Exploring Data: The Beast of Bias

Sources of Bias

A bit of revision. We've seen 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:

$$\text{outcome}_i = (b_1X_{1i} + b_2X_{2i} \cdots b_nX_{ni}) + \text{error}_i \qquad \text{Eq. 1}$$

Therefore, we predict an outcome variable, from one or more predictor variables (the *X*s) 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, we often estimate the parameters (*b*) using the method of least squares (known as **ordinary least squares** or **OLS**). We're not interested in our sample so much as a general population, 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 or confidence interval. We can test hypotheses about these parameters by computing test statistics and their associated probabilities (*p*-values). Therefore, when we think about bias, we need to think about it within three contexts:

1. Things that bias the parameter estimates.
2. Things that bias standard errors and confidence intervals.
3. Things that bias test statistics and *p*-values.

These situations are related because confidence intervals and test statistics both rely on the standard error. It is important that we identify and eliminate anything that might affect the information that we use to draw conclusions about the world, so we explore data to look for bias.

Outliers

An **outlier** is a score very different from the rest of the data. When I published my first book (Field, 2000), 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). The mean of these scores was 4.43. The only score that wasn't a 4 or 5 was the first rating of 2. This score is an example of an outlier—a weird and unusual person (sorry, I mean score) that deviates from the rest of humanity (I mean, data set). The mean of the scores when the outlier is not included is 4.83 (it increases by 0.4). This example shows how a single score, from 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.

Spotting outliers with graphs

A biologist was worried about the potential health effects of music festivals. So, one year she went to the Download Music Festival (http://www.downloadfestival.co.uk) and measured the hygiene of 810 concert goers over the three days of the festival. In theory each person was measured on each day but because it was difficult to track people down, there were some missing data on days 2 and 3. Hygiene was measured using a standardised technique (don't worry it *wasn't* licking the person's armpit) that results in a score ranging between 0 (you smell like a rotting corpse that's hiding up a skunk's arse) and 5 (you smell of sweet roses on a fresh spring day). Now I know from bitter experience that sanitation is not always great at these places (Reading festival seems particularly bad …) and so this researcher predicted that personal hygiene would go down dramatically over the three days of the festival. The data file is called **DownloadFestival.sav**.

To plot a histogram we use Chart Builder (see last week's handout) which is accessed through the Graphs Chart Builder... menu. Select *Histogram* in the list labelled *Choose from* to bring up the gallery shown in Figure 1. This gallery has four icons representing different types of histogram, and you should select the appropriate one either by double-clicking on it, or by dragging it onto the canvas in the Chart Builder:

- → **Simple histogram**: Use this option when you just want to see the frequencies of scores for a single variable (i.e. most of the time).
- → **Stacked histogram** and **Population Pyramid**: If you had a grouping variable (e.g. whether men or women attended the festival) you could produce a histogram in which each bar is split by group (stacked histogram) or the outcome (in this case hygiene) on the vertical axis and each group (i.e. men vs. women) on the horizontal (i.e., the histograms for men and women appear back to back on the graph).
- → **Frequency Polygon**: This option displays the same data as the simple histogram except that it uses a line instead of bars to show the frequency, and the area below the line is shaded.

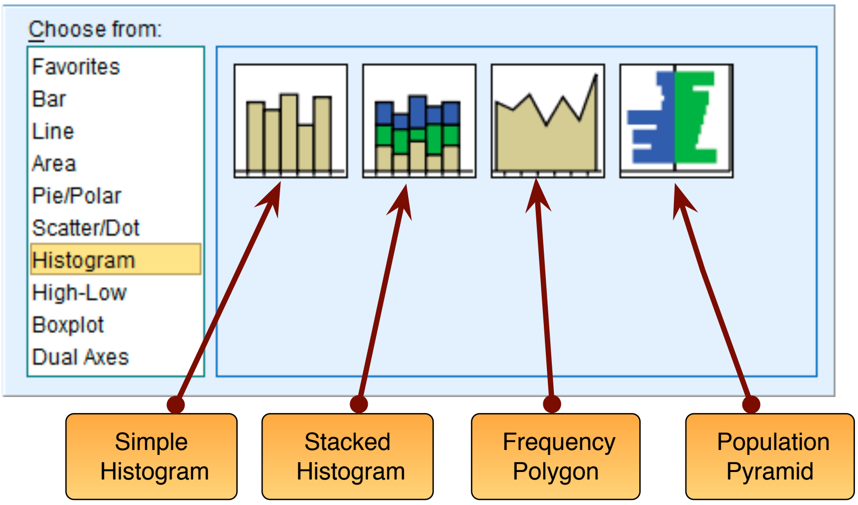


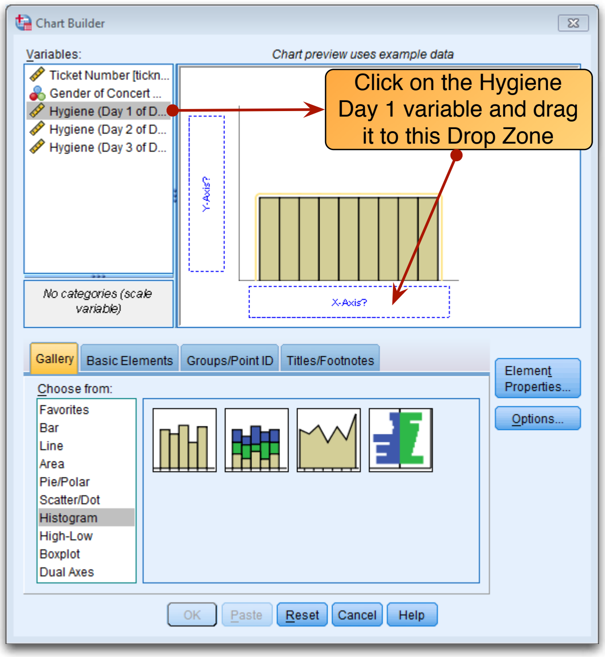
Figure 1: The histogram gallery

We are going to do a simple histogram so double-click on the icon for a simple histogram (Figure 1). The *Chart Builder* dialog box will now show a preview of the graph in the canvas area. At the moment it's not very exciting because we haven't told SPSS which variables we want to plot. Note that the variables in the data editor are listed on the left-hand side of the Chart Builder, and any of these variables can be dragged into any of the *drop zones* (spaces surrounded by blue dotted lines).

A histogram plots a single variable (*x*-axis) against the frequency of scores (*y*-axis), so all we need to do is select a variable from the list and drag it into X-Axis?. Let's do this for the hygiene scores on day 1 of the festival. Click on this variable in the list and drag it to X-Axis? as shown in Figure 2; you will now find the histogram previewed on the canvas. To draw the histogram click on OK.

The resulting histogram is shown in Figure 3 and the first thing that should leap out at you is that there appears to be one case that is very different to the others. All of the scores appear to be squished up one end of the distribution because they are all less than 5 (yielding what is known as a leptokurtic distribution!) except for one, which has a value of 20! This score is an outlier. What's odd about this outlier is that it has a score of 20, which is above the top of our scale (remember our hygiene scale ranged only from 0-5) and so it must be a mistake (or the person had obsessive compulsive disorder and had washed themselves into a state of extreme cleanliness). 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**.

You encountered boxplots or box-whisker diagrams last year. At the centre of the plot is the *median*, which is surrounded by a box the top and bottom of which are the limits within which the middle 50% of observations fall (the interquartile range). Sticking out of the top and bottom of the box are two whiskers which extend to the most and least extreme scores respectively. Outliers as shown as dots or stars (see my book for details).



Click on the Hygiene Day 1 variable and drag it to this Drop Zone

Figure 2: Plotting a histogram

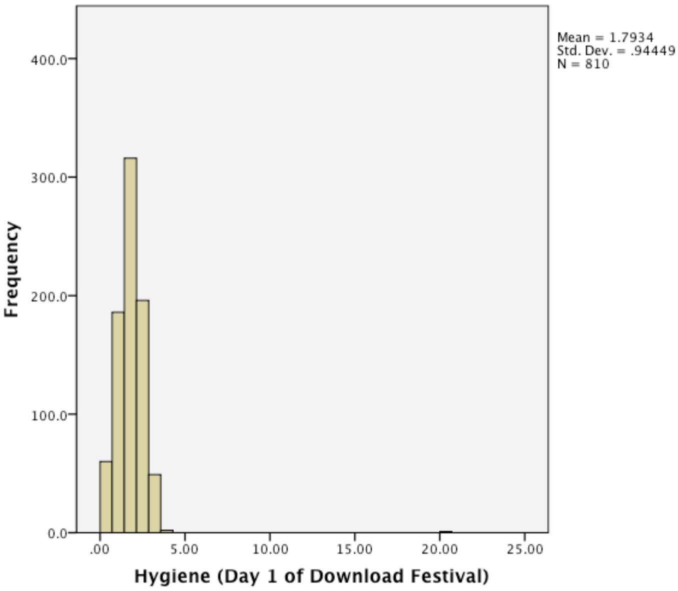


Figure 3: Histogram of the Day 1 Download Festival hygiene scores

Select the Graphs Chart Builder... menu, then select *Boxplot* in the list labelled *Choose from* to bring up the gallery shown in Figure 4. There are three types of boxplot you can choose:

→ **Simple boxplot**: Use this option when you want to plot a boxplot of a single variable, but you want different boxplots produced for different categories in the data (for these hygiene data we could produce separate boxplots for men and women).

→ **Clustered boxplot**: This option is the same as the simple boxplot except that you can select a second categorical variable on which to split the data. Boxplots for this second variable are produced in different colours. For example, we might have measured whether our festival-goer was staying in a tent or a nearby hotel during the festival. We could produce boxplots not just for men and women, but within men and women we could have different-coloured boxplots for those who stayed in tents and those who stayed in hotels.

- → **1-D Boxplot**: Use this option when you just want to see a boxplot for a single variable. (This differs from the simple boxplot only in that no categorical variable is selected for the x-axis.)

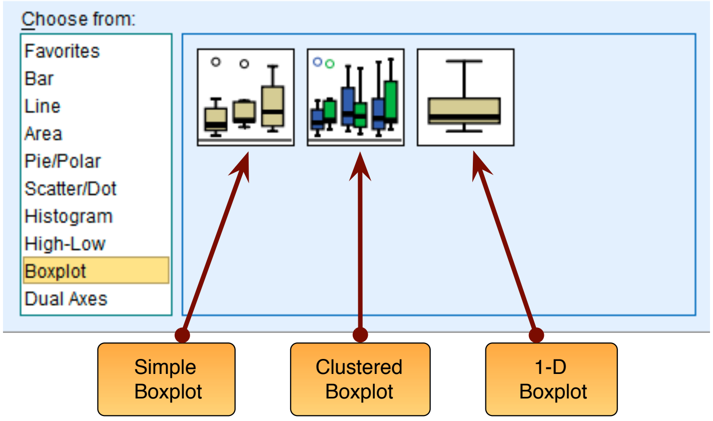


Figure 4: The boxplot gallery

To make our boxplot of the day 1 hygiene scores, double-click on the *simple boxplot* icon (Figure 4), then from the variable list select the hygiene day 1 score variable and drag it into Y-Axis?. The dialog should now look like Figure 5— click on OK to produce the graph.

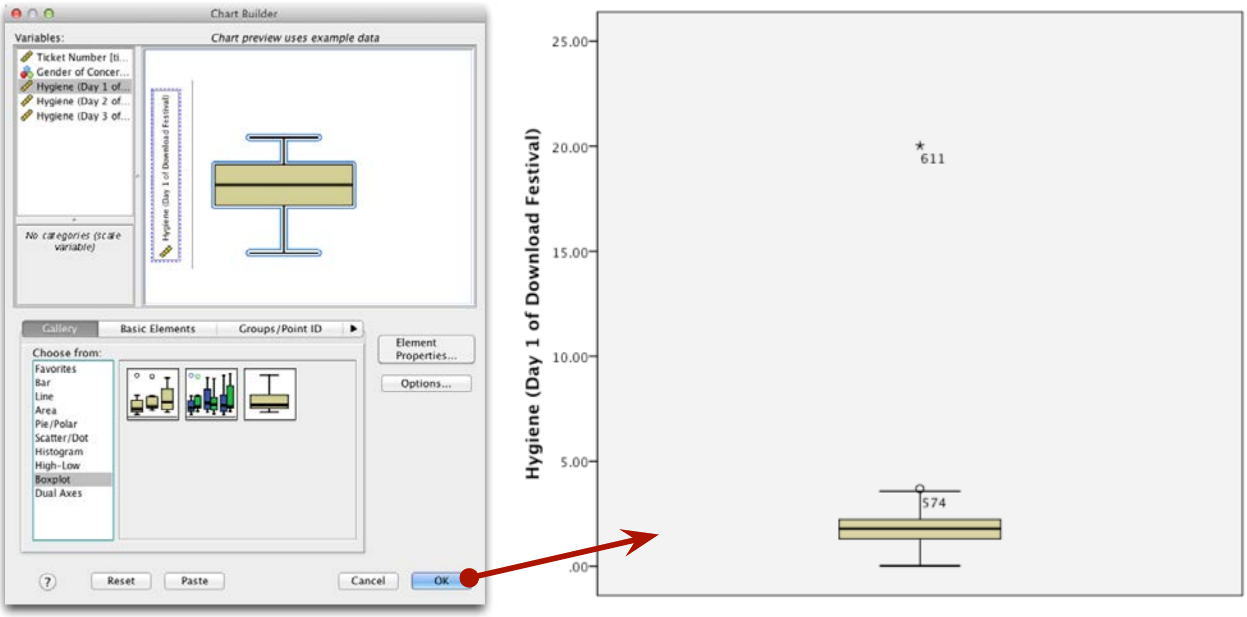


Figure 5: Boxplot for the download festival data

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

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

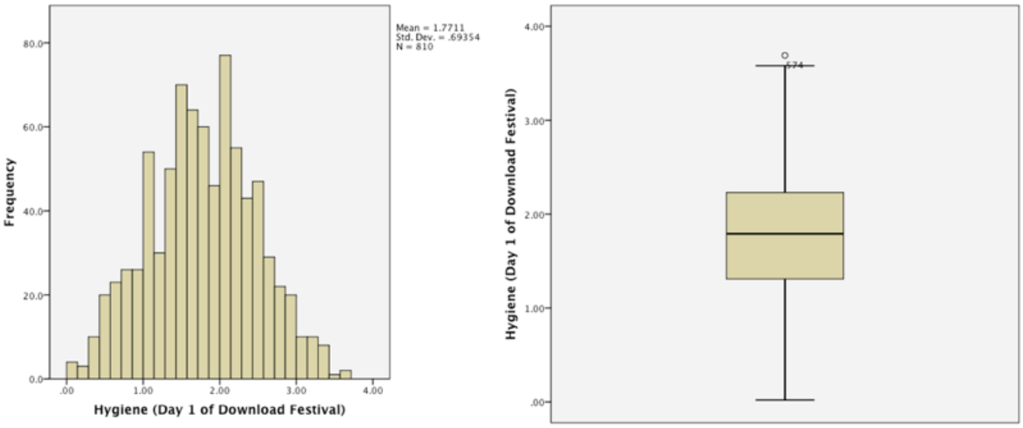


Figure 6: Histogram (left) and boxplot (right) of hygiene scores on day 1 of the Download Festival

Figure 6 shows the histogram and boxplot for the data after the extreme case has been corrected. The distribution 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

Assumptions

Most of our potential sources of bias come in the form of 'violations of assumptions'. 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 as though different statistical procedures have their own unique set of assumptions. However, because we're usually fitting variations of the linear model to our data, all of the tests in my book (Field, 2013) 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

Additivity and Linearity

The vast majority of statistical models in my book are based on the linear model, which takes this form:

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

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) 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:

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

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 it's behaviour isn't what you expect because even though you've a called it a dog, it is in fact a cat.

The assumption of normality

What does it mean?

Many people take the 'assumption of normality' to mean that your data need to be normally distributed. However, that isn't what it means. What it does mean is:

1. For confidence intervals around a parameter estimate (e.g., the mean, or a *b*) 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 *b*s in Eq. 1) 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 *b*s in Eq. 1) to be optimal (using the method of least squares) the residuals (the error_i in Eq. 1) in the population must be normally distributed.

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 also (and remember, we don't have direct access to the sampling distribution so we have to make educated guesses about its shape).

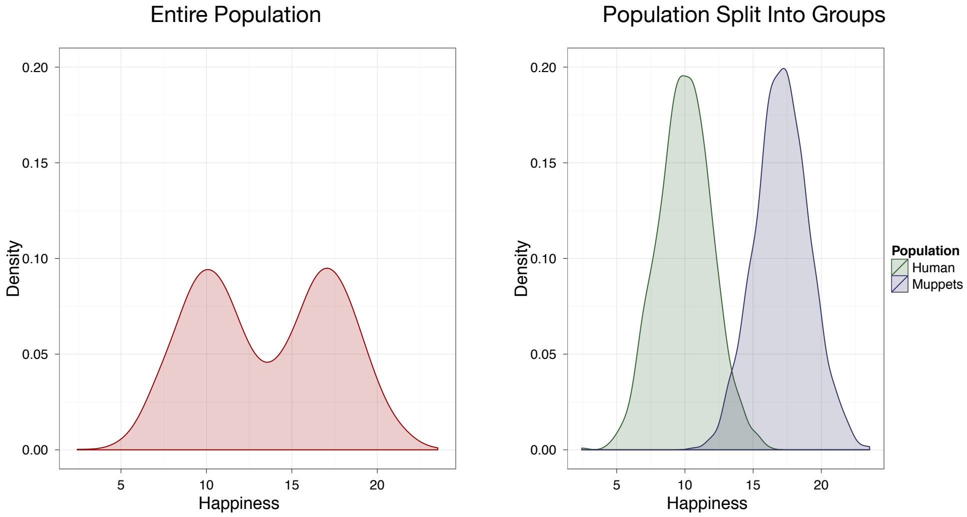


Figure 7: A distribution that looks non-normal (left) could be made up of different groups of normally-distributed scores

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 7 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 7) 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.

When does the assumption of normality matter?

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):

1. **Confidence intervals**: 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. **Significance tests**: the central limit theorem tells us that 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—for more information read Field (2013).
3. **Parameter estimates**: 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 Field (2013) for a discussion of what we might mean by a larger sample). In practical terms, as long as your sample is fairly large, outliers are a more pressing concern than normality.

Homogeneity of Variance/Homoscedasticity

The second assumption we'll explore relates to variance and it can impact on the two main things that we might do when we fit models to data:

- **Parameters**: If we use the method of least squares 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).

What is homoscedasticity/homogeneity of variance?

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.

When does homoscedasticity/homogeneity of variance matter?

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. They will be 'unbiased' (Hayes & Cai, 2007) but not optimal.

Unequal variances/heteroscedasticity also creates bias and inconsistency in the estimate of the standard error associated with the parameter estimates (Hayes & Cai, 2007). This basically means that confidence intervals and significance tests 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). Some test statistics are designed to be accurate even when this assumption is violated.

Independence

This assumption means that the errors in your model (the error_i in Eq. 1) 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 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, there are techniques you can use described in Chapter 20 of (Field, 2013).

Testing Assumptions

Testing normality

You can look for normality in three ways: (1) graphs; (2) numerically; and (3) significance tests. We can do all three using the *Explore* command in SPSS. In terms of graphs we can look at histograms (which we've already learnt about) and **P-P plots** and **Q-Q plots**. P-P and Q-Q plots basically show the same thing: a P-P plot plots the cumulative probability of a variable against the cumulative probability of a particular distribution (in this case a normal distribution). A Q-Q plot does the same thing but expressed as quantiles. With large data sets Q-Q plots are a bit easier to interpret. In a sense they plot the 'actual data' against 'what you'd expect to get from a normal distribution', so if the data are normally distributed then the actual scores will be the same as the expected scores 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.

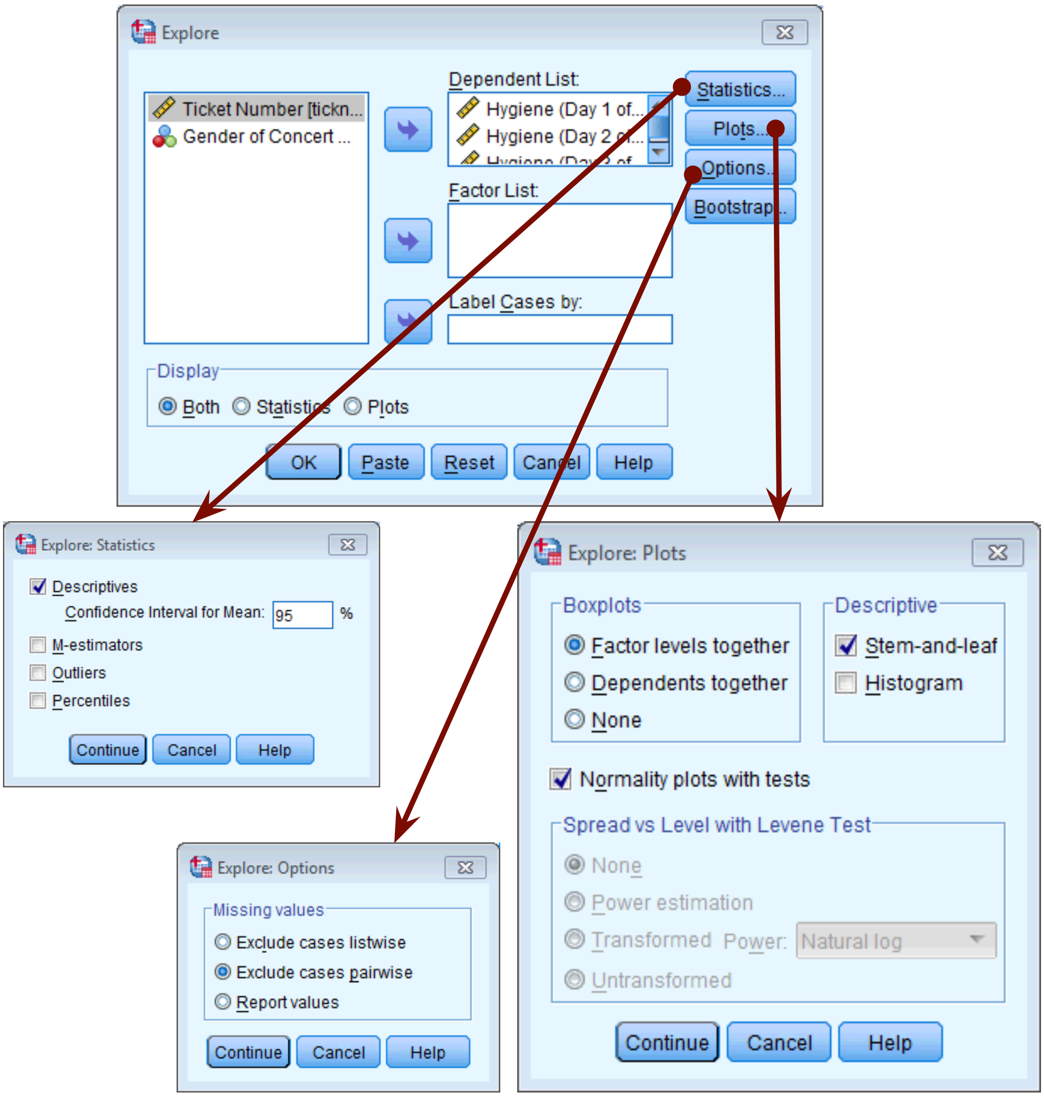
Numerically, SPSS uses methods to calculate skew and kurtosis (see Field (2013) if you have forgotten what these concepts are) that give values of zero in a normal distribution. Positive values of skewness indicate a pile-up of scores on the left of the distribution, whereas negative values indicate 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.

Finally, we can 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, the Jane Superbrain Box 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.

Figure 8 shows the dialog boxes for the *Explore* command (Analyze Descriptive Statistics ▸ Explore...). 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 Q-Q plots*. You can also split the analysis by a factor or grouping varaiable (for example, we could do a separete analysis for males and females by dragging **gender** to the *Factor List* box — we'll do this later in the handout).



Explore

Dependent List:
Hygiene (Day 1 of...
Hygiene (Day 2 of...
Hygiene (Day 3 of...

Factor List:

Label Cases by:

Display: ● Both ○ Statistics ○ Plots

Explore: Statistics

☑ Descriptives
Confidence Interval for Mean: 95 %
□ M-estimators
□ Outliers
□ Percentiles

Explore: Plots

Boxplots: ● Factor levels together ○ Dependents together ○ None

Descriptive: ☑ Stem-and-leaf □ Histogram

☑ Normality plots with tests

Spread vs Level with Levene Test: ● None ○ Power estimation ○ Transformed Power: Natural log ○ Untransformed

Explore: Options

Missing values: ○ Exclude cases listwise ● Exclude cases pairwise ○ Report values

Figure 8: Dialog boxes for the *explore* command

We also need to click on Options... to tell SPSS how to deal with misisng values. This is important because although we start off with 810 scores on day 1, by day two we have only 264 and this is reduced to 123 on day 3. By default SPSS will use only cases for which there are valid scores on all of thes elected 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 exlcuding cases *listwise*. However, we want it to use all of the scores it has on a given day, which is known as *pairwise*. 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.

SELF-TEST: We have already plotted a histogram of the day 1 scores, using what you leant before plot histograms for the hygiene scores for days 2 and 3 of the Download Festival.

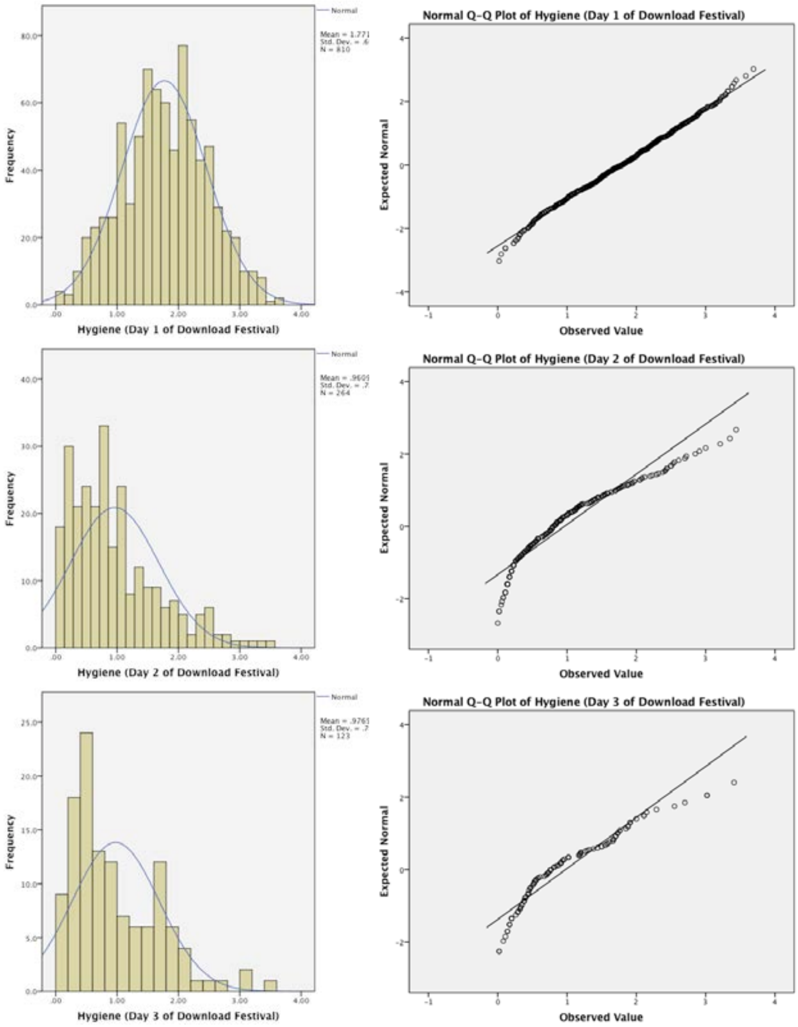


Figure 9: Histograms (left) and P-P plots (right)of the hygiene scores over the three days of the Download Festival

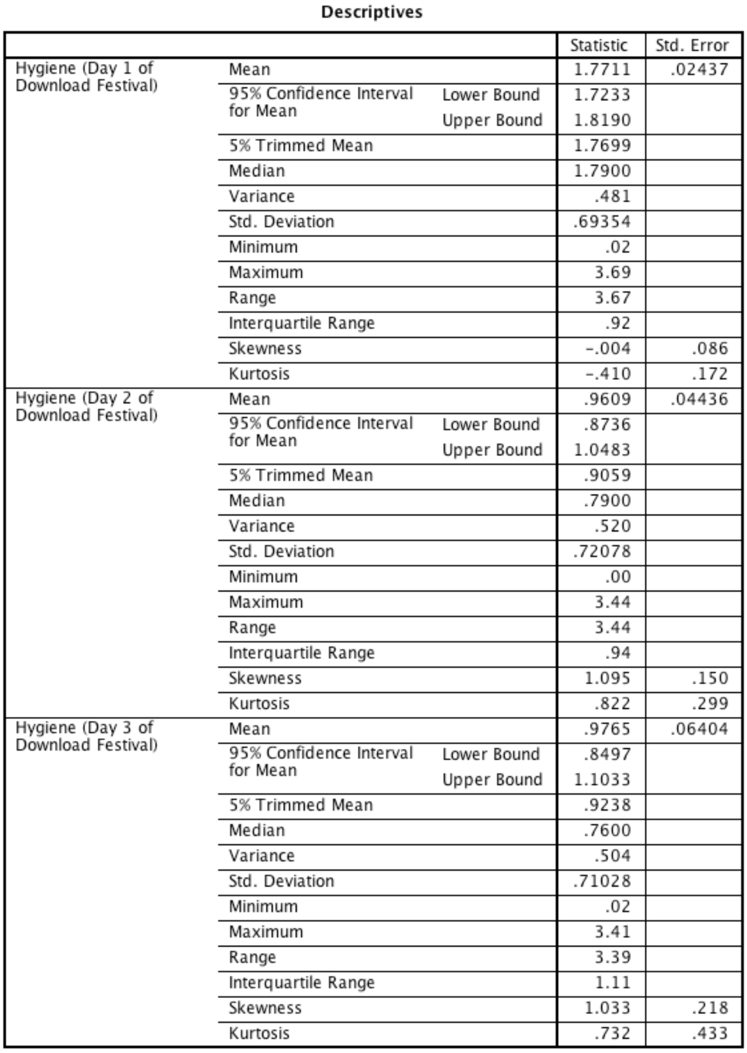
Figure 9 shows the histograms (from the self-test tasks) and the corresponding Q-Q plots. The day 1 scores look quite normal; The Q-Q plot echoes this view because the data points all fall very close to the 'ideal' diagonal line. However, the distributions for days 2 and 3 are not nearly as symmetrical as day 1: they both look positively skewed. Again, this can be seen in the Q-Q 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).

Output 1 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, both of which have an associated standard error. 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).

We can convert these values to *z*-scores 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), 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.

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

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. 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 look at 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).



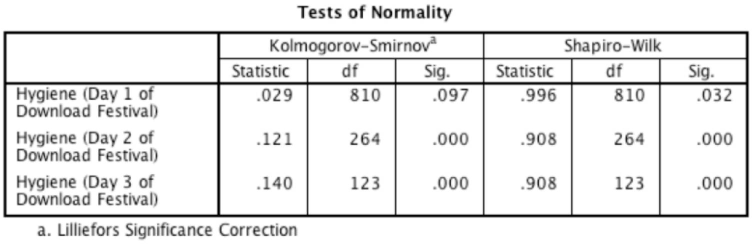
Descriptives

			Statistic	Std. Error
Hygiene (Day 1 of Download Festival)	Mean		1.7711	.02437
	95% Confidence Interval for Mean	Lower Bound	1.7233	
		Upper Bound	1.8190	
	5% Trimmed Mean		1.7699	
	Median		1.7900	
	Variance		.481	
	Std. Deviation		.69354	
	Minimum		.02	
	Maximum		3.69	
	Range		3.67	
	Interquartile Range		.92	
	Skewness		-.004	.086
	Kurtosis		-.410	.172
Hygiene (Day 2 of Download Festival)	Mean		.9609	.04436
	95% Confidence Interval for Mean	Lower Bound	.8736	
		Upper Bound	1.0483	
	5% Trimmed Mean		.9059	
	Median		.7900	
	Variance		.520	
	Std. Deviation		.72078	
	Minimum		.00	
	Maximum		3.44	
	Range		3.44	
	Interquartile Range		.94	
	Skewness		1.095	.150
	Kurtosis		.822	.299
Hygiene (Day 3 of Download Festival)	Mean		.9765	.06404
	95% Confidence Interval for Mean	Lower Bound	.8497	
		Upper Bound	1.1033	
	5% Trimmed Mean		.9238	
	Median		.7600	
	Variance		.504	
	Std. Deviation		.71028	
	Minimum		.02	
	Maximum		3.41	
	Range		3.39	
	Interquartile Range		1.11	
	Skewness		1.033	.218
	Kurtosis		.732	.433

Output 1

Using the values in Output 1, calculate the z-scores for skewness and Kurtosis for each day of the Download festival.

Your Answers:



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 2

Output 2 shows the K-S test. Remember that a significant value (*Sig.* less than .05) indicates a deviation from normality.

Are the results of the K-S tests surprising given the histograms we have already seen?

Your Answers:

For day 1 the K-S test is just about not significant (p = .097), which is surprisingly close to significant given now normal the day 1 scores looked in the histogram (Figure 3). 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). 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 but could again be down to the large sample (Figure 9).

Reporting the K-S test

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 2 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.

Jane Superbrain Box: Significance tests and assumptions

Throughout this handout 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 Kolmorgorov-Smirnoff and Shapiro-Wilk tests), tests of homogeneity of variances (Levene's test), and tests of significance of skew and kurtosis. All of these tests are 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.

The central limit theorem means that as sample sizes get larger, the less the assumption of normality matters 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 then pay attention if your significance tests are significant but resist being lulled into a false sense of security if they are not.

Testing homogeneity of variance/homoscedasticity

Like normality, you can look at the variances using graphs, numbers and significance tests. Graphically, we can create a scatterplot of the values of the residuals plotted 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[1] so this plot is sometimes referred to as **zpred vs. zresid**. 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 10 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 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.

Numerically, we can simply look at the values of the variances in different groups. Some people look at **Hartley's FMax** 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. The acceptable size of this ratio depends on the number of variances compared and the sample size— see (Field, 2013) for more detail.

More commonly used is **Levene's test** (Levene, 1960), which tests the null hypothesis that the variances in different groups are equal.

- If Levene's test is significant at $p \leq .05$ then you conclude that the variances are significantly different—therefore, the assumption of homogeneity of variances has been violated.

[1] Theses standardized errors are called standardized residuals.

- 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 you need to take Levene's test with a pinch of salt (Jane Superbrain Box).

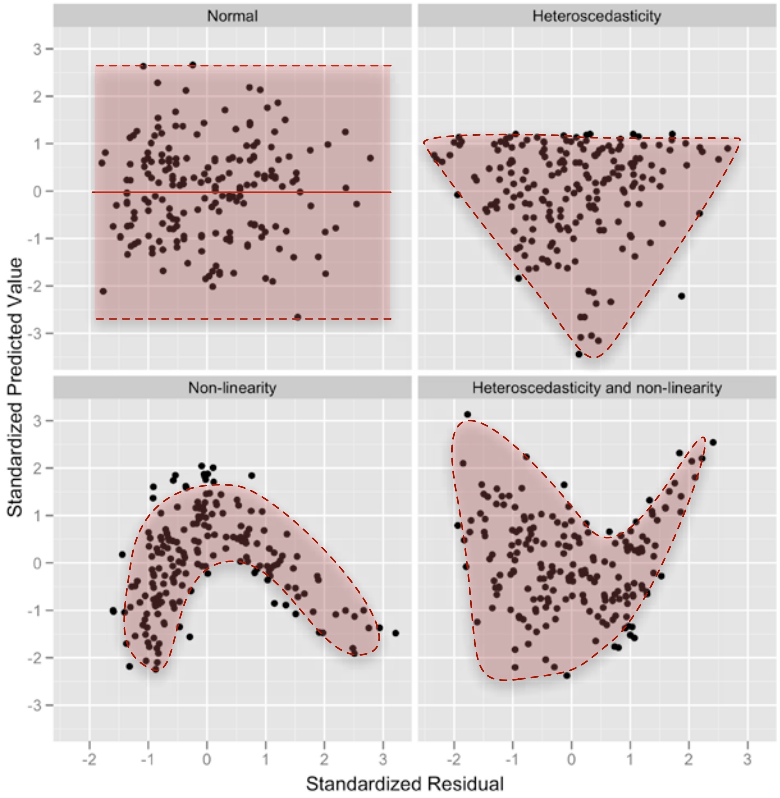
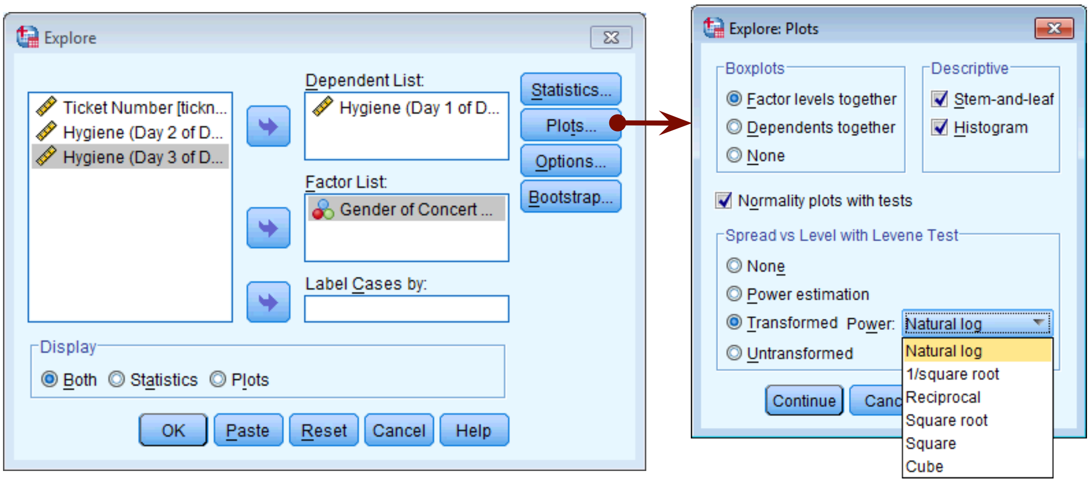


Figure 10: Plots of standardized residuals against predicted (fitted) values

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 of the festival. Use Analyze Descriptive Statistics ▸ Explore... to open the dialog box in Figure 11. 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 11. 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 3 shows the table for Levene's test. Levene's test can be based on differences between scores and the mean, and 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: 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 even this very small difference in variances is shown up as significant by Levene's test (Jane Superbrain Box). Hopefully this example convinces you to treat this test cautiously.



Explore

Ticket Number [tickn...
Hygiene (Day 2 of D...
Hygiene (Day 3 of D...

Dependent List: Hygiene (Day 1 of D...

Factor List: Gender of Concert ...

Label Cases by:

Display: ● Both ○ Statistics ○ Plots

Explore: Plots

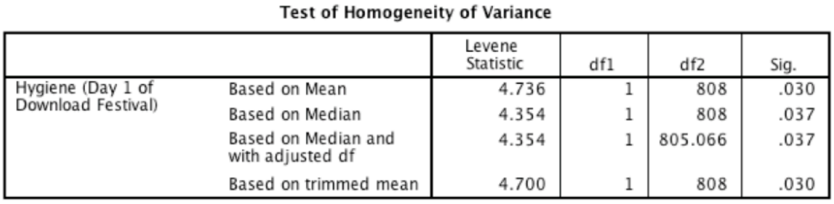
Boxplots: ● Factor levels together ○ Dependents together ○ None

Descriptive: ☑ Stem-and-leaf ☑ Histogram

☑ Normality plots with tests

Spread vs Level with Levene Test: ○ None ○ Power estimation ● Transformed Power: Natural log ○ Untransformed

Figure 11: 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 3

Reporting Levene's test

Levene's test can be denoted with the letter *F* and there are two different degrees of freedom. As such you can report it, in general form, as *F*(df1, df2) = value, *sig*:

✓ 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.

What is the assumption of homogeneity of variance?

Your Answers:

Reducing Bias

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:

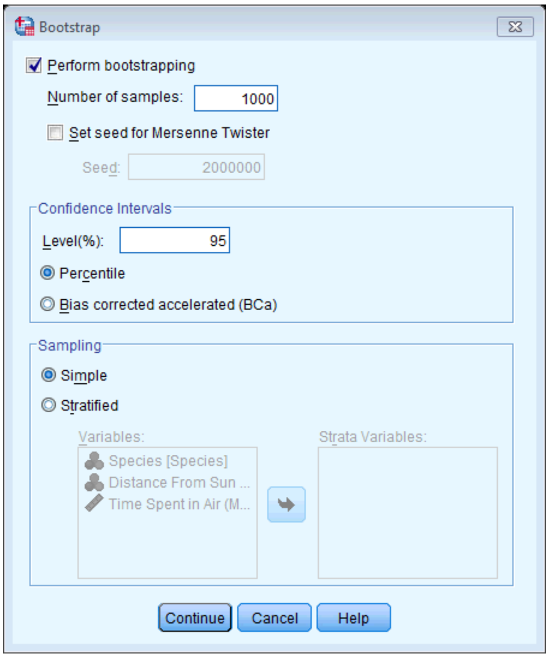
- **Trim the data**: delete a certain amount of scores from the extremes.
- **Windsorizing***:* substitute outliers with the highest value that isn't an outlier.
- **Analyse with Robust Methods**: this typically involves a technique known as bootstrapping.
- **Transform the data**: this involves applying 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. For the purposes of this handout we'll look at transforming data, and throughout the module we'll use bootstrapping (which is a robust method explained in your lecture), but you can find more detail on these techniques and the other in Chapter 5 of (Field, 2013).

Bootstrapping (robust methods)

Some SPSS procedures have a bootstrap option, which can be accessed by clicking on Bootstrap... to activate the dialog box in Figure 12. 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.

YouTube™ http://youtu.be/mNrxixgwA2M


Bootstrap

☑ Perform bootstrapping

Number of samples: 1000

□ Set seed for Mersenne Twister

Seed: 2000000

Confidence Intervals

Level(%): 95

● Percentile

○ Bias corrected accelerated (BCa)

Sampling

● Simple

○ Stratified

Variables: Species [Species], Distance From Sun ..., Time Spent in Air (M...

Strata Variables:

Figure 12: The standard bootstrap dialog box

Transforming Data

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. 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., 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.

Choosing a transformation

There are various transformations that you can do to the data that are helpful in correcting various problems. Table 1: shows some common transformations and their uses. 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.

Table 1: Data transformations and their uses

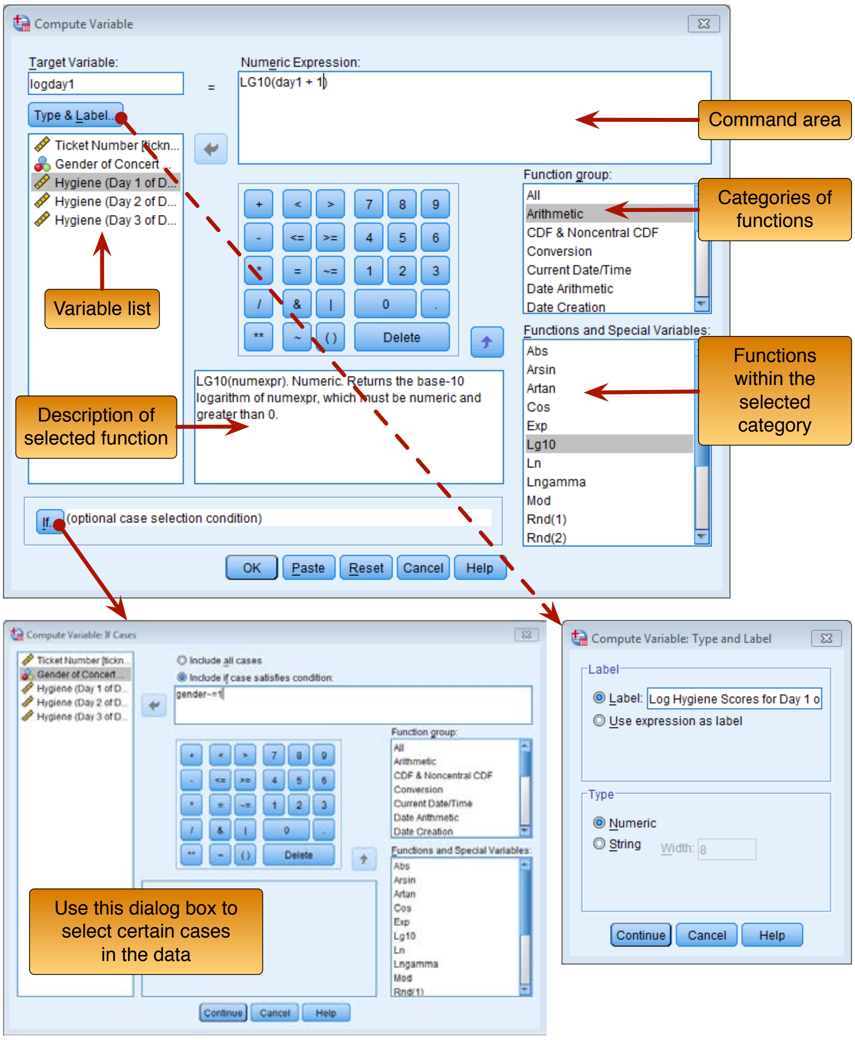
Data Transformation	Can Correct For
Log transformation ($\log(X_i)$): Taking the logarithm of a set of numbers squashes the right tail of the distribution. As such it's a good way to reduce positive skew. This transformation is also very useful if you have problems with linearity (it can sometimes make 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.	Positive skew, positive kurtosis, unequal variances, lack of linearity
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).	Positive skew, positive kurtosis, unequal variances, lack of linearity
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 big 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 bigger 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 $1/(X_{Highest}-X_i)$. Like the log transformation, you can't take the reciprocal of 0 (because 1/0 = infinity) so if you have zeros in the data you need to add a constant to all scores before doing the transformation.	Positive skew, positive kurtosis, unequal variances
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: big scores have become small and small scores have become big.	Negative skew

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. When we ran Levene's test (Figure 11) 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 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.

Using SPSS's Compute command

The *compute* command enables us to carry out various functions on columns of data in the data editor. Some typical functions are adding scores across several columns, taking the square root of the scores in a column, or calculating the mean of several variables. To access the *compute* dialog box, use the mouse to select Transform ▦ Compute Variable... . The resulting dialog box is shown in Figure 13; 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. The basic idea is that you type a name for a new variable in the area labelled *Target Variable* and then you write some kind of 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). There are hundreds of built-in functions that SPSS has grouped together. In the dialog box it lists these groups 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 box indicated in Figure 13. 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 Type & Label... 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 your handout from week 1). Then when you have written your command for SPSS to execute, click on OK to run the command and create the new variable. There are functions for calculating means, standard deviations and sums of columns. We're going to use the square root and logarithm functions, which are useful for transforming data that are skewed.



Compute Variable

Target Variable: logday1

Numeric Expression: LG10(day1 + 1)

LG10(numexpr). Numeric. Returns the base-10 logarithm of numexpr, which must be numeric and greater than 0.

If... (optional case selection condition)

Compute Variable: If Cases

○ Include all cases

● Include if case satisfies condition:

gender~=1

Compute Variable: Type and Label

Label

● Label: Log Hygiene Scores for Day 1 o

○ Use expression as label

Type

● Numeric

○ String Width: 8

Figure 13: Dialog box for the *Compute* command

Log Transformation

Let's return to our Download festival data. Open the main Compute dialog box by selecting Transform Compute Variable... . Enter the name **logday1** into the box labelled *Target Variable*, click on Type & Label... 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 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 1. The final dialog box should look like Figure 13. Note that the expression reads LG10(day1 + 1); that is, SPSS will add one to each of the day1 scores and then take the log of the resulting values. Click on OK 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.

Square Root Transformation

To use the square root transformation, we could run through the same process, by using a name such as **sqrtday1** and selecting the *SQRT(numexpr)* function from the list. This will appear in the box labelled *Numeric Expression:* as SQRT(?), and you can simply replace the question mark with the variable you want to change—in this case **day1**. The final expression will read *SQRT(day1)*.

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.

Reciprocal Transformation

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 /. Ordinarily you would select the variable name that you want to transform from the list and drag it across, click on ➡ 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 (). 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 ➡ (or type the name of the variable manually). Now click on + and then 1 (or type *+ 1* using your keyboard). The box labelled *Numeric Expression* should now contain the text *1/(day1 + 1)*. Click on OK 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.

The effect of transformations

Figure 14 shows the distributions for days 1 and 2 of the festival after the three different transformations. Compare these to the untransformed distributions in Figure 9. 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 with, they have now become slightly negatively skewed for the log and square root transformation, and positively skewed for the reciprocal transformation[2] 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☺

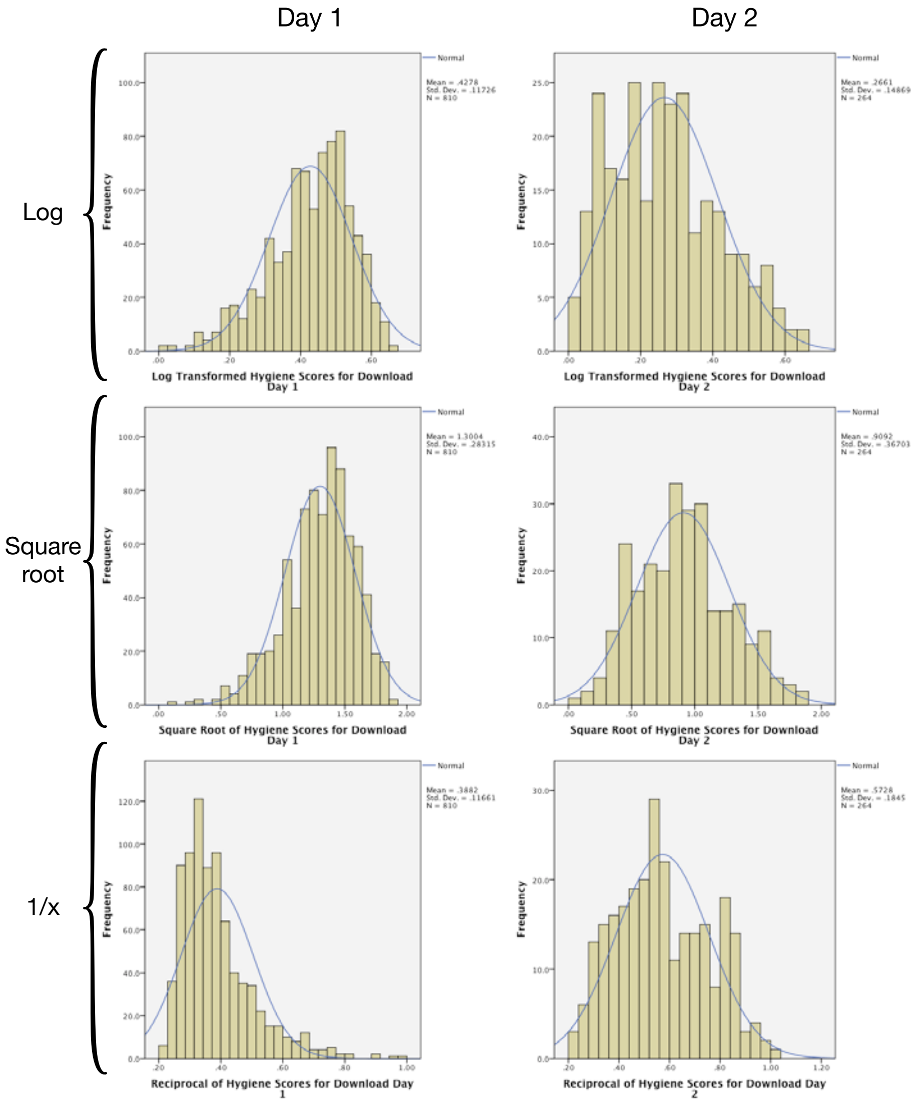


Figure 14: Distributions of the hygiene data on day 1 and day 2 after various transformations

Multiple Choice Test

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

Go to http://www.uk.sagepub.com/field4e/study/mcqs.htm and test yourself on the multiple choice questions for **Chapter 5**. If you get any wrong, re-read this handout (or Field, 2013, Chapter 5) and do them again until you get them all correct.

Acknowledgement

This handout contains material from:

Field, A. P. (2013). *Discovering statistics using SPSS: and sex and drugs and rock 'n' roll* (4th Edition). London: Sage.

References

Efron, B., & Tibshirani, R. (1993). *An introduction to the bootstrap*. Boca Raton, FL: Chapman & Hall/CRC.

Field, A. P. (2000). *Discovering statistics using SPSS for Windows: Advanced techniques for the beginner*. London: Sage.

Field, A. P. (2013). *Discovering statistics using IBM SPSS Statistics: And sex and drugs and rock 'n' roll* (4th ed.). London: Sage.

Gelman, A., & Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models.* Cambridge: Cambridge University Press.

Hayes, A. F., & Cai, L. (2007). Using heteroskedasticity-consistent standard error estimators in OLS regression: An introduction and software implementation. *Behavior Research Methods, 39*(4), 709-722.

Levene, H. (1960). Robust tests for equality of variances. In I. Olkin, S. G. Ghurye, W. Hoeffding, W. G. Madow & H. B. Mann (Eds.), *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling* (pp. 278-292). Stanford, CA: Stanford University Press.

Lumley, T., Diehr, P., Emerson, S., & Chen, L. (2002). The importance of the normality assumption in large public health data sets. *Annual Review of Public Health, 23*, 151-169.

Pearson, E. S., & Hartley, H. O. (1954). *Biometrika tables for statisticians, volume I*. New York: Cambridge University Press.

Wilcox, R. R. (2010). *Fundamentals of modern statistical methods: substantially improving power and accuracy*. New York: Springer.