15

Exploratory data analysis: graphical summaries

In the previous chapters we focused on probability models to describe random phenomena. Confronted with a new phenomenon, we want to learn about the randomness that is associated with it. It is common to conduct an experiment for this purpose and record observations concerning the phenomenon. The set of observations is called a *dataset*. By exploring the dataset we can gain insight into what probability model suits the phenomenon.

Frequently you will have to deal with a dataset that contains so many elements that it is necessary to condense the data for easy visual comprehension of general characteristics. In this chapter we present several graphical methods to do so. To graphically represent univariate datasets, consisting of repeated measurements of one particular quantity, we discuss the classical *histogram*, the more recently introduced *kernel density estimates* and the *empirical distribution function*. To represent a bivariate dataset, which consists of repeated measurements of two quantities, we use the *scatterplot*.

15.1 Example: the Old Faithful data

The Old Faithful geyser at Yellowstone National Park, Wyoming, USA, was observed from August 1st to August 15th, 1985. During that time, data were collected on the duration of eruptions. There were 272 eruptions observed, of which the recorded durations are listed in Table 15.1. The data are given in seconds.

The variety in the lengths of the eruptions indicates that randomness is involved. By exploring the dataset we might learn about this randomness. For instance: we like to know which durations are more likely to occur than others; is there something like "the typical duration of an eruption"; do the durations vary symmetrically around the center of the dataset; and so on. In order to retrieve this type of information, just listing the observed durations does not help us very much. Somehow we must summarize the observed data. We could

Table 15.1. Duration in seconds of 272 eruptions of the Old Faithful geyser.

216	108	200	137	272	173	282	216	117	261
110	235	252	105	282	130	105	288	96	255
108	105	207	184	272	216	118	245	231	266
258	268	202	242	230	121	112	290	110	287
261	113	274	105	272	199	230	126	278	120
288	283	110	290	104	293	223	100	274	259
134	270	105	288	109	264	250	282	124	282
242	118	270	240	119	304	121	274	233	216
248	260	246	158	244	296	237	271	130	240
132	260	112	289	110	258	280	225	112	294
149	262	126	270	243	112	282	107	291	221
284	138	294	265	102	278	139	276	109	265
157	244	255	118	276	226	115	270	136	279
112	250	168	260	110	263	113	296	122	224
254	134	272	289	260	119	278	121	306	108
302	240	144	276	214	240	270	245	108	238
132	249	120	230	210	275	142	300	116	277
115	125	275	200	250	260	270	145	240	250
113	275	255	226	122	266	245	110	265	131
288	110	288	246	238	254	210	262	135	280
126	261	248	112	276	107	262	231	116	270
143	282	112	230	205	254	144	288	120	249
112	256	105	269	240	247	245	256	235	273
245	145	251	133	267	113	111	257	237	140
249	141	296	174	275	230	125	262	128	261
132	267	214	270	249	229	235	267	120	257
286	272	111	255	119	135	285	247	129	265
109	268								

Source: W. Härdle. *Smoothing techniques with implementation in S.* 1991; Table 3, page 201. © Springer New York.

start by computing the mean of the data, which is 209.3 for the Old Faithful data. However, this is a poor summary of the dataset, because there is a lot more information in the observed durations. How do we get hold of this?

Just staring at the dataset for a while tells us very little. To see something, we have to rearrange the data somehow. The first thing we could do is order the data. The result is shown in Table 15.2. Putting the elements in order already provides more information. For instance, it is now immediately clear that all elements lie between 96 and 306.

Quick exercise 15.1 Which two elements of the Old Faithful dataset split the dataset in three groups of equal size?

A closer look at the ordered data shows that the two middle elements (the 136th and 137th elements in ascending order) are equal to 240, which is much closer to the maximum value 306 than to the minimum value 96. This seems to

Table 15.2. Ordered durations of eruptions of the Old Faithful geyser.

96	100	102	104	105	105	105	105	105	105
107	107	108	108	108	108	109	109	109	110
110	110	110	110	110	110	111	111	112	112
112	112	112	112	112	112	113	113	113	113
115	115	116	116	117	118	118	118	119	119
119	120	120	120	120	121	121	121	122	122
124	125	125	126	126	126	128	129	130	130
131	132	132	132	133	134	134	135	135	136
137	138	139	140	141	142	143	144	144	145
145	149	157	158	168	173	174	184	199	200
200	202	205	207	210	210	214	214	216	216
216	216	221	223	224	225	226	226	229	230
230	230	230	230	231	231	233	235	235	235
237	237	238	238	240	240	240	240	240	240
242	242	243	244	244	245	245	245	245	245
246	246	247	247	248	248	249	249	249	249
250	250	250	250	251	252	254	254	254	255
255	255	255	256	256	257	257	258	258	259
260	260	260	260	260	261	261	261	261	262
262	262	262	263	264	265	265	265	265	266
266	267	267	267	268	268	269	270	270	270
270	270	270	270	270	271	272	272	272	272
272	273	274	274	274	275	275	275	275	276
276	276	276	277	278	278	278	279	280	280
282	282	282	282	282	282	283	284	285	286
287	288	288	288	288	288	288	289	289	290
290	291	293	294	294	296	296	296	300	302
304	306								

indicate that the dataset is somewhat asymmetric, but even from the ordered dataset we cannot get a clear picture of this asymmetry. Also, geologists believe that there are two different kinds of eruptions that play a role. Hence one would expect two separate values around which the elements of the dataset would accumulate, corresponding to the typical durations of the two types of eruptions. Again it is not clear, not even from the ordered dataset, what these two typical values are. It would be better to have a plot of the dataset that reflects symmetry or asymmetry of the data and from which we can easily see where the elements accumulate. In the following sections we will discuss two such methods.

15.2 Histograms

The classical method to graphically represent data is the histogram, which probably dates from the mortality studies of John Graunt in 1662 (see West-

ergaard [39], p.22). The term *histogram* appears to have been used first by Karl Pearson ([22]). Figure 15.1 displays a histogram of the Old Faithful data. The picture immediately reveals the asymmetry of the dataset and the fact that the elements accumulate somewhere near 120 and 270, which was not clear from Tables 15.1 and 15.2.

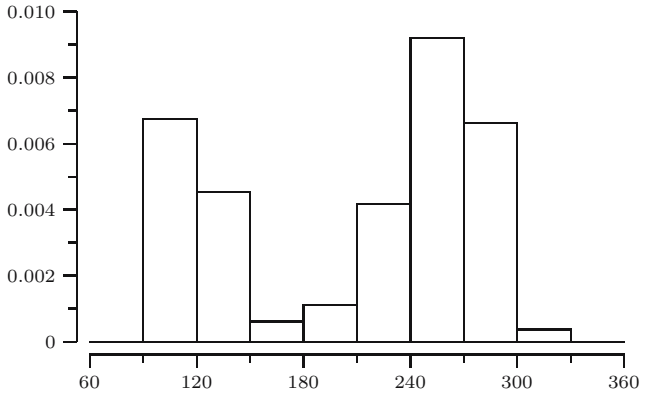

Fig. 15.1. Histogram of the Old Faithful data.

The construction of the histogram is as follows. Let us denote a generic (univariate) dataset of size n by

$$x_1, x_2, \ldots, x_n$$

and suppose we want to construct a histogram. We use the version of the histogram that is scaled in such a way that the total area under the curve is equal to one.[1]

First we divide the range of the data into intervals. These intervals are called *bins* and are denoted by

$$B_1, B_2, \ldots, B_m.$$

The length of an interval B_i is denoted by $|B_i|$ and is called the *bin width*. The bins do not necessarily have the same width. In Figure 15.1 we have eight bins of equal bin width. We want the area under the histogram on each bin B_i to reflect the number of elements in B_i. Since the total area 1 under the histogram then corresponds to the total number of elements n in the dataset, the area under the histogram on a bin B_i is equal to the proportion of elements in B_i:

$$\frac{\text{the number of } x_j \text{ in } B_i}{n}.$$

[1] The reason to scale the histogram so that the total area under the curve is equal to one is that if we view the data as being generated from some unknown probability density f (see Chapter 17), such a histogram can be used as a crude estimate of f.

The *height* of the histogram on bin B_i must then be equal to

$$\frac{\text{the number of } x_j \text{ in } B_i}{n|B_i|}.$$

QUICK EXERCISE 15.2 Use Table 15.2 to count how many elements fall into each of the bins (90, 120], (120, 150], ..., (300, 330] in Figure 15.1 and compute the height on each bin.

Choice of the bin width

Consider a histogram with bins of equal width. In that case the bins are of the form

$$B_i = (r + (i-1)b, r + ib] \quad \text{for } i = 1, 2, \ldots, m,$$

where r is some reference point smaller than the minimum of the dataset, and b denotes the bin width. In Figure 15.2, three histograms of the Old Faithful data of Table 15.2 are displayed with bin widths equal to 2, 30, and 90, respectively. Clearly, the choice of the bin width b, or the corresponding choice of the number of bins m, will determine what the resulting histogram will look like. Choosing the bin width too small will result in a chaotic figure with many isolated peaks. Choosing the bin width too large will result in a figure without much detail, at the risk of losing information about general characteristics. In Figure 15.2, bin width $b = 2$ is somewhat too small. Bin width $b = 90$ is clearly too large and produces a histogram that no longer captures the fact that the data show two separate modes near 120 and 270.

How does one go about choosing the bin width? In practice, this might boil down to picking the bin width by trial and error, continuing until the figure looks reasonable. Mathematical research, however, has provided some guidelines for a data-based choice for b or m. Formulas that may effectively be used are $m = 1 + 3.3 \log_{10}(n)$ (see [34]) or $b = 3.49\, sn^{-1/3}$ (see [29]; see also Remark 15.1), where s is the sample standard deviation (see Section 16.2 for the definition of the sample standard deviation).

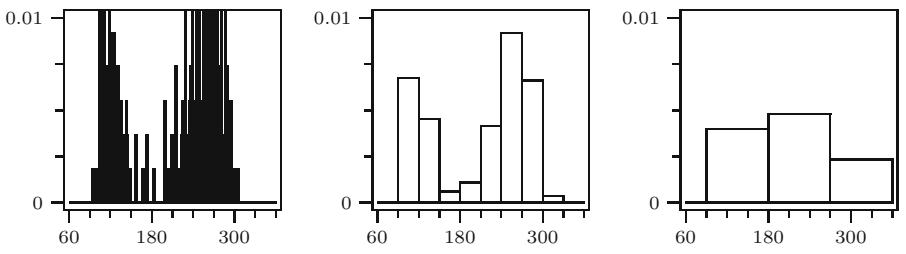

Fig. 15.2. Histograms of the Old Faithful data with different bin widths.

Remark 15.1 (Normal reference method for histograms). Let $H_n(x)$ denote the height of the histogram at x and suppose that we view our dataset as being generated from a probability distribution with density f. We would like to find the bin width that minimizes the difference between H_n and f, measured by the so-called mean integrated squared error (MISE)

$$\mathrm{E}\left[\int_{-\infty}^{\infty} (H_n(x) - f(x))^2 \, \mathrm{d}x\right].$$

Under suitable smoothness conditions on f, the value of b that minimizes the MISE as n goes to infinity is given by

$$b = C(f)n^{-1/3} \quad \text{where } C(f) = 6^{1/3}\left(\int_{-\infty}^{\infty} f'(x)^2 \, \mathrm{d}x\right)^{-1/3}$$

(see for instance [29] or [12]). A simple data-based choice for b is obtained by estimating the constant $C(f)$. The normal reference method takes f to be the density of an $N(\mu, \sigma^2)$ distribution, in which case $C(f) = (24\sqrt{\pi})^{1/3}\sigma$. Estimating σ by the sample standard deviation s (see Chapter 16 for a definition of s) would result in bin width

$$b = (24\sqrt{\pi})^{1/3} s n^{-1/3}.$$

For the Old Faithful data this would give $b = 36.89$.

Quick exercise 15.3 If we construct a histogram for the Old Faithful data with equal bin width $b = 3.49\, s n^{-1/3}$, how may bins will we need to cover the data if $s = 68.48$?

The main advantage of the histogram is that it is simple. Its disadvantage is the discrete character of the plot. In Figure 15.1 it is still somewhat unclear which two values correspond to the typical durations of the two types of eruptions. Another well-known artifact is that changing the bin width slightly or keeping the bin width fixed and shifting the bins slightly may result in a figure of a different nature. A method that produces a smoother figure and is less sensitive to these kinds of changes will be discussed in the next section.

15.3 Kernel density estimates

We can graphically represent data in a more variegated plot by a so-called kernel density estimate. The basic ideas of kernel density estimation first appeared in the early 1950s. Rosenblatt [25] and Parzen [21] provided the stimulus for further research on this topic. Although the method was introduced in the middle of the last century, until recently it remained unpopular as a tool for practitioners because of its computationally intensive nature.

Figure 15.3 displays a kernel density estimate of the Old Faithful data. Again the picture immediately reveals the asymmetry of the dataset, but it is much

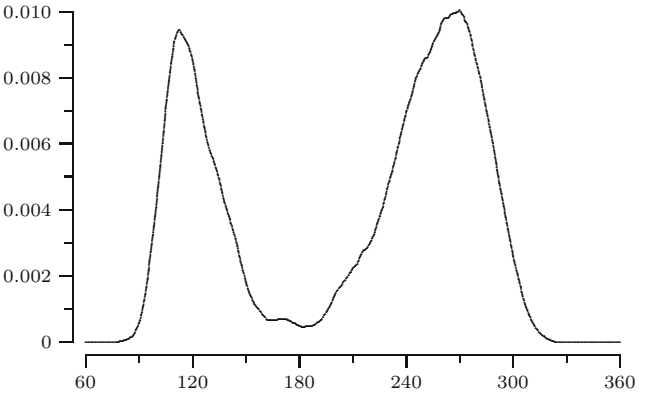

Fig. 15.3. Kernel density estimate of the Old Faithful data.

smoother than the histogram in Figure 15.1. Note that it is now easier to detect the two typical values around which the elements accumulate.

The idea behind the construction of the plot is to "put a pile of sand" around each element of the dataset. At places where the elements accumulate, the sand will pile up. The actual plot is constructed by choosing a *kernel* K and a *bandwidth* h. The kernel K reflects the shape of the piles of sand, whereas the bandwidth is a tuning parameter that determines how wide the piles of sand will be. Formally, a kernel K is a function $K : \mathbb{R} \to \mathbb{R}$. Figure 15.4 displays several well-known kernels. A kernel K typically satisfies the following conditions:

(K1) K is a probability density, i.e., $K(u) \geq 0$ and $\int_{-\infty}^{\infty} K(u)\,\mathrm{d}u = 1$;
(K2) K is symmetric around zero, i.e., $K(u) = K(-u)$;
(K3) $K(u) = 0$ for $|u| > 1$.

Examples are the *Epanechnikov kernel*:

$$K(u) = \frac{3}{4}\left(1 - u^2\right) \quad \text{for } -1 \leq u \leq 1$$

and $K(u) = 0$ elsewhere, and the *triweight kernel*

$$K(u) = \frac{35}{32}\left(1 - x^2\right)^3 \quad \text{for } -1 \leq u \leq 1$$

and $K(u) = 0$ elsewhere. Sometimes one uses kernels that do not satisfy condition (K3), for example, the *normal kernel*

$$K(u) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}u^2} \quad \text{for } -\infty < u < \infty.$$

Let us denote a kernel density estimate by $f_{n,h}$, and suppose that we want to construct $f_{n,h}$ for a dataset $x_1, x_2, \ldots, x_n$. In Figure 15.5 the construction is

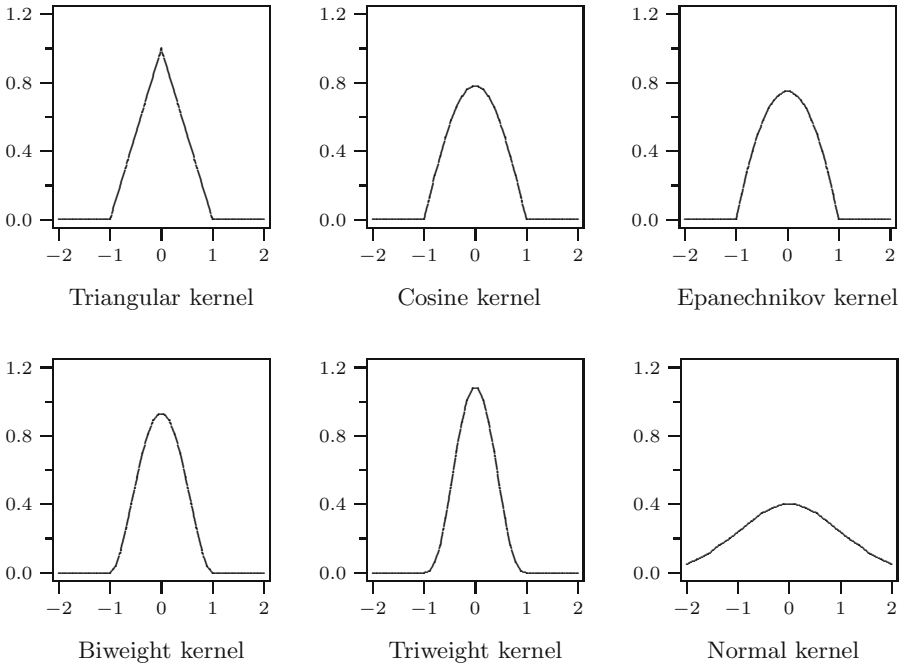

Fig. 15.4. Examples of well-known kernels K.

illustrated for a dataset containing five elements, where we use the Epanechnikov kernel and bandwidth $h = 0.5$. First we scale the kernel K (solid line) into the function

$$t \mapsto \frac{1}{h} K\left(\frac{t}{h}\right).$$

The scaled kernel (dotted line) is of the same type as the original kernel, with area 1 under the curve but is positive on the interval $[-h, h]$ instead of $[-1, 1]$ and higher (lower) when h is smaller (larger) than 1. Next, we put a scaled kernel around each element x_i in the dataset. This results in functions of the type

$$t \mapsto \frac{1}{h} K\left(\frac{t - x_i}{h}\right).$$

These shifted kernels (dotted lines) have the same shape as the transformed kernel, all with area 1 under the curve, but they are now symmetric around x_i and positive on the interval $[x_i - h, x_i + h]$. We see that the graphs of the shifted kernels will overlap whenever x_i and x_j are close to each other, so that things will pile up more at places where more elements accumulate. The kernel density estimate $f_{n,h}$ is constructed by summing the scaled kernels and dividing them by n, in order to obtain area 1 under the curve:

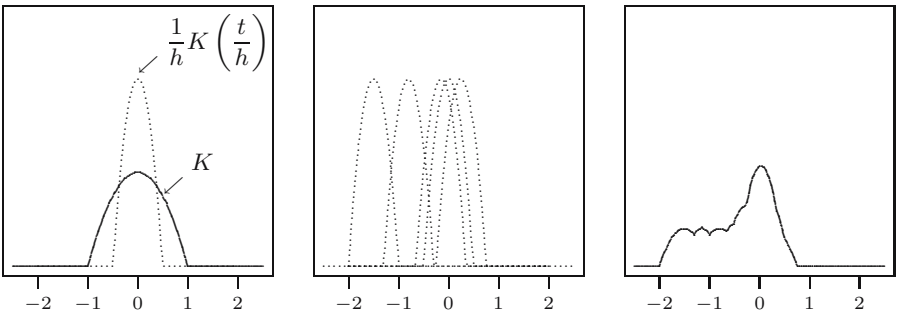

Fig. 15.5. Construction of a kernel density estimate $f_{n,h}$.

$$f_{n,h}(t) = \frac{1}{n}\left\{\frac{1}{h}K\left(\frac{t-x_1}{h}\right) + \frac{1}{h}K\left(\frac{t-x_2}{h}\right) + \cdots + \frac{1}{h}K\left(\frac{t-x_n}{h}\right)\right\}$$

or briefly,

$$f_{n,h}(t) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{t-x_i}{h}\right). \tag{15.1}$$

When computing $f_{n,h}(t)$, we assign higher weights to observations x_i closer to t, in contrast to the histogram where we simply count the number of observations in the bin that contains t. Note that as a consequence of condition (K1), $f_{n,h}$ itself is a probability density:

$$f_{n,h}(t) \geq 0 \text{ and } \int_{-\infty}^{\infty} f_{n,h}(t)\,\mathrm{d}t = 1.$$

Quick exercise 15.4 Check that the total area under the kernel density estimate is equal to one, i.e., show that $\int_{-\infty}^{\infty} f_{n,h}(t)\,\mathrm{d}t = 1$.

Note that computing $f_{n,h}$ is very computationally intensive. Its common use nowadays is therefore a typical product of the recent developments in computer hardware, despite the fact that the method was introduced much earlier.

Choice of the bandwidth

The bandwidth h plays the same role for kernel density estimates as the bin width b does for histograms. In Figure 15.6 three kernel density estimates of the Old Faithful data are plotted with the triweight kernel and bandwidths 1.8, 18, and 180. It is clear that the choice of the bandwidth h determines largely what the resulting kernel density estimate will look like. Choosing the bandwidth too small will produce a curve with many isolated peaks. Choosing the bandwidth too large will produce a very smooth curve, at the risk of smoothing away important features of the data. In Figure 15.6 bandwidth

$h = 1.8$ is somewhat too small. Bandwidth $h = 180$ is clearly too large and produces an oversmoothed kernel density estimate that no longer captures the fact that the data show two separate modes.

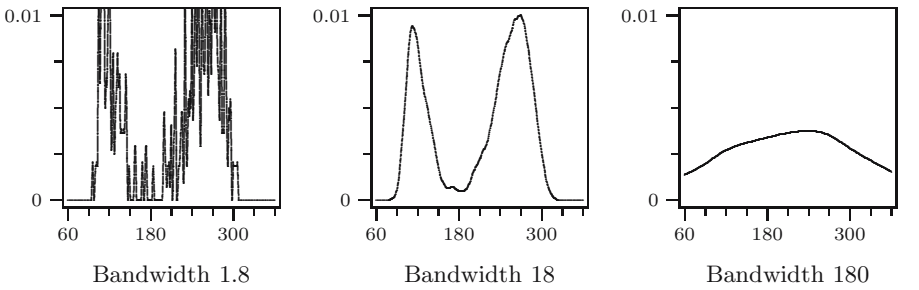

Fig. 15.6. Kernel estimates of the Old Faithful data.

How does one go about choosing the bandwidth? Similar to histograms, in practice one could do this by trial and error and continue until one obtains a reasonable picture. Recent research, however, has provided some guidelines for a data-based choice of h. A formula that may effectively be used is $h = 1.06\, sn^{-1/5}$, where s denotes the sample standard deviation (see, for instance, [31]; see also Remark 15.2).

Remark 15.2 (Normal reference method for kernel estimates). Suppose we view our dataset as being generated from a probability distribution with density f. Let K be a fixed chosen kernel and let $f_{n,h}$ be the kernel density estimate. We would like to take the bandwidth that minimizes the difference between $f_{n,h}$ and f, measured by the so-called mean integrated squared error (MISE)

$$\mathrm{E}\left[\int_{-\infty}^{\infty} (f_{n,h}(x) - f(x))^2 \, \mathrm{d}x\right].$$

Under suitable smoothness conditions on f, the value of h that minimizes the MISE, as n goes to infinity, is given by

$$h = C_1(f)C_2(K)n^{-1/5},$$

where the constants $C_1(f)$ and $C_2(K)$ are given by

$$C_1(f) = \left(\frac{1}{\int_{-\infty}^{\infty} f''(x)^2 \, \mathrm{d}x}\right)^{1/5} \quad \text{and} \quad C_2(K) = \frac{\left(\int_{-\infty}^{\infty} K(u)^2 \, \mathrm{d}u\right)^{1/5}}{\left(\int_{-\infty}^{\infty} u^2 K(u) \, \mathrm{d}u\right)^{2/5}}.$$

After choosing the kernel K, one can compute the constant $C_2(K)$ to obtain a simple data-based choice for h by estimating the constant $C_1(f)$. For instance, for the normal kernel one finds $C_2(K) = (2\sqrt{\pi})^{-1/5}$. As with

histograms (see Remark 15.1), the normal reference method takes f to be the density of an $N(\mu, \sigma^2)$ distribution, in which case $C_1(f) = (8\sqrt{\pi}/3)^{1/5}\sigma$. Estimating σ by the sample standard deviation s (see Chapter 16 for a definition of s) would result in bandwidth

$$h = \left(\tfrac{4}{3}\right)^{1/5} s n^{-1/5}.$$

For the Old Faithful data, this would give $h = 23.64$.

QUICK EXERCISE 15.5 If we construct a kernel density estimate for the Old Faithful data with bandwidth $h = 1.06sn^{-1/5}$, then on what interval is $f_{n,h}$ strictly positive if $s = 68.48$?

Choice of the kernel

To construct a kernel density estimate, one has to choose a kernel K and a bandwidth h. The choice of kernel is less important. In Figure 15.7 we have plotted two kernel density estimates for the Old Faithful data of Table 15.1: one is constructed with the triweight kernel (solid line), and one with the Epanechnikov kernel (dotted line), both with the same bandwidth $h = 24$. As one can see, the graphs are very similar. If one wants to compare with the normal kernel, one should set the bandwidth of the normal kernel at about $h/4$. This has to do with the fact that the normal kernel is much more spread out than the two kernels mentioned here, which are zero outside $[-1, 1]$.

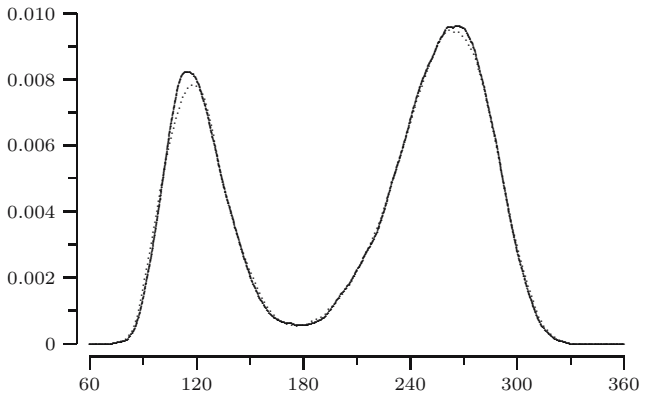

Fig. 15.7. Kernel estimates of the Old Faithful data with different kernels: triweight (solid line) and Epanechnikov kernel (dotted), both with bandwidth $h = 24$.

Boundary kernels

In order to estimate the parameters of a software reliability model, failure data are collected. Usually the most desirable type of failure data results when the

Table 15.3. Interfailure times between successive failures.

30	113	81	115	9	2	91	112	15	138
50	77	24	108	88	670	120	26	114	325
55	242	68	422	180	10	1146	600	15	36
4	0	8	227	65	176	58	457	300	97
263	452	255	197	193	6	79	816	1351	148
21	233	134	357	193	236	31	369	748	0
232	330	365	1222	543	10	16	529	379	44
129	810	290	300	529	281	160	828	1011	445
296	1755	1064	1783	860	983	707	33	868	724
2323	2930	1461	843	12	261	1800	865	1435	30
143	108	0	3110	1247	943	700	875	245	729
1897	447	386	446	122	990	948	1082	22	75
482	5509	100	10	1071	371	790	6150	3321	1045
648	5485	1160	1864	4116					

Source: J.D. Musa, A. Iannino, and K. Okumoto. *Software reliability: measurement, prediction, application.* McGraw-Hill, New York, 1987; Table on page 305.

failure times are recorded, or equivalently, the length of an interval between successive failures. The data in Table 15.3 are observed interfailure times in CPU seconds for a certain control software system. On the left in Figure 15.8 a kernel density estimate of the observed interfailure times is plotted. Note that to the left of the origin, $f_{n,h}$ is positive. This is absurd, since it suggests that there are negative interfailure times.

This phenomenon is a consequence of the fact that one uses a symmetric kernel. In that case, the resulting kernel density estimate will always be positive on the interval $[x_i - h, x_i + h]$ for every element x_i in the dataset. Hence, obser-

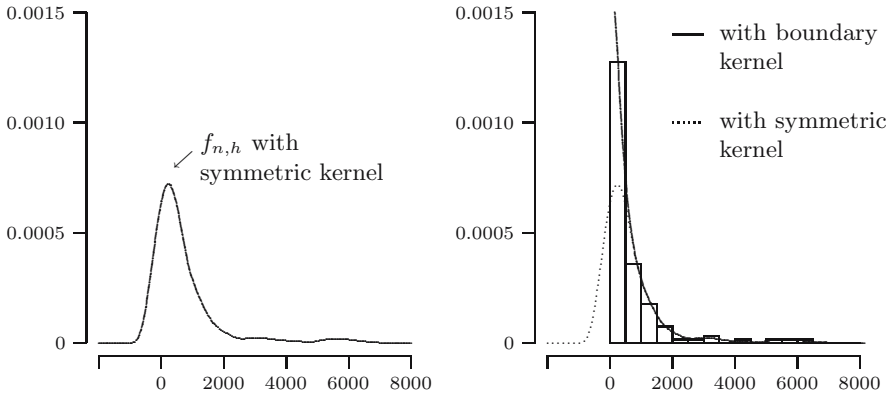

Fig. 15.8. Kernel density estimate of the software reliability data with symmetric and boundary kernel.

vations close to zero will cause the kernel density estimate $f_{n,h}$ to be positive to the left of zero. It is possible to improve the kernel density estimate in a neighborhood of zero by means of a so-called boundary kernel. Without going into detail about the construction of such an improvement, we will only show the result of this. On the right in Figure 15.8 the histogram of the interfailure times is plotted together with the kernel density estimate constructed with a symmetric kernel (dotted line) and with the boundary kernel density estimate (solid line). The boundary kernel density estimate is 0 to the left of the origin and is adjusted on the interval $[0, h)$. On the interval $[h, \infty)$ both kernel density estimates are the same.

15.4 The empirical distribution function

Another way to graphically represent a dataset is to plot the data in a cumulative manner. This can be done using the *empirical cumulative distribution function* of the data. It is denoted by F_n and is defined at a point x as the proportion of elements in the dataset that are less than or equal to x:

$$F_n(x) = \frac{\text{number of elements in the dataset} \leq x}{n}.$$

To illustrate the construction of F_n, consider the dataset consisting of the elements

$$4 \quad 3 \quad 9 \quad 1 \quad 7.$$

The corresponding empirical distribution function is displayed in Figure 15.9. For $x < 1$, there are no elements less than or equal to x, so that $F_n(x) = 0$. For $1 \leq x < 3$, only the element 1 is less than or equal to x, so that $F_n(x) = 1/5$. For $3 \leq x < 4$, the elements 1 and 3 are less than or equal to x, so that $F_n(x) = 2/5$, and so on.

In general, the graph of F_n has the form of a staircase, with $F_n(x) = 0$ for all x smaller than the minimum of the dataset and $F_n(x) = 1$ for all x greater than the maximum of the dataset. Between the minimum and maximum, F_n has a jump of size $1/n$ at each element of the dataset and is constant between successive elements. In Figure 15.9, the marks $\bullet$ and $\circ$ are added to the graph to emphasize the fact that, for instance, the value of $F_n(x)$ at $x = 3$ is 0.4, not 0.2. Usually, we leave these out, and one might also connect the horizontal segments by vertical lines.

In Figure 15.10 the empirical distribution functions are plotted for the Old Faithful data and the software reliability data. The fact that the Old Faithful data accumulate in the neighborhood of 120 and 270 is reflected in the graph of F_n by the fact that it is steeper at these places: the jumps of F_n succeed each other faster. In regions where the elements of the dataset are more stretched

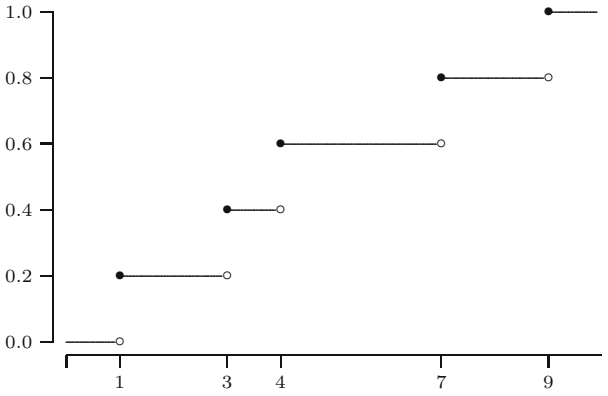

Fig. 15.9. Empirical distribution function.

out, the graph of F_n is flatter. Similar behavior can be seen for the software reliability data in the neighborhood of zero. The elements accumulate more close to zero, less as we move to the right. This is reflected by the empirical distribution function, which is very steep near zero and flattens out if we move to the right.

The graph of the empirical distribution function for the Old Faithful data agrees with the histogram in Figure 15.1 whose height is the largest on the bins (90, 120] and (240, 270]. In fact, there is a one-to-one relation between the two graphical summaries of the data: the area under the histogram on a single bin is equal to the relative frequency of elements that lie in that bin, which is also equal to the increase of F_n on that bin. For instance, the area under the histogram on bin (240, 270] for the Old Faithful data is equal to $30 \cdot 0.0092 =$

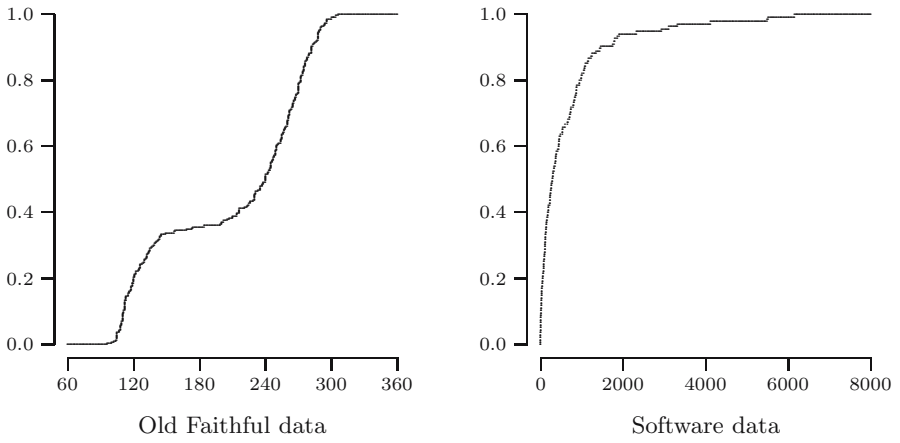

Fig. 15.10. Empirical distribution function of the Old Faithful data and the software reliability data.

0.276 (see Quick exercise 15.2). On the other hand, $F_n(270) = 215/272 = 0.7904$ and $F_n(240) = 140/272 = 0.5147$, whose difference $F_n(270) - F_n(240)$ is also equal to 0.276.

Quick exercise 15.6 Suppose that for a dataset consisting of 300 elements, the value of the empirical distribution function in the point 1.5 is equal to 0.7. How many elements in the dataset are strictly greater than 1.5?

Remark 15.3 (F_n as a discrete distribution function). Note that F_n satisfies the four properties of a distribution function: it is continuous from the right, $F_n(x) \to 0$ as $x \to -\infty$, $F_n(x) \to 1$ as $x \to \infty$ and F_n is nondecreasing. This means that F_n itself is a distribution function of some random variable. Indeed, F_n is the distribution function of the discrete random variable that attains values $x_1, x_2, \ldots, x_n$ with equal probability $1/n$.

15.5 Scatterplot

In some situations one wants to investigate the relationship between two or more variables. In the case of two variables x and y, the dataset consists of *pairs of observations*:

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n).$$

We call such a dataset a *bivariate* dataset in contrast to the *univariate* dataset, which consists of observations of one particular quantity. We often like to investigate whether the value of variable y depends on the value of the variable x, and if so, whether we can describe the relation between the two variables. A first step is to take a look at the data, i.e., to plot the points (x_i, y_i) for $i = 1, 2 \ldots, n$. Such a plot is called a *scatterplot*.

Drilling in rock

During a study about "dry" and "wet" drilling in rock, six holes were drilled, three corresponding to each process. In a dry hole one forces compressed air down the drill rods to flush the cutting and the drive hammer, whereas in a wet hole one forces water. As the hole gets deeper, one has to add a rod of 5 feet length to the drill. In each hole the time was recorded to advance 5 feet to a total depth of 400 feet. The data in Table 15.4 are in 1/100 minute and are derived from the original data in [23]. The original data consisted of drill times for each of the six holes and contained missing observations and observations that were known to be too large. The data in Table 15.4 are the mean drill times of the bona fide observations at each depth for dry and wet drilling.

One of the questions of interest is whether drill time depends on depth. To investigate this, we plot the mean drill time against depth. Figure 15.11 displays

Table 15.4. Mean drill time.

Depth	Dry	Wet	Depth	Dry	Wet
5	640.67	830.00	205	803.33	962.33
10	674.67	800.00	210	794.33	864.67
15	708.00	711.33	215	760.67	805.67
20	735.67	867.67	220	789.50	966.00
25	754.33	940.67	225	904.50	1010.33
30	723.33	941.33	230	940.50	936.33
35	664.33	924.33	235	882.00	915.67
40	727.67	873.00	240	783.50	956.33
45	658.67	874.67	245	843.50	936.00
50	658.00	843.33	250	813.50	803.67
55	705.67	885.67	255	658.00	697.33
60	700.00	881.67	260	702.50	795.67
65	720.67	822.00	265	623.50	1045.33
70	701.33	886.33	270	739.00	1029.67
75	716.67	842.50	275	907.50	977.00
80	649.67	874.67	280	846.00	1054.33
85	667.33	889.33	285	829.00	1001.33
90	612.67	870.67	290	975.50	1042.00
95	656.67	916.00	295	998.00	1200.67
100	614.00	888.33	300	1037.50	1172.67
105	584.00	835.33	305	984.00	1019.67
110	619.67	776.33	310	972.50	990.33
115	666.00	811.67	315	834.00	1173.33
120	695.00	874.67	320	675.00	1165.67
125	702.00	846.00	325	686.00	1142.00
130	739.67	920.67	330	963.00	1030.67
135	790.67	896.33	335	961.50	1089.67
140	730.33	810.33	340	932.00	1154.33
145	674.00	912.33	345	1054.00	1238.50
150	749.00	862.33	350	1038.00	1208.67
155	709.67	828.00	355	1238.00	1134.67
160	769.00	812.67	360	927.00	1088.00
165	663.00	795.67	365	850.00	1004.00
170	679.33	897.67	370	1066.00	1104.00
175	740.67	881.00	375	962.50	970.33
180	776.50	819.67	380	1025.50	1054.50
185	688.00	853.33	385	1205.50	1143.50
190	761.67	844.33	390	1168.00	1044.00
195	800.00	919.00	395	1032.50	978.33
200	845.50	933.33	400	1162.00	1104.00

Source: R. Penner and D.G. Watts. Mining information. *The American Statistician*, 45:4–9, 1991; Table 1 on page 6.

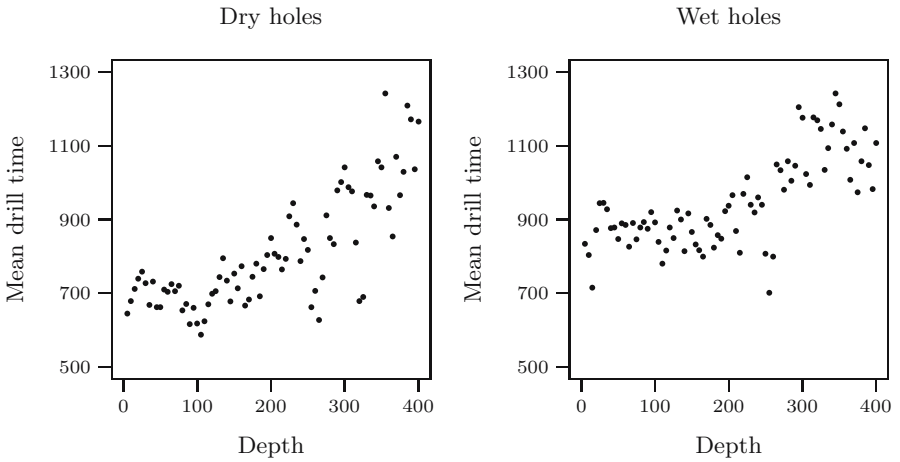

Fig. 15.11. Scatterplots of mean drill time versus depth.

the resulting scatterplots for the dry and wet holes. The scatterplots seem to indicate that in the beginning the drill time hardly depends on depth, at least up to, let's say, 250 feet. At greater depth, the drill time seems to vary over a larger range and increases somewhat with depth. A possible explanation for this is that the drill moved from softer to harder material. This was suggested by the fact that the drill hit an ore lens at about 250 feet and that the natural place such ore lenses occur is between two different materials (see [23] for details).

A more important question is whether one can drill holes faster using dry drilling or wet drilling. The scatterplots seem to suggest that dry drilling might be faster. We will come back to this later.

Predicting Janka hardness of Australian timber

The Janka hardness test is a standard test to measure the hardness of wood. It measures the force required to push a steel ball with a diameter of 11.28 millimeters (0.444 inch) into the wood to a depth of half the ball's diameter. To measure Janka hardness directly is difficult. However, it is related to the density of the wood, which is comparatively easy to measure. In Table 15.5 a bivariate dataset is given of density (x) and Janka hardness (y) of 36 Australian eucalypt hardwoods.

In order to get an impression of the relationship between hardness and density, we made a scatterplot of the bivariate dataset, which is displayed in Figure 15.12. It consists of all points (x_i, y_i) for $i = 1, 2, \ldots, 36$. The scatterplot might provide suggestions for the formula that describes the relationship between the variables x and y. In this case, a linear relationship between the two variables does not seem unreasonable. Later (Chapter 22) we will discuss

Table 15.5. Density and hardness of Australian timber.

Density	Hardness	Density	Hardness	Density	Hardness
24.7	484	39.4	1210	53.4	1880
24.8	427	39.9	989	56.0	1980
27.3	413	40.3	1160	56.5	1820
28.4	517	40.6	1010	57.3	2020
28.4	549	40.7	1100	57.6	1980
29.0	648	40.7	1130	59.2	2310
30.3	587	42.9	1270	59.8	1940
32.7	704	45.8	1180	66.0	3260
35.6	979	46.9	1400	67.4	2700
38.5	914	48.2	1760	68.8	2890
38.8	1070	51.5	1710	69.1	2740
39.3	1020	51.5	2010	69.1	3140

Source: E.J. Williams. *Regression analysis.* John Wiley & Sons Inc., New York, 1959; Table 3.1 on page 43.

how one can establish such a linear relationship by means of the observed pairs.

Quick exercise 15.7 Suppose we have a eucalypt hardwood tree with density 65. What would your prediction be for the corresponding Janka hardness?

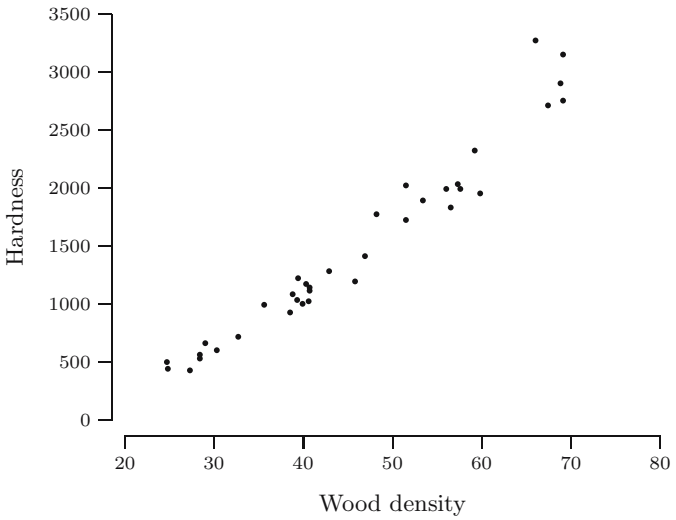

Fig. 15.12. Scatterplot of Janka hardness versus density of wood.

15.6 Solutions to the quick exercises

15.1 There are 272 elements in the dataset. The 91st and 182nd elements of the ordered data divide the dataset in three groups, each consisting of 90 elements. From a closer look at Table 15.2 we find that these two elements are 145 and 260.

15.2 In Table 15.2 one can easily count the number of observations in each of the bins $(90, 120], \ldots, (300, 330]$. The heights on each bin can be computed by dividing the number of observations in each bin by $272 \cdot 30 = 8160$. We get the following:

Bin	Count	Height	Bin	Count	Height
$(90, 120]$	55	0.0067	$(210, 240]$	34	0.0042
$(120, 150]$	37	0.0045	$(240, 270]$	75	0.0092
$(150, 180]$	5	0.0006	$(270, 300]$	54	0.0066
$(180, 210]$	9	0.0011	$(300, 330]$	3	0.0004

15.3 From Table 15.2 we see that we must cover an interval of length of at least $306 - 96 = 210$ with bins of width $b = 3.49 \cdot 68.48 \cdot 272^{-1/3} = 36.89$. Since $210/36.89 = 5.69$, we need at least six bins to cover the whole dataset.

15.4 By means of formula (15.1), we can write

$$\int_{-\infty}^{\infty} f_{n,h}(t)\,\mathrm{d}t = \frac{1}{nh}\sum_{i=1}^{n}\int_{-\infty}^{\infty} K\left(\frac{t - x_i}{h}\right)\,\mathrm{d}t.$$

For any $i = 1, \ldots, n$, we find by change of integration variables $t = hu + x_i$ that

$$\int_{-\infty}^{\infty} K\left(\frac{t - x_i}{h}\right)\,\mathrm{d}t = h\int_{-\infty}^{\infty} K\left(u\right)\,\mathrm{d}u = h,$$

where we also use condition (K1). This directly yields

$$\int_{-\infty}^{\infty} f_{n,h}(t)\,\mathrm{d}t = \frac{1}{nh}\cdot n \cdot h = 1.$$

15.5 The kernel density estimate will be strictly positive between the minimum minus h and the maximum plus h. The bandwidth equals $h = 1.06 \cdot 68.48 \cdot 272^{-1/5} = 23.66$. From Table 15.2, we see that this will be between $96 - 23.66 = 72.34$ and $306 + 23.66 = 329.66$.

15.6 By definition the number of elements less than or equal to 1.5 is $F_{300}(1.5) \cdot 300 = 210$. Hence 90 elements are strictly greater than 1.5.

15.7 Just by drawing a straight line that seems to fit the datapoints well, the authors predicted a Janka hardness of about 2700.

15.7 Exercises

15.1 In [33] Stephen Stigler discusses data from the *Edinburgh Medical and Surgical Journal* (1817). These concern the chest circumference of 5732 Scottish soldiers, measured in inches. The following information is given about the histogram with bin width 1, the first bin starting at 32.5.

Bin	Count	Bin	Count
(32.5, 33.5]	3	(40.5, 41.5]	935
(33.5, 34.5]	19	(41.5, 42.5]	646
(34.5, 35.5]	81	(42.5, 43.5]	313
(35.5, 36.5]	189	(43.5, 44.5]	168
(36.5, 37.5]	409	(44.5, 45.5]	50
(37.5, 38.5]	753	(45.5, 46.5]	18
(38.5, 39.5]	1062	(46.5, 47.5]	3
(39.5, 40.5]	1082	(47.5, 48.5]	1

Source: S.M. Stigler. *The history of statistics – The measurement of uncertainty before 1900.* Cambridge, Massachusetts, 1986.

a. Compute the height of the histogram on each bin.

b. Make a sketch of the histogram. Would you view the dataset as being symmetric or skewed?

15.2 Recall the example of the space shuttle *Challenger* in Section 1.4. The following list contains the launch temperatures in degrees Fahrenheit during previous takeoffs.

66 70 69 68 67 72 73 70 57 63 70 78
67 53 67 75 70 81 76 79 75 76 58

Source: Presidential commission on the space shuttle *Challenger* accident. Report on the space shuttle *Challenger* accident. Washington, DC, 1986; table on pages 129–131.

a. Compute the heights of a histogram with bin width 5, the first bin starting at 50.

b. On January 28, 1986, during the launch of the space shuttle *Challenger*, the temperature was 31 degrees Fahrenheit. Given the dataset of launch temperatures of previous takeoffs, would you consider 31 as a representative launch temperature?

15.3 ⊡ In an article in *Biometrika*, an example is discussed about mine disasters during the period from March 15, 1851, to March, 22, 1962. A dataset has been obtained of 190 recorded time intervals (in days) between successive coal mine disasters involving ten or more men killed. The ordered data are listed in Table 15.6.

Table 15.6. Number of days between successive coal mine disasters.

0	1	1	2	2	3	4	4	4	6
7	10	11	12	12	12	13	15	15	16
16	16	17	17	18	19	19	19	20	20
22	23	24	25	27	28	29	29	29	31
31	32	33	34	34	36	36	37	40	41
41	42	43	45	47	48	49	50	53	54
54	55	56	59	59	61	61	65	66	66
70	72	75	78	78	78	80	80	81	88
91	92	93	93	95	95	96	96	97	99
101	108	110	112	113	114	120	120	123	123
124	124	125	127	129	131	134	137	139	143
144	145	151	154	156	157	176	182	186	187
188	189	190	193	194	197	202	203	208	215
216	217	217	217	218	224	225	228	232	233
250	255	275	275	275	276	286	292	307	307
312	312	315	324	326	326	329	330	336	345
348	354	361	364	368	378	388	420	431	456
462	467	498	517	536	538	566	632	644	745
806	826	871	952	1205	1312	1358	1630	1643	2366

Source: R.G. Jarrett. A note on the intervals between coal mining disasters. *Biometrika*, 66:191-193, 1979; by permission of the Biometrika Trustees.

a. Compute the height on each bin of the histogram with bins $[0, 250]$, $(250, 500], \ldots, (2250, 2500]$.

b. Make a sketch of the histogram. Would you view the dataset as being symmetric or skewed?

15.4 ⊡ The ordered software data (see also Table 15.3) are given in the following list.

0	0	0	2	4	6	8	9	10	10
10	12	15	15	16	21	22	24	26	30
30	31	33	36	44	50	55	58	65	68
75	77	79	81	88	91	97	100	108	108
112	113	114	115	120	122	129	134	138	143
148	160	176	180	193	193	197	227	232	233
236	242	245	255	261	263	281	290	296	300
300	325	330	357	365	369	371	379	386	422
445	446	447	452	457	482	529	529	543	600
648	670	700	707	724	729	748	790	810	816
828	843	860	865	868	875	943	948	983	990
1011	1045	1064	1071	1082	1146	1160	1222	1247	1351
1435	1461	1755	1783	1800	1864	1897	2323	2930	3110
3321	4116	5485	5509	6150					

a. Compute the heights on each bin of the histogram with bins [0, 500], (500, 1000], and so on.

b. Compute the value of the empirical distribution function in the endpoints of the bins.

c. Check that the area under the histogram on bin (1000, 1500] is equal to the increase $F_n(1500) - F_n(1000)$ of the empirical distribution function on this bin. Actually, this is true for each single bin (see Exercise 15.11).

15.5 ⊡ Suppose we construct a histogram with bins [0,1], (1,3], (3,5], (5,8], (8,11], (11,14], and (14,18]. Given are the values of the empirical distribution function at the boundaries of the bins:

t	0	1	3	5	8	11	14	18
$F_n(t)$	0	0.225	0.445	0.615	0.735	0.805	0.910	1.000

Compute the height of the histogram on each bin.

15.6 ⊞ Given is the following information about a histogram:

Bin	Height
(0,2]	0.245
(2,4]	0.130
(4,7]	0.050
(7,11]	0.020
(11,15]	0.005

Compute the value of the empirical distribution function in the point $t = 7$.

15.7 In Exercise 15.2 a histogram was constructed for the *Challenger* data. On which bin does the empirical distribution function have the largest increase?

15.8 Define a function K by

$$K(u) = \cos(\pi u) \quad \text{for } -1 \leq u \leq 1$$

and $K(u) = 0$ elsewhere. Check whether K satisfies the conditions (K1)–(K3) for a kernel function.

15.9 On the basis of the duration of an eruption of the Old Faithful geyser, park rangers try to predict the waiting time to the next eruption. In Figure 15.13 a scatterplot is displayed of the duration and the time to the next eruption in seconds.

a. Does the scatterplot give reason to believe that the duration of an eruption influences the time to the next eruption?

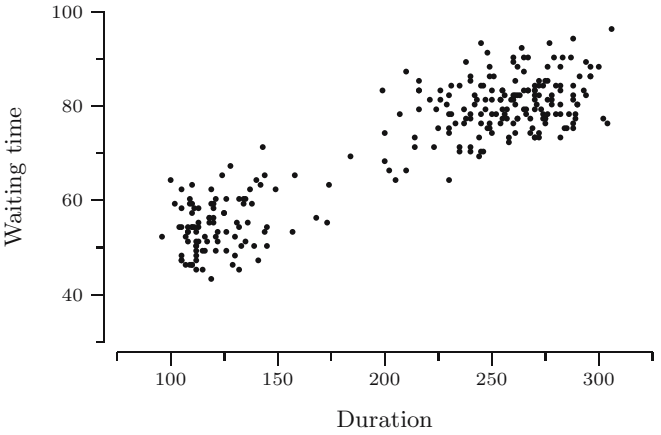

Fig. 15.13. Scatterplot of the Old Faithful data.

b. Suppose you have just observed an eruption that lasted 250 seconds. What would you predict for the time to the next eruption?

c. The dataset of durations shows two modes, i.e., there are two places where the data accumulate (see, for instance, the histogram in Figure 15.1). How many modes does the dataset of waiting times show?

15.10 Figure 15.14 displays the graph of an empirical distribution function of a dataset consisting of 200 elements. How many modes does the dataset show?

Fig. 15.14. Empirical distribution function.

15.11 ⊞ Given is a histogram and the empirical distribution function F_n of the same dataset. Show that the height of the histogram on a bin $(a, b]$ is

equal to

$$\frac{F_n(b) - F_n(a)}{b - a}.$$

15.12 ⊞ Let $f_{n,h}$ be a kernel estimate. As mentioned in Section 15.3, $f_{n,h}$ itself is a probability density.

a. Show that the corresponding expectation is equal to

$$\int_{-\infty}^{\infty} t f_{n,h}(t)\,\mathrm{d}t = \bar{x}_n.$$

Hint: you might consult the solution to Quick exercise 15.4.

b. Show that the second moment corresponding to $f_{n,h}$ satisfies

$$\int_{-\infty}^{\infty} t^2 f_{n,h}(t)\,\mathrm{d}t = \frac{1}{n}\sum_{i=1}^{n} x_i^2 + h^2 \int_{-\infty}^{\infty} u^2 K(u)\,\mathrm{d}u.$$