Chapter 5: The beast of bias

Self-test answers

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

First we need to compute the mean:

$$
\begin{aligned}
\bar{X} &= \frac{\sum_{i=1}^{n} x_i}{n} \\
&= \frac{1 + 3 + 10 + 3 + 2}{5} \\
&= \frac{19}{5} \\
&= 3.8
\end{aligned}
$$

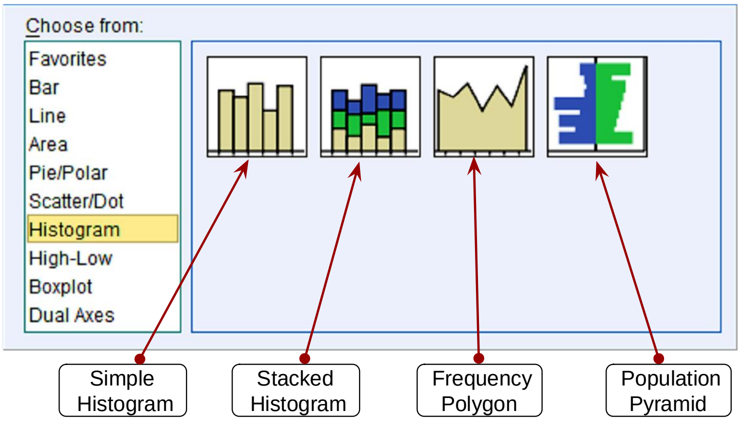
Then the sum of squared error:

Score	Mean	Error (Score – Mean)	Error Squared
1	3.8	–2.8	7.84
3	3.8	–0.8	0.64
10	3.8	6.2	38.44
3	3.8	–0.8	0.64
2	3.8	–1.8	3.24

Sum of squared error = 7.84 + 0.64 + 38.44 + 0.64 + 3.24 = 50.8.

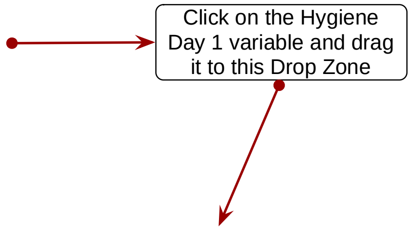
SELF-TEST Using what you learnt in Section 4.4, plot a histogram of the hygiene scores on day 1 of the festival.

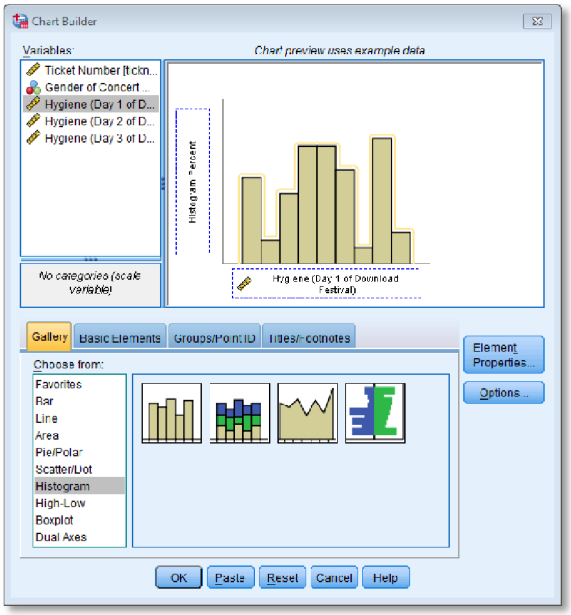
First, access the Chart Builder and then select *Histogram* in the list labelled *Choose from* to bring up the gallery shown below.



The histogram gallery

We are going to do a simple histogram, so double-click on the icon for a simple histogram. The *Chart Builder* dialog box will now show a preview of the graph in the canvas area. Click on the hygiene day 1 variable in the list and drag it to X-Axis? as shown below; you will now find the histogram previewed on the canvas. To draw the histogram click on OK.

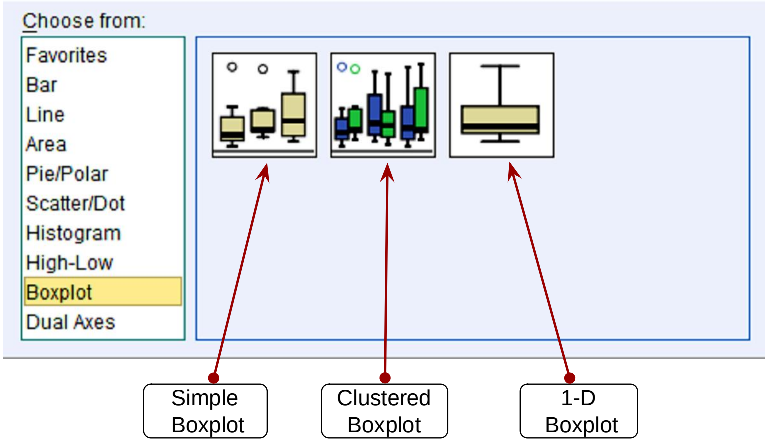

Click on the Hygiene Day 1 variable and drag it to this Drop Zone



Defining a histogram in the Chart Builder

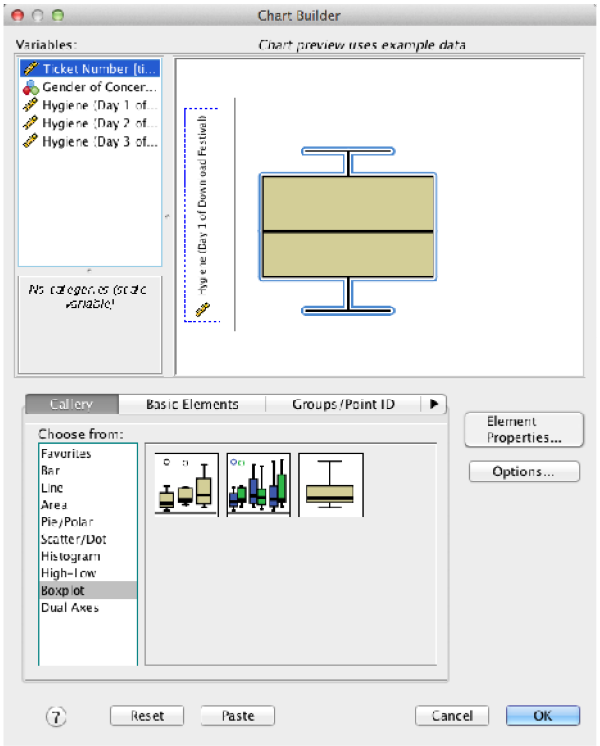
SELF-TEST Using what you learnt in Section 4.5, plot a boxplot of the hygiene scores on day 1 of the festival.

In the Chart Builder select *Boxplot* in the list labelled *Choose from* to bring up the gallery shown below.



The boxplot gallery

There are three types of boxplot you can choose; we just need a *simple boxplot* because we want to plot a boxplot of a single variable. Select this option by double-clicking on the *simple boxplot* icon, then from the variable list select the hygiene day 1 score variable and drag it into X-Axis? drop zone. The dialog should now look like the image below — note that the variable name is displayed in the drop zone, and the canvas now displays a preview of our graph. Click on OK to produce the graph.



Completed dialog box for a simple boxplot

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

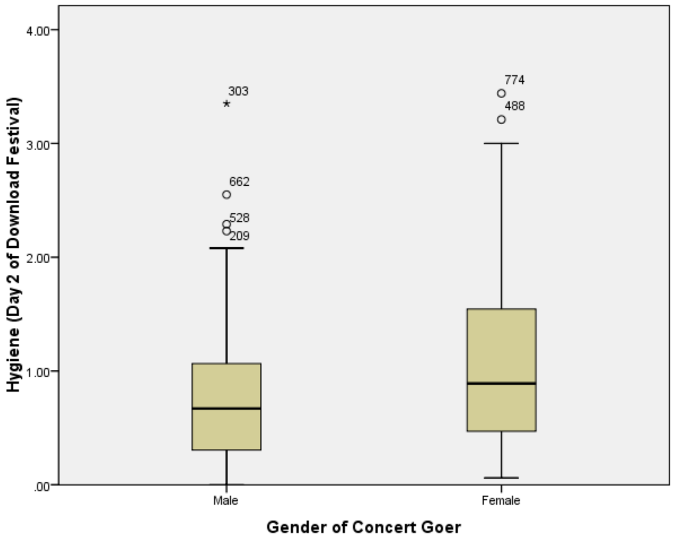
SELF-TEST Re-plot them but splitting by **Gender** along the *x*-axis. Are there differences between men and women?

The boxplots for days 2 and 3 should look like this:

Day 2 **Day 3**

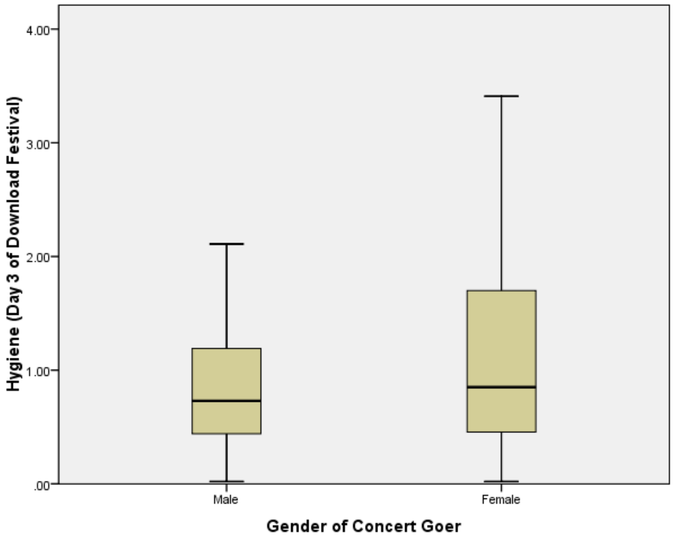
On day 2 there are 6 scores that are deemed to be mild outliers (greater than 1.5 times the interquartile range) and on day 3 there is only 1 score deemed to be a mild outlier (case 774). We should consider whether to take action to reduce the impact of these scores. More generally, the fact that the top whisker is longer than the bottom one for both graphs indicates skew in the distribution. There's more on that topic in the chapter.

After splitting by gender, the boxplot for the day 2 data should look like this:



Note that, as for day 1, the females are slightly more fragrant than males (look at the median line). However, if you compare these to the day 1 boxplots (in the book) scores are getting lower (i.e. people are getting less hygienic). In the males there are now more outliers (i.e. a rebellious few who have maintained their sanitary standards).

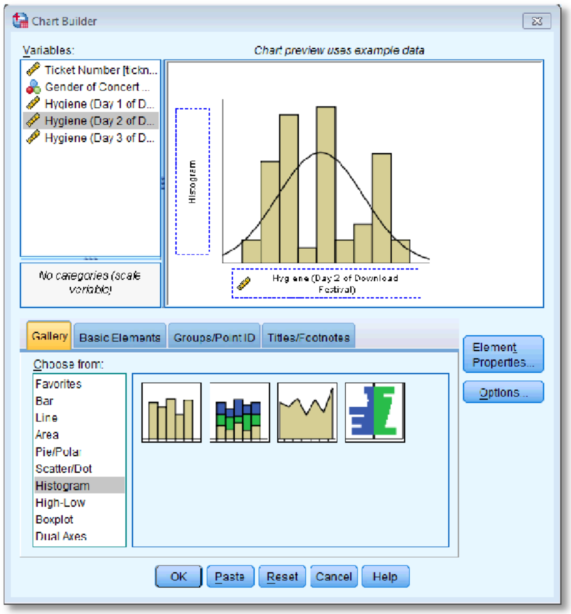
The boxplot for the day 3 data should look like this:



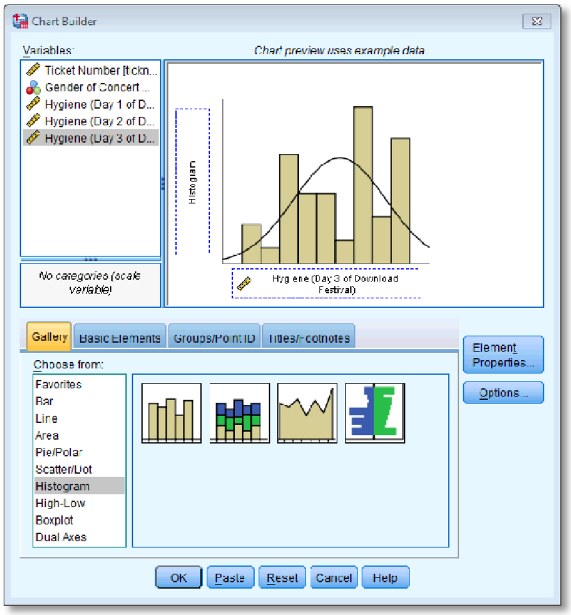
Note that compared to day 1 and day 2, the females are getting more like the males (i.e., smelly). However, if you look at the top whisker, this is much longer for the females. In other words, the top portion of females are more variable in how smelly they are compared to males. Also, the top score is higher than for males. So, at the top end females are better at maintaining their hygiene at the festival compared to males. Also, the box is longer for females, and although both boxes start at the same score, the top edge of the box is higher in females, again suggesting that above the median score more women are achieving higher levels of hygiene than men. Finally, note that for both days 1 and 2, the boxplots have become less symmetrical (the top whiskers are longer than the bottom whiskers). On day 1 (see the book chapter), which is symmetrical, the whiskers on either side of the box are of equal length (the range of the top and bottom scores is the same); however, on days 2 and 3 the whisker coming out of the top of the box is longer than that at the bottom, which shows that the distribution is skewed (i.e., the top portion of scores is spread out over a wider range than the bottom portion).

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

First, access the Chart Builder as in Chapter 4 of the book and then select *Histogram* in the list labelled *Choose from:* to bring up the gallery, which has four icons representing different types of histogram. We want to do a simple histogram, so double-click on the icon for a simple histogram. The *Chart Builder* dialog box will now show a preview of the graph in the canvas area. To plot the histogram of the day 2 hygiene scores select the hygiene day 2 variable from the list and drag it into the X-Axis? drop zone. To draw the histogram click on OK:



To plot the day 3 scores go back to the Chart Builder but this time select the hygiene day 3 variable from the list and drag it into the X-Axis? drop zone and click on OK:

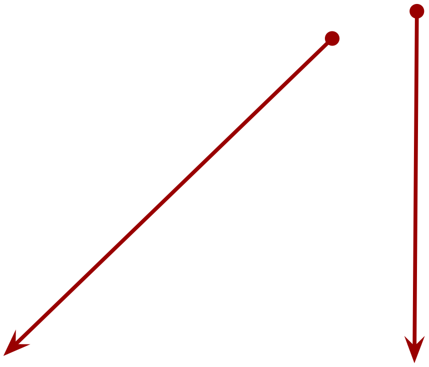


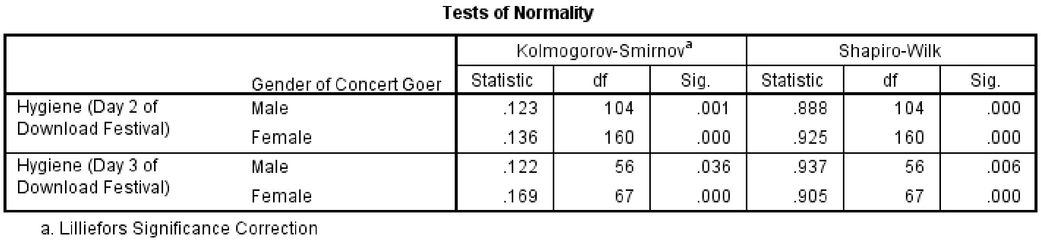
See Figure 5.12 in the book for the histograms of all three days of the festival.

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

The K-S test is accessed through the *explore* command (Analyze Descriptive Statistics ▸ Explore...). First, enter the hygiene scores for days 2 and 3 in the box labelled *Dependent List* by highlighting them and transferring them by clicking on ➡. The question asks us to look at the K-S test for males and females separately, therefore we need to select **Gender** and transfer it to the box labelled *Factor List* so that SPSS will produce exploratory analysis for each group – a bit like the *split file* command. Next, click on Plots... and select the option ☑ Normality plots with tests; this will produce both the K-S test *normal Q-Q plots*. A Q-Q plot plots the quantiles of the data set. If the data are normally distributed, then the observed values (the dots on the chart) should fall exactly along the straight line (meaning that the observed values are the same as you would expect to get from a normally distributed data set). Kurtosis is shown up by the dots sagging above or below the line, whereas skew is shown up by the dots snaking around the line in an 'S' shape.

We also need to click on Options... to tell SPSS how to deal with missing values. We want 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:





Tests of Normality

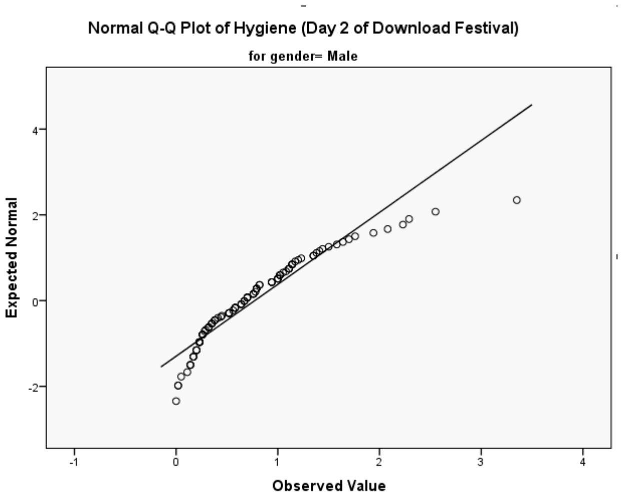
		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Gender of Concert Goer	Statistic	df	Sig.	Statistic	df	Sig.
Hygiene (Day 2 of Download Festival)	Male	.123	104	.001	.888	104	.000
	Female	.136	160	.000	.925	160	.000
Hygiene (Day 3 of Download Festival)	Male	.122	56	.036	.937	56	.006
	Female	.169	67	.000	.905	67	.000

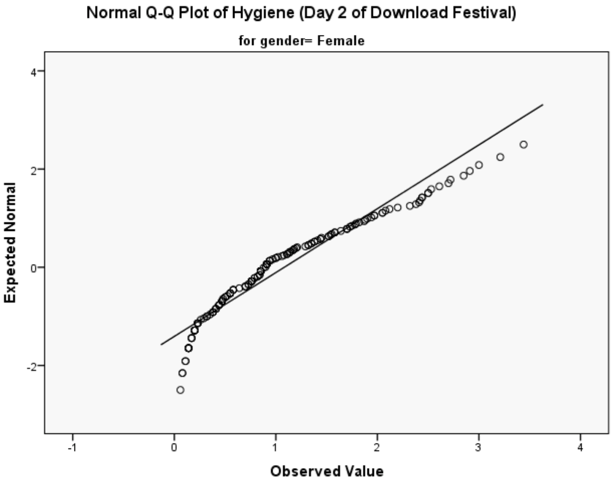
a. Lilliefors Significance Correction

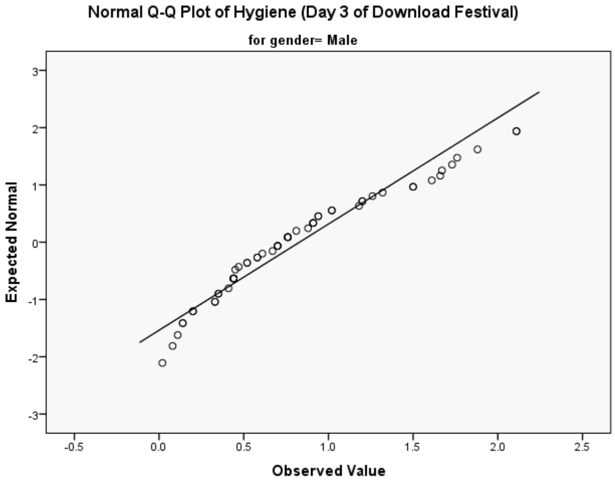
You should get the table above in your SPSS output, which shows that the distribution of hygiene scores on both days 2 and 3 of the Download Festival were significantly different from normal for both males and females (all values of *Sig.* are less than .05).

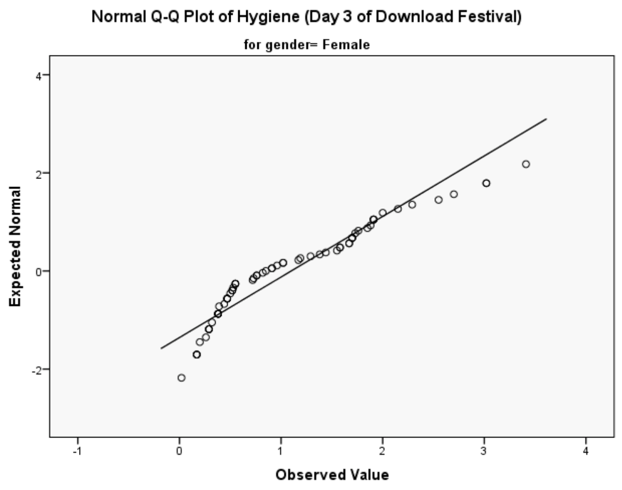
The normal Q-Q charts below plot the values you would expect to get if the distribution were normal (expected values) against the values actually seen in the data set (observed values). If we first look at the Q-Q plots for day 2, we can see that the plots for males and females are very similar: the quantiles do not fall close to the diagonal line, indicating a non-normal distribution; the quantiles sag below the line, suggesting a problem with kurtosis (this appears to be more of a problem for males than for females), and they have an ‘S’ shape, indicating skew. All this is not surprising given the significant K-S tests above.

The Q-Q plot for females on day 3 is very similar to that of day 2. However, for males the Q-Q plot for day 3 now indicates a more normal distribution. The quantiles fall closer to the line and there is less sagging below the line. This makes sense as the K-S test for males on day 3 was close to being non-significant, $D(56) = 0.12$, $p = .04$.

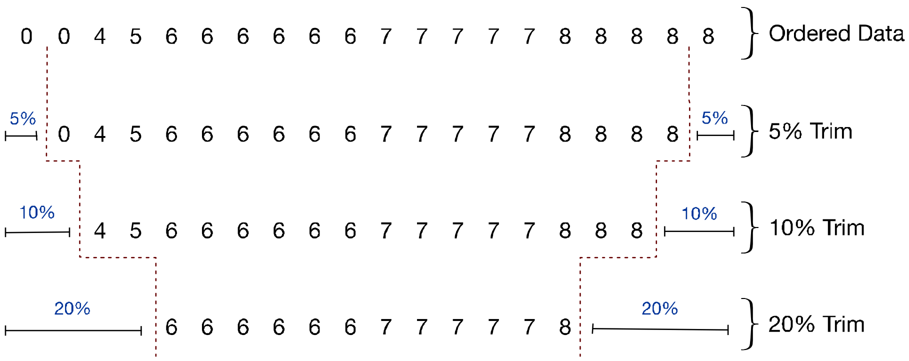








SELF-TEST Compute the mean and variance of the attractiveness ratings. Now compute them for the 5%, 10% and 20% trimmed data.



0 0 4 5 6 6 6 6 6 6 7 7 7 7 7 8 8 8 8 8 } Ordered Data

5% 0 4 5 6 6 6 6 6 6 7 7 7 7 7 8 8 8 8 5% } 5% Trim

10% 4 5 6 6 6 6 6 6 7 7 7 7 7 8 8 8 10% } 10% Trim

20% 6 6 6 6 6 6 7 7 7 7 7 8 20% } 20% Trim

To calculate the mean of the attractiveness ratings we use the equation:

$$
\begin{aligned}
\bar{X} &= \frac{\sum_{i=1}^{n} x_i}{n} \\
&= \frac{0+0+4+5+6+6+6+6+6+6+7+7+7+7+7+8+8+8+8+8}{20} \\
&= \frac{120}{20} \\
&= 6
\end{aligned}
$$

To calculate the variance we use the equation:

$$s^2 = \frac{\text{sum of squares}}{N-1}$$

The table below shows you how to calculate the squared error for each score.

Score	Error (Score – Mean)	Error Squared
0	–6	36
0	–6	36

4	–2	4
5	–1	1
6	0	0
6	0	0
6	0	0
6	0	0
6	0	0
6	0	0
7	1	1
7	1	1
7	1	1
7	1	1
7	1	1
8	2	4
8	2	4
8	2	4
8	2	4
8	2	4

Sum of squared errors = 36 + 36 + 4 + 1 + 0 + 0 + 0 + 0 + 0 + 0 + 1 + 1 + 1 + 1 + 1 + 4 + 4 + 4 + 4 + 4 = 102.

$$s^2 = \frac{\text{sum of squares}}{N - 1} = \frac{102}{19} = 5.37$$

Therefore the variance for the attractiveness data is 5.37.

Next, let's calculate the mean and variance for the 5% trimmed data. We basically do the same thing as before but delete 1 score at each extreme (there are 20 scores and 5% of 20 is 1). Therefore the mean would be:

$$\begin{aligned}\bar{X} &= \frac{\sum_{i=1}^{n} x_i}{n} \\ &= \frac{0 + 4 + 5 + 6 + 6 + 6 + 6 + 6 + 6 + 7 + 7 + 7 + 7 + 7 + 8 + 8 + 8 + 8}{18} \\ &= \frac{112}{18} \\ &= 6.22\end{aligned}$$

Variance:

$$s^2 = \frac{\text{sum of squares}}{N - 1}$$

Score	Error (Score – Mean)	Error Squared
0	–6.22	38.69
4	–2.22	4.93
5	–1.22	1.49
6	0.22	0.05
6	0.22	0.05
6	0.22	0.05
6	0.22	0.05
6	0.22	0.05
6	0.22	0.05
7	0.78	0.61
7	0.78	0.61
7	0.78	0.61
7	0.78	0.61
7	0.78	0.61
8	1.78	3.17
8	1.78	3.17
8	1.78	3.17
8	1.78	3.17

Sum of squared errors for 5% trimmed data = 61.11.

$$s^2 = \frac{\text{sum of squares}}{N - 1} = \frac{61.11}{17} = 3.59$$

Therefore the variance for the 5% trimmed data = 3.59.

Next, let's calculate the mean and variance for the 10% trimmed data. To do this we need to delete 2 scores from each extreme of the original data set (there are 20 scores and 10% of 20 is 2). Therefore the mean would be:

$$\begin{aligned}\bar{X} &= \frac{\sum_{i=1}^{n} x_i}{n} \\ &= \frac{4 + 5 + 6 + 6 + 6 + 6 + 6 + 6 + 7 + 7 + 7 + 7 + 7 + 8 + 8 + 8}{16} \\ &= \frac{104}{16} \\ &= 6.50\end{aligned}$$

To calculate the variance:

$$s^2 = \frac{\text{sum of squares}}{N - 1}$$

Score	Error (Score – Mean)	Error Squared
4	–2.5	6.25
5	–1.5	2.25
6	–0.5	0.25
6	–0.5	0.25
6	–0.5	0.25
6	–0.5	0.25
6	–0.5	0.25
6	–0.5	0.25
7	0.5	0.25
7	0.5	0.25
7	0.5	0.25
7	0.5	0.25
7	0.5	0.25
8	1.5	2.25
8	1.5	2.25
8	1.5	2.25

Sum of squared errors for 10% trimmed data = 18.

$$s^2 = \frac{\text{sum of squares}}{N - 1} = \frac{18}{15} = 1.20$$

Therefore the variance for the 10% trimmed data = 1.20.

Finally, let's calculate the mean and variance for the 20% trimmed data. To do this we need to delete 4 scores from each extreme of the original data set (there are 20 scores and 20% of 20 is 4). Therefore the mean would be:

$$\begin{aligned}\bar{X} &= \frac{\sum_{i=1}^{n} x_i}{n}\\ &= \frac{6+6+6+6+6+6+7+7+7+7+7+8}{12}\\ &= \frac{79}{12}\\ &= 6.58\end{aligned}$$

To calculate the variance:

$$s^2 = \frac{\text{sum of squares}}{N - 1}$$

Score	Error	Error Squared

	(Score – Mean)	
6	–0.58	0.34
6	–0.58	0.34
6	–0.58	0.34
6	–0.58	0.34
6	–0.58	0.34
6	–0.58	0.34
7	0.42	0.18
7	0.42	0.18
7	0.42	0.18
7	0.42	0.18
7	0.42	0.18
8	1.42	2.02

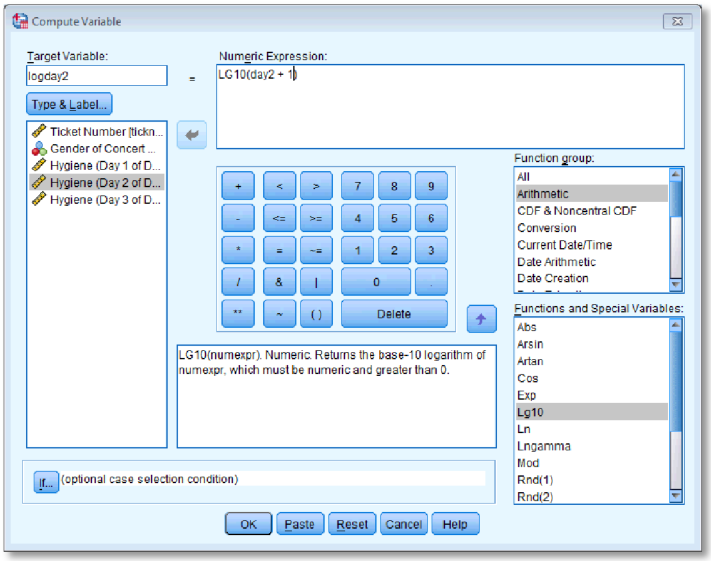
Sum of squared errors for 20% trimmed data = 4.92.

$$s^2 = \frac{\text{sum of squares}}{N - 1} = \frac{4.92}{11} = 0.45$$

Therefore the variance for the 20% trimmed data = 0.45.

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.

The completed *Compute Variable* dialog boxes for day 2 and day 3 should look as below:



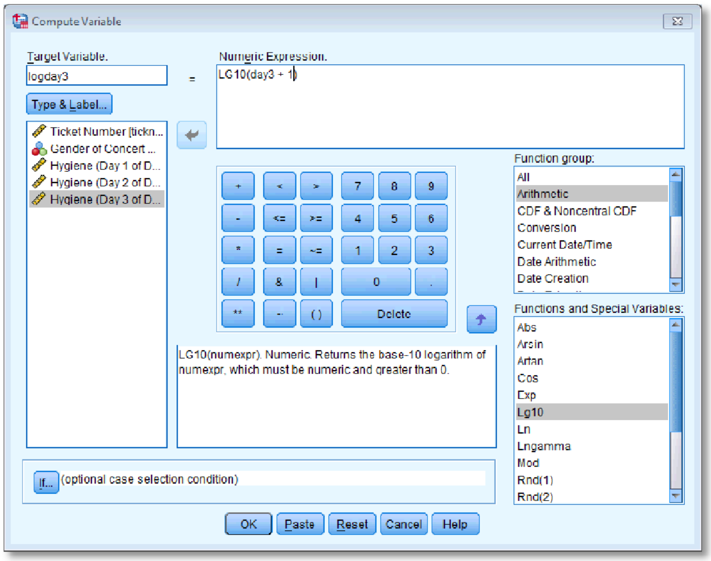
Target Variable: logday2

Numeric Expression: LG10(day2 + 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



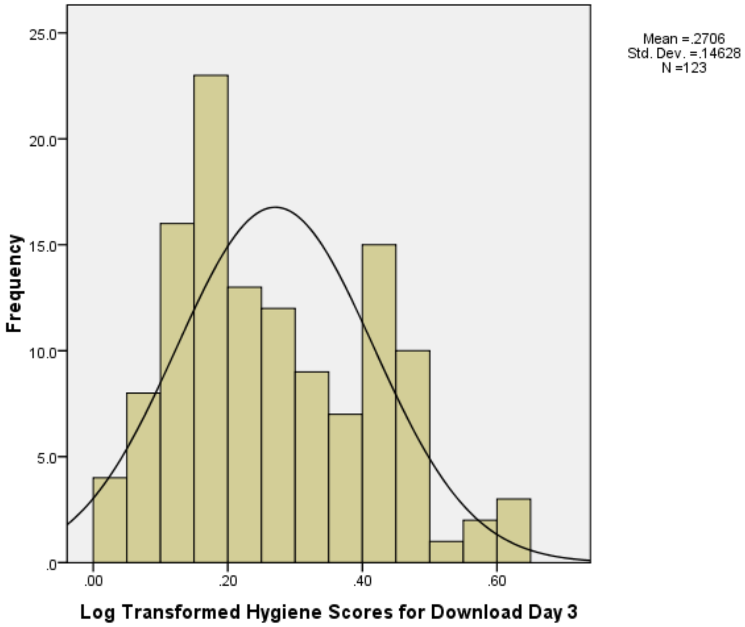
Target Variable: logday3

Numeric Expression: LG10(day3 + 1)

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

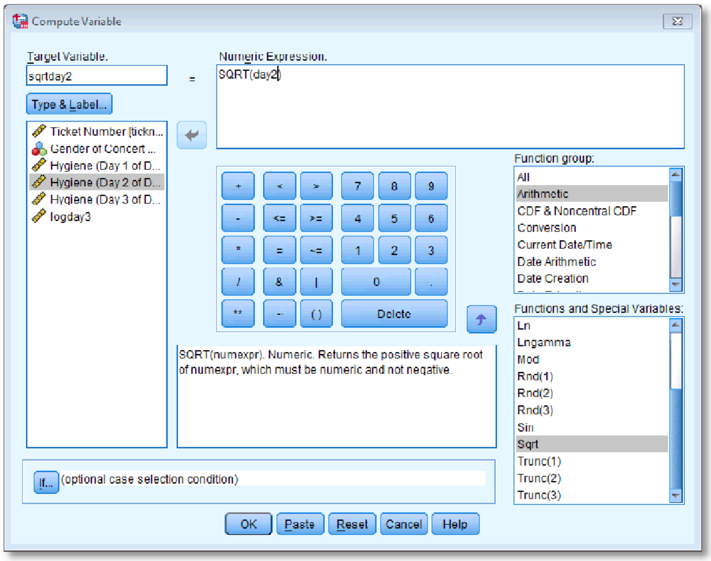
(optional case selection condition)

The histograms for days 1 and 2 are in the book, but for day 3 the histogram should look like this:



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.

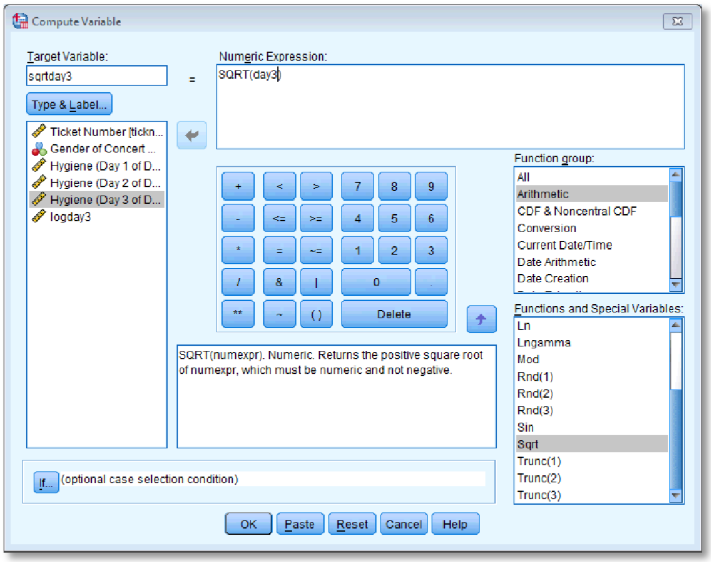
The completed *Compute Variable* dialog boxes for day 2 and day 3 should look as below:



Target Variable: sqrtday2

Numeric Expression: SQRT(day2)

SQRT(numexpr). Numeric. Returns the positive square root of numexpr, which must be numeric and not negative.

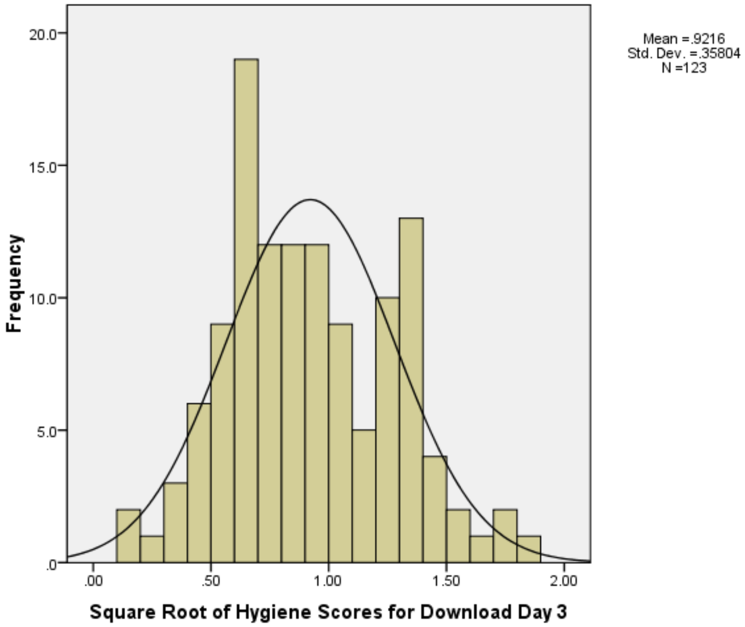


Target Variable: sqrtday3

Numeric Expression: SQRT(day3)

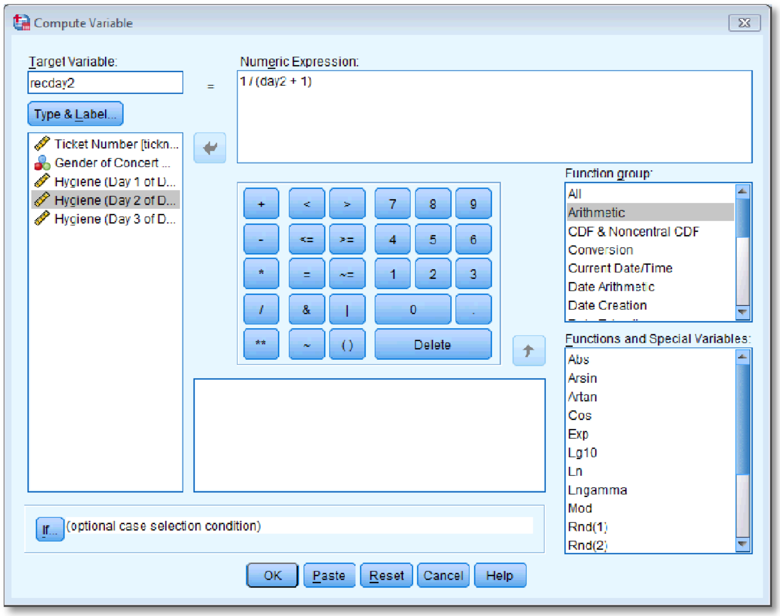
SQRT(numexpr). Numeric. Returns the positive square root of numexpr, which must be numeric and not negative.

The histograms for days 1 and 2 are in the book, but for day 3 the histogram should look like this:



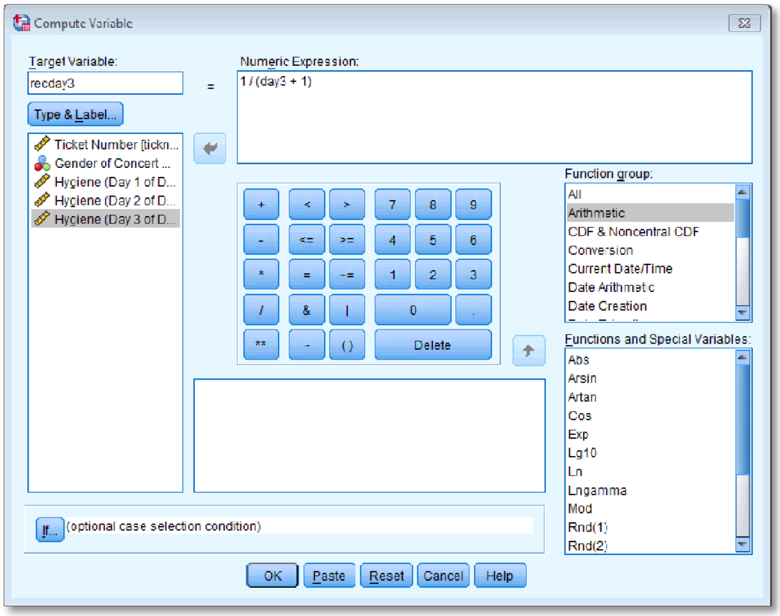
SELF-TEST Repeat this process for **day2** and **day3**. Plot histograms of the transformed scores for all three days.

The completed *Compute Variable* dialog boxes for day 2 and day 3 should look as below:



Target Variable: recday2

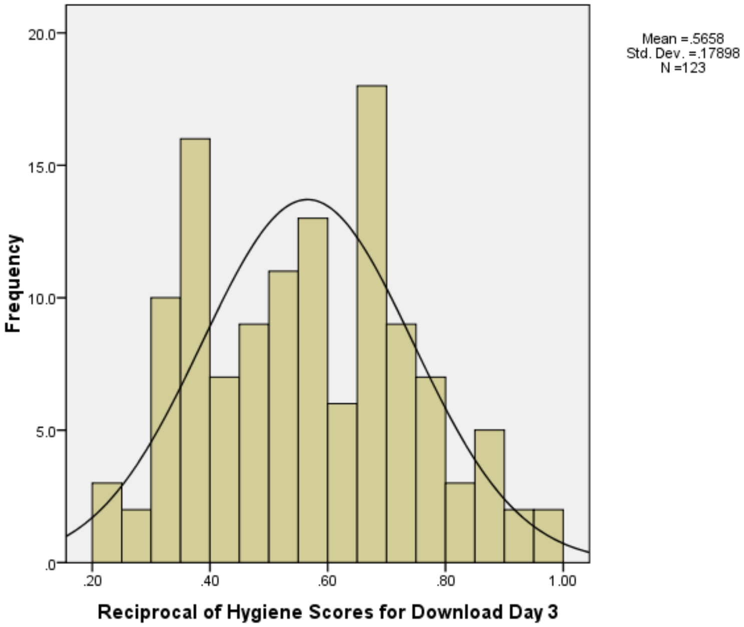
Numeric Expression: 1 / (day2 + 1)



Target Variable: recday3

Numeric Expression: 1 / (day3 + 1)

The histograms for days 1 and 2 are in the book, but for day 3 the histogram should look like this:



Mean =.5658
Std. Dev. =.17898
N =123
Frequency
Reciprocal of Hygiene Scores for Download Day 3