

In Chap. 1, we highlighted that different variables contain different levels of information. When summarizing or visualizing one or more variable(s), it is this information which determines the appropriate statistical methods to use.

Suppose we are interested in studying the employment opportunities and starting salaries of university graduates with a master's degree. Let the variable  $X$  denote the starting salaries measured in €/year. Now suppose 100 graduate students provide their initial salaries. Let us write down the salary of the first student as  $x_1$ , the salary of the second student as  $x_2$ , and so on. We therefore have 100 observations  $x_1, x_2, \dots, x_{100}$ . How can we summarize those 100 values best to extract meaningful information from them? The answer to this question depends upon several aspects like the nature of the recorded data, e.g. how many observations have been obtained (either small in number or large in number) or how the data was recorded (either exact values were obtained or the values were obtained in intervals). For example, the starting salaries may be obtained as exact values, say 51,500 €/year, 32,350 €/year, etc. Alternatively, these values could have been summarized in categories such as low income (<30,000 €/year), medium income (30,000–50,000 €/year), high income (50,000–70,000 €/year), and very high income (>70,000 €/year). Another approach is to ask whether the students were employed or not after graduating and record the data in terms of “yes” or “no”. It is evident that the latter classification is less detailed than the grouped income data which is less detailed than the exact data. Depending on which conceptualization of “starting salary” we use, we need to choose the approach to summarize the data, that is the 100 values relating to the 100 graduated students.

---

## 2.1 Absolute and Relative Frequencies

**Discrete Data.** Let us first consider a simple example to illustrate our notation.

*Example 2.1.1* Suppose there are ten people in a supermarket queue. Each of them is either coded as “F” (if the person is female) or “M” (if the person is male). The collected data may look like

M, F, M, F, M, M, M, F, M, M.

There are now two categories in the data: male (M) and female (F). We use  $a_1$  to refer to the male category and  $a_2$  to refer to the female category. Since there are seven male and three female students, we have 7 values in category  $a_1$ , denoted as  $n_1 = 7$ , and 3 values in category  $a_2$ , denoted as  $n_2 = 3$ . The number of observations in a particular category is called the **absolute frequency**. It follows that  $n_1 = 7$  and  $n_2 = 3$  are the absolute frequencies of  $a_1$  and  $a_2$ , respectively. Note that  $n_1 + n_2 = n = 10$ , which is the same as the total number of collected observations. We can also calculate the **relative frequencies** of  $a_1$  and  $a_2$  as  $f_1 = f(a_1) = \frac{n_1}{n} = \frac{7}{10} = 0.7 = 70\%$  and  $f_2 = f(a_2) = \frac{n_2}{n} = \frac{3}{10} = 0.3 = 30\%$ , respectively. This gives us information about the proportions of male and female customers in the queue.

We now extend these concepts to a general framework for the summary of **data on discrete variables**. Suppose there are  $k$  categories denoted as  $a_1, a_2, \dots, a_k$  with  $n_j$  ( $j = 1, 2, \dots, k$ ) observations in category  $a_j$ . The **absolute frequency**  $n_j$  is defined as the number of units in the  $j$ th category  $a_j$ . The sum of absolute frequencies equals the total number of units in the data:  $\sum_{j=1}^k n_j = n$ . The **relative frequencies** of the  $j$ th class are defined as

$$f_j = f(a_j) = \frac{n_j}{n}, \quad j = 1, 2, \dots, k. \quad (2.1)$$

The relative frequencies always lie between 0 and 1 and  $\sum_{j=1}^k f_j = 1$ .

**Grouped Continuous Data.** Data on continuous variables usually has a large number ( $k$ ) of different values. Sometimes  $k$  may even be the same as  $n$  and in such a case the relative frequencies become  $f_j = \frac{1}{n}$  for all  $j$ . However, it is possible to define intervals in which the observed values are contained.

*Example 2.1.2* Consider the following  $n = 20$  results of the written part of a driving licence examination (a maximum of 100 points could be achieved):

28, 35, 42, 90, 70, 56, 75, 66, 30, 89, 75, 64, 81, 69, 55, 83, 72, 68, 73, 16.

We can summarize the results in class intervals such as 0–20, 21–40, 41–60, 61–80, and 81–100, and the data can be presented as follows:

Class intervals	0–20	21–40	41–60	61–80	81–100
Absolute frequencies	$n_1 = 1$	$n_2 = 3$	$n_3 = 3$	$n_4 = 9$	$n_5 = 4$
Relative frequencies	$f_1 = \frac{1}{20}$	$f_2 = \frac{3}{20}$	$f_3 = \frac{3}{20}$	$f_4 = \frac{9}{20}$	$f_5 = \frac{5}{20}$

We have  $\sum_{j=1}^5 n_j = 20 = n$  and  $\sum_{j=1}^5 f_j = 1$ .

**Table 2.1** Frequency distribution for discrete data

Class intervals ( $a_j$ )	$a_1$	$a_2$	...	$a_k$
Absolute frequencies ( $n_j$ )	$n_1$	$n_2$	...	$n_k$
Relative frequencies ( $f_j$ )	$f_1$	$f_2$	...	$f_k$

Now, suppose the  $n$  observations can be classified into  $k$  class intervals  $a_1, a_2, \dots, a_k$ , where  $a_j (j = 1, 2, \dots, k)$  contains  $n_j$  observations with  $\sum_{j=1}^k n_j = n$ . The relative frequency of the  $j$ th class is  $f_j = n_j/n$  and  $\sum_{j=1}^k f_j = 1$ . Table 2.1 displays the **frequency distribution** of a discrete variable  $X$ .

*Example 2.1.3* Consider the pizza delivery service data (Example 1.4.2, Appendix A.4). We are interested in the pizza deliveries by branch and generate the respective frequency table, showing the distribution of the data, using the `table` command in *R* (after reading in and attaching the data) as

```
table(branch)           # absolute frequencies
table(branch)/length(branch) # relative frequencies
```

**R**

$a_j$	Centre	East	West
$n_j$	421	410	435
$f_j$	$\frac{421}{1266} \approx 0.333$	$\frac{410}{1266} \approx 0.323$	$\frac{435}{1266} \approx 0.344$

We have  $n = \sum_j n_j = 1266$  deliveries and  $\sum_j f_j = 1$ . We can see from this table that each branch has a similar absolute number of pizza deliveries and each branch contributes to about one-third of the total number of deliveries.

## 2.2 Empirical Cumulative Distribution Function

Another approach to summarize and visualize the (frequency) distribution of variables is the **empirical cumulative distribution function**, often abbreviated as “ECDF”. As the name itself suggests, it gives us an idea about the cumulative relative frequencies up to a certain point. For example, say we want to know how many people scored up to 60 points in Example 2.1.2. Then, this can be calculated by adding the number of people in the class intervals 0–20, 21–40, and 41–60, which corresponds to  $n_1 + n_2 + n_3 = 1 + 3 + 3 = 7$  and is the **cumulative frequency**. If we want to know the relative frequency of people obtaining up to 60 points, we have to add the relative frequencies of the people in the class intervals 0–20, 21–40, and 41–60 as  $f_1 + f_2 + f_3 = \frac{1}{20} + \frac{3}{20} + \frac{3}{20} = \frac{7}{20}$ .

Before discussing the empirical cumulative distribution function in a more general framework, let us first understand the concept of ordered values. Suppose the values of height of four people are observed as  $x_1 = 180$  cm,  $x_2 = 160$  cm,  $x_3 = 175$  cm, and  $x_4 = 170$  cm. We arrange these values in an order, say ascending order, i.e. first the smallest value (denoted as  $x_{(1)}$ ) and lastly the largest value (denoted as  $x_{(4)}$ ). We obtain

$$\begin{aligned}x_{(1)} &= x_2 = 160 \text{ cm}, & x_{(2)} &= x_4 = 170 \text{ cm}, \\x_{(3)} &= x_3 = 175 \text{ cm}, & x_{(4)} &= x_1 = 180 \text{ cm}.\end{aligned}$$

The values  $x_{(1)}$ ,  $x_{(2)}$ ,  $x_{(3)}$ , and  $x_{(4)}$  are called **ordered values** for which  $x_{(1)} < x_{(2)} < x_{(3)} < x_{(4)}$  holds. Note that  $x_1$  is not necessarily the smallest value but  $x_{(1)}$  is necessarily the smallest value. In general, if we have  $n$  observations  $x_1, x_2, \dots, x_n$ , then the ordered data is  $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ .

Consider  $n$  observations  $x_1, x_2, \dots, x_n$  of a variable  $X$ , which are arranged in ascending order as  $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$  (and are thus on an at least ordinal scale). The **empirical cumulative distribution function**  $F(x)$  is defined as the cumulative relative frequencies of all values  $a_j$ , which are smaller than, or equal to,  $x$ :

$$F(x) = \sum_{a_j \leq x} f(a_j). \quad (2.2)$$

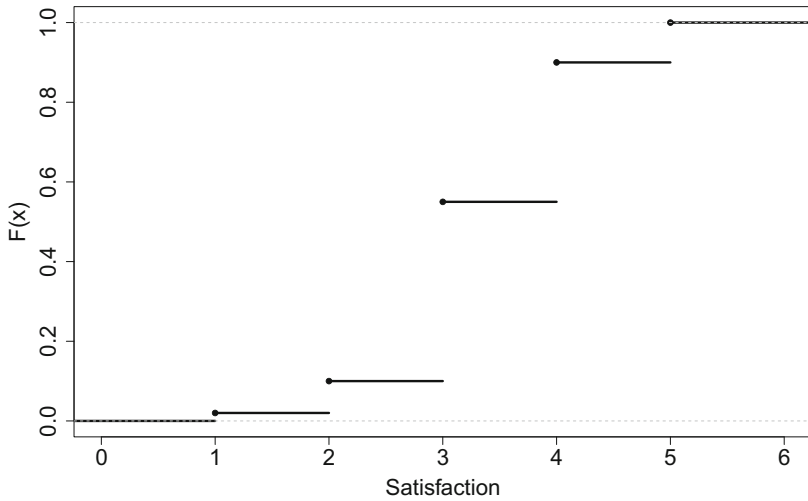
This definition implies that  $F(x)$  is a monotonically non-decreasing function,  $0 \leq F(x) \leq 1$ ,  $\lim_{x \rightarrow -\infty} F(x) = 0$  (the lower limit of  $F$  is 0),  $\lim_{x \rightarrow +\infty} F(x) = 1$  (the upper limit of  $F$  is 1), and  $F(x)$  is right continuous.

### 2.2.1 ECDF for Ordinal Variables

The empirical cumulative distribution function of ordinal variables is a **step function**.

*Example 2.2.1* Consider a customer satisfaction survey from a car service company. The 200 customers who had a car service done within the last 30 days were asked to respond regarding their overall level of satisfaction with the quality of the car service on a scale from 1 to 5 based on the following options: 1 = not satisfied at all, 2 = unsatisfied, 3 = satisfied, 4 = very satisfied, and 5 = perfectly satisfied. Based on the frequency of each option, we can calculate the relative frequencies and then plot the empirical cumulative distribution function, either manually (takes longer) or by using  $R$  (quick):

Satisfaction level ( $a_j$ )	$j = 1$	$j = 2$	$j = 3$	$j = 4$	$j = 5$
$n_j$	4	16	90	70	20
$f_j$	4/200	16/200	90/200	70/200	20/200
$F_j$	4/200	20/200	110/200	180/200	200/200



**Fig. 2.1** ECDF for the satisfaction survey

The  $F_j$ 's are calculated as follows:

$$\begin{aligned}
 F_1 &= f_1, & F_3 &= f_1 + f_2 + f_3, \\
 F_2 &= f_1 + f_2, & F_4 &= f_1 + f_2 + f_3 + f_4.
 \end{aligned}$$

The ECDF for this data can be obtained by summarizing the data in a vector and using the `plot.ecdf()` function in *R*, see Fig. 2.1:

```
sv <- c(rep(1,4), rep(2,16), rep(3,90), rep(4,70), rep(5,20))
plot.ecdf(sv)
```

The ECDF can be used to obtain the relative frequencies for values contained in certain intervals as

$$H(c \leq x \leq d) = \text{relative frequency of values } x \text{ with } c \leq x \leq d.$$

It further follows that:

$$H(x \leq a_j) = F(a_j) \tag{2.3}$$

$$H(x < a_j) = H(x \leq a_j) - f(a_j) = F(a_j) - f(a_j) \tag{2.4}$$

$$H(x > a_j) = 1 - H(x \leq a_j) = 1 - F(a_j) \tag{2.5}$$

$$H(x \geq a_j) = 1 - H(x < a_j) = 1 - F(a_j) + f(a_j) \tag{2.6}$$

$$H(a_{j_1} \leq x \leq a_{j_2}) = F(a_{j_2}) - F(a_{j_1}) + f(a_{j_1}) \tag{2.7}$$

$$H(a_{j_1} < x \leq a_{j_2}) = F(a_{j_2}) - F(a_{j_1}) \tag{2.8}$$

$$H(a_{j_1} < x < a_{j_2}) = F(a_{j_2}) - F(a_{j_1}) - f(a_{j_2}) \tag{2.9}$$

$$H(a_{j_1} \leq x < a_{j_2}) = F(a_{j_2}) - F(a_{j_1}) - f(a_{j_2}) + f(a_{j_1}) \tag{2.10}$$

*Example 2.2.2* Suppose, in Example 2.2.1, we want to know how many customers are not satisfied with their car service. Then, using the data relating to the responses “1” and “2”, we observe from the ECDF that  $(16 + 4)/200 \% = 10 \%$  of the customers were not satisfied with the car service. This relates to using rule (2.3):  $H(X \leq 2) = F(2) = 0.1$  or 10 %. Similarly, the proportion of customers who are more than satisfied can be obtained using (2.5) as  $H(X > 3) = 1 - H(x \leq 3) = 1 - 110/200 = 0.45 \%$  or 45 %.

## 2.2.2 ECDF for Continuous Variables

In general, we can apply formulae (2.2)–(2.10) to continuous data as well. However, before demonstrating their use, let us consider a somewhat different setting. Let us assume that a continuous variable of interest is only available in the form of grouped data. We may assume that the observations within each group, i.e. each category or each interval, are distributed uniformly over the entire interval. The ECDF then consists of straight lines connecting the lower and upper values of the ECDF in each of the intervals. To understand this concept in more detail, we introduce the following notation:

$k$	number of groups (or intervals),
$e_{j-1}$	lower limit of $j$ th interval,
$e_j$	upper limit of $j$ th interval,
$d_j = e_j - e_{j-1}$	width of the $j$ th interval,
$n_j$	number of observations in the $j$ th interval.

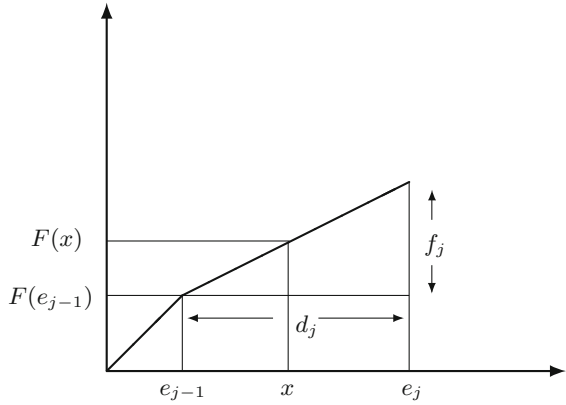
Under the assumption that all values in a particular interval are distributed uniformly within this interval, the empirical cumulative distribution function relates to a **polygonal chain** connecting the points  $(0, 0)$ ,  $(e_1, F(e_1))$ ,  $(e_2, F(e_2))$ ,  $\dots$ ,  $(e_k, 1)$ . The ECDF can then be defined as

$$F(x) = \begin{cases} 0, & x < e_0 \\ F(e_{j-1}) + \frac{f_j}{d_j}(x - e_{j-1}), & x \in [e_{j-1}, e_j] \\ 1, & x \geq e_k \end{cases} \quad (2.11)$$

with  $F(e_0) = 0$ . The idea behind (2.11) is presented in Fig. 2.2. For any interval  $[e_{j-1}, e_j]$ , the respective lower and upper limits of the ECDF are  $F(e_j)$  and  $F(e_{j-1})$ . If we assume values to be distributed uniformly over this interval, we can connect  $F(e_j)$  and  $F(e_{j-1})$  with a straight line. To obtain  $F(x)$  with  $x > e_{j-1}$  and  $x < e_j$ , we simply add the height of the ECDF between  $F(e_{j-1})$  and  $F(x)$  to  $F(e_{j-1})$ .

*Example 2.2.3* Consider Example 2.1.3 of the pizza delivery service. Suppose we are interested in determining the distribution of the pizza delivery times. Using the function `plot.ecdf()` in  $R$ , we obtain the ECDF of the continuous data, see Fig. 2.3a. Note that the structure of the curve is a step function but now almost looks like a continuous curve. The reason for this is that when the number of observations is large, then the lengths of class intervals become small. When these small lengths are

**Fig. 2.2** Illustration of the ECDF for continuous data available in groups/intervals\*



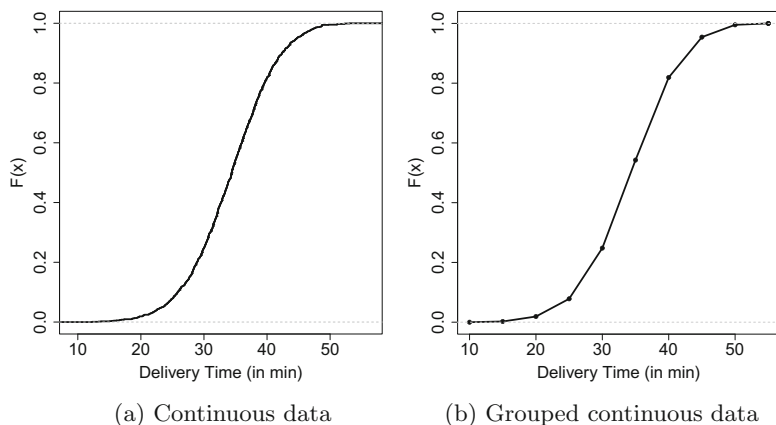
joined together, they appear like a continuous curve. As the number of observations increases, the smoothness of the curve increases too. If the number of observations is not large, e.g. suppose the data is reported as a summary from the drivers, i.e. whether the delivery took <15 min, between 15 and 20 min, between 20 and 25 min, and so on, then we can construct the ECDF by creating a table summarizing the data features as in Table 2.2.

Figure 2.3b shows the ECDF based on the grouped data evaluated in Table 2.2. It is interesting to see that the graphs emerging from the use of the grouped data and ungrouped data are similar in this specific example.

Suppose we are interested in calculating how many deliveries were completed within the desired time limit of 30 min, with a tolerance of maximum 10 % deviation, i.e. a deviation of 3 min. We can evaluate the ECDF at  $x = 33$  min.

**Table 2.2** The values needed to calculate the ECDF for the grouped pizza delivery time data in Example 2.2.3

Delivery time	$j$	$e_{j-1}$	$e_j$	$n_j$	$f_j$	$F(e_j)$
[0; 10]	1	0	10	0	0.0000	0.0000
(10; 15]	2	10	15	3	0.0024	0.0024
(15; 20]	3	15	20	21	0.0166	0.0190
(20; 25]	4	20	25	75	0.0592	0.0782
(25; 30]	5	25	30	215	0.1698	0.2480
(30; 35]	6	30	35	373	0.2946	0.5426
(35; 40]	7	35	40	350	0.2765	0.8191
(40; 45]	8	40	45	171	0.1351	0.9542
(45; 50]	9	45	50	52	0.0411	0.9953
(50; 55]	10	50	55	6	0.0047	1.0000



**Fig. 2.3** Empirical cumulative distribution function for pizza delivery time

Based on (2.11), we calculate  $H(X \leq 33) = F(33) = F(30) + f(6)/5(33 - 30) = 0.2480 + 0.2946/5 \cdot 3 = 0.42476$ . Thus, we conclude, based on the grouped data, that only about 42 % of the deliveries were completed in the desired time frame.

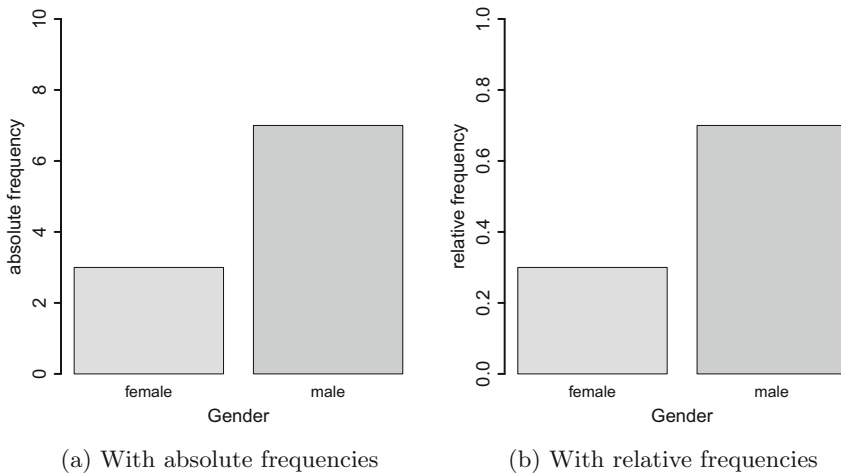
## 2.3 Graphical Representation of a Variable

Frequency tables and empirical cumulative distribution functions are useful in providing a numerical summary of a variable. Graphs are an alternative way to summarize a variable's information. In many situations, they have the advantage of conveying the information hidden in the data more compactly. Similarly, someone's mood can be more easily understood when looking at a smiley ☺ than by reading an essay about one's mood in a long paragraph.

### 2.3.1 Bar Chart

A simple tool to visualize the relative or absolute frequencies of observed values of a variable is a **bar chart**. A bar chart can be used for nominal and ordinal variables, as long as the number of categories is not very large. It consists of one bar for each category. The height of each bar is determined by either the absolute frequency or the relative frequency of the respective category and is shown on the  $y$ -axis. If the variable is measured on an ordinal level, then it is recommended to arrange the bars on the  $x$ -axis according to their ranks or values. If the number of categories is large, then the number of bars will be large too and the bar chart, in turn, may not remain informative.





**Fig. 2.4** Bar charts

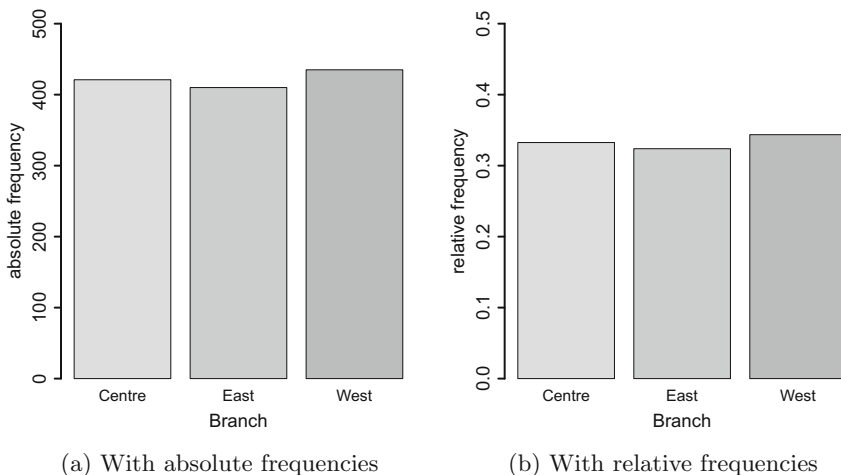
*Example 2.3.1* Consider Example 2.1.1 in which ten people, queuing in a supermarket, were classified as being either male (M) or female (F). The absolute frequencies for males and females are  $n_1 = 7$  and  $n_2 = 3$ , respectively. Since there are two categories, M and F, two bars are needed to construct the chart—one for the male category and another for the female category. The heights of the bars are determined as either  $n_1 = 7$  and  $n_2 = 3$  or  $f_1 = 0.7$  and  $f_2 = 0.3$ . These graphs are shown in Fig. 2.4.

*Example 2.3.2* Consider the data in Example 2.1.3, where the pizza delivery times for each branch are recorded over a period of 1 month. The frequency table forms the basis for the bar chart, either using the absolute or relative frequencies on the y-axis. Figure 2.5 shows the bar charts for the number and proportion of pizza deliveries per branch. The graphs can be produced in *R* by applying the `barplot` command to a frequency table:

```
barplot(table(branch))
barplot(table(branch)/length(branch))
```

**R**

*Remark 2.3.1* Instead of vertical bars, horizontal bars can be drawn using the optional argument `horiz=TRUE` in the `barplot` command.



**Fig. 2.5** Bar charts for the pizza deliveries per branch

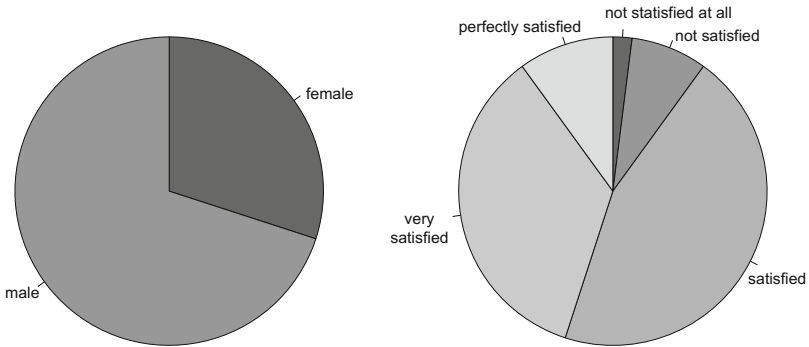
### 2.3.2 Pie Chart

Pie charts are another option to visualize the absolute and relative frequencies of nominal and ordinal variables. A pie chart is a circle partitioned into segments, where each of the segments represents a category. The size of each segment depends upon the relative frequency and is determined by the angle  $f_j \cdot 360^\circ$ .

*Example 2.3.3* To illustrate the construction of a pie chart, let us consider again Example 2.1.1 in which ten people in a supermarket queue were classified as being either male (M) or female (F): M, F, M, F, M, M, M, F, M, M. The pie chart for this data will have two segments: one for males and another one for females. The relative frequencies are  $f_1 = 7/10$  and  $f_2 = 3/10$ , respectively. The size of the segment for the first category (M) is  $f_1 \cdot 360^\circ = (7/10) \cdot 360^\circ = 252^\circ$ , and the size of the segment for the second category (F) is  $f_2 \cdot 360^\circ = (3/10) \cdot 360^\circ = 108^\circ$ . The pie chart is shown in Fig. 2.6a.

*Example 2.3.4* Consider again Example 2.2.1, where 200 customers were asked about their level of satisfaction (5 categories) with their car service. The pie chart for this example consists of five segments representing the categories 1, 2, 3, 4, and 5. The size of the  $j$ th segment is  $f_j \cdot 360^\circ$ ,  $j = 1, 2, 3, 4, 5$ . For example, for category 1, there are 4 out of 200 customers, who are not satisfied at all. The angle of the segment “not satisfied at all” therefore is  $f_1 \cdot 360^\circ = 4/200 \cdot 360^\circ = 7.2^\circ$ . Similarly, we can calculate the angle of the other segments and obtain a pie chart as shown in Fig. 2.6b using the `pie` command in R

```
pie(table(sv))
```



(a) For gender of people queuing

(b) For satisfaction with the car service

**Fig. 2.6** Pie charts

*Remark 2.3.2* Note that the area of a segment is *not* proportional to the absolute frequency of the respective category. Instead, the area of the segment is proportional to the angle  $f_j \cdot 360^\circ$  (and depends also on the radius of the whole circle). It has been argued that this may cause improper interpretations as the human eye may catch the segment's area more easily than the angle of a segment. Pie charts should therefore be used with caution.

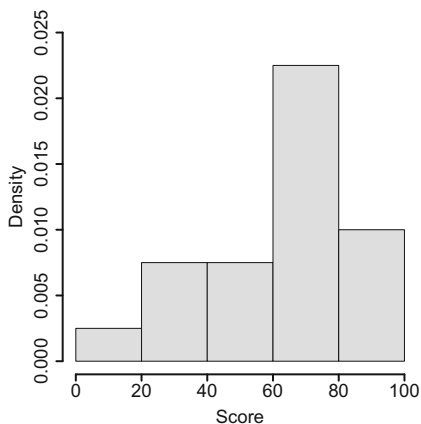
### 2.3.3 Histogram

If a variable consists of a large number of different values, the number of categories used to construct bar charts will consequently be large too. A bar chart may thus not give a clear summary when applied to a continuous variable. Instead, a **histogram** is the appropriate choice to represent the distribution of values of continuous variables. It is based on the idea to categorize the data into different groups and plot the bars for each category with height  $h_j = f_j/d_j$ , where  $d_j = e_j - e_{j-1}$  denotes the width of the  $j$ th class interval or category. An important consideration for this concept is that the area of the bars (=height  $\times$  width) is proportional to the relative frequency. This means that the widths of the bars need not necessarily to be the same because different widths can be adjusted with different heights of the bars.

*Example 2.3.5* Consider Example 2.1.2, where  $n = 20$  people were divided into five class intervals 0–20, 21–40, 41–60, 61–80, and 81–100 based on their performance in a written driving licence examination. The frequency table is given as

Class intervals	0–20	21–40	41–60	61–80	81–100
Absolute freq	$n_1 = 1$	$n_2 = 3$	$n_3 = 3$	$n_4 = 9$	$n_5 = 4$
Relative freq	$f_1 = \frac{1}{20}$	$f_2 = \frac{3}{20}$	$f_3 = \frac{3}{20}$	$f_4 = \frac{9}{20}$	$f_5 = \frac{5}{20}$
Height $f_j/d_j$	$h_1 = \frac{1}{400}$	$h_2 = \frac{3}{400}$	$h_3 = \frac{3}{400}$	$h_4 = \frac{9}{400}$	$h_5 = \frac{4}{400}$

**Fig. 2.7** Histogram for the scores of the people



The histogram for this grouped data set has five categories and therefore it has five bars. Since the widths of class intervals are the same, the heights of the bars are proportional to the relative frequency of the respective category. The resulting histogram is displayed in Fig. 2.7.

*Example 2.3.6* Recall Example 2.2.3 and the variable “pizza delivery time”. Table 2.3 shows the summary of the grouped data and the values needed to calculate the histogram. Figure 2.8a shows the histogram with equal widths of delivery time intervals. We see a symmetric distribution of the pizza delivery times, but many delivery times exceeding the target time of 30 min. If the histogram is required to have different widths for different bars, i.e. different delivery time intervals for different categories, then it can also be constructed as shown in Fig. 2.8b. This representation is different from Fig. 2.8a. The following commands in *R* are used to construct the histograms for absolute and relative frequencies, respectively:

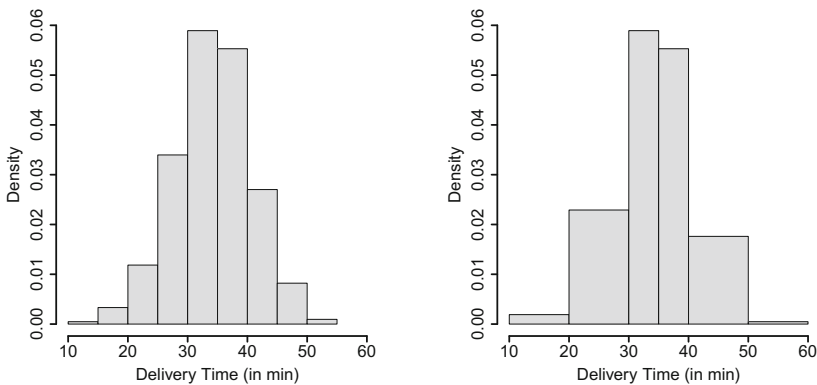
```
hist(time)           # show abs. frequencies
hist(time, freq=F)  # show rel. frequencies
```



*Remark 2.3.3* The *R* command `truehist()` from the library `MASS` presents an alternative to the `hist()` command. The default specifications are somewhat different, and many users prefer it to the command `hist`.

**Table 2.3** Values needed to calculate the histogram for the grouped pizza delivery time data

Delivery time	$j$	$e_{j-1}$	$e_j$	$d_j$	$f_j$	$h_j$
[0; 10]	1	0	10	10	0.0000	0.00000
(10; 15]	2	10	15	5	0.0024	0.00047
(15; 20]	3	15	20	5	0.0166	0.00332
(20; 25]	4	20	25	5	0.0592	0.01185
(25; 30]	5	25	30	5	0.1698	0.03397
(30; 35]	6	30	35	5	0.2946	0.05893
(35; 40]	7	35	40	5	0.2765	0.05529
(40; 45]	8	40	45	5	0.1351	0.02701
(45; 50]	9	45	50	5	0.0411	0.00821
(50; 55]	10	50	55	5	0.0047	0.00094

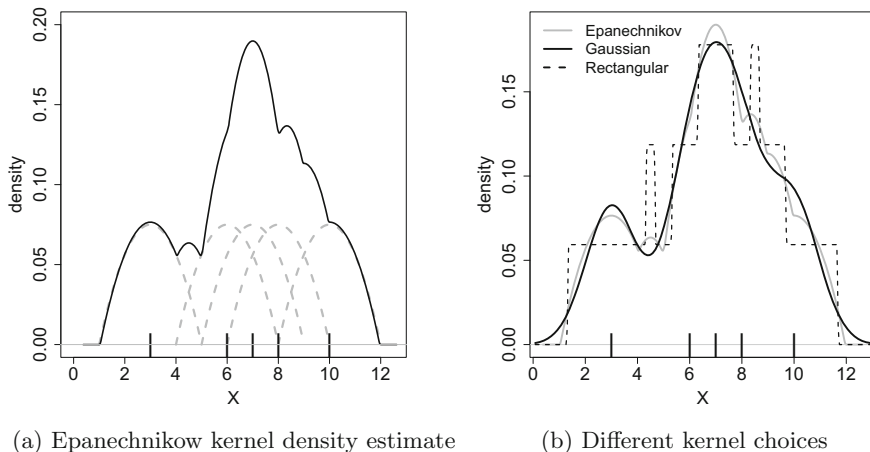


(a) With same width for each category (b) With different widths per category

**Fig. 2.8** Histogram for pizza delivery time

### 2.4 Kernel Density Plots

A disadvantage of histograms is that continuous data is categorized artificially. The choice of the class intervals is crucial for the final look of the graph. A more elegant way to deal with this problem is to smooth the histogram in the sense that each observation may contribute to different classes with different weights, and the distribution is represented by a continuous function rather than a step function. A **kernel density plot** can be produced by using the following function:



**Fig. 2.9** Construction of kernel density plots

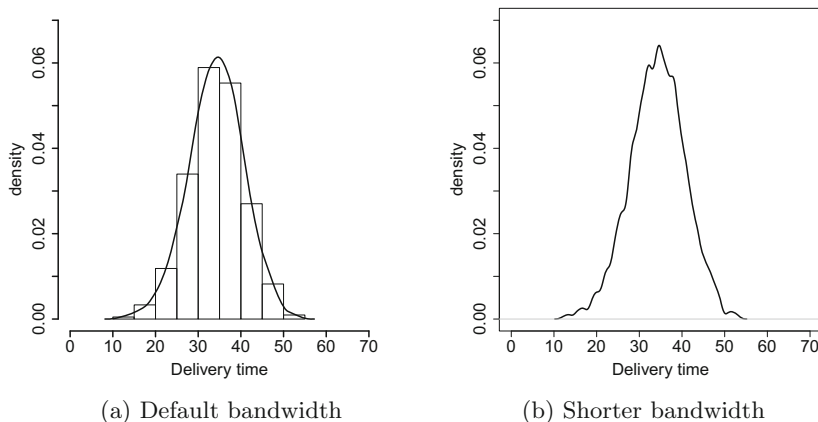
$$\hat{f}_n(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - x_i}{h}\right), \quad h > 0, \quad (2.12)$$

where  $n$  is the sample size,  $h$  is the bandwidth, and  $K$  is a kernel function, for example

$$K(x) = \begin{cases} \frac{1}{2} & \text{if } -1 \leq x \leq 1 \\ 0 & \text{elsewhere} \end{cases} \quad (\text{rectangular kernel})$$

$$K(x) = \begin{cases} \frac{3}{4}(1 - x^2) & \text{if } |x| < 1 \\ 0 & \text{elsewhere.} \end{cases} \quad (\text{Epanechnikov kernel})$$

To better understand this concept, consider Fig. 2.9a. The tick marks on the  $x$ -axis represent five observations: 3, 6, 7, 8, and 10. On each observation  $x_i$  as well as its surrounding values, we apply a kernel function, which is the Epanechnikov kernel in the figure. This means that we have five functions (grey, dashed lines), which refer to the five observations. These functions are largest at the observation itself and become gradually smaller as the distance from the observation increases. Summing up the functions, as described in Eq. (2.12), yields the solid black line, which is the kernel density plot of the five observations. It is a smooth curve, which represents the data distribution. The degree of smoothness can be controlled by the bandwidth  $h$ , which is chosen as 2 in Fig. 2.9a.



**Fig. 2.10** Kernel density plot for delivery time

The choice of the kernel may affect the overall look of the plot. Above, we have given the functions for the rectangular and Epanechnikov kernels. However, another common function for kernel density plots is the normal distribution function, which is introduced in Sect. 8.2.2, see Fig. 2.9b for a comparison of different kernels. The kernel which is based on the normal distribution is called the “Gaussian kernel” and is the default in *R*, where a kernel density plot can be produced combining the `plot` and `density` commands:

```
example <- c(3,6,7,8,10)
plot(density(example, kernel='gaussian'))
```

**R**

Please note that kernel functions are not defined arbitrarily and need to satisfy certain conditions, such as those required for probability density functions as explained in Chap. 7, Theorem 7.2.1.

*Example 2.4.1* Let us consider the pizza data which we introduced earlier and in Appendix A.4. We can summarize the delivery time by using a kernel density plot using the *R* command `plot(density(time))` and compare it with a histogram, see Fig. 2.10a. We see that the delivery times are symmetric around 35 min. If we shorten the bandwidth to a half of the default bandwidth (option `adjust=0.5`), the kernel density plot becomes more wiggly, which is illustrated in Fig. 2.10b.

## 2.5 Key Points and Further Issues

### Note:

- ✓ Bar charts and histograms are not the same graphical tools. Bar charts visualize the categories of nominal or ordinal variables whereas histograms visualize the distribution of continuous variables. A bar chart does not require to have ordered values on the  $x$ -axis, but a histogram always requires the values on the  $x$ -axis to be on a continuous scale and to be ordered. The interpretation of a histogram is simplified if the class intervals are equally sized, since then the heights of the rectangles of the histogram are proportional to the absolute or relative frequencies.
- ✓ The ECDF can be used only for ordinal and continuous variables, see Sect. 7.2 for the theoretical background of the cumulative distribution function.
- ✓ A pie chart summarizes observations from a discrete (nominal, ordinal or grouped continuous) variable. It is only useful if the number of different values (categories) is small. It is to be kept in mind that the area of each segment is not proportional to the absolute frequency of the respective category. The angle of the segment is proportional to the relative frequency of the respective category.
- ✓ Other possibilities to visualize the distribution of variables are, for example, box plots (Sect. 3.3) and stratified plots (Sects. 4.1.3, 4.3.1, and 4.4).

## 2.6 Exercises

*Exercise 2.1* Consider the results of the national elections in South Africa in 2014 and 2009:

Party	Results 2014 (%)	Results 2009 (%)
ANC (African National Congress)	62.15	65.90
DA (Democratic Alliance)	22.23	16.66
EFF (Economic Freedom Fighters)	6.35	–
IFP (Inkatha Freedom Party)	2.40	4.55
COPE (Congress of the People)	0.67	7.42
Others	6.20	5.47

- (a) Summarize the results of the 2014 elections in a bar chart. Do it manually and by using  $R$ .
- (b) How would you compare the results of the 2009 and 2014 elections? Offer a simple solution that can be represented in a single plot. Construct this plot in  $R$ .



*Exercise 2.2* Consider a variable  $X$  describing the time until the first goal was scored in the matches of the 2006 football World Cup competition. Only matches with at least one goal are considered, and goals during the  $x$ th minute of extra time are denoted as  $90 + x$ :

6	24	90+1	8	4	25	3	83	89	34	25	24	18	6
23	10	28	4	63	6	60	5	40	2	22	26	23	26
44	49	34	2	33	9	16	55	23	13	23	4	8	26
70	4	6	60	23	90+5	28	49	6	57	33	56	7	

- (a) What is the scale of  $X$ ?
- (b) Write down the frequency table of  $X$  based on the following categories:  $[0, 15)$ ,  $[15, 30)$ ,  $[30, 45)$ ,  $[45, 60)$ ,  $[60, 75)$ ,  $[75, 90)$ ,  $[90, 96)$ .
- (c) Draw the histogram for  $X$  with intervals relating to the groups from the frequency table.
- (d) Now use  $R$  to reproduce the histogram. Compare the histogram to a kernel density plot of your choice.
- (e) Calculate the empirical cumulative distribution function for the grouped data.
- (f) Use  $R$  to plot the ECDF (via a step function) for
  - (i) the original data and
  - (ii) the grouped data.
- (g) Consider the grouped data. Now assume that the values within each interval are distributed uniformly. Determine the proportion of first goals which occurred
  - (i) in the first half, i.e. during the first 45 min,
  - (ii) in the last 10 min or during the extra time,
  - (iii) between the 20th and 65th min, i.e. what is  $H(20 \leq X \leq 65)$ ?
- (h) Determine the time point at which in 80 % of the matches the first goal was scored at or before this time point.

*Exercise 2.3* Suppose we have the following information to construct a histogram for a continuous variable with 2000 observations:

$j$	$e_{j-1}$	$e_j$	$d_j$	$h_j$
1	0	1	1	0.125
2	1	4	3	0.125
3	4	7	3	0.125
4	7	8	1	0.125

- (a) Determine the relative frequencies for each interval.
- (b) Determine the absolute frequencies.

*Exercise 2.4* A university survey was conducted on 500 first-year students to obtain knowledge about the size of their accommodation (in square metres).

$j$	Size of accommodation ( $\text{m}^2$ ) $e_{j-1} \leq x \leq e_j$	$F(x)$
1	8–14	0.25
2	14–22	0.40
3	22–34	0.75
4	34–50	0.97
5	50–82	1.00

- Determine the absolute frequencies for each category.
- What proportion of people live in a flat of at least  $34 \text{ m}^2$ ?

*Exercise 2.5* Consider a survey in which 100 people were asked to rate on a scale from 1 to 10 how much they agree with the statement that “there is too much football on television”. The results are summarized below:

Score	0	1	2	3	4	5	6	7	8	9	10
Responses	0	1	3	8	8	27	30	11	6	4	2

- Calculate and draw the ECDF of the scores.
- Determine  $F(3)$  and  $F(9)$ .
- Consider the situation, where the data is summarized in the two categories “disagree” (score  $\leq 5$ ) and “agree” (score  $> 5$ ). What would the ECDF look like under the approach outlined in (2.11)? Determine  $F(3)$  and  $F(9)$  for the summarized data.
- Explain the differences between (b) and (c).

*Exercise 2.6* It is possible to produce professional graphics in *R*. However, it is advantageous to go beyond the default options. To demonstrate this, consider Example 2.1.3 about the pizza delivery data, which is described in Appendix A.4.

- Set the working directory in *R* (`setwd()`), read in the data (`read.csv()`), and attach the data. Draw a histogram of the variable “temperature”. Type `?hist`, and view the options. Adjust the histogram so that you are satisfied with (i) axes labelling, (ii) axes range, and (iii) colour. Now use the `lines()` command to add a dashed vertical line at  $65^\circ\text{C}$  (which is the minimum temperature the pizza should have at the time of delivery).
- Consider a different approach, which constructs plots by means of multiple layers using `ggplot2`. You need an Internet connection to install the package using the command `install.packages('ggplot2')`. Browse through the help

pages on <http://docs.ggplot2.org/current/>. Look specifically at the examples for `ggplot`, `qplot`, `scale_histogram`, and `scale_y_continuous`. Try to understand the roles of “aesthetics” and “geoms”. Now, after loading the library via `library(ggplot2)`, create a `ggplot` object for the pizza data, which declares “temperature” to be the  $x$ -variable. Now add a layer with `geom_histogram` to create a histogram with interval width of 2.5 and dark grey bars which are 50 % transparent. Change the  $y$ -axis labelling by adding the relevant layer using `scale_y_continuous`. Plot the graph.

- (c) Now create a normal bar chart for the variable “driver” in *R*. Type `?barplot` and `?par` to see the options one can pass on to `barchart()` to adjust the graph. Make the graph look good.
- (d) Now create the same bar chart with `ggplot2`. Use `qplot` instead of `ggplot` to create the plot. Use an option which makes each bar to consist of segments relating to the day of delivery, so that one can see the number of deliveries by driver to highlight during which days the drivers delivered most often. Browse through “themes” and “scales” on the help page, and add layers that make the background black and white and the bars on a grey scale.

→ Solutions to all exercises in this chapter can be found on p. 325

\*Source Toutenburg, H., Heumann, C., *Deskriptive Statistik*, 7th edition, 2009, Springer, Heidelberg



<http://www.springer.com/978-3-319-46160-1>

Introduction to Statistics and Data Analysis  
With Exercises, Solutions and Applications in R  
Heumann, C.; Schomaker, M.; Shalabh  
2016, XIII, 456 p. 89 illus., Hardcover  
ISBN: 978-3-319-46160-1