrm{e}^{-\lambda x} \quad \text{and} \quad f(x) = \lambda \mathrm{e}^{-\lambda x} \quad \text{for } x \geq 0.$$

Even if we are interested in the curves of $F$ and $f$, our problem would reduce to estimating a *single parameter* on the basis of our dataset.

In other cases we may not be interested in the distribution as a whole, but only in a specific feature of the model distribution that represents the quantity of interest. For instance, in a physical experiment, such as the one performed by Michelson, one usually thinks of each measurement as

$$\text{measurement} = \text{quantity of interest} + \text{measurement error}.$$

The quantity of interest, in this case the speed of light, is thought of as being some (unknown) constant and the measurement error is some random fluctuation. In the absence of systematic error, the measurement error can be modeled by a random variable with zero expectation and finite variance. In that case the measurements are modeled by a random sample from a distribution with some unknown expectation and finite variance. The speed of light is represented by the expectation of the model distribution. Our problem would be to estimate the *expectation of the model distribution* on the basis of our dataset.

In the remaining chapters, we will develop several statistical methods to infer knowledge about the "true" distribution or about a specific feature of it, by means of a dataset. In the remainder of this chapter we will investigate how the graphical and numerical summaries of our dataset can serve as a first indication of what an appropriate choice would be for this distribution or for a specific feature, such as its expectation.

## 17.2 Distribution features and sample statistics

In Chapters 15 and 16 we have discussed several empirical summaries of datasets. They are examples of numbers, curves, and other objects that are a

function

$$h(x_1, x_2, \ldots, x_n)$$

of the dataset $x_1, x_2, \ldots, x_n$ only. Since datasets are modeled as realizations of random samples $X_1, X_2, \ldots, X_n$, an object $h(x_1, x_2, \ldots, x_n)$ is a realization of the corresponding random object

$$h(X_1, X_2, \ldots, X_n).$$

Such an object, which depends on the random sample $X_1, X_2, \ldots, X_n$ only, is called a *sample statistic*.

If a statistical model adequately describes the dataset at hand, then the sample statistics corresponding to the empirical summaries should somehow reflect corresponding features of the model distribution. We have already seen a mathematical justification for this in Chapter 13 for the sample statistic

$$\bar{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n},$$

based on a sample $X_1, X_2, \ldots, X_n$ from a probability distribution with expectation $\mu$. According to the law of large numbers,

$$\lim_{n\to\infty} \mathrm{P}\big(|\bar{X}_n - \mu| > \varepsilon\big) = 0$$

for every $\varepsilon > 0$. This means that for large sample size $n$, the sample mean of most realizations of the random sample is close to the expectation of the corresponding distribution. In fact, all sample statistics discussed in Chapters 15 and 16 are close to corresponding distribution features. To illustrate this we generate an artificial dataset from a normal distribution with parameters $\mu = 5$ and $\sigma = 2$, using a technique similar to the one described in Section 6.2. Next, we compare the sample statistics with corresponding features of this distribution.

**The empirical distribution function**

Let $X_1, X_2, \ldots, X_n$ be a random sample from distribution function $F$, and let

$$F_n(a) = \frac{\text{number of } X_i \text{ in } (-\infty, a]}{n}$$

be the empirical distribution function of the sample. Another application of the law of large numbers (see Exercise 13.7) yields that for every $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathrm{P}(|F_n(a) - F(a)| > \varepsilon) = 0.$$

This means that for most realizations of the random sample the empirical distribution function $F_n$ is close to $F$:

$$F_n(a) \approx F(a).$$

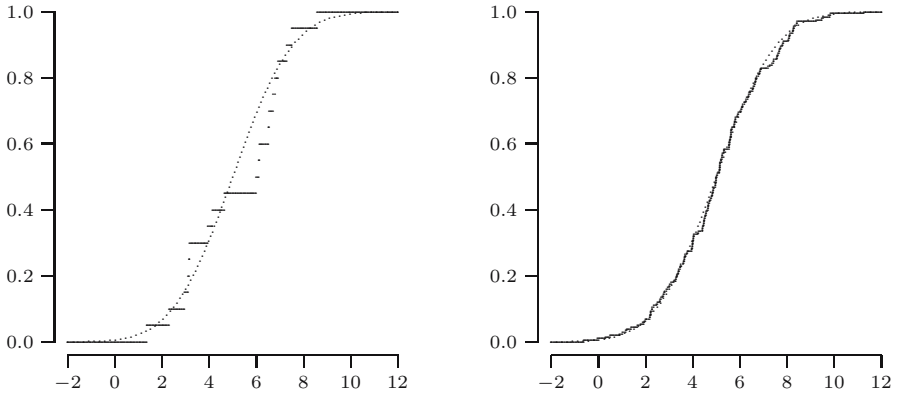


**Fig. 17.1.** Empirical distribution functions of normal samples.

Hence the empirical distribution function of the normal dataset should resemble the distribution function

$$F(a) = \int_{-\infty}^{a} \frac{1}{2\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}\left(\frac{x-5}{2}\right)^2} \,\mathrm{d}x$$

of the $N(5, 4)$ distribution, and the fit should become better as the sample size $n$ increases. An illustration of this can be found in Figure 17.1. We displayed the empirical distribution functions of datasets generated from an $N(5, 4)$ distribution together with the "true" distribution function $F$ (dotted lines), for sample sizes $n = 20$ (left) and $n = 200$ (right).

### The histogram and the kernel density estimate

Suppose the random sample $X_1, X_2, \ldots, X_n$ is generated from a continuous distribution with probability density $f$. In Section 13.4 we have seen yet another consequence of the law of large numbers:

$$\frac{\text{number of } X_i \text{ in } (x-h, x+h]}{2hn} \;\approx\; f(x).$$

When $(x-h, x+h]$ is a bin of a histogram of the random sample, this means that the height of the histogram approximates the value of $f$ at the midpoint of the bin:

$$\text{height of the histogram on } (x-h, x+h] \;\approx\; f(x).$$

Similarly, the kernel density estimate of a random sample approximates the corresponding probability density $f$:

$$f_{n,h}(x) \approx f(x).$$

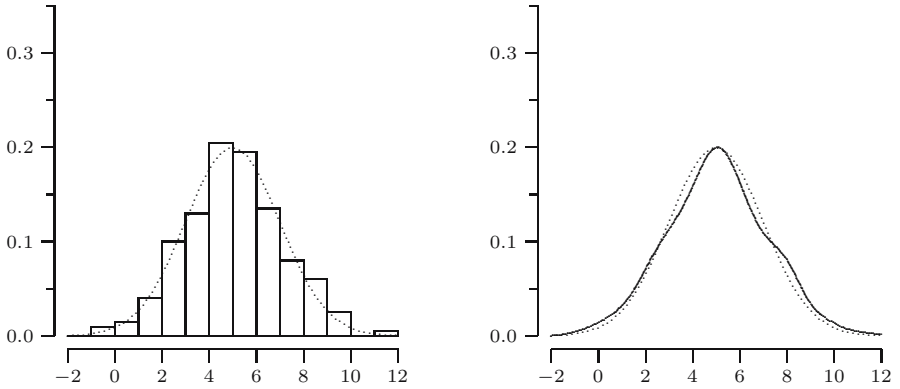


**Fig. 17.2.** Histogram and kernel density estimate of a sample of size 200.

So the histogram and kernel density estimate of the normal dataset should resemble the graph of the probability density

$$f(x) = \frac{1}{2\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}\left(\frac{x-5}{2}\right)^2}$$

of the $N(5, 4)$ distribution. This is illustrated in Figure 17.2, where we displayed a histogram and a kernel density estimate of our dataset consisting of 200 values generated from the $N(5, 4)$ distribution. It should be noted that with a smaller dataset the similarity can be much worse. This is demonstrated in Figure 17.3, which is based on the dataset consisting of 20 values generated from the same distribution.

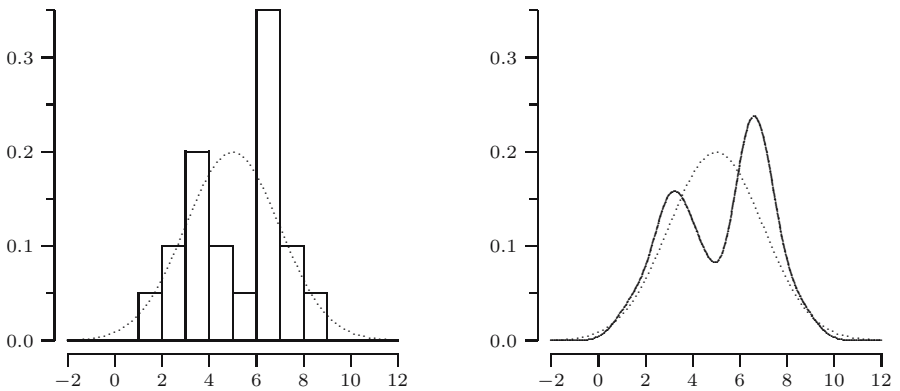


**Fig. 17.3.** Histogram and kernel density estimate of a sample of size 20.

**Remark 17.1 (About the approximations).** Let $H_n$ be the height of the histogram on the interval $(x-h, x+h]$, which is assumed to be a bin of the histogram. Direct application of the law of large numbers merely yields that $H_n$ converges to

$$\frac{1}{2h}\int_{x-h}^{x+h} f(u)\,\mathrm{d}u.$$

Only for small $h$ this is close to $f(x)$. However, if we let $h$ tend to 0 as $n$ increases, a variation on the law of large numbers will guarantee that $H_n$ converges to $f(x)$: for every $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathrm{P}(|H_n - f(x)| > \varepsilon) = 0.$$

A possible choice is the optimal bin width mentioned in Remark 15.1. Similarly, direct application of the law of large numbers yields that a kernel density estimator with fixed bandwidth $h$ converges to

$$\int_{-\infty}^{\infty} f(x+hu)K(u)\,\mathrm{d}u.$$

Once more, only for small $h$ this is close to $f(x)$, provided that $K$ is symmetric and integrates to one. However, by letting the bandwidth $h$ tend to 0 as $n$ increases, yet another variation on the law of large numbers will guarantee that $f_{n,h}(x)$ converges to $f(x)$: for every $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathrm{P}(|f_{n,h}(x) - f(x)| > \varepsilon) = 0.$$

A possible choice is the optimal bandwidth mentioned in Remark 15.2.

### The sample mean, the sample median, and empirical quantiles

As we saw in Section 5.5, the expectation of an $N(\mu, \sigma^2)$ distribution is $\mu$; so the $N(5, 4)$ distribution has expectation 5. According to the law of large numbers: $\bar{X}_n \approx \mu$. This is illustrated by our dataset of 200 values generated from the $N(5, 4)$ distribution for which we find

$$\bar{x}_{200} = 5.012.$$

For the sample median we find

$$\mathrm{Med}(x_1, \ldots, x_{200}) = 5.018.$$

This illustrates the fact that the sample median of a random sample from $F$ approximates the median $q_{0.5} = F^{\mathrm{inv}}(0.5)$. In fact, we have the following general property for the $p$th empirical quantile:

$$q_n(p) \approx F^{\mathrm{inv}}(p) = q_p.$$

In the special case of the $N(\mu, \sigma^2)$ distribution, the expectation and the median coincide, which explains why the sample mean and sample median of the normal dataset are so close to each other.

**The sample variance and standard deviation, and the MAD**

As we saw in Section 5.5, the standard deviation and variance of an $N(\mu, \sigma^2)$ distribution are $\sigma$ and $\sigma^2$; so for the $N(5, 4)$ distribution these are 2 and 4. Another consequence of the law of large numbers is that

$$S_n^2 \approx \sigma^2 \quad \text{and} \quad S_n \approx \sigma.$$

This is illustrated by our normal dataset of size 200, for which we find

$$s_{200}^2 = 4.761 \quad \text{and} \quad s_{200} = 2.182$$

for the sample variance and sample standard deviation.

For the MAD of the dataset we find 1.334, which clearly differs from the standard deviation 2 of the $N(5, 4)$ distribution. The reason is that

$$\text{MAD}(X_1, X_2, \ldots, X_n) \approx F^{\text{inv}}(0.75) - F^{\text{inv}}(0.5),$$

for any distribution that is symmetric around its median $F^{\text{inv}}(0.5)$. For the $N(5, 4)$ distribution $F^{\text{inv}}(0.75) - F^{\text{inv}}(0.5) = 2\Phi^{\text{inv}}(0.75) = 1.3490$, where $\Phi$ denotes the distribution function of the standard normal distribution (see Exercise 17.10).

**Relative frequencies**

For continuous distributions the histogram and kernel density estimates of a random sample approximate the corresponding probability density $f$. For discrete distributions we would like to have a sample statistic that approximates the probability mass function. In Section 13.4 we saw that, as a consequence of the law of large numbers, relative frequencies based on a random sample approximate corresponding probabilities. As a special case, for a random sample $X_1, X_2, \ldots, X_n$ from a discrete distribution with probability mass function $p$, one has that

$$\frac{\text{number of } X_i \text{ equal to } a}{n} \approx p(a).$$

This means that the relative frequency of $a$'s in the sample approximates the value of the probability mass function at $a$. Table 17.2 lists the sample statistics and the corresponding distribution features they approximate.

## 17.3 Estimating features of the "true" distribution

In the previous section we generated a dataset of 200 elements from a probability distribution, and we have seen that certain features of this distribution are approximated by corresponding sample statistics. In practice, the situation is reversed. In that case we have a dataset of $n$ elements that is modeled as the realization of a random sample with a probability distribution that is unknown to us. Our goal is to use our dataset to estimate a certain feature of this distribution that represents the quantity of interest. In this section we will discuss a few examples.

**Table 17.2.** Some sample statistics and corresponding distribution features.

| Sample statistic | Distribution feature |
|---|---|
| **Graphical** | |
| Empirical distribution function $F_n$ | Distribution function $F$ |
| Kernel density estimate $f_{n,h}$ and histogram | Probability density $f$ |
| (Number of $X_i$ equal to $a$)/$n$ | Probability mass function $p(a)$ |
| **Numerical** | |
| Sample mean $\bar{X}_n$ | Expectation $\mu$ |
| Sample median $\text{Med}(X_1, X_2, \ldots, X_n)$ | Median $q_{0.5} = F^{\text{inv}}(0.5)$ |
| $p$th empirical quantile $q_n(p)$ | $100p$th percentile $q_p = F^{\text{inv}}(p)$ |
| Sample variance $S_n^2$ | Variance $\sigma^2$ |
| Sample standard deviation $S_n$ | Standard deviation $\sigma$ |
| $\text{MAD}(X_1, X_2, \ldots, X_n)$ | $F^{\text{inv}}(0.75) - F^{\text{inv}}(0.5)$, for symmetric $F$ |

### The Old Faithful data

We stick to the assumptions of Section 17.1: by lack of knowledge on this phenomenon we prefer not to specify a particular parametric type of distribution, and we model the Old Faithful data as the realization of a random sample of size 272 from a continuous probability distribution. From the previous section we know that the kernel density estimate and the empirical distribution function of the dataset approximate the probability density $f$ and the distribution function $F$ of this distribution. In Figure 17.4 a kernel density estimate (left) and the empirical distribution function (right) are displayed. Indeed, neither graph resembles the probability density function or distribution function of any of the familiar parametric distributions. Instead of viewing both graphs

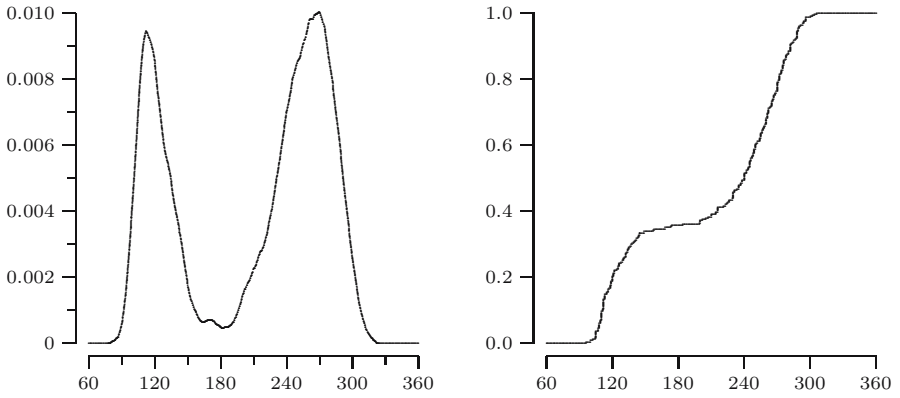


**Fig. 17.4.** Nonparametric estimates for $f$ and $F$ based on the Old Faithful data.

only as graphical summaries of the data, we can also use both curves as estimates for $f$ and $F$. We estimate the model probability density $f$ by means of the kernel density estimate and the model distribution function $F$ by means of the empirical distribution function. Since neither estimate assumes a particular parametric model, they are called *nonparametric* estimates.

### The software data

Next consider the software reliability data. As motivated in Section 17.1, we model interfailure times as the realization of a random sample from an exponential distribution. To see whether an exponential distribution is indeed a reasonable model, we plot a histogram and a kernel density estimate using a boundary kernel in Figure 17.5.

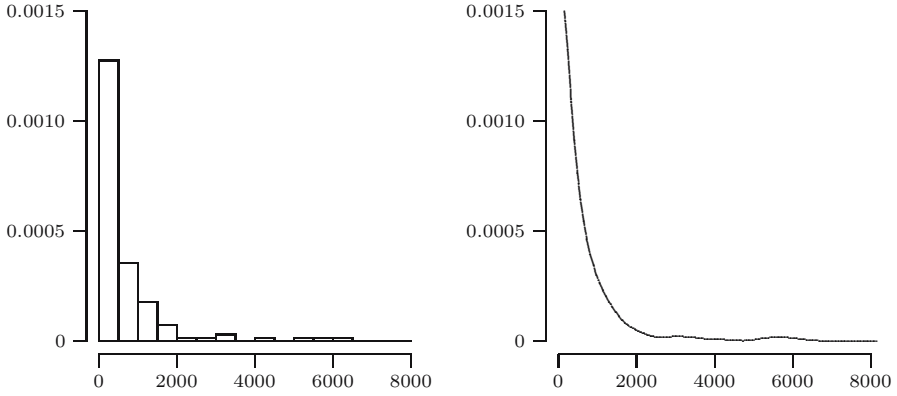


**Fig. 17.5.** Histogram and kernel density estimate for the software data.

Both seem to corroborate the assumption of an exponential distribution. Accepting this, we are left with estimating the parameter $\lambda$. Because for the exponential distribution $\mathrm{E}[X] = 1/\lambda$, the law of large numbers suggests $1/\bar{x}$ as an estimate for $\lambda$. For our dataset $\bar{x} = 656.88$, which yields $1/\bar{x} = 0.0015$. In Figure 17.6 we compare the estimated exponential density (left) and distribution function (right) with the corresponding nonparametric estimates. Note that the nonparametric estimates do *not* assume an exponential model for the data. But, *if* an exponential distribution were the right model, the kernel density estimate and empirical distribution function should resemble the estimated exponential density and distribution function. At first sight the fit seems reasonable, although near zero the data accumulate more than one might perhaps expect for a sample of size 135 from an exponential distribution, and the other way around at the other end of the data range. The question is whether this phenomenon can be attributed to chance or is caused by the fact that the exponential model is the wrong model. We will return to this type of question in Chapter 25 (see also Chapter 18).

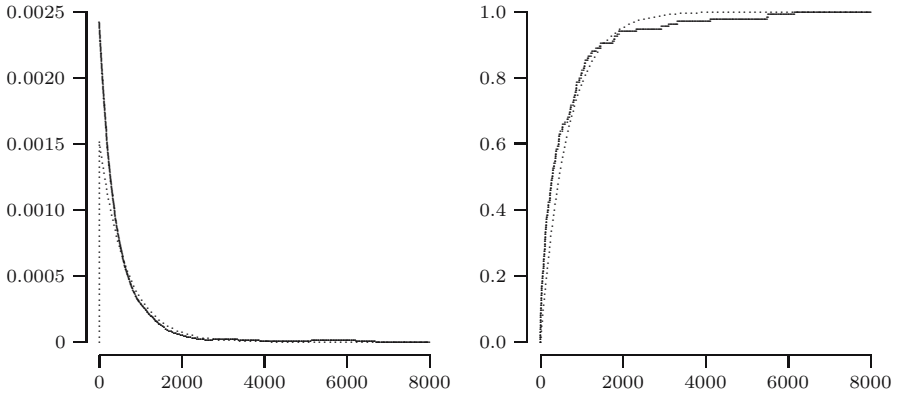


**Fig. 17.6.** Kernel density estimate and empirical cdf for software data (solid) compared to $f$ and $F$ of the estimated exponential distribution.

### Michelson data

Consider the Michelson data on the speed of light. In this case we are not particularly interested in estimation of the "true" distribution, but solely in the expectation of this distribution, which represents the speed of light. The law of large numbers suggests to estimate the expectation by the sample mean $\bar{x}$, which equals 852.4.

## 17.4 The linear regression model

Recall the example about predicting Janka hardness of wood from the density of the wood in Section 15.5. The idea is, of course, that Janka hardness is related to the density: the higher the density of the wood, the higher the value of Janka hardness. This suggests a relationship of the type

$$\text{hardness} = g(\text{density of timber})$$

for some increasing function $g$. This is supported by the scatterplot of the data in Figure 17.7. A closer look at the bivariate dataset in Table 15.5 suggests that randomness is also involved. For instance, for the value 51.5 of the density, different corresponding values of Janka hardness were observed. One way to model such a situation is by means of a *regression model*:

$$\text{hardness} = g(\text{density of timber}) + \text{random fluctuation}.$$

The important question now is *what sort of function* $g$ fits well to the points in the scatterplot?

In general, this may be a difficult question to answer. We may have so little knowledge about the phenomenon under study, and the data points may be

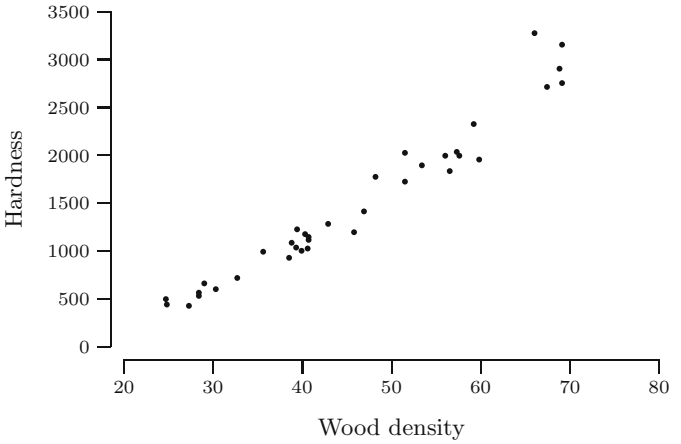


**Fig. 17.7.** Scatterplot of Janka hardness versus wood density.

scattered in such a way, that there is no reason to assume a specific type of function for $g$. However, for the Janka hardness data it makes sense to assume that $g$ is increasing, but this still leaves us with many possibilities. Looking at the scatterplot, at first sight it does not seem unreasonable to assume that $g$ is a straight line, i.e., Janka hardness depends linearly on the density of timber. The fact that the points are not exactly on a straight line is then modeled by a random fluctuation with respect to the straight line:

$$\text{hardness} = \alpha + \beta \cdot (\text{density of timber}) + \text{random fluctuation}.$$

This is a loose description of a simple linear regression model. A more complete description is given below.

SIMPLE LINEAR REGRESSION MODEL. In a *simple linear regression model* for a bivariate dataset $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, we assume that $x_1, x_2, \ldots, x_n$ are nonrandom and that $y_1, y_2, \ldots, y_n$ are realizations of random variables $Y_1, Y_2, \ldots, Y_n$ satisfying

$$Y_i = \alpha + \beta x_i + U_i \quad \text{for } i = 1, 2, \ldots, n,$$

where $U_1, \ldots, U_n$ are *independent* random variables with $\mathrm{E}[U_i] = 0$ and $\mathrm{Var}(U_i) = \sigma^2$.

The line $y = \alpha + \beta x$ is called the *regression line*. The parameters $\alpha$ and $\beta$ represent the *intercept* and *slope* of the regression line. Usually, the $x$-variable is called the *explanatory variable* and the $y$-variable is called the *response variable*. One also refers to $x$ and $y$ as *independent* and *dependent* variables. The random variables $U_1, U_2, \ldots, U_n$ are assumed to be independent when the different measurements do not influence each other. They are assumed to have

expectation zero, because the random fluctuation is considered to be around the regression line $y = \alpha + \beta x$. Finally, because each random fluctuation is supposed to have the same amount of variability, we assume that all $U_i$ have the same variance. Note that by the propagation of independence rule in Section 9.4, independence of the $U_i$ implies independence of $Y_i$. However, $Y_1, Y_2, \ldots, Y_n$ *do not* form a random sample. Indeed, the $Y_i$ have different distributions because every $Y_i$ has a different expectation

$$\mathrm{E}[Y_i] = \mathrm{E}[\alpha + \beta x_i + U_i] = \alpha + \beta x_i + \mathrm{E}[U_i] = \alpha + \beta x_i.$$

QUICK EXERCISE 17.3 Consider the simple linear regression model as defined earlier. Compute the variance of $Y_i$.

The parameters $\alpha$ and $\beta$ are unknown and our task will be to estimate them on the basis of the data. We will come back to this in Chapter 22. In Figure 17.8 the scatterplot for the Janka hardness data is displayed with the estimated

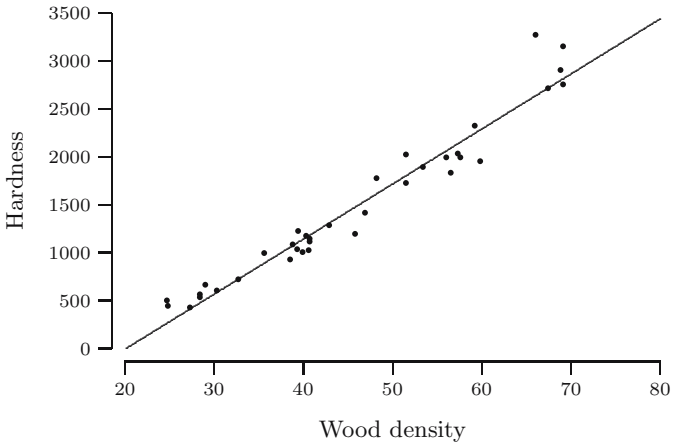


**Fig. 17.8.** Estimated regression line for the Janka hardness data.

regression line

$$y = -1160.5 + 57.51x.$$

Taking a closer look at Figure 17.8, you might wonder whether

$$y = \alpha + \beta x + \gamma x^2$$

would be a more appropriate model. By trying to answer this question we enter the area of *multiple* linear regression. We will not pursue this topic; we restrict ourselves to *simple* linear regression.

## 17.5 Solutions to the quick exercises

**17.1** Because $X_1, X_2$ form a random sample, they are independent. Using the rule about the variance of the sum of independent random variables, this means that $\text{Var}(X_1 + X_2) = \text{Var}(X_1) + \text{Var}(X_2) = 1 + 1 = 2$.

**17.2** The result of each toss of a coin can be modeled by a Bernoulli random variable taking values 1 (heads) and 0 (tails). In the case when it is known that we are tossing a *fair* coin, heads and tails occur with equal probability. Since it is reasonable to assume that the tosses do not influence each other, the outcomes of the ten tosses are modeled as the realization of a random sample $X_1, \ldots, X_{10}$ from a Bernoulli distribution with parameter $p = 1/2$. In this case the model distribution is completely specified and coincides with the "true" distribution: a $Ber(\frac{1}{2})$ distribution.

In the case when we are dealing with a *possibly unfair* coin, the outcomes of the ten tosses are still modeled as the realization of a random sample $X_1, \ldots, X_{10}$ from a Bernoulli distribution, but we cannot specify the value of the parameter $p$. The model distribution is a Bernoulli distribution. The "true" distribution is a Bernoulli distribution with one particular value for $p$, unknown to us.

**17.3** Note that the $x_i$ are considered nonrandom. By the rules for the variance, we find $\text{Var}(Y_i) = \text{Var}(\alpha + \beta x_i + U_i) = \text{Var}(U_i) = \sigma^2$.

## 17.6 Exercises

**17.1** ⊡ Figure 17.9 displays several histograms, kernel density estimates, and empirical distribution functions. It is known that all figures correspond to datasets of size 200 that are generated from normal distributions $N(0, 1)$, $N(0, 9)$, and $N(3, 1)$, and from exponential distributions $Exp(1)$ and $Exp(1/3)$. Report for each figure from which distribution the dataset has been generated.

**17.2** ⊡ Figure 17.10 displays several boxplots. It is known that all figures correspond to datasets of size 200 that are generated from the same five distributions as in Exercise 17.1. Report for each boxplot from which distribution the dataset has been generated.

**17.3** ⊞ At a London underground station, the number of women was counted in each of 100 queues of length 10. In this way a dataset $x_1, x_2, \ldots, x_{100}$ was obtained, where $x_i$ denotes the observed number of women in the $i$th queue. The dataset is summarized in the following table and lists the number of queues with 0 women, 1 woman, 2 women, etc.

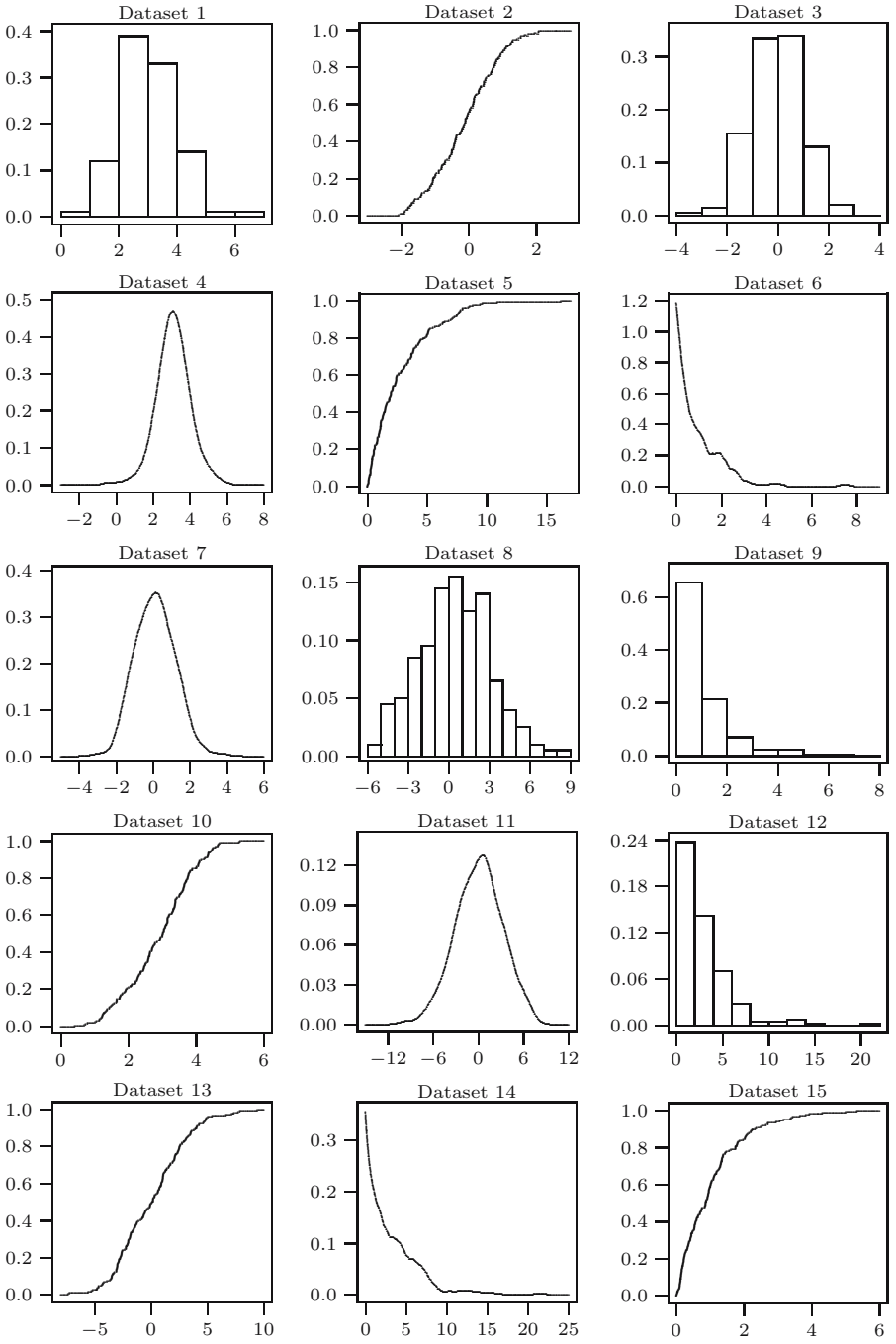


**Fig. 17.9.** Graphical representations of different datasets from Exercise 17.1.

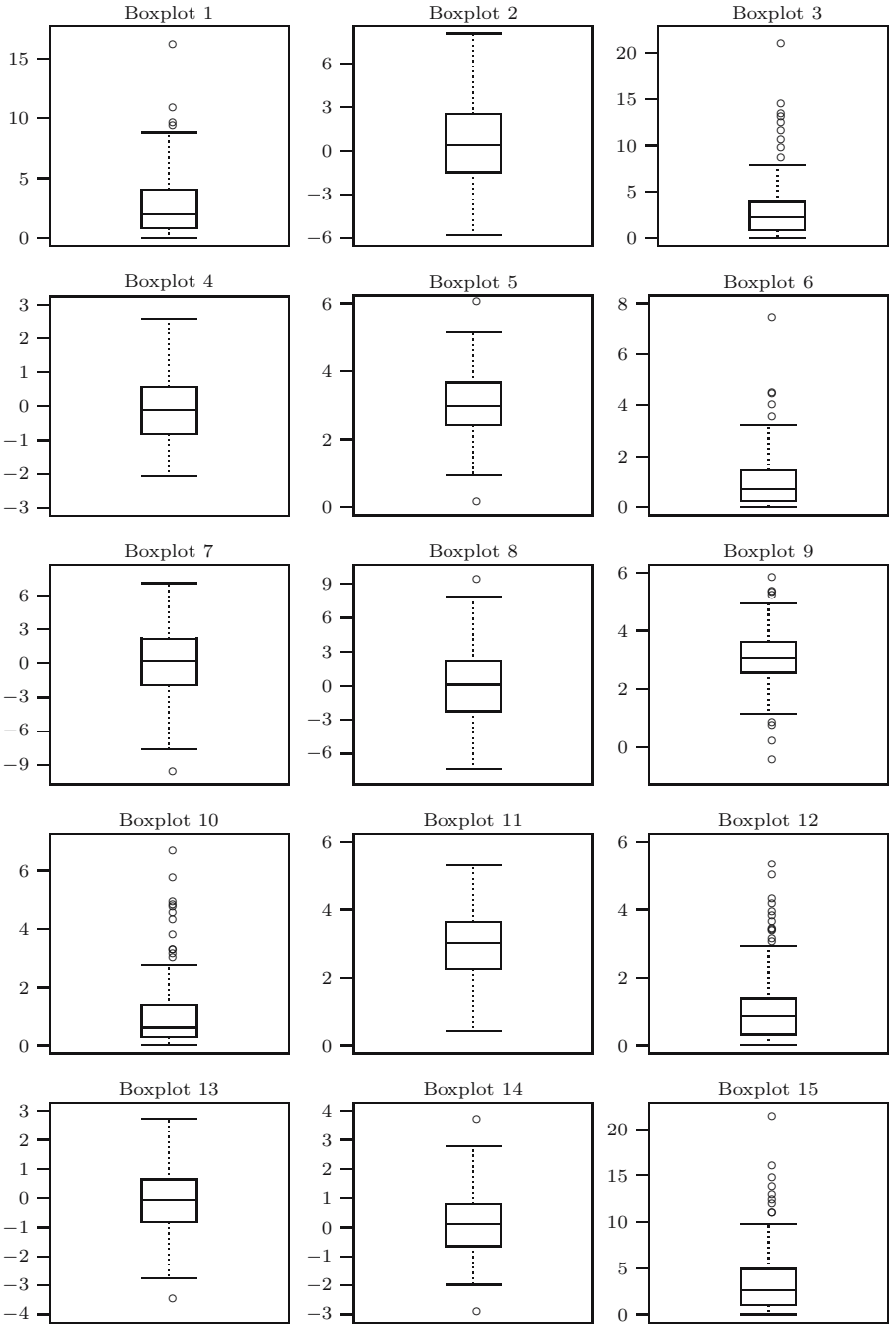


**Fig. 17.10.** Boxplot of different datasets from Exercise 17.2.

| Count | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Frequency | 1 | 3 | 4 | 23 | 25 | 19 | 18 | 5 | 1 | 1 | 0 |

*Source:* R.A. Jinkinson and M. Slater. Critical discussion of a graphical method for identifying discrete distributions. *The Statistician*, 30:239–248, 1981; Table 1 on page 240.

In the statistical model for this dataset, we assume that the observed counts are a realization of a random sample $X_1, X_2, \ldots, X_{100}$.

**a.** Assume that people line up in such a way that a man or woman in a certain position is independent of the other positions, and that in each position one has a woman with equal probability. What is an appropriate choice for the model distribution?

**b.** Use the table to find an estimate for the parameter(s) of the model distribution chosen in part **a**.

**17.4** During the Second World War, London was hit by numerous flying bombs. The following data are from an area in South London of 36 square kilometers. The area was divided into 576 squares with sides of length 1/4 kilometer. For each of the 576 squares the number of hits was recorded. In this way we obtain a dataset $x_1, x_2, \ldots, x_{576}$, where $x_i$ denotes the number of hits in the $i$th square. The data are summarized in the following table which lists the number of squares with no hits, 1 hit, 2 hits, etc.

| Number of hits | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|---|
| Number of squares | 229 | 211 | 93 | 35 | 7 | 0 | 0 | 1 |

*Source:* R.D. Clarke. An application of the Poisson distribution. *Journal of the Institute of Actuaries*, 72:48, 1946; Table 1 on page 481. © Faculty and Institute of Actuaries.

An interesting question is whether London was hit in a completely random manner. In that case a Poisson distribution should fit the data.

**a.** If we model the dataset as the realization of a random sample from a Poisson distribution with parameter $\mu$, then what would you choose as an estimate for $\mu$?

**b.** Check the fit with a Poisson distribution by comparing some of the observed relative frequencies of 0's, 1's, 2's, etc., with the corresponding probabilities for the Poisson distribution with $\mu$ estimated as in part **a**.

**17.5** ⊡ We return to the example concerning the number of menstrual cycles up to pregnancy, where the number of cycles was modeled by a geometric random variable (see Section 4.4). The original data concerned 100 smoking and 486 nonsmoking women. For 7 smokers and 12 nonsmokers, the exact number of cycles up to pregnancy was unknown. In the following tables we only

incorporated the 93 smokers and 474 nonsmokers, for which the exact number of cycles was observed. Another analysis, based on the complete dataset, is done in Section 21.1.

**a.** Consider the dataset $x_1, x_2, \ldots, x_{93}$ corresponding to the smoking women, where $x_i$ denotes the number of cycles for the $i$th smoking woman. The data are summarized in the following table.

| Cycles | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Frequency | 29 | 16 | 17 | 4 | 3 | 9 | 4 | 5 | 1 | 1 | 1 | 3 |

*Source:* C.R. Weinberg and B.C. Gladen. The beta-geometric distribution applied to comparative fecundability studies. *Biometrics*, 42(3):547–560, 1986.

The table lists the number of women that had to wait 1 cycle, 2 cycles, etc. If we model the dataset as the realization of a random sample from a geometric distribution with parameter $p$, then what would you choose as an estimate for $p$?

**b.** Also estimate the parameter $p$ for the 474 nonsmoking women, which is also modeled as the realization of a random sample from a geometric distribution. The dataset $y_1, y_2, \ldots, y_{474}$, where $y_j$ denotes the number of cycles for the $j$th nonsmoking woman, is summarized here:

| Cycles | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Frequency | 198 | 107 | 55 | 38 | 18 | 22 | 7 | 9 | 5 | 3 | 6 | 6 |

*Source:* C.R. Weinberg and B.C. Gladen. The beta-geometric distribution applied to comparative fecundability studies. *Biometrics*, 42(3):547–560, 1986.

You may use that $y_1 + y_2 + \cdots + y_{474} = 1285$.

**c.** Compare the estimates of the probability of becoming pregnant in three or fewer cycles for smoking and nonsmoking women.

**17.6** Recall Exercise 15.1 about the chest circumference of 5732 Scottish soldiers, where we constructed the histogram displayed in Figure 17.11. The histogram suggests modeling the data as the realization of a random sample from a normal distribution.

**a.** Suppose that for the dataset $\sum x_i = 228377.2$ and $\sum x_i^2 = 9124064$. What would you choose as estimates for the parameters $\mu$ and $\sigma$ of the $N(\mu, \sigma^2)$ distribution?
*Hint:* you may want to use the relation from Exercise 16.15.

**b.** Give an estimate for the probability that a Scottish soldier has a chest circumference between 38.5 and 42.5 inches.

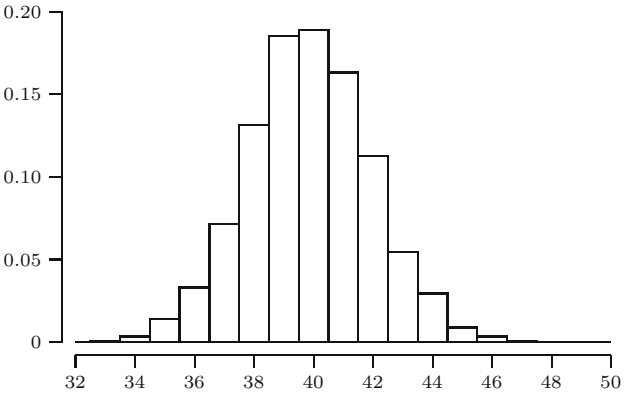


**Fig. 17.11.** Histogram of chest circumferences.

**17.7** ⊞ Recall Exercise 15.3 about time intervals between successive coal mine disasters. Let us assume that the rate at which the disasters occur is constant over time and that on a single day a disaster takes place with small probability independently of what happens on other days. According to Chapter 12 this suggests modeling the series of disasters with a Poisson process. Figure 17.12 displays a histogram and empirical distribution function of the observed time intervals.

**a.** In the statistical model for this dataset we model the 190 time intervals as the realization of a random sample. What would you choose for the model distribution?

**b.** The sum of the observed time intervals is 40 549 days. Give an estimate for the parameter(s) of the distribution chosen in part **a**.

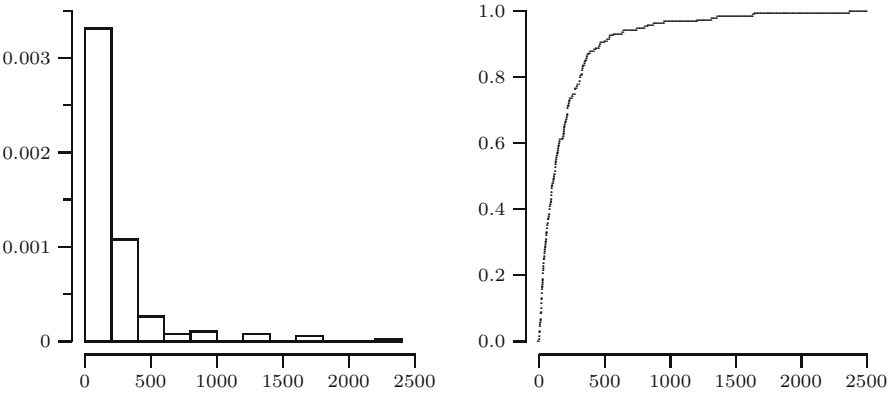


**Fig. 17.12.** Histogram of time intervals between successive disasters.

**17.8** The following data represent the number of revolutions to failure (in millions) of 22 deep-groove ball-bearings.

| | | | | |
|---|---|---|---|---|
| 17.88 | 28.92 | 33.00 | 41.52 | 42.12 |
| 45.60 | 48.48 | 51.84 | 51.96 | 54.12 |
| 55.56 | 67.80 | 68.64 | 68.88 | 84.12 |
| 93.12 | 98.64 | 105.12 | 105.84 | 127.92 |
| 128.04 | 173.40 | | | |

*Source:* J. Lieblein and M. Zelen. Statistical investigation of the fatigue-life of deep-groove ball-bearings. *Journal of Research, National Bureau of Standards*, 57:273–316, 1956; specimen worksheet on page 286.

Lieblein and Zelen propose modeling the dataset as a realization of a random sample from a Weibull distribution, which has distribution function

$$F(x) = 1 - \mathrm{e}^{-(\lambda x)^{\alpha}} \quad \text{for } x \geq 0,$$

and $F(x) = 0$, for $x < 0$, where $\alpha, \lambda > 0$.

**a.** Suppose that $X$ is a random variable with a Weibull distribution. Check that the random variable $Y = X^{\alpha}$ has an exponential distribution with parameter $\lambda^{\alpha}$ and conclude that $\mathrm{E}[X^{\alpha}] = 1/\lambda^{\alpha}$.

**b.** Use part **a** to explain how one can use the data in the table to find an estimate for the parameter $\lambda$, if it is given that the parameter $\alpha$ is estimated by 2.102.

**17.9** ⊞ The volume (i.e., the effective wood production in cubic meters), height (in meters), and diameter (in meters) (measured at 1.37 meter above the ground) are recorded for 31 black cherry trees in the Allegheny National Forest in Pennsylvania. The data are listed in Table 17.3. They were collected to find an estimate for the volume of a tree (and therefore for the timber yield), given its height and diameter. For each tree the volume $y$ and the value of $x = d^2h$ are recorded, where $d$ and $h$ are the diameter and height of the tree. The resulting points $(x_1, y_1), \ldots, (x_{31}, y_{31})$ are displayed in the scatterplot in Figure 17.13.

We model the data by the following linear regression model (without intercept)

$$Y_i = \beta x_i + U_i$$

for $i = 1, 2, \ldots, 31$.

**a.** What physical reasons justify the linear relationship between $y$ and $d^2h$? *Hint:* how does the volume of a cylinder relate to its diameter and height?

**b.** We want to find an estimate for the slope $\beta$ of the line $y = \beta x$. Two natural candidates are the average slope $\bar{z}_n$, where $z_i = y_i/x_i$, and the

**Table 17.3.** Measurements on black cherry trees.

| Diameter | Height | Volume |
|---|---|---|
| 0.21 | 21.3 | 0.29 |
| 0.22 | 19.8 | 0.29 |
| 0.22 | 19.2 | 0.29 |
| 0.27 | 21.9 | 0.46 |
| 0.27 | 24.7 | 0.53 |
| 0.27 | 25.3 | 0.56 |
| 0.28 | 20.1 | 0.44 |
| 0.28 | 22.9 | 0.52 |
| 0.28 | 24.4 | 0.64 |
| 0.28 | 22.9 | 0.56 |
| 0.29 | 24.1 | 0.69 |
| 0.29 | 23.2 | 0.59 |
| 0.29 | 23.2 | 0.61 |
| 0.30 | 21.0 | 0.60 |
| 0.30 | 22.9 | 0.54 |
| 0.33 | 22.6 | 0.63 |
| 0.33 | 25.9 | 0.96 |
| 0.34 | 26.2 | 0.78 |
| 0.35 | 21.6 | 0.73 |
| 0.35 | 19.5 | 0.71 |
| 0.36 | 23.8 | 0.98 |
| 0.36 | 24.4 | 0.90 |
| 0.37 | 22.6 | 1.03 |
| 0.41 | 21.9 | 1.08 |
| 0.41 | 23.5 | 1.21 |
| 0.44 | 24.7 | 1.57 |
| 0.44 | 25.0 | 1.58 |
| 0.45 | 24.4 | 1.65 |
| 0.46 | 24.4 | 1.46 |
| 0.46 | 24.4 | 1.44 |
| 0.52 | 26.5 | 2.18 |

*Source:* A.C. Atkinson. Regression diagnostics, trend formations and constructed variables (with discussion). *Journal of the Royal Statistical Society, Series B*, 44:1–36, 1982.

slope of the averages $\bar{y}/\bar{x}$. In Chapter 22 we will encounter the so-called least squares estimate:

$$\frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}.$$

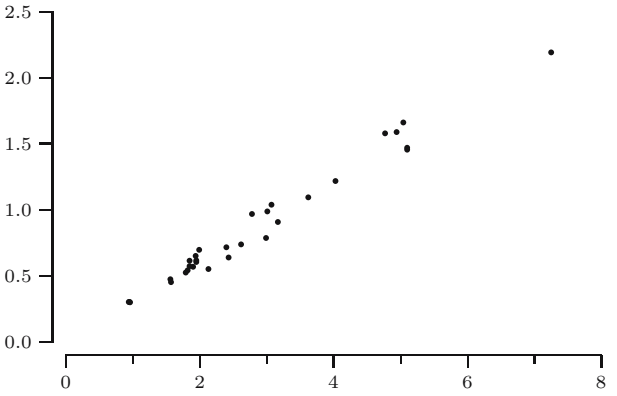


**Fig. 17.13.** Scatterplot of the black cherry tree data.

Compute all three estimates for the data in Table 17.3. You need at least 5 digits accuracy, and you may use that $\sum x_i = 87.456$, $\sum y_i = 26.486$, $\sum y_i/x_i = 9.369$, $\sum x_i y_i = 95.498$, and $\sum x_i^2 = 314.644$.

**17.10** Let $X$ be a random variable with (continuous) distribution function $F$. Let $m = q_{0.5} = F^{\text{inv}}(0.5)$ be the median of $F$ and define the random variable

$$Y = |X - m|.$$

**a.** Show that $Y$ has distribution function $G$, defined by

$$G(y) = F(m + y) - F(m - y).$$

**b.** The MAD of $F$ is the median of $G$. Show that if the density $f$ corresponding to $F$ is symmetric around its median $m$, then

$$G(y) = 2F(m + y) - 1$$

and derive that

$$G^{\text{inv}}(\tfrac{1}{2}) = F^{\text{inv}}(\tfrac{3}{4}) - F^{\text{inv}}(\tfrac{1}{2}).$$

**c.** Use **b** to conclude that the MAD of an $N(\mu, \sigma^2)$ distribution is equal to $\sigma\Phi^{\text{inv}}(3/4)$, where $\Phi$ is the distribution function of a standard normal distribution. Recall that the distribution function $F$ of an $N(\mu, \sigma^2)$ can be written as

$$F(x) = \Phi\left(\frac{x - \mu}{\sigma}\right).$$

You might check that, as stated in Section 17.2, the MAD of the $N(5, 4)$ distribution is equal to $2\Phi^{\text{inv}}(3/4) = 1.3490$.

**17.11** In this exercise we compute the MAD of the $Exp(\lambda)$ distribution.

**a.** Let $X$ have an $Exp(\lambda)$ distribution, with median $m = (\ln 2)/\lambda$. Show that $Y = |X - m|$ has distribution function

$$G(y) = \frac{1}{2}\left(\mathrm{e}^{\lambda y} - \mathrm{e}^{-\lambda y}\right).$$

**b.** Argue that the MAD of the $Exp(\lambda)$ distribution is a solution of the equation $\mathrm{e}^{2\lambda y} - \mathrm{e}^{\lambda y} - 1 = 0$.

**c.** Compute the MAD of the $Exp(\lambda)$ distribution.
*Hint:* put $x = \mathrm{e}^{\lambda y}$ and first solve for $x$.