Univariate Data Analysis

3

3.1 First Steps in Data Analysis

Let us return to our students from the previous chapter. After completing their survey of bread spreads, they have now coded the data from the 850 respondents and entered them into a computer. In the first step of data assessment, they investigate each variable – for example, average respondent age – separately. This is called *univariate analysis* (Fig. 3.1). By contrast, when researchers analyze the relationship between two variables – for example, between gender and choice of spread – this is called *bivariate analysis* (see Sect. 4). With relationships between more than two variables, one speaks of *multivariate analysis* (see Sect. 5.3).

How can the results of 850 responses be "distilled" to create a realistic and accurate impression of the surveyed attributes and their relationships? Here the importance of statistics becomes apparent. Recall the professor who was asked about the results of the last final exam. The students expect distilled information, e.g. "*the average score was 75 %*" or "*the failure rate was 29.4 %*". Based on this information, students believe they can accurately assess general performance: "*an average score of 75 % is worse than the 82 % average on the last final exam*". A single distilled piece of data – in this case, the average score – appears sufficient to sum up the performance of the entire class.[1]

This chapter and the next will describe methods of distilling data and their attendant problems. The above survey will be used throughout as an example.

Chapter 3 Translated from the German original, Cleff, T. (2011). 3 Vom Datensatz zur Information. In *Deskriptive Statistik und moderne Datenanalyse* (pp. 31–77) © Gabler Verlag, Springer Fachmedien Wiesbaden GmbH, 2011.

[1] It should be noted here that the student assessment assumes a certain kind of distribution. An average score of 75 % is obtained whether all students receive a score of 75 %, or whether half score 50 % and the other half score 100 %. Although the average is the same, the qualitative difference in these two results is obvious. Average alone, therefore, does not suffice to describe the results.

DOI 10.1007/978-3-319-01517-0_3,

	index	gender	age	Bodyweight	spread	offer
1	1	male	31	63.1	butter	very poor
2	2	male	73	77.5	butter	very poor
3	5	male	45	82.1	butter	very poor
4	6	male	57	61.7	butter	very poor
5	9	male	38	36.5	butter	very poor
6	11	male	27	64.0	butter	very poor
7	12	male	36	70.9	butter	very poor
8	13	male	60	70.4	butter	very poor
9	15	male	21		butter	very poor
10	16	male	26			
11	18	male	55			
12	22	male	27		butter	very poor
13	25	male	30	72.7	butter	very poor
14	26	male	33	77.8	butter	very poor
15	27	male	33	90.8	butter	very poor
16	28	male	58	62.4	butter	very poor
17	29	male	23	91.2	butter	very poor

Analysis of only one variable: Univariate Analysis

Note: Using SPSS or Stata: The data editor can usually be set to display the codes or labels for the variables, though the numerical values are stored

Fig. 3.1 Survey data entered in the data editor

Graphical representations or frequency tables can be used to create an overview of the univariate distribution of nominal- and ordinal-scaled variables. In the *frequency table* in Fig. 3.2, each variable trait receives its own line, and each line intersects the columns *absolute frequency*, *relative frequency [in %]*,[2] *valid percentage values*, and *cumulative percentage*. The relative frequency of trait x_i is abbreviated algebraically by $f(x_i)$. Any missing values are indicated in a separate line with a percentage value. Missing values are not included in the calculations of *valid percentage values*[3] and *cumulative percentage*. The cumulative percentage reflects the sum of all rows up to and including the row in question. The figure of 88.1 % given for the rating *average* in Fig. 3.2 indicates that 88.1 % of the respondents described the selection as average or worse. Algebraically, the cumulative frequencies are expressed as a *distribution function*, abbreviated F(x), and calculated as follows:

$$F(x_p) = f(x_1) + f(x_2) + \cdots + f(x_p) = \sum_{i=1}^{p \le n} f(x_i) \tag{3.1}$$

These results can also be represented graphically as a *pie chart*, a *horizontal bar chart*, or a *vertical bar chart*. All three diagram forms can be used with nominal and ordinal variables, though pie charts are used mostly for nominal variables.

[2] Relative frequency ($f(x_i)$) equals the absolute frequency ($h(x_i)$) relative to all valid and invalid observations ($N = N_{valid} + N_{invalid}$): $f(x_i) = h(x_i)/N$.

[3] Valid percentage ($gf(x_i)$) equals the absolute frequency ($h(x_i)$) relative to all valid observations (N_{valid}): $g(x_i) = h(x_i)/N_{valid}$.

	Absolute frequency	Relative frequency [in %]	Valid percentage values	Cumulative percentage
Poor	391	46.0	46.0	46.0
Fair	266	31.3	31.3	77.3
Average	92	10.8	10.8	88.1
Good	62	7.3	7.3	95.4
Excellent	39	4.6	4.6	100.0
Total	850	100.0	100.0	

Fig. 3.2 Frequency table for selection ratings

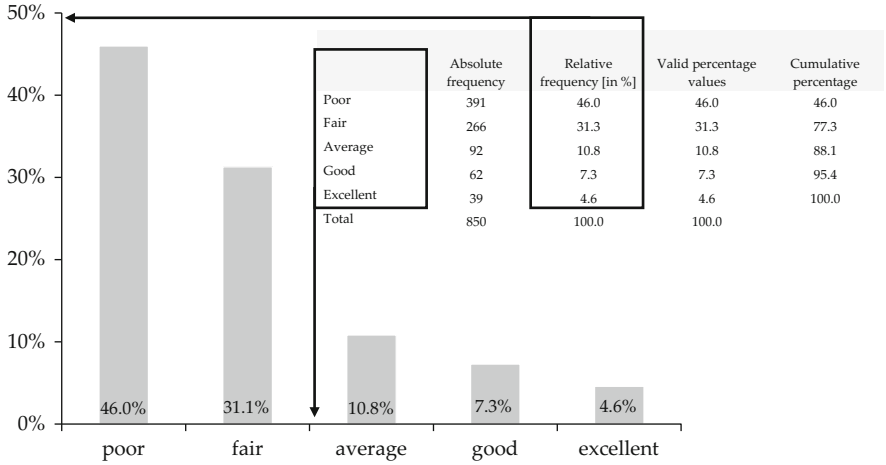


Fig. 3.3 Bar chart/Frequency distribution for the selection variable

The traits of the frequency table in the bar chart (poor, fair, average, good, excellent) are assigned to the x-axis and the relative or absolute frequency to the y-axis. The height of a bar equals the frequency of each x-value. If the relative frequencies are assigned to the y-axis, a graph of the frequency function is obtained (see Fig. 3.3).

In addition to the frequency table, we can also represent the distribution of an ordinally scaled variable (or higher) using the F(x) distribution function. This function leaves the traits of the x-variables in question on the x-axis, and assigns the cumulative percentages to the y-axis, generating a *step function*. The data representation is analogous to the column with cumulative percentages in the frequency table (Fig. 3.4).

In many publications, the scaling on the y-axis of a vertical bar chart begins not with 0 but with some arbitrary value. As Fig. 3.5 shows, this can lead to a misunderstanding at first glance. Both graphs represent the same content – the relative frequency of male and female respondents (49 % and 51 %, respectively).

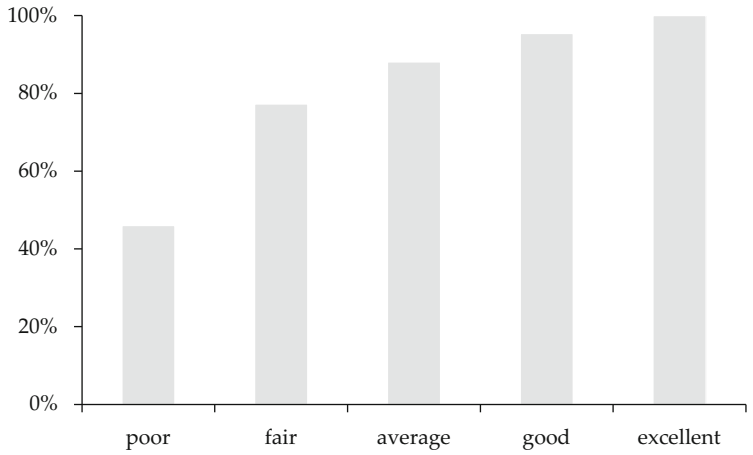
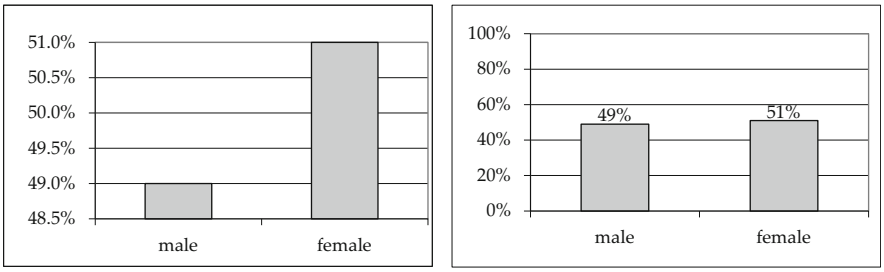


Fig. 3.4 Distribution function for the selection variable



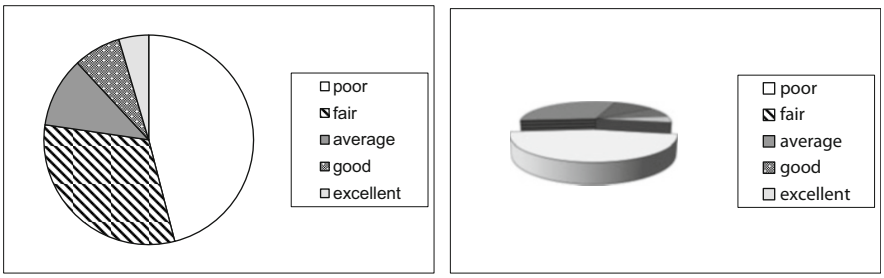
Part 1 Part 2

Fig. 3.5 Different representations of the same data (1). . .

But because the y-axis is cut off in the first graph, the relative frequency of the genders appears to change. The first graph appears to show a relationship of five females to one male, suggesting that there are five times as many female observations as male observations in the sample. The interval in the first graph is misleading – a problem we'll return to below – so that the difference of 2 % points seems larger than it actually is. For this reason, the second graph in Fig. 3.5 is the preferable form of representation.

Similar distortions can arise when two alternate forms of a pie chart are used. In the first chart in Fig. 3.6, the size of each wedge represents relative frequency. The chart is drawn by weighting the circle segment angles such that each angle $\alpha_i = f(x_i) \cdot 360°$.

Since most viewers read pie charts clockwise from the top, the traits to be emphasized should be placed in the 12 o'clock position whenever possible. Moreover, the chart shouldn't contain too many segments – otherwise the graph will be hard to read. They should also be ordered by some system – for example, by size or content.



Part 1 Part 2

Fig. 3.6 Different representations of the same data (2). . .

The second graph in Fig. 3.6, which is known as a "perspective" or "3D" pie chart, looks more modern, but the downside is that the area of each wedge no longer reflects relative frequency. The representation is thus somewhat misleading. The pie chart segments in the foreground seem larger. The edge of the pie segments in the front can be seen, but not those in the back. The "lifting up" of a particular wedge can amplify this effect even more.

And what of cardinal variables? How should they be represented? The novice might attempt to represent bodyweight using a vertical bar diagram – as shown in graph 1 of Fig. 3.7. But the variety of possible traits generates too many bars, and their heights rarely vary. Frequently, a trait appears only once in a collection of cardinal variables. In such cases, the goal of presenting all the basic relationships at a glance is destined to fail. For this reason, the individual values of cardinal variables should be grouped in classes, or classed. Bodyweight, for instance, could be assigned to the classes shown in Fig. 3.7.[4]

By standard convention, the upper limit value in a class belongs to that class; the lower limit value does not. Accordingly, persons who are 60 kg belong to the 50–60 kg group, while those who are 50 kg belong to the class below. Of course, it is up to the persons assessing the data to determine class size and class membership at the boundaries. When working with data, however, one should clearly indicate the decisions made in this regard.

A *histogram* is a classed representation of cardinal variables. What distinguishes the histogram from other graphic representations is that it expresses relative class frequency not by height but by area (height $\times$ width). The height of the bars represents frequency density. The denser the bars are in the bar chart in part 1 of Fig. 3.7, the more observations there are for that given class and the greater its frequency density. As the frequency density for a class increases, so too does its area (height $\times$ width). The histogram obeys the principle that the intervals in a diagram should be selected so that the data are not distorted. In the histogram, the share of area for a specific class relative to the entire area of all classes equals the relative frequency of the specific class. To understand why the selection of

[4] For each ith class, the following applies: $x_i < X \leq x_{i+1}$ with i $\in \{1, 2, \ldots, k\}$.

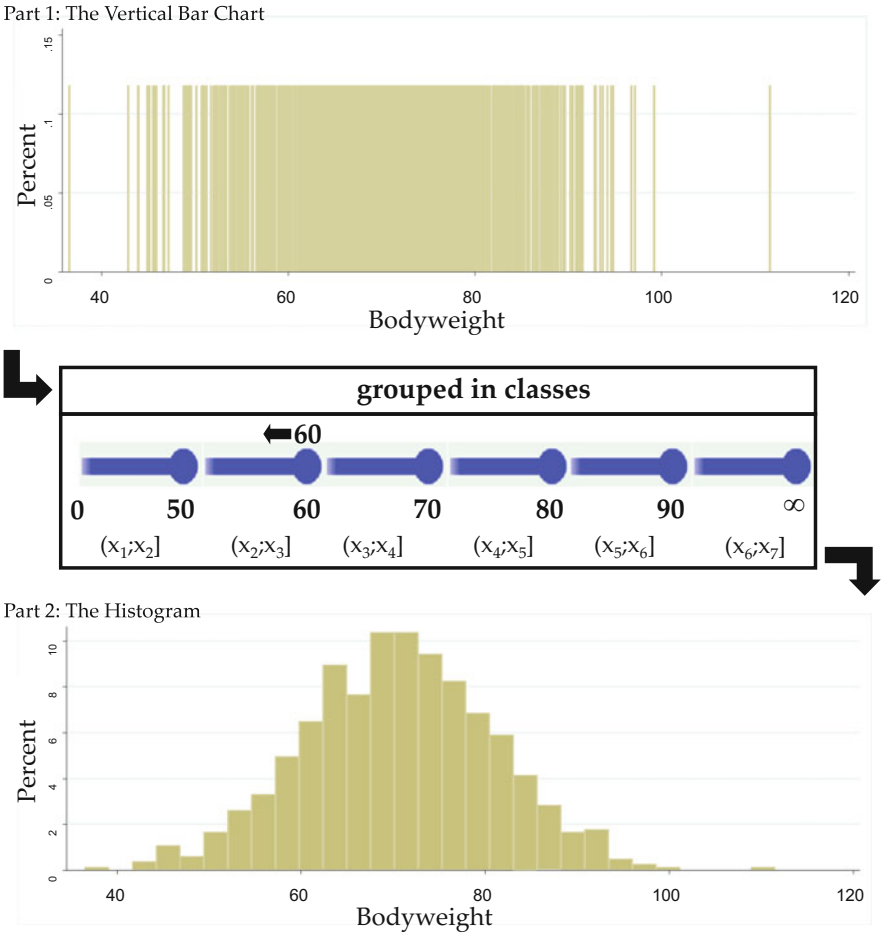


Fig. 3.7 Using a histogram to classify data

suitable intervals is so important consider part 1 of Fig. 3.8, which represents the same information as Fig. 3.7 but uses unequal class widths. In a vertical bar chart, height represents relative frequency. The white bars in the figure represent relative frequency. The graph appears to indicate that a bodyweight between 60 and 70 kg is the most frequent class. Above this range, frequency drops off before rising again slightly for the 80–90 kg class. This impression is created by the distribution of the 70–80 kg group into two classes, each with a width of 5 kg, or half that of the others. If the data are displayed without misleading intervals, the frequency densities can be derived from the grey bars. With the same number of observations in a class, the bars would only be the same height if the classes were equally wide. By contrast, with a class half as large and the same number of observations, the observations will be twice as dense. Here we see that, in terms of class width, the density for the 70–75 kg range is the largest.

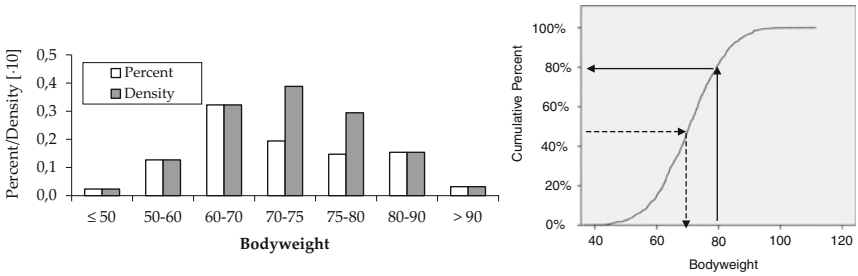


Fig. 3.8 Distorting interval selection with a distribution function

It would be useful if the histogram's differences in class width were indicated to scale by different widths on the x-axis. Unfortunately, no currently available statistics or graphical software can perform this function. Instead, they avoid the problem by permitting equal class widths only.

The distribution function of a cardinal variable can be represented as unclassed. Here too, the frequencies are cumulative as one moves along the x-axis. The values of the distribution function rise evenly and remain between 0 and 1. The distribution function for the bodyweight variable is represented in part 2 of Fig. 3.8. Here, one can obtain the cumulated percentages for a given bodyweight and vice versa. Some 80 % of the respondents are 80 kg or under, and 50 % of the respondents are 70 kg or under.

3.2 Measures of Central Tendency

The previous approach allowed us to reduce the diversity of information from the questionnaires – in our sample there were 850 responses – by creating graphs and tables with just a few lines, bars, or pie wedges. But how and under which conditions can this information be reduced to a single number or measurement that summarizes the distinguishing features of the dataset and permits comparisons with others? Consider again the student who, to estimate the average score on the last final exam, looks for a single number – the average grade or failure rate. The average score for two final exams is shown in Fig. 3.9.[5]

Both final exams have an identical distribution; in the second graph (part 2), this distribution is shifted one grade to the right on the x-axis. This shift represents a mean value one grade higher than the first exam. Mean values or similar parameters that express a general trend of a distribution are called *measures of central tendency*. Choosing the most appropriate measure usually depends on context and the level of measurement.

[5] The grade scale is taken here to be cardinal scaled. This assumes that the difference in scores between A and B is identical to the difference between B and C, etc. But because this is unlikely in practice, school grades, strictly speaking, must be seen as ordinal scaled.

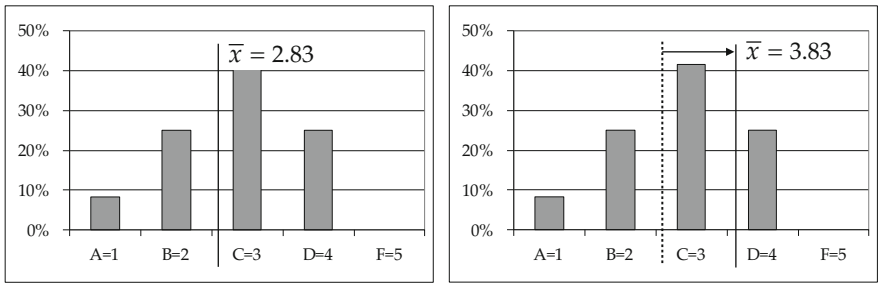


Fig. 3.9 Grade averages for two final exams

3.2.1 Mode or Modal Value

The most basic measure of central tendency is known as the *mode* or *modal value*. The mode identifies the value that appears most frequently in a distribution. In part 1 of Fig. 3.9 the mode is the grade C. The mode is the "champion" of the distribution. Another example is the item selected most frequently from five competing products. This measure is particularly important with voting, though its value need not be clear. When votes are tied, there can be more than one modal value. Most software programmes designate only the smallest trait. When values are far apart this can lead to misinterpretation. For instance, when a cardinal variable for age and the traits 18 and 80 appear in equal quantities and more than all the others, many software packages still indicate the mode as 18.

3.2.2 Mean

The *arithmetic mean* – colloquially referred to as the *average* – is calculated differently depending on the nature of the data. In empirical research, data most frequently appears in a raw data table that includes all the individual trait values. For raw data tables, the mean is derived from the formula:

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \ldots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{3.2}$$

All values of a variable are added and divided by n. For instance, given the values 12, 13, 14, 16, 17, and 18 the mean is $\bar{x} = \frac{1}{6}(12 + 13 + 14 + 16 + 17 + 18) = 15$.

The mean can be represented as a balance scale (see Fig. 3.10), and the deviations from the mean can be regarded as weights. If, for example, there is a deviation of (−3) units from the mean, then a weight of 3 g is placed on the left side of the balance scale. The further a value is away from the mean, the heavier the weight. All negative deviations from the mean are placed on the left side of

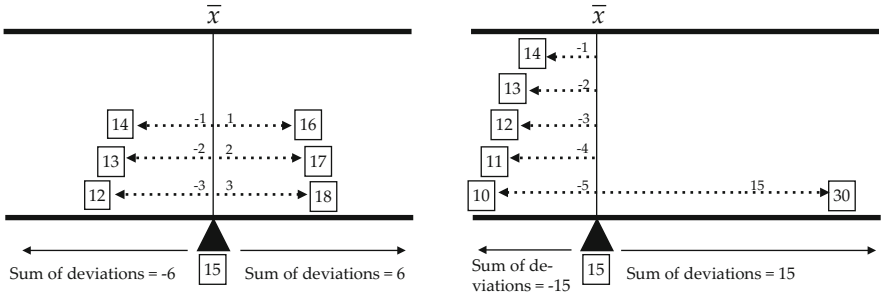


Fig. 3.10 Mean expressed as a balanced scale

the mean, and all positive deviations on the right. The scale is exactly balanced. With an arithmetic mean, the sum of negative deviations equals the sum of positive deviations:

$$\sum_{i=1}^{n} (x_i - \bar{x}) = 0 \tag{3.3}$$

In real life, if a heavy weight is on one side of the scale and many smaller weights are on the other, the scale can still be balanced (cf. Fig. 3.10). But the mean is not a good estimate for this kind of distribution: it could over- or underestimate the many smaller weights. We encountered this problem in Sect. 2.5; in such cases, an outlier value is usually responsible for distorting the results. Assume you want to calculate the average age of animals in a zoo terrarium containing five snakes, nine spiders, five crocodiles, and one turtle. The last animal – the turtle – is 120 years old, while all the others are no older than four (Fig. 3.11).

Based on these ages, the mean would be 7.85 years. To "balance" the scale, the ripe old turtle would have to be alone on the right side, while all the other animals are on the left side. We find that the mean value is a poor measure to describe the average age in this case because only one other animal is older than three. To reduce or eliminate the outlier effect, practitioners frequently resort to a *trimmed mean*. This technique "trims" the smallest and largest 5 % of values before calculating the mean, thus partly eliminating outliers. In our example, the 5 % trim covers both the youngest and oldest observation (the terrarium has 20 animals), thereby eliminating the turtle's age from the calculation. This results in an average age of 2 years, a more realistic description of the age distribution. We should remember, however, that this technique eliminates 10 % of the observations, and this can cause problems, especially with small samples.

Let us return to the "normal" mean, which can be calculated from a frequency table (such as an overview of grades) using the following formula:

$$\bar{x} = \frac{1}{n} \sum_{v=1}^{k} x_v \cdot n_v = \sum_{v=1}^{k} x_v \cdot f_v \tag{3.4}$$

		Age					
		1	2	3	4	120	Total
Animal	Snake	2	1	1	1	0	5
	Turtle	0	0	0	0	1	1
	Crocodile	1	2	2	0	0	5
	Spider	4	4	1	0	0	9
Total		7	7	4	1	1	20

Note: Mean = 7.85 years; 5 % trimmed mean = 2 years

Fig. 3.11 Mean or trimmed mean using the zoo example

We will use the frequency table in Fig. 3.2 as an example. Here the index v runs through the different traits of the observed ordinal variables for selection (*poor*, *fair*, *average*, *good*, *excellent*). The value n_v equals the absolute number of observations for a trait. The trait *good* yields a value of $n_v = n_4 = 62$. The variable x_v assumes the trait value of the index v. The trait *poor* assumes the value $x_1 = 1$, the trait *fair* the value $x_2 = 2$, etc. The mean can be calculated as follows:

$$\bar{x} = \frac{1}{850} \cdot (391 \cdot 1 + 266 \cdot 2 + 92 \cdot 3 + 62 \cdot 4 + 39 \cdot 5) = 1.93 \tag{3.5}$$

The respondents gave an average rating of 1.93, which approximately corresponds to *fair*. The mean could also have been calculated using the relative frequencies of the traits f_v:

$$\bar{x} = (0.46 \cdot 1 + 0.313 \cdot 2 + 0.108 \cdot 3 + 0.073 \cdot 4 + 0.046 \cdot 5) = 1.93 \tag{3.6}$$

Finally, the mean can also be calculated from traditional classed data according to this formula:

$$\bar{x} = \frac{1}{n} \sum_{v=1}^{k} n_v m_v = \sum_{v=1}^{k} f_v m_v, \tag{3.7}$$

where m_v is the mean of class number v.

Students often confuse this with the calculation from frequency tables, as even the latter contain classes of traits. With classed data, the mean is calculated from cardinal variables that are summarized into classes by making certain assumptions. In principle the mean can be calculated this way from a histogram. Consider again Fig. 3.7. The calculation of the mean bodyweight in part 1 agrees with the calculation from the raw data table. But what about when there is no raw data table, only the information in the histogram, as in part 2 of Fig. 3.7? Figure 3.12 shows a somewhat more simplified representation of a histogram with only six classes.

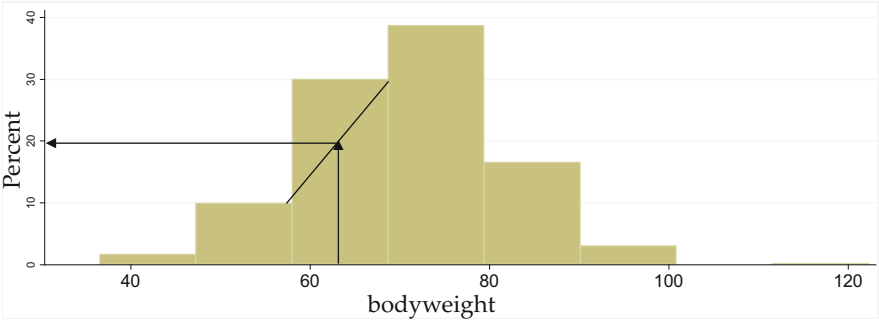


Fig. 3.12 Calculating the mean from classed data

Table 3.1 Example of mean calculation from classed data

Water use [in l]	0–200	200–400	400–600	600–1,000
Rel. frequency	0.2	0.5	0.2	0.1

Source: Schwarze (2008, p. 16), translated from the German

We start from the implicit assumption that all observations are distributed evenly within a class. Accordingly, cumulated frequency increases linearly from the lower limit to the upper limit of the class. Here class frequency average necessarily equals the mean. To identify the total mean, add all products from the class midpoint and the attendant relative frequencies.

Here is another example to illustrate the calculation. Consider the following information on water use by private households (Table 3.1):

The water-use average can be calculated as follows:

$$\bar{x} = \sum_{v=1}^{k} f_v m_v = \sum_{v=1}^{4} f_v m_v = 0.2 \cdot 100 + 0.5 \cdot 300 + 0.2 \cdot 500 + 0.1 \cdot 800 = 350 \quad (3.8)$$

With all formulas calculating the mean, we assume equidistant intervals between the traits. This is why the mean cannot be determined for nominal variables. This is also why, strictly speaking, no mean can be calculated for ordinal variables. But this is only true if one takes a dogmatic position. Practically minded researchers who possess sufficiently large samples (approx. $n > 99$) often calculate the mean by assuming equidistance.

The informational value of the mean was previously demystified in Sect. 3.2 using the example of average test grades. An average grade of C occurs when all students receive C. The same average results when half of the students receive an A and the other half an F. The same kind of problem could result by selecting travel destinations based on temperature averages. Beijing, Quito, and Milan all have an average temperature of 12 °C, but the experience of temperature in the three cities varies greatly. The winter in Beijing is colder than in Stockholm and the summer is hotter than in Rio de Janeiro. In Milan the temperatures are Mediterranean, fluctuating seasonally, while the altitude in Quito ensures that the temperature stays pretty much the same the whole year over (Swoboda 1971, p. 36).

The average is not always an information-rich number that uncovers all that remains hidden in tables and figures. When no information can be provided on distribution (e.g. average deviation from average) or when weightings and reference values are withheld, the average can also be misleading. The list of amusing examples is long, as described by Krämer (2005, p. 61). Here are a few:

- Means rarely result in whole numbers. For instance, what do we mean by the decimal place when we talk of 1.7 children per family or 3.5 sexual partners per person?
- When calculating the arithmetic mean, all values are treated equally. Imagine a proprietor of an eatery in the Wild West who, when asked about the ingredients of his stew, says: *Half and half. One horse and one jackrabbit*. It is not always accurate to consider the values as equal in weight. The cook might advertise his concoction as a *wild game stew*, but if the true weights of the inputs were taken into account, it would be more accurately described as horse goulash. Consider an example from the economy: if the average female salary is 20 MUs (monetary units) and the average male salary is 30 MUs, the average employee salary is not necessary 25 MUs. If males constitute 70 % of the workforce, the average salary will be: $0.7 \cdot 30\ \text{MU} + 0.3 \cdot 20\ \text{MU} = 27\ \text{MU}$. One speaks here of a weighted arithmetic mean or a scaled arithmetic mean. The Federal Statistical Office of Germany calculates the rate of price increase for products in a basket of commodities in a similar fashion. The price of a banana does not receive the same weight as the price of a vehicle; its weight is calculated based on its average share in a household's consumption.
- The choice of *reference base* – i.e. the dominator for calculating the average – can also affect the interpretation of data. Take the example of traffic deaths. Measured by deaths per passenger-kilometres travelled, trains have a rate of nine traffic deaths per 10 billion kilometres travelled and planes three deaths per ten billion kilometres travelled. Airlines like to cite these averages in their ads. But if we consider traffic deaths not in relation to distance but in relation to time of travel, we find completely different risks. For trains there are seven fatalities per 100 million passenger-hours and for planes there are 24 traffic deaths per 100 million passenger-hours. Both reference bases can be asserted as valid. The job of empirical researchers is to explain their choice. Although I have a fear of flying, I agree with Krämer (2005, p. 70) when he argues that passenger-hours is a better reference base. Consider the following: Few of us are scared of going to bed at night, yet the likelihood of dying in bed is nearly 99 %. Of course, this likelihood seems less threatening when measured against the time we spend in bed.

3.2.3 Geometric Mean

The above problems frequently result from a failure to apply weightings or by selecting a wrong or poor reference base. But sometimes the arithmetic mean as a measure of general tendency can lead to faulty results even when the weighting and reference base are appropriate. This is especially true in economics when measuring

Year	Sales [mio.]	Rate of change [in %]	Changes in sales when using arithm. mean	Changes in sales when using geom. mean
2002	€20,000.00		€20,000.00	€20,000.00
2003	€22,000.00	1.000%	€20,250.00	€20,170.56
2004	€20,900.00	-5.000%	€20,503.13	€20,342.57
2005	€18,810.00	-10.000%	€20,759.41	€20,516.04
2006	€20,691.00	10.000%	€21,018.91	€20,691.00
Arithmetic mean		1.250%		
Geometric mean		0.853%		

Fig. 3.13 An example of geometric mean

rates of change or growth. These rates are based on data observed over time, which is why such data are referred to as time series. Figure 3.13 shows an example of sales and their rates of change over 5 years.

Using the arithmetic mean to calculate the average rate of change yields a value of 1.25 %. This would mean that yearly sales have increased by 1.25 %. Based on this growth rate, the €20,000 in sales in 2002 should have increased to €21,018.91 by 2006, but actual sales in 2006 were €20,691.00. Here we see how calculating average rates of change using arithmetic mean can lead to errors. This is why the *geometric mean for* rates of change is used. In this case, the parameter links initial sales in 2002 with the subsequent rates of growth each year until 2006. The result is:

$$\begin{aligned} U_6 &= U_5 \cdot (1 + 0.1) = (U_4 \cdot (1 - 0.1)) \cdot (1 + 0.1) = \ldots \\ &= (U_2 \cdot (1 + 0.1)) \cdot (1 - 0.05) \cdot (1 - 0.1) \cdot (1 + 0.1). \end{aligned} \tag{3.9}$$

To calculate the average change in sales from this chain, the four rates of change (1 + 0.1)·(1–0.05)·(1–0.1)·(1 + 0.1) must yield the same value as the fourfold application of the average rate of change:

$$\left(1 + \overline{p}_{geom}\right) \cdot \left(1 + \overline{p}_{geom}\right) \cdot \left(1 + \overline{p}_{geom}\right) \cdot \left(1 + \overline{p}_{geom}\right) = \left(1 + \overline{p}_{geom}\right)^4 \tag{3.10}$$

For the geometric mean, the yearly rate of change is thus:

$$\overline{p}_{geom} = \sqrt[4]{(1 + 0.1)\ (1 - 0.05)\ (1 - 0.1)\ (1 + 0.1)} - 1 = 0.853 \tag{3.11}$$

The last column in Fig. 3.13 shows that this value correctly describes the sales growth between 2002 and 2006. Generally, the following formula applies for identifying *average rates of change*:

$$\overline{p}_{geom} = \sqrt[n]{(1 + p_1) \cdot (1 + p_2) \cdot \cdot (1 + p_n)} - 1 = \sqrt[n]{\prod_{i=1}^{n} (1 + p_i)} - 1 \tag{3.12}$$

The geometric mean for rates of change is a special instance of the *geometric mean*, and is defined as follows:

$$\bar{x}_{geom} = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n} = \sqrt[n]{\prod_{i=1}^{n} x_i} \tag{3.13}$$

The geometric mean equals the arithmetic mean of the logarithms[6] and is only defined for positive values. For observations of different sizes, the geometric mean is always smaller than the arithmetic mean.

3.2.4 Harmonic Mean

A measure seldom required in economics is the so-called harmonic mean. Because of the rarity of this measure, researchers tend to forget it, and instead use the arithmetic mean. However, sometimes the arithmetic mean produces false results. The harmonic mean is the appropriate method for averaging ratios consisting of numerators and denominators (unemployment rates, sales productivity, kilometres per hour, price per litre, people per square metre, etc.) when the values in the numerator are not identical. Consider, for instance, the sales productivity (as measured in revenue per employee) of three companies with differing headcounts but identical revenues. The data are given in Table 3.2.

To compare the companies, we should first examine the sales productivity of each firm regardless of its size. Every company can be taken into account with a simple weighted calculation. We find average sales per employee as follows:

$$\bar{x} = \frac{1}{3}\left(\frac{S_1}{E_1} + \frac{S_2}{E_2} + \frac{S_3}{E_3}\right) = €433.33 \tag{3.14}$$

If this value were equally applicable to all employees, the firms – which have 16 employees together – would have sales totalling 16·€433.33 ≈ €6,933, but the above table shows that actual total sales are only €3,000. When calculating company sales, it must be taken into account that the firms employ varying numbers of employees and that the employees contribute in different ways to total productivity. This becomes clear from the fact that companies with equal sales (identical numerators) have different headcounts and hence different values in the denominator. To identify the contribution made by each employee to sales, one must weight the individual observations ($i = 1,\ldots, 3$) of sales productivity (SP_i) with the number of employees (n_i), add them and then divide by the total number of employees. The result is an arithmetic mean weighted by the number of employees:

[6] If all values are available in logarithmic form, the following applies to the arithmetic mean: $\frac{1}{n}\left(\ln(x_1) + \ldots + \ln(x_n)\right) = \frac{1}{n}\ln(x_1 \cdot \ldots \cdot x_n) = \ln\left(x_1 \cdot \ldots \cdot x_n\right)^{\frac{1}{n}} = \sqrt[n]{\prod_{i=1}^{n} x_i} = \bar{x}_{geom.}$

Table 3.2 Harmonic mean

	Sales	Employees	Sales per employee (SP)	Formula in Excel
	€1,000	10	€100.00	
	€1,000	5	€200.00	
	€1,000	1	€1,000.00	
Sum	€3,000	16	€1,300.00	SUM(D3:D5)
Arithmetic mean			€433.33	AVERAGE(D3:D5)
Harmonic mean			€187.50	HARMEAN(D3:D5)

$$\frac{n_1 \cdot SP_1 + n_2 \cdot SP_2 + n_3 \cdot SP_3}{n} = \frac{10 \cdot €100}{16} + \frac{5 \cdot €200}{16} + \frac{1 \cdot €1{,}000}{16} \approx €187.50 \quad (3.15)$$

Using this formula, the 16 employees generate the real total sales figure of €3,000. If the weighting for the denominator (i.e. the number of employees) is unknown, the value for k = 3 sales productivity must be calculated using *an unweighted harmonic mean*:

$$\bar{x}_{harm} = \frac{k}{\sum_{i=1}^{k} \frac{1}{x_i}} = \frac{k}{\sum_{i=1}^{k} \frac{1}{SP_i}} = \frac{3}{\frac{1}{€100} + \frac{1}{€200} + \frac{1}{€1{,}000}} = \frac{€187.50}{\text{Employee}} \quad (3.16)$$

Let's look at another example that illustrates the harmonic mean. A student must walk 3 km to his university campus by foot. Due to the nature of the route, he can walk the first kilometre at 2 km/h, the second kilometre at 3 km/h, and the last kilometre at 4 km/h. As in the last example, the arithmetic mean yields the wrong result:

$$\bar{x} = \frac{1}{3}\left(2\frac{km}{h} + 3\frac{km}{h} + 4\frac{km}{h}\right) = 3\frac{km}{h}, \text{or 1 hour} \quad (3.17)$$

But if we break down the route by kilometre, we get 30 min for the first kilometre, 20 min for the second kilometre, and 15 min for the last kilometre. The durations indicated in the denominator vary by route segment, resulting in a total of 65 min. The weighted average speed is thus 2.77 km/h.[7] This result can also be obtained using the harmonic mean formula and k = 3 for the route segments:

$$\bar{x}_{harm} = \frac{k}{\sum_{i=1}^{k} \frac{1}{x_i}} = \frac{3}{\frac{1}{2\,\frac{\text{km}}{\text{h}}} + \frac{1}{3\frac{\text{km}}{\text{h}}} + \frac{1}{4\frac{\text{km}}{\text{h}}}} = 2.77\frac{\text{km}}{\text{h}} \quad (3.18)$$

[7] (30 min · 2 km/h + 20 min · 3 km/h + 15 min · 4 km/h) /65 min = 2.77 km/h.

In our previous examples the values in the numerator were identical for every observation. In the first example, all three companies had sales of €1,000 and in the second example all route segments were 1 km. If the values are not identical, the *unweighted harmonic mean* must be calculated. For instance, if the k = 3 companies mentioned previously had sales of n_1 = €1,000, n_2 = €2,000, and n_3 = €5,000, we would use the following calculation:

$$\bar{x}_{harm} = \frac{n}{\sum_{i=1}^{k} \frac{n_i}{x_i}} = \frac{n}{\sum_{i=1}^{k} \frac{n_i}{SP_i}} = \frac{€1,000 + €2,000 + €5,000}{\frac{€1,000}{€100} + \frac{€2,000}{€200} + \frac{€5,000}{€1,000}} = \frac{€500}{\text{Employee}} \tag{3.19}$$

As we can see here, the unweighted harmonic mean is a special case of the weighted harmonic mean.

Fractions do not always necessitate the use of the harmonic mean. For example, if the calculation involving the route to the university campus included different times instead of different segments, the arithmetic mean should be used to calculate the average speed. If one student walked an hour long at 2 km/h, a second hour at 3 km/h, and the last hour at 4 km/h, the arithmetic mean yields the correct the average speed. Here the size of the denominator (time) is identical and yields the value of the numerator (i.e. the length of the partial route):

$$\bar{x} = \frac{1}{3}\left(2\frac{km}{h} + 3\frac{km}{h} + 4\frac{km}{h}\right) = 3\frac{km}{h} \tag{3.20}$$

The harmonic mean must be used when: (1) ratios are involved and (2) relative weights are indicated by numerator values (e.g. km). If the relative weights are given in the units of the denominator (e.g. hours), the arithmetic mean should be used. It should also be noted that the harmonic mean – like the geometric mean – is only defined for positive values greater than 0. For unequally sized observations, the following applies:

$$\bar{x}_{harm} < \bar{x}_{geom} < \bar{x} \tag{3.21}$$

3.2.5 The Median

As the mean is sometimes not "representative" of a distribution, an alternative is required to identify the central tendency. Consider the following example: You work at an advertising agency and must determine the average age of diaper users for a diaper ad. You collect the following data (Table 3.3):

Based on what we learned above about calculating the mean using the class midpoint of classed data, we get: $\bar{x}$ = 0.3·0.5 + 0.15·1.5 + 0.25·3.5 + 0.04·8

Table 3.3 Share of sales by age class for diaper users

Age class	Under 1	1	2–4	5–10	11–60	61–100
Relative frequency (%)	30	15	25	4	3	23
Cumulated: F(x) (%)	30	45	70	74	77	100

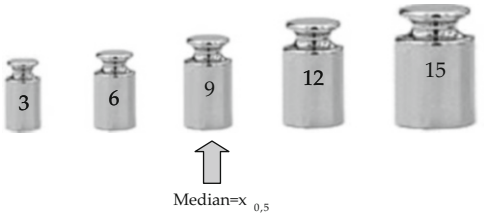


Fig. 3.14 The median: The central value of unclassed data

$+0.03{\cdot}36 + 0.23{\cdot}81 \approx 21$ years.[8] This would mean that the average diaper user is college age! This is doubtful, of course, and not just because of the absence of baby-care rooms at universities. The high values on the outer margins – classes 0–1 and 61–100 – create a bimodal distribution and paradoxically produce a mean in the age class in which diaper use is lowest.

So what other methods are available for calculating the average age of diaper users? Surely one way would be to find the modal value of the most important age group: 0–1. This value, the so-called *median,* not only offers better results in such cases. The median is also the value that divides the size-ordered dataset into two equally large halves. Exactly 50 % of the values are smaller and 50 % of the values are larger than the median.[9]

Figure 3.14 shows five weights ordered by *heaviness*. The median is $\tilde{x} = x_{0,5} = x_{(3)} = 9$, as 50 % of the weights are to the left and right of weight number 3.

There are several formula for calculating the median. When working with a raw data table – i.e. with unclassed data – most statistics textbooks suggest these formula:

$$\tilde{x} = x\left(\frac{n+1}{2}\right) \text{ for an odd number of observations (n)} \tag{3.22}$$

and

[8] To find the value for the last class midpoint, take half the class width – (101–61)/2 = 20 – and from that we get 61 + 20 = 81 years for the midpoint.

[9] Strictly speaking, this only applies when the median lies between two observations, which is to say, only when there are an even number of observations. With an odd number of observations, the median corresponds to a single observation. In this case, 50 % of (n-1) observations are smaller and 50 % of (n-1) observations are larger than the median.

$$\tilde{x} = \frac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right) \text{ for an even number of observations.} \tag{3.23}$$

If one plugs in the weights from the example into the first formula, we get:

$$\tilde{x} = x_{\left(\frac{n+1}{2}\right)} = x_{\left(\frac{5+1}{2}\right)} = x_{(3)} = 9 \tag{3.24}$$

The trait of the weight in the third position of the ordered dataset equals the median. If the median is determined from a classed dataset, as in our diaper example, the following formula applies:

$$\tilde{x} = x_{0.5} = x_{i-1}^{UP} + \frac{0.5 - F\left(x_{i-1}^{UP}\right)}{f(x_i)}\left(x_i^{UP} - x_i^{LOW}\right) \tag{3.25}$$

First we identify the class in which 50 % of observations are just short of being exceeded. In our diaper example this corresponds to the 1 year olds. The median is above the upper limit x_{i-1}^{UP} of the class, or 1 year. But how many years above the limit? There is a difference of 5 % points between the postulated value of 0.5 and the upper limit value of $F(x_{i-1}^{UP}) = 0.45$:

$$0.5 - F\left(x_{i-1}^{UP}\right) = 0.5/0.45 = 0.05 \tag{3.26}$$

This 5 % points must be accounted for from the next largest (ith) class, as it must contain the median. The 5 % points are then set in relation to the relative frequency of the entire class:

$$\frac{0.5 - F\left(x_{i-1}^{UP}\right)}{f(x_i)} = \frac{0.5 - 0.45}{0.25} = 0.2 \tag{3.27}$$

Twenty per cent of the width of the age class that contains the median must be added on by age. This results in a Δi of 3 years, as the class contains all persons who are 2, 3, and 4 years old. This produces a median of $\tilde{x} = 2 + 20\% \cdot 3 = 2.6$ years. This value represents the "average user of diapers" better than the value of the arithmetic mean. Here I should note that the calculation of the median in a bimodal distribution can, in principle, be just as problematic as calculating the mean. The more realistic result here has almost everything to do with the particular characteristics of the example. The median is particularly suited when many outliers exist (see Sect. 2.5). Figure 3.15 traces the steps for us once more.

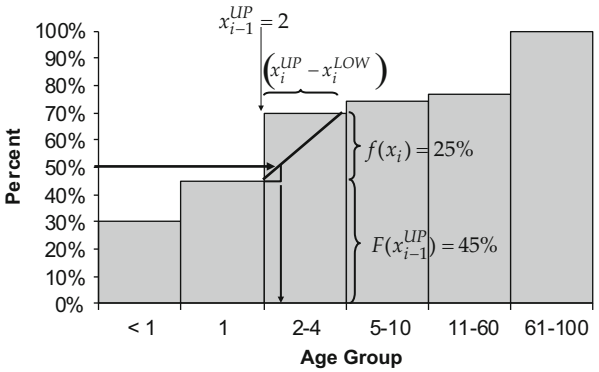


Fig. 3.15 The median: The middle value of classed data

3.2.6 Quartile and Percentile

In addition to the median, there are several other important measures of central tendency that are based on the quantization of an ordered dataset. These parameters are called *quantiles*. When quantiles are distributed over 100 equally sized intervals, they are referred to as *percentiles*. Their calculation requires an ordinal or cardinal scale and can be defined in a manner analogous to the median. In an ordered dataset, the p percentile is the value at which no less than p per cent of the observations are smaller or equal in value and no less than (1-p) per cent of the observations are larger or equal in value. For instance, the 17th percentile of age in our grocery store survey is 23 years old. This means that 17 % of the respondents are 23 years or younger, and 83 % are 23 years old or older. This interpretation is similar to that of the median. Indeed, the median is ultimately a special case ($p = 50$ %) of a whole class of measures that partitions the ordered dataset into parts, i.e. quantiles.

In practical applications, one particular important group of quantiles is known as the quartiles. It is based on an ordered dataset divided into four equally sized parts. These are called the first quartile (the lower quartile or 25th percentile), the second quartile (the median or 50th percentile), and the third quartile (the upper quartile or 75th percentile).

Although there are several methods for calculating quantiles from raw data tables, the weighted average method is considered particularly useful and can be found in many statistics programmes. For instance, if the ordered sample has a size of $n = 850$, and we want to calculate the lower quartile ($p = 25$ %), we first have to determine the product $(n + 1)\cdot p$. In our example, $(850 + 1)\cdot 0.25$ produces the value 212.75. The result consists of an integer before the decimal mark ($i = 212$) and a decimal fraction after the decimal mark ($f = 0.75$). The integer (i) helps indicate the values between which the desired quantile lies – namely, between the observations (i) and $(i + 1)$, assuming that (i) represents the ordinal numbers of the ordered dataset. In our case, this is between rank positions 212 and 213. Where exactly does the quantile in question lie between these ranks? Above we saw that

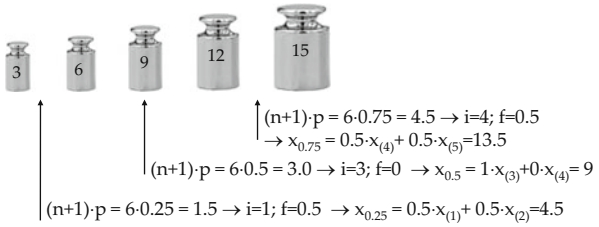


Fig. 3.16 Calculating quantiles with five weights

the total value was 212.75, which is to say, closer to 213 than to 212. The figures after the decimal mark can be used to locate the position between the values with the following formula:

$$(1-f)\cdot x_{(i)} + f\cdot x_{(i+1)} \tag{3.28}$$

In our butter example, the variable bodyweight produces these results:

$$(1-0.75)\cdot x_{(212)} + 0.75\cdot x_{(213)} = 0.25\cdot 63.38 + 0.75\cdot 63.44 = 63.43 \text{ kg} \tag{3.29}$$

Another example for the calculation of the quartile is shown in Fig. 3.16.

It should be noted here that the weighted average method cannot be used with extreme quantiles. For example, to determine the 99 % quantile for the five weights in Fig. 3.16 a sixth weight is needed, since $(n + 1)\cdot p = (5 + 1)\cdot 0.99 = 5.94$. This weight does not actually exist. It is fictitious, just like a weight of 0 for determining the 1 % quantile ($(n + 1)\cdot p = (5 + 1)\cdot 0.01 = 0.06$). In such cases, software programmes indicate the largest and smallest variable traits as quantiles. In the example case, we thus have: $x_{0.99} = 15$ and $x_{0.01} = 3$.

3.3 The Boxplot: A First Look at Distributions

We have now seen some basic measures of central tendency. All of these measures attempt to reduce dataset information to a single number expressing a general tendency. We learned that this reduction does not suffice to describe a distribution that contains outliers or special forms of dispersion. In practice, so-called boxplots are used to get a general sense of dataset distributions.

The boxplot combines various measures. Let's look at an example: Imagine that over a 3 year period researchers recorded the weekly sales of a certain brand of Italian salad dressing, collecting a total of 156 observations.[10] Part 1 of Fig. 3.17 shows the boxplot of weekly sales. The plot consists of a central box whose lower edge indicates the lower quartile and whose upper edge indicates the upper quartile. The values are chartered along the y-axis and come to 51,093 bottles sold for the

[10] The data can be found in the file *salad_dressing.sav* at springer.com.

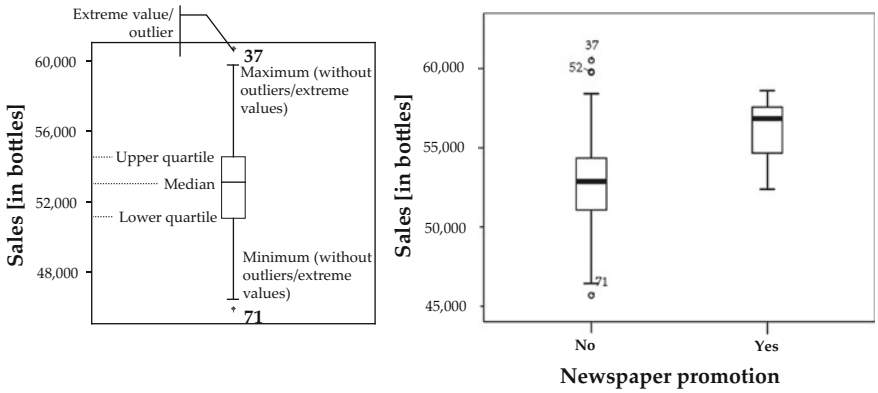


Fig. 3.17 Boxplot of weekly sales

lower quartile and 54,612 bottles sold for the upper quartile. The edges frame the middle 50 % of all observations, which is to say: 50 % of all observed weeks saw no less than 51,093 and no more than 54,612 bottles sold. The difference between the first and third quartile is called the *interquartile range*. The line in the middle of the box indicates the median position (53,102 bottles sold). The lines extending from the box describe the smallest and largest 25 % of sales. Known as whiskers, these lines terminate at the lowest and highest observed values, provided they are no less than 1.5 times the box length (interquartile range) below the lower quartile or no more than 1.5 times the box length (interquartile range) above the upper quartile. Values beyond these ranges are indicated separately as potential *outliers*. Some statistical packages like SPSS differentiate between *outliers* and *extreme values* – i.e. values that are less than three times the box length (interquartile range) below the lower quartile or more than three times the box length (interquartile range) above the upper quartile. These extreme values are also indicated separately. It is doubtful whether this distinction is helpful, however, since both outliers and extreme values require separate analysis (see Sect. 2.5).

From the boxplot in Part 1 of Fig. 3.17 we can conclude the following:

- Observations 37 and 71 are outliers above the maximum (60,508 bottles sold) and below the minimum (45,682 bottles sold), respectively. These values are fairly close to the edges of the whiskers, indicating weak outliers.
- Some 15,000 bottles separate the best and worst sales weeks. The smallest observation (45,682 bottles) represents a deviation from the best sales week of more than 30 %.
- In this example the median lies very close to the centre of the box. This means that the central 50 % of the dataset is symmetrical: the interval between the lower quartile and the median is just as large as the interval between the median and the upper quartile. Another aspect of the boxplot's symmetry is the similar length of the whiskers: the range of the lowest 25 % of sales is close to that of the highest 25 %.

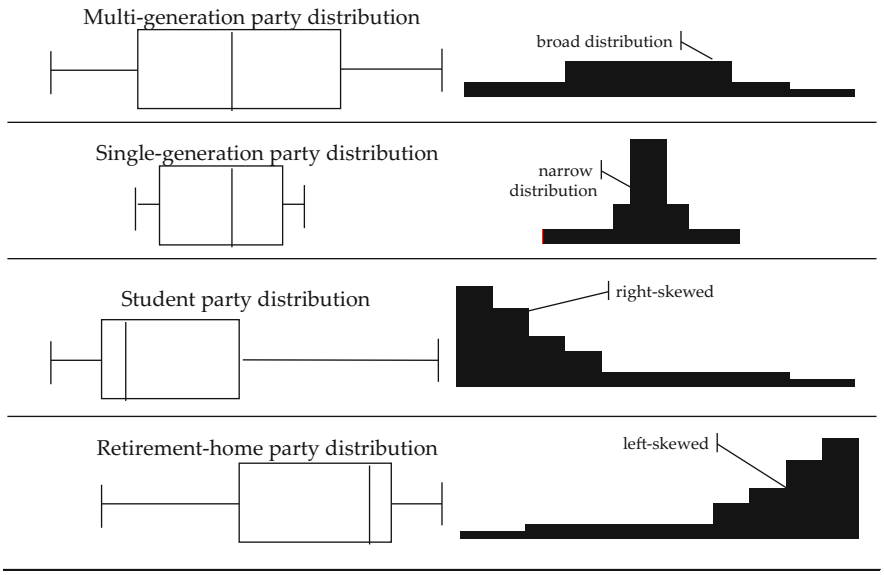


Fig. 3.18 Interpretation of different boxplot types

Figure 3.18 summarizes different boxplot types and their interpretations. The boxplots are presented horizontally, not vertically, though both forms are common in practice. In the vertical form, the values are read from the y-axis; in the horizontal form, they are read from the x-axis.

If the boxplot is symmetrical – i.e. with the median in the centre of the box and whiskers of similar length – the distribution is symmetrical. When the value spread is large, the distribution is flat and lacks a clear-cut modal value. Such a distribution results, for instance, when plotting ages at a party with guests from various generations. If the value spread is small – i.e. with a compact box and whiskers – the distribution is narrow. This type of distribution results when plotting ages at a party with guests from a single generation. Boxplots can also express asymmetrical datasets. If the median is shifted to the left and the left whisker is short, then the middle 50 % falls within a narrow range of relatively low values. The remaining 50 % of observations are mostly higher and distributed over a large range. The resulting histogram is right-skewed and has a peak on the left side. Such a distribution results when plotting the ages of guests at a student party. Conversely, if the median is shifted to the right and the right whisker is relatively short, then the distribution is skewed left and has a peak on the right side. Such a distribution results when plotting the ages of guests at a retirement-home birthday party.

In addition to providing a quick overview of distribution, boxplots allow comparison of two or more distributions or groups. Let us return again to the salad dressing example. Part 2 of Fig. 3.17 displays sales for weeks in which ads appeared in daily newspapers compared with sales for weeks in which no ads appeared. The boxplots show which group (i.e. weeks with or without newspaper

ads) has a larger median, a larger interquartile range, and a greater dispersion of values. Since the median and the boxplot box is larger in weeks with newspaper ads, one can assume that these weeks had higher average sales. In terms of theory, this should come as no surprise, but the boxplot also shows a left-skewed distribution with a shorter spread and no outliers. This suggests that the weeks with newspaper ads had relatively stable sales levels and a concentration of values above the median.

3.4 Dispersion Parameters

The boxplot provides an indication of the value spread around the median. The field of statistics has developed parameters to describe this spread, or dispersion, using a single measure. In the last section we encountered our first dispersion parameter: the *interquartile range*, i.e. the difference between the upper and lower quartile, which is formulated as

$$\text{IQR} = (x_{0.75} - x_{0.25}) \tag{3.30}$$

The larger the range, the further apart the upper and lower values of the midspread. Some statistics books derive from the IQR the *mid-quartile range*, or the IQR divided by two, which is formulated as

$$\text{MQR} = 0.5 \cdot (x_{0.75} - x_{0.25}) \tag{3.31}$$

The easiest dispersion parameter to calculate is one we've already encountered implicitly: *range*. This parameter results from the difference between the largest and smallest values:

$$\text{Range} = \text{Max}(\text{x}_\text{i}) - \text{Min}(\text{x}_\text{i}) \tag{3.32}$$

If the data are classed, the range results from the difference between the upper limit of the largest class of values and the lower limit of the smallest class of values. Yet we can immediately see why range is problematic for measuring dispersion. No other parameter relies so much on external distribution values for calculation, making range highly susceptible to outliers. If, for instance, 99 values are gathered close together and a single value appears as an outlier, the resulting range predicts a high dispersion level. But this belies the fact that 99 % of the values lie very close together. To calculate dispersion, it makes sense to use as many values as possible, and not just two.

One alternative parameter is the *median absolute deviation*. Using the median as a measure of central tendency, this parameter is calculated by adding the absolute deviations of each observation and dividing the sum by the number of observations:

$$\text{MAD} = \frac{1}{n} \sum_{i=1}^{n} |x_i - \tilde{x}| \tag{3.33}$$

In empirical practice, this parameter is less important than that of variance, which we present in the next section.

3.4.1 Standard Deviation and Variance

An accurate measure of dispersion must indicate average deviation from the mean. The first step is to calculate the deviation of every observation. Our intuition tells us to proceed as with the arithmetic mean – that is, by adding the values of the deviations and dividing them by the total number of deviations:

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}) \tag{3.34}$$

Here, however, we must recall a basic notion about the mean. In an earlier section we likened the mean to a balance scale: the sum of deviations on the left side equals the sum of deviations on the right. Adding together the negative and positive deviations from the mean always yields a value of 0. To prevent the substitution of positive with negative values, we can add the absolute deviation amounts and divide these by the total number of observations:

$$\frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}| \tag{3.35}$$

Yet statistics always make use of another approach: squaring both positive and negative deviations, thus making all values positive. The squared values are then added and divided by the total number of observations. The resulting dispersion parameter is called *empirical variance*, or *population variance,* and represents one of the most important dispersion parameters in empirical research:

$$Var(x)_{emp} = S^2_{emp} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{3.36}$$

The root of the variance yields the *population standard deviation,* or the *empirical standard deviation*:

$$S_{emp} = \sqrt{Var(x)_{emp}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{3.37}$$

Its value equals the average deviation from the mean. The squaring of the values gives a few large deviations more weight than they would have otherwise.

To illustrate, consider the observations 2, 2, 4, and 4. Their mean is three, or $\bar{x} = 3$. Their distribution has four deviations of one unit each. The squared sum of the deviations is:

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 = 1^2 + 1^2 + 1^2 + 1^2 = 4 \text{ units} \tag{3.38}$$

Another distribution contains the observations 2, 4, 4, and 6. Their mean is four, or $\bar{x} = 4$, and the total sum of deviations again is $2 + 2 = 4$ units. Here, two observations have a deviation of 2 and two observations have a deviation of 0. But the sum of the squared deviation is larger:

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 = 2^2 + 0^2 + 0^2 + 2^2 = 8 \text{ units} \tag{3.39}$$

Although the sum of the deviations is identical in each case, a few large deviations lead to a larger empirical variance than many small deviations with the same quantity ($S^2_{emp} = 1$ versus $S^2_{emp} = 2$). This is yet another reason to think carefully about the effect of outliers in a dataset.

Let us consider an example of variance. In our grocery store survey, the customers have an average age of 38.62 years and an empirical standard deviation of 17.50 years. This means that the average deviation from the mean age is 17.50 years.

Almost all statistics textbooks contain a second and slightly modified formula for variance or standard deviation. Instead of dividing by the total number of observations (n), one divides by the total number of observations minus 1 (n−1). Here one speaks of *unbiased sample variance*, or of *Bessel's corrected variance*:

$$Var(x) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{3.40}$$

Unbiased sample variance can then be used to find the *unbiased sample standard deviation*:

$$S = \sqrt{Var(x)} = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{3.41}$$

This is a common cause of confusion among students, who frequently ask "What's the difference?" Unbiased sample variance is used when we want to infer a population deviation from a sample deviation. This method of measuring variance is necessary to make an unbiased estimation of a population deviation

from a sample distribution when the mean of the population is unknown. If we use the empirical standard deviation (S_{emp}) of a sample instead, we invariably underestimate the true standard deviation of the population. Since, in practice, researchers work almost exclusively from samples, many statistics textbooks even forgo discussions of empirical variance. When large samples are being analyzed, it makes little difference whether the divisor is n or (n-1). Ultimately, this is why many statistics packages indicate only the values of unbiased sample variance (standard deviation), and why publications and statistics textbooks mean unbiased sample variance whenever they speak of variance, or S^2. Readers should nevertheless be aware of this fine distinction.

3.4.2 The Coefficient of Variation

Our previous example of customer age shows that, like the mean, the standard deviation has a unit – in our survey sample, years of age. But how do we compare dispersions measured in different units? Figure 3.19 shows the height of five children in centimetres and inches. Body height is dispersed $S_{emp} = 5.1$ cm – or $S_{emp} = 2.0$ in – around the mean. Just because the standard deviation for the inches unit is smaller than the standard deviation for the centimetres unit does not mean the dispersion is any less. If two rows are measured with different units, then the values of the standard deviation cannot be used as the measure of comparison for the dispersion. In such cases, the *coefficient of variation* is used. It is equal to the quotient of the (empirical or unbiased) standard deviation and the absolute value of the mean:

$$V = \frac{S}{|\bar{x}|}, \text{provided the mean does not have the value } \bar{x} = 0 \tag{3.42}$$

The coefficient of variation has no unit and expresses the dispersion as a percentage of the mean. Figure 3.19 shows that the coefficient of variation – 0.04 – has the same value regardless of whether body height is measured in inches or centimetres.

Now, you might ask, why not just convert the samples into a single unit (for example, centimetres) so that the standard deviation can be used as a parameter for comparison? The problem is that there are always real-life situations in which conversion either is impossible or demands considerable effort. Consider the differences in dispersion when measuring…

- …the consumption of different screws, if one measure counts the number of screws used, and the other total weight in grammes;
- …the value of sales for a product in countries with different currencies. Even if the average exchange rate is available, conversion is always approximate.

In such – admittedly rare – cases, the coefficient of variation should be used.

		Child no.					Mean	S_{emp}	Coefficient of variation
		1	2	3	4	5			
cm	x	120	130	125	130	135	128.0	5.1	0.04
in	y	48	52	50	52	54	51.2	2.0	0.04

Fig. 3.19 Coefficient of variation

3.5 Skewness and Kurtosis

The boxplot in Fig. 3.18 not only provides information about central tendency and dispersion, but also describes the symmetry of the distribution. Recall for a moment that the student party produced a distribution that was right-skewed (peak on the left), and the retirement-home birthday party produced a distribution that was left-skewed (peak on the right). Skewness is a measure of distribution asymmetry. A simple parameter from *Yule* & *Pearson* uses the difference between median and mean in asymmetric distributions. Look again at the examples in Fig. 3.20: In the right-skewed distribution there are many observations on the left side and few observations on the right. The student party has many young students (ages 20, 21, 22, 23, 24) but also some older students and young professors (ages 41 and 45). The distinguishing feature of the right-skewed distribution is that the mean is always to the right of the median, which is why $\bar{x} > \tilde{x}$. The few older guests pull the mean upward, but leave the median unaffected. In the left-skewed distribution, the case is reversed. There are many older people at the retirement-home birthday party, but also a few young caregivers and volunteers. The latter pull the mean downwards, moving it to the left of the median ($\bar{x} < \tilde{x}$). *Yule* & *Pearson* express the difference between median and mean as a degree of deviation from symmetry:

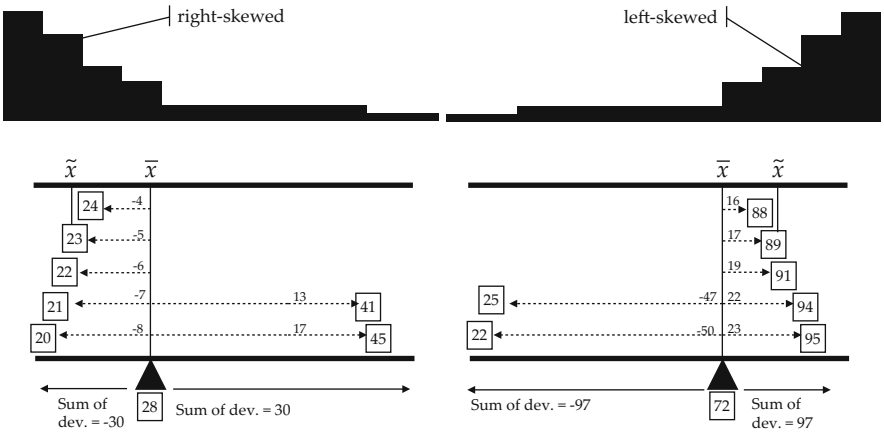
$$\text{Skew} = \frac{3 \cdot (\bar{x} - \tilde{x})}{S} \tag{3.43}$$

Values larger than 0 indicate a right-skewed distribution, values less than 0 indicate a left-skewed distribution, and values that are 0 indicate a symmetric distribution.

The most common parameter to calculate the skewness of a distribution is the so-called *third central moment*:

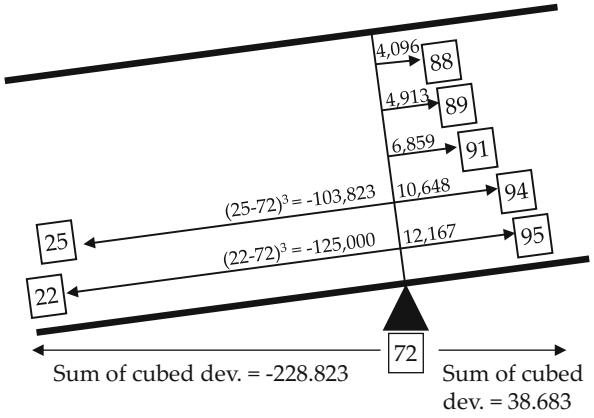
$$\text{Skew} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{S^3} \tag{3.44}$$

To understand this concept, think again about the left-skewed distribution of the retirement-home birthday party in Fig. 3.21. The mean is lowered by the young caregivers, moving it from around 91 years to 72 years. Nevertheless, the sum of deviations on the left and right must be identical. The residents of the



Note: The numbers in the boxes represent ages. The mean is indicated by the arrow. Like a balance scale, the deviations to the left and right of the mean are in equilibrium.

Fig. 3.20 Skewness



Note: The numbers in the boxes represent ages. The mean is indicated by the triangle. Like a balance scale, the cubed deviations to the left and right of the mean are in disequilibrium.

Fig. 3.21 The third central moment

retirement home create many small upward deviations on the right side of the mean (16, 17, 19, 22, 23). The sum of these deviations – 97 years – corresponds exactly to the few large deviations on the left side of the mean caused by the young caregivers (47 and 50 years).

But what happens if the deviations from the mean for each observation are cubed $\left((x_i - \bar{x})^3\right)$ before being summed? Cubing produces a value for caregiver ages of

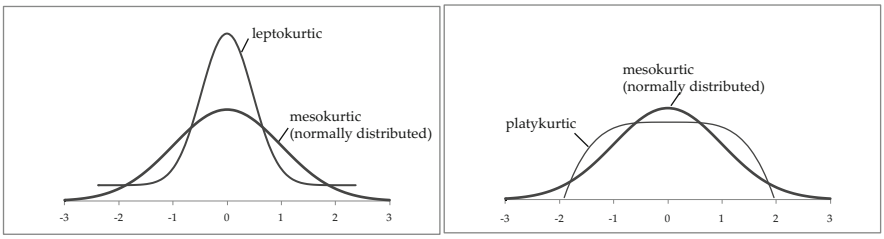


Fig. 3.22 Kurtosis distributions

−228,823 and a value for resident ages of 38,683. While the sums of the basic deviations are identical, the sums of the cubed deviations are different. The sum on the side with many small deviations is smaller than the sum on the side with a few large deviations. This disparity results from the mathematical property of exponentiation: relatively speaking, larger numbers raised to a higher power increase more than smaller numbers raised to a higher power. One example of this is the path of a parabolic curve.

The total sum of the values from the left and right hand sides results in a negative value of −190,140 (= −228,823 + 38,683) for the left-skewed distribution. For a right-skewed distribution, the result is positive, and for symmetric distributions the result is close to 0. A value is considered different than 0 when the absolute value of the skewness is more than twice as large as the *standard error* of the skew. This means that a skewness of 0.01 is not necessary different than 0. The standard error is always indicated in statistics programmes and does not need to be discussed further here.

Above we described the symmetry of a distribution with a single parameter. Yet what is missing is an index describing the bulge (pointy or flat) of a distribution. Using the examples in Fig. 3.18, the contrast is evident between the wide distribution of a multi-generation party and the narrow distribution of a single-generation party. *Kurtosis* is used to help determine which form is present. Defined as the *fourth central moment*, kurtosis is described by the following formula:

$$\text{Kurt} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{S^4} \tag{3.45}$$

A unimodal normal distribution as shown in Fig. 3.22 has a kurtosis value of three. This is referred to as a *mesokurtic* distribution. With values larger than three, the peak of the distribution becomes steeper, provided the edge values remain the same. This is called a *leptokurtic* distribution. When values are smaller than three, a flat peak results, also known as a *platykurtic* distribution. Figure 3.22 displays the curves of leptokurtic, mesokurtic, and platykurtic distributions.

When using software such as Excel or SPSS, similar parameters are sometimes calculated and displayed as an *excess*. But they normalize to a value of 0, not 3. The user must be aware of which formula is being used when calculating kurtosis.

Parameter	Level of Measurement			robust?
	nominal	ordinal	cardinal	
Mean	not permitted	not permitted	permitted	not robust
Median	not permitted	permitted	permitted	robust
Quantile	not permitted	permitted	permitted	robust
Mode	permitted	permitted	permitted	robust
Sum	not permitted	not permitted	permitted	not robust
Variance	not permitted	not permitted	permitted	not robust
Interquartile range	not permitted	not permitted	permitted	robust
Range	not permitted	not permitted	permitted	not robust
Skewness	not permitted	not permitted	permitted	not robust
Kurtosis	not permitted	not permitted	permitted	not robust

Note: Many studies use mean, variance, skewness, and kurtosis with ordinal scales as well. Section 2.2 describes the conditions necessary for this to be possible.

Fig. 3.23 Robustness of parameters

3.6 Robustness of Parameters

We previously discussed the effects of outliers. Some parameters, such as mean or variance, react sensitively to outliers; others, like the median in a bigger sample, don't react at all. The latter group are referred to as *robust* parameters. If the data include only robust parameters, there is no need to search for outliers. Figure 3.23 provides a summary of the permitted scales for each parameter and its robustness.

3.7 Measures of Concentration

The above measures of dispersion dominate empirical research. They answer (more or less accurately) the following question: To what extent do observations deviate from a location parameter? Occasionally, however, another question arises: How concentrated is a trait (e.g. sales) within a group of particular statistical units (e.g. a series of firms). For instance, the EU's Directorate General for Competition may investigate whether a planned takeover will create excessively high concentration in a given market. To this end, indicators are needed to measure the concentration of sales, revenues, etc.

The simplest way of measuring concentration is by calculating the *concentration ratio*. Abbreviated as CR_g, the concentration ratio indicates the percentage of a quantity (e.g. revenues) achieved by g statistical units with the highest trait values. Let's assume that five companies each have a market share of 20 %. The market concentration ratio CR_2 for the two largest companies is $0.2 + 0.2$, or 0.4. The other concentration rates can be calculated in a similar fashion: $CR_3 = 0.2 + 0.2 + 0.2 = 0.6$, etc. The larger the concentration ratio is for a given g, the greater the market share controlled by the g largest companies, and the larger the concentration. In Germany, g has a minimum

value of three in official statistics. In the United States, the minimum value is four. Smaller values are not published because they would allow competitors to determine each other's market shares with relative precision, thus violating confidentiality regulations.

Another very common measure of concentration is the *Herfindahl index*. First proposed by O.C. Herfindahl in a 1950 study of concentration in the U.S. steel industry, the index is calculated by summing the squared shares of each trait:

$$\mathrm{H} = \sum_{i=1}^{n} f(x_i)^2 \tag{3.46}$$

Let us take again the example of five equally sized companies (an example of low concentration in a given industry). Using the above formula, this produces the following results:

$$\mathrm{H} = \sum_{i=1}^{n} f(x_i)^2 = 0.2^2 + 0.2^2 + \ 0.2^2 + 0.2^2 + 0.2^2 = 0.2 \tag{3.47}$$

Theoretically, a company with 100 % market share would have a Herfindahl index value of

$$\mathrm{H} = \sum_{i=1}^{n} f(x_i)^2 = 1^2 + 0^2 + 0^2 + 0^2 + 0^2 = 1 \tag{3.48}$$

The value of the Herfindahl index thus varies between 1/n (provided all statistical units display the same shares and there is no concentration) and 1 (only one statistical unit captures the full value of a trait for itself; i.e. full concentration).

A final and important measure of concentration can be derived from the graphical representation of the *Lorenz curve*. Consider the curve in Fig. 3.25 with the example of a medium level of market concentration in Fig. 3.24. Each company represents 20 % of the market, or 1/5 of all companies. The companies are then ordered by the size of the respective trait variable (e.g. sales), from smallest to largest, on the x-axis. In Fig. 3.25, the x-axis is spaced at 20 % point intervals, with the corresponding cumulative market shares on the y-axis. The smallest company (i.e. the lowest 20 % of companies) generates 10 % of sales. The two smallest companies (i.e. the lowest 40 % of the companies) generate 20 % of sales, while the three smallest companies generate 30 % of sales, and so on.

The result is a "sagging" curve. The extent to which the curve sags depends on market concentration. If the market share is distributed equally (i.e. five companies, each representing 20 % of all companies), then every company possesses 20 % of the market. In this case, the Lorenz curve precisely bisects the coordinate plane. This 45-degree line is referred to the *line of equality*. As concentration increases or deviates from the uniform distribution, the Lorenz curve sags more, and the area between it and the bisector increases. If one sets the area in relationship to the entire area below the bisector, an index results between 0 (uniform distribution, since

	Concentration		
	Minimum	Medium	Maximum
Share of Company 1	20%	50%	100%
Share of Company 2	20%	20%	0%
Share of Company 3	20%	10%	0%
Share of Company 4	20%	10%	0%
Share of Company 5	20%	10%	0%
CR_2	40%	70%	100%
CR_3	60%	80%	100%
Herfindahl	0.20	0.32	1.00
GINI	0	0.36	0.80
$GINI_{norm.}$	0	0.45	1

Fig. 3.24 Measure of concentration

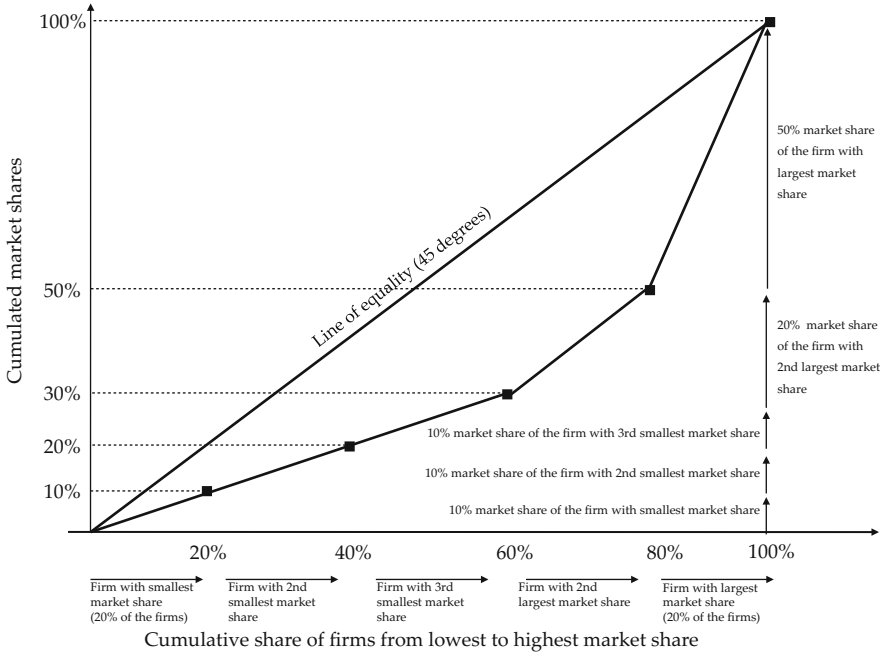


Fig. 3.25 Lorenz curve

otherwise the area between the bisector and the Lorenz curve would be 0) and (n−1)/n (full possession of all shares by a statistical unit):

$$\text{GINI} = \frac{\text{Area between bisector and the Lorenz curve}}{\text{Entire area below the bisector}} \tag{3.49}$$

This index is called the *Gini coefficient*. The following formulas are used to calculate the Gini coefficient:

(a) For unclassed ordered raw data:

$$\text{GINI} = \frac{2\sum_{i=1}^{n} i \cdot x_i - (n+1)\sum_{i=1}^{n} x_i}{n\sum_{i=1}^{n} x_i} \tag{3.50}$$

(b) For unclassed ordered relative frequencies:

$$\text{GINI} = \frac{2\sum_{i=1}^{n} i \cdot f_i - (n+1)}{n} \tag{3.51}$$

For the medium level of concentration shown in Fig. 3.24, the Gini coefficient can be calculated as follows:

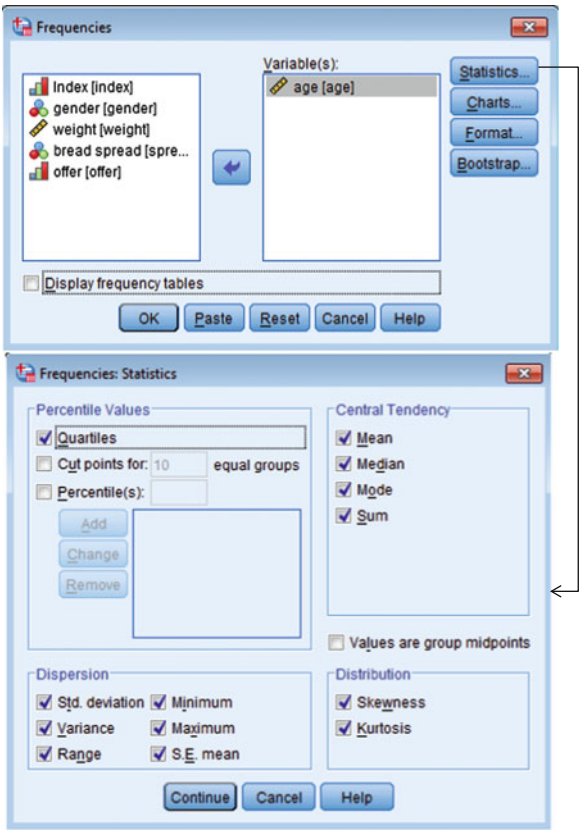
$$\begin{aligned} \text{GINI} &= \frac{2\sum_{i=1}^{n} i \cdot f_i - (n+1)}{n} \\ &= \frac{2 \cdot (1 \cdot 0.1 + 2 \cdot 0.1 + 3 \cdot 0.1 + 4 \cdot 0.2 + 5 \cdot 0.5) - (5+1)}{5} \\ &= 0.36 \end{aligned} \tag{3.52}$$

In the case of full concentration, the Gini coefficient depends on the number of observations (n). The value GINI = 1 can be approximated only when a very large number of observations (n) are present. When there are few observation numbers ($n < 100$), the Gini coefficient must be normalized by multiplying each of the above formulas by $n/(n-1)$. This makes it possible to compare concentrations among different observation quantities. A full concentration always yields the value $\text{GINI}_{\text{norm.}} = 1$.

3.8 Using the Computer to Calculate Univariate Parameters

3.8.1 Calculating Univariate Parameters with SPSS

This section uses the sample dataset *spread.sav*. There are two ways to calculate univariate parameters with SPSS. Most descriptive parameters can be calculated by clicking the menu items *Analyze* → *Descriptive Statistics* → *Frequencies*. In the menu that opens, first select the variables that are to be calculated for the univariate statistics. If there's a cardinal variable among them, deactivate the option *Display frequency tables*. Otherwise, the application will calculate contingency tables that don't typically produce meaningful results for cardinal variables. Select *Statistics*. . . from the submenu to display the univariate parameters for calculation.



Statistics

age

N	Valid	854
	Missing	0
Mean		38.58
Std. Error of Mean		0.598
Median		30.00
Mode		25
Std. Deviation		17.472
Variance		305.262
Skewness		.823
Std. Error of Skewness		.084
Kurtosis		-.694
Std. Error of Kurtosis		.167
Range		74
Minimum		18
Maximum		92
Sum		32946
Percentiles	25	25.00
	50	30.00
	75	55.00

Note: Applicable syntax commands: Frequencies; Descriptives

Fig. 3.26 Univariate parameters with SPSS

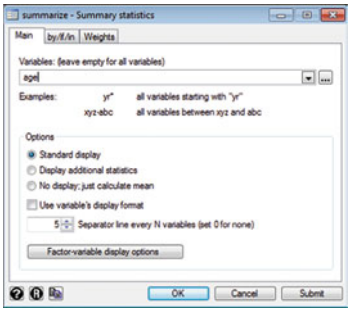
SPSS uses a standard kurtosis of 0, not 3. Figure 3.26 shows the menu and the output for the age variable from the sample dataset.

Another way to calculate univariate statistics can be obtained by selecting *Analyze* → *Descriptive Statistics* → *Descriptives*.... Once again, select the desired variables and indicate the univariate parameters in the submenu *Options*.

Choose *Graphs* → *Chart Builder*... to generate a boxplot or other graphs.

3.8.2 Calculating Univariate Parameters with Stata

Let's return again to the file *spread.dta*. The calculation of univariate parameters with Stata can be found under *Statistics* → *Summaries, tables, and tests* → *Summary and descriptive statistics* → *Summary statistics*. From the menu select the variables to be calculated for univariate statistics. To calculate the entire range of




```
. summarize age

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         age |       850    38.61765    17.50163        18         92

. summarize age, detail

                            alter
-------------------------------------------------------------
      Percentiles      Smallest
 1%           18            18
 5%           20            18
10%           22            18       Obs                 850
25%           25            18       Sum of Wgt.         850

50%           30                     Mean            38.61765
                       Largest       Std. Dev.       17.50163
75%           55            83
90%           66            85       Variance        306.3071
95%           71            89       Skewness        .8151708
99%           80            92       Kurtosis        2.290657
```

Note: Applicable syntax commands for univariate parameters: ameans; centile; inspect; mean;pctile; summarize; mean; tabstat; tabulate summarize.

Fig. 3.27 Univariate parameters with Stata

descriptive statistics, make sure to select *Display additional statistics*, as otherwise only the mean, variance, and smallest and greatest values will be displayed. Figure 3.27 shows the menu and the output for the variable age in the sample dataset.

To see the graphs (boxplot, pie charts, etc.) select *Graphics* from the menu.

3.8.3 Calculating Univariate Parameters with Excel 2010

Excel contains a number of preprogrammed statistical functions. These functions can be found under *Formulas* → *Insert Function*. Select the category *Statistical* to set the constraints. Figure 3.28 shows the Excel functions applied to the dataset *spread.xls*. It is also possible to use the Add-in Manager[11] to permanently activate the *Analysis ToolPak* and the *Analysis ToolPak VBA* for Excel 2010. Next, go to *Data* → *Data Analysis* → *Descriptive Statistics*. This function can calculate the most important parameters. Excel's graphing functions can also generate the most important graphics. The option to generate a boxplot is the only thing missing from the standard range of functionality.

Go to http://www.reading.ac.uk/ssc/n/software.htm for a free non-commercial, Excel statistics add-in (SSC-Stat) download. In addition to many other tools, the add-in allows you to create boxplots.

Excel uses a special calculation method for determining quantiles. Especially with small samples, it can lead to implausible results. In addition, Excel scales the kurtosis to the value 0 and not 3, which equals a subtraction of 3.

[11] The Add-In Manager can be accessed via *File* → *Options* → *Add-ins* → Manage: Excel Add-ins → Go...

Example: Calculation of univariate parameters of the dataset *spread.xls*

Variable Age

Parameter	Symbol	Result	Excel Command/Function
Count	N	850	=COUNT(Data!$C$2:$C$851)
Mean	$\overline{x}$	38.62	=AVERAGE(Data!$C$2:$C$851)
Median	$\tilde{x}$	30.00	=MEDIAN(Data!$C$2:$C$851)
Mode	x_{mod}	25.00	=MODALWERT(Data!$C$2:$C$851)
Trimmed Mean	x_{trim}	37.62	=TRIMMEAN(Data!$C$2:$C$851;0,1)
Harmonic Mean	x_{harm}	32.33	=HARMEAN(Data!$C$2:$C$851)
25th percentile	$x_{0.25}$	25.00	=PERCENTILE(Data!$C$2:$C$851;0,25)
50th percentile	$x_{0,5}$	30.00	=PERCENTILE(Data!$C$2:$C$851;0,5)
75th percentile	$x_{0,75}$	55.00	=PERCENTILE(Data!$C$2:$C$851;0,75)
Minimum	MIN	18.00	=MIN(Data!$C$2:$C$851)
Maximum	MAX	92.00	=MAX(Data!$C$2:$C$851)
Sum	Σ	32,825.00	=SUM(Data!$C$2:$C$851)
Standard Deviation	S_{emp}	17.50	=STDEVP(Data!$C$2:$C$851)
Standard Deviation	S	17.49	=STDEV(Data!$C$2:$C$851)
Empirical Variance	VAR_{emp}	306.31	=VARP(Data!$C$2:$C$851)
Unbiased Variance	VAR	305.95	=VAR(Data!$C$2:$C$851)
Skewness		0.82	=SKEW(Data!$C$2:$C$851)
Kurtosis		-0.71	=KURT(Data!$C$2:$C$851)

Fig. 3.28 Univariate parameters with Excel

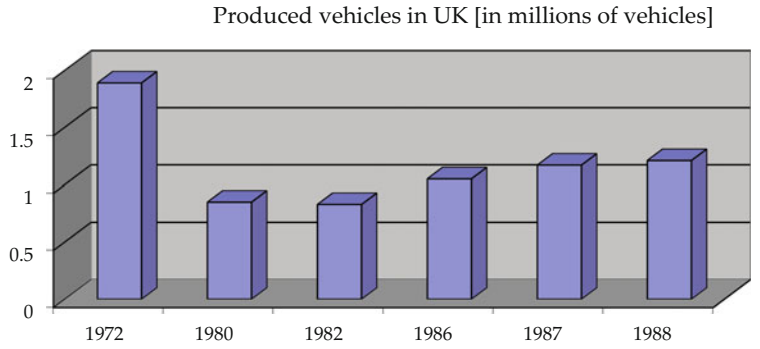
3.9 Chapter Exercises

Exercise 4:
A spa resort in the German town of Waldbronn conducts a survey of their hot spring users, asking how often they visit the spa facility. This survey results in the following absolute frequency data:

first time	rarely	regularly	frequently	every day
15	75	45	35	20

1. Identify the trait (level of measurement).
2. Sketch the relative frequency distribution of the data.
3. Identify the two location parameters that can be calculated and determine their size.
4. Identify one location parameter that can't be calculated. Why?

Exercise 5:
Supposed the following figure appears in a market research study. What can be said about it?



Exercise 6:
Using the values 4, 2, 5, 6, 1, 6, 8, 3, 4, and 9 calculate...
(a) The median
(b) The arithmetic mean
(c) The mean absolute deviation from the median
(d) The empirical variance
(e) The empirical standard deviation
(f) The interquartile range

Exercise 7:
The arithmetic mean $\bar{x} = 10$ and the empirical standard deviation $S_{emp} = 2$ were calculated for a sample ($n = 50$). Later the values $x_{51} = 18$ und $x_{52} = 28$ were added to the sample. What is the new arithmetic mean and empirical standard deviation for the entire sample ($n = 52$)?

Exercise 8:
You're employed in the marketing department of an international car dealer. Your boss asks you to determine the most important factors influencing car sales. You receive the following data:

Country	Sales [in 1,000 s of units]	Number of dealerships	Unit price [in 1,000 s of MUs]	Advertising budget [in 100,000 s of MUs]
1	6	7	32	45
2	4	5	33	35
3	3	4	34	25
4	5	6	32	40
5	2	6	36	32
6	2	3	36	43
7	5	6	31	56
8	1	9	39	37
9	1	9	40	23
10	1	9	39	34

(a) What are the average sales (in 1,000 s of units)?
(b) What is the empirical standard deviation and the coefficient of variation?
(c) What would be the coefficient of variation if sales were given in a different unit of quantity?
(d) Determine the lower, middle, and upper quartile of sales with the help of the "weighted average method".
(e) Draw a boxplot for the variable sales.
(f) Are sales symmetrically distributed across the countries? Interpret the boxplot.
(g) How are company sales concentrated in specific countries? Determine and interpret the Herfindahl index.
(h) Assume that total sales developed as follows over the years: 1998: +2 %; 1999: +4 %; 2000: +1 %. What is the average growth in sales for this period?

Exercise 9:

(a) A used car market contains 200 vehicles across the following price categories:

Car price (in €)	Number
Up to 2,500	2
Between 2,500 and 5,000	8
Between 5,000 and 10,000	80
Between 10,000 and 12,500	70
Between 12,500 and 15,000	40

(a) Draw a histogram for the relative frequencies. How would you have done the data acquisition differently?
(b) Calculate and interpret the arithmetic mean, the median, and the modal class.
(c) What price is reached by 45 % of the used cars?
(d) 80% of used cars in a different market are sold for more than €11,250. Compare this value with the market figures in the above table.

Exercise 10:

Unions and employers sign a 4-year tariff agreement. In the first year, employees' salaries increase by 4 %, in the second year by 3 %, in the third year by 2 %, and in the fourth year by 1 %. Determine the average salary increase to four decimal places.

Exercise 11:

A company has sold €30 m worth of goods over the last 3 years. In the first year they sold €8 m, in the second year €7 m, in the third year €15 m. What is the concentration of sales over the last 3 years? Use any indicator to solve the problem.