Chapter 2
One-Dimensional Frequency Distributions

2.1 One-Dimensional Distribution

The collection of information about class boundaries and relative or absolute frequencies constitutes the frequency distribution. For a single variable (e.g., height) we have a one-dimensional frequency distribution. If more than one variable is measured for each statistical unit (e.g., height and weight), we may define a two-dimensional frequency distribution. We use the notation X to denote the observed variable.

2.1.1 Frequency Distributions for Discrete Data

Suppose the variable X can take on k distinct values $x_j, j = 1, .., k$. Note that we index these distinct values or classes using the subscript j. We will denote n observations on the random variable by $x_i, i = 1, \ldots, n$. The context will usually make it clear whether we are referring to the k distinct values or the n observations. We will assume that $n > k$.

Frequency Table

For a discrete variable X, the frequency table displays the distribution of frequencies over the given categories. From now on we will speak of discrete variables to encompass categorical variables and discrete metric variables with few possible observations. Note that the sum of the frequencies across the various categories equals the number of observations, i.e., $\sum_{j=1}^{k} x_j = n$ (Table 2.1).

W.K. Härdle et al., *Introduction to Statistics*, DOI 10.1007/978-3-319-17704-5_2

Table 2.1 A frequency table

Values	Absolute frequencies	Relative frequencies
x_1	$h(x_1)$	$f(x_1)$
x_2	$h(x_2)$	$f(x_2)$
$\vdots$	$\vdots$	$\vdots$
x_j	$h(x_j)$	$f(x_j)$
$\vdots$	$\vdots$	$\vdots$
x_k	$h(x_k)$	$f(x_k)$
Total	n	1

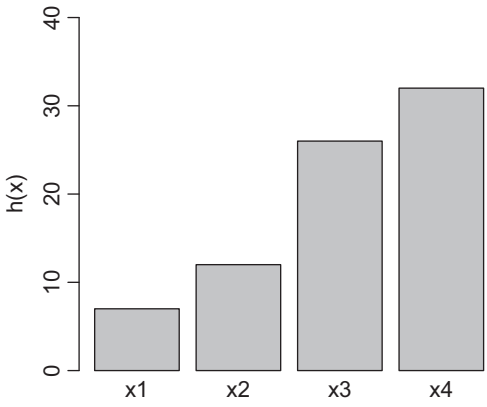


Fig. 2.1 Example of a bar graph

2.1.2 Graphical Presentation

Several graph types exist for displaying frequency distributions of discrete data.

Bar Graph

In a bar graph, frequencies are represented by the height of bars vertically drawn over the categories depicted on the horizontal axis. Since the categories do not represent intervals as in the case of grouped continuous data, the width of the bars cannot be interpreted meaningfully. Consequently, the bars are drawn with equal width (Fig. 2.1).

Stacked Bar Chart

Sometimes one wants to compare relative frequencies in different samples (different samples may arise at different points in time or from different populations). This can be done by drawing one bar graph for each sample. An alternative is the stacked bar

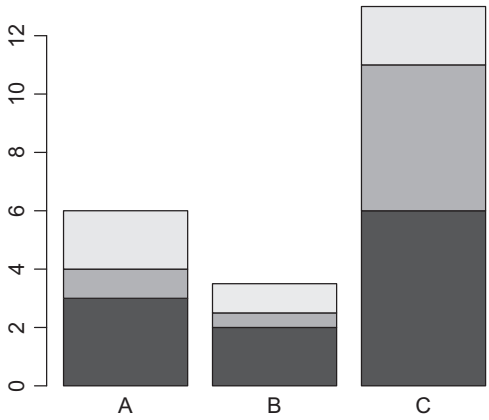


Fig. 2.2 Example of a stacked bar chart

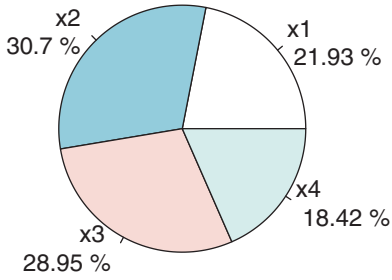


Fig. 2.3 Example of a pie chart

chart. It consists of as many segmented bars as there are samples. Each segment of a bar chart represents a relative frequency (Fig. 2.2).

Pie Chart

In pie charts, frequencies are displayed as segments of a pie. The area of each segment is proportional to the corresponding relative frequency (Fig. 2.3).

Pictograph

In a pictograph, the size or number of pictorial symbols is proportional to observed frequencies (Fig. 2.4).

Statistical Map

Different relative frequencies in different areas are visualized by different colors, shadings, or patterns (Fig. 2.5).

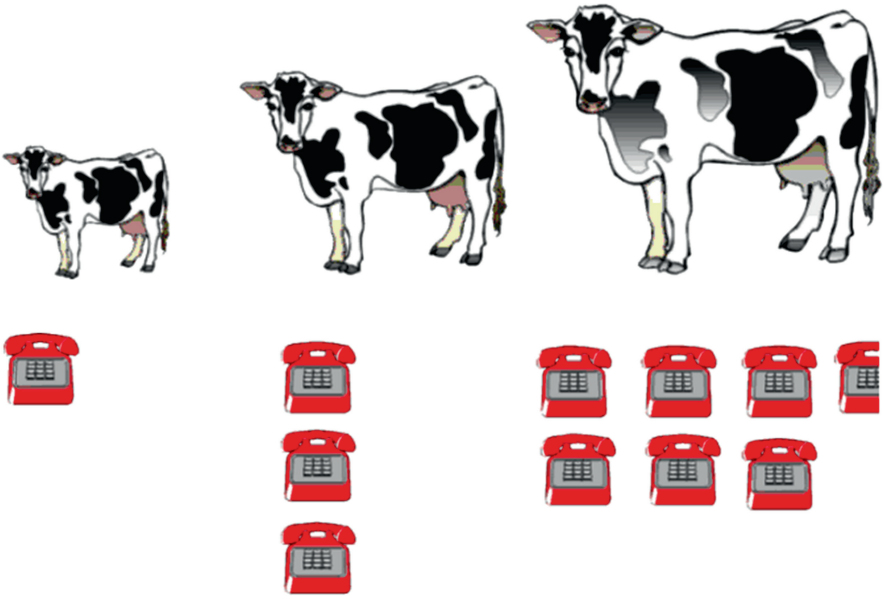


Fig. 2.4 Two examples of pictographs

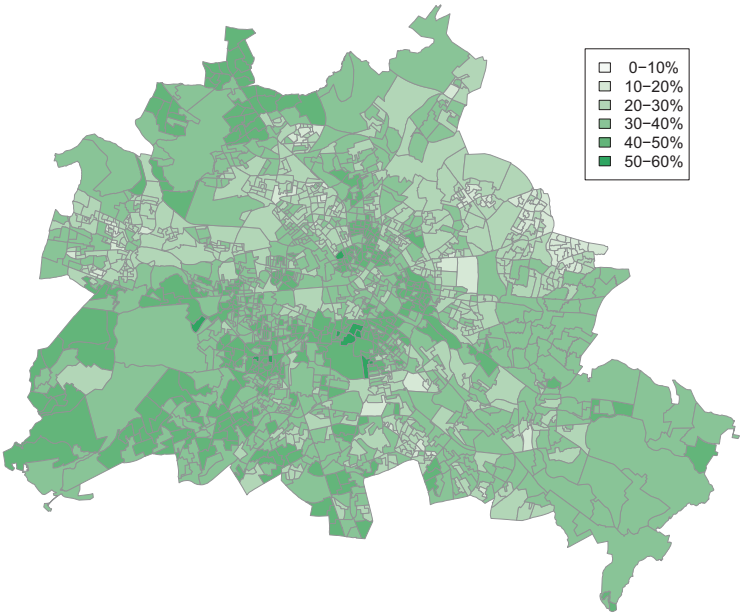


Fig. 2.5 Example of a statistical map

Table 2.2 Frequency table on employed population in Germany

j	Status x_j	$h(x_j)$ $(1000's)$	$f(x_j)$
1	Wage-earners	14568	0.389
2	Salaried	16808	0.449
3	Civil servants	2511	0.067
4	Self employed	3037	0.081
5	Family employed	522	0.014
	Total	37466	1.000

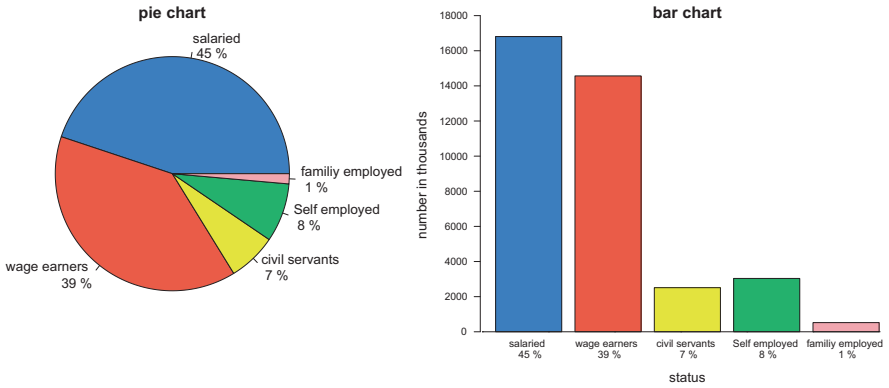


Fig. 2.6 Pie chart and bar graph on employed population in Germany

Explained: Job Proportions in Germany

In April 1991, Germany's employed population was surveyed with respect to type of employment. Table 2.2 summarizes the data. Visualizing the proportions helps us to analyze the data. In Fig. 2.6 you can clearly see the high proportion of wage-earners and salaried in contrast to the other categories.

Enhanced: Evolution of Household Sizes

The evolution of household sizes over the twentieth century can be studied using data compiled at various points in time.

Statistical elements: households

Statistical variable: size of household (metric, discrete)

Table 2.3 contains relative frequencies measured in percent for various years.

The structural shift in the pattern of household sizes towards the end of the century becomes visible if we draw bar charts for each year. The graphics in Fig. 2.7 display a clear shift towards smaller families during the twentieth century.

Table 2.3 Frequency table on the evolution of household sizes over the twentieth century

Household size X	1900	1925	1950	1990
1	7.1	6.7	19.4	35.0
2	14.7	17.7	25.3	30.2
3	17.0	22.5	23.0	16.7
4	16.8	19.7	16.2	12.8
≥ 5	44.4	33.3	16.1	5.3
Total	100	100	100	100

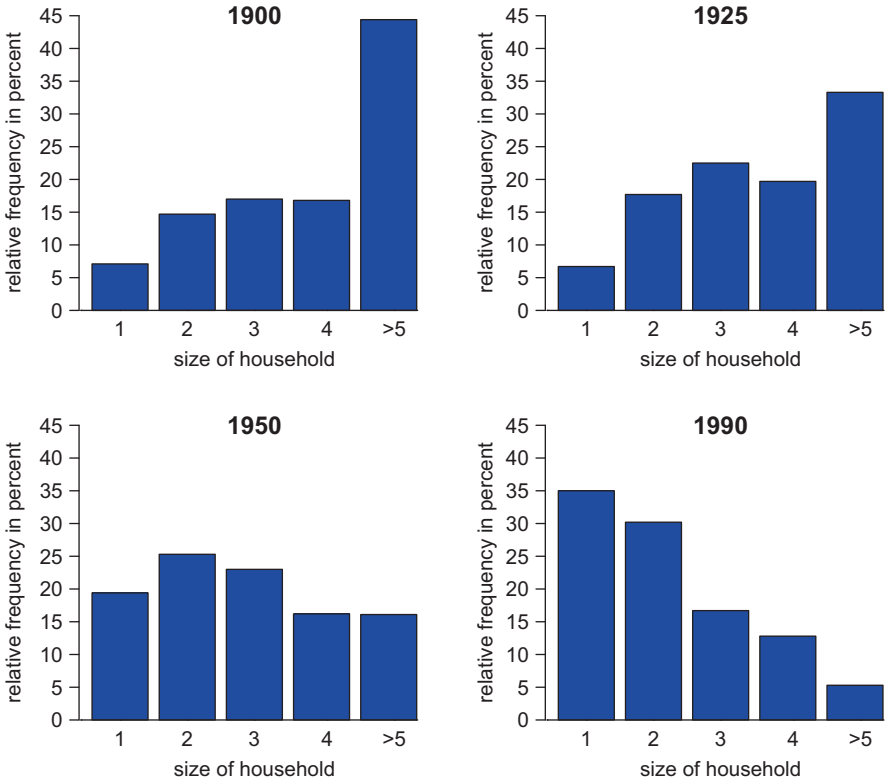


Fig. 2.7 Histograms on the evolution of household sizes over the twentieth century

2.2 Frequency Distribution for Continuous Data

Given a sample $x_1, x_2, \ldots, x_n$ on a continuous variable X, we may group the data into k classes with class boundaries denoted by $x_1^l, x_1^u = x_2^l, x_2^u = x_3^l, \ldots, x_k^u$ and class widths $\Delta x_j = x_j^u - x_j^l$ $(j = 1, \ldots, k)$. Note that the upper boundary for a given class is equal to the lower boundary for the succeeding class.

An observation x_i belongs to class j, if $x_j^l \leq x_i < x_j^u$. Since within a category, there are a range of possible values we will focus on the midpoint and denote it

Table 2.4 Structure of a frequency table

Class #	Classes	Absolute frequencies	Relative frequencies
1	$x_1^l \leq X < x_1^u$	$h(x_1)$	$f(x_1)$
2	$x_2^l \leq X < x_2^u$	$h(x_2)$	$f(x_2)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
j	$x_j^l \leq X < x_j^u$	$h(x_j)$	$f(x_j)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
k	$x_k^l \leq X < x_k^u$	$h(x_k)$	$f(x_k)$
	Total	n	1

by x_j. (Contrast this with the discrete data case where x_j denotes the value for the category.) Once again the subscript j corresponds to categories x_j, $j = 1, \ldots, k$ and the subscript i denotes observations x_i, $i = 1, \ldots, n$.

Frequency Table

A *frequency table* for continuous data provides the distribution of frequencies over the given classes. The structure of a frequency table is shown in Table 2.4.

Graphical Presentation

Histogram

In a histogram, continuous data that have been grouped into categories are represented by rectangles. Class boundaries are marked on the horizontal axis. As they can be of varying width, we cannot simply represent frequencies by the heights of bars as we did for bar graphs. Rather, we must correct for class widths. The rectangles are constructed so that their areas are equal to the corresponding absolute or relative frequencies.

$$\hat{h}(x_j) \cdot \Delta x_j = \frac{h(x_j)}{x_j^u - x_j^l} \cdot \left(x_j^u - x_j^l\right) = h(x_j)$$

or

$$\hat{f}(x_j) \cdot \Delta x_j = \frac{f(x_j)}{x_j^u - x_j^l} \cdot \left(x_j^u - x_j^l\right) = f(x_j)$$

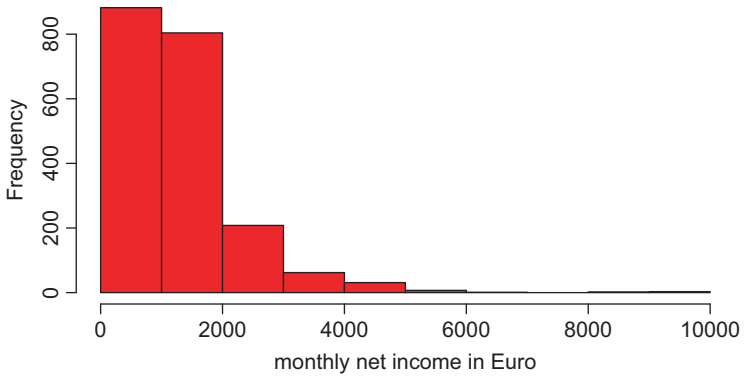


Fig. 2.8 Example of histogram—716 observations on monthly income (Euro)

If the class widths are identical, then the frequencies are also proportional to the heights of the rectangles. The rectangles are drawn contiguous to each other, reflecting common class boundaries $x_j^u = x_{j+1}^l$ (Fig. 2.8).

Stem-and-Leaf Display

In stem-and-leaf displays (plots), the data are not summarized using geometric objects. Rather, the actual values are arranged to give a rough picture of the data structure. The principle is similar to that of the bar chart, but values belonging to a particular class are recorded horizontally rather than being represented by vertical bars. Classes are set up by splitting the numerical observations into two parts: One or more of the leading digits make up the stem, the remaining (trailing) digits are called leaves. All observations with the same leading digits, i.e., the same stem, belong to one class. Typically, class frequencies are proportional to the lengths of the lines.

The principle is best understood by applying it to real data. Consider the following collection of observations :

$$32, 32, 35, 36, 40, 44, 47, 48, 53, 57, 57, 100, 105$$

The "stems" consist of the following "leading digits": $3, 4, 5, 10$. They correspond to the number of times that "ten" divides into the observation. The resulting stem-and-leaf diagram is displayed below.

Frequency	Stems	Leaves
4	3	2256
4	4	0478
3	5	377
2	10	05

Displaying data graphically (or, as is the case here, quasi-graphically), we can extract more relevant information than we could otherwise. (The human brain is comparatively efficient at storing and comparing visual patterns.)

The above stem-and-leaf plot appears quite simple. We can refine this by splitting the lines belonging to one stem in two, the first one for the trailing digits in the range one to four, the second for five to nine. We label the first group with *l* for low, the second with *h* for high. In the resulting stem-and-leaf plot the data appears approximately evenly distributed:

Frequency	Stems		Leaves
2	3	l	22
2	3	h	56
2	4	l	04
2	4	h	78
1	5	l	3
2	5	h	77
1	10	l	0
1	10	h	5

Yet there is an apparent gap between stems 5 and 10. It is indeed one of the advantages of stem-and-leaf plots that they are helpful in both giving insights into concentration of data in specific regions and spotting extraordinary or extreme observations. By labeling 100 and 105 as outliers we obtain a useful enhancement to the stem-and-leaf plot:

Frequency	Stems		Leaves
2	3	l	22
2	3	h	56
2	4	l	04
2	4	h	78
1	5	l	3
2	5	h	77
2	Extremes: 100, 105		

For an example with data conveying a richer structure of concentration and a more detailed stem structure have a look at the following examples for grouped continuous data.

Dotplots

Dotplots are used to graphically display small datasets. For each observation, a "dot" (a point, a circle or any other symbol) is plotted. Some data will take on the same values. Such ties would result in "overplotting" and thus would distort the display of the frequencies.

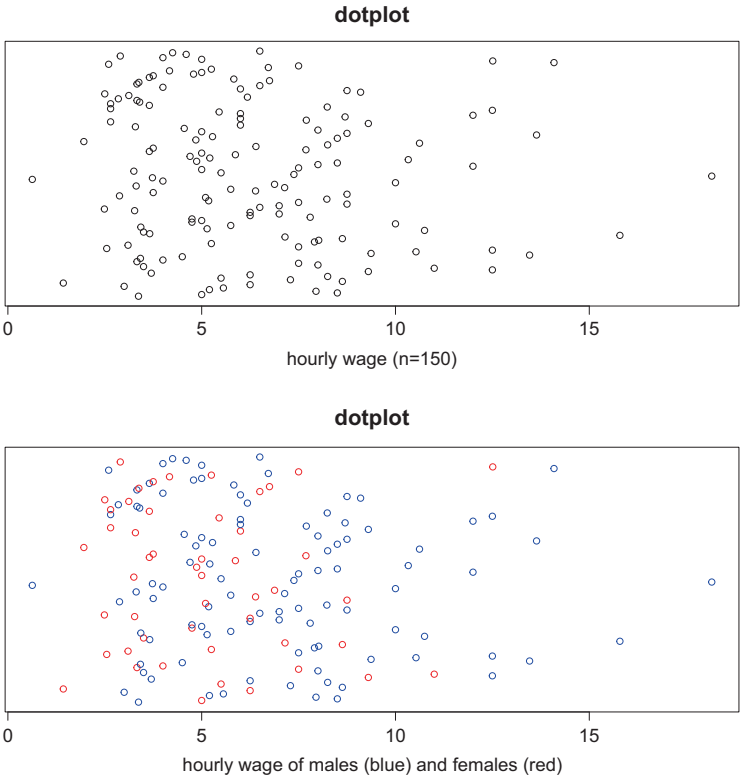


Fig. 2.9 Example of dotplot—student salaries in the USA

The dots are therefore spread out into the vertical dimension in a random fashion. The y-axis thus contains uniformly spread random numbers over the [0, 1] interval. Provided, the size of each symbol is sufficiently small for a given sample size, the dots are then unlikely to overlap each other.

Example The data in Fig. 2.9 consist of 150 observations on student salaries in the USA. In the upper part panel, we display a dot plot for all 150 observations. In the lower part, we use color to distinguish the gender of the students. Since the random perturbations in the vertical dimension are different for the two panels, the points are located in slightly different positions.

Explained: Petrol Consumption of Cars

Petrol consumption of 74 cars has been measured in miles per gallon (MPG). The measurements are displayed in a frequency table shown in Table 2.5. Using the same

Table 2.5 Petrol consumption of 74 cars in miles per gallon (MPG)

X: Petrol consumption (MPG)	Absolute frequencies $h(x_j)$	Relative frequencies $f(x_j)$
$12 \leq X < 15$	8	0.108
$15 \leq X < 18$	10	0.135
$18 \leq X < 21$	20	0.270
$21 \leq X < 24$	13	0.176
$24 \leq X < 27$	12	0.162
$27 \leq X < 30$	4	0.054
$30 \leq X < 33$	3	0.041
$33 \leq X < 36$	3	0.041
$36 \leq X < 39$	0	0.000
$39 \leq X < 42$	1	0.013
Total	74	1.000

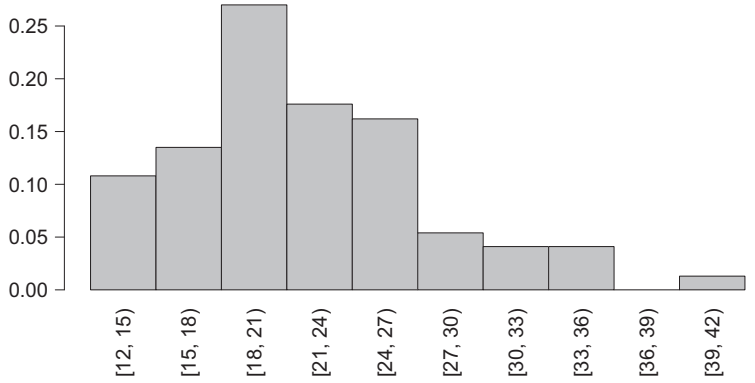


Fig. 2.10 Histogram for petrol consumption of 74 cars in miles per gallon (MPG)

constant class width of 3 MPG, the frequency distribution is displayed in a histogram in Fig. 2.10. As is evident from both, the frequency table and the histogram, the largest proportion of cars lies in the category 18–21 MPG.

Explained: Net Income of German Nationals

Data

Statistical elements:	German nationals, residing in private households, minimum age 18
Statistical variable:	monthly net income
sample size n	716

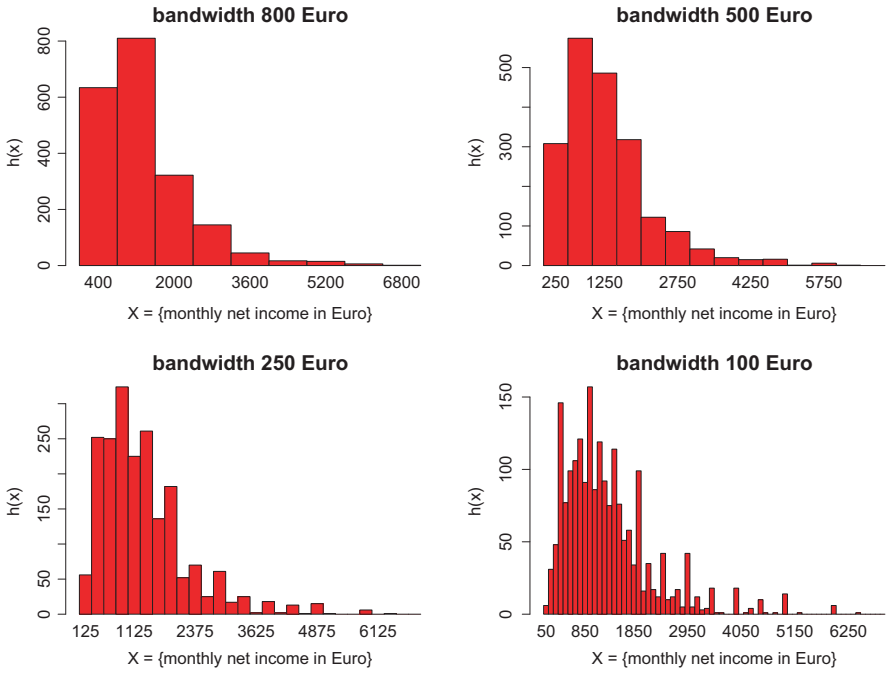


Fig. 2.11 Histograms of monthly net income in Euro for different bandwidths

Histogram

In the histograms shown in Fig. 2.11, the classes are income brackets of equal width. Reducing the common class size (and hence increasing the number of classes) yields a more detailed picture of the income distribution. Observe how the absolute frequencies decline as the class widths become more narrow.

Furthermore, increasing the number of classes decreases the smoothness of the graph. Additional gaps become visible as more information about the actual data is displayed. In choosing a class width we are striking a balance between two criteria: the essential information about the population which might be more strikingly conveyed in a smoother graph, and greater detail contained in a histogram with a larger number of classes.

We can also separate histograms by gender, using a bin width of 500 Euro, as shown in Fig. 2.12.

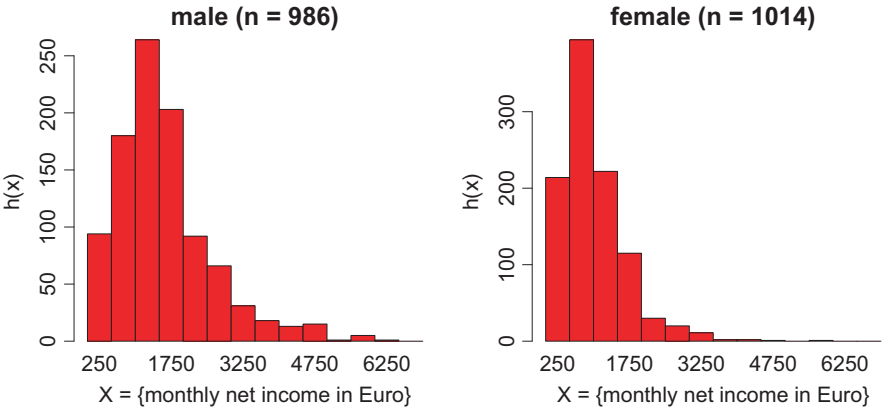


Fig. 2.12 Histograms of monthly net income in Euro for males and females

Stem-and-Leaf Display

The stem-and-leaf plot provided in Table 2.6 displays all 716 income figures. It is more detailed than the stem-and-leaf plots we have previously drawn. The stems, specified by the first leading digit, are divided into five subclasses corresponding to different values in the first trailing, i.e., leaf digit: The first line of each stem, denoted by *, lists all leaves starting with 0 or 1, the second (t) those starting with 2 or 3, and so on. As the stem width is specified to be 1000, the first leaf digit counts the hundreds. To condense exposition, each two observations belonging to the same class (i.e., being the same leaf) are represented by just one number (leaf). For example, six of the 716 surveyed persons earn between 2400 and 2500 Euros, denoted by "444" in the "2 f" line.

The ampersand (&) denotes pairs of observations covering both leaves represented by one line. For example, 4 persons earn between 4200 and 4400 Euros. Following the convention of each leaf representing two cases, there are two persons with net earnings in the interval [4200, 4300). The other two persons, symbolized by &, would be displayed by the sequence "23," if one leaf represented one observation. Thus, one of the two persons belongs to the income bracket [4200, 4300), the other to the [4300, 4400)-bracket.

Observe, that the 17 "extreme" values are displayed separately to highlight their distance from the other more heavily populated classes.

Table 2.6 Stem-and-leaf plot

Frequency	Stem and Leaf		
2	0	*	1
21	0	t	2233333333
35	0	f	44444444555555555
47	0	s	66666666666666667777777
41	0	.	88888888888899999999
45	1	*	000000000000000111111
38	1	t	2222222222222233333
63	1	f	44444444444555555555555555555
45	1	s	6666666666667777777777
72	1	.	88888888888888888888888889999999999
78	2	*	00000000000000000000000000000001111111
46	2	t	22222222222222333333333
32	2	f	444555555555555
28	2	s	66666667777777
23	2	.	88888889999
28	3	*	00000000000011
10	3	t	2233
16	3	f	44555555
8	3	s	6677
5	3	.	88
12	4	*	00000&
4	4	t	2&
10	Extremes:		(4400), (4500), (5000),(5500), (5600),(5900), (6400), (6500), (7000), (15000)
Stem width: 1000			
Each leaf: 2 case(s), & denotes fractional leaves			

2.3 Empirical Distribution Function

Empirical distribution functions can be constructed for data that have a natural numerical ordering. If $h\left(x_j\right)$ is the absolute frequency of observations on a discrete variable, then the absolute frequency (or number) of observations not exceeding that value is called the *absolute cumulated frequency*:

$$H\left(x_j\right) = \sum_{s=1}^{j} h\left(x_s\right), \qquad j = 1, \ldots, k$$

The *relative cumulative frequency* is calculated as:

$$F\left(x_j\right) = \frac{H\left(x_j\right)}{n} = \sum_{s=1}^{j} f\left(x_s\right), \qquad j = 1, \ldots, k$$

If the variable is continuous and the data are grouped into k classes, then the above definitions apply except that we interpret $H(x_j)$ as the frequency of observations not exceeding the upper boundary of the j-th class.

2.3.1 Empirical Distribution Function for Discrete Data

For the *relative cumulative frequency* we have

$$F(x) = \begin{cases} 0 & \text{if } x < x_1 \\ \sum_{s=1}^{j} f\left(x_s\right) & \text{if } x_j \leq x < x_{j+1}, \quad j = 2, \ldots, k \\ 1 & \text{if } x_k \leq x \end{cases}$$

The graph of an empirical distribution function is a monotonically increasing step function, the step size corresponds to the relative frequency at the "jump" points x_j (Table 2.7; Fig. 2.13).

In creating empirical distribution functions we are not losing information about relative frequencies of observations, as we can always reverse the cumulation process:

$$f\left(x_j\right) = F\left(x_j\right) - F\left(x_{j-1}\right), \quad \text{for } j = 1, \ldots, k; F\left(x_0\right) = 0$$

Suppose $x_l < x_u$ are two values that the discrete variable can take. Then the number or frequency of observations taking on values between x_l and x_u can be calculated as follows:

$$F\left(x_{u-1}\right) - F\left(x_l\right)$$

Table 2.7 Example of cumulative frequencies for number of persons in a household—data from 1990

# persons per household	$f\left(x_j\right)$	$F\left(x_j\right)$
1	0.350	0.350
2	0.302	0.652
3	0.167	0.819
4	0.128	0.947
≥ 5	0.053	1.000

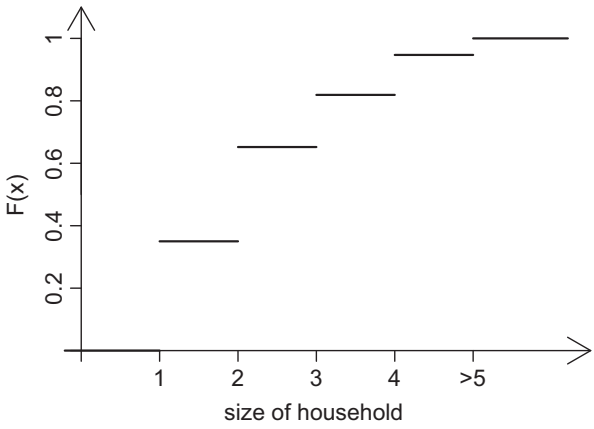


Fig. 2.13 Distribution function for the number of persons in a household—data from 1990

2.3.2 *Empirical Distribution Function for Grouped Continuous Data*

As for discrete data, the empirical distribution function for grouped continuous data is a function of relative cumulative frequencies. But in this case, rather than using a step function, one plots the cumulative frequencies against the upper boundaries of each class, then joints the points with straight lines. Mathematically, the empirical distribution function may be written as:

$$F(x) = \begin{cases} 0 & \text{if } x < x_1^l \\ \sum_{i=1}^{j-1} f(x_i) + \frac{x - x_j^l}{x_j^u - x_j^l} \cdot f\left(x_j\right) & \text{if } x_j^l \le x < x_j^u, \quad j = 1, \ldots, k \\ 1 & \text{if } x_k^u \le x \end{cases}$$

The rationale for interpolating with straight lines is that one might expect the distribution of points within classes to be approximately uniform.

An Example is provided in Table 2.8. The corresponding distribution function is given in Fig. 2.14.

As mentioned earlier, the straight lines connecting class boundaries reflect linear interpolations motivated by the assumption that observations are evenly distributed within classes. We will illustrate this by drawing the variable part of the distribution function for $x_j^l \le x < x_j^u$, $\sum_{i=1}^{j-1} f(x_i) + \frac{x - x_j^l}{x_j^u - x_j^l} f(x_j)$, for a fixed interval (class) $\left[x_j^l, x_j^u\right)$.

Evaluating at a lower class boundary we obtain $F\left(x_j^l\right) = \sum_{i=1}^{j-1} f(x_i) + \frac{x_j^l - x_j^l}{x_j^u - x_j^l} f(x_j) = \sum_{i=1}^{j-1} f(x_i)$. We can thus substitute $F\left(x_j^l\right)$ for $\sum_{i=1}^{j-1} f(x_i)$ in the

Table 2.8 Example—lives of 100 light bulbs

Statistical elements	Light bulbs			
Statistical variable	Life in hours, metric variable			
sample size n	100			
X: Life (hours)	$h(x_j)$	$f(x_j)$	$H(x_j)$	$F(x_j)$
$0 \leq X < 100$	1	0.01	1	0.01
$100 \leq X < 500$	24	0.24	25	0.25
$500 \leq X < 1000$	45	0.45	70	0.70
$1000 \leq X < 2000$	30	0.30	100	1.00
Total	100	1.00		

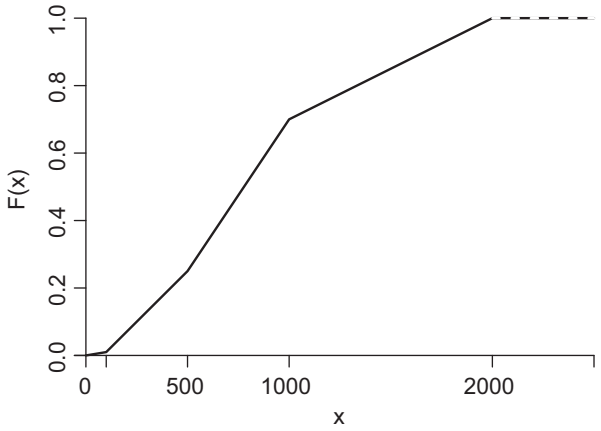


Fig. 2.14 Cumulative distribution function for lives of 100 light bulbs

formula for the distribution function and get

$$F(x) = F\left(x_j^l\right) + \frac{x - x_j^l}{x_j^u - x_j^l} \quad \text{if } x_j^l \leq x < x_j^u\,, \quad j = 1, \ldots, k$$

Figure 2.15 depicts the linear intra-class segment.

Explained: Petrol Consumption of Cars

The petrol consumption of 74 cars has been measured in miles per gallon (MPG). The measurements are displayed in an augmented frequency table shown in Table 2.9. The corresponding empirical distribution function is given in Fig. 2.16.

Again, the linear interpolation of lower class boundaries follows from the assumption of an even distribution of observations within classes. Class widths and boundaries are in turn constructed to approximate this assumption as closely as possible. This allows us to retain as much information as possible about the shape of the data.

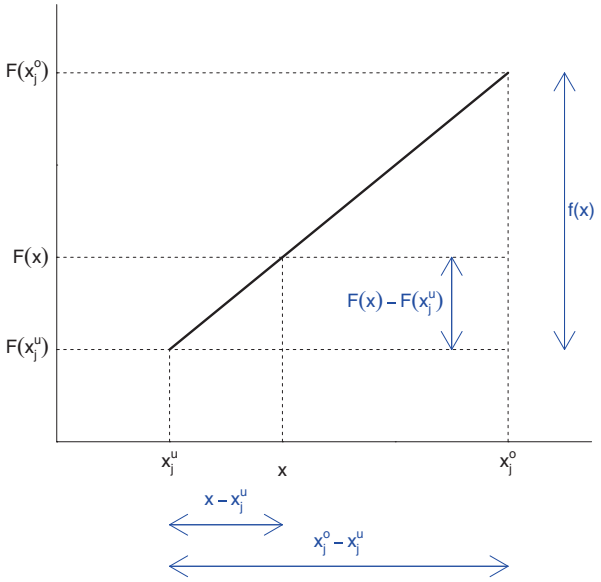


Fig. 2.15 Linear intra-class segment for distribution function

Table 2.9 Augmented frequency table for petrol consumption of 74 cars measured in miles per gallon (MPG)

X: Petrol consumption (MPG)	Absolute frequencies $h(x_j)$	Relative frequencies $f(x_j)$	Relative cumulative frequencies $F(x_j)$
$12 \leq X < 15$	8	0.108	0.108
$15 \leq X < 18$	10	0.135	0.243
$18 \leq X < 21$	20	0.270	0.513
$21 \leq X < 24$	13	0.176	0.689
$24 \leq X < 27$	12	0.162	0.851
$27 \leq X < 30$	4	0.054	0.905
$30 \leq X < 33$	3	0.041	0.946
$33 \leq X < 36$	3	0.041	0.987
$36 \leq X < 39$	0	0.000	0.987
$39 \leq X < 41$	1	0.013	1.000
Total	74	1.000	

Various statements can be extracted from Table 2.9, e.g.: 68.9 % of cars cannot travel more than 24 miles per gallon.

Explained: Grades in Statistics Examination

These are the grades 20 students have achieved in a Statistics examination:

$$\{2, 2, 4, 1, 3, 2, 5, 4, 2, 4, 3, 2, 5, 1, 3, 2, 2, 3, 5, 4\}$$

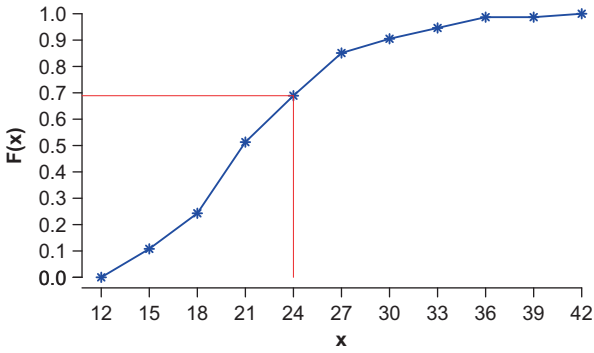


Fig. 2.16 Empirical distribution function for petrol consumption of 74 cars measured in miles per gallon (MPG)

Table 2.10 Frequency table of grades in statistics examination

X: Mark	Absolute frequency $h(x_j)$	Relative frequency $f(x_j)$	Relative cumulative frequency $F(x_j)$
1	2	0.10	0.10
2	7	0.35	0.45
3	4	0.20	0.65
4	4	0.20	0.85
5	3	0.15	1.00

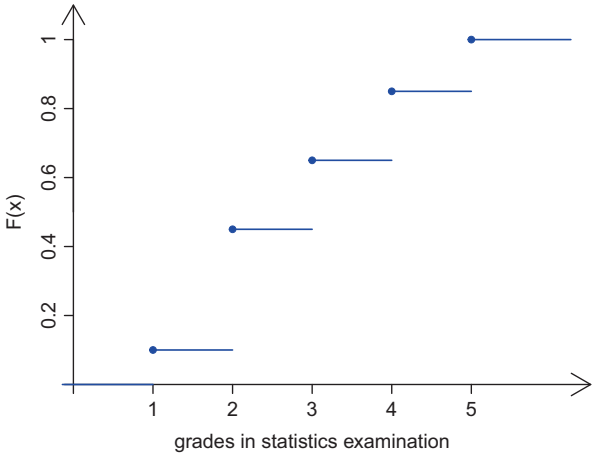


Fig. 2.17 Relative cumulative frequencies of grades in statistics examination

Table 2.10 summarizes the information about the distribution of the given data. The graph of the relative cumulative frequencies is depicted in Fig. 2.17. We observe that the graph of the relative cumulative frequency (and hence the function) is continuous from the right. Each bullet indicates the value of the distribution function at a jump point. In the figure, the x-axis covers all real numbers within the grade range, even though the random variable cannot take other values than $\{1,\ 2,\ 3,\ 4,\ 5\}$. For theoretical reasons, the definition of the distribution function

also assigns numbers (zero and one, respectively) to values outside [1, 5]. Various statements can be deduced from the data summarized in the frequency table, e.g.

- 65 % of students have achieved a grade of at least 3.
- 15 % (1.00 − 0.85) of students achieved a grade of 5.

2.4 Numerical Description of One-Dimensional Frequency Distributions

Statistics are numbers which summarize particular features of the data. Formally, a statistic is a function of the data. They can be used to measure different features, such as where the data are generally located (measures of location), the degree to which they are dispersed (measures of dispersion or scale), whether they are symmetrically distributed, the degree to which they are correlated, and so on. In the following sections we will consider various measures of location and dispersion. These measures can then be used to compare different datasets.

Measures of Location

In addition to summarizing where the data are located or concentrated, location measures provide a benchmark against which individual observations can be assessed.

Mode

The value occurring most frequently in a dataset is called the mode or the modal value. If the variable is discrete, the mode is simply the value with the greatest frequency. For continuous data measured with sufficient accuracy, however, most observations are likely to be distinct, rendering the idea meaningless. However, by grouping the data, we can determine the *modal class*, i.e., the class with the highest frequency.

Mode for qualitative or discrete data is given by

$$\arg\max_{x_j} \{f(x_j)\}$$

Mode for Grouped Continuous Data The modal class is the class with the highest class frequency. As a class interval consists of infinitely many numbers, we have to introduce a convention according to which a single number within this class is determined to represent the mode. The simplest convention is to use the midpoint of

the modal class. An alternative and more technical adjustment involves selecting a point which moves towards the neighboring cell with the higher density of observations. It is defined as follows:

$$x_D = x_j^l + \frac{\hat{f}\left(x_j\right) - \hat{f}\left(x_{j-1}\right)}{2 \cdot \hat{f}\left(x_j\right) - \hat{f}\left(x_{j-1}\right) - \hat{f}\left(x_{j+1}\right)} \cdot \left(x_j^u - x_j^l\right),$$

where

x_j^l,x_j^u	lower/upper boundary of modal class
$\hat{f}\left(x_j\right)$	frequency distribution for modal class
$\hat{f}\left(x_{j-1}\right)$	frequency distribution for class preceding modal class
$\hat{f}\left(x_{j+1}\right)$	frequency distribution for class succeeding modal class

The modal class is given by: $[500, 1000)$. We can calculate the mode approximated by the midpoint of the modal class which is just the arithmetic average of the class boundaries: $0.5 \cdot \left(x_j^u + x_j^l\right) = 750$ h. Using the above formula which moves the mid-point in the direction of the neighboring cell with the higher density of observations one obtains: $x_D = 500 + \frac{9-6}{18-6-3} \cdot 500 = 666\frac{2}{3}$ (Table 2.11).

Quantiles

Given data $x_1, x_2, \ldots, x_n$, suppose we order or rank the data in increasing order to obtain the ordered sequence $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$..We call the elements of this sequence the order statistics of the data. From the order statistics we can immediately read off the third largest value, the smallest value, and so on.

Let p be a number between zero and one and think of p as a proportion of the data. A value which divides the sequence of order statistics into the two subsequences containing the first $(p \cdot n)$ and the last $((1-p) \cdot n)$ observations is called the p-quantile. We will denote it by x_p. Equivalently, we may think of x_p as a value such that $100p\,\%$ of the data lie below it and $100(1-p)\,\%$ of the data lie above.

Table 2.11 Example—Lives of 100 light bulbs

j	X: Life (hours)	$h\left(x_j\right)$	$f\left(x_j\right)$	$\hat{f}\left(x_j\right) \cdot 10^{-4}$	$F\left(x_j\right)$
1	$0 \leq X < 100$	1	0.01	1	0.01
2	$100 \leq X < 500$	24	0.24	6	0.25
3	$500 \leq X < 1000$	45	0.45	9	0.70
4	$1000 \leq X < 2000$	30	0.30	3	1.00
	Total	100	1.00		

Quantiles for Ungrouped Data

- If $n \cdot p$ is not an integer and k the smallest integer satisfying $k > n \cdot p$, then we define $x_p = x_{(k)}$. The quantile is thus the observation with rank k, $x_{(k)}$.
- If, $k = n \cdot p$ is an integer, we will take x_p to be the midpoint between $x_{(k)}$.and $x_{(k+1)}$.

Quantiles for Grouped Data For data that are grouped in classes, we will carry out interpolations between class boundaries to obtain a p-quantile:

$$x_p = x_j^l + \frac{p - F\left(x_j^l\right)}{f\left(x_j\right)} \cdot \left(x_j^u - x_j^l\right)$$

Here, x_j^l, x_j^u and $f\left(x_j\right)$ are the lower boundary, upper boundary, and the relative frequency of the class containing the p-th quantile. The cumulative relative frequency up to and including the class preceding the quantile class is denoted by $F\left(x_j^l\right)$.

The quantile x_p can be defined using interpolation. The principle of interpolation for the quantity $p = F(x_p)$ can be easily understood from Fig. 2.18.

Some special quantiles:

- deciles (tenths)—the ordered observations are divided into ten equal parts. $p = s/10, s = 1, \ldots, 9$—deciles: $x_{0.1}, x_{0.2}, \ldots, x_{0.9}$
- quintiles—the ordered observations are divided into five equal parts. $p = r/5, r = 1, 2, 3, 4$—quintiles: $x_{0.2}, x_{0.4}, x_{0.6}, x_{0.8}$
- quartiles—the ordered observations are divided into four equal parts. $p = q/4, q = 1, 2, 3$—quartiles: $x_{0.25}, x_{0.5}, x_{0.75}$

Median (Central Value)

The value which divides the ordered observations into two equal parts is called the median $x_z = x_{0.5}$. The median is much less sensitive to outlying or extreme observations than other measures such as the mean which we study below. The median x_z corresponds to the second quartile $x_{0.5}$.

Median for Ungrouped Data

- for n odd: $x_{0.5} = x_{(\frac{n+1}{2})}$
- for n even: $x_{0.5} = (x_{(n/2)} + x_{(n/2+1)})/2$. This is simply the mid-point of the two center-most observations.

Median for Grouped Variables

- The median for grouped data is defined as the mid-point of the class which contains the central portion of the data.

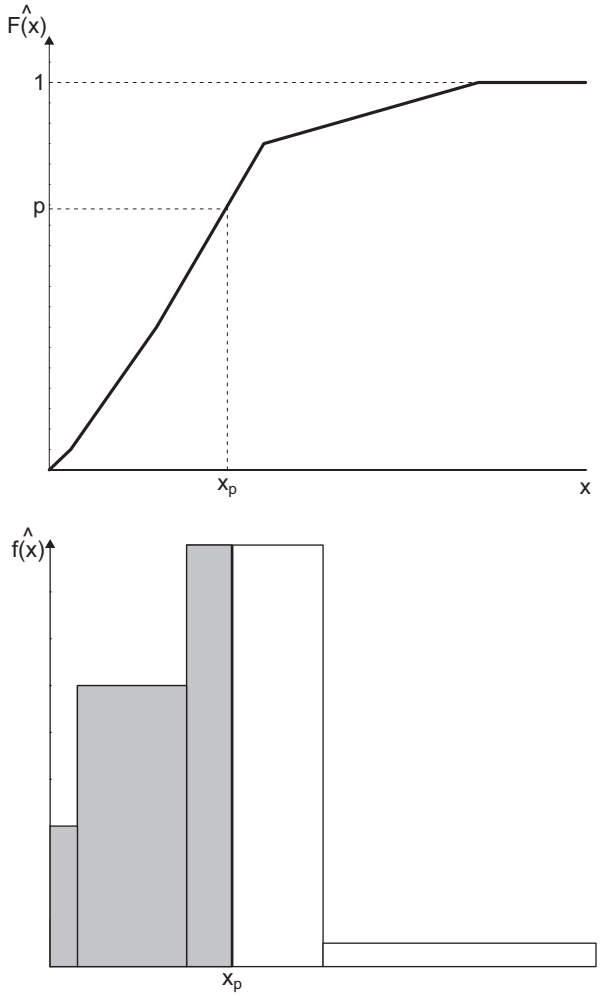


Fig. 2.18 Quantiles of grouped data

- Formally, let x_j^l and x_j^u be the lower and upper boundaries of the class for which $F(x_{j-1}^u) = F(x_j^l) \leq 0.5$ and $F(x_j^u) \geq 0.5$. Then,

$$x_{0.5} = x_j^l + \frac{0.5 - F(x_j^l)}{f(x_j)} \cdot (x_j^u - x_j^l)$$

- The median can be easily determined from the graph of the distribution function since $F(x_{0.5}) = 0.5$, see Fig. 2.19.

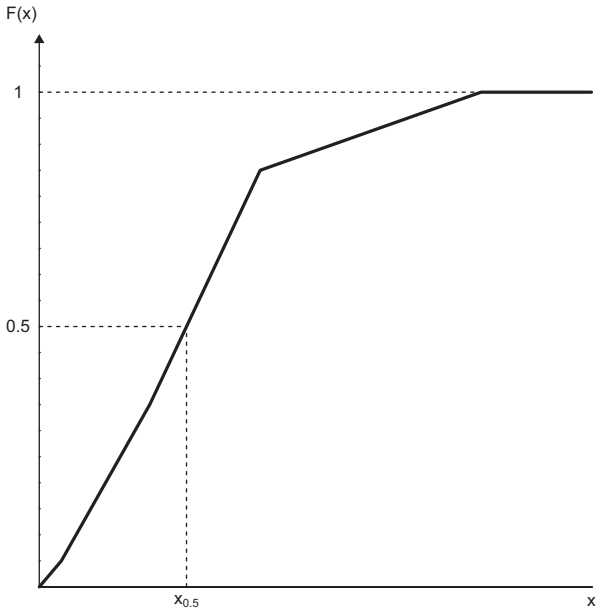


Fig. 2.19 Median for grouped continuous data

Properties of the Median (of Numerical Variables)

- optimality

$$\sum_{i=1}^{n} |x_i - x_{0.5}| = \sum_{j=1}^{k} |x_j - x_{0.5}| \cdot f(x_j) \rightarrow min.$$

  The median is optimal in the sense that it minimizes the sum of absolute deviations of the observations from a point that lies in the midst of the data (Fig. 2.19).
- linear transformation $y_i = a + bx_i \longrightarrow y_{0.5} = a + bx_{0.5}$

  If the data are transformed linearly, then the median is shifted by that same linear transformation.

Calculation of Quartiles The empirical distribution function (third column of the Table 2.12) implies that both the first quartile $x_{0.25},\ p = 0.25$ and the second quartile $x_{0.5},\ p = 0.50$ belong to third group (3000–5000 EUR). By interpolation we find

Table 2.12
Example—Monthly net income of households (up to 25000 EUR)

Income range (EUR)	Proportion of households: $f(x)$	Empirical distribution function: $F(x)$
1–800	0.044	0.044
800–1400	0.166	0.210
1400–3000	0.471	0.681
3000–5000	0.243	0.924
5000–25000	0.076	1.000

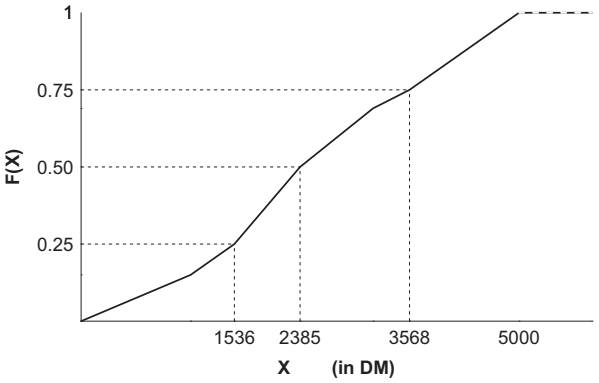


Fig. 2.20 Graph of the empirical distribution function and quartiles

the following (Fig. 2.20).

$$x_{0.25} = 1400 + 1600 \cdot \frac{0.25 - 0.21}{0.471} = 1535.88\,\text{EUR}$$

$$x_{0.50} = 1400 + 1600 \cdot \frac{0.50 - 0.21}{0.471} = 2385.14\,\text{EUR}$$

$$x_{0.75} = 3000 + 2000 \cdot \frac{0.75 - 0.681}{0.243} = 3567.90\,\text{EUR}$$

The Interpretation 25 % of the households has net monthly income not exceeding 1535.88 EUR and 75 % of the households has income higher than 1535.88 EUR (first quartile). 50 % of the households have income smaller than 2385.14 EUR and 50 % of the households have income higher than 2385.14 EUR (second quartile). 75 % of the households have income less than 3567.90 EUR and 25 % of the households have income exceeding 3567.90 EUR (third quartile).

The above also implies that 50 % of the households has net income between 1535.88 EUR and 3567.90 EUR.

Arithmetic Mean

The arithmetic mean or average, denoted $\bar{X}$, is obtained by summing all observations and dividing by n. The arithmetic mean is sensitive to outliers. In particular, an extreme value tends to "pull" the arithmetic mean in its direction.

The mean can be calculated in various ways, using the original data, using the frequency distribution and using the relative frequency distribution. For discrete data, each method yields a numerically identical answer.

Calculation using original data:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Calculation using the frequency and relative frequency distribution:

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{k} x_j h(x_j) = \sum_{j=1}^{k} x_j f(x_j)$$

Properties of the Arithmetic Mean

- Center of gravity: The sum of the deviations of the data from the arithmetic mean is equal to zero.

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0 \quad \Leftrightarrow \quad \sum_{j=1}^{k}(x_j - \bar{x})h(x_j) = 0$$

- Minimum sum of squares: The sum of squares of the deviations of the data from the arithmetic mean is smaller than the sum of squares of deviations from any other value c.

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 < \sum_{i=1}^{n}(x_i - c)^2$$

$$\sum_{j=1}^{k}(x_j - \bar{x})^2 h(x_j) < \sum_{j=1}^{k}(x_j - c)^2 h(x_j)$$

- Pooled data: Assume that the observed data are in disjoint sets $D_1, D_2, \ldots, D_r$, and that the arithmetic mean $\bar{x}_p$ for each of the sets is known. Then the arithmetic mean of all observed values (considered as one set) can be calculated using the formula

$$\bar{x} = \frac{1}{n}\sum_{p=1}^{r} \bar{x}_p n_p \qquad n = \sum_{p=1}^{r} n_p$$

where n_p denotes the number of observations in p-th group ($p = 1, \ldots, r$).

Table 2.13
Example 1—Monthly income of households (MIH)

MIH (EUR)	Proportion of households $f(x)$	Cumulative distribution function $F(x)$
1–800	0.044	0.044
800–1400	0.166	0.210
1400–3000	0.471	0.681
3000–5000	0.243	0.924
5000–25000	0.076	1.000

Table 2.14
Example 2—Monthly income of 716 people

$\bar{x} = 1881.40$ EUR
$x_{0.25} = 1092.50$ EUR
$x_{0.50} = 1800.00$ EUR
$x_{0.75} = 2400.00$ EUR
'mode' = 2000.00 EUR

- Linear transformation:

$$y_i = a + bx_i \longrightarrow \bar{y} = a + b\bar{x}$$

- Sum:

$$z_i = x_i + y_i \longrightarrow \bar{z} = \bar{x} + \bar{y}$$

From the data of Example 1 given in Table 2.13 we can calculate the arithmetic mean using the mid-points of the groups:

$$\bar{x} = 400 \cdot 0.044 + 1100 \cdot 0.166 + 2200 \cdot 0.471 + 4000 \cdot 0.243 + 15000 \cdot 0.076$$
$$= 17.6 + 182.6 + 1036.2 + 972 + 1140 = 3348.4\,\text{EUR}.$$

The arithmetic mean 3348.4 EUR is higher than the median calculated above (2385.14 EUR). This can be explained by the fact that the arithmetic mean is more sensitive to the relatively small number of large incomes. The high values shift the arithmetic mean but do not influence the median (Table 2.14).

Explained: Average Prices of Cars

This dataset contains prices (in USD) of 74 cars. The distribution of prices is displayed using a dotplot below. The price variable is on the horizontal axis. The data are randomly scattered in the vertical direction for better visualization.

In Fig. 2.21, the median is displayed in red and the arithmetic mean in magenta. As can be seen, the two values almost coincide.

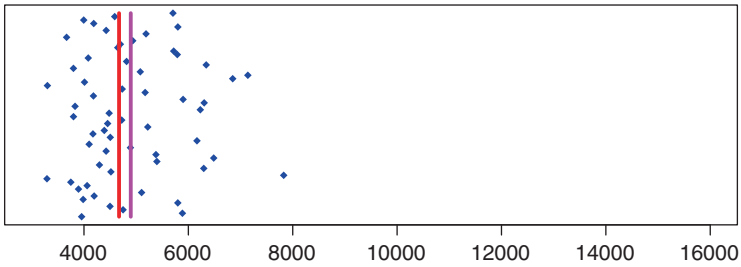


Fig. 2.21 Prices for 74 cars (USD)—arithmetic mean: 4896.417 (*magenta*) and median: 4672.000 (*red*)

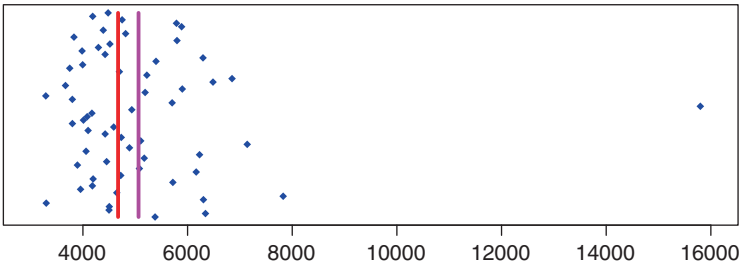


Fig. 2.22 Corrected prices for 74 cars (USD)—arithmetic mean: 5063.083 (*magenta*) and median: 4672.000 (*red*)

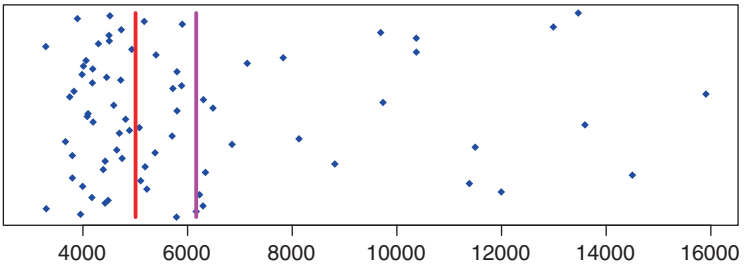


Fig. 2.23 Repeated measurements of car prices—arithmetic mean: 5063.083 (*magenta*) and median: 5006.500 (*red*)

For symmetric distributions, the median and arithmetic mean are identical. This is almost true for our example.

However, during a check of the data, it was discovered that one value had not been entered correctly. The value 15962 USD was incorrectly changed to 5962 USD. Figure 2.22 contains corrected values:

The median (because it is robust) did not change. On the other hand, the arithmetic mean has increased significantly, as it is sensitive to extreme values. The miscoded observation takes on a value well outside the main body of the data.

The measurements were repeated after some time with the results shown in Fig. 2.23.

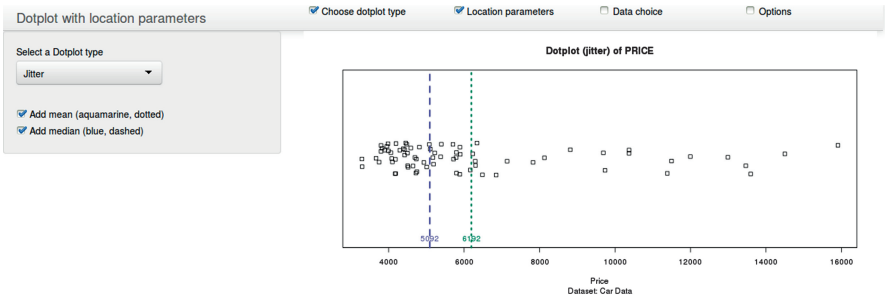


Fig. 2.24 Screenshot of the interactive example, available at http://u.hu-berlin.de/men_dot1

Now, there are a number of relatively more expensive cars. The distribution of prices is now skewed to the right. These more extreme observations pull the mean to the right much more so than the median. Thus for right-skewed distributions, the arithmetic mean is larger than the median.

Interactive: Dotplot with Location Parameters

The interactive example includes a number of sidebar panels. You can access the panels by setting a mark at the corresponding check box on the upper right.

Please select

- a dotplot type, e.g., jitter
- if you like the mean and median to be included in the plot

The last two panels allow you to choose a dataset or variable and to change the font size. For a detailed explanation on datasets and variables, please refer to Appendix A.

Output

The interactive example allows us to display a one-dimensional frequency distribution in the form of a dotplot for a variety of variables. Possible values are displayed along the horizontal axis. For easier visualization, the observations may be randomly shifted (jitter) in the vertical direction. The median and the arithmetic mean can be displayed graphically and numerically (Fig. 2.24).

Interactive: Simple Histogram

The interactive example includes a number of sidebar panels. You can access the panels by setting a mark at the corresponding check box on the upper right.

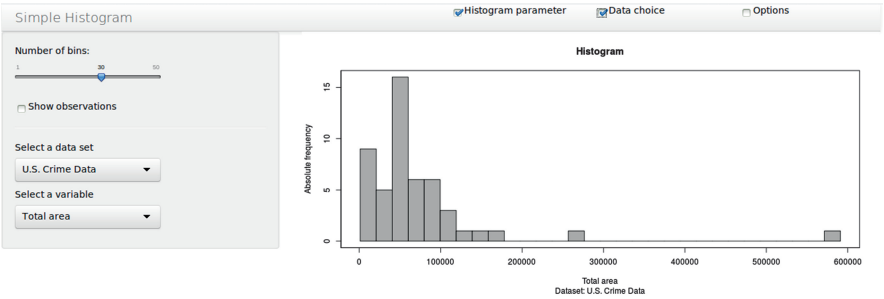


Fig. 2.25 Screenshot of the interactive example, available at http://u.hu-berlin.de/men_hist

Please select

- the number of bins
- if you like the observations to be shown

The last two panels allow you to choose a dataset or variable and to change the font size. For a detailed explanation on datasets and variables, please refer to Appendix A.

Output

The graphic displays all observations of a variable summarized in a histogram (Fig. 2.25).

2.5 Location Parameters: Mean Values—Harmonic Mean, Geometric Mean

If the observed variables are ratios, then the arithmetic mean may not be appropriate.

Harmonic Average

The harmonic average, denoted $\bar{x}_H$, is useful for variables which are ratios. We assume that all data points are not equal to zero, i.e., $x_i \neq 0$. As a consequence the $x_j \neq 0$.

$$\bar{x}_H = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

$$\bar{x}_H = \frac{\sum_{j=1}^{k} g_j}{\sum_{j=1}^{k} \frac{g_j}{x_j}}, \quad j = 1, \ldots, k$$

In the latter formula, g_j provides additional information which will become clear in the example below.

Example 1

Part of the road j	1	2	3	4
Distance g_j in km	2	4	3	8
Speed x_j in km/h	40	50	80	100

We would like to calculate the average speed of the car during the period of travel. It is inappropriate to simply average the speeds since they are measured over differing periods of time. In the table, g_j is the distance traveled in each segment. Using the above formula we calculate:

Total time: $\sum_{j=1}^{k} \frac{g_j}{x_j} = 0.2475\,\text{h}$

Total distance: $\sum_{j=1}^{k} g_j = 17\,\text{km}$

Average: $\bar{x}_H = \frac{17}{0.2475} = \frac{2+4+3+8}{\frac{2}{40}+\frac{4}{50}+\frac{3}{80}+\frac{8}{100}} = 68.687\,\text{km/h}$

The arithmetic mean would lead to an incorrect result 67.5 km/h, because it does not account for the varying lengths of the various parts of the road. Correct use of the arithmetic mean would involve calculating the time spent along each segment. In the above example these times are denoted by $h_j = g_j/x_j$ for each segment.

$$h_1 = g_1/x_1 = 0.05;\ h_2 = g_2/x_2 = 0.08;$$

$$h_3 = g_3/x_3 = 0.0375;\ h_4 = g_4/x_4 = 0.08;$$

$$\bar{x} = \frac{40 \cdot 0.05 + 50 \cdot 0.08 + 80 \cdot 0.0375 + 100 \cdot 0.08}{0.05 + 0.08 + 0.0375 + 0.08} = 68.687\,\text{km/h}$$

Thus, in order to calculate the average of ratios using additional information for the **numerator** (in our case x_j with the additional information g_j) we use the **harmonic average**. In order to calculate the average from ratios using additional information on the **denominator**, we choose the **arithmetic average**.

Example 2 Four students, who have part time jobs, have the hourly (respectively weekly) salaries given in Table 2.15.

We are supposed to find the average hourly salary. This calculation cannot be done using only the arithmetic average of the hourly salaries, because that would

Table 2.15 Hourly and weekly salary of four students

Student	Euro/h	Weekly salary in Euro
A	18	180
B	20	300
C	15	270
D	19	380

Table 2.16 Hourly salary and working hours of four students

Student	Euro/h	Working hours
A	18	10
B	20	15
C	15	18
D	19	20

not take into account the different times spent in the job. The variable of interest is a ratio (Euro/h) and the additional information (weekly salary in Euro) is related to the numerator of this ratio. Hence, we will use the harmonic average.

$$\bar{x}_H = \frac{\sum\limits_j g_j}{\sum\limits_j \frac{g_j}{x_j}} = \frac{180 + 300 + 270 + 380}{\frac{180}{18} + \frac{300}{20} + \frac{270}{15} + \frac{380}{19}} = \frac{1130}{63} = 17.94$$

These four students earn on average **17.94 Euro/h** (Table 2.15). The situation changes if we are given the number of hours worked per week (instead of the weekly salary).

Now, the additional information (weekly working hours) is related to the denominator of the ratio. Hence, we can use an arithmetic average, in this case the **weighted arithmetic average**.

$$\bar{x} = \frac{18 \cdot 10 + 20 \cdot 15 + 15 \cdot 18 + 19 \cdot 20}{10 + 15 + 18 + 20} = \frac{1130}{63} = 17.94$$

The average salary is again **17.94 Euro/h**.

Geometric Average

The geometric mean, denoted $\bar{x}_G$, is used to calculate the mean value of variables which are positive, are ratios (e.g., rate of growth) and are multiplicatively related.

$$\bar{x}_G = \sqrt[n]{x_1 \cdot x_2 \cdot \cdots \cdot x_n}$$

The logarithm of the geometric average is equal to the arithmetic average of the logarithms of the observations:

$$\log \bar{x}_G = \frac{1}{n} \sum_{i=1} n \log x_i$$

Mean Growth Rate and Forecast

Let $x_0, x_1, \ldots, x_n$ be the measurements ordered according to the time of observation from 0 to n. The growth rates can be calculated as

$$i_t = x_t / x_{t-1}$$
$$i_1 \cdot i_2 \cdot \cdots \cdot i_n = x_n / x_0$$

The product of all growth rates is equal to the total growth from time 0 to n. **The average growth rate** will be obtained as a geometric average of the growth rates in distinct time periods:

$$\bar{i}_g = \sqrt{n}\, i_1 \cdot i_2 \cdot \cdots \cdot i_n = \sqrt{n} \frac{x_n}{x_0}$$

Knowing the mean growth rate and the value in time n, we can **forecast** the value in time $n + T$.

$$x^{\star}_{n+T} = x_n \cdot (\bar{i}_G)^T$$

Solving this equation with respect to T, we obtain a formula for the time which is necessary to reach the given value:

$$T = \frac{\log(x_{n+T}) - \log(x_n)}{\log(\bar{i}_G)}$$

Example 1

Now we calculate:

- mean value (geometric average)
- forecast for 1990
- time (year), when GDP reaches the value 2500.

$$\bar{i}_G = \sqrt[8]{\frac{1971.8}{1733.8}} = 1.0162$$
$$x^{\star}_{1990} = 1971.8 \cdot 1.0162^2 = 2036.2 \,\text{bn DM}$$
$$T = \frac{\log(2500) - \log(1971.8)}{\log(1.0162)} = 14.77 \,\text{years}.$$

Table 2.17 Gross domestic product (GDP) for Germany in 1985 prices (bn DM)

Year	t	GDP x_t	i_t
1980	0	1733.8	–
1981	1	1735.7	1.0011
1982	2	1716.5	0.9889
1983	3	1748.4	1.0186
1984	4	1802.0	1.0307
1985	5	1834.5	1.0180
1986	6	1874.4	1.0217
1987	7	1902.3	1.0149
1988	8	1971.8	1.0365

Table 2.18 German stock index (DAX) during the period 1990–1997

Year	1989	1990	1991	1992	1993	1994	1995	1996	1997
DAX (end of the year)	1791	1399	1579	1546	2268	2107	2254	2889	4250
DAX (change)		−21.9 %	12.9 %	−2.1 %	46.7 %	−7.1 %	7.0 %	28.0 %	47.1 %

The value of GDP of 2500 is forecasted in year $1988 + 15 = 2003$ (Table 2.17).

Example 2

The German stock index (DAX) was changing during the period 1990–1997, as shown in Table 2.18.

We want to find the average yearly change in the DAX over the period. Use of the arithmetic average leads to an **incorrect result** as illustrated below.

- $\bar{x} = \frac{(-21.9)+(12.9)+(-2.1)+(46.7)+(-7.1)+(7.0)+(28.2)+(47.1)}{8} = \frac{110.80}{8} = 13.85\,\%$
- Starting in the year 1989 and using the "average change of DAX" to calculate the value of the DAX in 1997, one obtains:

  1990 $1791 \cdot 1.1385 = 2093$
  1991 $2093 \cdot 1.1385 = 2383$
  ... ...
  1997 $4440 \cdot 1.1385 = \mathbf{5055}$
- The result **5055** is much higher than the actual value of the DAX in 1997 which was **4250**.

The correct mean value is, in this case, the geometric mean, because it measure the growth during a certain period. The value of DAX in 1990 can be calculated from the value in 1989 and the relative change as follows:

$$\begin{aligned} \mathrm{DAX}_{1990} &= (1 + (-0.219)) \cdot \mathrm{DAX}_{1989} \\ &= (1 + (-0.219)) \cdot 1791 = 0.781 \cdot 1791 = 1399 \end{aligned}$$

Analogously, we can "forecast" the value for 1991 from the relative change and the value of DAX in 1990:

$$\begin{aligned}\text{DAX}_{1991} &= (1 + 0.129) \cdot \text{DAX}_{1990} \\ &= (1 + 0.129)) \cdot 1399 = 1.129 \cdot 1399 = 1579\end{aligned}$$

The values are multiplicatively related. The geometric mean yields the following:

$$\begin{aligned}X_G &= \sqrt[8]{0.781 \cdot 1.129 \cdot 0.979 \cdot 1.467 \cdot 0.929 \cdot 1.070 \cdot 1.282 \cdot 1.417} \\ &= 1.1141\end{aligned}$$

The average growth rate per year of the DAX over the period 1990–1997 was **11.41 %**. Using this geometric mean and the value of DAX in 1989 to predict the value of DAX in 1997, we obtain the correct result:

1990	$1791 \cdot 1.1141 = 1995$
1991	$1995 \cdot 1.1141 = 2223$
...	...
1997	$3815 \cdot 1.1141 =$ **4250**

The average growth rate of DAX in 1990–1997 can be used also to forecast the value of at the end of year 1999. We obtain the prediction:

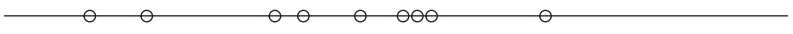
$$\text{DAX}_{1999} = \text{DAX}_{1997} \cdot 1.1141 \cdot 1.1141 = 4250 \cdot 1.1141^2 = 5275$$

2.6 Measures of Scale or Variation

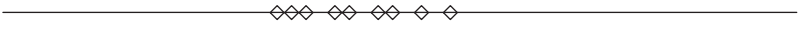
The various measures of location outlined in the previous sections are not sufficient for a good description of one-dimensional data. An illustration of this follows:

Monthly expenditures for free time and holidays (in EUR):

- data from 10 two person households: 210, 250, 340, 360, 400, 430, 440, 450, 530, 630 displayed on the axis:



- data from 10 four person households: 340, 350, 360, 380, 390, 410, 420, 440, 460, 490 displayed on the axis:



The arithmetic average $\bar{X}$ is in both cases is equal to 404 EUR, but the graphs show visible differences between the two distributions. For households with four people the values are more concentrated around the center (in this case the mean) than for households with two people, i.e., the spread or variation is smaller.

Measures of scale measure the variability of data. Together with measures of location (such as means, medians, and modes) they provide a reasonable description of one-dimensional data. Intuitively one would want measures of dispersion to have the property that if the same constant was added to each of the data-points, the measure would be unaffected. A second property is that if the data were spread further apart, for example through multiplication by a constant greater than one, the measure should increase.

Range

The range is the simplest measure of scale:

Range for Ungrouped Data

- The range, denoted R, is defined as the difference between the largest and the smallest observed value

$$R = x_{\max} - x_{\min} = x_{(n)} - x_{(1)}$$

  where $x_{(1)}, \ldots, x_{(n)}$ are the ordered data, i.e., the order statistics.

Range for Grouped Data

- For grouped data, the range R is defined as the difference between the upper bound of the last (highest) class x_k^u and the lower bound of the first (smallest) class x_1^l:

$$R = x_k^u - x_1^l$$

Properties

- For a linear transformation we have: $y_i = a + bx_i \longrightarrow R_y = |b|R_x$

  Note that addition of the constant a which merely shifts the data does not affect the measure of variability.

Interquartile Range

The interquartile range is the difference between the third quartile $x_{0,75}$ and the first quartile $x_{0,25}$:

$$QA = x_{0.75} - x_{0.25}$$

The interquartile range is the width of the central region which captures 50 % of the observed data. The interquartile range relative to the median is defined as $QA_r = QA/x_{0.5}$.

Properties

- Robust towards extreme values (outliers)
- Linear transformation: $y_i = a + bx_i \longrightarrow QA_y = |b|QA_x$
  Again addition of the constant a does not affect the measure of variability.

Mean Absolute Deviation

The mean of the absolute deviations of the observed values from a fixed point c is called the mean absolute deviation (MAD) and it is denoted by d. The fixed point c can be any value. Usually, it is chosen to be one of the measures of location; typically the mean $\bar{x}$ or median $x_{0.5}$.

As with the range and the interquartile range, adding the same constant to all the data. Multiplication by a constant rescales the measure by the absolute value of that same constant. Each of the formulas below may be used for ungrouped data. If the data have been grouped, then one would use the second formula where the x_j are mid-points of the classes, and $h(x_j)$ and $f(x_j)$ are the absolute and relative frequencies:

$$d = \frac{1}{n}\sum_{i=1}^{n}|x_i - c|$$

$$d = \frac{1}{n}\sum_{j=1}^{k}|x_j - c|h(x_j) = \sum_{j=1}^{k}|x_j - c|f(x_j)$$

Properties

- The optimality property of the median implies that the median is the value which minimizes the mean absolute deviation. Thus any other value substituted for c above would yield a larger value of this measure.
- For a linear transformation of the data: $y_i = a + bx_i \longrightarrow d_y = |b|d_x$

Example

- Observed values: 2, 5, 9, 20, 22, 23, 29
  $x_{0.5} = 20,\ d(x_{0.5}) = 8,29$
  $\bar{x} = 15.71,\ d(\bar{x}) = 8.90$

The Variance and the Standard Deviation

The mean of the squared deviations of the observed values from a certain fixed point c is called the mean squared error (MSE) or the mean squared deviation. The point c can be chosen ad libitum.

$$MQ(c) = \frac{1}{n}\sum_{i=1}^{n}(x_i - c)^2$$

$$MQ(c) = \frac{1}{n}\sum_{j=1}^{k}(x_j - c)^2 h(x_j) = \sum_{j=1}^{k}(x_j - c)^2 f(x_j)$$

The Variance If we choose the point c to be the mean $\bar{x}$, then the MSE is called the variance. The variance of the observed values will be denoted as s^2 and may be computed as follows.

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - \bar{x}^2$$

$$s^2 = \frac{1}{n}\sum_{j=1}^{k}(x_j - \bar{x})^2 h(x_j) = \sum_{j=1}^{k}(x_j - \bar{x})^2 f(x_j)$$

Standard Deviation The standard deviation (s) is defined as the square root of the variance.

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$s = \sqrt{\frac{1}{n}\sum_{j=1}^{k}(x_j - \bar{x})^2 h(x_j)} = \sqrt{\sum_{j=1}^{k}(x_j - \bar{x})^2 f(x_j)}$$

The variance s^2 (and therefore also the standard deviation s) is always greater than or equal to 0. Zero variance implies that the observed data are all identical and consequently do not have any spread.

Properties

- The mean squared error with respect to $\bar{x}$ (the variance) is smaller than the mean square error with respect to any other point c. This result can be proved as follows:

$$\begin{aligned} MSE(c) &= \frac{1}{n}\sum_{i=1}^{n}(x_i - c)^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x} + \bar{x} - c)^2 \\ &= \frac{1}{n}\left[\sum_{i=1}^{n}(x_i - \bar{x})^2 + 2(\bar{x} - c)\sum_{i=1}^{n}(x_i - \bar{x}) + n(\bar{x} - c)^2\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 + (\bar{x} - c)^2 = s^2 + (\bar{x} - c)^2 \end{aligned}$$

  The middle term of the middle line vanishes since $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$. These formulas imply that the mean square error $MSE(c)$ is always greater than or equal to the variance. Obviously equality holds only if $c = \bar{x}$.
- For linear transformations we have: $y_i = a + bx_i \longrightarrow s_y^2 = b^2 s_x^2,\ s_y = |b| s_x$
- Standardization: by subtracting the mean and dividing by the standard deviation one creates a new dataset for which the mean is zero and the variance is one. Let: $z_i = a + bx_i$, where $a = -\bar{x}/s_x,\ b = 1/s_x$, then

$$z_i = \frac{x_j - \bar{x}}{s_x}$$

$$\Rightarrow \bar{z} = 0, \quad s_z^2 = 1$$

Example

- Observed values: 2, 5, 9, 20, 22, 23, 29
- $x_{0.5} = 20 \qquad MSE(x_{0.5}) = 109.14$
- $\bar{x} = 15.71 \qquad MSE(\bar{x}) = Variance = 90.78$

Theorem (pooling) *Let us assume that the observed values (data) are divided into r groups with n_i $i = 1, .., r$ observations. Assume also that the means and variances in these groups are known. To obtain the variance s^2 of the pooled data we may use:*

$$s^2 = \sum_{i=1}^{r}\frac{n_i}{n}s_i^2 + \sum_{i=1}^{r}\frac{n_i}{n}(\bar{x_i} - \bar{x})^2$$

$\bar{x_1}, \ldots, \bar{x_r}$ are the arithmetic averages in the groups
$s_1^2, \ldots, s_r^2$ are the variances in the groups
$n_1, \ldots, n_r$ are numbers of observations in the groups, $n = n_1 + \cdots + n_r$

Variance Decomposition The above formula illustrates that the variance can be decomposed into two parts:

Total variance = variance *within* the groups + variance *between* the groups.

Coefficient of Variation In order to compare the standard deviations for different distributions, we introduce a relative measure of scale (relative to the mean), the so-called coefficient of variation. The coefficient of variation expresses variation as a percentage of the mean:

$$v = s/\bar{x} \quad \bar{x} > 0$$

Example The mean values and the standard deviations of two sets of observations are:

$\bar{x}_1 = 250 \quad s_1 = 10$
$\bar{x}_2 = 750 \quad s_2 = 30$

By comparing the standard deviations, we conclude that the variation in the second dataset is three times higher than the variation in the first. But, in this case it would be more appropriate to compare the coefficients of variation since the data have very different means:

$v_1 = 10/250 = 0.04$
$v_2 = 30/750 = 0.04$

The relative spread of both datasets is the same.

Explained: Variations of Pizza Prices

The price (in EUR) of Dr. Oetker pizza was collected in 20 supermarkets in Berlin (Fig. 2.26):

3.99; 4.50; 4.99; 4.79; 5.29; 5.00; 4.19; 4.90; 4.99; 4.79; 4.90; 4.69; 4.89; 4.49; 5.09; 4.89; 4.99; 4.29; 4.49; 4.19

- The average price for a pizza in these 20 supermarkets is **4.27 Euro (= mean)**
- The median price is **4.84 Euro (= median)**
- The difference between the highest and smallest price is **1.30 Euro (= range)**
- If the MAD is calculated around the mean it is **0.29 Euro (= MAD)** if calculated around the median it is **0.28 Euro (= MAD)**.
- 50 % of all prices lie in the interval between **4.49 Euro (quartile $x_{0.25}$) and 4.99 Euro (quartile $x_{0.75}$)**, this interval is of width **0.50 Euro (= interquartile range)**.2
- Mean square error around the mean is **0.12241 Euro^2 (= variance)**, the square root of the variance is **0.34987 Euro (= standard deviation)**.

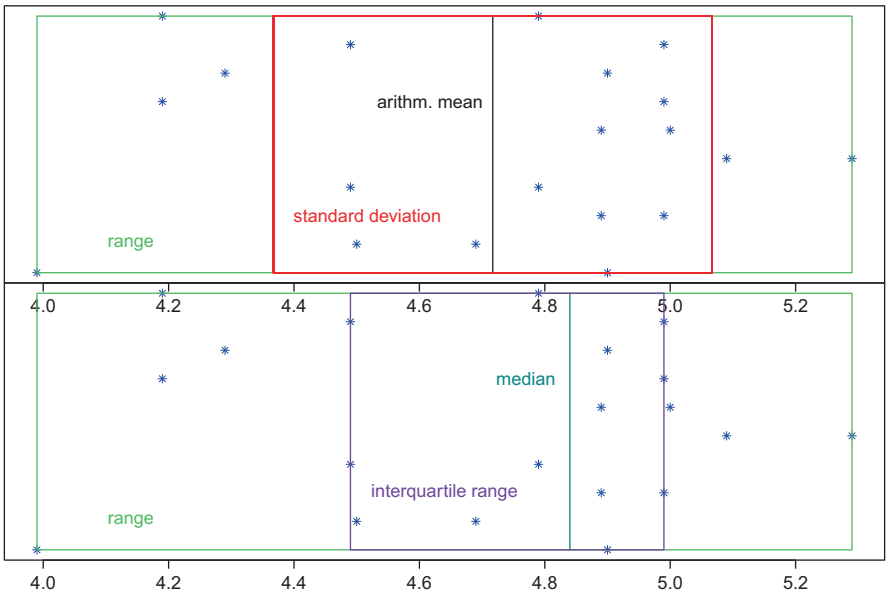


Fig. 2.26 Prices for pizza in 20 supermarkets—parameters of scale

Enhanced: Parameters of Scale for Cars

The price of 74 types of cars in USD was collected in 1985. The data are displayed in Fig. 2.27. The upper panel displays the range (green), arithmetic average (black), and the standard deviation (red). The lower panel displays the range (green), median (mint green), and the interquartile range (magenta).

Arithmetic average:	4896.417
Median:	4672
Range	4536
Interquartile range	1554.75
Standard deviation	991.2394

During a check of the data, it was discovered that there was an input error. The correct value of 15962 USD was incorrectly recorded as 5962 USD. Figure 2.28 contains the corrected results.

Arithmetic average:	5063.083
Median:	4672
Range	12508
Interquartile range	1554.75
Standard deviation	1719.064

It is clear that the range increased, because it is a function of the extreme values. The value of interquartile range did not change since no prices within this range were altered. The standard deviation increased significantly. The reason is that

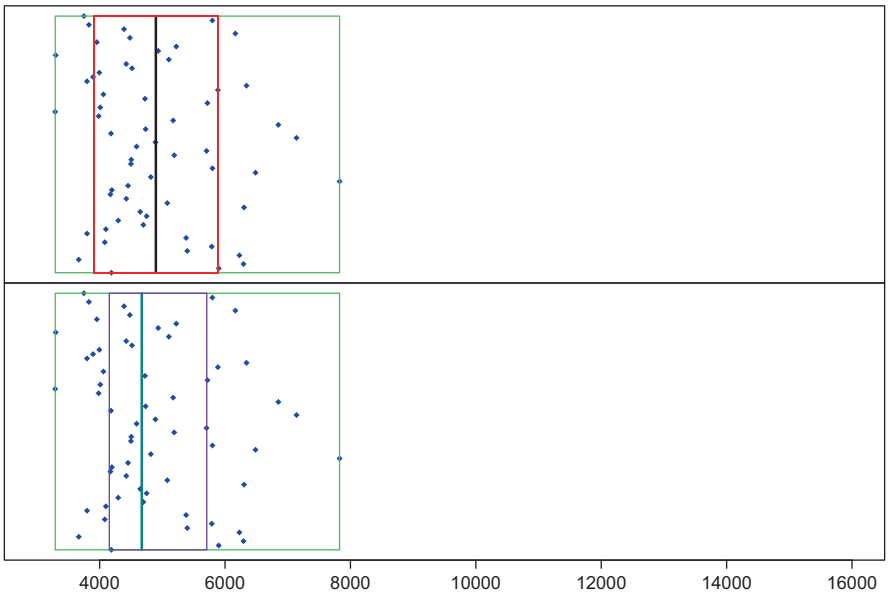


Fig. 2.27 Prices of 74 cars in USD—upper panel: range (*green*), arithmetic average (*black*), and the standard deviation (*red*); lower panel: range (*green*), median (*mint green*), and the interquartile range (*magenta*)

standard deviation is calculated from all observed prices and involves the squares of deviations which causes it to be particularly sensitive to extreme values (outliers).

The investigation was repeated after some time. The results are presented in Fig. 2.29.

Arithmetic average:	6165.257
Median:	5006.5
Range	12615
Interquartile range	2112
Standard deviation	2949.496

Now, there are a number of expensive vehicles whose prices are substantially different from the lower priced cars. Thus the price are skewed to the right. For skewed distributions, the standard deviation is typically higher than the interquartile range. This feature is demonstrated in the above example.

Interactive: Dotplot with Scale Parameters

The interactive example includes a number of sidebar panels. You can access the panels by setting a mark at the corresponding check box on the upper right.

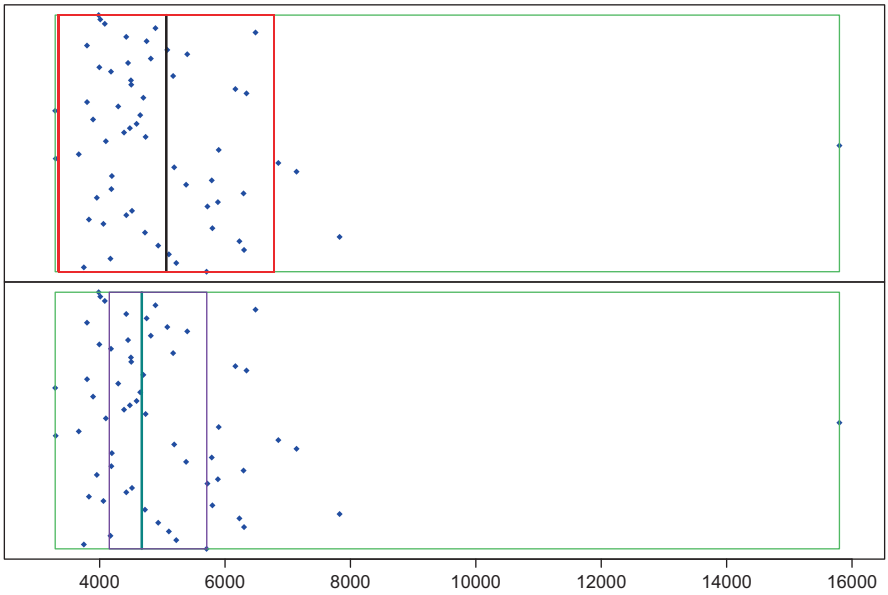


Fig. 2.28 Corrected prices of 74 cars in USD—upper panel: range (*green*), arithmetic average (*black*), and the standard deviation (*red*); lower panel: range (*green*), median (*mint green*), and the interquartile range (*magenta*)

Please select

- a dotplot type, e.g., jitter
- if you like the mean, median, range, or interquartile range to be included in the plot

The last two panels allow you to choose a dataset or variable and to change the font size. For a detailed explanation on datasets and variables, please refer to Appendix A.

Output

The interactive example in Fig. 2.30 allows us to display a one-dimensional frequency distribution in the form of a dotplot for a variety of variables. Possible values are displayed along the horizontal axis. For easier visualization, the observations may be randomly shifted (jitter) in the vertical direction. Furthermore, the median, the arithmetic mean, range, and interquartile range can be included.

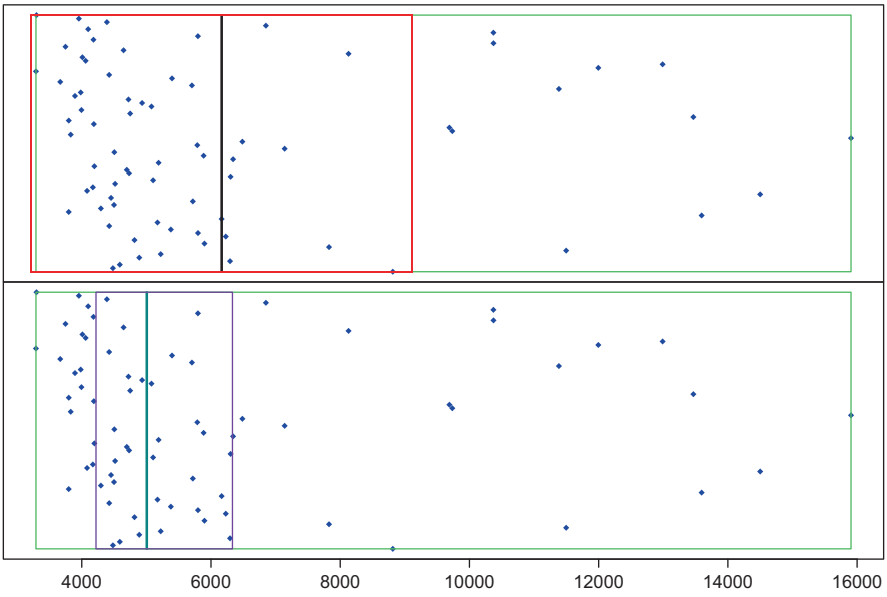


Fig. 2.29 Repeated investigation of prices of 74 cars in USD—upper panel: range (*green*), arithmetic average (*black*), and the standard deviation (*red*); lower panel: range (*green*), median (*mint green*), and the interquartile range (*magenta*)

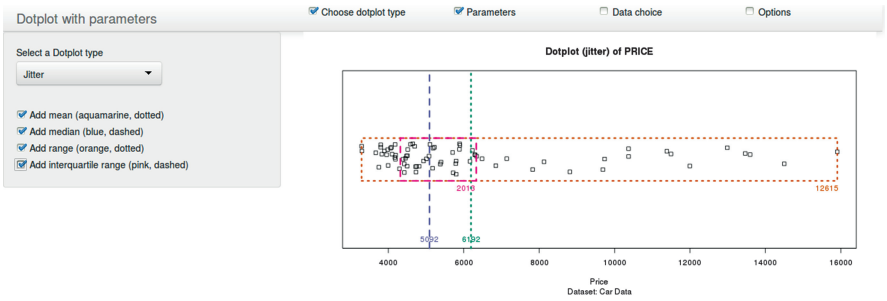


Fig. 2.30 Screenshot of the interactive example, available at http://u.hu-berlin.de/men_dot2

2.7 Graphical Display of the Location and Scale Parameters

Boxplot (Box-Whisker-Plot)

Unlike the stem-and-leaf diagram, the boxplot does not contain information about all observed values. It displays only the most important information about the frequency distribution. Specifically, the boxplot contains the smallest and the largest

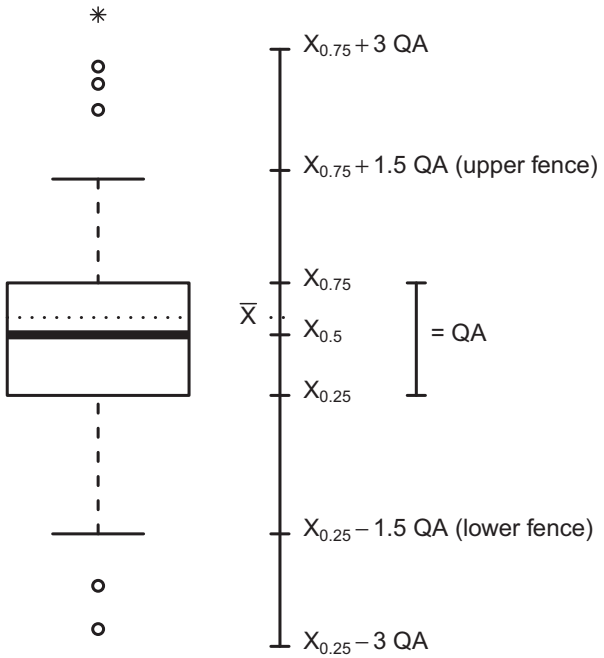
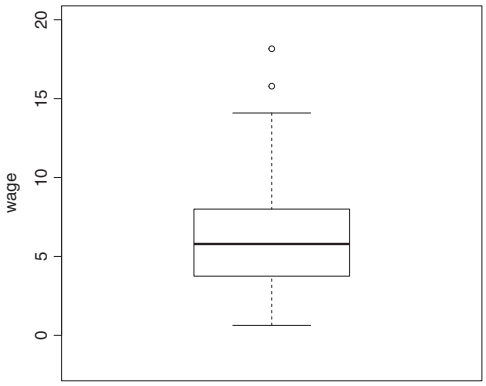


Fig. 2.31 The structure of a boxplot

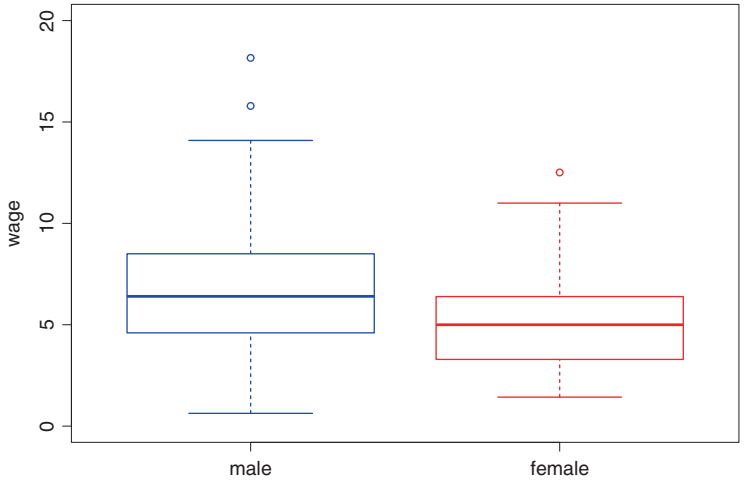
observed values $x_{(1)}$ and $x_{(n)}$ and three quartiles $x_{0.25}, x_{0.5}ax_{0.75}$. The second quartile $x_{0.5}$ is of course the median (Fig. 2.31).

The quartiles are denoted by a line and the first and third quartile are connected so that we obtain a box. The line inside this box denotes the median. The height of this box is the interquartile range which is the difference between the third and the first quartile: $x_{0.75}$ and $x_{0.25}$. Inside this box, one finds the central 50 % of all observed values.

The whiskers show the smallest and largest values within a 1.5 multiple of the interquartile range calculated from the boundary of the box. The bounds $x_{0.25} - 1.5 \cdot QA$ and $x_{0.75} + 1.5 \cdot QA$ are called the lower and upper fence, respectively. The values lying outside the fences are marked as outliers with a different symbol. Usually, the boxplot also displays the mean as a dashed line. The boxplot provides quick insight into the location, scale, shape, and structure of the data.



Example—boxplot of student salaries in USD



Example—boxplot of student salaries in USD; males and females separated

Explained: Boxplot of Car Prices

The prices of 74 types of cars were obtained in 1983. The results are displayed in Fig. 2.32.

The upper panels of the graphs contain dotplots. The lower panels show boxplots. The values lying outside a 1.5 multiple (resp. 3 multiple) of the interquartile range are denoted as extreme (outlying) observations. These outlying observations produce a large difference between the median (solid line) and the mean (dashed line).

Table 2.19
Example—Student salaries in USD

Total	Men	Women
$x_{\min} = 1$	$x_{\min} = 1$	$x_{\min} = 1.74997$
$x_{\max} = 44.5005$	$x_{\max} = 26.2903$	$x_{\max} = 44.5005$
$R = 43.5005$	$R = 25.2903$	$R = 42.7505$
$x_{0.25} = 5.24985$	$x_{0.25} = 6.00024$	$x_{0.25} = 4.74979$
$x_{0.5} = 7.77801$	$x_{0.5} = 8.92985$	$x_{0.5} = 6.79985$
$x_{0.75} = 11.2504$	$x_{0.75} = 12.9994$	$x_{0.75} = 10.0001$
$QA = 6.00065$	$QA = 9.99916$	$QA = 5.25031$
$\bar{x} = 9.02395$	$\bar{x} = 9.99479$	$\bar{x} = 7.87874$
$s^2 = 26.408$	$s^2 = 27.9377$	$s^2 = 22.2774$
$s = 5.13887$	$s = 5.28562$	$s = 4.7199$
$v = 0.57$	$v = 0.53$	$v = 0.60$

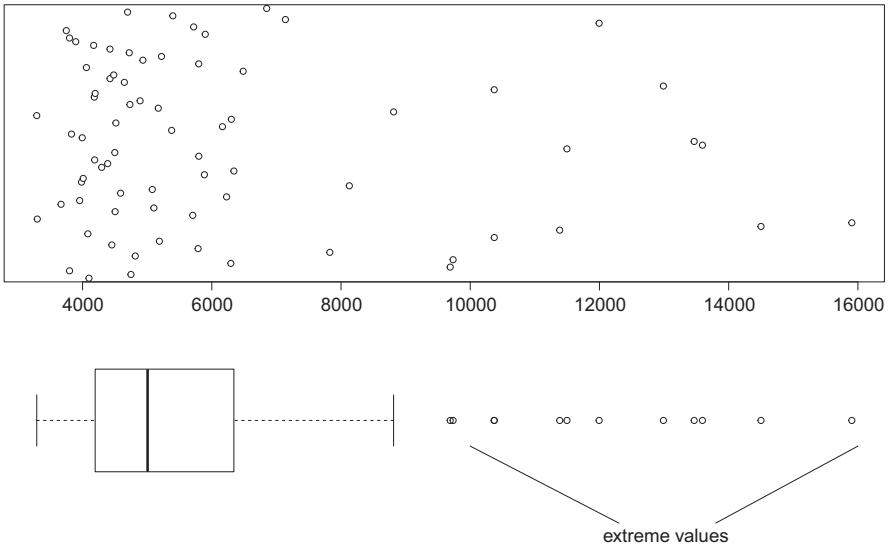


Fig. 2.32 Boxplot of prices of 74 cars

Interactive: Visualization of One-Dimensional Distributions

The interactive example includes a number of sidebar panels. You can access the panels by setting a mark at the corresponding check box on the upper right.

Please choose

- a dotplot type, e.g., jitter
- the number of bins for the histogram
- if you like the mean and median to be included in the plots

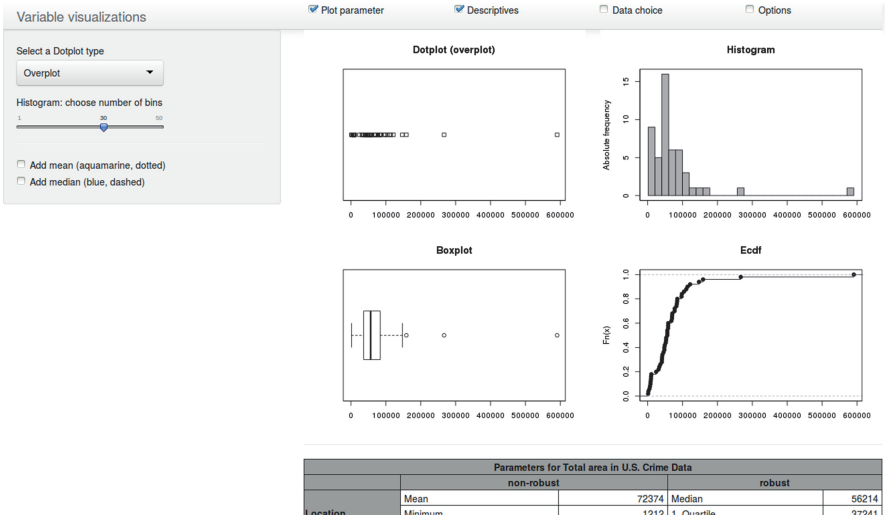


Fig. 2.33 Screenshot of the interactive example, available at http://u.hu-berlin.de/men_vis

The last two panels allow you to choose a dataset or variable and to change the font size. For a detailed explanation on datasets and variables, please refer to Appendix A.

Output

The interactive example in Fig. 2.33 allows us to display a one-dimensional frequency distribution in the form of a dotplot, a histogram, a boxplot, and cumulative distribution function for a variety of variables. Possible values are displayed along the horizontal axis. For easier visualization, the observations may be randomly shifted (jitter) in the vertical direction. Furthermore, the median and the arithmetic mean can be included. You also receive a table showing the numerical values of certain parameters.

http://www.springer.com/978-3-319-17703-8

Introduction to Statistics
Using Interactive MM*Stat Elements
Härdle, W.K.; Klinke, S.; Rönz, B.
2015, XX, 516 p. 205 illus., 173 illus. in color.,
Hardcover
ISBN: 978-3-319-17703-8