Exploring Data

Assumptions of Parametric Data

If you use a parametric test and certain assumptions are not met then the results are likely to be inaccurate. Therefore, it is very important that you check the assumptions before deciding which statistical test is appropriate. The assumptions of parametric tests based on the normal distribution are as follows:

- **Normally distributed data**: This assumption is tricky and misunderstood because it means different things in different contexts (see Field, 2009). In short, the rationale behind hypothesis testing relies on having something that is normally distributed: in some cases it's the sampling distribution, in others the errors in the model and so on. If this assumption is not met then your statistical model or test is usually flawed in some way. The reason why this assumption is misunderstood is because people often believe that the data themselves need to be normally distributed; actually, the raw data rarely need to be normally distributed, but if they are then it usually means that the things that we actually want to be normally distributed are too. Confused? Well, it is confusing. In many statistical tests (e.g. the *t*-test) we assume that the sampling distribution is normally distributed. This is a problem because we don't have access to this distribution — we can't simply look at its shape and see whether it is normally distributed. However, we know from the central limit theorem (see my book if you don't know what that is) that if the sample data are approximately normal then the sampling distribution will be also. Therefore, people tend to look at their sample data to see if they are normally distributed. If so, then they have a little party to celebrate and assume that the sampling distribution (which is what actually matters) is also. We also know from the central limit theorem that in big samples the sampling distribution tends to be normal anyway — regardless of the shape of the data we actually collected (and the sampling distribution will tend to be normal regardless of the population distribution in samples of 30 or more). As our sample gets bigger then, we can be more confident that the sampling distribution is normally distributed.
- **Homogeneity of variance**: This assumption means that the variances should be the same throughout the data. In designs in which you test several groups of participants this assumption means that each of these samples come from populations with the same variance. In correlational designs, this assumption means that the variance of one variable should be stable at all levels of the other variable.
- **Interval data**: Data should be measured at least at the interval level. This means that the distance between points of your scale should be equal at all parts along the scale. For example, if you had a 10-point anxiety scale, then the difference in anxiety represented by a change in score from 2 to 3 should be the same as that represented by a change in score from 9 to 10.
- **Independence**: This assumption is that data from different participants are independent, which means that the behaviour of one participant does not influence the behaviour of another. In repeated measures designs (in which participants are measured in more than one experimental condition), we expect scores in the experimental conditions to be non-independent for a given participant, but behaviour between different participants should be independent.

The assumptions of interval data and independent measurements are, unfortunately, tested only by common sense. These assumptions of homogeneity of variance and Normal Distribution can be tested more objectively. The remainder of this handout explain how to explore data and to test these assumptions.

Using Histograms to Spot the Obvious Mistakes

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**: 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. In the example of gender, this is a good way to compare the relative frequency of scores across groups (e.g. were there more smelly women than men?).
- → **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.
- → **Population Pyramid**: Like a stacked histogram this shows the relative frequency of scores in two populations. It plots the variable (in this case hygiene) on the vertical axis and the frequencies for each population on the horizontal: the populations appear back to back on the graph. If the bars either side of the dividing line are equally long then the distributions have equal frequencies.

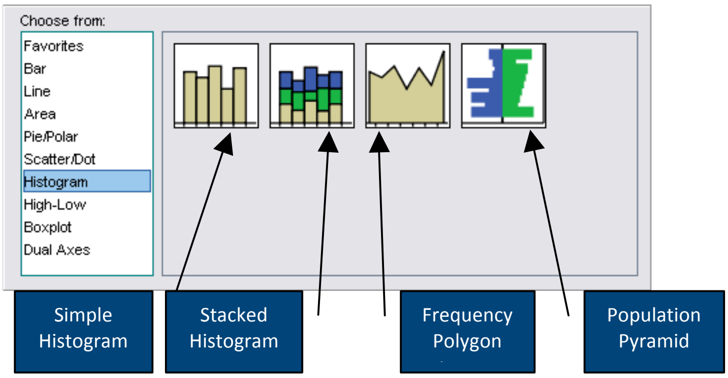


Figure 1: The histogram gallery

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 (top of Figure) because we haven't told SPSS/PASW 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 is an **outlier**: a score very different to the rest. Outliers bias the mean and inflate the standard deviation and screening data is an important way to detect them. 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**.

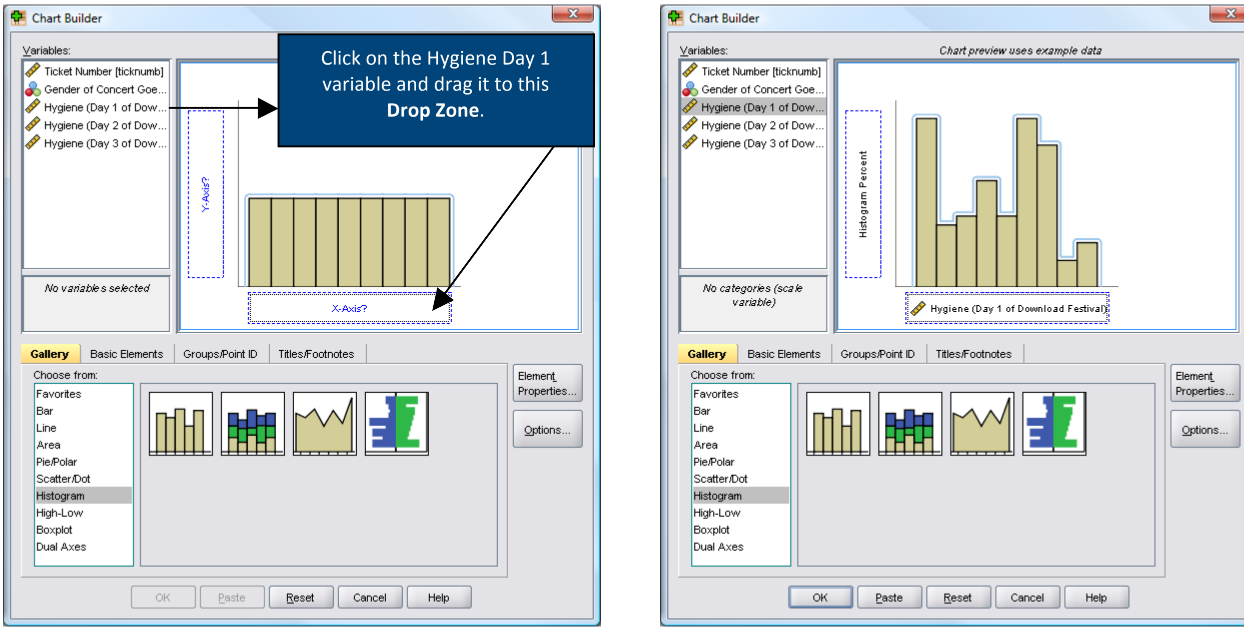

Click on the Hygiene Day 1 variable and drag it to this **Drop Zone**.

Figure 2: Defining a histogram in the Chart Builder

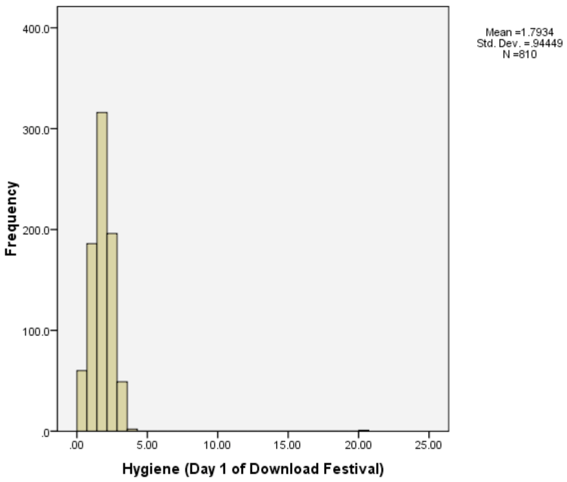


Figure 3

Boxplots

You encountered boxplots or box-whisker diagrams in first 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.

Access the Chart Builder (Graphs Chart Builder...), 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.)

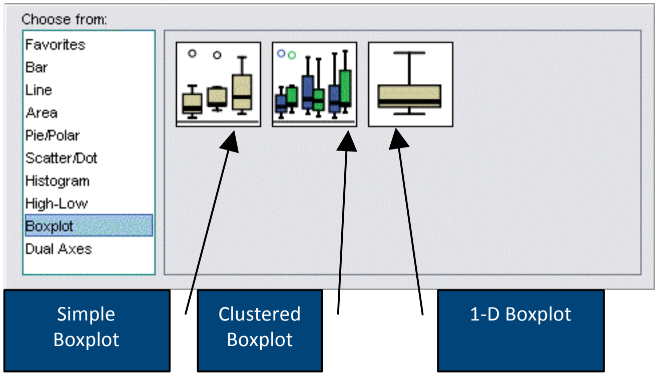
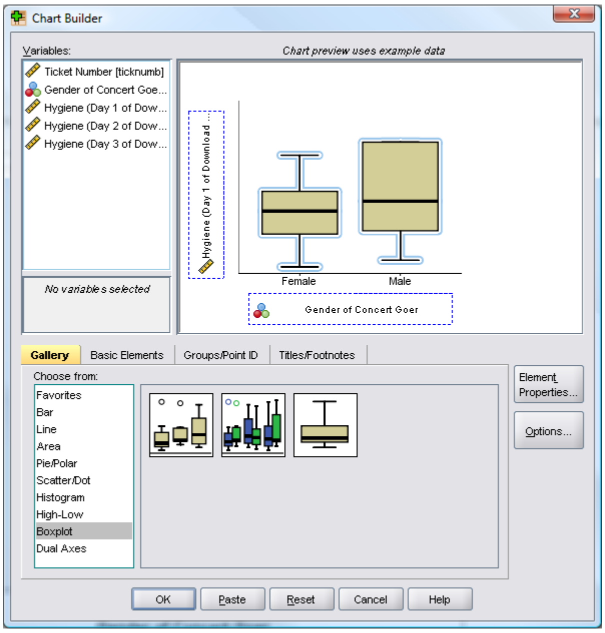


Figure 4: The boxplot gallery

In the data file of hygiene scores we also have information about the gender of the concert-goer. Let's plot this information as well. To make our boxplot of the day 1 hygiene scores for males and females, 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? and select the variable gender and drag it to X-Axis?. The dialog should now look like Figure 5—note that the variable names are displayed in the drop zones, and the canvas now displays a preview of our graph (e.g. there are two boxplots representing each gender). Click on OK to produce the graph.



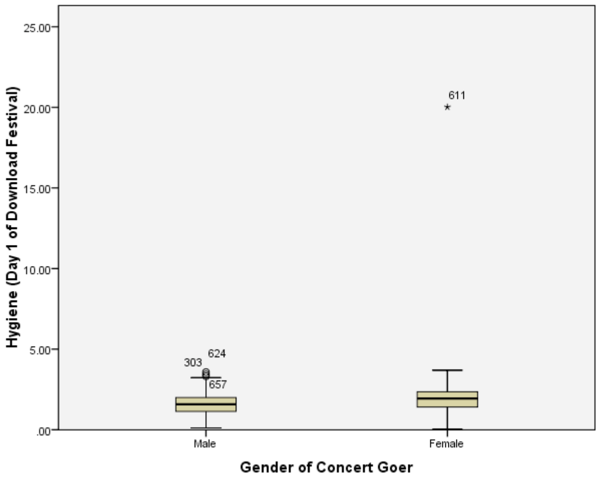


Figure 5: Completed dialog box for a simple boxplot

Figure 6: Boxplot of hygiene scores on day 1 of the Download Festival split by gender

The resulting boxplot is shown in Figure 6. It shows a separate boxplot for the men and women in the data. You may remember that the whole reason that we got into this boxplot malarkey was to help us to identify an outlier from our histogram (if you have skipped straight to this section then you might want to backtrack a bit). The important thing to note is that the outlier that we detected in the histogram is shown up as an asterisk (*) on the boxplot and next to it is the number of the case (611) that's producing this outlier. (We can also tell that this case was a female.) 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 it should be 2.02, so replace the value 20.02 with the value 2.02 before we continue this example.

Now that we've removed the mistake let's re-plot the histogram. While we're at it we should plot histograms for the data from day 2 and day 3 of the festival as well.

Figure 7 shows the resulting three histograms. The first thing to note is that the data from day 1 look a lot more healthy now we've removed the data point that was mis-typed. In fact the distribution is amazingly normal-looking: it is nicely symmetrical and doesn't seem too pointy or flat—these are good things! However, the distributions for days 2 and 3 are not nearly so symmetrical. In fact, they both look positively skewed. This suggests that 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). However, these skewed distributions might cause us a problem if we want to use parametric tests. In the next section we'll look at ways to try to quantify the skewness and kurtosis (i.e. how pointy or flat the distribution is) of these distributions.

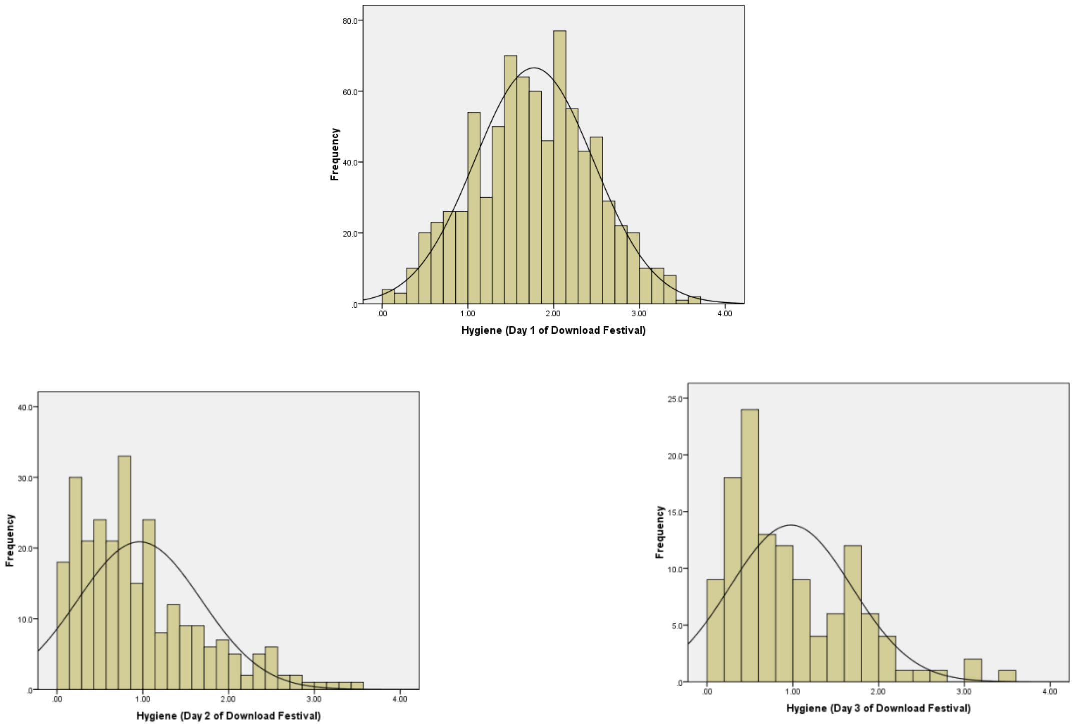


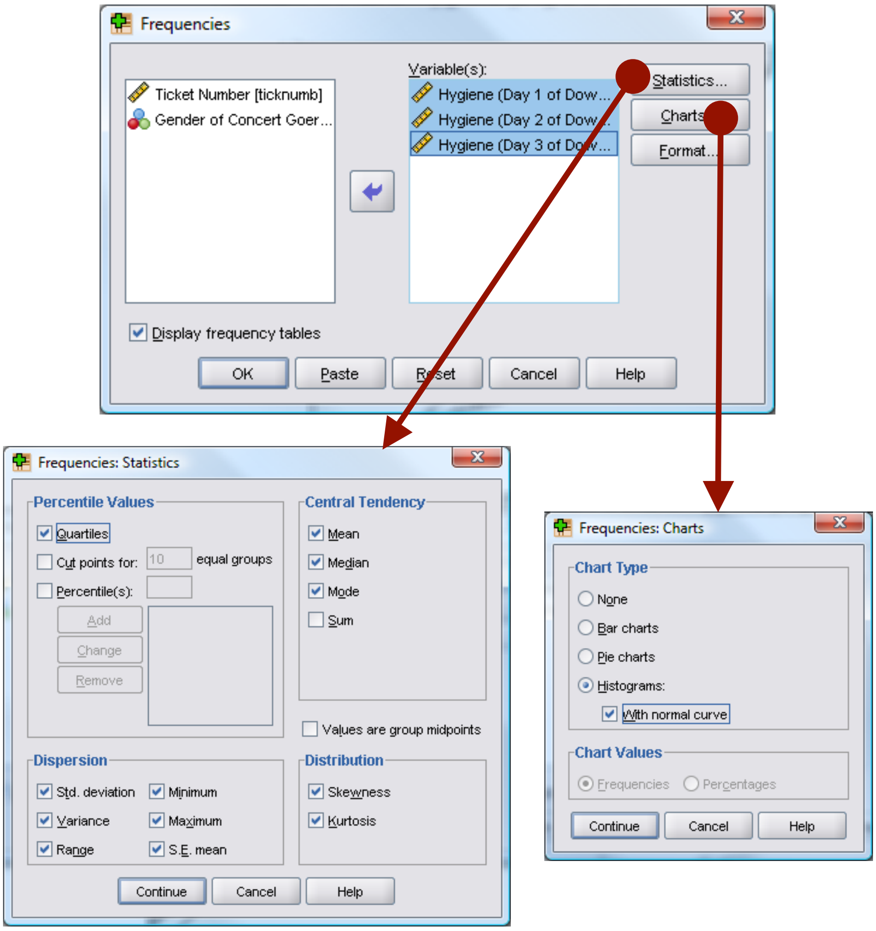
Figure 7: Histograms of the hygiene scores over the three days of the Download Festival

Quantifying normality with numbers

Skew and Kurtosis

To further explore the distribution of the variables, we can use the *frequencies* command (Analyze Descriptive Statistics ▸ 123 Frequencies...). The main dialog box is shown in Figure 8. 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 using the mouse, 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/PASW 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 several ways in which a distribution of scores can be described, such as measures of central tendency (mean, mode, median), measures of variability (range, standard deviation, variance, quartile splits), measures of shape (kurtosis and skewness). To describe the characteristics of the data we should select the mean, mode, median, standard deviation, variance and range. To check that a distribution of scores is normal, we need to look at the values of kurtosis and skewness. 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.



Frequencies

Ticket Number [ticknumb]
Gender of Concert Goer...

Variable(s):
Hygiene (Day 1 of Dow...
Hygiene (Day 2 of Dow...
Hygiene (Day 3 of Dow...

☑ Display frequency tables

Frequencies: Statistics

Percentile Values
☑ Quartiles
☐ Cut points for: 10 equal groups
☐ Percentile(s):

Central Tendency
☑ Mean
☑ Median
☑ Mode
☐ Sum

☐ Values are group midpoints

Dispersion
☑ Std. deviation ☑ Minimum
☑ Variance ☑ Maximum
☑ Range ☑ S.E. mean

Distribution
☑ Skewness
☑ Kurtosis

Frequencies: Charts

Chart Type
○ None
○ Bar charts
○ Pie charts
● Histograms:
☑ With normal curve

Chart Values
● Frequencies ○ Percentages

Figure 8

SPSS/PASW Output 1 shows the table of descriptive statistics for the three variables in this example. From this table, we can see that, 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 day 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.

- ✓ The values of skewness and kurtosis should be 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 distribution whereas negative values indicate a flat distribution.
- ✓ The further the value is from zero, the more likely it is that the data are not normally distributed.

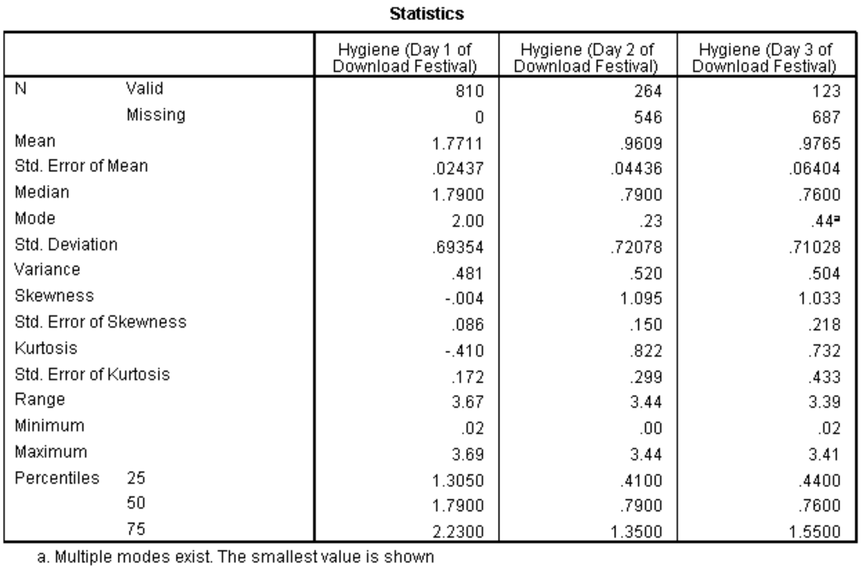
However, the actual values of skewness and kurtosis are not, in themselves, informative. Instead, we need to take the value and convert it to a *z*-score by subtracting the mean (which is, in both cases zero) and dividing by the standard error.

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

In the above equations, the values of *S* (skewness) and *K* (kurtosis) and their respective standard errors are produced by SPSS/PASW. These *z*-scores can be compared against known values for the normal distribution shown in Field (2009). An absolute value greater than 1.96 is significant at $p < .05$, above 2.58 is significant at $p < .01$ and absolute values above about 3.29 are significant at $p < .001$. Large samples will give rise to small standard errors and so when sample sizes are big significant values arise from even small deviations from normality. In most samples it's Ok to look for values above 1.96, however, in large samples this criterion should be increased to 2.58 or 3.29 and in very large samples, because of the problem of small standard errors that I've described, no criterion should be applied at all!

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

Your Answers:



Statistics

		Hygiene (Day 1 of Download Festival)	Hygiene (Day 2 of Download Festival)	Hygiene (Day 3 of Download Festival)
N	Valid	810	264	123
	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

SPSS/PASW Output 1

The Kolmogorov-Smirnov Test

The **Kolmogorov-Smirnov** and **Shapiro-Wilk** tests 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, they have their limitations because with large sample sizes it is very easy to get significant results from small deviations from normality. The take home message is that by all means use these tests, but plot your data as well and try to make an informed decision about the extent of non-normality.

The Kolmogorov-Smirnov (K-S from now on) test can be accessed through the explore command (Analyze Descriptive Statistics ▸ Explore...). Figure 9 shows the dialog boxes for the explore 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 ➡ (or dragging them across). For this example, select the hygiene scores for each day. It is also possible to select a factor (or grouping variable) by which to split the output (so, we could transfer gender into the box labelled Factor List and SPSS/PASW will produce exploratory analysis for each group—a bit like the split file command). If you click on Statistics... a dialog box appears, but the default option is fine. The more interesting option for our 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. By default, SPSS/PASW will produce boxplots (split according to group if a factor has been specified) and stem and leaf diagrams as well. If you click on Options... you'll see that, by default, SPSS/PASW 'excludes cases listwise'. This means that if a case of data is missing for any of the variables we have selected it will exclude that case entirely! This is not desirable here because for hygiene on day 1 (for example) we want to see the distribution of all people, not just the people who were tested on all three days. So, change this option to 'Exclude cases pairwise' and then for each variable we analyse, SPSS/PASW will exclude people only if there is not data for them on that particular variable. Click on Continue to return to the main dialog box and then click OK to run the analysis.

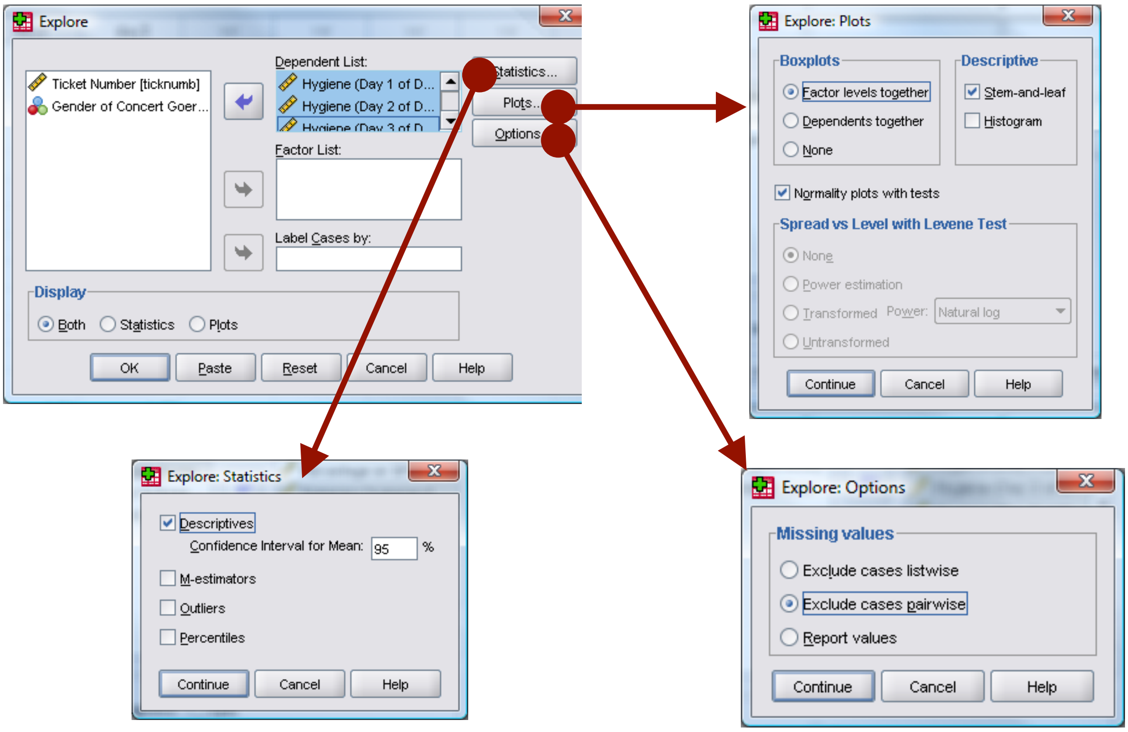
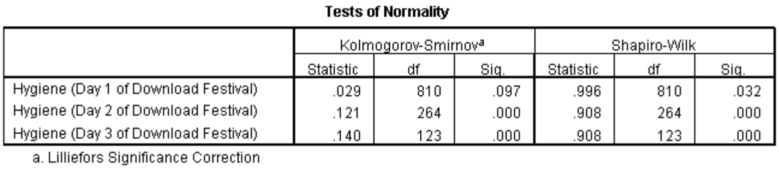


Figure 9

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

Your Answers:

The first table produced by SPSS/PASW contains descriptive statistics (mean etc.) and should have the same values as the tables obtained using the frequencies procedure earlier. The important table is that of the Kolmogorov-Smirnov test (SPSS/PASW Output 2). 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. The K-S test is highly significant for the last two distributions indicating that they are not normal.



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

SPSS/PASW Output 2

The test statistic for the K-S test is denoted by *D* and so we can report these results as: The hygiene scores on day 1, *D*(810) = 0.03, *ns*, did not significantly deviate from normality; however, day 2, *D*(264) = 0.12, *p* < .001, and day 3, *D*(123) = 0.14, *p* < .001 scores were significantly non-normal. The numbers in brackets are the degrees of freedom (*df*) from the table.

Homogeneity of Variance

Different statistical procedures have their own unique way to look for homogeneity of variance: in correlational analysis such a regression we tend to use graphs (see lectures/handouts on Multiple Regression) and for groups of data we tend to use a test called **Levene's test**. Levene's test tests the hypothesis that the variances in the groups are equal (i.e. the difference between the variances is zero). Therefore,

- If Levene's test is significant at $p \leq .05$ then we can 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 we must accept the null hypothesis that the difference between the variances is zero—the variances are roughly equal and the assumption is tenable.

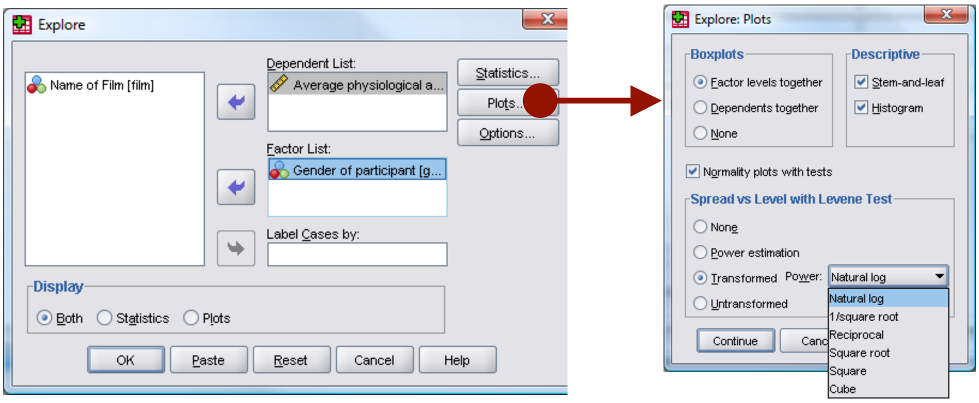
Although Levene's test can be selected as an option in many of the statistical tests that require it, it can also be examined when you're exploring data (and strictly speaking it's better to examine it now than wait until your main analysis). However, as with the K-S test (and other tests of normality), when the samples size is large, small differences in group variances can produce a Levene's test that is significant when the variances are not particularly different.

What is the assumption of homogeneity of variance?

Your Answers:

We can get Levene's test using the *explore* menu that we used in the previous section. For this example, well use the chick flick data from last week (if you didn't save your own file last week, you can find the data in the file

ChickFlick.sav). Once the data are loaded, use Analyze Descriptive Statistics ▸ Explore... to open the dialog box in Figure 7. Transfer the **arousal** variable from the list on the left-hand side to the box labelled *Dependent List:* by clicking the ➡ next to this box, and 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 10. To get Levene's test we need to select one of the options where it says *Spread vs. Level with Levene's test*. If you select ○ Untransformed the 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 OK to run the analysis.



Explore

Name of Film [film]

Dependent List: Average physiological a...

Factor List: Gender of participant [g...

Label Cases by:

Display: ● Both ○ Statistics ○ Plots

Explore: Plots

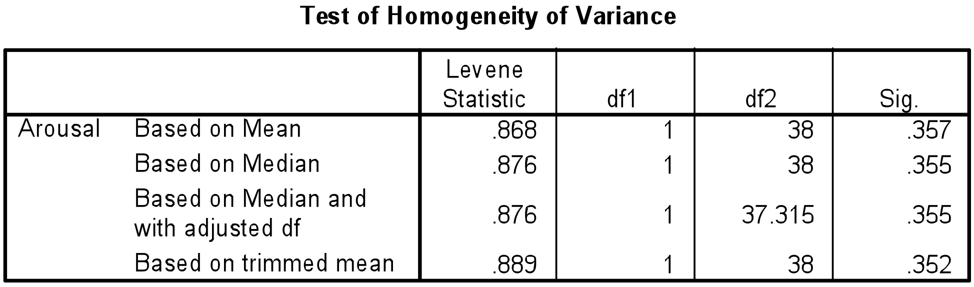
Boxplots: ● Factor levels together ○ Dependents together ○ None

Descriptive: ☑ Stem-and-leaf ☑ Histogram

☑ Normality plots with tests

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

Figure 10



Test of Homogeneity of Variance

		Levene Statistic	df1	df2	Sig.
Arousal	Based on Mean	.868	1	38	.357
	Based on Median	.876	1	38	.355
	Based on Median and with adjusted df	.876	1	37.315	.355
	Based on trimmed mean	.889	1	38	.352

SPSS/PASW Output 3

SPSS/PASW Output 3 shows the results. I'm not going to dwell on this because we will come across Levene's test in future seminars, so we'll just note that Levene's test is nonsignificant (values in the column labelled *Sig.* are more than .05) indicating that the variances are not significantly different (i.e. they are similar and the homogeneity of variance assumption is tenable).

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*. So, we could say the variances are equal, *F*(1, 38) = 0.87, *ns*.

Correcting problems in the Data

If your data are not normally distributed, there are things you can do to try to correct the problem. The main thing is transforming the data. 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 data. As such, transforming the data won't change the relationships between variables (the relative differences between people for a given variable stay the same), however, it does change the differences between different variables (because it changes the units of measurement). Therefore, even if you have only one variable that has a skewed distribution, you should still transform any other variables in your data set if you're going to compare differences between that variable and others that you intend to transform. So, for example, with our hygiene data, we might want to look at how hygiene levels changed across the three days (that is, compare the mean on day 1 to the means on days 2 and 3).

The data for day 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 it isn't 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 simply be due to us transforming one variable and not the others.

When you have positively skewed data (as we do) that needs correcting, there are three transformations that can be useful (and as you'll see these can be adapted to deal with negatively skewed data too):

1. **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. However, you can't get a Log value of zero or negative numbers, so if your data include any zeros 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.

2. **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 a positively-skewed data; however, you still have the same problem with negative numbers (negative numbers don't have a square root).

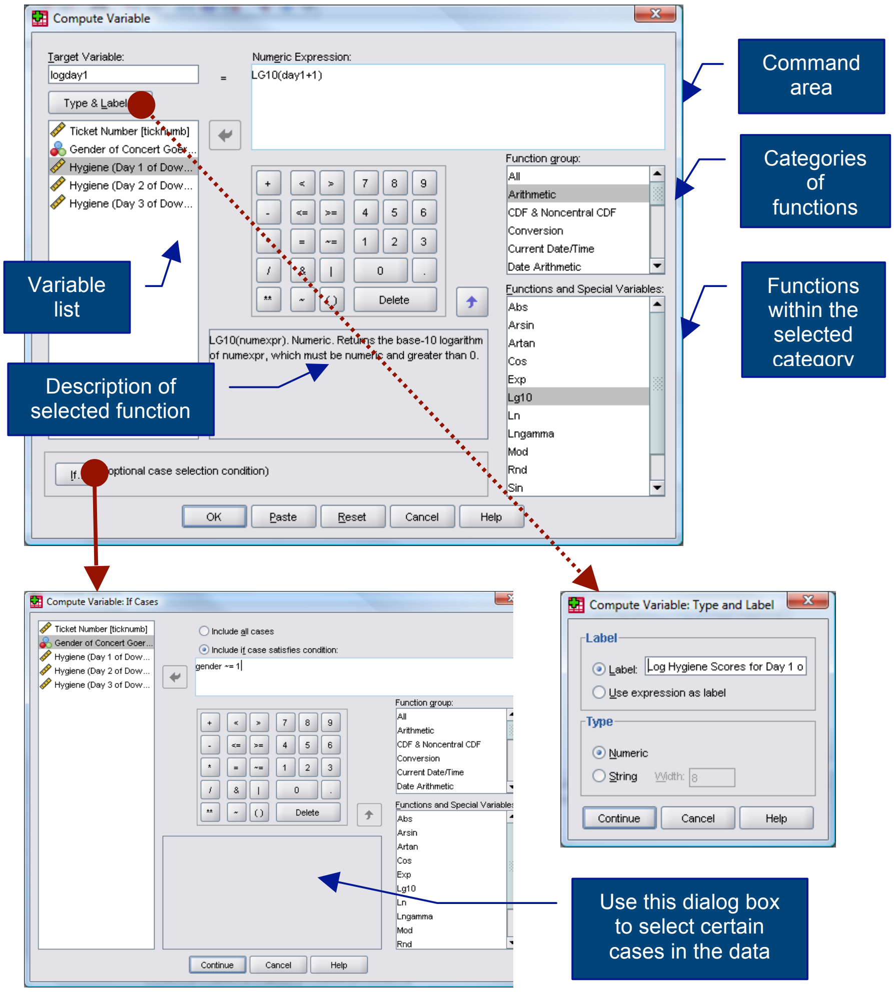
3. **Reciprocal Transformation (1/X_i)**: Dividing 1 by each score also reduces the impact of large scores. The transformed variable will have a lower limit of 0 (very large numbers will become close to zero). One thing to bear in mind with this transformation is that it reverses the scores in the sense that 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/($X_{Highest} - X_i$).

Any one of these transformations can also be used to correct negatively skewed data, but first you'd have to reverse the scores (that is, subtract each score from the highest score obtained)—if you do this, don't forget to reverse the scores back afterwards!

Using SPSS/PASW'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 11; 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/PASW 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/PASW 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 grey box indicated in Figure 11. 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 ↑.

The basic procedure is to first type a variable name in the box labelled *Target Variable*. You can then click on Type & Label... and another dialog box appears, where you can give the variable a descriptive label, and where you can specify whether it is a numeric or string variable (see your handout from week 1). Then when you have written your command for SPSS/PASW 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 (see Field, 2009, for more detail).



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.

Function group: Arithmetic

Functions and Special Variables: Lg10

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 11: Dialog box for the *compute* function

Log Transformation

Let's return to our Download festival data. To transform these data, first open the main compute dialog box by selecting Transform Compute Variable... . Enter the name **logday1** into the box labelled Target Variable and then 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.

There is a value of 0 in the original data (on day 2), and there is no logarithm of the value 0. To overcome this we should add a constant to our original scores before we take the log of those scores. Any constant will do, provided that it makes all of the scores greater than 0. In this case our lowest score is 0 in the data set so we can simply add 1 to all of the scores and that will ensure that all scores are greater than zero. To do this, make sure the cursor is still inside the brackets and click on + and then 1. The final dialog box should look like Figure 11. Note that the expression reads LG10(day1 + 1); that is, SPSS/PASW 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.

Now you have a go at creating similar variables **logday2** and **logday3** for the day 2 and day 3 data!

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)*.

Try repeating this for **day2** and **day3** to create variables called **sqrtday2** and **sqrtday3**.

Reciprocal Transformation

Finally, if we wanted to use a reciprocal transformation on the data from day 1, we could use a name such as **recday1** and then 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). As such 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.

Try repeating this for **day2** and **day3** to create variables called **recday2** and **recday3**.

The Effect of Transforming Data

Plot Histograms of your transformed values!

Let's have a look at our new distributions (Figure 12). Compare these with the untransformed distributions in Figure 7. Now, you can see that all three transformations have cleaned up the hygiene scores for day 2: the positive skew is gone (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[1]! If we're using scores from day 2 alone then we could use the transformed scores, however, if we want 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 outweighs the problems it creates in the day 1 scores—data analysis can be frustrating sometimes!

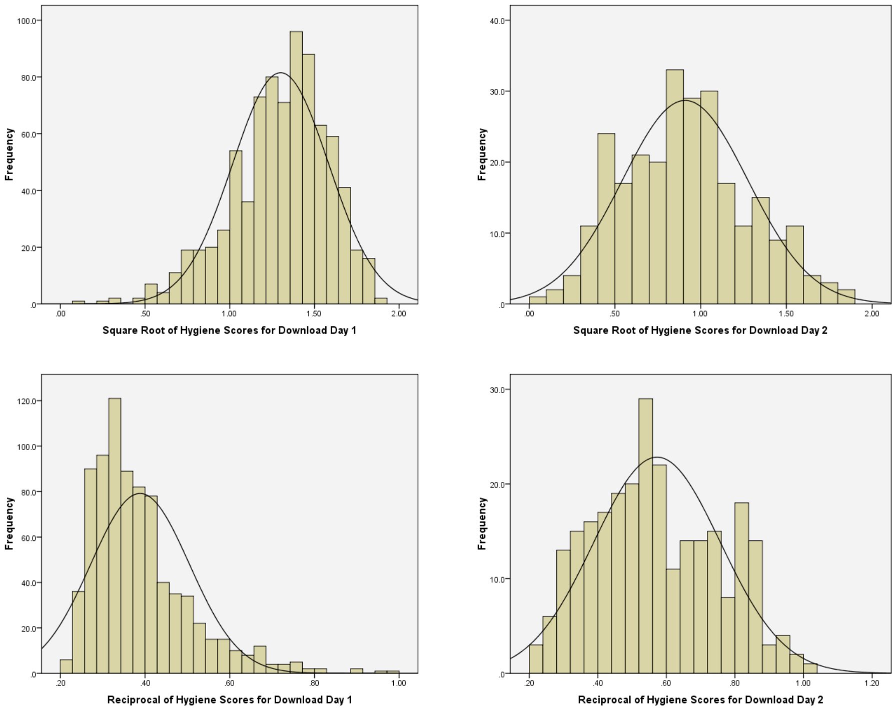


Figure 12: Day 1 (left) and day 22 (right) hygiene scores from the Download Festival

Multiple Choice Test

Go to http://www.uk.sagepub.com/field3e/MCQ.htm and test yourself on the multiple choice questions for **Chapter 5**. If you get any wrong, re-read this handout (or Field, 2009, Chapter 5) and do them again until you get them all correct.

This handout contains material from:

Field, A. P. (2009). *Discovering statistics using SPSS: and sex and drugs and rock 'n' roll (3rd Edition)*. London: Sage.

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