Chapter 6

Quantitative data: univariate and bivariate analysis

In this chapter, we start with more formal notation for key concepts such as variables and respondents. Next, in section 6.2 measurement levels are discussed, which are important for understanding the information content of variables, and necessary to take into consideration for choosing one's analysis technique. Subsequently, we discuss a number of descriptive measures for univariate analysis, such as the mean and the standard deviation. We continue with measures to describe the association between variables, that is to say, bivariate analysis. We discuss two important classes of measures to investigate association: through correlation coefficients and through the analysis of cross-tabulations. We give some examples of graphs that can be used to display findings attractively. In section 6.4.3 we sketch univariate and bivariate analysis of duration variables, such as time to settlement of disputes.[1]

6.1 Analysis units versus variables

When we carry out research we usually – as stated on a number of occasions above – collect information on entities (citizens, lawyers, villages, yoghurt desserts, light bulbs, court cases) and we generally do so with regard to a number of their properties (religious affiliation, population structure, creaminess of texture, lifetime, time to completion). The entities are in methodology-speak called 'respondents' or 'analysis units'. Although the term 'respondents' is a universally used term, as the examples show, respondents can also be non-people, such as files that contain information, or structures such as buildings, or conglomerates such as cities, occurrences such as charges or victim impact statements, or even household appliances.

Often, when we are not studying persons, we indicate the entities as the entities they are or simply as 'analysis units' or 'replications' – as 'respondent' then has an

[1] This section can be skipped by readers with no particular interest in the analysis of such variables as it is not essential for understanding the subsequent chapters.

odd feel to it. Thus we say that we have studied bankruptcy with 200 replications over bankruptcy cases, or that we have studied the life-time of 500 light bulbs. In general, the number of replications over entities is indicated with the symbol 'N'. In the bankruptcy cases we just gave, N equals 200, and for the light bulbs N equals 500.

What we are interested in is the properties of these entities: respondents' political affiliations, profession, age, gender, income, intelligence, result in the half-marathon. When the entities are not people, we are still interested in their properties, such as time to failure of light bulbs, amounts paid out to creditors in bankruptcy cases, creaminess for yoghurt desserts. These properties are called *variables*. Thus creaminess is a variable, as is lifetime for the light bulb and amount paid out in bankruptcy settlements. We may measure more than one variable per respondent. In general for the light bulb, we would not have a lot to measure other than lifetime, as light bulbs are fabricated in identical fashion, although we might want to measure weight, and the time when the bulb was produced (it might be that bulbs tend to be less solidly produced and to function for a shorter period when the machines need oiling or maintenance and repair). In general, whenever we are investigating 'richer' phenomena, such as court cases, we tend to register more than one variable.

Indeed, often we are not only interested in one, but in several properties of our units of observation. This is so because generally we not only want to describe a variable, but also to describe its association with other variables, or even *explain* it. We might for instance want to know what factors determine amounts paid out in bankruptcy cases, or what the profile is of bankruptcy cases in which we typically find suspicions (or proof) of an *actio Pauliana*. The number of variables measured is generally referred to as 'k'. In some cases of social research, several hundreds of variables may be measured depending on the research question at hand and the richness of the available material.

Variables are usually, if we write things more formally and mathematically, given symbols – such as X instead of 'age' or Y instead of 'amount paid out'. When we have several variables, they are generally indicated with a subscript, such as X_1, X_2, X_3, etc., to indicate the first, second and third variable. As we have often at least several replications over respondents, the respondents can also be indicated with an index. The notation then becomes in purely abstract notation X_{ij}, with X_{ij} indicating the score of the i-th person on the j-th variable. See Table 6.1, which has data for four respondents on three variables: age, gender and type of legal dispute the respondent is involved in.

Table 6.1: Example of notation of variables and respondents

respondent (number)	age	gender	type of legal dispute
John (1)	46	male	unemployment benefits
Beth (2)	30	female	tort
Pete (3)	45	male	bankruptcy
Mary (4)	21	female	medical malpractice

In Table 6.1, X_{11} is 46 and X_{12} is 'male'. Pete's score on 'age' is X_{31}, being the score of the third person on the first variable, and Mary's score 'female' is indicated as X_{42}, and X_{43} is 'medical malpractice', being the legal dispute Mary is involved in.

6.2 Measurement levels

6.2.1 Nominal to absolute measurement level

Variables can have varying information content. When a variable measures age, for instance, the values of that variable contain a lot of information. If one respondent is 50 years old, and another is 25 years old, we can safely say that the first respondent is twice as old as the second. And whether we measure age in weeks, or years, or seconds, that conclusion remains the same. If however our variable measures religion, we can say no such thing. If a first respondent has the category 'Buddhist' and the second has the category 'Christian', the only inference we can draw is that the second person's religion is different from the first person's. We cannot say that the first person has more or less of the property 'religion' or that the second has twice as much, or half, the religion that the second one has. The only information content that the variable 'religion' has is equality and inequality: we know whether our analysis units have the same or a different score on that variable, but we cannot order the categories, let alone infer whether the categories are a multiple of each other. The information content that categories of a variable have is commonly referred to as a variable's *measurement level*.

In general, five measurement levels are distinguished, which can be rank-ordered as to the information content of the categories. These measurement levels are: *nominal* measurement level, *ordinal*, *interval*, *ratio* and *absolute*. We will discuss each of them in turn.

(1) *Nominal measurement level*

At the nominal measurement level ('nominal' from the Latin *nomen* meaning 'name'), the categories of the variables are nothing more than distinct categories, which have no other relation than that they are different. The variable 'Type of legal dispute' is such a variable. The legal disputes could be coded as 1='unemployment benefits', 2 = 'tort', 3 = 'bankruptcy', and 4 = 'medical malpractice'. But they could also have been coded as 100 = 'unemployment benefits', -12.2 = 'tort', 27 = 'bankruptcy', and 0.004 = 'medical malpractice'. The numbers have no other meaning than to distinguish the different categories. As such, it would not make sense to compute a mean, or some such statistic.

As such, a nominal variable is a variable that classifies the respondents into distinct sets of different classes. Other examples of nominal variables are 'type of court', 'hair colour', 'gender', etc.

(2) *Ordinal measurement level*

When we move up to the ordinal measurement level, the categories of the variable reflect an *ordering* – the name 'ordinal' says it. This means that if John has score '1', Pete has score '2' and Mary has score '3', John has the least of

property that the variable measures, next comes Pete, and Mary has the most. If the property is for instance likeability, this means that John is the least likeable, next Pete, and that Mary is the most likeable of the three. As the variable is an ordinal variable, John, Pete and Mary could however as well have had scores of -187, +2 and 236,000. As the scores of an ordinal variable reflect an ordering only, from these values we would also draw the same conclusion: John is least likeable and Pete follows him, with Mary having the highest score and therefore being the most likeable.

An example of such a variable is the ordering of winners in a half-marathon. In that case we would know that Jet ended first and Klara second and Bibi third, but we would not know their actual times. It could be that Jet ran in 1:59:03 and Klara in 1:59:59, and Bibi in 2:11:04, in which case Jet and Klara are very close and Bibi came in quite a bit later. However, it could also be that Jet ran in 1:52:45, Klara in 2:13:33 and Bibi in 2:13:58, in which case Klara and Bibi ended with only a small difference and Jet was way ahead of these two. That information is not contained in an ordinal variable, however.

Ordinal variables are encountered often in so-called 'preference judgements', for example when respondents are asked to say which of certain foodstuffs or smells they like best, which one comes next, which afterwards, or to compare such items pairwise (i.e. A against B, A against C, A against D, B against C, etc.).

When variables are ordinal, special techniques must be used to analyse them. Clearly, also for ordinal variables computing a mean score is not informative. What we could do, obviously, is compute a mean *rank* score, and carry out our computations on that in a sense derived variable. Such approaches will be discussed below in section 6.3.1.

(3) *Interval measurement level*

Moving up the measurement level ladder, we arrive at the interval level. When a variable has been measured at interval level, we can compute a mean score, and get a meaningful summary of the scores on that variable. As such, the category values of an interval level variable contain more information than just equality and difference (which a nominal variable had) or ordering (which an ordinal variable had).

If a variable is an interval variable, we know that the intervals are equidistant: the difference between a score of 4 and 5 is just as large as the difference between a score of 1 and 2 or between 3 and 4. Examples are rating scales, where we ask respondents to rate to what extent they agree with a certain statement, for instance on a scale from 1 to 7 or 1 to 4.

Does this mean that if person A has scored '10' on the interval level variable and person B has scored '5', that we may conclude that person A has twice as much of the property the variable is supposed to measure? This cannot be done because an interval level variable has no *fixed origin*.

Let us explain this with an example. Temperature can be measured in degrees Celsius but also in degrees Fahrenheit. Suppose that on one day it is 24 degrees Celsius in Amsterdam, and 48 degrees Celsius in Khartoum. While 48 is twice 24, it does not make sense to say that it is twice as warm in Khartoum as in Amsterdam, because those temperatures could be equally well have been measured in degrees Fahrenheit (namely as 118 and 75 respectively) but would then have a very different ratio. That this is so is because Celsius and Fahrenheit have different points of origin.

(4) *Ratio measurement level*

A ratio measurement level variable is a variable that does have a fixed origin. Examples are weight or price. Prices for instance may be given in US dollars, British pounds or euros, so the scale may vary. But if something is twice as expensive in dollars, it is also twice as expensive when the price is calculated in euros. And a product that is free is free in whatever currency prices are given. This means that from a ratio variable onwards, we can make statements such as 'product A is twice as expensive as product B'. The same goes for height, which may be expressed in yards or metres, but also here little John is half his father's height, regardless of the metric in which these heights are measured. This means that we may – formally speaking – change the scores in the following manner: $X' = aX$.

(5) *Absolute measurement level*

Once we have arrived at the highest measurement level, a variable's value contains all information there is. This also implies that we can no longer alter or transform the scores without affecting the meaning of the score: the scores we have measured are the final, ultimate, only measurements.

An example of an absolute variable is a variable that measures the frequency of occurrence of some phenomenon, for instance the number of children that respondents have. If someone reports that she has four children, we cannot transform this to eight children, or two. She simply has four children and that's it.

If we summarize the above, we note first that the various measurement levels are ordered in the sense that as we move up from nominal to absolute, variable values contain increasingly more information at each subsequent level. At nominal interval level, a variable does not do more than disaggregate the respondents into different categories. Once we are at ordinal level, these categories are ordered; we now know who has more and who has less of the property that the variable measures. At interval level, the ordering becomes more structured, and we may transform the data to present them on a different scale. Once at ratio level, the information content of the data is already so high that the only transformation that is allowed is multiplication or division. At absolute measurement level, the scores are exactly the information content of the variable and may thus not be altered any more.

The terminology employed here is commonly used. We may also encounter nominal variables as *categorical variables*; sometimes the same is said of ordinal variables.

By contrast, interval and ratio level variables are often referred to as *continuous* variables; they are also encountered under the label *numerical* variables. Variables with only two categories are labelled *dichotomous* variables.

6.2.2 Implications for analysis

Now why is it important at all to know what measurement level a variable has? It is important because the measurement level dictates the analysis method that can be used. To give a very simple example: if we have measured a variable at nominal level, it does not make sense to compute a mean score: there is for instance no such thing as a mean 'religion'. Conversely, if we have measured a variable at ratio level (for instance 'age'), there are more efficient ways of describing the measurement than through a frequency tabulation that gives the number of cases with ages 18, 19, 20, ..., 65.

This applies to the choice of very simple descriptive statistics, as well as to the choice of more sophisticated methods. Knowing the measurement level is therefore important to be able to decide what methods can be used to summarize and analyse the data.

The interval measurement level, in a sense, constitutes a watershed, as from interval measurement level onwards we can use standard quantitative statistical techniques. Ordinal variables need special methods and so do nominal variables. Dichotomous variables (variables with just two categories) are a special case, as technically speaking they are equivalent to interval variables, but in research practice they command special analysis techniques. Examples of such variables are 'gender', 'agree–disagree' judgements, 'guilty–not guilty' verdicts, etc.

6.3 Describing the variables

Suppose we interviewed 100 secondary school students. We asked these students to tell us whether they committed during the past school year any of the three following offences: theft, assault, or vandalism. We noted the gender of these respondents, the type of secondary school they are in, and we asked them how much pocket money they received.

If we want to report on our findings, we need to summarize them. What the reader would want to know is things like: how many students committed an offence last year? How many property or violent offences were committed on average? Did students differ a lot in how much pocket money they had?

If we want to provide the answers to such questions, we need simple summarizing measures. Such summarizing measures come in two classes. The first are *measures of central tendency*, which tell us what scores respondents tend to have, what the average of the students is. Second, what a reader would want to know is whether all respondents tended to score on or around the average, or whether there was quite a bit of fluctuation in respondents' scores. In other words, the reader would want to have a measure of *variability* as well.

These measures, measures of central tendency and of variation, are labelled 'statistics', not because they incorporate complicated statistical reasoning, but because they

in the etymological sense summarize a set of numerical values into one number: a statistic.

In the following we will discuss both groups of measures: measures of central tendency and measures of variability. With such measures we do a *univariate analysis*: we present the results for just one variable.

6.3.1 Central tendency: the mean, median and mode

If we want to give the reader an overview of our findings, we generally start by giving mean values, or so-called measures of central tendency. The average or 'arithmetic mean' is the run-of-the-mill average that everyone employs automatically when we ask after his or her average grade or alcohol consumption. In formula it is written as:

$$\sum_{i=1}^{N} \frac{X_i}{N},$$

which is nothing more than the sum of all observations X_i from the first to the N-th respondent, divided by the total number of respondents N – or our standard run-of-the-mill mean. The average or mean is the most often used measure of central tendency. It is attractive because it is easily and intuitively understood by readers. The mean is written as $\bar{X}$, as M_X, or simply as M.

A disadvantage of the mean is, however, is that it is sensitive to so-called outliers, i.e., extreme scores. Suppose for instance that we have a list of observations 1-2-5-7-3-6-6, of which the average is 4.29. But suppose now that the highest score was not 7 but 27 – an extreme score. The average would then suddenly be 7.14! The value 7.14 is obviously not a good summary of the scores. In fact, all scores but the extreme 27 lie below that summary value. This means that the reader does not get the right impression, a good overview of the scores 1-2-3-5-6-6-27.

In such cases, the *median* is a good alternative. The median is that number below which 50% of the values lie (and above which thus also 50%). It is that number that cuts the distribution of scores in two. In the example we just gave, the median is equal to 5, with three values below, and three values above it. Note that the median is indeed insensitive to outliers: whether the highest score is 7, 27, or 27,000, the median remains 5. The median is therefore a good measure of central tendency whenever there are outliers, or whenever the data have a skewed distribution. In empirical legal studies, we often encounter skewed variables, such as amount of damages awarded or sentence length.

Lastly, for summarizing the scores on a variable, the *mode* may be used. The mode is the score that occurs most often (as reflected in 'modal'). The mode is not often used as the only measure for central tendency, one reason for which may be that distributions may have more than one mode (they are then called bimodal as opposed to unimodal). Suppose for instance that in the example of the 100 students given above, an equal number were to own up to vandalism and property offending: we would then report two modes.

6.3.2 Measures of variability

Central tendency is not the only relevant measure to summarize the distribution of scores. In practice, we do not only want to know what the average number of offences was, but we would also want to know whether all respondents scored the mean or close to the mean or whether on the other hand there was a lot of variation in scores. In practice two measures are used to express such variability. The first is the *variance*, often used in computations but not aiding the reader so much in understanding what is going on. The second is the *standard deviation*, which is more insightful to gauge the extent to which respondents vary around the mean. We will detail them both.

The first measure is the *variance*; it is denoted with the symbol s^2. The variance is computed as the mean sum of squared deviations of the individual scores from the mean.[2] If you do not understand this right away, that illustrates why the variance is of limited conceptual use. A problem with the variance as a measure of variability is namely that because of the squaring of the differences between means and scores, the relation with the original scaling of the variable is gone. It is hard to interpret the number that is generated by the formula.

For that reason, in most situations, the so-called standard deviation is used, which is simply the square root of the variance:

$$s = \sqrt{s^2}.$$

The main advantage of the standard deviation compared to the variance is that it is, because of the square root, expressed again in the original scale of the variable. This means that the standard deviation can easily be interpreted. If for instance the mean number of offences of secondary schoolers is 1.3 and the standard deviation is 0.7, this can be rephrased by saying that on average the secondary schoolers varied by 0.7 offences around the arithmetic mean. It can also be envisaged as the scores hovering – on average – between 0.6 and 2.0. The advantage of the standard deviation is that it gives you a 'feel' for how much fluctuation there was. The standard deviation is often abbreviated as 'sd' or 'SD'.

[2] s^2 is computed as:

$$\sum_{i=1}^{N} \frac{(X_i - \bar{X})^2}{N\text{-}1}.$$

In this formula the following happens. For each respondent i we take his or her score X_i and compute how much it deviates from the mean $\bar{X}$: $(X_i - \bar{X})$. We square all these deviations and then add them up (if we not square them first, they would always add to zero). The sum of squared deviations, which is in the numerator, reflects how much people vary around the mean: if everybody had scores really close to the mean (a case with little variability), the sum of the squared deviations would be small. As people's individual scores are further away from the mean, the deviations become larger, the squared deviations do so too, and so does the sum. So, the numerator is small when the scores cluster around the mean, and it gets larger as they are further away from that mean, which reflects that overall they vary more. Next, we divide that sum by the number of respondents minus 1: (N-1). This serves to standardize the measure: if we did not do so, the variability would become higher with sample size, which would be strange.

6.3.3 Graphs for visualizing the variables

As stated, the mean is meaningful only for variables from interval level up. The median can be used for rank-ordered variables, and variables from interval level up. The mode can be used for any kind of variable. If we want to give an indication of the variability in scores, the standard deviation is generally used, but again this is only for variables from interval level up. So, what if we want to give an overview of variables that are at lower than interval levels, such as a nominal level variable?

There are two approaches to describing the distribution of scores for such variables. The first is through simple *frequency tabulation*, such as that produced in any computational software package. See Table 6.2, which gives an example of the distribution over the categories of the variable 'employment status' from a study among 240 volunteers. The first column gives the values of the categories, the second gives the absolute numbers, the third gives the percentages. So we can deduce from Table 6.2 that a total of 240 volunteers took part in the study, that almost two-thirds were employed, and two in nine were students, of whom half had some kind of employment. Almost 10% were retired, a few unemployed and one person did not want to give their employment status.

Table 6.2: Frequency tabulation of employment status

employment status	number	percentage
employed	154	64.2
working student	27	11.3
student	28	11.7
retired	23	9.6
unemployed	7	2.9
other	1	.4
total	240	100.0

A second option for describing nominal data is through the use of graphs. See Figure 6.1, which is an example of a so-called *pie chart* for the same data as in Table 6.2. In a pie chart the data are represented as a pie, and the chunks represent the proportion of cases scoring in each of the respective categories. The data were collected in a study on office misbehaviour (Wesselius et al., 2023). Pie charts are very easily understood by readers, including the non-academically trained, and informative at a glance. Pie charts are often used when the various categories to be represented graphically are categories of a nominal variable.

Graphs have the huge advantage that they are inspected quickly and in general also understood well by readers. In more lengthy, dry texts, they have the added advantage that they in a sense 'break' the text, making the look of the pages more lively, and thus adding to comprehension and retention of the material.

Figure 6.1: Pie chart of employment status

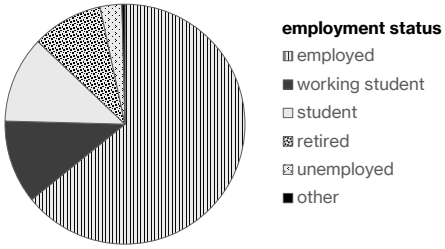
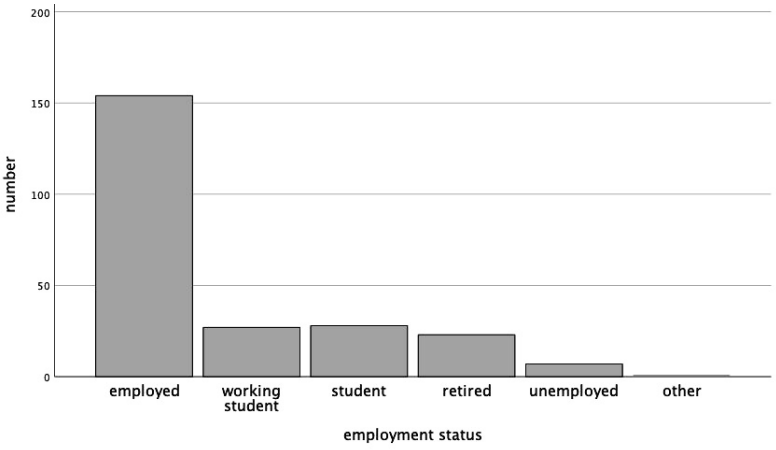


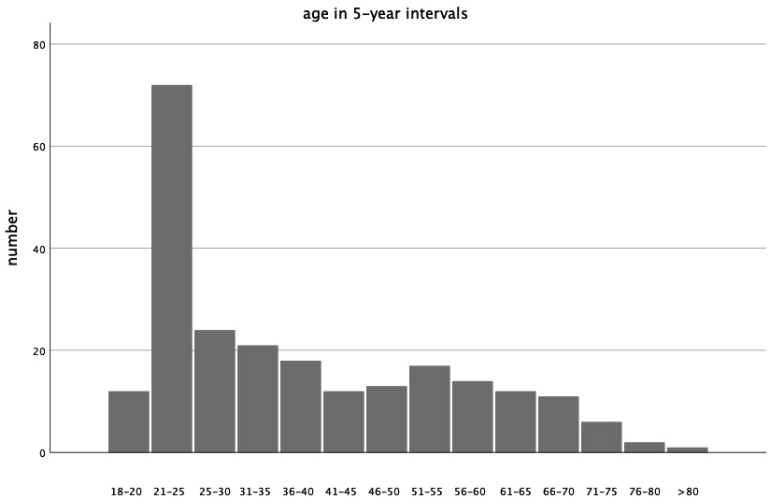
Figure 6.2: Bar chart of employment status



A second way to graphically display these data is by using a bar chart. An example is given in Figure 6.2, in which the categories of employment status have been ordered along the x-axis, and the y-axis gives the number of cases for each category (obviously, the percentage could have been taken just as well). In a bar chart, the categories of the variable along the x-axis are unordered, indicated by the fact that the bars are not connected.

When a variable has been measured at interval measurement level or higher, we generally use means and standard deviations to summarize the variable for the reader. However, also for ordered or continuous variables, graphs would have the advantage that they are easier to inspect and interpret. In addition, they can show how a variable may have a *skewed distribution*, something which means and variances do not show. See Figure 6.3, an example of a so-called *histogram*, that shows the distribution of the age of respondents in the vignette study on office misbehaviour mentioned before (Wesselius et al., 2023). Mean age is 38.24 (sd = 16.62, skewness = 0.738), but the

Figure 6.3: Histogram for distribution of age in 5-year categories



graph is much more informative. Age, categorized into 5-year intervals, is clearly quite skewed because of a large number of respondents in the 21–25 age range, and numbers tailing off towards the right. The graph also shows that there are actually two peaks: in addition to the peak in the 21–35 age range, there is also a peak between 51 and 55 years of age. The data collectors in this research were students, and they likely, as the authors report, sampled persons from their own network as well as from that of their parents.

6.4 Describing the association between variables

So far, we have talked about ways to describe the variables that have been measured. Such descriptions generally entail properties such as the mean or standard deviation, or distribution – either through frequency tabulations or using graphs. But often we want to go further and we want to investigate not only variables one by one, but to look as well at how variables are interrelated. We might want to assess for instance whether older citizens have more trust in the criminal justice system than younger citizens, or whether judges' gender matters in the assessment of sex discrimination claims.

We need tools to summarize such interrelations, just like the tools of mean (or median) and standard deviation for summarizing central tendency and variability. Many of such tools are available. In general, they are called coefficients of association or coefficients of correlation. What tool is appropriate in a given situation depends on two criteria. Firstly, it depends on the *measurement level* of the variables. If we want to report on the association between two continuous variables, a different measure must

be employed than if we want to investigate the association between two nominal variables. Secondly, the choice of measure depends on the *distribution* of the variables. If this is for instance very skewed, different association measures are recommended.

While there is a host of different measures to represent association or correlation, measures are generally interpreted in a similar way. As such, not knowing each and every one of them in detail does not in practice hamper interpretation of their meaning. The class of coefficients for the association between two nominal variables stands a little apart however, so we will treat those in more detail below.

When we investigate the association between two variables, it is often said we are carrying out a *bivariate analysis*: we present the interrelation of two variables. This is different from the univariate analyses we just talked about. There, we may also analyse more than one variable. But then, in each case, we would be doing that for one variable by itself; it is a variable by variable analysis. Below, however, we will each time investigate the variables in conjunction. As such, it is not variable A and variable B we are investigating; rather, it is the relation between variables A and B that is the object of interest.

6.4.1 Correlation coefficients

The most often-used measure for association is the Pearson product moment correlation coefficient. Often, when speaking of 'the correlation', people refer to this correlation coefficient,[3] with the coefficient for the correlation between X and Y written as r_{XY}.

To understand what the correlation coefficient expresses, it is easiest to use scatterplots, that is, plots of the scores of the respondents on the variables X and Y. To make these, we take the score on the X variable as an X-coordinate, and take the score on the Y variable as the Y-coordinate. The scores on the variables X and Y then become points in a two-dimensional space. In Figure 6.4 we give four such *scatterplots*.

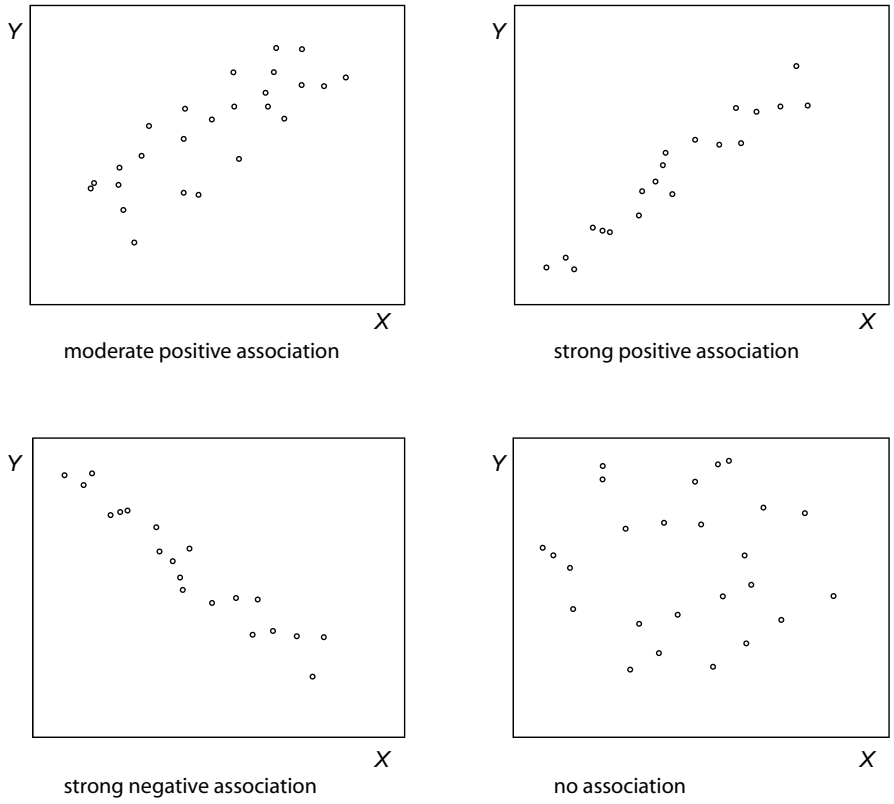
The first scatterplot depicts a situation of moderate to strong positive association. When X is high, Y tends to be high too; when X is low, Y is also generally low. The second (upper right) scatterplot depicts essentially the same situation but now the cloud of points has become much narrower. For this plot as well, if we know that the score on X is high, we know that the score on Y will likely be high too, but we know that now with a much narrower margin: while in the upper left plot, there could still be quite a range in possible values for Y given a certain X, the picture has become much 'sharper' in the upper right one. Knowing X, I know within a fairly narrow band what Y will be. The thick Cuban cigar-shaped cloud in the first picture points therefore to a weaker

[3] In formula the Pearson product moment correlation coefficient is:

$$r_{XY} = \sum_{i=1}^{N} \frac{(X_i - \bar{X})\,(Y_i - \bar{Y})}{(N\text{-}1)s_X s_Y},$$

which is translated in words as: the correlation coefficient is the sum of the product of each respondent's score X_i in deviation from the mean of that variable X and each respondent's score Y_i in deviation from variable Y's mean, divided by the product of the respective standard deviations of X and Y, times a correction factor for sample size. This may not all be immediately obvious to the reader, so we will now first look at the correlation in a more intuitive way, and then investigate whether this formula indeed gives us a quantity that corresponds to our intuitive idea of association.

Figure 6.4: Scatterplots for associations of various direction and strength



association between X and Y. As the cigar becomes slimmer and turns into a ladylike thin cigar or even cigarette, the association becomes stronger. If the cloud of points form a straight line, the association is perfect. The third picture (bottom left) gives a situation of strong negative association – in fact it is the exact mirror of the upper right one. So the association is just as strong, only in the reverse direction: if the score on X is high, the score on Y will likely be low, and vice versa. The last picture gives the situation of almost no association: knowing X does not tell us anything about Y.

Thus, in a situation where people with higher than average scores on X also have higher than average scores on Y and vice versa (those with very low scores on X also have very low scores on Y), we would say that X and Y resemble each other; they are similar. In that case everyone would say, without needing to do any computations, that these two aspects are correlated. And we would expect r_{XY} to turn out to be high. That is exactly what r_{XY} does.[4]

[4]To understand how the formula works, suppose that a particular respondent i's score is much higher than the mean on X: $X_i - \bar{X}$ will then be high. And suppose that that respondent i's score on Y is also much higher than the mean on Y: $Y_i - \bar{Y}$ will then also be high. Simply said: this respondent scores higher than

Now what if we have the opposite situation? That is, people with higher than average scores on X have lower than average scores on Y (and vice versa). Now X and Y are not similar, they are each other's mirror in a way: X is the opposite of Y.[5] In that case, we end up with a negative correlation coefficient. When a correlation coefficient is negative, say r_{XY} = -0.7, the association is just as strong as when it is 0.7; the negative sign indicates that the variables now mirror each other.

The correlation coefficient is a scale- and sample size-independent measure for the association between two numerical variables X and Y.[6] The correlation coefficient varies between -1 and +1. When r_{XY} equals 1, there is perfect correlation; when r_{XY} equals -1, there is also perfect correlation, but the association is negative – the variables are each other's opposite. When r_{XY} equals 0, there is no association. Correlation coefficients between -0.3 and 0.3 we label 'weak'; if $0.3 < |r_{XY}| < 0.7$ the correlation coefficient is judged as 'moderate'; correlation coefficients with an absolute value larger than 0.7 are labelled 'strong'. So, the absolute value of the correlation coefficient tells us something about the strength of the association, while the sign of the association tells us something about the direction (positive or negative). The correlation coefficient is the most widely used measure for association. Many other measures are variations on the structure of r_{XY}.

We take this a little further in Figure 6.5. This graph depicts the situation of moderate association from Figure 6.4, and sketches the idea of prediction. Suppose that I do not have any information about a person's score on Y or on X. Then, if I were asked to predict a person's score on Y, all I could say is that it is between 2 and 18, the total range of the scores. I could also give one number, 11, which is the arithmetic mean of Y. Without any additional information, my best estimate would therefore be the overall mean of Y (see Figure 6.5 along the y-axis). If however, I knew that person's X score, I could make my estimate of Y much more precise! If I knew for instance that X equals 8, then the bandwidth for Y becomes 6–11, and a best estimate would now be 9 (see Figure 6.5). Thus, the fact that X and Y are associated helps me to make predictions

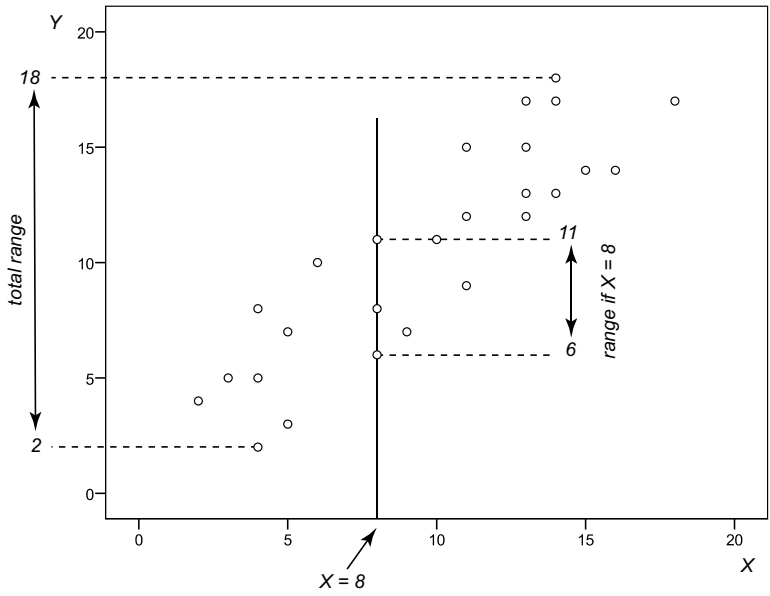
---

average on X and higher than average on Y. The product of these two terms – that is in the denominator of the formula – will then for this respondent also be high. If this applies to all respondents, i.e. that those who score higher than average on X also score higher than average on Y (and vice versa: those who score lower than average on X also score lower than average on Y), we get for each respondent a large product term $(X_i - \bar{X}) \times (Y_i - \bar{Y})$ and summing all these product terms a large value in the denominator, and thus eventually – after the further divisions in the denominator – a large value for r_{XY}.

[5] In that case, in the formula, each case where X_i is higher than the average and $X_i - \bar{X}$ is positive will then couple with the situation that Y_i is lower than the average and $\bar{Y}$ and $Y_i - \bar{Y}$ is negative. Their product will then be negative. Then, the sum in the nominator does not become a large positive number, but a large negative number.

[6] The correlation coefficient has a term (N-1) in the denominator to correct for sample size – if sample size increases or decreases the correlation remains equally strong (why that term is N-1 and not simply N is discussed in more depth in statistics textbooks). Secondly, the denominator contains terms for the standard deviation of X, s_X, and the standard deviation of Y, s_Y. Why this is necessary is best understood as follows. Suppose that we are investigating whether shoe size X and height in metres Y are associated. Suppose that we have computed a correlation coefficient and that that coefficient equals 0.95, meaning that shoe size and height are correlated positively and (very) strongly. Suppose now that we decide to measure height not in metres, but in centimetres, or inches. Intuitively, this should not change the association between X and Y. However, if we change the values for Y – moving from metres to centimetres – they would be multiplied by a factor 100, and the r_{XY} would also become 100 times as high! Dividing the denominator by the standard deviations solves for such undesirabilities.

Figure 6.5: Relation between association and prediction

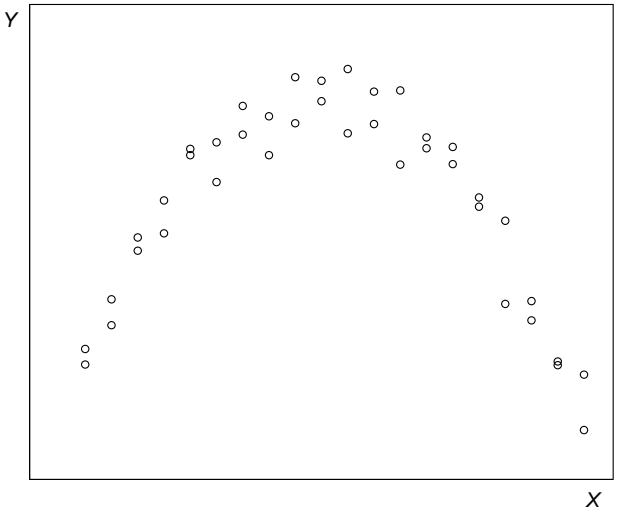


about one variable, using information from the other. Knowing X tells you something about Y.

Now, we said above that a cigar-shaped cloud like we have in Figure 6.5 is indicative of moderate to strong association, and that as the cigar becomes slimmer the association becomes stronger. If that is the case, then we would expect our prediction to improve as an association gets stronger. Imagine, therefore, in Figure 6.5 a similarly positioned but slimmer cigar. Again, suppose that I know that X equals 8. Now, however, because the cigar is slimmer, the possible range for Y is narrower: it is possibly 8–10 instead of the wider 6–11 that we had. If I were required to give one number again for Y, again I would estimate Y at 9. So, a stronger correlation between X and Y does not lead to a different estimate for Y on the basis of X, but it leads to a *more certain* estimate, reflected in the smaller bandwidth of the possible values for Y.

There are two important issues to heed when using the Pearson product moment correlation coefficient. The first is that the correlation coefficient is a measure for *linear* association. The Cuban and ladies' cigars we drew were all clouds around a *straight* line. But what if we have the situation as depicted in Figure 6.6? In this case there is a – strong – *non-linear* association: if we know X we can predict within a fairly narrow bandwidth what Y is. However, if we were to compute the correlation coefficient for these data, it would turn out to be almost zero. That is not a wrong result, as the Pearson product moment correlation coefficient is a measure of *linear* association – and indeed there is no linear association here. However, from that we cannot conclude that there is no association at all. Other measures would be needed to

Figure 6.6: Non-linear association between X and Y

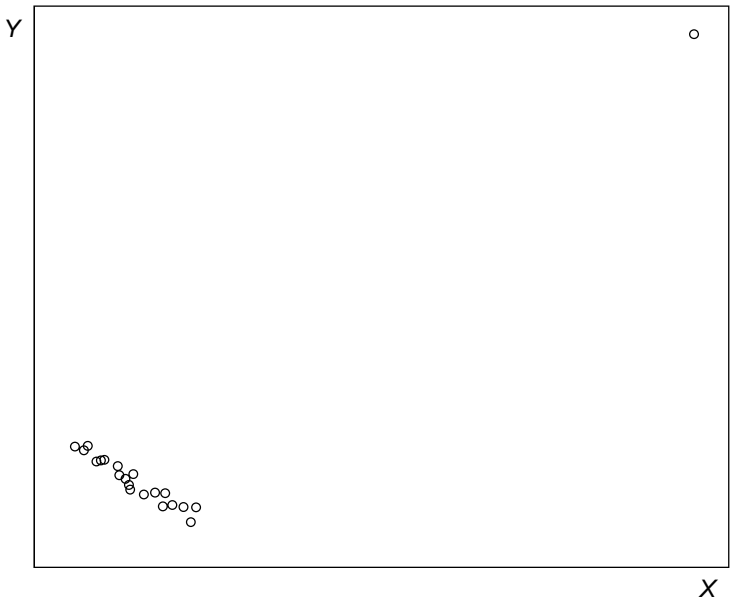


capture the non-linear association. Several such measures (the best known is probably η) are discussed in standard statistics textbooks. For now, it is sufficient to know that the ordinary correlation coefficient is useful for assessing linear association between continuous variables only.

Eyeballing the graphs is important for the second issue too. The correlation coefficient is namely sensitive to so-called *outliers*. See Figure 6.7, in which an example is given of a sample in which the scores on X and Y are negatively associated: a cloud of points in the lower left corner forms a ladies' cigar, tilted to the left, indicative of a (fairly strong) negative correlation. However, because there is one person who combines a huge X with a huge Y, the correlation coefficient if we were to simply compute it would come up strongly positive here. That correlation coefficient is, however, based on just one out of N sample members, so not a result one would want to report like that. Inspecting the scatterplots therefore also prevents mistakes one could make if one were to simply – blindly – compute the correlation coefficient. Outliers like the one depicted here are generally removed from quantitative analysis: they are considered an anomaly that distort our view of the general picture.

Most researchers compute their correlation coefficients with a software package or a calculator. Understanding the formula for the coefficient is noble, but not crucial. What is however crucial is to understand the scatterplots we discussed and to realize how much can be gleaned from them. Anyone – including the maths-averse – can read from a simple scatterplot whether there is a linear or a non-linear association between X and Y. Also, anyone can see whether the association is positive or negative. Lastly, anyone can get at least an idea of the strength of the association. It is only when we want to have an exact number that we need to resort to computations. All the other

Figure 6.7: Outlier



much more relevant and much more interesting stuff we can all do by simply eyeballing the graphs.

6.4.2 Other measures of association

As we said earlier, the ordinary Pearson product moment correlation coefficient is but one measure of association: for the linear association between two continuous variables that – we hinted at this but did not make it explicit – are distributed pretty 'normally'. The moment one of the variables is not interval level or higher, if the association is not linear, or if one of the two variables has a 'funny' distribution (like a skewed distribution), we must resort to other statistical measures. This is important to know.

We will however not discuss all those measures: mainly because it is all pretty tedious and not really necessary for our purposes, as most coefficients are interpreted along similar lines. The reader who needs to decide which coefficient to use can look up the various options in a textbook, for instance Hinkle, Wiersma, & Jurs (2003). We only briefly mention two other fairly commonly used correlation coefficients.

The first of these is Spearman's ρ (or 'rho') for variables X and Y both measured at ordinal level. Spearman's ρ varies – just like the ordinary correlation coefficient – between -1 and 1, and the interpretation given to these numbers is also identical.

The second is Kendall's τ (or 'tau'), which is a measure for the association between two ordinal variables. It is less often used for ordinal variables than Spearman's ρ. It

Table 6.3: Observed frequencies of sentence type and gender

	imprisonment	community service order	
males	40	60	100
females	10	90	100
	50	150	200

Table 6.4: Expected frequencies if sentence type and gender were unassociated

	imprisonment	community service order	
males	25	75	100
females	25	75	100
	50	150	200

is however regularly used when the variables are interval level or higher, but their distribution is skewed. The value of 'tau' in that case tends to be somewhat lower than that of the ordinary correlation coefficient, because it takes into account only ranking information of the scores.

When we are interested in the association between two nominal or categorical variables, numerous measures are available. Each has weaknesses and strengths. In research practice, however, two are most often used: the χ^2 (or chi-square) and the odds ratio. The first is not really a measure of association (although it is often interpreted as such); it is a statistical test, which does not actually tell us how strong the association between the two nominal variables is, but rather how likely our data is to be found if the variables were unassociated. The second tells us something about the strength of the association between two dichotomous variables (i.e. variables with two categories each only). It expresses this association in terms of risk. We will first discuss the chi-square (or χ^2), which is very often used. The examples that we use are, for ease of explaining, simple two-way tabulations; for the chi-square, all computations can be simply extended to variables with more categories.

Suppose therefore that we have measured two nominal variables for a number of respondents, and that our findings are as given in Table 6.3. In this table we see that a total of 200 respondents have been investigated: the sample consisted of 100 males and 100 females. Of these 200 respondents, a quarter, or 50, were sentenced to imprisonment, and 150 received a community service order. Now the question is: are the two variables we are investigating here associated? Are gender and type of sentence associated? In other words, does it make a difference for sentencing whether a respondent is male or female? We can see without any computational fuss that this is the case. Of the women, only one in ten was sentenced to imprisonment, while for the males this was 40%. Clearly, gender matters. But how strong is the association? Can we express that in a number, a statistic?

For that, the observed frequencies (Table 6.3) are compared with the frequencies we would expect to find if the two variables were not associated. In Table 6.4 these are given: we see that if gender played no role in sentencing we would expect to find equal proportions of men and women sentenced to imprisonment (25) and to a community service order (75). Formulated differently, with the data as in Table 6.4, knowing what someone's gender is does not give you any information on the type of sentence they likely received. For computing the chi-square, the differences between the observed and expected frequencies are used.[7]

As the chi-square is computed as a sum of squared values, its lowest possible value is zero: a value of zero indicates that the frequencies are exactly as expected under the null hypothesis (see chapter 8) of no association. We then conclude that the association between the two variables is insignificant. The higher the chi-square that is computed, the more likely it is that we can conclude that the two investigated variables are associated. Almost all software packages that compute the chi-square also generate the significance level. At this point we reiterate that the chi-square can be computed for nominal variables with more than two categories (for instructive purposes we used only dichotomous variables here). Attention should be paid to whether the cells of the cross-tabulated variables are adequately filled: if the sample size is small or too many expected frequencies are low, the test does not perform well and other tests may be recommended (such as Fisher's exact test).

In our example, the chi-square value equals 24, which is highly significant ($p < .001$). The significance of the chi-square is in research practice often equated with the strength of the association between the two investigated variables. In that sense, formulations of the χ^2 are often slightly sloppy.[8] While this may seem nit-picking, it should be noted – as we will discuss in chapter 8 – that significance also depends on sample size. So, if we have a very large sample, a tiny value for the chi-square test may also emerge as significant even though the association between the two nominal variables may be weak.

Various measures are available that do give an indication of the strength of association between nominal variables – just like the correlation coefficient does for continuous variables. Probably the best of these, under some conditions, is the so-called *Cramer's V*, which gives us a number between 0 and 1, with 0 reflecting no association, and 1 perfect association (meaning that from the score on one nominal variable,

[7] The chi-square is computed as:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - Ei)^2}{E_i},$$

with k the total number of cells, O_i the observed frequency in the i-th cell of the cross-tabulation, and E_i the frequency we would have expected in that cell if – in this example – gender and sentence type were unassociated, as given in Table 6.4. The chi-square then becomes:

$$\frac{(40–25)^2}{25} + \frac{(60–75)^2}{75} + \frac{(10–25)^2}{25} + \frac{(90–75)^2}{75} = 24.$$

[8] A second manner in which wording of the test is generally sloppy, is that most researchers speak of the value of the χ^2, while the statistic computed is written with a Latin letter X as X^2, and in the hypothesis the Greek letter χ is employed.

Table 6.5: Observed frequencies of sentence type and gender

	imprisonment	community service order	
males	40	60	100
females	10	90	100
	50	150	200

the score on the other nominal variable can be predicted perfectly). Other measures are also available, such as the ϕ-coefficient for two-by-two tables, or λ.

However, whichever measure one uses, it is important to realize that we should always inspect the cross-tabulation to understand what the substantive interpretation should be. As the correlation coefficient deals with continuous variables only, it is easy to interpret: if the coefficient is high and positive, high scores on X imply high scores on Y. For nominal variables, the interpretation is not immediately obvious. In our sentencing and gender example, the χ^2 tells us that the two are associated but it does not tell us whether that means that being a man increases the likelihood of imprisonment, or whether it is being female that increases the likelihood of imprisonment. So, we always need to refer back to a cross-tabulation to be able to interpret the association between the two nominal variables. We refer the reader to a statistics textbook such as for instance Hinkle, Wiersma, & Jurs (2003) for more information.

For dichotomous variables, another very often used measure to inspect association is the *odds ratio*. The odds ratio is not easy conceptually, but very easy in use and interpretation. The reasoning behind it is best explained with an example. Suppose that we have the same data we had before. See Table 6.5, which is repeated here for inspection purposes but is the same as Table 6.3.

The reasoning of the odds ratio is as follows. In the example, we have data on gender and sentence type. For females the probability of receiving a prison sentence is 0.1, and the probability of a community service order 0.9. This means that for females the probability of being sentenced to a community service order is 9 times that of a prison sentence. For males, we can apply the same reasoning and then we find that the probability of receiving a community service order is 1.5 times that of a prison sentence. The odds ratio now is the ratio of these two ratios or odds,[9] which equals 6.

How can we interpret this? The literal interpretation is that the ratio of imprisonment to community service order is for males 6 times that for females. This is hard to master conceptually, and a jumble of words. In practice the two terms in the nominator and denominator are perceived of and formulated as 'risk'. The odds ratio is conceptualized as a ratio of risks (or indeed 'odds'). In practice, interpretation is unproblematic with formulations such as: 'the risk of imprisonment for males is 6 times the risk for females'.

[9]
$$OR = \frac{\frac{\text{male imprisonment}}{\text{male community service order}}}{\frac{\text{female imprisonment}}{\text{community service order}}} = \frac{\frac{40}{60}}{\frac{10}{90}} = \frac{\frac{1}{1.5}}{\frac{1}{9}} = 6.00$$

Note firstly that a risk is not a chance: risks can be larger than 1 while probabilities are always between 0 and 1. Secondly, it is important to note that if there were no difference in risk between males and females, the odds ratio would be 1 (not 0!). Compare again the data in Table 6.4. This table reflects the situation where there is no relation between gender and sentence type – in other words, for the scores on sentence type it does not make any difference whether someone is male or female. Thus, we would also expect no difference in the odds. And if we wanted to compare the risk for men and women for this dataset, we would indeed arrive at an odds ratio of 1.[10]

So, an odds ratio equal to 1 means no difference in risk, and an odds ratio larger than one means an increased risk. Obviously, if we had in our example in Table 6.5 reversed men and women in the equation, we would have arrived at an odds ratio of 0.1667,[11] and we would have reported that females have a 6 times *lower* risk of imprisonment. So, with the odds ratio, it is important to keep in mind how this risk is specified (i.e. what is in the nominator and what is in the denominator). Also, because of this, an odds ratio of 2 (twice the risk) reflects just as big a difference in risk as an odds ratio of 0.5 (half the risk).

Now what is generally perceived as an increased risk, or as a not-so-strongly increased risk? Just like for the correlation coefficient, here also there are rules of thumb. In general, an odds ratio larger than 2 or smaller than 0.5 is judged to constitute a sizeable risk increase or decrease. This is a rule of thumb though, and for some research situations an odds ratio of 1.3 may already be judged to constitute a relevant increase in risk. It is also possible to test whether an odds ratio differs significantly from 1.

In an exploratory study, Tollenaar (2018) investigated differences between culpability as assessed by municipalities and as assessed by administrative courts when deciding on administrative fines in the Netherlands. Data were collected from the website `rechtspraak.nl`, the official website of the Dutch judiciary, using the search terms *bestuurlijke boete* (administrative fine) and *verwijtbaarheid* (culpability) for cases dealt with by the Centrale Raad voor Beroep (Central Appeal Tribunal) from late 2014 to mid-2017. This led to a usable sample of 125 cases in which an administrative body – municipal or judiciary – had imposed a fine and an administrative court had later judged culpability in the case.

Tollenaar first described how many citizens had defended themselves by stating that they were not, or only to a limited degree, culpable. Second, he found that many administrative authorities had assumed that the transgression was intentional (28% of cases). Third, Tollenaar notes that administrative courts often judged the behaviour to be non-intentional, but instead constituting normal culpability (70%), or reduced or in fact no culpability (16%); the percentages for the municipal authorities had been 50% and 9% respectively. The data are in Table 6.6.

10
$$OR = \frac{\frac{\text{male imprisonment}}{\text{male community service order}}}{\frac{\text{female imprisonment}}{\text{female community service order}}} = \frac{\frac{25}{75}}{\frac{25}{75}} = 1.00$$

11
$$OR = \frac{\frac{\text{female imprisonment}}{\text{female community service order}}}{\frac{\text{male imprisonment}}{\text{male community service order}}} = \frac{\frac{10}{90}}{\frac{40}{60}} = 0.166667$$

Table 6.6: Positions of the administrative authorities and court judgment

Degree of culpability	Position of administrative authority	Judgment of administrative court
Intentional	28%	6%
Gross negligence	11%	7%
Normal culpability	51%	70%
Reduced culpability	8%	14%
Absence of culpability	1%	2%

Source: Tollenaar (2018)

Next, Tollenaar sets out to explain this difference. For this, he compares the judgments of the administrative authority against those of the courts on a case-by-case basis. He notes that in 62% of cases, the court and administrative authority judge identically. In 38% of cases they judged differently, and the court always found lower culpability.

Tollenaar identifies two possible explanations. The first is a temporal one: he states that it may have taken the administrative authority a while to absorb new rules that took effect from late 2014. The data do not support this explanation, however. It is not the case that differences were larger after the new rules took effect. A second explanation is that smaller municipalities may have lacked the knowledge and expertise to apply previous judgments of the Central Appeal Tribunal. Tollenaar states that this is a much more likely explanation, as the Association of Netherlands Municipalities, which normally supports municipalities with the implementation of new regulations, only issued a position paper in April 2017, well after the new rules came into effect.

Additional analyses lend further support to this explanation. Disaggregating the cases by whether they had been dealt with in first instance by a municipal administrative body or by a national administrative agency, it became clear that decisions made by the national agency had much more often been upheld by the courts (88%) and rarely altered (11%). For the municipalities this pattern was reversed: a majority of cases were altered by the courts. See Table 6.7.

The odds ratio equals 8.83 (computed as the odds of confirmation versus alteration for national agencies, divided by the odds of confirmation versus alteration for municipalities), showing how indeed the risk that the court confirms is very much increased in national agencies as compared to municipalities.

These results do not prove that it is municipalities' lack of expertise that causes the difference. Tollenaar argues that municipalities and national agencies typically deal with different, likely not well-comparable cases: municipalities decide mainly in social benefits cases, while national agencies decide in employee insurance schemes. Second, municipalities must pay social benefits out of their municipal budgets, which may be an incentive to educate offenders about the need to observe the relevant provisions.

Last, we note that the sample of cases does not constitute a representative sample of all cases in which administrative fines have been imposed, as not all citizens who are

Table 6.7: Type of administrative authority and court judgment

	National agency	Municipality
Court confirms	88%	48%
Court alters	11%	53%
$\chi^2 = 19.79$, $p < .001$		

Source: Tollenaar (2018)

handed such a fine will object and take their case to court. It also does not constitute a representative sample of all cases that were decided by the courts.

6.4.3 Graphs for describing the association between variables

What we said in section 6.3.3 about the ease of visual representation for simple descriptives – such as the distribution of a variable – also applies when it comes to investigating the association between variables. Here also, graphs are an efficient tool to make the reader see and interpret association. Given however that association is investigated, graphs are bound to be more complicated. We already showed a number of scatterplots in section 6.4.1, and stressed how important it is not to rely on computations only, but to actually inspect the data. Doing so will reveal non-linear associations and outliers. This means that the use of graphs is not only pleasant for readers, but supports the analysis as well.

Another way to graphically display association is by using a special kind of bar chart, a so-called *stacked bar chart*. See Figure 6.8, in which we show several things, with the data again stemming from the study on office misbehaviour. Firstly, we show how many respondents did and did not personally know someone who had experienced office misbehaviour. The graph shows that approximately 100 did, and 140 did not. Next, the graph also shows that the majority of those who knew someone who had experienced office misbehaviour was female, while for those who did not know such a person the distribution was approximately even. There is all in all a clear association between the two variables.

When a variable is at interval measurement level or higher, we can also visualize association. Using a so-called *clustered histogram*, we can inspect whether the distribution of one variable differs by the categories of another variable. See Figure 6.9, in which we depict scores on a dependent variable, disaggregated by age category of the respondents.

Figure 6.9 breaks up the distribution of age into two groups: one below 40, and one above 40. We see that the distribution of agreement is mirrored: the distribution in the younger group is right skewed and in the older group it is left skewed. Again, this pattern is much easier to grasp and understand for readers than the same data in a table would be.

Figure 6.8: Example of stacked bar chart for type of abuse and recidivism

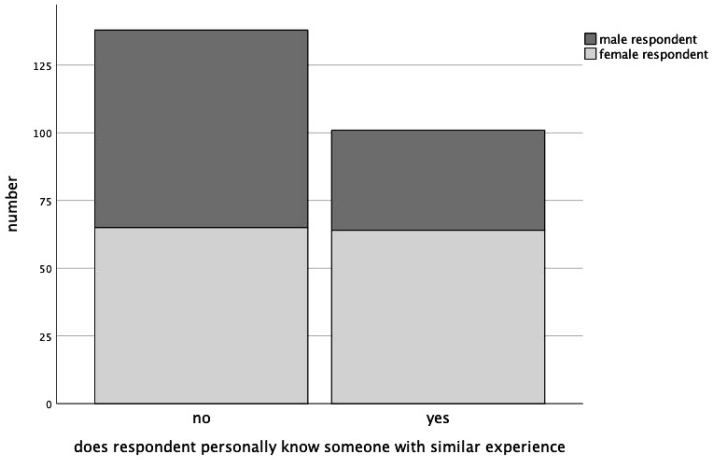
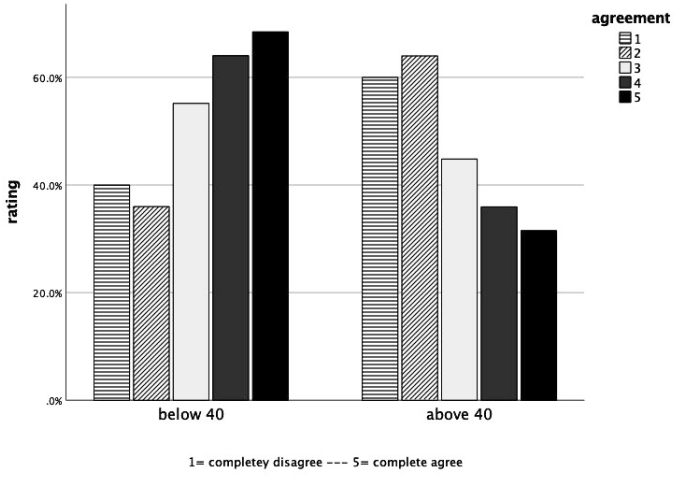


Figure 6.9: Clustered histogram for distribution of ratings by age group



6.5 Duration variables: survival analysis

In some types of research, we encounter special variables. The properties of these variables are such that we cannot do regular computations with them. One example of such a special variable we encounter regularly in empirical legal research is a so-called

duration variable or a *time-to-event variable*. This may sound fairly obscure, so let us illustrate what we mean through an example.

Suppose we want to study recidivism in released prisoners. We want to assess how long it takes, on average, for prisoners to recidivate in the sense that they are re-sentenced to imprisonment and return to prison. Now let us assume that we have a sample of 10 prisoners who were released on various days over the course of one year. The day that a prisoner is released we count as day zero, and from that day onwards we start counting. For a prisoner who is reconvicted and is re-imprisoned after 30 days, we set the variable 'recidivism' we are interested in at 30; another prisoner, back in prison after 90 days, gets a value of 90 for the variable 'recidivism'. If we were to calculate the average time to recidivism for these two prisoners, we would arrive at a value of 60 days. And if we have another two prisoners who are reconvicted and returned to prison, one after 120 days and the other after 320 days, the mean time to recidivism for all four would be 140 days. So far, so good, it appears.

But what are we to do with six other prisoners who did not recidivate within our observation window of one year? If we report that average time to recidivism is 140 days, we are well off the mark; we do not reckon with the fact that there are six others who did not recidivate for all those 365 days we observed them... The average of 140 days is not a good summary of what happened. Should we then set the recidivism time for these six at 366 days? This is also not a good solution as that is not what happened; it is in fact quite likely that some will not recidivate at all, or possibly only after a very long time. And suppose that one prisoner dies after 200 days, not having recidivated. Do we remove this prisoner from our dataset? That would also be distorting as then we are omitting someone who did not recidivate and are then left with a more 'serious' subsample.

Survival analysis is a technique that can deal with these issues. It is a method that is very often used in epidemiological research, where for instance survival after treatment for a certain disease is studied. Survival analysis is encountered under different names, such as *event history analysis* or *failure time analysis*. The dependent variable in survival analysis is survival time (in epidemiology real physical survival, so time until death) or time until a certain event takes place. That event can be recidivism or time until a court case is dispositioned. Survival analysis is also used in other disciplines, such as in demography, where it can be used to study survival but also time until divorce.

Survival analysis has a method to deal with the specific properties of the time-to-event variable. A first property is that the values of this variable can only be positive (prisoners cannot return to prison before they were released). This by itself does not generate massive issues, as it is not problematic to compute averages for variables with positive values only. What is problematic though is the phenomenon we just described. As we conduct our study over only a certain stretch of time, we do not have values on the duration variable for those who do not recidivate before the end of our observation period. While we know that they did not experience the event we are interested in, we have no way we can attach a value to their duration variable.

The duration variable is for these persons, as it is called, *censored*. It may be that they do recidivate at some point after our study has ended, but we do not know whether that is the case, or when. This phenomenon is also referred to as *censoring*. Survival

Table 6.8: Months, recidivism and censoring, number of persons *at risk*, event chance, survival chance and cumulative survival chances

month	event	*at risk*	P_t(event)	P_t(survival)	S_t
1	-	5	0.00	1.00	1.00
2	recidivism	5	0.20	0.80	0.80
3	-	4	0.00	1.00	0.80
4	censoring	4	0.00	1.00	0.80
5	recidivism	3	0.33	0.67	0.54
6	recidivism	2	0.50	0.50	0.27
7	-	1	0.00	1.00	0.27

analysis can deal with this censoring. What it does is to incorporate the measures for all those who have not recidivated into the analysis, until the point where we are unable to observe them anymore, either because our window of observation ends, or because they have died or emigrated or moved out of our sight in some other way. This may sound pretty abstract, so we will demonstrate survival analysis using a small example.

In its simplest form, survival analysis computes a function of survival against time. The curve of this function (the 'survival function') shows, for every observation point in time, the percentage of respondents that has survived until then, that is, that has not experienced the event we are interested in. As said, our respondents can be people and the event can be criminal recidivism, or disease, or employment after obtaining a law degree. But our 'respondents' can also be court cases or tort claims, and then we would be interested to see how soon these are dispositioned or settled in some way.

Let us construct a very small example of five fictitious persons who were sentenced to community service. We want to analyse how soon they recidivate. The day community service is completed is day zero, after which the clock 'starts ticking'. We observe these five persons for seven months. At the start of the first month, nobody has recidivated yet. All five respondents are, as we say, *at risk* of experiencing the event. Suppose that nobody recidivates in the first month. The chance of survival is then 1, and the chance of recidivism 0. See the first row of Table 6.8.

At the beginning of the second month, all five respondents are therefore still at risk; each of them can still undergo the event. Now suppose that in month 2, one person indeed recidivates. This respondent then has not 'survived'. The average survival chance in month 2 P_t(survival) is then equal to 0.8 (and the chance of undergoing the event P_t(event) is therefore equal to 0.2). At the beginning of month 3, we are therefore left with four persons who are still at risk of experiencing the event; see the third column of Table 6.8. The last column of this table gives the so-called cumulative survival probability S_t, the proportion of respondents who have survived until time point t. S_t is calculated as $S_{t-1} \times P_t$(survival). One of the assumptions of survival analysis is that the survival chances are independent across the time intervals, so that we may multiply them, to end with a proportion of survivors.

Now suppose that in month 3 no one recidivates. The chance of undergoing the event is then 0, and the survival chance 1. The cumulative survival probability remains

at 0.80. In month 4, a respondent emigrates; this means that we will not have information on this respondent's recidivism for the coming months, hence we now have censoring. At the end of month 4, the cumulative survival probability is still 0.80. However, at the beginning of month 5, we have only three respondents who are still at risk. While the emigrated respondent may undergo the event, we have no way of registering this as the respondent has disappeared from our sight.

In month 5, a second respondent recidivates. The chance that the event occurs is in this month P_5(event) equal to 0.33: one out of three respondents at risk has recidivated. The survival probability P_5(survival) is therefore equal to 0.67. The cumulative survival probability S_5 now equals $S_4 \times P_5(\text{survival}) = 0.80 \times 0.67 = 0.536$. Note how the cumulative survival probability incorporates the effect of censoring: even though two out of five respondents have recidivated, we do not calculate the cumulative survival as 0.60 but arrive at a lower figure of 0.54, because we now take into account that one respondent has 'disappeared' from our calculations. See Table 6.8.

At the beginning of month 6, only two persons are still at risk of undergoing the event. If in that month one person recidivates, the probability of undergoing the event is in that month 0.50, the survival chance is the same, and the cumulative survival probability is 0.27. At the beginning of month 7, only one respondent is still at risk.

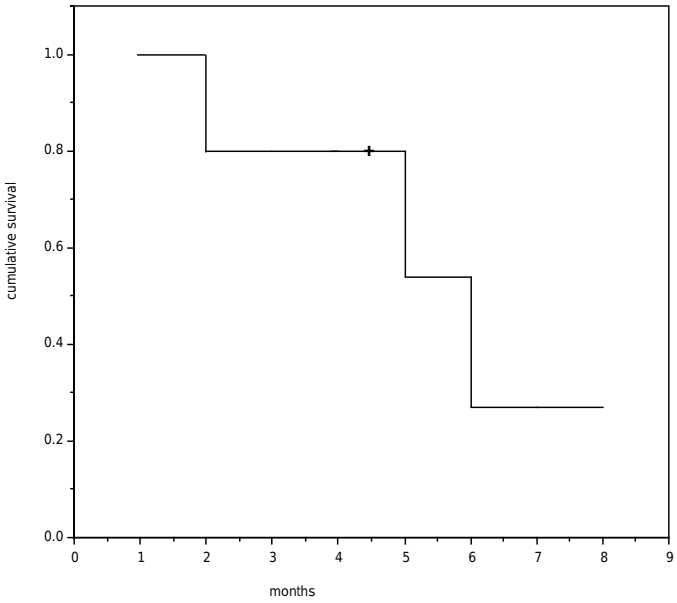
The cumulative survival probabilities constitute the survival function. Figure 6.10 shows the survival function from the example we just discussed. As we are dealing with a small number of respondents, the curve is quite jagged; if we analyse larger samples, the survival curve will be smoother. This curve is also called the *Kaplan–Meier curve* (Kaplan & Meier, 1958). It is customary to indicate with a small cross or dot when censoring took place (see the little cross for our emigrated person in month 4 in Figure 6.10).

Using survival analysis, it is also possible to compare the survival curves of one or more groups. Suppose for instance we have two groups of former prisoners, one that had a normal incarceration regime, and another group that was administered a special training to facilitate reintegration. We would expect those who had undergone the training to recidivate less. We would expect their survival curve to differ from the survival curve of the regular prisoners. If that is the case, we may conclude that there is an association between the reintegration training and recidivism: those with the training have different patterns than those without the training.

Bruns, Pullmann, Wiggins, & Watterson (2011) investigated the King County Family Treatment Court (KCFTC), an effort to address the special needs of families involved in the child welfare system due to child abuse and neglect charges related to parental substance abuse. Family treatment courts (FTC) were modelled on drug treatment courts, but differ in their aims. While drug treatment courts aim to keep offenders free from the influence of drugs and alcohol so that they do not have repeat involvement with the criminal justice system, FTCs aim to strengthen family bonds, and promote children's health and safety. The child welfare system is involved in the judicial process, and court-appointed special advocates may be enrolled. The authors studied a host of outcomes for one of these courts, the King County Family Treatment Court.

For this, they collected data on a total of 258 parents, of whom 76 were KCFTC participants, and 182 were parents in principle eligible for KCFTC but were not admitted for various reasons. This means that we have a quasi-experimental design, with

Figure 6.10: Kaplan–Meier survival curve

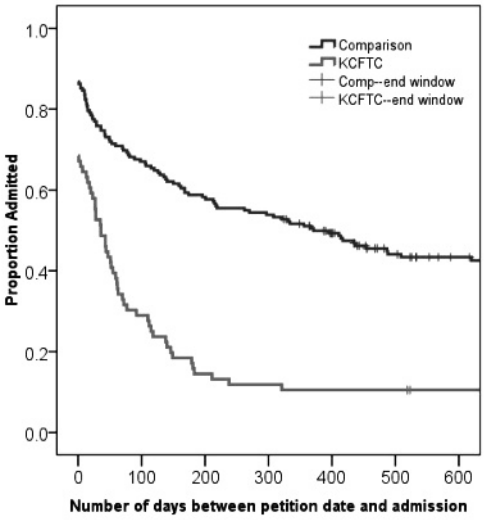


some doubt on whether the KCFTC and non-KCFTC parents were comparable (in fact, the groups were indeed not comparable as the KCFTC parents had fewer risk factors). The authors investigated a large number of questions, of which we will focus on two, namely whether – as compared to the control group – KCFTC parents were admitted to treatment more quickly than the comparison group, and whether the involvement in the child welfare system of children of KCFTC participants ended sooner.

For the first question, the authors conducted a Kaplan–Meier survival analysis. The dependent variable is here 'time until admission to treatment'. The survival curves are depicted in Figure 6.11, with the top curve representing the comparison group and the lower curve the KCFTC group.

If we inspect the figure, we see immediately that the survival curve for the KCFTC parents is always (much) lower than that of the comparison group. For instance, after 100 days, we see that only approximately 30% of KCFTC parents has survived, that is, have not undergone the event. This means that after 100 days or about 3 months, 70% of the KCFTC parents are enrolled in treatment. The comparable figure for the control parents is 70%, meaning that only 30% of control parents are in treatment after 100 days. This difference remains, and even increases. For instance, after 200 days, a little over 15% of KCFTC parents have not experienced the event, so that 85% are enrolled in treatment; of the control parents, a little over 40% are by then enrolled in treatment. The figure shows two more interesting things. First, we see how enrolment in treatment stagnates for both groups. After about a year we see no real improvement anymore: within the KCFTC parents about 10% do not enrol anymore, and within the control

Figure 6.11: Kaplan–Meier survival analysis of days until treatment entry for all parents



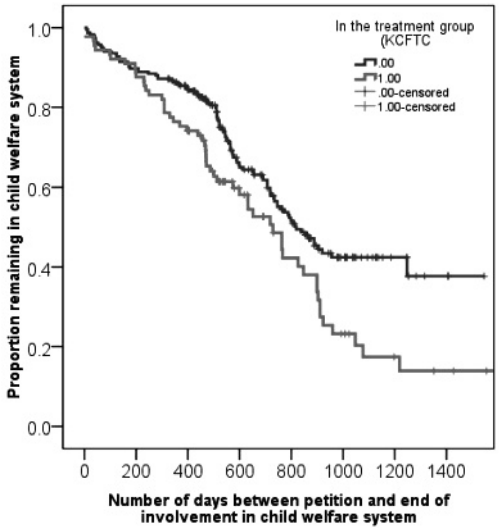
Source: Bruns et al. (2011)

parents this is a little under 50% – a huge difference. Second, we see how from the start KCFTC parents were already doing better (underlining that the two groups were not randomly chosen). On the day the 'clock started ticking' for the survival analysis (the day a so-called 'index petition' was handed in), a little over 10% of control parents were already following treatment. KCFTC were from day 1 already faring better: 30% were already enrolled.

For the second question, the authors also conducted a Kaplan–Meier survival analysis, with as the dependent variable now 'time until child is no longer involved in child welfare system'. The survival curves are depicted in Figure 6.12, with the top curve again representing the comparison group and the lower curve the KCFTC group.

Here, the interpretation is identical. After developing similarly to the comparison children for the first 200 days, their survival curves start descending faster, meaning that more of them leave the child welfare system. While the differences are less marked than in the previous picture, the KCFTC children are also here doing consistently better. After 500 days, 40% of KCFTC children have left the child welfare system, while 75% of control children are still in the child welfare system. In this figure, we see a similar 'stagnation', which turns out to be much less favourable for the control children. Towards the end of the evaluation period, it appears that 40% remain 'stuck' in the sense that they seem unable to leave the child welfare system, but for a few who did so

Figure 6.12: Kaplan–Meier survival analysis of length of time until child's end of involvement in the child welfare system



Source: Bruns et al. (2011)

eventually. This applies to a much smaller percentage of KCFTC children, just a little over 15%. All in all, a – roughly speaking – 25% difference between the two groups.

Do these results mean that we can be sure that these differences are attributable to the KCFT Court? Strictly speaking not, as parents were not allocated to the Court randomly. As the figures also showed and as other tables in the report indicate, there is reason to suspect that the KCFTC parents and children were doing better from the outset. To eliminate such *confounding*, the authors performed a number of supplementary statistical analyses. The findings from these analyses also pointed in the direction of the FTC special courts indeed having a beneficial effect.

Numerous extensions of survival analysis exist. We showed only examples where the survival curves of two groups are compared (for instance KCFTC and non-KCFTC). Other comparisons are possible too; it is for instance also possible to investigate not just the occurrence (or not) of one event, but the occurrence of several events (for instance settlement, withdrawal, judgment). An accessible introduction to survival analysis is Allison (1984); a more elaborate classic is Kalbfleisch & Prentice (2002).

6.6 Wrapping up

In this chapter, we introduced concepts such as respondents, variables and measurement levels. We gave a number of statistics to summarize variables, such as the mean for central tendency and measures for variability. After such univariate statistics, we discussed bivariate measures, such as the correlation coefficient for numerical variables, the χ^2 for nominal variables and the odds ratio for dichotomous variables. We also illustrated how to use graphs to visualize such bivariate associations. We discussed a special kind of analysis, survival analysis, for a dependent variable that is a time-to-event variable. Even though we each time gave the formulas in footnotes for the measures we discussed, we also stressed that these formulas are not necessary to understand how the statistics work, nor needed if one has a statistical software package at hand (even Excel will do quite a few statistics).

Chapter questions

1. What are the five measurement levels? (section 6.2.1)
2. Describe in what ways they are ordered (section 6.2.1)
3. Argue why outliers are particularly problematic for measures of association and central tendency for interval level variables (section 6.3)
4. List the measures of association that are available for nominal, ordinal and interval level variables (section 6.4)
5. Under what conditions can the Pearson product moment correlation coefficient be used? (section 6.4)
6. What is meant by censoring in duration variables? What complications does it cause? (section 6.5)
7. Argue whether regular correlation coefficients could be used if a duration variable had no censoring (section 6.5)