



The beast of bias

5



FIGURE 5.1
My first failed career choice was a soccer star

5.1. What will this chapter tell me? ①

Like many young boys in the UK my first career choice was to become a soccer star. My granddad (Harry) had been something of a local soccer hero in his day, and I wanted nothing more than to be like him. Harry had a huge influence on me: he had been a goalkeeper, and consequently I became a goalkeeper too. This decision, as it turned out, wasn't a great one because I was a bit short for my age, which meant that I never got picked to play in goal for my school. Instead, a taller boy was always chosen. I was technically a better goalkeeper than the other boy, but the trouble was that the opposition could just lob the ball over my head (so, technique aside, I was a worse goalkeeper). Instead, I typically got played at left back ('left back in the changing room' as the joke used to go) because, despite being right footed, I could kick with my left one too. The trouble was, having spent years trying to emulate my granddad's goal-keeping skills, I didn't really have a clue what a left back was supposed to do.¹ Consequently,

¹ In the 1970s at primary school, no one actually bothered to teach you anything about how to play soccer; they just shoved 11 boys onto a pitch and hoped for the best.





I didn't exactly shine in the role, and that put an end for many years to my belief that I could play soccer. This example shows that a highly influential thing (like your granddad) can bias the conclusions you come to and that this can lead to quite dramatic consequences. The same thing happens in data analysis: sources of influence and bias lurk within the data, and unless we identify and correct for them we'll end up becoming goalkeepers despite being too short. Or something like that.

5.2. What is bias? ①

You will all be familiar with the term 'bias'. For example, if you've ever watched a sports game you'll probably have accused a referee of being 'biased' at some point, or perhaps you've watched a TV show like *The X Factor* and felt that one of the judges was 'biased' towards the acts that they mentored. In these contexts, bias means that someone isn't evaluating the evidence (e.g., someone's singing) in an objective way: there are other things affecting their conclusions. Similarly, when we analyse data there can be things that lead us to the wrong conclusions.

A bit of revision. We saw in Chapter 2 that, having collected data, we usually fit a model that represents the hypothesis that we want to test. This model is usually a linear model, which takes the form of equation (2.4). To remind you, it looks like this:

$$\text{outcome}_i = (b_1X_{1i} + b_2X_{2i} + \dots + b_nX_{ni}) + \text{error}_i$$

Therefore, we predict an outcome variable from some kind of model. That model is described by one or more predictor variables (the X s in the equation) and parameters (the b s in the equation) that tell us something about the relationship between the predictor and the outcome variable. Finally, the model will not predict the outcome perfectly, so for each observation there will be some error.

When we fit a model to the data, we estimate the parameters and we usually use the method of least squares (Section 2.4.3). We're not interested in our sample so much as a more general population to which we don't have access, so we use the sample data to estimate the value of the parameters in the population (that's why we call them estimates rather than values). When we estimate a parameter we also compute an estimate of how well it represents the population such as a standard error (Section 2.5.1) or confidence interval (Section 2.5.2). We also can test hypotheses about these parameters by computing test statistics and their associated probabilities (p -values, Section 2.6.1). Therefore, when we think about bias, we need to think about it within three contexts:

- 1 things that bias the parameter estimates (including effect sizes);
- 2 things that bias standard errors and confidence intervals;
- 3 things that bias test statistics and p -values.

These situations are related: firstly, if the standard error is biased then the confidence interval will be too because it is based on the standard error; secondly, test statistics are usually based on the standard error (or something related to it), so if the standard error is biased test statistics will be too; and thirdly, if the test statistic is biased then so too will its p -value. It is important that we identify and eliminate anything that might affect the information that we use to draw conclusions about the world: if our test statistic is inaccurate (or biased) then our conclusions will be too.



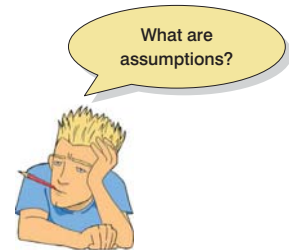
Sources of bias come in the form of a two-headed, fire-breathing, green-scaled beast that jumps out from behind a mound of blood-soaked moss to try to eat us alive. One of its heads goes by the name of unusual scores, or ‘outliers’, while the other is called ‘violations of assumptions’. These are probably names that led to it being teased at school, but then it could breath fire from both heads so it could handle that. Onward into battle ...

5.2.1. Assumptions ①

Most of our potential sources of bias come in the form of violations of assumptions, and you will often hear or read about ‘assumptions’ of statistical tests. An assumption is a condition that ensures that what you’re attempting to do works. For example, when we assess a model using a test statistic, we have usually made some assumptions, and if these assumptions are true then we know that we can take the test statistic (and, therefore, p -value) associated with a model at face value and interpret it accordingly. Conversely, if any of the assumptions are not true (usually referred to as a violation) then the test statistic and p -value will be inaccurate and could lead us to the wrong conclusion if we interpret them at face value.

Assumptions are often presented so that it seems like different statistical procedures have their own unique set of assumptions. However, because we’re usually fitting variations of the linear model to our data (see Section 2.4), all of the tests in this book basically have the same assumptions. These assumptions relate to the quality of the model itself, and the test statistics used to assess it (which are usually **parametric tests** based on the normal distribution). The main assumptions that we’ll look at are:

- additivity and linearity;
- normality of something or other;
- homoscedasticity/homogeneity of variance;
- independence.

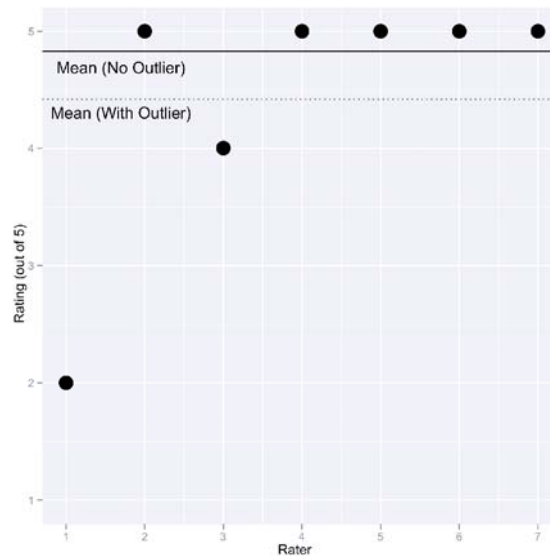


5.2.2. Outliers ①

I mentioned that the first head of the beast of bias is called ‘outliers’. An **outlier** is a score very different from the rest of the data. Let’s look at an example. When I published my first book (the first edition of this book), I was very excited and I wanted everyone in the world to love my new creation and me. Consequently, I obsessively checked the book’s ratings on amazon.co.uk. Customer ratings can range from 1 to 5 stars, where 5 is the best. Back in 2002, my first book had seven ratings (in the order given) of 2, 5, 4, 5, 5, 5, and 5. All but one of these ratings are fairly similar (mainly 5 and 4) but the first rating was quite different from the rest – it was a rating of 2 (a mean and horrible rating). Figure 5.2 plots seven reviewers on the horizontal axis and their ratings on the vertical axis. There is also a horizontal line that represents the mean rating (4.43, as it happens). It should be clear that all of the scores except one lie close to this line. The score of 2 is very different and lies some way below the mean. This score is an example of an outlier – a weird and unusual person (I mean, score) that deviates from the rest of humanity (I mean, data set). The dashed horizontal line represents the mean of the scores when the outlier is not included (4.83). This line is higher than the original mean, indicating that by ignoring this score the mean increases (it increases by 0.4). This example shows how a single score, from

FIGURE 5.2

The first seven customer ratings of this book on www.amazon.co.uk (in about 2002). The first score biases the mean



some mean-spirited badger turd, can bias a parameter such as the mean: the first rating of 2 drags the average down. Based on this biased estimate, new customers might erroneously conclude that my book is worse than the population actually thinks it is. Although I am consumed with bitterness about this whole affair, it has at least given me a great example of an outlier.

The example illustrates that outliers can bias a parameter estimate, but it has an even greater influence on the error associated with that estimate. Back in Section 2.4.1 we looked at example of the number of friends that 5 statistics lecturers had. The data were 1, 3, 4, 3, 2, the mean was 2.6 and the sum of squared error was 5.2. Let's replace one of the scores with an outlier by changing the 4 to a 10. The data are now: 1, 3, 10, 3, and 2.



SELF-TEST Compute the mean and sum of squared error for the new data set.

If you did the self-test, you should find that the mean of the data set with the outlier is 3.8 and the sum of squared error is 50.8. Figure 5.3 shows these values; like Figure 2.7 it shows the sum of squared error (y-axis) associated with different potential values of the mean (the parameter we're estimating, b). For both the original data set and the one with the outlier the estimate for the mean is the optimal estimate: it is the one with the least error, which you can tell by the fact the curve converges on the values of the mean (2.6 and 3.8). The presence of the outlier, however, pushes the curve to the right (i.e., it makes the mean higher) and pushes it upwards too (i.e., it makes the sum of squared error larger). By comparing how far the curves shift horizontally compared to vertically you should (I hope) get a clear sense that the outlier affects the sum of squared error more dramatically than it affects the parameter estimate itself. This is because we use squared errors, so any bias created by the outlier is magnified by the fact that deviations are squared.²

² In this example, the difference between the outlier and the mean (the deviance) is $10 - 3.8 = 6.2$. The deviance squared is $6.2^2 = 38.44$. Therefore, of the 50.8 units of error that we have, a whopping 38.44 are attributable to the outlier.

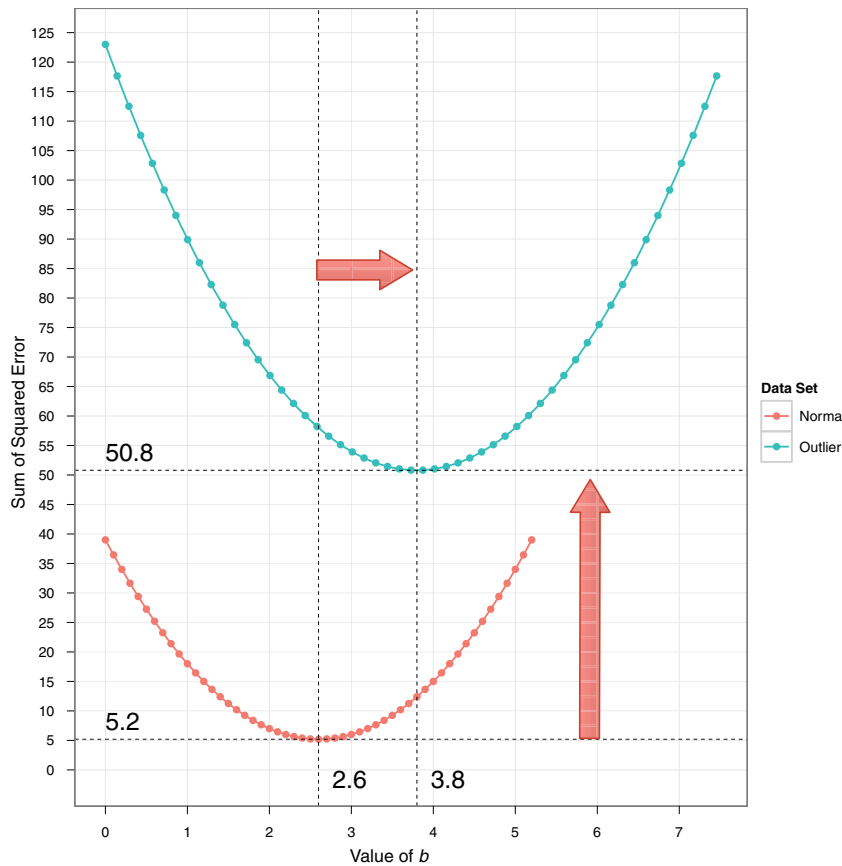


FIGURE 5.3 The effect of an outlier on a parameter estimate (the mean) and its associated estimate of error (the sum of squared errors)

We have seen that outliers can bias estimates of parameters (such as the mean), and also dramatically affect the sum of squared errors. This latter point is important because the sum of squared errors is used to compute the standard deviation, which in turn is used to estimate the standard error, which itself is used to calculate confidence intervals around the parameter estimate. Therefore, if the sum of squared errors is biased, so are the standard error and the confidence intervals associated with the parameter estimate. In addition, most test statistics are based on sums of squares so these will be biased too by outliers.

5.2.3. Additivity and linearity ①

The second head of the beast of bias is called ‘violation of assumptions’. The first assumption we’ll look at is additivity and linearity. The vast majority of statistical models in this book are based on the linear model, which takes this form:

$$\text{outcome}_i = (b_1X_{1i} + b_2X_{2i} + \dots + b_nX_{ni}) + \text{error}_i$$

The assumption of additivity and linearity means that the outcome variable is, in reality, linearly related to any predictors (i.e., their relationship can be summed up by a straight line – think back to Jane Superbrain Box 2.1), and that if you have several predictors then their combined effect is best described by adding their effects together. In other words, it means that the process we’re trying to model can be accurately described as:

$$b_1X_{1i} + b_2X_{2i} + \dots + b_nX_{ni}$$



This assumption is the most important because if it is not true then even if all other assumptions are met, your model is invalid because you have described it incorrectly. It's a bit like calling your pet cat a dog: you can try to get it to go in a kennel, or to fetch sticks, or to sit when you tell it to, but don't be surprised when its behaviour isn't what you expect because even though you've called it a dog, it is in fact a cat. Similarly, if you have described your statistical model inaccurately it won't behave itself and there's no point in interpreting its parameter estimates or worrying about significance tests or confidence intervals: the model is wrong.

5.2.4. Normally distributed something or other ①

The second assumption relates to the normal distribution, which we encountered in Chapter 1 and so we know what it looks like and we (hopefully) understand it. The normal distribution is relevant to many of the things we want to do when we fit models to data and assess them:

- **Parameter estimates:** The mean is a parameter, and we saw in the previous section (the Amazon ratings) that extreme scores can bias it. This illustrates that estimates of parameters are affected by non-normal distributions (such as those with outliers). Parameter estimates differ in how much they are biased in a non-normal distribution: the median, for example, is less biased by skewed distributions than the mean.
- **Confidence intervals:** We use values of the standard normal distribution to compute the confidence interval (Section 2.5.2.1) around a parameter estimate (e.g., the mean, or a b in equation (2.4)). Using values of the standard normal distribution makes sense only if the parameter estimates actually come from one.
- **Null hypothesis significance testing:** If we want to test a hypothesis about a model (and, therefore, the parameter estimates within it) using the framework described in Section 2.6.1 then we assume that the parameter estimates have a normal distribution. We assume this because the test statistics that we use (which we will learn about in due course) have distributions related to the normal distribution (such as the t , F and chi-square distributions), so if our parameter estimate is normally distributed then these test statistics and p -values will be accurate.
- **Errors:** We've seen that any model we fit will include some error (it won't predict the outcome variable perfectly). We also saw that we could calculate the error for each case of data (called the deviance or residual). If these residuals are normally distributed in the population then using the method of least squares to estimate the parameters (the b s in equation (2.4)) will produce better estimates than other methods.

5.2.4.1. The assumption of normality ②

Many people take the 'assumption of normality' to mean that your data need to be normally distributed. However, that isn't what it means. In fact, there is an awful lot of confusion about what it does mean. We have just looked at ways in which normality might introduce bias, and this list hints that the 'assumption of normality' might mean different things in different contexts:

- 1 For confidence intervals around a parameter estimate (e.g., the mean, or a b in equation (2.4)) to be accurate, that estimate must come from a normal distribution.
- 2 For significance tests of models (and the parameter estimates that define them) to be accurate the *sampling distribution* of what's being tested must be normal.



For example, if testing whether two means are different, the data do not need to be normally distributed, but the sampling distribution of means (or differences between means) does. Similarly, if looking at relationships between variables, the significance tests of the parameter estimates that define those relationships (the bs in equation (2.4)) will be accurate only when the sampling distribution of the estimate is normal.

- 3 For the estimates of the parameters that define a model (the bs in equation (2.4)) to be optimal (have the least possible error given the data) the residuals (the error, ϵ_i in equation 2.4) in the population must be normally distributed. This is true mainly if we use the method of least squares (Section 2.4.3), which we often do.

The misconception that people often have about the data themselves needing to be normally distributed probably stems from the fact that if the data are normally distributed then it's reasonable to assume that the errors in the model and the sampling distribution are too (and remember, we don't have direct access to the sampling distribution, so we have to make educated guesses about its shape). Therefore, the assumption of normality tends to get translated as 'your data need to be normally distributed', even though that's not really what it means (see Jane Superbrain Box 5.1 for some more information).

5.2.4.2. The central limit theorem revisited ③

To understand when and if we need to worry about the assumption of normality we need to revisit the central limit theorem,³ which we encountered in Section 2.5.1. Imagine



JANE SUPERBRAIN 5.1

The assumption of normality with categorical predictors ②

Although it is often the shape of the sampling distribution that matters, researchers tend to look at the scores on the outcome variable (or the residuals) when assessing normality. An important thing to remember is that when you have a categorical predictor variable (such as people falling into different groups) you wouldn't expect the overall distribution of the outcome (or residuals) to be normal. For example, if you have seen the movie *The*

Muppets, you will know that muppets live among us. Imagine you predicted that muppets are happier than humans (on TV they seem to be). You collect happiness scores in some muppets and some humans and plot the frequency distribution. You get the graph on the left of Figure 5.4 and decide that your data are not normal: you think that you have violated the assumption of normality. However, you haven't because you predicted that humans and muppets will differ in happiness; in other words, you predict that they come from different populations. If we plot separate frequency distributions for humans and muppets (right of Figure 5.4) you'll notice that within each group the distribution of scores is very normal. The data are as you predicted: muppets are happier than humans and so the centre of their distribution is higher than that of humans. When you combine all of the scores this gives you a bimodal distribution (i.e., two humps). This example illustrates that it is not the normality of the outcome (or residuals) overall that matters, but normality at each unique level of the predictor variable.

³ The 'central' in the name refers to the theorem being important and far-reaching and has nothing to do with centres of distributions.

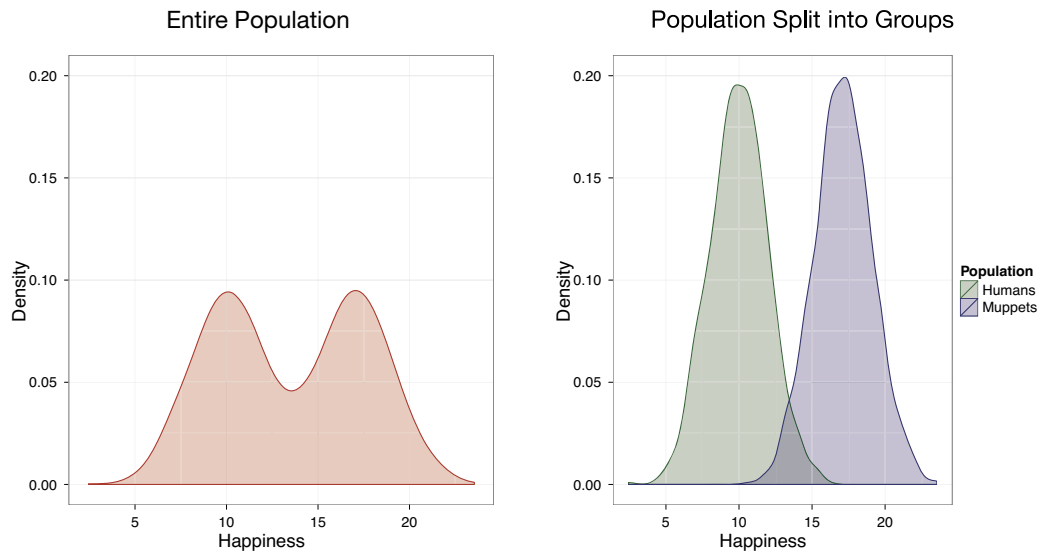


FIGURE 5.4 A distribution that looks non-normal (left) could be made up of different groups of normally distributed scores

we have a population of scores that is not normally distributed. Figure 5.5 shows such a population containing scores of how many friends statistics lecturers have: it is very skewed, with most lecturers having only one friend, and the frequencies declining as the number of friends increases to the maximum score of 7 friends. I'm not tricking you; this population is as far removed from the bell-shaped normal curve as it looks. Imagine that I took samples of 5 scores from this population and in each sample I estimated a parameter (let's say I computed the mean) and then replaced the scores. In fact, I took 5000 samples, and consequently I have 5000 values of the parameter estimate (each one from a different sample). Let's look what happens when we plot these 5000 values in a frequency distribution. The frequency distribution of the 5000 parameter estimates from the 5000 samples is on the far left of Figure 5.5. This is the sampling distribution of the parameter estimate. Note that it is quite skewed, but not as skewed as the population. Imagine now that I repeated the sampling process, but this time my samples each contained 30 scores instead of only 5. The resulting distribution of the 5000 parameter estimates is in the centre of Figure 5.5. There is still skew in this sampling distribution but it is a lot more normal than when the samples were based on only 5 scores. Finally, I repeated the whole process but this time took samples of 100 scores rather than 30. The resulting distribution of the 5000 parameter estimates is basically normal (right of Figure 5.5). As our sample sizes got bigger the sampling distributions became more normal, up to point at which the sample is big enough that the sampling distribution is normal – despite the fact that the population of scores was very non-normal indeed. This is the central limit theorem: regardless of the shape of the population, parameter estimates of that population will have a normal distribution provided the samples are 'big enough' (see Jane Superbrain Box 5.2).

5.2.4.3. When does the assumption of normality matter? ②

The central limit theorem means that *there are a variety of situations in which we can assume normality regardless of the shape of our sample data* (Lumley, Diehr, Emerson, & Chen, 2002). Let's think back to the things affected by normality:

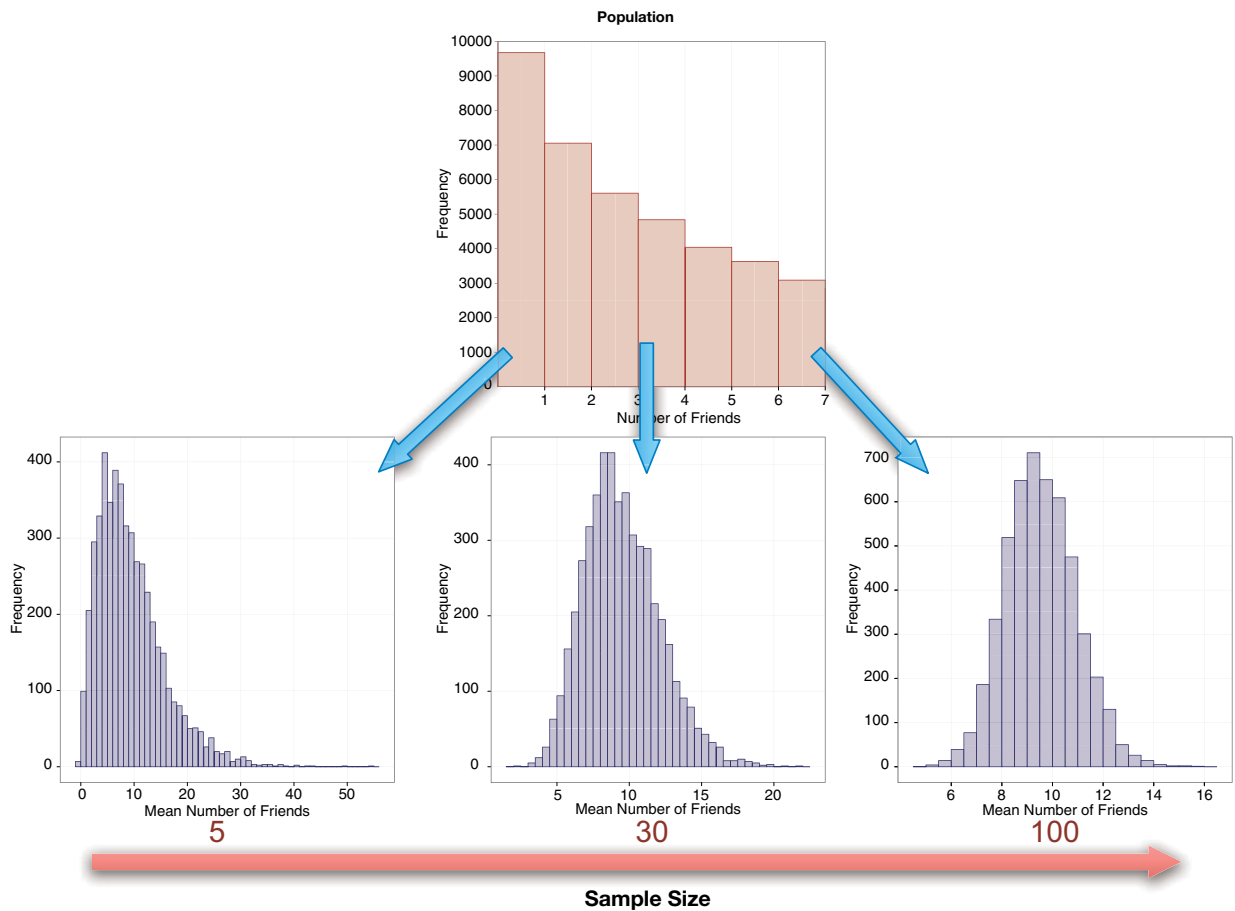


FIGURE 5.5 Parameter estimates sampled from a non-normal population. As the sample size increases, the distribution of those parameters becomes increasingly normal



ODITI'S LANTERN

The central limit theorem

'I, Odi, believe that the central limit theorem is key to unlocking the hidden truths that the cult strives to find. The true wonder of the CLT cannot be understood by a static diagram and the ramblings of a damaged mind. Only by staring into my lantern can you see the CLT at work in all its glory. Go forth and look into the abyss.'

- 1 For confidence intervals around a parameter estimate (e.g., the mean, or a b in equation (2.4)) to be accurate, that estimate must come from a normal distribution. The central limit theorem tells us that in large samples the estimate will have come from a normal distribution regardless of what the sample or population data look like. Therefore, if we are interested in computing confidence intervals then we don't need to worry about the assumption of normality if our sample is large enough.
- 2 For significance tests of models to be accurate the sampling distribution of what's being tested must be normal. Again, the central limit theorem





JANE SUPERBRAIN 5.2

Size really does matter ②

How big is 'big enough' for the central limit theorem to kick in? The widely accepted value is a sample size of 30, and we saw in Figure 5.4 that with samples of

this size we started to get a sampling distribution that approximated normal. However, we also saw that with samples of 100 we got a better approximation of normal. As with most things in statistics, there isn't a simple answer: how big is 'big enough' depends on the distribution of the population. In light-tailed distributions (where outliers are rare) an N as small as 20 can be 'large enough', but in heavy-tailed distributions (where outliers are common) then up to 100 or even 160 might be necessary. If the distribution has a lot of skew and kurtosis you might need a very large sample indeed for the central limit theorem to work. It also depends on the parameter that you're trying to estimate (Wilcox, 2010, discusses this issue in detail).

tells us that in large samples this will be true no matter what the shape of the population. Therefore, the shape of our data shouldn't affect significance tests *provided our sample is large enough*. However, the extent to which test statistics perform as they should do in large samples varies across different test statistics, and we will deal with these idiosyncratic issues in the appropriate chapter.

- 3 For the estimates of model parameters (the bs in equation (2.4)) to be optimal (using the method of least squares) the residuals in the population must be normally distributed. The method of least squares will always give you an estimate of the model parameters that minimizes error, so in that sense you don't need to assume normality of anything to fit a linear model and estimate the parameters that define it (Gelman & Hill, 2007). However, there are other methods for estimating model parameters, and if you happen to have normally distributed errors then the estimates that you obtained using the method of least squares will have less error than the estimates you would have got using any of these other methods.

To sum up then, if all you want to do is estimate the parameters of your model then normality doesn't really matter. If you want to construct confidence intervals around those parameters, or compute significance tests relating to those parameters, then the assumption of normality matters in small samples, but because of the central limit theorem we don't really need to worry about this assumption in larger samples (but see Jane Superbrain Box 5.2). In practical terms, as long as your sample is fairly large, outliers are a more pressing concern than normality. Although we tend to think of outliers as isolated very extreme cases, you can have outliers that are less extreme but are not isolated cases. These outliers can dramatically reduce the power of significance tests (Jane Superbrain Box 5.3).

5.2.5. Homoscedasticity/homogeneity of variance ②

The second assumption we'll explore relates to variance (Section 1.6.3), which can affect the two main things that we might do when we fit models to data:



JANE SUPERBRAIN 5.3

Stealth outliers ③

Although we often think of outliers as one or two very extreme scores, sometimes they soak themselves in radar-absorbent paint and contort themselves into strange shapes so as to avoid detection. These ‘stealth outliers’ (that’s my name for them; no one else calls them that) hide undetected in data sets, radically affecting analyses. Imagine you collected happiness scores, and when you plotted the frequency distribution it looked like Figure 5.6 (left). You might decide that this distribution is normal, because it has the characteristic bell-shaped curve. However, it is not: it is a **mixed normal distribution** or **contaminated normal distribution** (Tukey, 1960). The happiness scores on the left of Figure 5.6 are made up of two distinct populations: 90% of scores are from

humans, but 10% are from muppets (we saw in Jane Superbrain Box 5.1 that they live among us). Figure 5.6 (right) reproduces this overall distribution (the blue one), but also shows the unique distributions for the humans (red) and muppets (Kermit-coloured green) that contribute to it.

The human distribution is a perfect normal distribution, but the curve for the muppets is flatter and heavier in the tails, showing that muppets are more likely than humans to be extremely happy (like Kermit) or extremely miserable (like Statler and Waldorf). When these populations combine, the muppets contaminate the perfectly normal distribution of humans: the combined distribution (blue) has slightly more scores in the extremes than a perfect normal distribution (red). The muppet scores have affected the overall distribution even though (1) they make up only 10% of the scores; and (2) their scores are more frequent at the extremes of ‘normal’ and not radically different like you might expect an outlier to be. These extreme scores inflate estimates of the population variance (think back to Jane Superbrain Box 1.5). Mixed normal distributions are very common and they reduce the power of significance tests – see Wilcox (2010) for a thorough account of the problems associated with these distributions.

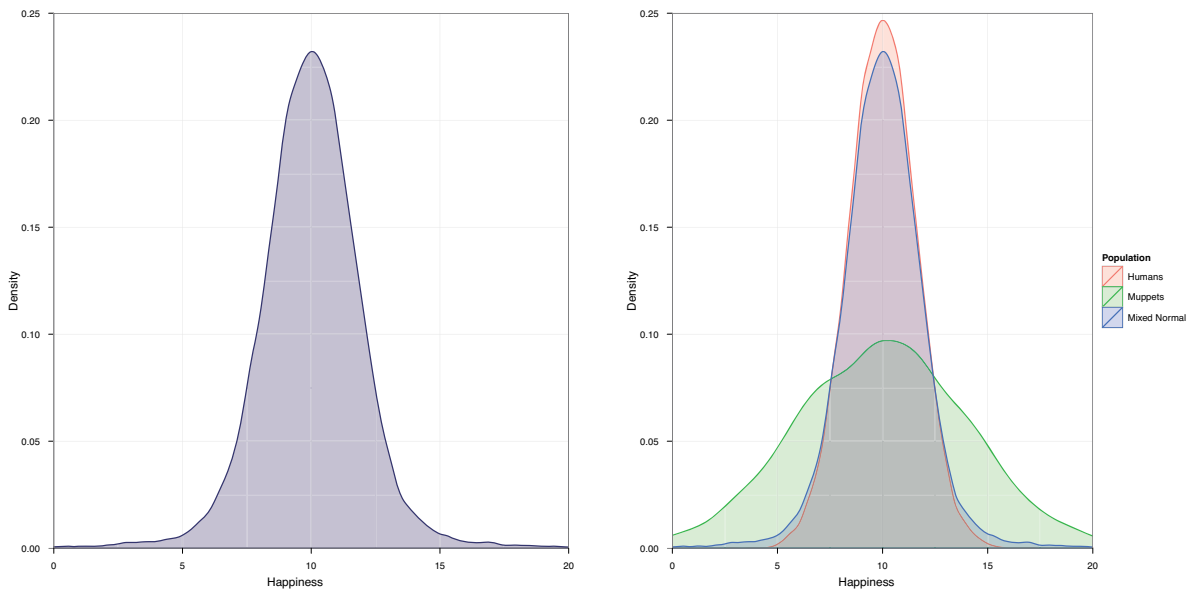


FIGURE 5.6 An apparently normal distribution (left), which is actually a ‘mixed normal’ distribution made up of two populations (right)



- **Parameters:** If we use the method of least squares (Section 2.4.3) to estimate the parameters in the model, then this will give us optimal estimates if the variance of the outcome variable is equal across different values of the predictor variable.
- **Null hypothesis significance testing:** Test statistics often assume that the variance of the outcome variable is equal across different values of the predictor variable. If this is not the case then these test statistics will be inaccurate.

Therefore, to make sure our estimates of the parameters that define our model and significance tests are accurate we have to assume homoscedasticity (also known as homogeneity of variance).

5.2.5.1. What is homoscedasticity/homogeneity of variance? ②

In designs in which you test several groups of participants this assumption means that each of these samples comes from populations with the same variance. In correlational designs, this assumption means that the variance of the outcome variable should be stable at all levels of the predictor variable. In other words, as you go through levels of the predictor variable, the variance of the outcome variable should not change. Let's illustrate this idea with an example. An audiologist was interested in the effects of loud concerts on people's hearing. She sent 10 people on tour with the loudest band she could find, Motörhead. These people went to concerts in Brixton (London), Brighton, Bristol, Edinburgh, Newcastle, Cardiff and Dublin, and the audiologist measured for how many hours after the concert these people had ringing in their ears.

The top of Figure 5.7 shows the number of hours that each person (represented by a circle) had ringing in his or her ears after each concert. The squares show the average number of hours of ringing in the ears after each concert. A line connects these means so that we can see the general trend. For each concert, the circles are the scores from which the mean is calculated. We can see in both graphs that the means increase as the people go to more concerts: there is a cumulative effect of the concerts on ringing in the ears. The graphs don't differ with respect to the means (which are roughly the same), but do differ in the *spread* of scores around the mean. The bottom of Figure 5.7 removes the data and replaces it with a bar that shows the range of the scores displayed in the top figure. In the left-hand graphs, the green bars are roughly the same length, which tells us that the spread of scores around the mean was roughly the same at each concert. This is what we mean by **homogeneity of variance** or **homoscedasticity**:⁴ the spread of scores for hearing loss is the same at each level of the concert variable (i.e., the spread of scores is the same at Brixton, Brighton, Bristol, Edinburgh, Newcastle, Cardiff and Dublin). The right-hand side of Figure 5.7 shows a different scenario: the scores after the Brixton concert (which are again displayed by the green lines in the bottom part of the figure) are quite tightly packed around the mean (the vertical distance from the lowest score to the highest score is small), but after the Dublin show (for example) the scores are very spread out around the mean (the vertical distance from the lowest score to the highest score is large). In general, the green bars on the right differ in length, showing that the spread of scores was different at each concert. This scenario is an example of **heterogeneity of variance** or **heteroscedasticity**: at some levels of the concert variable the variance of scores is different than at other levels (graphically, the vertical distance from the lowest to highest score is different after different concerts).

⁴ My explanation is a bit simplified because usually we're making the assumption about the errors in the model and not the data themselves, but the two things are related.



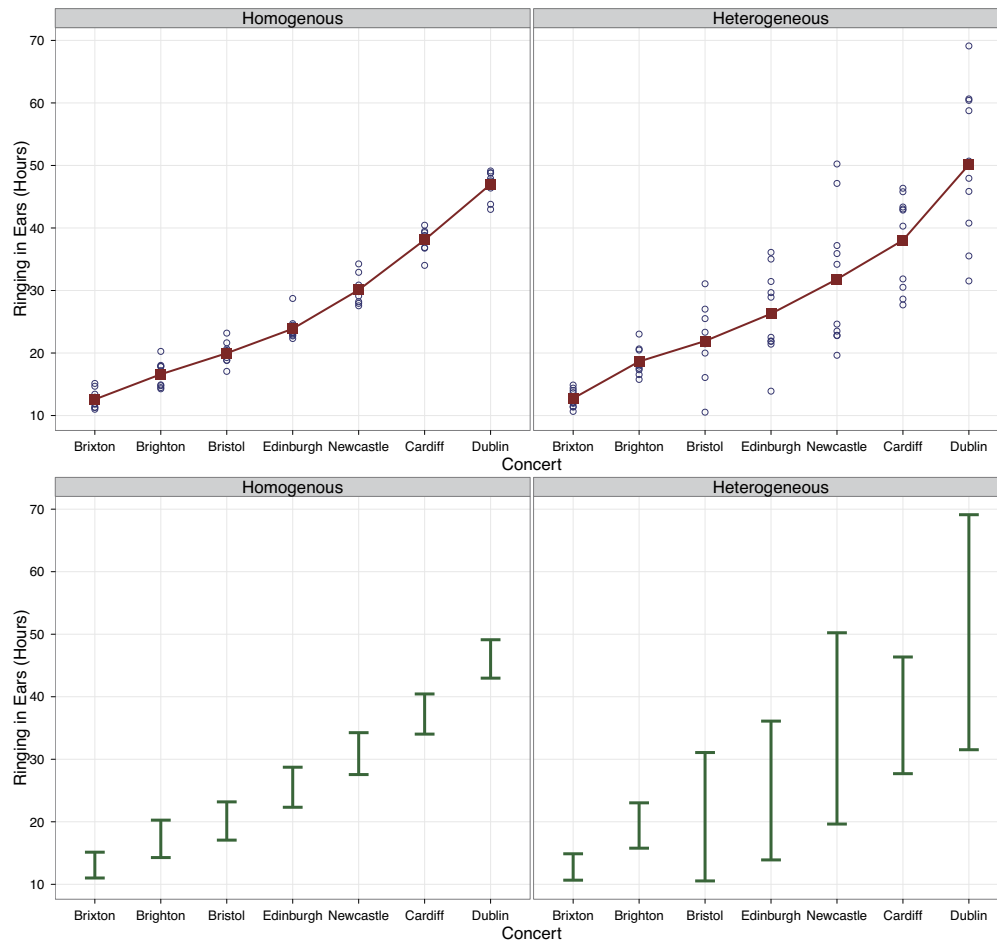
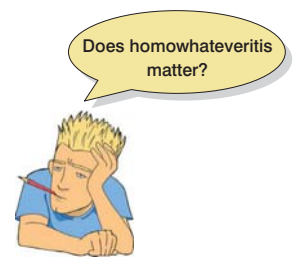


FIGURE 5.7 Graphs illustrating data with homogeneous (left) and heterogeneous (right) variances

5.2.5.2. When does homoscedasticity/homogeneity of variance matter? ②

In terms of estimating the parameters within a linear model, if we assume equality of variance then the estimates we get using the method of least squares will be optimal. If variances for the outcome variable differ along the predictor variable then the estimates of the parameters within the model will not be optimal. The method of least squares will produce ‘unbiased’ estimates of parameters even when homogeneity of variance can’t be assumed, but better estimates can be achieved using different methods, for example, by using **weighted least squares** in which each case is weighted by a function of its variance. Therefore, if all you care about is estimating the parameters of the model in your sample then you don’t need to worry about homogeneity of variance in most cases: the method of least squares will produce unbiased estimates (Hayes & Cai, 2007).

However, unequal variances/heteroscedasticity creates a bias and inconsistency in the estimate of the standard error associated with the parameter estimates in your model (Hayes & Cai, 2007). As such, your confidence intervals and significance tests for the parameter estimates will be biased, because they are computed using the standard error. Confidence intervals can be ‘extremely inaccurate’ when homogeneity of variance/homoscedasticity cannot be assumed (Wilcox, 2010). Therefore, if you want to look at the confidence intervals around your model parameter estimates or to test the significance of the





model or its parameter estimates then homogeneity of variance matters. Some test statistics are designed to be accurate even when this assumption is violated, and we'll discuss these in the appropriate chapters.

5.2.6. Independence ②

This assumption means that the errors in your model (the error_{*i*} in equation (2.4)) are not related to each other. Imagine Paul and Julie were participants in an experiment where they had to indicate whether they remembered having seen particular photos. If Paul and Julie were to confer about whether they'd seen certain photos then their answers would *not* be independent: Julie's response to a given question would depend on Paul's answer. We know already that if we estimate a model to predict their responses, there will be error in those predictions and because Paul and Julie's scores are not independent the errors associated with these predicted values will also not be independent. If Paul and Julie were unable to confer (if they were locked in different rooms) then the error terms should be independent (unless they're telepathic): the error in predicting Paul's response should not be influenced by the error in predicting Julie's response.

The equation that we use to estimate the standard error (equation (2.8)) is valid only if observations are independent. Remember that we use the standard error to compute confidence intervals and significance tests, so if we violate the assumption of independence then our confidence intervals and significance tests will be invalid. If we use the method of least squares, then model parameter estimates will still be valid but not optimal (we could get better estimates using a different method). In general, if this assumption is violated, we should apply the techniques covered in Chapter 20, so it is important to identify whether the assumption is violated.

5.3. Spotting bias ②

5.3.1. Spotting outliers ②

When they are isolated, extreme cases and outliers are fairly easy to spot using graphs such as histograms and boxplots; it is considerably trickier when outliers are more subtle (using *z*-scores may be useful – Jane Superbrain Box 5.4). Let's look at an example. A biologist was worried about the potential health effects of music festivals. She went to the Download Music Festival⁵ (those of you outside the UK can pretend it is Roskilde Festival, Ozzfest, Lollapalooza, Wacken or something) and measured the hygiene of 810 concert-goers over the three days of the festival. She tried to measure every person on every day but, because it was difficult to track people down, there were missing data on days 2 and 3. Hygiene was measured using a standardized technique (don't worry, it *wasn't* licking the person's armpit) that results in a score ranging between 0 (you smell like you've bathed in sewage) and 4 (you smell of sweet roses on a fresh spring day). I know from bitter experience that sanitation is not always great at these places (the Reading Festival seems particularly bad) and so the biologist predicted that personal hygiene would go down dramatically over the three days of the festival. The data can be found in **DownloadFestival.sav**.

⁵ <http://www.downloadfestival.co.uk>



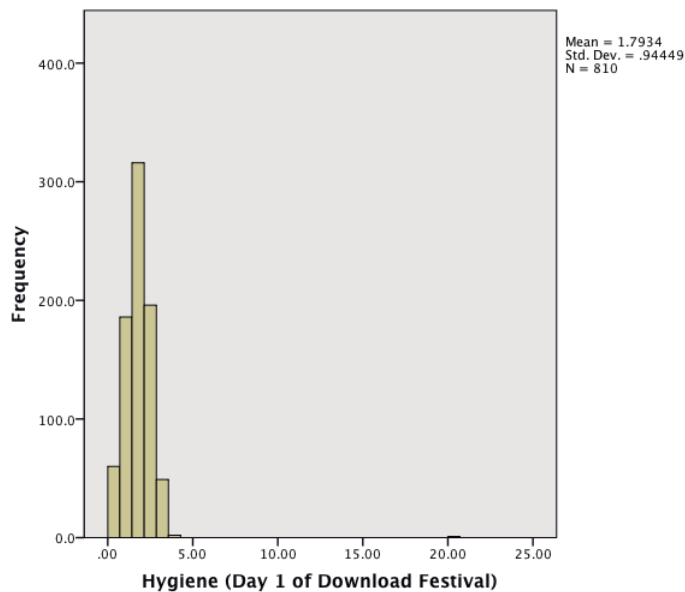


FIGURE 5.8
Histogram of the
day 1 Download
Festival hygiene
scores



SELF-TEST Using what you learnt in Section 4.4, plot a histogram of the hygiene scores on day 1 of the festival.

The resulting histogram is shown in Figure 5.8. The first thing that should leap out at you is that there is one case that is very different from the others. All of the scores appear to be squashed up at one end of the distribution because they are all less than 5 (yielding a very pointy distribution) except for one, which has a value of 20. This score is an obvious outlier because it is above the top of our scale (remember our hygiene scale ranged only from 0 to 4). It must be a mistake. However, with 810 cases, how on earth do we find out which case it was? You could just look through the data, but that would certainly give you a headache, and so instead we can use a boxplot (see Section 4.5), which is another very useful way to spot outliers.



SELF-TEST Using what you learnt in Section 4.5, plot a boxplot of the hygiene scores on day 1 of the festival.


The resulting boxplot is shown in Figure 5.9. The outlier that we detected in the histogram has shown up as an extreme score (*) on the boxplot. SPSS helpfully tells us the number of the case (611) that's producing this outlier. If we go to the data editor (data view), we can locate this case quickly by clicking on  and typing 611 in the dialog box that appears. That takes us straight to case 611. Looking at this case reveals a score of 20.02, which is probably a mistyping of 2.02. We'd have to go back to the raw data and check. We'll assume we've checked the raw data and this score should be 2.02, so replace the value 20.02 with the value 2.02 before we continue this example.

FIGURE 5.9

Boxplot of hygiene scores on day 1 of the Download Festival

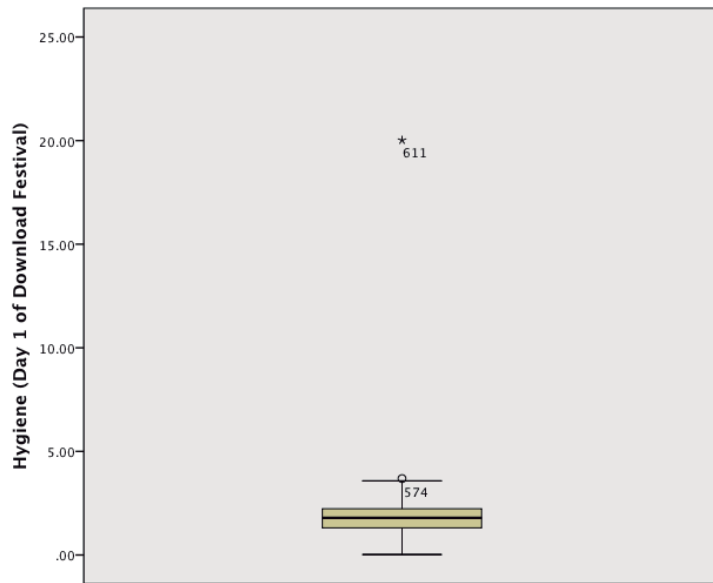
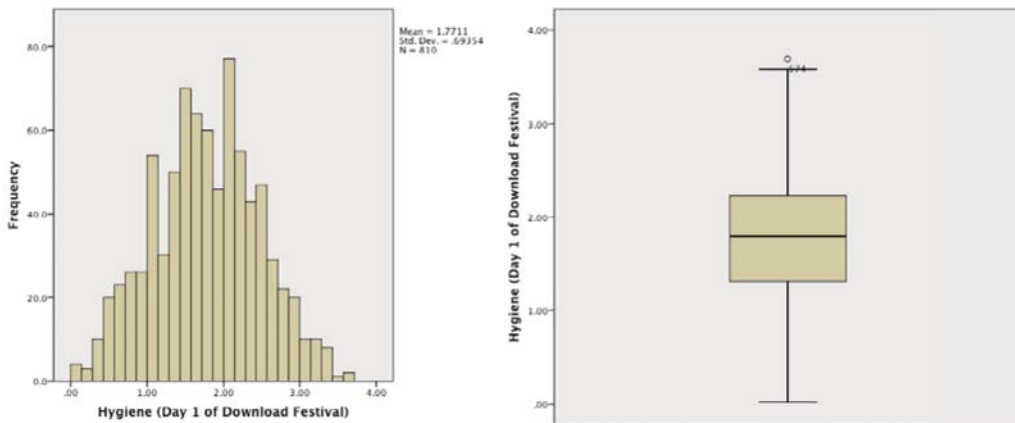


FIGURE 5.10

Histogram (left) and boxplot (right) of hygiene scores on day 1 of the Download Festival after removing the extreme score



SELF-TEST Now we have removed the outlier in the data, re-plot the histogram and boxplot.

Figure 5.10 shows the histogram and boxplot for the data after the extreme case has been corrected. The distribution looks amazingly normal: it is nicely symmetrical and doesn't seem too pointy or flat. Neither plot indicates any particularly extreme scores: the boxplot suggests that case 574 is a mild outlier, but the histogram doesn't seem to show any cases as being particularly out of the ordinary.



SELF-TEST Produce boxplots for the day 2 and day 3 hygiene scores and interpret them.

SELF-TEST Re-plot these scores but splitting by **Gender** along the x-axis. Are there differences between men and women?

5.3.2. Spotting normality ①

5.3.2.1. Using graphs to spot normality ①

Frequency distributions are not only good for spotting outliers; they are the natural choice for looking at the shape of the distribution as a whole. We have already plotted a histogram of the day 1 scores (Figure 5.10). The **P-P plot** (probability–probability plot) is another useful graph for checking normality; it plots the cumulative probability of a variable against the cumulative probability of a particular distribution (in this case we would specify a



JANE SUPERBRAIN 5.4

Using z-scores to find outliers ③

We saw in Section 1.6.4 that z-scores express scores in terms of a distribution with a mean of 0 and a standard deviation of 1. By converting our data to z-scores we can use benchmarks that we can apply to any data set (regardless of what its original mean and standard deviation were) to search for outliers. We can get SPSS to do this conversion using the **Analyze Descriptive Statistics** dialog box. Select the variable(s) to convert (such as day 2 of the hygiene data as in the diagram) and tick the *Save standardized values as variables* option (Figure 5.11). SPSS will create a new variable in the data editor (with the same name prefixed with the letter z).

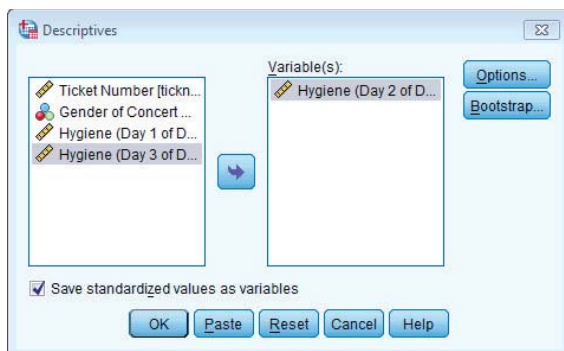


FIGURE 5.11 Saving z-scores

To look for outliers we can count how many z-scores fall within certain important limits. If we ignore whether

the z-score is positive or negative (called the 'absolute value'), then in a normal distribution we'd expect about 5% to be greater than 1.96 (we often use 2 for convenience), 1% to have absolute values greater than 2.58, and none to be greater than about 3.29. To get SPSS to do the counting for you, use the syntax file **Outliers (Percentage of Z-scores).sps** (on the companion website), which will produce a table for day 2 of the Download Festival hygiene data. Load this file and run the syntax (see Section 3.9). It uses the following commands:

```
DESCRIPTIVES
VARIABLES= day2/SAVE.
COMPUTE zday2= abs(zday2).
EXECUTE.
```

These commands use the *descriptives* function on the variable **day2** to save the z-scores in the data editor (as a variable called **zday2**). We then use the *compute* command to change **zday2** so that it contains the absolute values.

```
RECODE
zday2 (3.29 thru highest = 1)(2.58 thru highest = 2)
(1.96 thru highest = 3)(Lowest thru 1.95 = 4).
EXECUTE.
```

These commands recode the variable **zday2** so that if a value is greater than 3.29 it's assigned a code of 1, if it's greater than 2.58 it's assigned a code of 2, if it's greater than 1.96 it's assigned a code of 3, and if it's less than 1.95 it gets a code of 4.

```
VALUE LABELS zday2
4 'Normal range' 3 'Potential Outliers (z > 1.96)' 2
'Probable Outliers (z > 2.58)' 1 'Extreme (z-score > 3.29)'.
```

This syntax assigns appropriate labels to the codes we defined above.

```
FREQUENCIES
VARIABLES= zday2
/ORDER=ANALYSIS.
```

Finally, this syntax uses the *frequencies* command to produce a table (Output 5.1) telling us the percentage of 1s, 2s, 3s and 4s found in the variable **zday2**. Thinking about what we know about the absolute values of z-scores, we would expect to see only 5% (or less) with an values greater than 1.96, 1% (or less) with values greater than 2.58, and very few cases above 3.29. The column labelled *Cumulative Percent* tells us the corresponding percentages for the hygiene




scores on day 2: 0.8% of cases were above 3.29 (extreme cases), 2.3% (compared to the 1% we'd expect) had values greater than 2.58, and 6.8% (compared to the 5% we would expect) had values greater than 1.96. The remaining cases (which, if you look at the *Valid Percent*, constitute 93.2%) were in the normal range. All in all these percentages are broadly consistent with what we'd expect in a normal distribution (around 95% were in the normal range).

Zscore: Hygiene (Day 2 of Download Festival)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Extreme (z-score > 3.29)	2	.2	.8	.8
	Probable Outliers (z > 2.58)	4	.5	1.5	2.3
	Potential Outliers (z > 1.96)	12	1.5	4.5	6.8
	Normal range	246	30.4	93.2	100.0
	Total	264	32.6	100.0	
Missing	System	546	67.4		
	Total	810	100.0		

OUTPUT 5.1

normal distribution). The data are ranked and sorted, then for each rank the corresponding z-score is calculated to create an 'expected value' that the score should have in a normal distribution. Next, the score itself is converted to a z-score (see Section 1.6.4). The actual z-score is plotted against the expected z-score. If the data are normally distributed then the actual z-score will be the same as the expected z-score and you'll get a lovely straight diagonal line. This ideal scenario is helpfully plotted on the graph and your job is to compare the data points to this line. If values fall on the diagonal of the plot then the variable is normally distributed; however, when the data sag consistently above or below the diagonal then this shows that the kurtosis differs from a normal distribution, and when the data points are S-shaped, the problem is skewness.

To get a P-P plot use **Analyze Descriptive Statistics**  **P-P Plots...** to access the dialog box in Figure 5.12.⁶ There's not a lot to say about this dialog box because the default options will compare any variables selected to a normal distribution, which is what we want (although note that there is a drop-down list of different distributions against which you could compare your data). Select the three hygiene score variables in the variable list (click on the day 1 variable, then hold down *Shift* and select the day 3 variable and the day 2 scores will be selected as well). Transfer the selected variables to the box labelled *Variables* by clicking on . Click on  to draw the graphs.



SELF-TEST Using what you learnt in Section 4.4, plot histograms for the hygiene scores for days 2 and 3 of the Download Festival.

Figure 5.13 shows the histograms (from the self-test tasks) and the corresponding P-P plots. We've looked at the day 1 scores in the previous section and concluded that they

⁶ You'll notice in the same menu something called a Q-Q plot, which is very similar and which we'll discuss later.

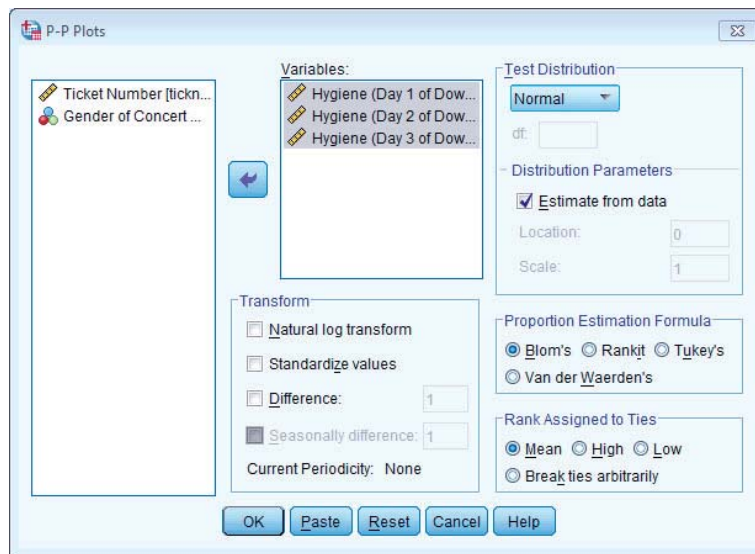


FIGURE 5.12
Dialog box for
obtaining P-P
plots

looked quite normal. The P-P plot echoes this view because the data points all fall very close to the ‘ideal’ diagonal line. However, the distributions for days 2 and 3 are not nearly as symmetrical as day 1: they both look positively skewed. Again, this can be seen in the P-P plots by the data points deviating away from the diagonal. In general, this seems to suggest that by days 2 and 3, hygiene scores were much more clustered around the low end of the scale. Remember that the lower the score, the less hygienic the person is, so generally people became smellier as the festival progressed. The skew occurs because a substantial minority insisted on upholding their levels of hygiene (against all odds) over the course of the festival (baby wet-wipes are indispensable, I find).

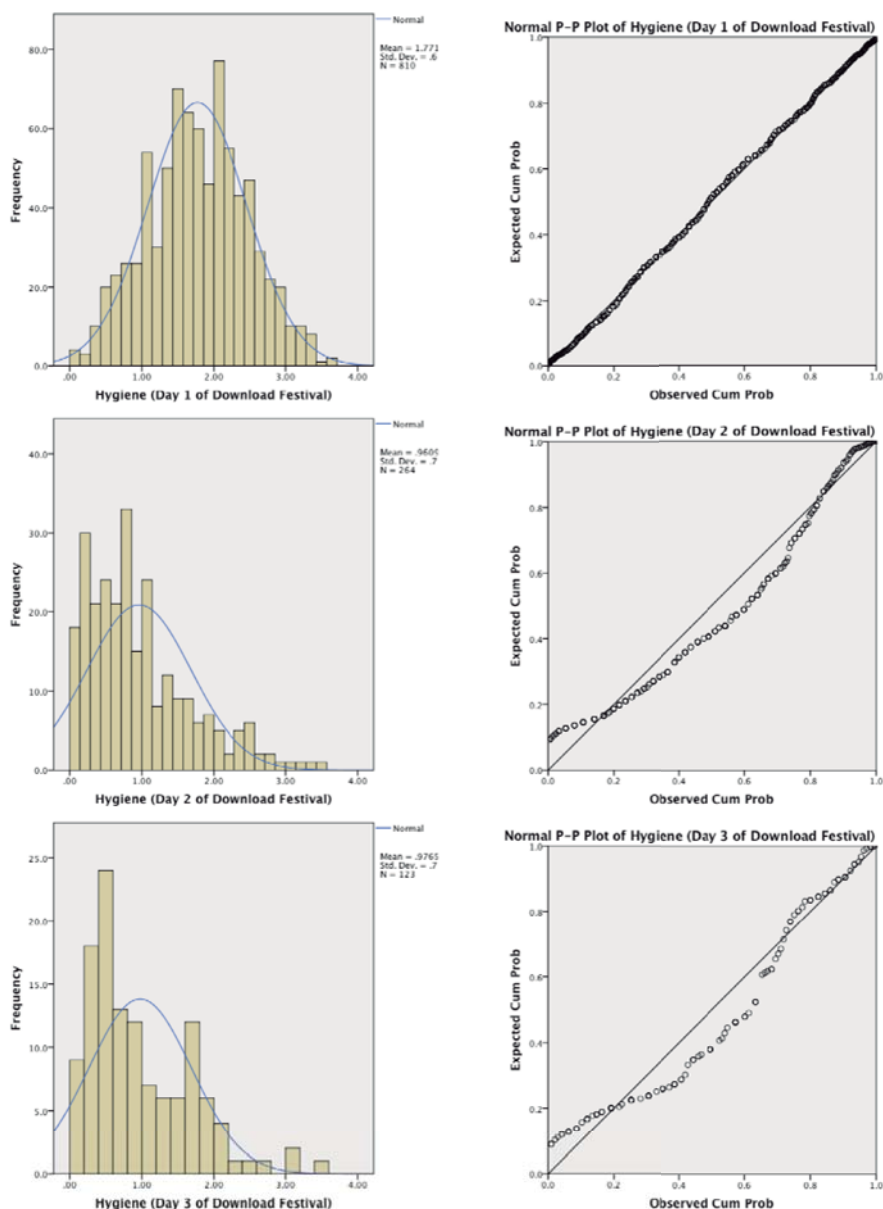
5.3.2.2. Using numbers to spot normality ①

Graphs are particularly useful for looking at normality in big samples; however, in smaller samples it can be useful to explore the distribution of the variables using the *frequencies* command (**Analyze** > **Descriptive Statistics** > **Frequencies...**). The main dialog box is shown in Figure 5.14. The variables in the data editor are listed on the left-hand side, and they can be transferred to the box labelled *Variable(s)* by clicking on a variable (or highlighting several with the mouse) and then clicking on **➤**. If a variable listed in the *Variable(s)* box is selected, it can be transferred back to the variable list by clicking on the arrow button (which should now be pointing in the opposite direction). By default, SPSS produces a frequency distribution of all scores in table form. However, there are two other dialog boxes that can be selected that provide other options. The *Statistics* dialog box is accessed by clicking on **Statistics...**, and the *Charts* dialog box is accessed by clicking on **Charts...**.

The *Statistics* dialog box allows you to select ways to describe a distribution, such as measures of central tendency (mean, mode, median), measures of variability (range, standard deviation, variance, quartile splits), measures of shape (kurtosis and skewness). Select the mean, mode, median, standard deviation, variance and range. To check that a distribution of scores is normal, we can look at the values of kurtosis and skewness (see Section 1.6.1). The *Charts* option provides a simple way to plot the frequency distribution of scores (as a bar chart, a pie chart or a histogram). We’ve already plotted histograms of our data so we don’t need to select these options, but you could use these options in future analyses. When you have selected the appropriate options, return to the main dialog box by clicking on **Continue**. Once in the main dialog box, click on **OK** to run the analysis.



FIGURE 5.13
Histograms
(left) and P-P
plots (right)
of the hygiene
scores over the
three days of
the Download
Festival



Output 5.2 shows the table of descriptive statistics for the three variables in this example. On average, hygiene scores were 1.77 (out of 5) on day 1 of the festival, but went down to 0.96 and 0.98 on days 2 and 3, respectively. The other important measures for our purposes are the skewness and the kurtosis (see Section 1.6.1), both of which have an associated standard error.

There are different ways to calculate skewness and kurtosis, but SPSS uses methods that give values of zero in a normal distribution. Positive values of skewness indicate a pile-up of scores on the left of the distribution, whereas negative values indicate a pile-up on the right. Positive values of kurtosis indicate a pointy and heavy-tailed distribution, whereas negative values indicate a flat and light-tailed distribution. The further the value is from zero, the more likely it is that the data are not normally distributed. For day 1 the skew value is very close to zero (which is good) and kurtosis is a little negative. For days 2 and 3, though, there is a skewness of around 1 (positive skew).



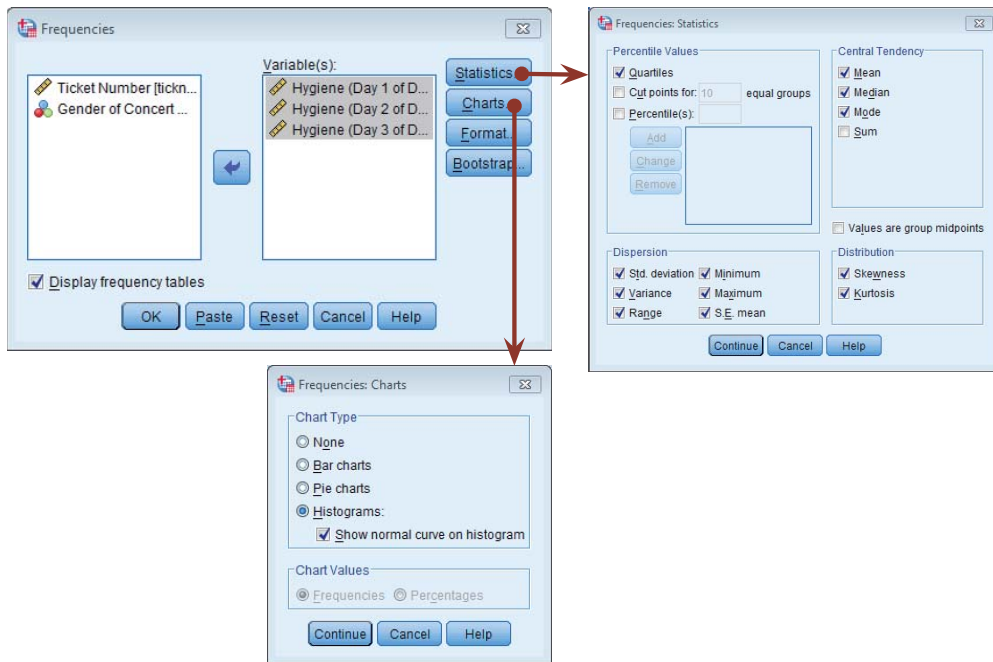


FIGURE 5.14
Dialog boxes for the *frequencies* command

Statistics

	Hygiene (Day 1 of Download Festival)	Hygiene (Day 2 of Download Festival)	Hygiene (Day 3 of Download Festival)
N	810	264	123
Valid			
Missing	0	546	687
Mean	1.7711	.9609	.9765
Std. Error of Mean	.02437	.04436	.06404
Median	1.7900	.7900	.7600
Mode	2.00	.23	.44 ^a
Std. Deviation	.69354	.72078	.71028
Variance	.481	.520	.504
Skewness	-.004	1.095	1.033
Std. Error of Skewness	.086	.150	.218
Kurtosis	-.410	.822	.732
Std. Error of Kurtosis	.172	.299	.433
Range	3.67	3.44	3.39
Minimum	.02	.00	.02
Maximum	3.69	3.44	3.41
Percentiles			
25	1.3050	.4100	.4400
50	1.7900	.7900	.7600
75	2.2300	1.3500	1.5500

a. Multiple modes exist. The smallest value is shown

OUTPUT 5.2

We can convert these values to *z*-scores (Section 1.6.4), which enables us to (1) compare skew and kurtosis values in different samples that used different measures, and (2) calculate a *p*-value that tells us if the values are significantly different from 0 (i.e., normal). Although there are good reasons not to do this (see Jane Superbrain Box 5.5), if you want to you can do it by subtracting the mean of the distribution (in this case zero) from the score and then dividing by the standard error of the distribution.

$$z_{\text{skewness}} = \frac{S - 0}{SE_{\text{skewness}}} \quad z_{\text{kurtosis}} = \frac{K - 0}{SE_{\text{kurtosis}}}$$

In the above equations, the values of *S* (skewness) and *K* (kurtosis) and their respective standard errors are produced by SPSS. These *z*-scores can be compared against values that you would expect to get if skew and kurtosis were not different from 0 (see Section 1.6.4).



JANE SUPERBRAIN 5.5

Significance tests and assumptions ②

Throughout this section we will look at various significance tests that have been devised to look at whether assumptions are violated. These include tests of whether a distribution is normal (the Kolmogorov–Smirnov and Shapiro–Wilk tests), tests of homogeneity of variances (Levene’s test), and tests of significance of skew and kurtosis. Although I cover these tests because people expect to see these sorts of things in introductory statistics books, there is a fundamental problem with using them. They are all based on null hypothesis significance testing, and this means that (1) in large samples

they can be significant even for small and unimportant effects, and (2) in small samples they will lack power to detect violations of assumptions (Section 2.6.1.10).

We have also seen in this chapter that the central limit theorem means that as sample sizes get larger, the assumption of normality matters less because the sampling distribution will be normal regardless of what our population (or indeed sample) data look like. So, the problem is that in large samples, where we don’t need to worry about normality, a test of normality is more likely to be significant, and therefore likely to make us worry about and correct for something that doesn’t need to be corrected or worried about. Conversely, in small samples, where we might want to worry about normality, a significance test won’t have the power to detect non-normality and so is likely to encourage us not to worry about something that we probably ought to. Therefore, the best advice is that if your sample is large then don’t use significance tests of normality; in fact, don’t worry too much about normality at all. In small samples pay attention if your significance tests are significant but resist being lulled into a false sense of security if they are not.

So, an absolute value greater than 1.96 is significant at $p < .05$, above 2.58 is significant at $p < .01$ and above 3.29 is significant at $p < .001$. However, you really should use these criteria only in small samples: in larger samples examine the shape of the distribution visually, interpret the value of the skewness and kurtosis statistics, and possibly don’t even worry about normality at all (Jane Superbrain Box 5.5).

For the hygiene scores, the z -score of skewness is $-0.004/0.086 = 0.047$ on day 1, $1.095/0.150 = 7.300$ on day 2 and $1.033/0.218 = 4.739$ on day 3. It is pretty clear then that although on day 1 scores are not at all skewed, on days 2 and 3 there is a very significant positive skew (as was evident from the histogram). The kurtosis z -scores are: $-0.410/0.172 = -2.38$ on day 1, $0.822/0.299 = 2.75$ on day 2 and $0.732/0.433 = 1.69$ on day 3. These values indicate significant problems with skew, kurtosis or both (at $p < .05$) for all three days; however, because of the large sample, this isn’t surprising and so we can take comfort from the central limit theorem.

Another way of looking at the problem is to see whether the distribution of scores deviates from a comparable normal distribution. The **Kolmogorov–Smirnov test** and **Shapiro–Wilk test** do this: they compare the scores in the sample to a normally distributed set of scores with the same mean and standard deviation. If the test is non-significant ($p > .05$) it tells us that the distribution of the sample is not significantly different from a normal distribution (i.e., it is probably normal). If, however, the test is significant ($p < .05$) then the distribution in question is significantly different from a normal distribution (i.e., it is non-normal). These tests seem great: in one easy procedure they tell us whether our scores are normally distributed (nice!). However, Jane Superbrain

Box 5.5 explains some really good reasons not to use them. If you insist on using them, bear Jane’s advice in mind and always plot your data as well and try to make an informed decision about the extent of non-normality based on converging evidence.



Did someone say Smirnov? Great, I need a drink after all this data analysis!



CRAMMING SAM'S TIPS

Skewness and kurtosis

- To check that the distribution of scores is approximately normal, we need to look at the values of skewness and kurtosis in the output.
- Positive values of skewness indicate too many low scores in the distribution, whereas negative values indicate a build-up of high scores.
- Positive values of kurtosis indicate a pointy and heavy-tailed distribution, whereas negative values indicate a flat and light-tailed distribution.
- The further the value is from zero, the more likely it is that the data are not normally distributed.
- You can convert these scores to z-scores by dividing by their standard error. If the resulting score (when you ignore the minus sign) is greater than 1.96 then it is significant ($p < .05$).
- Significance tests of skew and kurtosis should not be used in large samples (because they are likely to be significant even when skew and kurtosis are not too different from normal).



OLIVER TWISTED

Please, Sir, can I have some more ... frequencies?

In your output you will also see tabulated frequency distributions of each variable. This table is reproduced in the additional online material along with a description.

The Kolmogorov–Smirnov (K-S; Figure 5.15) test is accessed through the *explore* command (**Analyze Descriptive Statistics** ▶ **Explore...**). Figure 5.16 shows the dialog boxes for this command. First, enter any variables of interest in the box labelled *Dependent List* by highlighting them on the left-hand side and transferring them by clicking on **▶**. For this example, select the hygiene scores for the three days. If you click on **Statistics...** a dialog box appears, but the default option is fine (it will produce means, standard deviations and so on). The more interesting option for our current purposes is accessed by clicking on **Plots...**. In this dialog box select the option **Normality plots with tests**, and this will produce both the K-S test and some *normal quantile–quantile (Q-Q) plots*. A **Q-Q plot** is very similar to the P-P plot that we encountered in Section 5.3.2 except that it plots the quantiles (Section 1.6.3) of the data instead of every individual score in the data. The expected quantiles are a straight diagonal line, whereas the observed quantiles are plotted as individual points. The Q-Q plot can be interpreted in the same way as a P-P plot: any deviation of the dots from the diagonal line represents a deviation from normality. Kurtosis is shown up by the dots sagging above or below the line, whereas skew is shown up by the dots snaking around the line in an ‘S’ shape. If you have a lot of scores, Q-Q plots can be easier to interpret than P-P plots because they will display fewer values.

By default, SPSS will produce boxplots (split according to group if a factor has been specified) and stem-and-leaf diagrams as well. We also need to click on **Options...** to tell SPSS how to deal with missing values. This is important because although we start off with 810 scores on

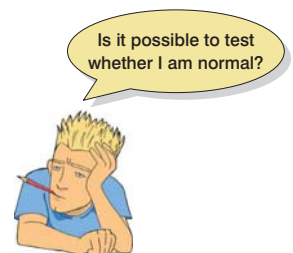


FIGURE 5.15

Andrei Kolmogorov, wishing he had a Smirnov

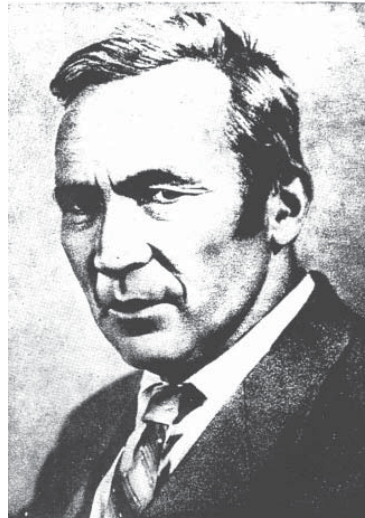
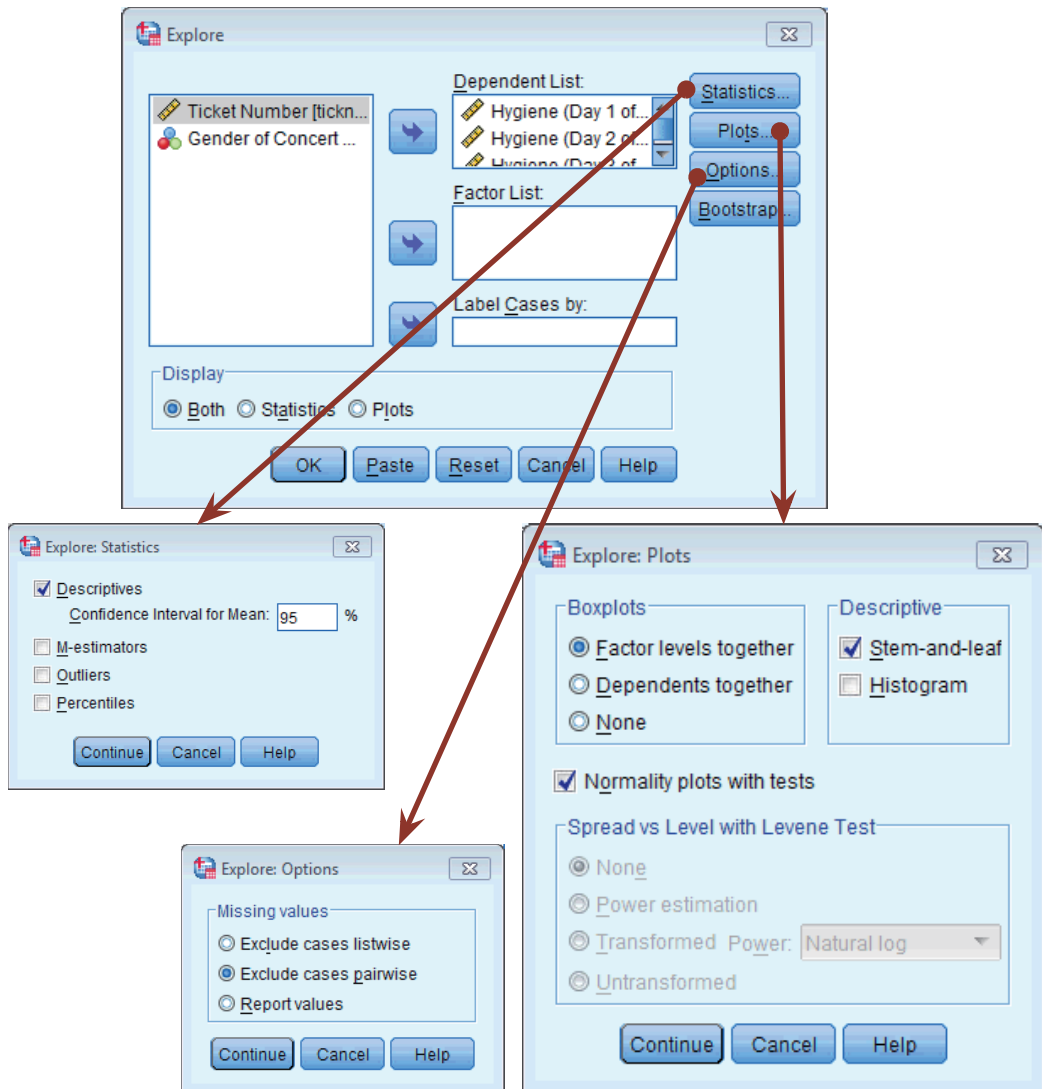


FIGURE 5.16

Dialog boxes for the *explore* command



day 1, by day 2 we have only 264 and by day 3 only 123. By default, SPSS will use only cases for which there are valid scores on all of the selected variables. This would mean that for day 1, even though we have 810 scores, it will use only the 123 cases for which there are scores on all three days. This is known as excluding cases *listwise*. However, we want it to use all of the scores it has on a given day, which is known as *pairwise*. There's more information on these two methods in SPSS Tip 5.1. Once you have clicked on **Options...**, select *Exclude cases pairwise*, then click on **Continue** to return to the main dialog box and click on **OK** to run the analysis.

SPSS will produce a table of descriptive statistics (mean, etc.) that should have the same values as the tables obtained using the frequencies procedure. The important table is that of the K-S test (Output 5.3). This table includes the test statistic itself, the degrees of freedom (which should equal the sample size) and the significance value of this test. Remember that a significant value (*Sig.* less than .05) indicates a deviation from normality. For day 1 the K-S test is just about non-significant ($p = .097$), which is surprisingly close to significant given how normal the day 1 scores looked in the histogram (Figure 5.13). However, the sample size on day 1 is very large ($N = 810$) and the significance of the K-S test for these data shows how in large samples even small and unimportant deviations from normality might be deemed significant by this test (Jane Superbrain Box 5.5). For days 2 and 3 the test is highly significant, indicating that these distributions are not normal, which is likely to reflect the skew seen in the histograms for these data (Figure 5.13).



SPSS TIP 5.1

Pairwise or listwise? ①

Many of the analyses in this book have additional options that can be accessed by clicking on **Options...**. Often the resulting *Options* dialog box will ask you if you want to exclude cases 'pairwise', 'analysis by analysis' or 'listwise'. Let's imagine we wanted to use our hygiene scores to compare mean scores on days 1 and 2, days 1 and 3, and days 2 and 3. First, we can exclude cases listwise, which means that if a case has a missing value for any variable, then they are excluded from the whole analysis. So, for example, if we had the hygiene score for a person (let's call her Melody) at the festival on days 1 and 2, but not day 3, then Melody's data will be excluded for all of the comparisons mentioned above. Even though we have her data for days 1 and 2, we won't use them for that comparison – *they would be completely excluded from the analysis*. Another option is to excluded cases on a *pairwise* (a.k.a. *analysis-by-analysis* or *test-by-test*) basis, which means that Melody's data will be excluded only for analyses for which she has missing data: so her data would be used to compare days 1 and 2, but would be excluded for the other comparisons (because we don't have her score on day 3).

Tests of Normality

	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Hygiene (Day 1 of Download Festival)	.029	810	.097	.996	810	.032
Hygiene (Day 2 of Download Festival)	.121	264	.000	.908	264	.000
Hygiene (Day 3 of Download Festival)	.140	123	.000	.908	123	.000

a. Lilliefors Significance Correction

OUTPUT 5.3



OLIVER TWISTED

Please, Sir, can I have some more ... normality tests?

'There is another test reported in the table (the Shapiro–Wilk test)', whispers Oliver as he creeps up behind you, knife in hand, 'and a footnote saying that the "Lilliefors significance correction" has been applied. What the hell is going on?' (If you do the K-S test through the Nonparametric Tests menu rather than the Explore menu this correction is not applied.) Well, Oliver, all will be revealed in the additional material for this chapter on the companion website: you can find out more about the K-S test, and information about the Lilliefors correction and Shapiro–Wilk test. What are you waiting for?

5.3.2.3. Reporting the K-S test ①

The test statistic for the K-S test is denoted by D , and we should report the degrees of freedom (df) from the table in brackets after the D . We can report the results in Output 5.3 in the following way:

- ✓ The hygiene scores on day 1, $D(810) = 0.029$, $p = .097$, did not deviate significantly from normal; however, day 2, $D(264) = 0.121$, $p < .001$, and day 3, $D(123) = 0.140$, $p < .001$, scores were both significantly non-normal.



CRAMMING SAM'S TIPS

Normality tests

- The K-S test can be used to see if a distribution of scores significantly differs from a normal distribution.
- If the K-S test is significant (*Sig.* in the SPSS table is less than .05) then the scores are significantly different from a normal distribution.
- Otherwise, scores are approximately normally distributed.
- The Shapiro–Wilk test does much the same thing, but it has more power to detect differences from normality (so this test might be significant when the K-S test is not).
- **Warning:** In large samples these tests can be significant even when the scores are only slightly different from a normal distribution. Therefore, I don't particularly recommend them and they should always be interpreted in conjunction with histograms, P-P or Q-Q plots, and the values of skew and kurtosis.

5.3.2.4. Normality within groups and the split file command ①

We saw earlier that when predictor variables are formed of categories, if you decide that you need to check the assumption of normality then you need to do it within each group separately (Jane Superbrain Box 5.1). For example, for the hygiene scores we have data for males and females (in the variable **Gender**). If we made some prediction about there being differences in hygiene between males and females at a music festival then we should look at normality within males and females separately. There are several ways to produce basic descriptive statistics for separate groups. First, I will introduce you to the *split file* function. This function allows you to specify a grouping variable (remember, these variables are used to specify categories of cases). Any subsequent procedure in SPSS is then carried out on *each category of cases separately*.

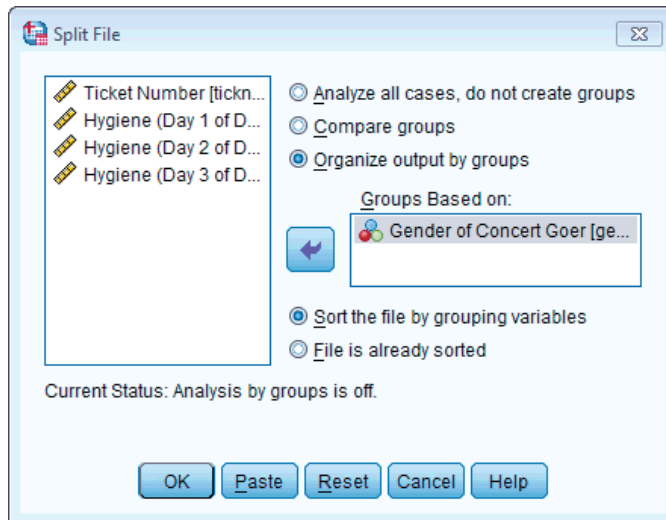
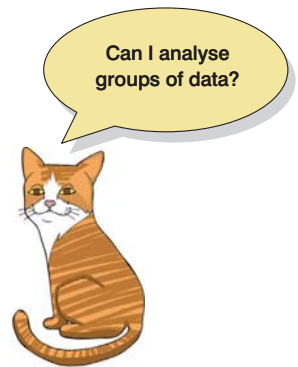


FIGURE 5.17
Split File dialog box

If we want to obtain separate descriptive statistics for males and females in our festival hygiene scores, we can split the file, and then proceed using the *frequencies* command described in the previous section. To split the file, select **Data** > **Split File...** or click on . In the resulting dialog box (Figure 5.17) select the option *Organize output by groups*. Once this option is selected, the *Groups Based on* box will activate. Select the variable containing the group codes by which you wish to repeat the analysis (in this example select **Gender**), and drag it to the box or click on . By default, SPSS will sort the file by these groups (i.e., it will list one category followed by the other in the data editor). Once you have split the file, use the *frequencies* command (see the previous section). Let's request statistics for all three days as in Figure 5.14.



Output 5.4 shows the results, which have been split into two tables: the results for males and the results for females. Males scored lower than females on all three days of the festival (i.e., they were smellier). The values of skew and kurtosis are similar for males and females on days 2 and 3, but differ a little on day 1: as already indicated, males show a very slight positive skew (0.200) but for females the skew is slightly negative (-0.176). In both cases the skew on day 1 is very small. Figure 5.18 shows the histograms of hygiene scores split according to the gender of the

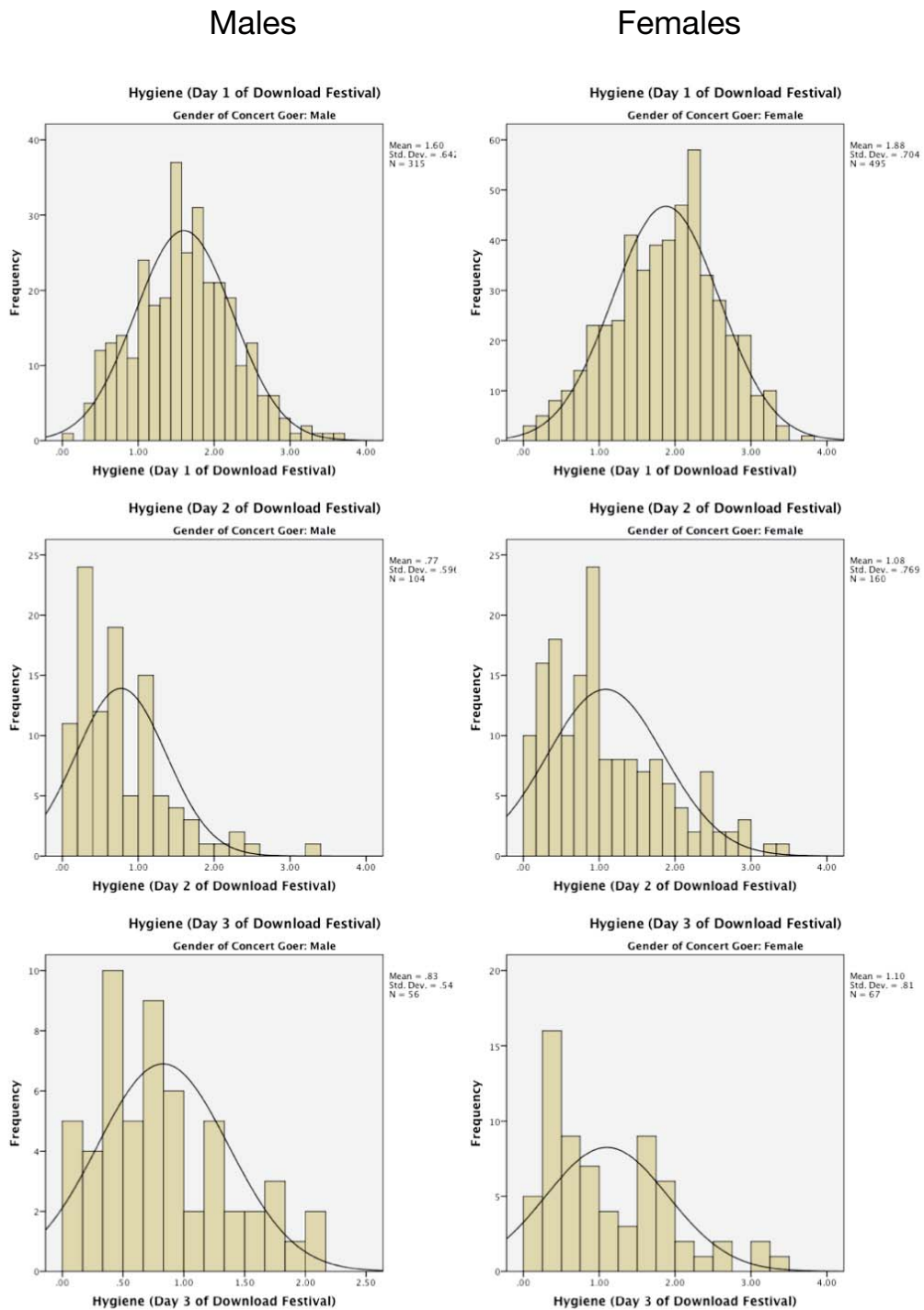
Male					Female				
Statistics ^a					Statistics ^a				
		Hygiene (Day 1 of Download Festival)	Hygiene (Day 2 of Download Festival)	Hygiene (Day 3 of Download Festival)			Hygiene (Day 1 of Download Festival)	Hygiene (Day 2 of Download Festival)	Hygiene (Day 3 of Download Festival)
N	Valid	315	104	56	N	Valid	495	160	67
	Missing	0	211	259		Missing	0	335	428
Mean		1.6021	.7733	.8291	Mean		1.8787	1.0829	1.0997
Std. Error of Mean		.03620	.05847	.07210	Std. Error of Mean		.03164	.06078	.09896
Median		1.5800	.6700	.7300	Median		1.9400	.8900	.8500
Mode		2.00	.23	.44	Mode		2.02	.85	.38
Std. Deviation		.64241	.59630	.53954	Std. Deviation		.70396	.76876	.81001
Variance		.413	.356	.291	Variance		.496	.591	.656
Skewness		.200	1.476	.719	Skewness		-.176	.870	.869
Std. Error of Skewness		.137	.237	.319	Std. Error of Skewness		.110	.192	.293
Kurtosis		-.101	3.134	-.268	Kurtosis		-.397	.089	.069
Std. Error of Kurtosis		.274	.469	.628	Std. Error of Kurtosis		.219	.381	.578
Range		3.47	3.35	2.09	Range		3.67	3.38	3.39
Minimum		.11	.00	.02	Minimum		.02	.06	.02
Maximum		3.58	3.35	2.11	Maximum		3.69	3.44	3.41

a. Gender of Concert Goer = Male

a. Gender of Concert Goer = Female

OUTPUT 5.4

FIGURE 5.18
Distributions of hygiene scores for males (left) and females (right) over three days (top to bottom) of a music festival



festival-goer. Male and female scores have similar distributions. On day 1 they are fairly normal (although females perhaps show a very slight negative skew, which indicates a higher proportion of them were at the higher end of hygiene scores than males). On days 2 and 3 both males and females show the characteristic positive skew that we saw in the sample as a whole. It looks as though proportionally more females are in the skewed end of the distribution (i.e., at the hygienic end).

We can also do K-S tests within the different groups by repeating the analysis we did earlier (Figure 5.16); because the *split file* command is switched on, we'd get the K-S test performed on males and females separately. An alternative method is to split the analysis by group from within the *explore* command itself. First, switch *split file* off by clicking on **Data** **Split File...** (or click on to activate the dialog box in Figure 5.17. Select *Analyze all cases, do not create groups* and click on **OK**. The *split file* function is now off and analyses will be conducted on the data as whole. Next, activate the *explore* command just as we did before: **Analyze** **Descriptive Statistics** **Explore...** We can ask for separate tests for males and females by placing **Gender** in the box labelled *Factor List* as in Figure 5.21 and selecting the same options as described earlier. Let's do this for the day 1 hygiene scores. You should see the table in Output 5.5, which shows that the distribution of hygiene scores was normal for males (the value of *Sig.* is greater than .05) but not for females (the value of *Sig.* is smaller than .05).

Tests of Normality

		Kolmogorov-Smirnov ^a			Shapiro-Wilk		
		Statistic	df	Sig.	Statistic	df	Sig.
Hygiene (Day 1 of Download Festival)	Male	.035	315	.200	.993	315	.119
	Female	.053	495	.002	.993	495	.029

^a. This is a lower bound of the true significance.
a. Lilliefors Significance Correction

OUTPUT 5.5

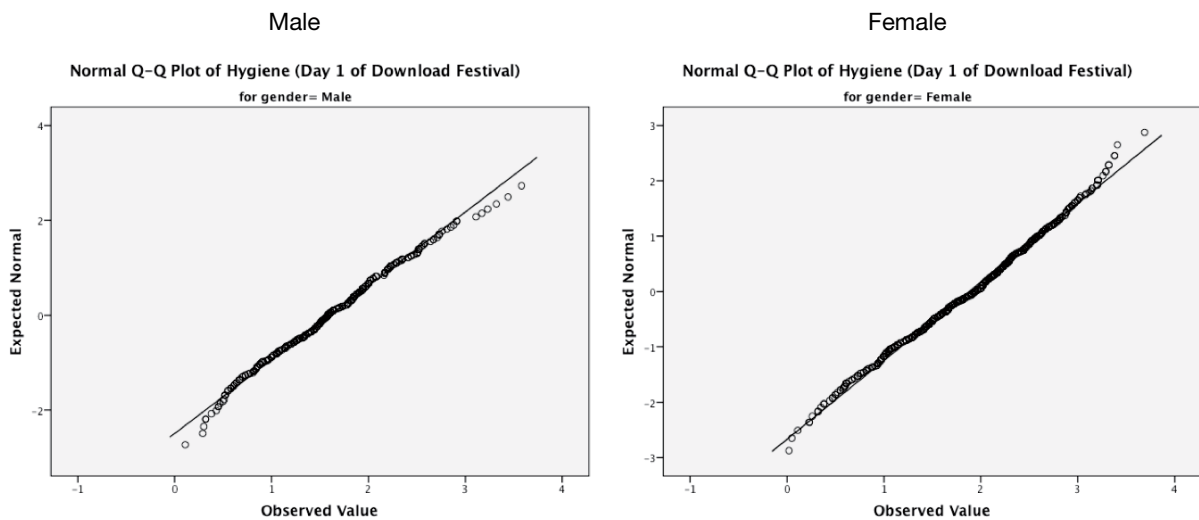


FIGURE 5.19 Normal Q-Q plots of hygiene scores for day 1 of the music festival

SPSS also produces a normal Q-Q plot (see Figure 5.19). Despite the K-S having completely different outcomes for males and females, the Q-Q plots are remarkably similar: there is no sign of a major problem with kurtosis (the dots do not particularly sag above or below the line) and there is some slight skew (the female graph in particular has a slight S-shape). However, both graphs show that the quantiles fall very close to the diagonal line, which, let's not forget, represents a perfect normal distribution. For the females the graph is at odds with the significant K-S test, and this illustrates my earlier point that if you have a large sample then tests like K-S will lead you to conclude that even very minor deviations from normality are 'significant'.



SELF-TEST Compute and interpret a K-S test and Q-Q plots for males and females for days 2 and 3 of the music festival.

5.3.3. Spotting linearity and heteroscedasticity/heterogeneity of variance ②

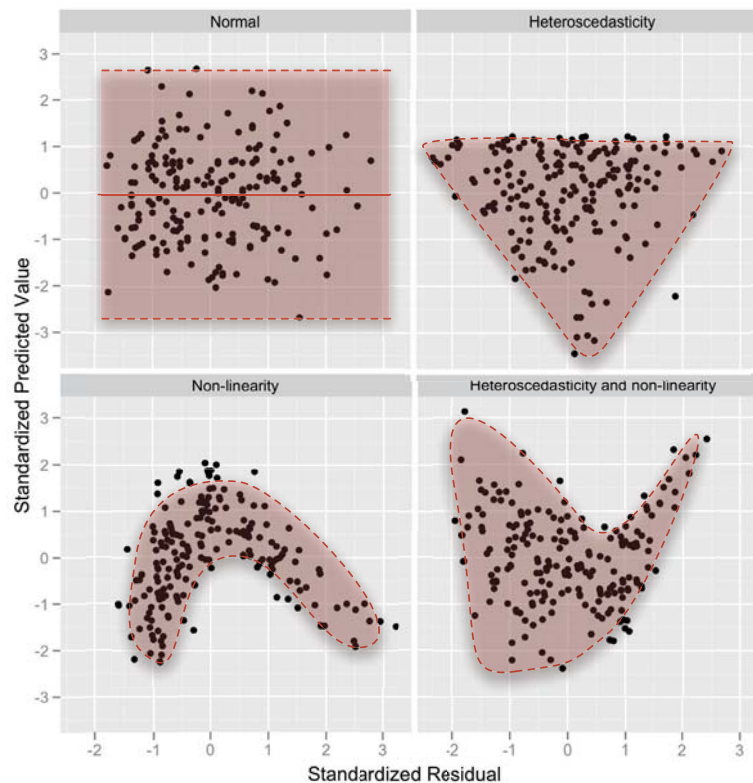
5.3.3.1. Using graphs to spot problems with linearity or homoscedasticity ②

It might seem odd that I have chosen to look at the assumption of linearity and homoscedasticity together. However, there is a graph that shows up problems with both of these assumptions. These assumptions both relate to the errors (a.k.a. residuals) in the model we fit to the data. We can create a scatterplot of the values of the residuals against the values of the outcome predicted by our model. In doing so we're looking at whether there is a systematic relationship between what comes out of the model (the predicted values) and the errors in the model. Normally we convert the predicted values and errors to z -scores,⁷ so this plot is sometimes referred to as z_{pred} vs. z_{resid} . If linearity and homoscedasticity hold true then there should be no systematic relationship between the errors in the model and what the model predicts. Looking at this graph can, therefore, kill two birds with one stone. If this graph funnels out, then the chances are that there is heteroscedasticity in the data. If there is any sort of curve in this graph then the chances are that the data have broken the assumption of linearity.

Figure 5.20 shows several examples of the plot of standardized residuals against standardized predicted values. The top left panel shows a situation in which the assumptions of linearity and homoscedasticity have been met. The top right panel shows a similar plot for a data

FIGURE 5.20

Plots of standardized residuals against predicted (fitted) values



⁷ These standardized errors are called standardized residuals, which we'll discuss in Chapter 8.

set that violates the assumption of homoscedasticity. Note that the points form a funnel: they become more spread out across the graph. This funnel shape is typical of heteroscedasticity and indicates increasing variance across the residuals. The bottom left panel shows a plot of some data in which there is a non-linear relationship between the outcome and the predictor: there is a clear curve in the residuals. Finally, the bottom right panel illustrates data that not only have a non-linear relationship, but also show heteroscedasticity. Note first the curved trend in the residuals, and then also note that at one end of the plot the points are very close together whereas at the other end they are widely dispersed. When these assumptions have been violated you will not see these exact patterns, but hopefully these plots will help you to understand the general anomalies you should look out for. We'll look at an example of how this graph is used in Chapter 8, but for the time being just be aware of the patterns to look out for.

5.3.3.2. Spotting heteroscedasticity/heterogeneity of variance using numbers ②

Remember that homoscedasticity/homogeneity of variance means that as you go through levels of one variable, the variance of the other should not change. If you've collected groups of data then this means that the variance of your outcome variable or variables should be the same in each of these groups. You'll sometimes come across **Levene's test** (Levene, 1960), which tests the null hypothesis that the variances in different groups are equal. It's a very simple and elegant test that works by doing a one-way ANOVA (see Chapter 11) on the deviation scores; that is, the absolute difference between each score and the mean of the group from which it came (see Glass, 1966, for a very readable explanation).⁸ For now, all you need to know is that if Levene's test is significant at $p \leq .05$ then you conclude that the null hypothesis is incorrect and that the variances are significantly different – therefore, the assumption of homogeneity of variances has been violated. If, however, Levene's test is non-significant (i.e., $p > .05$) then the variances are roughly equal and the assumption is tenable. Although Levene's test can be selected as an option in many of the statistical tests that require it, it's best to look at it when you're exploring data because it informs the model you fit. As with the K-S test (and other tests of normality), when the sample size is large, small differences in group variances can produce a Levene's test that is significant (Jane Superbrain Box 5.5). There are also other very strong arguments for not using it (Jane Superbrain Box 5.6).

Some people also look at **Hartley's F_{\max}** , also known as the **variance ratio** (Pearson & Hartley, 1954). This is the ratio of the variances between the group with the biggest variance and the group with the smallest variance. This ratio was compared to critical values in a table published by Hartley. Although this ratio isn't used very often, if you want the critical values (for a .05 level of significance) see *Oliver Twisted*. The critical values depend on the number of cases per group, and the number of variances being compared. For example, with sample sizes (n) of 10 per group, an F_{\max} of less than 10 is more or less always going to be non-significant, with 15–20 per group the ratio needs to be less than about 5, and with samples of 30–60 the ratio should be below about 2 or 3.

5.3.3.3. If you still decide to do Levene's test ②

We can get Levene's test using the *Explore* menu that we used in the previous section. Sticking with the hygiene scores, we'll compare the variances of males and females on day 1

⁸ We haven't covered ANOVA yet, so this explanation won't make much sense to you now, but in Chapter 11 we will look in more detail at how Levene's test works.



JANE SUPERBRAIN 5.6

Is Levene's test worth the effort? ②

Statisticians used to recommend testing for homogeneity of variance using Levene's test and, if the assumption was violated, using an adjustment to correct for it. However, people have stopped using this approach for two reasons. First, when you have violated this assumption it only matters if you have unequal group sizes: if

you don't have unequal group sizes, this assumption is pretty much irrelevant, and can be ignored. Second, the tests of homogeneity of variance like Levene's tend to work very well when you have equal group sizes and large samples (when it doesn't matter as much if you have violated the assumption) and don't work as well with unequal group sizes and smaller samples (which is exactly when it does matter). Plus, there are adjustments to correct for violations of this assumption that can often be applied (as we shall see) which would be a right nuisance if you had to do them by hand, but are very easy to do if you have a computer. In most cases, if you have violated the assumption then a correction is made – and if you haven't violated the assumption, a correction is not made. So, you might as well always do the adjustment and forget about the assumption. If you're really interested in this, I like the article by Zimmerman (2004).



OLIVER TWISTED

Please, Sir, can I have some more ... Hartley's F_{max} ?

Oliver thinks that it's stupid to talk about the variance ratio without the critical values. 'No critical values?' he laughed. 'That's the most stupid thing I've seen since I was at Sussex Uni and I saw my statistics lecturer, Andy Fie...'. Well, go choke on your gruel, you Dickensian bubo, because the full table of critical values is in the additional material for this chapter on the companion website.

of the festival. Use **Analyze Descriptive Statistics** > **Explore...** to open the dialog box in Figure 5.21. Transfer the **day1** variable from the list on the left-hand side to the box labelled **Dependent List** by clicking on the **▶** next to this box; because we want to split the output by the grouping variable to compare the variances, select the variable **Gender** and transfer it to the box labelled **Factor List** by clicking on the appropriate **▶**. Then click on **Plots...** to open the other dialog box in Figure 5.21. To get Levene's test we need to select one of the options where it says *Spread vs. level with Levene test*. If you select **Untransformed**, Levene's test is carried out on the raw data (a good place to start). When you've finished with this dialog box click on **Continue** to return to the main **Explore** dialog box and then click on **OK** to run the analysis.

Output 5.6 shows the table for Levene's test. The test can be based on differences between scores and the mean, and between scores and the median. The median is slightly preferable (because it is less biased by outliers). When using both the mean ($p = .030$) and the median ($p = .037$) the significance values are less than .05, indicating a significant difference between the male and female variances. To calculate the variance ratio, we need to divide the largest variance by the smallest. You should find the variances in your output, but if not, we obtained these values in Output 5.4. The male variance was 0.413 and the female one 0.496; the variance ratio is, therefore, $0.496/0.413 = 1.2$. In essence the variances are practically equal. So, why does Levene's test tell us they are significantly different? The answer is because the sample sizes are so large: we had 315 males and 495 females, so

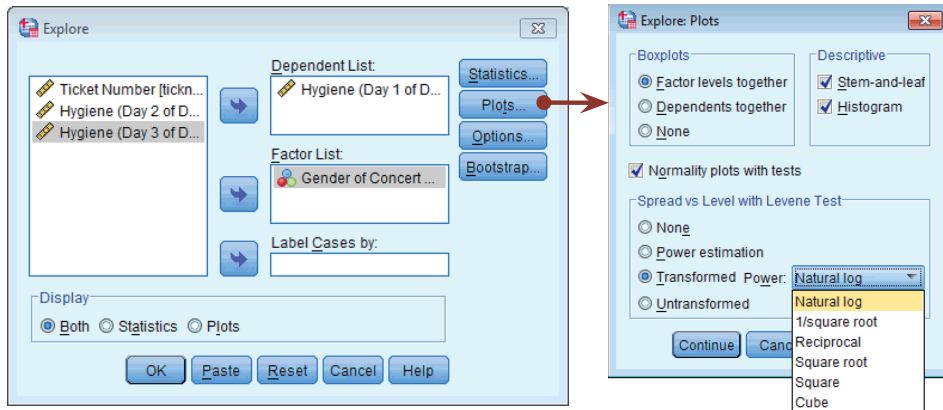


FIGURE 5.21 Exploring groups of data and obtaining Levene’s test

Test of Homogeneity of Variance

		Levene Statistic	df1	df2	Sig.
Hygiene (Day 1 of Download Festival)	Based on Mean	4.736	1	808	.030
	Based on Median	4.354	1	808	.037
	Based on Median and with adjusted df	4.354	1	805.066	.037
	Based on trimmed mean	4.700	1	808	.030

OUTPUT 5.6

even this very small difference in variances is shown up as significant by Levene’s test (Jane Superbrain Box 5.5). Hopefully this example convinces you to treat these tests cautiously.

5.3.3.4. Reporting Levene’s test ①

Levene’s test can be denoted by the letter F and there are two different degrees of freedom. As such you can report it, in general form, as $F(df_1, df_2) = \text{value}, p = p\text{-value}$. So, for the results in Output 5.6 we could say:

- ✓ For the hygiene scores on day 1 of the festival, the variances were unequal for for males and females, $F(1, 808) = 4.74, p = .03$.



CRAMMING SAM’S TIPS

Homogeneity of variance

- Homogeneity of variance/homoscedasticity is the assumption that the spread of outcome scores is roughly equal at different points on the predictor variable.
- This can be tested by looking at a plot of the standardized predicted values from your model against the standardized residuals (z_{pred} vs. z_{resid}).
- When comparing groups, this assumption can be tested with Levene’s test and the variance ratio (Hartley’s F_{max}).
 - If Levene’s test is significant ($Sig.$ in the SPSS table is less than .05) then the variances are significantly different in different groups.
 - Otherwise, homogeneity of variance can be assumed.
 - The variance ratio is the largest group variance divided by the smallest. This value needs to be smaller than the critical values in the additional material.
- **Warning:** There are good reasons not to use tests like Levene’s test. In large samples Levene’s test can be significant even when group variances are not very different. Therefore, it should be interpreted in conjunction with the variance ratio.



5.4. Reducing bias ②

Having looked at potential sources of bias, the next issue is how to reduce the impact of bias. Essentially there are four methods for correcting problems with the data, which can be remembered with the handy acronym of TWAT (or WATT, if you prefer):

- **Trim the data:** Delete a certain amount of scores from the extremes.
- **Winsorizing:** Substitute outliers with the highest value that isn't an outlier.
- **Analyse with robust methods:** This typically involves a technique known as bootstrapping.
- **Transform the data:** This involves applying a mathematical function to scores to try to correct any problems with them.

Probably the best of these choices is to use **robust tests**, which is a term applied to a family of procedures to estimate statistics that are reliable even when the normal assumptions of the statistic are not met (Section 5.4.3). Let's look at each technique in more detail.

5.4.1. Trimming the data ②

Trimming the data means deleting some scores from the extremes, and it takes many forms. In its simplest form it could be deleting the data from the person who contributed the outlier. However, this should be done only if you have good reason to believe that this case is not from the population that you intended to sample. For example, if you were investigating factors that affected how much cats purr and one cat didn't purr at all, this would likely be an outlier (all cats purr). Upon inspection, if you discovered that this cat was actually a dog wearing a cat costume, then you'd have grounds to exclude this case because it comes from a different population (dogs who like to dress as cats) than your target population (cats).

More often, trimming involves removing extreme scores using one of two rules: (1) a percentage based rule; and (2) a standard deviation based rule. A percentage based rule would be, for example, deleting the 10% of highest and lowest scores. Let's look at an example. Meston and Frohlich (2003) report a study showing that heterosexual people rate a picture of someone of the opposite sex as more attractive after riding a roller coaster compared to before. Imagine we took 20 people as they came off of the Rockit rollercoaster at Universal Studios in Orlando⁹ and asked them to rate the attractiveness of someone in a photograph on a scale of 0 (looks like Jabba the Hutt) to 10 (my eyes have just exploded because they weren't designed to gaze upon such beauty). Figure 5.22 shows these scores. As you can see, most people gave ratings above the mid-point of the scale: they were pretty positive in their ratings. However, there were two people who gave zeros. If we were to trim 5% of the data from either end, this would mean deleting one score at each extreme (there are 20 scores and 5% of 20 is 1). Figure 5.22 shows that this involves deleting a 0 and an 8. We could compute a 5% trimmed mean by working out the mean for this trimmed data set. Similarly, Figure 5.22 shows that with 20 scores, a 10% trim would mean deleting two scores from each extreme, and a 20% trim would entail deleting four scores from each extreme. If you take trimming to its extreme then you get the median, which is the value left when you have trimmed all but the middle score. If we calculate the

⁹ I have a video of my wife and me on this rollercoaster during our honeymoon. I swear quite a lot on it, but I might stick it on my YouTube channel so you can laugh at what a cissy I am.



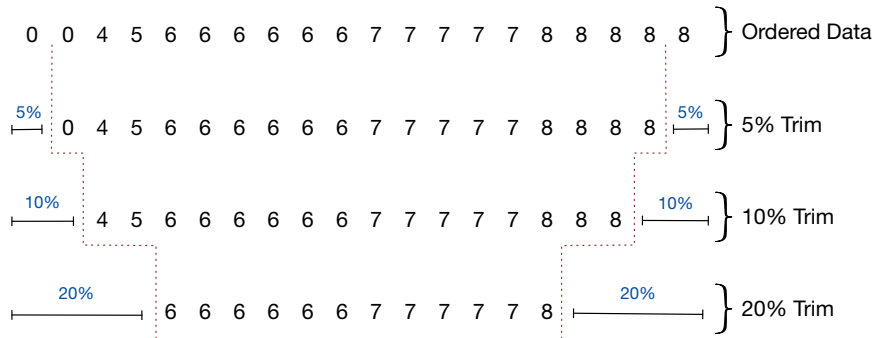


FIGURE 5.22
Illustration of
trimmed data

mean in a sample that has been trimmed in this way, it is called (unsurprisingly) a **trimmed mean**. A similar robust measure of location is the **M-estimator**, which differs from a trimmed mean in that the amount of trimming is determined empirically. In other words, rather than the researcher deciding before the analysis how much of the data to trim, an M-estimator determines the optimal amount of trimming necessary to give a robust estimate of, say, the mean. This has the obvious advantage that you never over- or under-trim your data. However, the disadvantage is that it is not always possible to reach a solution.



SELF-TEST Compute the mean and variance of the attractiveness ratings. Now compute them for the 5%, 10% and 20% trimmed data.

If you do the self-test you should find that the mean rating was 6 with a variance of 5.37. The 5% trimmed mean is 6.22, the 10% trimmed mean is 6.50, and the 20% trimmed mean is 6.58. The means get higher in this case because the trimming is reducing the impact of the few scores that were very small (two miserable people who gave ratings of 0). What happens to the variances? For the overall sample it is 5.37, but for the 5%, 10%, and 20% trimmed data you get 3.59, 1.20 and 0.45, respectively. The variances get smaller (and more stable) because, again, the outliers have less impact. We saw earlier that the accuracy of the mean and variance depends on a symmetrical distribution, but a trimmed mean (and variance) will be relatively accurate even when the distribution is not symmetrical, because by trimming the ends of the distribution we remove outliers and skew that bias the mean. Some robust methods work by taking advantage of the properties of the trimmed mean.

Standard deviation based rules involve calculating the mean and standard deviation of a set of scores, and then removing values that are a certain number of standard deviations greater than the mean. For example, when analysing reaction time data (which is notoriously messy) it is very common to remove any reaction times greater than (or below) 2.5 standard deviations above the mean (Ratcliff, 1993). For the roller coaster data the standard deviation is 2.32, so 2.5 times the standard deviation is 5.8. The mean was 6, therefore, we would delete scores greater than $6 + 5.8 = 11.8$, of which there were none (it was only a 10-point scale); we would also delete scores less than $6 - 5.8 = 0.2$, which means deleting the two scores of zero because they are the only scores less than 0.2. If we recalculate the mean excluding these two zeros we get 6.67 and a variance of 1.29. Again, you can see that this method reduces the impact of extreme scores. However, there is one fundamental problem with standard deviation based trimming, which is that the mean and standard deviation are both highly influenced by outliers (see Section 5.2.2); therefore, if you have outliers in the data the criterion you use to reduce their impact has already been biased by them.

When it comes to implementing these methods in SPSS, there isn't a simple way to do it. Although SPSS will calculate a 5% trimmed mean for you if you use the *explore* command (Figure 5.16), it won't remove the actual cases from the data set, so to do tests based on a trimmed sample you would need to manually trim the data (or do it using syntax commands) or use the *select cases* command (see Oditi's Lantern).

5.4.2. Winsorizing ①

Winsorizing the data involves replacing outliers with the next highest score that is *not* an outlier. It's perfectly natural to feel uncomfortable at the idea of changing the scores you collected to different values. It feels a bit like cheating. However, you need to bear in mind that if the score you're changing is very unrepresentative of the sample as a whole and biases your statistical model then it's not cheating at all; it's improving your accuracy.¹⁰ What *is* cheating is not dealing with extreme cases that bias the results in favour of your hypothesis, or changing scores in a systematic way other than to reduce bias (again, perhaps to support your hypothesis).

There are some subtle variations on winsorizing, such as replacing extreme scores with a score 3 standard deviations from the mean. A z -score of 3.29 constitutes an outlier (see 5.3.1) so we can calculate what score would give rise to a z -score of 3.29 (or perhaps 3) by rearranging the z -score equation, which gives us $X = (z \times s) + \bar{X}$. All we're doing is calculating the mean (\bar{X}) and standard deviation (s) of the data and, knowing that z is 3 (or 3.29 if you want to be exact), adding three times the standard deviation to the mean and replacing our outliers with that score. As with trimming, this is something you would need to do manually in SPSS or use the *select cases* command (see Oditi's Lantern).



ODITI'S LANTERN

Select Cases

'I, Oditi, believe that those who would try to prevent our cult from discovering the truths behind the numbers have placed dead herrings within the data. These rotting numerical fish permeate our models and infect the nostrils of understanding with their putrid stench. We must banish them; we must select only the good data, the pure data, the data uncontaminated by piscene putrefaction. You, the trooper of truth, must stare into my lantern to discover how to select cases using SPSS.'

5.4.3. Robust methods ③

By far the best option if you have irksome data (other than throwing your hands in the air and having a good scream) is to use a test that is robust to violations of assumptions and outliers. In other words, tests that are relatively unaffected by irksome data. The first set of tests are ones that do not rely on the assumption of normally distributed data (see

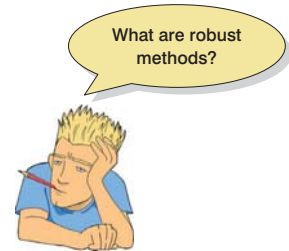
¹⁰ It is worth making the point that having outliers is interesting in itself, and if you don't think they represent the population then you need to ask yourself why they are different. The answer to the question might be a fruitful topic of more research.

Chapter 6).¹¹ One thing that you will quickly discover about non-parametric tests is that they have been developed for only a fairly limited range of situations. So, happy days if you want to compare two means, but sad and lonely days listening to Joy Division if you have a complex experimental design.

A much more promising approach is to use robust methods, which I mentioned earlier. These tests have developed as computers have got more sophisticated (doing these tests without computers would be only marginally less painful than ripping off your skin and diving into a bath of salt). How these tests work is beyond the scope of this book (and my brain), but two simple concepts will give you the general idea. The first we have already looked at: robust measures of the centre of the distribution such as the trimmed mean and M-estimators. The second is the **bootstrap** (Efron & Tibshirani, 1993), which is a very simple and elegant idea. The problem that we have is that we don't know the shape of the sampling distribution, but normality in our data allows us to infer that the sampling distribution is normal (and hence we can know the probability of a particular test statistic occurring). Lack of normality prevents us from knowing the shape of the sampling distribution unless we have big samples. Bootstrapping gets around this problem by estimating the properties of the sampling distribution from the sample data. Figure 5.23 illustrates the process: in effect, the sample data are treated as a population from which smaller samples (called bootstrap samples) are taken (putting each score back before a new one is drawn from the sample). The parameter of interest (e.g., the mean) is calculated in each bootstrap sample. This process is repeated perhaps 2000 times. The end result is that we have 2000 parameter estimates, one from each bootstrap sample. There are two things we can do with these estimates: the first is to order them and work out the limits within which 95% of them fall. For example, in Figure 5.23, 95% of bootstrap sample means fall between 2 and 9. We can use these values as an estimate of the limits of the 95% confidence interval of the parameter. The result is known as a percentile bootstrap confidence interval (because it is based on the values between which 95% of bootstrap sample estimates fall). The second thing we can do is to calculate the standard deviation of the parameter estimates from the bootstrap samples and use it as the standard error of parameter estimates. Therefore, when we use bootstrapping, we're effectively getting the computer to use our sample data to mimic the sampling process described in Section 2.5. An important point to remember is that because bootstrapping is based on taking random samples from the data you've collected, the estimates you get will be slightly different every time. This is nothing to worry about. For a fairly gentle introduction to the concept of bootstrapping, see Wright, London, and Field (2011).

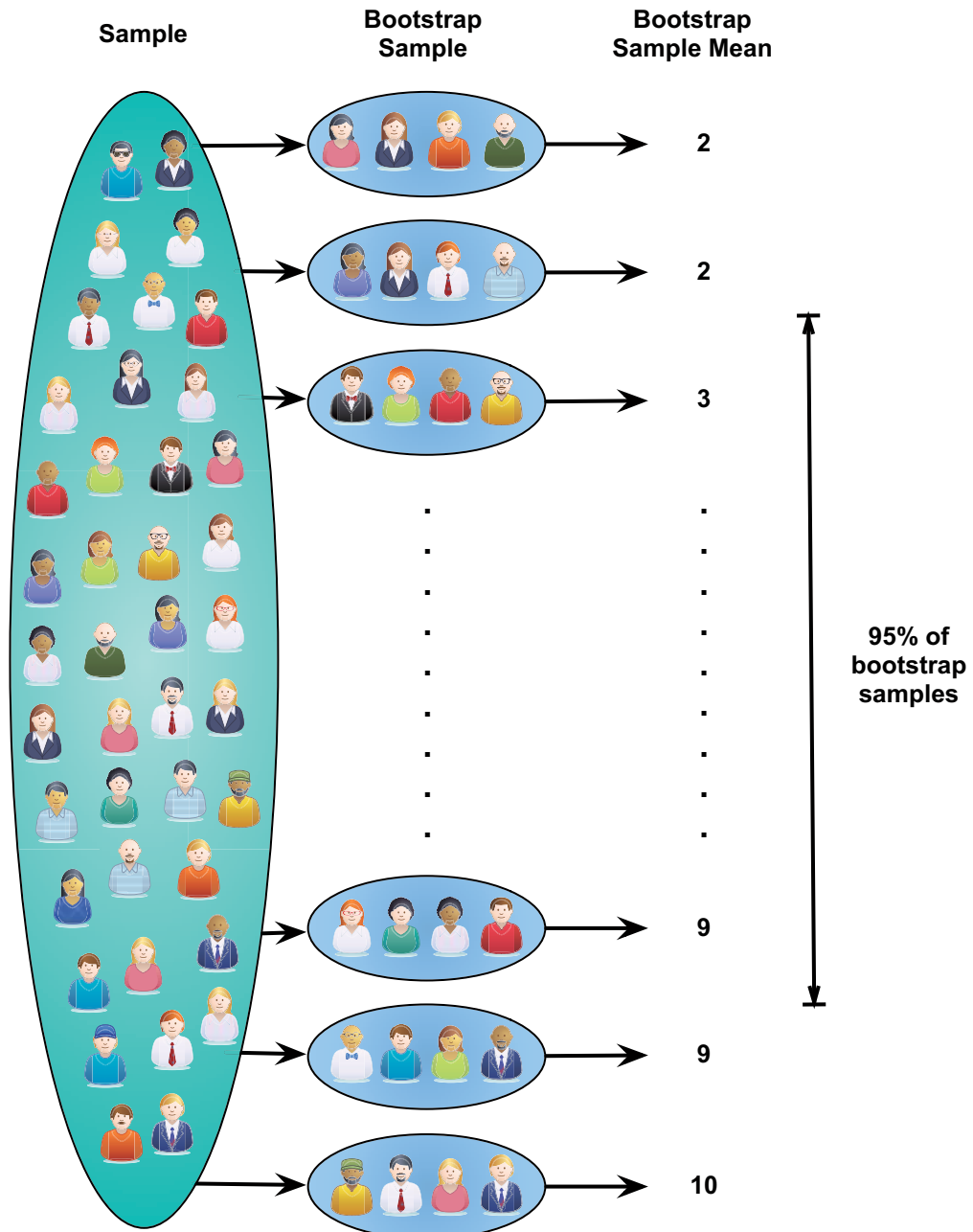
SPSS implements bootstrapping in some contexts, which we'll encounter as we go through various chapters. Some procedures have a bootstrap option, which can be accessed by clicking on **Bootstrap...** to activate the dialog box in Figure 5.24 (see Odit's Lantern). Select **Perform bootstrapping** to activate bootstrapping for the procedure you're currently doing. In terms of the options, SPSS will compute a 95% percentile confidence interval (**Percentile**), but you can change the method to a slightly more accurate one (Efron & Tibshirani, 1993) called a bias corrected and accelerated confidence interval (**Bias corrected accelerated (BCa)**). You can also change the confidence level by typing a number other than 95 in the box labelled **Level(%)**. By default, SPSS uses 1000 bootstrap samples, which is a reasonable number, and you certainly wouldn't need to use more than 2000.

There are versions of common procedures such as ANOVA, ANCOVA, correlation and multiple regression based on trimmed means and bootstrapping that enable you to ignore



¹¹ For convenience a lot of textbooks refer to these tests as *non-parametric tests* or *assumption-free tests* and stick them in a separate chapter. Actually neither of these terms is particularly accurate (none of these tests is assumption-free), but in keeping with tradition I've put them in Chapter 6, on their own, feeling lonely and ostracized from their 'parametric' counterparts.

FIGURE 5.23
Illustration of
the percentile
bootstrap



everything we have discussed about bias in this chapter. That’s a happy story, but one with a tragic ending because you can’t implement them directly in SPSS. The definitive guide to these tests is Wilcox’s (2012) outstanding book. Thanks to Wilcox, these tests can be implemented using a free statistics program called R (www.r-project.org). There is a plug-in for SPSS that enables you to use R via the SPSS interface, but it’s fiddly to get working and once it is working all it really does is allow you to type the commands that you would type into R. Therefore, I find it much easier just to use R. If you want to go down that route, then I have written a version of this textbook for R that covers these robust tests in some detail (Field, Miles, & Field, 2012). (Sorry, that was a shameless plug.)

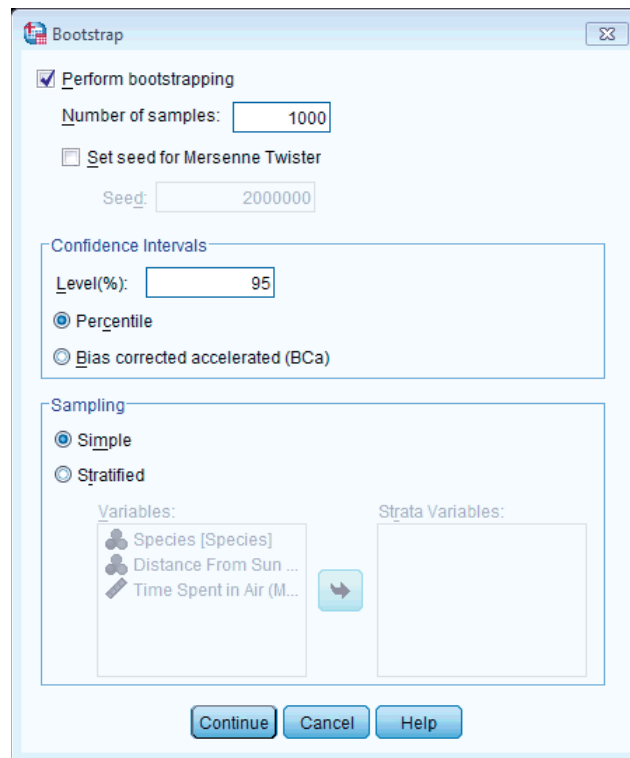


FIGURE 5.24
Dialog box for
the standard
bootstrap



ODITI'S LANTERN

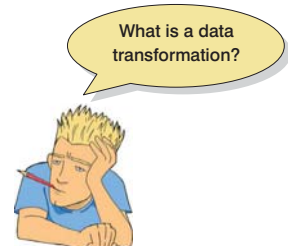
Bootstrapping

'I, Odit, believe that R is so-called because it makes you shout "Arrghhh!!?" You, my followers, are precious to me and I would not want you to place your sensitive body parts into that guillotine. Instead, stare into my lantern to see how we can use bootstrapping within SPSS.'

5.4.4. Transforming data ②

The final thing that you can do to combat problems with normality and linearity is to transform your data. The idea behind transformations is that you do something to every score to correct for distributional problems, outliers, lack of linearity or unequal variances. Although some students often (understandably) think that transforming data sounds dodgy (the phrase 'fudging your results' springs to some people's minds!), in fact it isn't because you do the same thing to all of your scores. As such, transforming the data changes the form of the relationships between variables but the relative differences between people for a given variable stay the same, so we can still quantify those relationships. However, it does change the differences between different variables (because it changes the units of measurement). Therefore, if you are looking at relationships between variables (e.g., regression) just transform the problematic variable, but if you are looking at differences between variables (e.g., change in a variable over time) then you need to transform all of those variables.

For example, our festival hygiene data were not normal on days 2 and 3 of the festival. Now, we might want to look at how hygiene levels changed across the three days (i.e.,





JANE SUPERBRAIN 5.7

*To transform or not to transform,
that is the question* ③

Not everyone thinks that transforming data is a good idea: Glass, Peckham, and Sanders (1972) commented in a review that 'the payoff of normalizing transformations in terms of more valid probability statements is low, and they are seldom considered to be worth the effort' (p. 241). The issue is quite complicated (especially for this early in the book), but essentially we need to know whether the statistical models we apply perform better on transformed data than they do when applied to data that violate the assumption that the transformation corrects. The question of whether to transform is linked to what test you are performing on your data and whether it is robust (see Section 5.4).

A good case in point is the F -test in ANOVA (see Chapter 11), which is often claimed to be robust (Glass et al., 1972). Early findings suggested that F performed as it should in skewed distributions and that transforming the data helped as often as it hindered the accuracy of F (Games & Lucas, 1966). However, in a lively but informative exchange, Levine and Dunlap (1982) showed that transformations of skew did improve the performance of F . In response, Games (1983) argued

that their conclusion was incorrect, which Levine and Dunlap (1983) contested in a response to the response. Finally, in a response to the response to the response, Games (1984) pointed out several important issues:

- 1 As we've seen, the central limit theorem (Section 5.2.4.2) tells us that in large samples the sampling distribution will be normal regardless. Lots of early research did show that with samples of 40 the sampling distribution was, as predicted, normal. However, this research focused on distributions with light tails, and with heavy-tailed distributions larger samples would be necessary to invoke the central limit theorem (Wilcox, 2012). Transformations might be useful for such distributions.
- 2 By transforming the data you change the hypothesis being tested (when using a log transformation and comparing means you change from comparing arithmetic means to comparing geometric means). Transformation also means that you're now addressing a different construct to the one originally measured, and this has obvious implications for interpreting that data (Grayson, 2004).
- 3 In small samples it is tricky to determine normality one way or another (see Jane Superbrain Box 5.5).
- 4 The consequences for the statistical model of applying the 'wrong' transformation could be worse than the consequences of analysing the untransformed scores.

Given these issues, unless you're correcting for a lack of linearity I would use robust procedures, where possible, in preference to transforming the data.

compare the mean on day 1 to the means on days 2 and 3 to see if people got smellier). The data for days 2 and 3 were skewed and need to be transformed, but because we might later compare the data to scores on day 1, we would also have to transform the day 1 data (even though scores were not skewed). If we don't change the day 1 data as well, then any differences in hygiene scores we find from day 1 to day 2 or 3 will be due to us transforming one variable and not the others. However, if we were going to look at the relationship between day 1 and day 2 scores (not the difference between them) we could transform only the day 2 scores and leave the day 1 scores alone.

5.4.4.1. Choosing a transformation ②

There are various transformations that you can do to the data that are helpful in correcting various problems. However, whether these transformations are necessary or useful is quite a complex issue (see Jane Superbrain Box 5.7).¹² Nevertheless, because they *are* used,

¹² Although there aren't statistical consequences of transforming data, there may be empirical or scientific implications that outweigh the statistical benefits (see Jane Superbrain Box 5.7).

TABLE 5.1 Data transformations and their uses

Data Transformation	Can Correct For
<p>Log transformation ($\log(X_i)$): Taking the logarithm of a set of numbers squashes the right tail of the distribution. As such it's a good way to reduce positive skew. This transformation is also very useful if you have problems with linearity (it can sometimes make a curvilinear relationship linear). However, you can't get a log value of zero or negative numbers, so if your data tend to zero or produce negative numbers you need to add a constant to all of the data before you do the transformation. For example, if you have zeros in the data then do $\log(X_i + 1)$, or if you have negative numbers add whatever value makes the smallest number in the data set positive.</p>	<p>Positive skew, positive kurtosis, unequal variances, lack of linearity</p>
<p>Square root transformation ($\sqrt{X_i}$): Taking the square root of large values has more of an effect than taking the square root of small values. Consequently, taking the square root of each of your scores will bring any large scores closer to the centre – rather like the log transformation. As such, this can be a useful way to reduce positive skew; however, you still have the same problem with negative numbers (negative numbers don't have a square root).</p>	<p>Positive skew, positive kurtosis, unequal variances, lack of linearity</p>
<p>Reciprocal transformation ($1/X_i$): Dividing 1 by each score also reduces the impact of large scores. The transformed variable will have a lower limit of 0 (very large numbers will become close to 0). One thing to bear in mind with this transformation is that it reverses the scores: scores that were originally large in the data set become small (close to zero) after the transformation, but scores that were originally small become large after the transformation. For example, imagine two scores of 1 and 10; after the transformation they become $1/1 = 1$, and $1/10 = 0.1$: the small score becomes larger than the large score after the transformation. However, you can avoid this by reversing the scores before the transformation, by finding the highest score and changing each score to the highest score minus the score you're looking at. So, you do a transformation $1/(X_{\text{Highest}} - X_i)$. Like the log transformation, you can't take the reciprocal of 0 (because $1/0 = \text{infinity}$) so if you have zeros in the data you need to add a constant to all scores before doing the transformation.</p>	<p>Positive skew, positive kurtosis, unequal variances</p>
<p>Reverse score transformations: Any one of the above transformations can be used to correct negatively skewed data, but first you have to reverse the scores. To do this, subtract each score from the highest score obtained, or the highest score + 1 (depending on whether you want your lowest score to be 0 or 1). If you do this, don't forget to reverse the scores back afterwards, or to remember that the interpretation of the variable is reversed: large scores have become small and small scores have become large.</p>	<p>Negative skew</p>

Table 5.1 shows some common transformations and their uses.¹³ The way to decide which transformation to use is by good old fashioned trial and error: try one out, see if it helps and if it doesn't then try a different one.

Trying out different transformations can be quite time-consuming. However, if heterogeneity of variance is your issue then we can see the effect of a transformation quite quickly. In Section 5.3.3.3 we saw how to use the *explore* function to get Levene's test. In

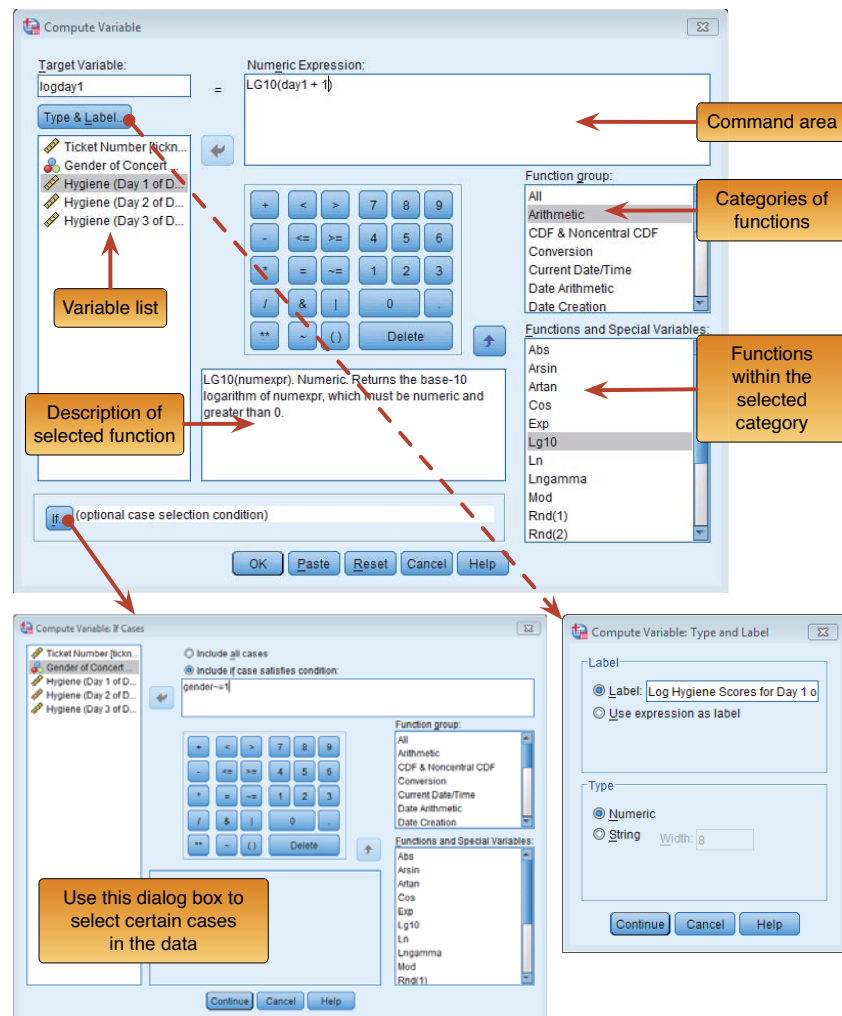
¹³ You'll notice in this section that I keep writing X_i . We saw in Chapter 1 that this refers to the observed score for the i th person (so the i could be replaced with the name of a particular person, thus for Graham, $X_i = X_{\text{Graham}}$ is Graham's score, and for Carol, $X_i = X_{\text{Carol}}$ is Carol's score).



that section we ran the analysis selecting the raw scores (*Untransformed*). However, if the variances turn out to be unequal, as they did in our example, you can use the same dialog box (Figure 5.21) but select *Transformed*. When you do this you should notice a drop-down list that becomes active and if you click on this you'll notice that it lists several transformations including the ones that I have just described. If you select a transformation from this list (*Natural log* perhaps or *Square root*) then SPSS will calculate what Levene's test would be if you were to transform the data using this method. This can save you a lot of time trying out different transformations.

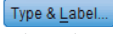

5.4.4.2. The *compute* function ②

To do transformations on SPSS we use the *compute* command, which enables us to carry out functions (such as adding or multiplying) on columns of data in the data editor. To access the *Compute Variable* dialog box, select **Transform** *Compute Variable...*. Figure 5.25 shows the main dialog box; it has a list of functions on the right-hand side, a calculator-like keyboard in the centre and a blank space that I've labelled the command area. You type a name for a new variable in the area labelled *Target Variable* and then you write some kind of








FIGURE 5.25
Compute Variable dialog box command











command in the command area to tell SPSS how to create this new variable. You use a combination of existing variables selected from the list on the left, and numeric expressions. So, for example, you could use it like a calculator to add variables (i.e., add two columns in the data editor to make a third). However, you can also use it to generate data without using existing variables too. There are hundreds of built-in functions that SPSS has grouped together. In the dialog box these groups are listed in the area labelled *Function group*; upon selecting a function group, a list of available functions within that group will appear in the box labelled *Functions and Special Variables*. If you select a function, then a description of that function appears in the white box indicated in Figure 5.25. You can enter variable names into the command area by selecting the variable required from the variables list and then clicking on . Likewise, you can select a certain function from the list of available functions and enter it into the command area by clicking on .

First type a variable name in the box labelled *Target Variable*, then click on  and another dialog box appears, where you can give the variable a descriptive label and specify whether it is a numeric or string variable (see Section 3.5.2). When you have written your command for SPSS to execute, click on  to run the command and create the new variable. If you type in a variable name that already exists in the data editor then SPSS will tell you and ask you whether you want to replace this existing variable. If you respond with *Yes* then SPSS will replace the data in the existing column with the result of the *compute* command; if you respond with *No* then nothing will happen and you will need to rename the target variable. If you're computing a lot of new variables it can be quicker to use syntax (see SPSS Tip 5.2).


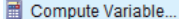
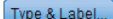


Let's first look at some of the simple functions:

-  **Addition:** This button places a plus sign in the command area. For example, with our hygiene data, 'day1 + day2' creates a column in which each row contains the hygiene score from the column labelled day1 added to the score from the column labelled day2 (e.g., for participant 1: $2.65 + 1.35 = 4$).
-  **Subtraction:** This button places a minus sign in the command area. For example, if we wanted to calculate the change in hygiene from day 1 to day 2 we could type 'day2 - day1'. This creates a column in which each row contains the score from the column labelled day1 subtracted from the score from the column labelled day2 (e.g., for participant 1: $2.65 - 1.35 = 1.30$).
-  **Multiply:** This button places a multiplication sign in the command area. For example, 'day1*day2' creates a column that contains the score from the column labelled day1 multiplied by the score from the column labelled day2 (e.g., for participant 1: $2.65 \times 1.35 = 3.58$).
-  **Divide:** This button places a division sign in the command area. For example, 'day1/day2' creates a column that contains the score from the column labelled day1 divided by the score from the column labelled day2 (e.g., for participant 1: $2.65/1.35 = 1.96$).
-  **Exponentiation:** This button raises the preceding term to the power of the succeeding term. So, 'day1**2' creates a column that contains the scores in the day1 column raised to the power of 2 (i.e., the square of each number in the day1 column: for participant 1, $2.65^2 = 7.02$). Likewise, 'day1**3' creates a column with values of day1 cubed.
-  **Less than:** This operation is usually used for 'include case' functions. If you click on the  button, a dialog box appears that allows you to select certain cases on which to carry out the operation. So, if you typed 'day1 < 1', then SPSS would carry out the compute function only for those participants whose hygiene score on day 1 of the festival was less than 1 (i.e., if day1 was 0.99 or less). So, we might use this if we wanted to look only at the people who were already smelly on the first day of the festival.

- 
Less than or equal to: This operation is the same as above except that in the example above, cases that are exactly 1 would be included as well.
- 
More than: This operation is used to include cases above a certain value. So, if you clicked on  and then typed 'day1 > 1' then SPSS will carry out any analysis only on cases for which hygiene scores on day 1 of the festival were greater than 1 (i.e., 1.01 and above). This could be used to exclude people who were already smelly at the start of the festival. We might want to exclude them because these people will contaminate the data (not to mention our nostrils) because they reek of putrefaction to begin with so the festival cannot further affect their hygiene.
- 
More than or equal to: This operation is the same as above but will include cases that are exactly 1 as well.
- 
Equal to: You can use this operation to include cases for which participants have a specific value. So, if you clicked on  and typed 'day1 = 1' then only cases that have a value of exactly 1 for the day1 variable are included. This is most useful when you have a coding variable and you want to look at only one of the groups. For example, if we wanted to look only at females at the festival we could type 'gender = 1', and then the analysis would be carried out on only females (who are coded as 1 in the data).
- 
Not equal to: This operation will include all cases except those with a specific value. So, 'gender \neq 1' (as in Figure 5.25) will carry out the compute command only on the males and exclude females (because they have a 1 in the gender column).

Some of the most useful functions are listed in Table 5.2, which shows the standard form of the function, the name of the function, an example of how the function can be used and what SPSS would output if that example were used. There are several basic functions for calculating means, standard deviations and sums of columns. There are also functions such as the square root and logarithm that are useful for transforming data that are skewed, and we will use these functions now. For the interested reader, the SPSS help files have details of all of the functions available through the *Compute Variable* dialog box (click on  when you're in the dialog box).




5.4.4.3. The log transformation in SPSS ②

Let's use *compute* to transform our data. Open the main *compute* dialog box by selecting  . Enter the name **logday1** into the box labelled *Target Variable*, click on  and give the variable a more descriptive name such as *Log transformed hygiene scores for day 1 of Download festival*. In the list box labelled *Function group* click on *Arithmetic* and then in the box labelled *Functions and Special Variables* click on *Lg10* (this is the log transformation to base 10; *Ln* is the natural log) and transfer it to the command area by clicking on . When the command is transferred, it appears in the command area as 'LG10(?)' and the question mark should be replaced with a variable name (which can be typed manually or transferred from the variables list). So replace the question mark with the variable **day1** by either selecting the variable in the list and dragging it across, clicking on , or just typing 'day1' where the question mark is.

For the day 2 hygiene scores there is a value of 0 in the original data, and there is no logarithm of the value 0. To overcome the problem we add a constant to our original scores before we take the log of those scores. Any constant will do (although sometimes it can matter), provided that it makes all of the scores greater than 0. In this case our lowest score

TABLE 5.2 Some useful compute functions





Function	Name	Example Input	Output
MEAN(?,?,...)	Mean	Mean(day1, day2, day3)	For each row, SPSS calculates the average hygiene score across the three days of the festival
SD(?,?,...)	Standard deviation	SD(day1, day2, day3)	Across each row, SPSS calculates the standard deviation of the values in the columns labelled <i>day1</i> , <i>day2</i> and <i>day3</i>
SUM(?,?,...)	Sum	SUM(day1, day2)	For each row, SPSS adds the values in the columns labelled <i>day1</i> and <i>day2</i>
SQRT(?)	Square root	SQRT(day2)	Produces a column containing the square root of each value in the column labelled <i>day2</i>
ABS(?)	Absolute value	ABS(day1)	Produces a variable that contains the absolute value of the values in the column labelled <i>day1</i> (i.e., the signs are ignored, so -5 becomes $+5$ and $+5$ stays as $+5$)
LG10(?)	Base 10 logarithm	LG10(day1)	Produces a variable that contains the logarithmic values (to base 10) of the variable <i>day1</i> .
RV.NORMAL (mean, stddev)	Normal random numbers	Normal(20, 5)	Produces a variable of pseudo-random numbers from a normal distribution with a mean of 20 and a standard deviation of 5.

is 0 in the data so we could add 1 to all of the scores to ensure that all scores are greater than zero. Even though this problem affects the day 2 scores, we need to be consistent and do the same to the day 1 scores as we will do with the day 2 scores. Therefore, make sure the cursor is still inside the brackets and click on  and then . The final dialog box should look like Figure 5.25. Note that the expression reads LG10(day1 + 1); that is, SPSS will add one to each of the day 1 scores and then take the log of the resulting values. Click on  to create a new variable **logday1** containing the transformed values.



SELF-TEST Have a go at creating similar variables **logday2** and **logday3** for the day 2 and day 3 data. Plot histograms of the transformed scores for all three days.

5.4.4.4. The square root transformation on SPSS ②

To do a square root transformation, we run through the same process, by using a name such as **sqrtday1** in the box labelled *Target Variable* (and click on  to give the variable a more descriptive name). In the list box labelled *Function group* click on *Arithmetic* and then in the box labelled *Functions and Special Variables* click on *Sqrt* and drag it to the command area or click on . When the command is transferred, it appears in the command area as SQRT(?). Replace the question mark with the variable **day1** by selecting the variable in the list and dragging it, clicking on , or just typing 'day1' where the question mark is. The final expression will read **SQRT(day1)**. Click on  to create the variable.



SELF-TEST Repeat this process for **day2** and **day3** to create variables called **sqrtday2** and **sqrtday3**. Plot histograms of the transformed scores for all three days.

5.4.4.5. The reciprocal transformation on SPSS ②

To do a reciprocal transformation on the data from day 1, we could use a name such as **recday1** in the box labelled *Target Variable*. Then we can simply click on **1** and then **1**. Ordinarily you would select the variable name that you want to transform from the list and drag it across, click on **1** or just type the name of the variable. However, the day 2 data contain a zero value and if we try to divide 1 by 0 then we'll get an error message (you can't divide by 0). We need to add a constant to our variable just as we did for the log transformation. Any constant will do, but 1 is a convenient number for these data. So, instead of selecting the variable we want to transform, click on **0**; this places a pair of brackets into the box labelled *Numeric Expression*. Then make sure the cursor is between these two brackets and select the variable you want to transform from the list and transfer it across by clicking on **1** (or type the name of the variable manually). Now click on **+** and then **1** (or type '+ 1')



SPSS TIP 5.2

Using syntax to compute new variables ③

If you're computing a lot of new variables it can be quicker to use syntax. I've written the file **Transformations.sps** to do all nine of the transformations that we've discussed. Open this file and you'll see these commands in the syntax window (see Section 3.9):

```
COMPUTE logday1 = LG10(day1 + 1).
COMPUTE logday2 = LG10(day2 + 1).
COMPUTE logday3 = LG10(day3 + 1).
COMPUTE sqrtday1 = SQRT(day1).
COMPUTE sqrtday2 = SQRT(day2).
COMPUTE sqrtday3 = SQRT(day3).
COMPUTE recday1 = 1/(day1+1).
COMPUTE recday2 = 1/(day2+1).
COMPUTE recday3 = 1/(day3+1).
EXECUTE.
```

Each *compute* command above does the equivalent of what you'd do using the *Compute Variable* dialog box in Figure 5.25. So, the first three lines ask SPSS to create three new variables (**logday1**, **logday2** and **logday3**), which are the log transformations of the variables **day1**, **day2** and **day3** plus 1. The next three lines create new variables called **sqrtday1**, **sqrtday2** and **sqrtday3** by using the *SQRT* function to take the square root of **day1**, **day2** and **day3**, respectively. The next three lines do the reciprocal transformation in much the same way. The final line has the command *execute* without which none of the *compute* commands beforehand will be executed. Note also that every line ends with a full stop.

using your keyboard). The box labelled *Numeric Expression* should now contain the text $1/(day1 + 1)$. Click on to create a new variable containing the transformed values.



SELF-TEST Repeat this process for **day2** and **day3**. Plot histograms of the transformed scores for all three days.

5.4.4.6. The effect of transformations ②

Figure 5.26 shows the distributions for days 1 and 2 of the festival after the three different transformations. Compare these to the untransformed distributions in Figure 5.13. Now, you can see that all three transformations have cleaned up the hygiene scores for day 2: the positive skew is reduced (the square root transformation in particular has been useful). However, because our hygiene scores on day 1 were more or less symmetrical to begin

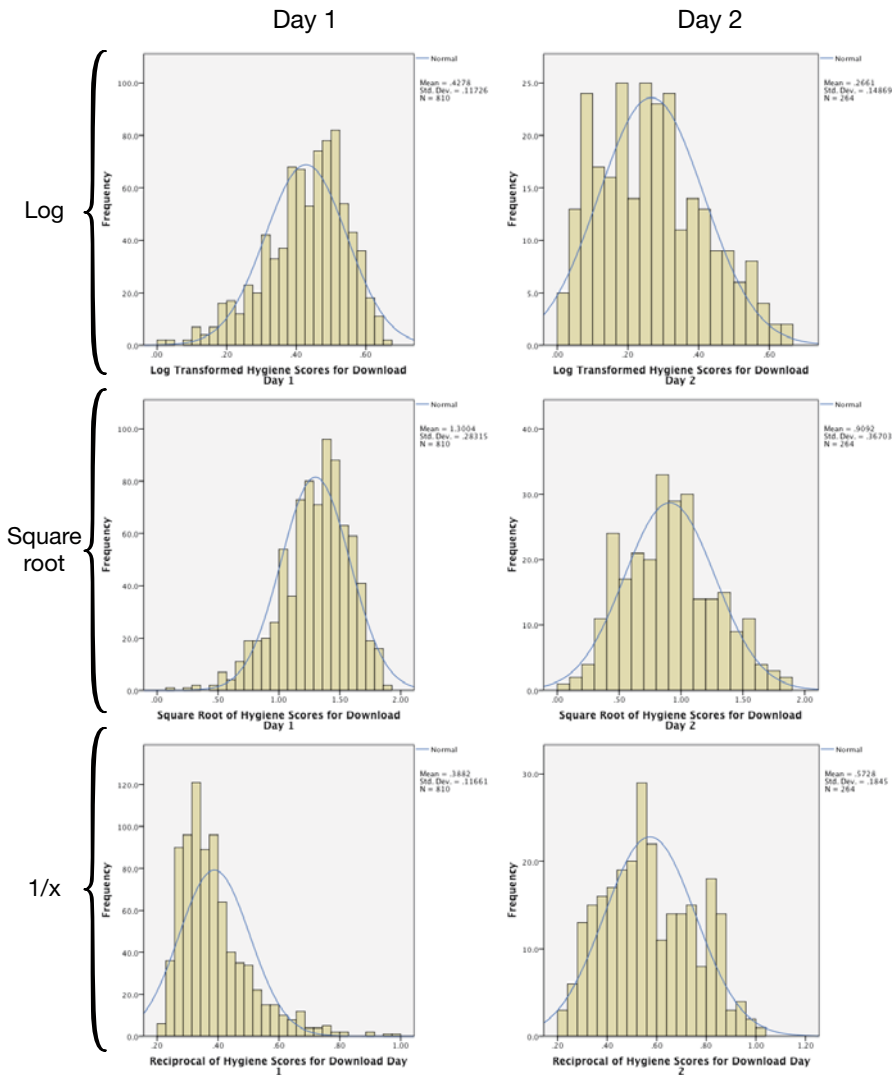
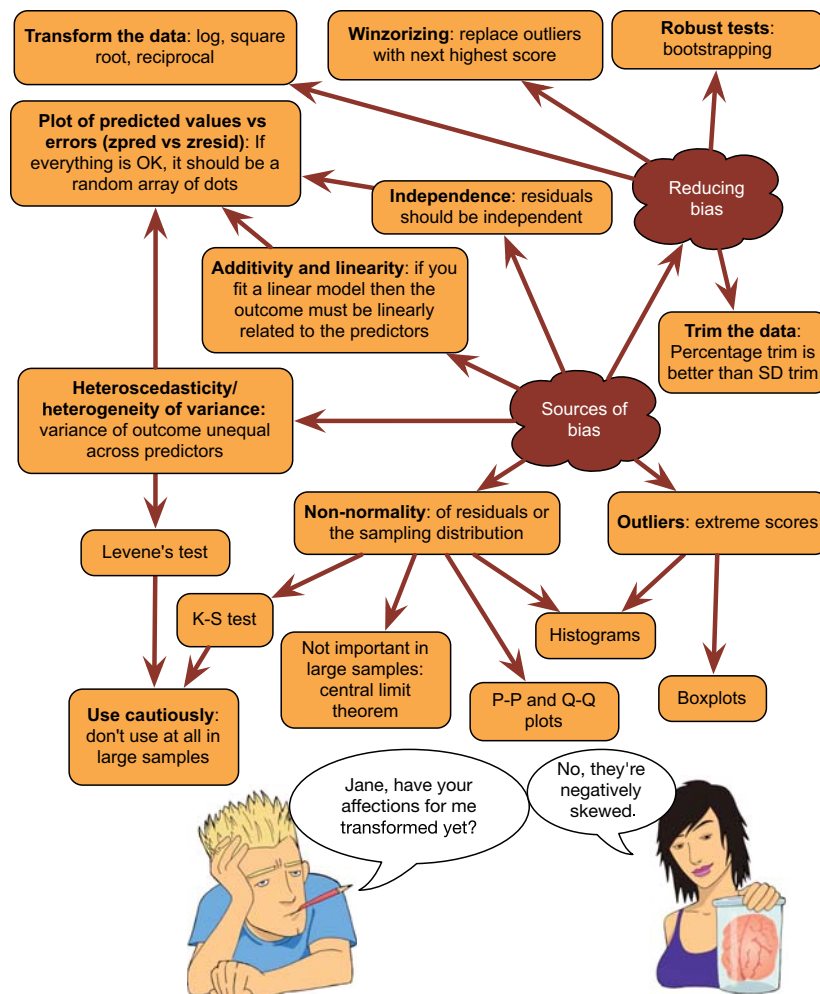


FIGURE 5.26 Distributions of the hygiene data on day 1 and day 2 after various transformations

with, they have now become slightly negatively skewed for the log and square root transformation, and positively skewed for the reciprocal transformation.¹⁴ If we're using scores from day 2 alone or looking at the relationship between day 1 and day 2, then we could use the transformed scores; however, if we wanted to look at the *change* in scores then we'd have to weigh up whether the benefits of the transformation for the day 2 scores outweigh the problems it creates in the day 1 scores – data analysis can be frustrating sometimes.☺

5.5. Brian's attempt to woo Jane ①

FIGURE 5.27
What Brian learnt from this chapter



5.6. What next? ①

This chapter has taught us how to identify bias. Had I read this chapter I might have avoided being influenced by my idolization of my granddad¹⁵ and instead realized that I

¹⁴ The reversal of the skew for the reciprocal transformation is because, as I mentioned earlier, the reciprocal has the effect of reversing the scores.

¹⁵ Oddly enough, despite absolutely worshipping the ground my granddad walked on, I ended up supporting a different team than him: he supported a certain north London team close to where we grew up and I support their local rivals.

could be a useful midfield player. From there a successful career in soccer would undoubtedly have unfolded in front of me. Or, as anyone who has seen me play will realize, perhaps not. Still, I sort of had the last laugh on the goalkeeping front. At the end of my time at primary school we had a five-a-side tournament between local schools so that kids from different schools could get to know each other before going to secondary school together. My goalkeeping nemesis was, of course, chosen to play and I was the substitute. In the first game he had a shocker, and I was called up to play in the second game during which I made a series of dramatic and acrobatic saves (at least that's how I remember them). I did likewise in the next game, and my nemesis had to sit out the whole of the rest of the tournament. Perhaps this should have encouraged me to pursue being goalkeeper at my new school. However, five-a-side goals are shorter than normal goals, so my height wasn't an issue and that was my last time trying to get into the school football team – I just gave up. Years later when I started playing again, I regretted this decision: not because I could have been a professional soccer player, but just because I missed many years of enjoying playing. Instead, I read books and immersed myself in music. Unlike my cleverer older brother who was reading Albert Einstein's papers (well, Isaac Asimov) as an embryo, my literary preferences were more in keeping with my intellect ...

5.7. Key terms that I've discovered

Bootstrap	Independence	Parametric test
Contaminated normal distribution	Kolmogorov–Smirnov test	Q-Q plot
Hartley's F_{\max}	Levene's test	Robust test
Heterogeneity of variance	M-estimator	Shapiro–Wilk test
Heteroscedasticity	Mixed normal distribution	Transformation
Homogeneity of variance	Normally distributed data	Trimmed mean
Homoscedasticity	Outlier	Variance ratio
	P-P plot	Weighted least squares

5.8. Smart Alex's tasks

- **Task 1:** Using the **ChickFlick.sav** data from Chapter 4, check the assumptions of normality and homogeneity of variance for the two films (ignore **Gender**): are the assumptions met? ①
- **Task 2:** The file **SPSSExam.sav** contains data regarding students' performance on an SPSS exam. Four variables were measured: **exam** (first-year SPSS exam scores as a percentage), **computer** (measure of computer literacy in percent), **lecture** (percentage of SPSS lectures attended) and **numeracy** (a measure of numerical ability out of 15). There is a variable called **uni** indicating whether the student attended Sussex University (where I work) or Duncetown University. Compute and interpret descriptive statistics for **exam**, **computer**, **lecture**, and **numeracy** for the sample as a whole. ①
- **Task 3:** Calculate and interpret the z -scores for skewness for all variables. ①
- **Task 4:** Calculate and interpret the z -scores for kurtosis for all variables. ①
- **Task 5:** Use the *split file* command to look at and interpret the descriptive statistics for **numeracy** and **exam**. ①
- **Task 6:** Repeat Task 5 but for the computer literacy and percentage of lectures attended. ①





- **Task 7:** Conduct and interpret a K-S test for **numeracy** and **exam**. ①
- **Task 8:** Conduct and interpret a Levene's test for **numeracy** and **exam**. ①
- **Task 9:** Transform the **numeracy** scores (which are positively skewed) using one of the transformations described in this chapter. Do the data become normal? ②
- **Task 10:** Use the *explore* command to see what effect a natural log transformation would have on the four variables measured in **SPSSExam.sav**.

Answers can be found on the companion website.

5.9. Further reading

- Tabachnick, B. G., & Fidell, L. S. (2012). *Using multivariate statistics* (6th ed.). Boston: Allyn & Bacon. (They have the definitive guide to screening data.)
- Wilcox, R. R. (2005). *Introduction to robust estimation and hypothesis testing* (2nd ed.). Burlington, MA: Elsevier. (Quite technical, but this is the definitive book on robust methods.)
- Wilcox, R. R. (2010). *Fundamentals of modern statistical methods: Substantially improving power and accuracy*. New York: Springer-Verlag. (A fantastic book on bias in statistical methods that expands upon many of the points in this chapter and is written by someone who actually knows what he's talking about.)

