Inferential Statistics IV: Choosing a Hypothesis Test

LEARNING OBJECTIVES

At the end of this chapter, you should be able to:

O7.A *Interpret a quantile–quantile plot of sample data against a theoretical distribution and produce one using MATLAB*
O7.B *Compute the probability plot correlation coefficient between two distributions using MATLAB*
O7.C *Compute skew values by hand and using MATLAB, and interpret them to decide how close sample data are to a normal distribution*
O7.D *Compute z-values of a sample by hand and using MATLAB, and use them to decide how close the sample is to a normal distribution*
O7.E *Apply the Shapiro–Wilk test to sample data by hand and using MATLAB, to test if the sample fits a normal distribution*
O7.F *Apply the chi-square test to sample data by hand and using MATLAB, to test if the sample fits a normal distribution*
O7.G *Make appropriate use of tools to decide which hypothesis test to use*

7.1 INTRODUCTION

We have now covered most of the main concepts of parametric and nonparametric hypothesis testing. We know that it is important to choose an appropriate test to ensure that we gain the maximum statistical power but that we do not violate any assumptions of the test we choose. So far we have just used simple "intuitive" ways of assessing our data to decide which type of test to use. For instance, we have plotted histograms of the sample data and visually inspected their distribution. However, sometimes this approach gives unclear results, and we would like to be more precise in our assessment. In this chapter, we look at more powerful tools that we can use to help us choose the best hypothesis test. As it is the most common criterion for choosing hypothesis tests, we focus on tools that can help us to decide whether our data fit a normal distribution. In particular, we would like to demonstrate that the population variable of inter-

Statistics for Biomedical Engineers and Scientists. https://doi.org/10.1016/B978-0-08-102939-8.00016-5

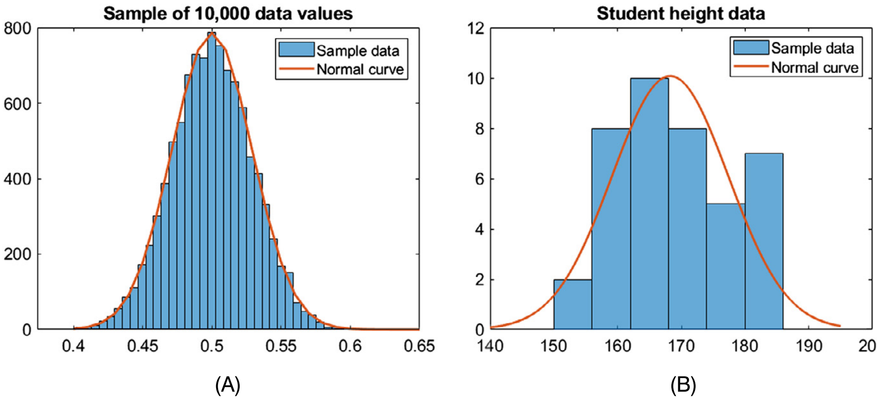


FIGURE 7.1
(A) Histogram of a sample of 10,000 data values with $\bar{x} = 0.5$ and $s = 0.029$. The curve shows the probability distribution function from a normal distribution with $\mu = 0.5$ and $\sigma = 0.029$. (B) Histogram of the student height data with sample size 40 and $\bar{x} = 168.25$, $s = 9.1083$. The curve shows a normal probability distribution function with $\mu = 168.25$ and $\sigma = 9.1083$. Note that both distribution curves have been scaled along the y-axis, so that they can be directly compared with the histograms.

est is normally distributed, as this is a key assumption of the Student's t-test. However, since we typically do not have full access to the population data, we can only assess our sample data, and we make the assumption that its distribution approximates the population distribution, which is reasonable since it is a random representative sample of the population.

7.2 VISUAL METHODS TO INVESTIGATE WHETHER A SAMPLE FITS A NORMAL DISTRIBUTION

7.2.1 Histograms

We have already seen the use of histograms to assess sample distributions several times in this book, so they are one obvious way to assess if data fit a normal distribution. Figs. 7.1A and 7.1B show histograms of sample data together with normal probability distribution functions with mean and standard deviation equal to the sample mean and standard deviation. Fig. 7.1A shows a sample with 10,000 data values. The sample distribution seems to fit well to a normal distribution. In this case, we might feel confident in applying parametric hypothesis tests to answer questions about the data. However, Fig. 7.1B shows the student height data that we first saw in Chapter 1. Would we be justified in assuming that this sample comes from a normal distribution? We will now consider further ways to help us to answer such questions.

7.2.2 Quantile–Quantile Plots

Histograms such as those shown in Fig. 7.1 are one visual way to assess similarity to a normal distribution. An alternative way is the *quantile–quantile plot*. We came across quartiles in Chapter 1 when discussing measures of variation for skewed data. We calculated the upper and lower quartiles and the median. These were defined as the data values which had:

- $\frac{1}{4}$ of the data values below it: *lower quartile*
- $\frac{2}{4}$ of the data values below it: *median*
- $\frac{3}{4}$ of the data values below it: *upper quartile*

The concept of a *quantile* is related to that of a *quartile*. A *quantile* is the generic name for a value which has $\frac{k}{q}$th of the data values below it (where q is the number of quantiles calculated, and $k = 1, \ldots, q-1$). For example, we can compute the 95% quantile, which is the data value at which $\frac{95}{100}$ of the data values lie below that value (i.e. $q = 100$ and $k = 95$). The lower quartile, median, and upper quartile correspond to the 25%, 50%, and 75% quantiles, respectively.

For example, consider Fig. 7.2. Fig. 7.2A shows a uniform distribution between -5 and $+5$ (in blue). The black vertical lines indicate the values of the quantiles for $10\%, 20\%, \ldots, 90\%$ or, in other words, the values that have $\frac{10}{100}$, $\frac{20}{100}, \ldots, \frac{90}{100}$ of the distribution below them. Fig. 7.2B shows a standard normal distribution (i.e. a normal distribution with $\mu = 0$ and $\sigma = 1$) with the same quantiles. Notice how the quantiles for the uniform distribution are evenly spaced because the height of the distribution does not change, whereas those for the normal distribution are more closely spaced near the center of the distribution where the curve is higher. *The area under the curve between successive quantiles is the same in both figures*, that is, 10% of the total area.

We can produce a quantile–quantile plot (often called a *Q–Q plot*) by calculating the quantile values from two samples, or a sample and a theoretical distribution (e.g. normal distribution), and plotting the corresponding quantile values against each other. If the two samples come from the same distribution, then the plot should approximately show a straight line along $y = x$. Fig. 7.3 shows Q–Q plots for the two distributions shown in Fig. 7.1, both against a standard normal distribution.

How should we interpret these Q–Q plots? As noted before, if the two distributions are the same, then the Q–Q plot will show a straight line through the origin at 45°. The closer we are to this case, the closer the distributions are. However, in addition to this, certain shapes in the Q–Q plot imply particular types of difference between the distributions:

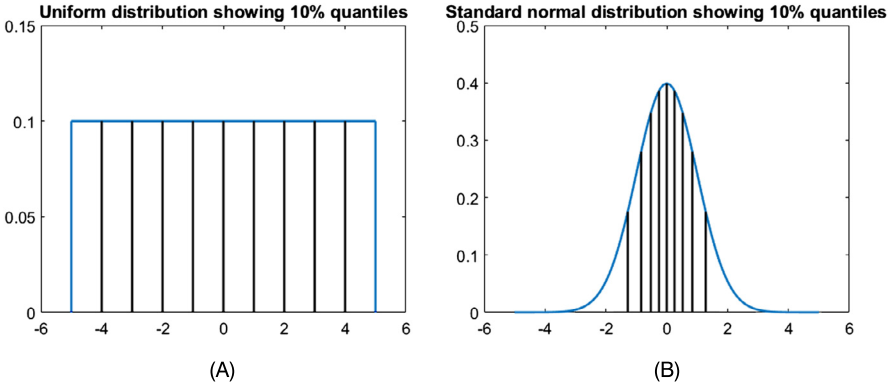


FIGURE 7.2
(A) A uniform distribution between -5 and $+5$ with quantiles shown at 10% intervals, that is, 10%, 20%, etc. (B) A standard normal distribution (i.e. $\mu = 0, \sigma = 1$) showing the same quantile values.

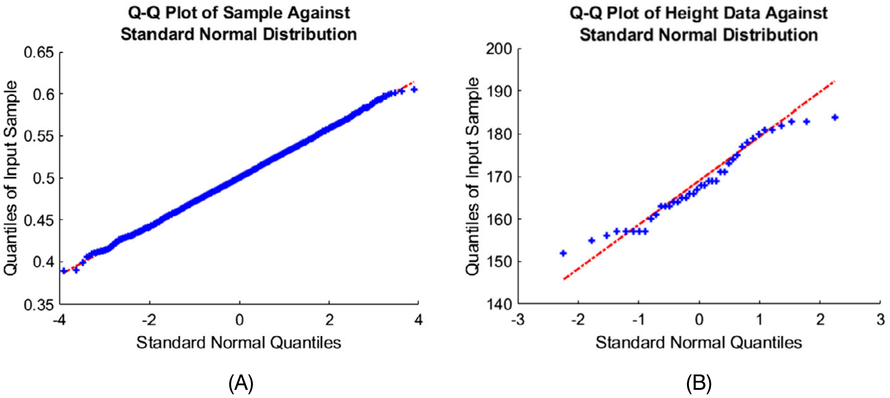


FIGURE 7.3
Q–Q plots for sample data against a standard normal distribution. (A) Sample data from Fig. 7.1A; (B) student height data from Fig. 7.1B.

- A straight line that is not at 45° and/or does not go through the origin implies that the distributions would be the same if a linear transformation was applied to one sample. If we are comparing a sample to a normal distribution, then this means that we have chosen the wrong mean and standard deviation for the normal distribution: if we chose the correct mean and standard deviation, then the Q–Q plot would show an ideal match.
- An arc shape in the Q–Q plot is the result of one distribution having a higher skew (see Section 1.5.3) than the other. This is illustrated by the example of a Q–Q plot between samples from normal and gamma distributions shown in Fig. 7.4.

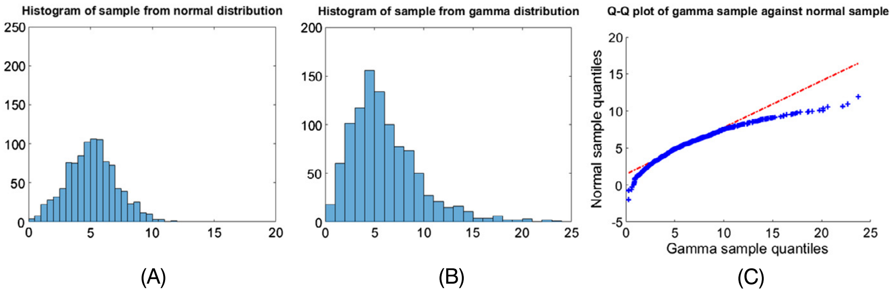


FIGURE 7.4
(A) A sample from a normal distribution with $\mu = 5$ and $\sigma = 2$; (B) a sample from a gamma distribution with shape parameter 3 and scale parameter 2; and (C) a Q–Q plot of the gamma sample against the normal sample. Notice how the higher skew of the gamma distribution causes an arc shape in the Q–Q plot.

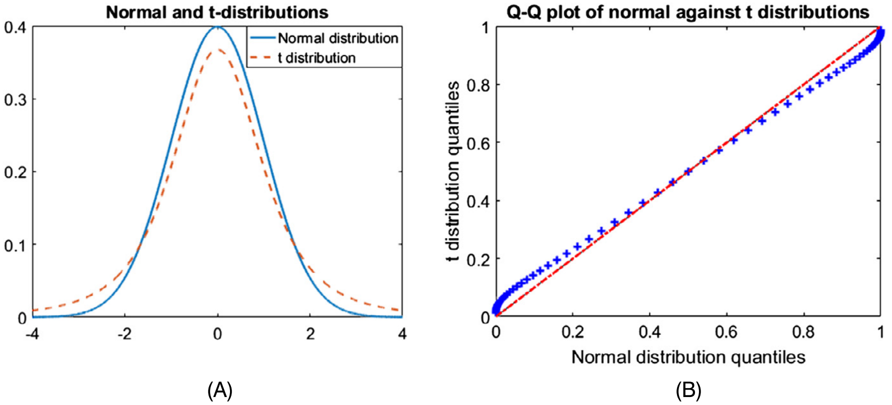


FIGURE 7.5
(A) Probability distribution functions for the normal distribution and a t-distribution with 3 degrees of freedom; (B) Q–Q plot for the same two distributions. Notice how the longer tails of the t-distribution cause an "S" shape in the Q–Q plot.

- An "S" shape in the Q–Q plot indicates that one distribution has longer tails than the other. For example, Fig. 7.5 shows a Q–Q plot of samples from a normal distribution and a t-distribution with 3 degrees of freedom (see Section 5.4). The curves at the ends of the Q–Q plot make an "S" shape, which is caused by the different quantile values at the outer ends of the distribution.

We will now introduce a case study for this chapter and use it to illustrate the Q–Q plot. We return to Professor A and her attempts to revolutionize the

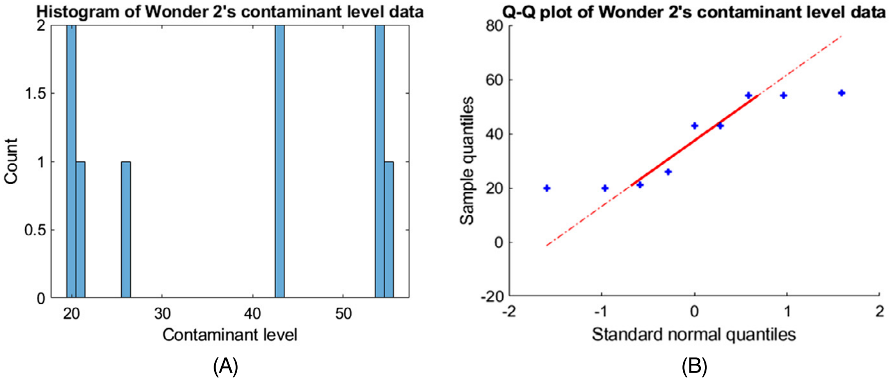


FIGURE 7.6
Analysis of Professor A's new "Wonder 2" contaminant level data: (A) histogram and (B) Q–Q plot. The Q–Q plot is against theoretical values from a standard normal distribution.

world of drug production. She has now given her researcher more freedom and asked him to come up with his own drug production technique. The result is a new drug, known as "Wonder 2". Nine batches of "Wonder 2" have been produced, and the contaminant level data are as follows: (21 54 20 55 20 54 26 43 43).

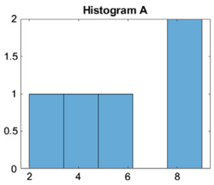
Fig. 7.6 shows a histogram and a Q–Q plot of these sample data against a standard normal distribution. As before, the Q–Q plot shows a cross for each quantile value, but note that this plot only shows 9 different quantile values. This is because there are only 9 values in the sample; therefore each data value is itself a quantile. For example, the 3rd smallest value has two values below it, 6 values above it, and 1 value equal to it (i.e. itself). So we can view this as the $\frac{2.5}{9}$th quantile, and we can find the corresponding quantile value from the standard normal distribution to plot it against. Similarly, the other data values represent the $\frac{0.5}{9}$th, $\frac{1.5}{9}$th, etc. quantiles.

Recall that any straight line in the Q–Q plot will indicate that the data come from a normal distribution (possibly with different mean and standard deviation). However, for these data, this does not seem to be the case, as is confirmed by the histogram which shows three distinct peaks.

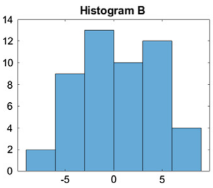
Although Q–Q plots allow us to make a more informed assessment of whether data fit a normal distribution, there is still a significant element of subjectivity in this assessment, and this will be the case for any visual method. In the next section, we will introduce methods that can *quantify* how well the data fit the distribution, and even test hypotheses about the goodness-of-fit to the distribution.

Activity 7.1

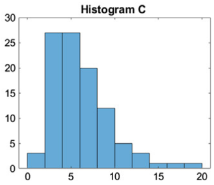
Four samples have been drawn from unknown population distributions. The images below show their histograms and Q–Q plots. Match each histogram to its corresponding Q–Q plot. *O7.A*



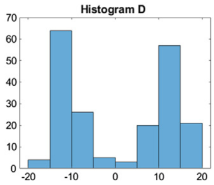
Histogram A



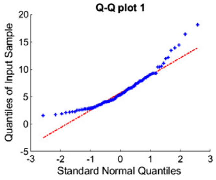
Histogram B



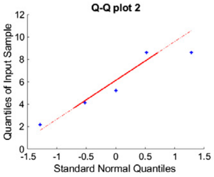
Histogram C



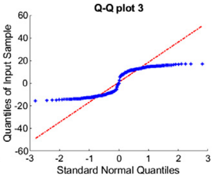
Histogram D



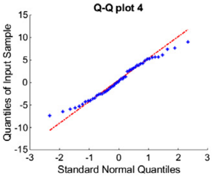
Q-Q plot 1



Q-Q plot 2



Q-Q plot 3



Q-Q plot 4

7.3 NUMERICAL METHODS TO INVESTIGATE WHETHER A SAMPLE FITS A NORMAL DISTRIBUTION

Numerical techniques have clear advantages over the visual methods that we have seen in the previous section. By quantifying (i.e. putting numbers to) how well data fit a particular distribution comparisons can be made between different samples in terms of how well they fit the distribution, and we can even think about ways of testing particular hypotheses about the fit.

In this section, we consider a number of numerical techniques that can be used to quantify some aspects of the sample distribution to help us to make more informed decisions about how close they are to a normal distribution.

7.3.1 Probability Plot Correlation Coefficient

The first numerical technique is actually derived from the Q–Q plot. The *probability plot correlation coefficient* (PPCC) is the Pearson's correlation coefficient of the data plotted in the Q–Q plot. We introduced the Pearson's correlation coefficient (or Pearson's r) in Section 2.4.1 as a measure of linear dependence between two variables. Recall that the Pearson's r value will be equal to 1 (or -1) whenever there is a linear relationship between two sets of data. For our Q–Q plot, any straight line means that our two distributions fit perfectly (or would do if an appropriate linear transform were applied to one of the data sets). Therefore, the Pearson's r value of the quantile data should be a good measure of how well the distributions match.

Let us return to our Professor A case study to illustrate this. For the sample data, the quantiles are the data values themselves: (20 20 21 26 43 43 54 54 55).

How can we compute the corresponding quantiles for the standard normal distribution? What we want to know is what values of the standard normal distribution have the same proportions of values below them as the quantile values for our sample. As explained in the previous section, for a sample of size n, the proportions are $\frac{0.5}{n}, \frac{1.5}{n}, \ldots, \frac{n-0.5}{n}$. Based on these proportions and the equation for a normal distribution, we can determine that the standard normal distribution quantiles are (−1.593 −0.967 −0.59 −0.282 0 0.282 0.59 0.967 1.593).

Calculating the Pearson's r value between these two sets of quantile values results in a PPCC of 0.9205. But what does this mean? What is an acceptable value for PPCC? In truth, there is no clear answer to this question, but a very high value (e.g. 0.99) would suggest that the sample distribution and the test distribution *were* similar. On the other hand, a much lower value (e.g. 0.75) would definitely suggest that they were *not*. Our value of 0.9205 is probably inconclusive, so it would be useful to perform some further analysis before making a final decision about the distribution of our sample.

7.3.2 Skew Values

Computing the *skew* of a sample is a quick test that we can use to test how *symmetric* its distribution is. We already saw one way to calculate skew in Section 1.5.3, which we reproduce here for convenience:

$$\text{skew} = \frac{3(\bar{x} - \text{ median})}{s}.$$

If the skew is not between −1 and 1, then the distribution is not very symmetric and so may well not be normal.

Let us compute the skew of Professor A's latest contaminant level data. We have $\bar{x} = 37.33$, median $= 43$, $s = 15.52$, and therefore skew $= \frac{3(37.33-43)}{15.52} = -1.0951$. From this it seems as if our distribution has a significant negative skew, so it is still not looking good for it to fit to a normal distribution.

7.3.3 z-values

A further numerical test can be carried out by converting our sample data into *z-values*, as defined by the equation

$$z_i = \frac{x_i - \bar{x}}{s}. \tag{7.1}$$

This equation transforms our sample so that it has a mean of 0 and a standard deviation of 1. Therefore the z-values indicate how many standard deviations each value is away from the mean of the sample.[1] If our data are actually from a normal distribution, then from the properties of this distribution (see Fig. 4.5B) we know the percentages of values that should be less than 1, 2, 3, and 4 standard deviations from the mean, which are 68.3%, 95.4%, 99.7%, and 99.994%, respectively.

A quick check to see if our data come from a normal distribution is to look at the maximum magnitude of the z-values (i.e. ignoring their signs). It is very unlikely that we would obtain many z-values with magnitude 3 or more from a normal distribution unless we have a very large sample (we would only expect 3 data points with a z-value magnitude above 3 in a sample of 1,000, and we would need samples of 10,000 before we would expect to obtain a z-value magnitude above 4). Therefore large z-scores from small samples are an indication that our data may not be from a normal distribution.

Let us return to our case study. Our original data were

$$(21\ 54\ 20\ 55\ 20\ 54\ 26\ 43\ 43)$$

We transform these data to z-values using Eq. (7.1), resulting in

$$(-1.052\ 1.074\ -1.117\ 1.138\ -1.117\ 1.074\ -0.73\ 0.365\ 0.365)$$

We can see straight away that we do not have any very large z-values. This seems promising at first. However, for our sample size of 9, we would expect to see $(1 - 0.68) \times 9 = 2.88$ values beyond 1 standard deviation. We have 6 values that are more than 1 standard deviation from the mean. This suggests that the distribution from which our sample was taken may have longer tails than a normal distribution.

■ Activity 7.2

A company has developed a new surgical technique for inserting hip implants. The angular error of the placement can be assessed postoperatively using x-ray imaging. Error data (in degrees) have been gathered from 9 patients who underwent the new surgery. The data are shown in the table below.

O7.C, O7.D

Angular error (degrees)								Mean	Std. dev.	Median
8	3	2	2	4	1	1	2	2.88	2.3	2

[1] Note the similarity of Eq. (7.1) with Eq. (5.5) in Section 5.9. Whereas the statistic in the z-test computes the number of *standard errors* our mean is from the expected mean, z-values compute how many *standard deviations* each sample value is from the sample mean.

1. Compute a skew value for the alignment error data. Comment on the result.
2. Compute z-values for the same data. Again, comment on the results with regard to whether the data are likely to be normally distributed.

■

7.3.4 Shapiro–Wilk Test

All of the above techniques can provide useful information to help us to decide whether or not sample data were likely to have come from a normal distribution. However, sometimes the answer is still unclear, and it would be useful to have a way to make a formal decision as to whether this is likely to be the case or not. There are a number of more sophisticated methods to help us to answer such questions. A commonly used one for small samples is the *Shapiro–Wilk test*.

The Shapiro–Wilk test is a hypothesis test, and so we will follow the checklist for hypothesis testing that we first introduced in Section 5.2.

- *Form null and alternative hypotheses and choose a degree of confidence*: For the Shapiro–Wilk test, the null hypothesis is that the sample comes from a normal distribution, and the alternative hypothesis is that it does not. We can choose any degree of confidence, but common choices are 95% and 99%.
- *Compute a test statistic*: We do this by first ranking the sample values in increasing order – we denote these ranked data by $(x_{(1)}\ x_{(2)}\ \ldots\ x_{(n-1)}\ x_{(n)})$, where n is the sample size. Next, we calculate $b = a_1(x_{(n)} - x_{(1)}) + a_2(x_{(n-1)} - x_{(2)})\ldots$, where $a_1, a_2, \ldots$ are the coefficients from Table A.6 (see Appendix). Finally, the test statistic is computed as Calc $W = \frac{b^2}{(n-1)s^2}$, where s is the sample standard deviation.
- *Compare the test statistic with a critical value*: The critical W values for a Shapiro–Wilk test are shown in Table A.7 in the Appendix. We denote the critical value by Tab W.
- If Calc $W >$ Tab W, then we do *not* reject the null hypothesis.

We will illustrate this process using the contaminant level data for Professor A's new drug "Wonder 2". We will work to a 95% degree of confidence, or 0.05 significance. The test statistic is calculated by following the steps outlined above:

- Rank the data: (20 20 21 26 43 43 54 54 55).
- Using coefficients from Table A.6 for $n = 9$, we calculate $b = 0.5888(55 - 20) + 0.3244(54 - 20) + 0.1976(54 - 21) + 0.0947(43 - 26) = 39.7683$.
- Calculate the test statistic: Calc $W = \frac{39.7683^2}{(9-1)15.52^2} = 0.8203$.

- Look up the critical W value in Table A.7: for $n = 9$, 0.05 significance, Tab $W = 0.829$.
- Because Calc W < Tab W (i.e. 0.8203 < 0.829), we *reject the null hypothesis*, that is, we conclude with 95% confidence that our data are *not* from a normal distribution.

This conclusion confirms our earlier suspicions based on the Q–Q plot, PPCC, skew, and z-values.

The Intuition. Shapiro–Wilk Test

The Shapiro–Wilk test is essentially a *goodness-of-fit* test. That is, it examines how close the sample data fit to a normal distribution. It does this by ordering and *standardizing* the sample (*standardizing* refers to converting the data to a distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$). If the sample data perfectly fit a normal distribution, then after this ordering and standardization process the sample values would be regularly spaced quantile values of the standard normal distribution (see Section 7.2.2, in particular, the explanation of how to compute quantiles for Professor A's contaminant level data). The Shapiro–Wilk test statistic (Calc W) is basically a measure of how well the ordered and standardized sample quantiles fit the standard normal quantiles. The statistic will take a value between 0 and 1 with 1 being a perfect match. This is why a small value of Calc W will result in rejection of the null hypothesis of normality. The equation for Calc W given above, which is based on the coefficients in Table A.6, basically performs the standardization and calculation of the goodness-of-fit. Therefore, there is a close link between the Shapiro–Wilk test and Q–Q plots, which offer a visual assessment of this goodness-of-fit using quantiles. ■

Note that the result of the Shapiro–Wilk test should not be taken as being 100% reliable (no hypothesis test should). There will be cases where the test will give a misleading result. Therefore the Shapiro–Wilk test should never be used on its own – it should always be interpreted along with the results of graphical and numerical tools. In many cases, it will be obvious from the graphical and numerical tools that the data are (or are not) normally distributed, so the Shapiro–Wilk test will be unnecessary. It can be useful, however, when the results of the graphical and numerical tools are inconclusive.

Activity 7.3

In Activity 7.2, we introduced the error data for a new surgical technique for inserting hip implants. For convenience, the data are shown again in the following table. O7.E

Angular error (degrees)								Mean	Std. dev.	Median
8	3	2	2	4	1	1	2	2.88	2.3	2

Perform a Shapiro–Wilk test to determine, with 95% confidence, whether the data fit a normal distribution. Clearly state your hypotheses and show all working. ■

7.3.5 Chi-Square Test for Normality

The Shapiro–Wilk test was designed for use with small sample sizes. Although there is no hard-and-fast rule, a rule of thumb is that it is suitable when dealing with sample sizes of 50 or less. One alternative to the Shapiro–Wilk test that may be more powerful for larger sample sizes is the χ^2 test. In Section 6.5, we saw the use of the χ^2 test for testing hypotheses about categorical data. A slightly different form of the test can be used to test the goodness-of-fit of a sample against any expected distribution. We will demonstrate its use for testing against a normal distribution.

To illustrate the use of the χ^2 test, we introduce a new case study. A team of biomedical engineers has developed a technique for automatically estimating gestational age from a magnetic resonance (MR) scan of a fetus. They have tested their technique on 300 fetal MR scans for which the "gold standard" gestational age was known. Based on these data, they have computed the errors in gestational age estimation for their technique. These errors are summarized in Table 7.1. To perform further statistical analysis on the error figures, the team would like to know if their data are normally distributed or not. We will work to a 95% degree of confidence.

The null hypothesis for the χ^2 goodness-of-fit test is that there is no significant difference between the sample data and the expected distribution (in this case, a normal distribution). The alternative hypothesis is that there is a difference. To decide whether or not we can reject the null hypothesis, we need to compute the χ^2 test statistic Calc χ^2. In a similar way to the χ^2 tests that we saw in Section 6.5, the χ^2 goodness-of-fit test computes the test statistic by comparing observed frequencies with expected frequencies. This time, we compare observed frequencies of sample values within particular ranges (or *bins*) with those that would be expected if the sample were from a normal distribution with the same mean and standard deviation as the sample. This comparison is summarized in Table 7.2. The first column shows the bins (or ranges) of errors used (i.e. the same as in Table 7.1). The second column shows the probabilities of sample values from these bins (assuming that the sample was normally distributed). These probabilities can be computed from the areas under a normal distribution with the same mean and standard deviation as the sample (e.g. see Fig. 4.5B). Based on these probabilities and the sample size, we can compute expected frequencies E for each bin. These values are shown in the third column of the table. For example, the value for the < -10 bin is 3.39,

Table 7.1 Errors in MR-based gestational age estimation for 300 fetuses.

Error in gestational age estimation (days)	Number of cases
Less than -10	2
Between -10 and -5	36
Between -5 and 0	123
Between 0 and 5	97
Between 5 and 10	38
More than 10	4

Table 7.2 Computation of Calc χ^2 for the gestational age error data.

Gest. age error	Prob	E		O		$\frac{(O-E)^2}{E}$
< -10	0.0113	3.39	} 38.09	2	} 38	0.0002
≥ -10 and < -5	0.1157	34.7		36		
≥ -5 and < 0	0.3723	111.7		123		1.14
≥ 0 and < 5	0.373	111.9		97		2.29
≥ 5 and < 10	0.1163	34.89	} 38.31	38	} 42	0.36
≥ 10	0.0114	3.42		4		
Totals:	1.0	300.0		300		*Calc* $\chi^2 = 3.79$

which is equal to the probability 0.0113 multiplied by the sample size 300. The fourth column shows the observed frequencies O, which are reproduced from Table 7.1. Finally, the fifth column shows the χ^2 statistic for each row. This is computed as the square of the difference between the observed and expected frequencies divided by the expected frequency. The final test statistic, Calc χ^2, is the sum of all of these χ^2 statistics. Note that this is the same formula as we used for the χ^2 tests that we saw in Section 6.5, that is, Eq. (6.4), which is reproduced here for convenience:

$$\chi^2 = \sum \frac{(O-E)^2}{E}.$$

Note that the frequencies for the first two bins (<-10, ≥ -10 and <-5) have been combined. We should always do this when the frequency of any bin is less than or equal to 5. In our case, the first bin (<-10) has both observed and expected frequencies that are less than or equal to 5. Therefore, we have to combine the first two bins to ensure that both expected and observed frequencies are greater than 5. We perform the same combination for the last two bins for the same reason.

After we have calculated our test statistic, we simply compare it to a critical value from Table A.5. To look up the critical value, we must know the number of degrees of freedom of the test. For a χ^2 goodness-of-fit test, the number of degrees of freedom is the number of bins minus 3. We subtract 3 because we already know that the sums of the expected and observed frequencies are the same, and we also know the mean and standard deviation of the distribution. For our example, we have 4 bins (after the first two and the last two have been combined). Therefore, we have $4 - 3 = 1$ degree of freedom. From Table A.5 we see that our critical value Tab χ^2 for a 0.05 significance level (i.e. 95% confidence) is equal to 3.841. Because Calc $\chi^2 = 3.79$ is not bigger than Tab $\chi^2 = 3.841$, we do not reject the null hypothesis that the data are from a normal distribution.

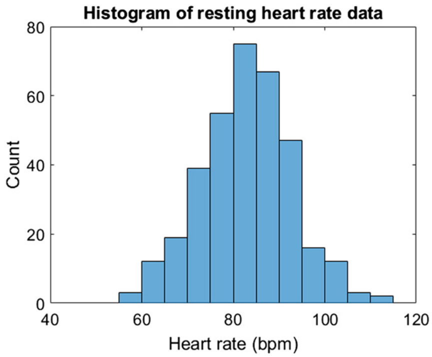
■ The Intuition. The Chi-Square Goodness-of-Fit Test

The intuition behind the χ^2 goodness-of-fit test is similar to that described for the χ^2 tests that we saw in Section 6.5. The χ^2 test statistic has a known distribution as shown in Fig. 6.4. The null hypothesis of normality is rejected when the statistic is large enough to go beyond the critical χ^2 value Tab χ^2 for the given significance level, that is, it becomes unlikely that we would get a value this large or larger by chance. ■

■ Activity 7.4

O7.F Resting heart rate data (in beats per minute, bpm) have been gathered from a cohort of 350 volunteers. The data are summarized in the table and histogram below.

Heart rate	Probability	Expected frequency, E	Observed frequency, O	$\frac{(O-E)^2}{E}$
<60	0.0096		3	
60–65	0.0238		12	
65–70	0.0594		19	
70–75	0.115		39	
75–80	0.1726		55	
80–85	0.2009		75	
85–90	0.1814		67	
90–95	0.127		47	
95–100	0.0689		16	
100–105	0.029		12	
105–110	0.0095		3	
>110	0.0029		2	
Totals:	1.0		350	Calc $\chi^2 =$



The mean heart rate is 82.9 bpm, and the standard deviation is 9.8 bpm. The table also shows the probabilities of the different heart rate bins based on a normal distribution with the same mean and standard deviation. The expected frequency and χ^2 test statistic columns have been left blank for you to fill in.

Use the values in the table to perform the χ^2 test by hand to test if the heart rate data come from a normal distribution. ■

7.4 SHOULD WE USE A PARAMETRIC OR NONPARAMETRIC TEST?

We have now seen a number of different statistical hypothesis tests, both parametric and nonparametric. When making use of tests such as these, it is important to choose an appropriate test. This enables us to be sure that our conclusions are legitimate and also that we have used the maximum statistical power at our disposal. It can sometimes be difficult to choose between parametric and nonparametric tests, and in fact, statisticians often disagree about when certain tests can be applied and how to get the maximum power.

However, we can outline a few basic considerations that are worth remembering when choosing which test to apply. First, we should definitely choose a nonparametric test if we have ranked data (e.g. a grade such as A, B, C, etc.). We should also choose a nonparametric test if we know that our data were not drawn from a normal distribution. The following methods can be used to help us to decide if this is the case or not:

- Examine the data, for example, by using histograms or Q–Q plots.
- Use numerical measures, such as PPCC, skew, or z-values.
- Use a formal test of normality, such as the Shapiro–Wilk test or the χ^2 goodness-of-fit test.

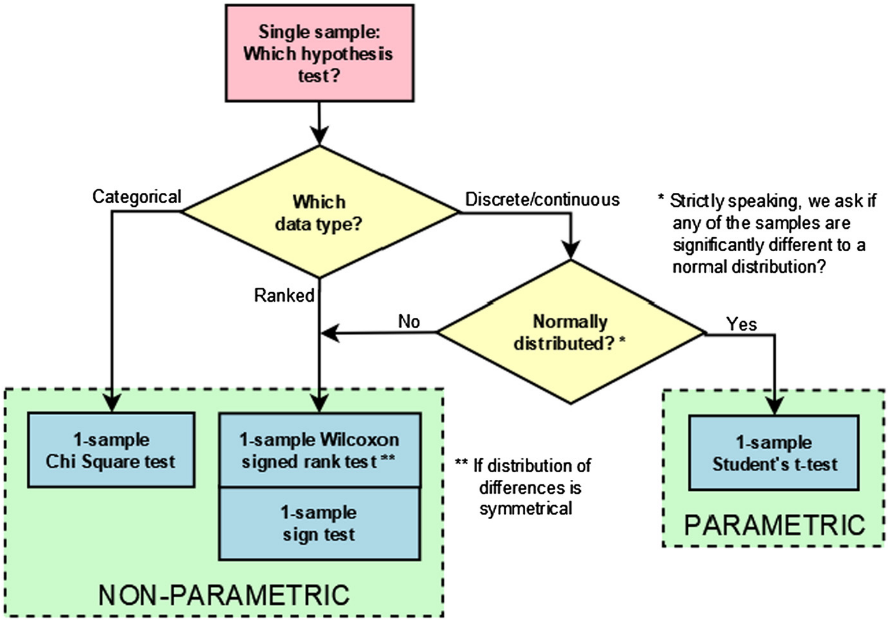

Single sample: Which hypothesis test?
Which data type?
Categorical
Discrete/continuous
Ranked
* Strictly speaking, we ask if any of the samples are significantly different to a normal distribution?
Normally distributed? *
No
Yes
1-sample Chi Square test
1-sample Wilcoxon signed rank test **
1-sample sign test
** If distribution of differences is symmetrical
1-sample Student's t-test
PARAMETRIC
NON-PARAMETRIC

FIGURE 7.7
Summary of considerations when choosing an appropriate one sample hypothesis test.

- Use information from other studies from the same population. Were these shown to be normally distributed?
- Think about the causes of variation in our data: if the data are likely to be the result of lots of random factors then the resulting distribution is likely to be normal.

Figs. 7.7 and 7.8 present simplified flow diagrams summarizing some of the factors that should be considered when choosing a hypothesis test. Note that we only include in these diagrams the tests that we have covered in this book. In reality, there are many more tests to choose from, each with their own assumptions and strengths/weaknesses. But we hope that these diagrams will enable the interested reader to get started on analyzing and answering questions about their data and will act as a springboard from which to explore other types of test that may be suitable for their needs.

7.5 DOES IT MATTER IF WE USE THE WRONG TEST?

We have mentioned several times that parametric tests require the population variable(s) from which the sample data were drawn to be normally distributed. This is a widely applied and very useful rule. However, in hypothesis testing, we are typically interested in assessing a measure of central tendency, such as the mean. We saw in Chapter 4 that the central limit theorem states that the

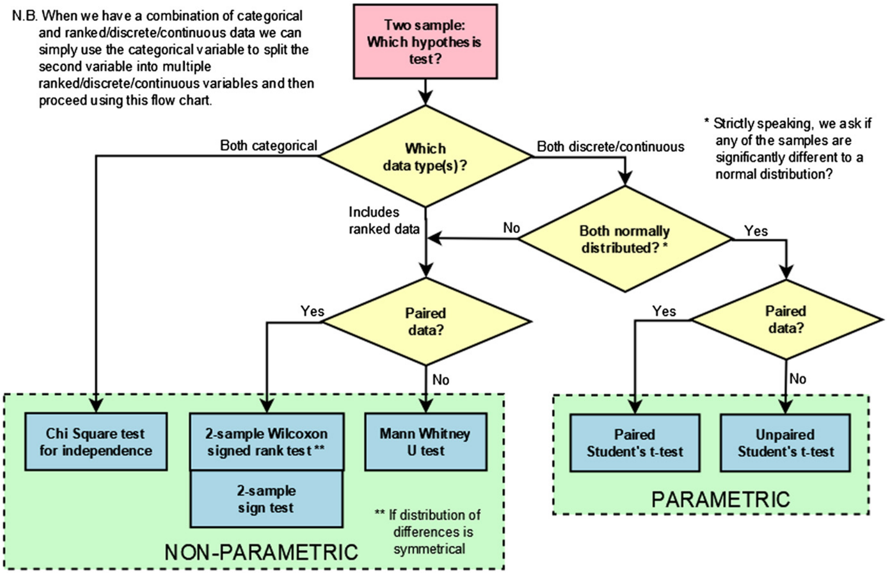

N.B. When we have a combination of categorical and ranked/discrete/continuous data we can simply use the categorical variable to split the second variable into multiple ranked/discrete/continuous variables and then proceed using this flow chart.
Two sample: Which hypothesis test?
Which data type(s)?
Both categorical
Both discrete/continuous
Includes ranked data
* Strictly speaking, we ask if any of the samples are significantly different to a normal distribution?
Both normally distributed? *
No
Yes
Paired data?
Yes
No
Paired data?
Yes
No
Chi Square test for independence
2-sample Wilcoxon signed rank test **
2-sample sign test
Mann Whitney U test
** If distribution of differences is symmetrical
NON-PARAMETRIC
Paired Student's t-test
Unpaired Student's t-test
PARAMETRIC

FIGURE 7.8

Summary of considerations when choosing an appropriate two sample hypothesis test.

mean of a sample will *always* be approximately normally distributed so long as the sample size is large enough. So a legitimate question to ask is: if we are interested answering questions about the mean value, why not always use a parametric test? The answer to this question is that the assumption of normality is required to derive the formulation of the t-test. In reality the t-test does tend to be quite robust to nonnormally distributed data, so it may not matter that much if we apply a t-test with data whose distribution is only slightly different from a normal distribution. However, we should also bear in mind which measure of central tendency we are interested in – for highly skewed data, the mean is not an appropriate measure, and so the t-test should not be used.

Overall, it is certainly safest to only use a parametric test if we are sure that the variable(s) of interest are normally distributed. In fact, for large sample sizes, parametric and nonparametric tests will give similar results. However, for smaller sample sizes, the use of a parametric test on data from a clearly nonnormal population may result in an inaccurate result. The use of a nonparametric test on data from a clearly normal population will reduce the power of our statistics.

7.6 SUMMARY

To be able to apply an appropriate hypothesis test, it is important to be able to test assumptions about the distribution that the sample data were drawn from. One of the most common assumptions is that a sample was drawn from a normal distribution. A variety of visual and numerical techniques can be utilized to help us to decide if this assumption is valid.

Visual methods include plotting histograms or Q–Q plots. Q–Q plots illustrate the relationship between corresponding quantiles between two samples or between a sample and a theoretical distribution. A quantile is a value below which a certain proportion of the data values will lie. If the relationship between a sample's quantile values and those of a normal distribution is linear, then the sample is likely to have been drawn from a normal distribution.

Numerical methods include the probability plot correlation coefficient (PPCC), which is the correlation coefficient of the quantile values. A PPCC value of 1 indicates a linear relationship, and hence that the sample came from the theoretical distribution. Skew values can also be calculated, with a magnitude greater than 1 indicating a significant skew. z-values involve transforming the data to a standard normal distribution and enable us to assess the proportions of data values that lie beyond certain multiples of the standard deviation from the mean. These proportions can be compared against known values for a normal distribution.

The Shapiro–Wilk test and the chi-square (χ^2) test are formal tests, which can be used to test whether samples were likely to have been drawn from a normal distribution. The Shapiro–Wilk test is suitable for small samples, typically $\leq$50, whereas the χ^2 test is suitable for larger samples.

7.7 ASSESSING DATA DISTRIBUTIONS USING MATLAB

In this section, we detail the MATLAB functions that are provided for assessing the distribution of data samples. Where no built-in function is available, we provide basic implementations ourselves.

7.7.1 Visual Methods

Histograms:

```
histogram(x,nbins)
```

Produces a histogram from the array `x`. The optional parameter `nbins` specifies the number of bins to use in the histogram.

Q–Q Plot:

```
qqplot(x)
```

Displays a quantile–quantile plot of sample data `x` against a standard normal distribution.

7.7.2 Numerical Methods

Probability Plot Correlation Coefficient:

There is no built-in MATLAB function to compute the PPCC. The implementation given below will compute the PPCC of an array `data` against a standard normal distribution.

ppcc.m:

```
function coef = ppcc(data)
% compute probability plot correlation coefficient (PPCC)
% usage:
%   coef = ppcc(data)
%     data: sample data
%     coef: PPCC value

% make sure it's a column vector
data = data(:);

n=length(data);  % sample size
data=sort(data); % sort data

% compute PPCC
for i=1:n % loop over all the data values
    %calculate the fraction of data we need below
    % each quantile value
    frac = (i-0.5)/n;
    % calculate the quantile values from a standard
    % normal distribution, and save in an array
    normq(i) = norminv(frac);
end

coef = corr(data,normq'); % Pearson's r
% note because input is a column vector, while
% normq is a row vector we need to transpose
% one using '

end
```

Skew:

As noted in Chapter 1, the skew of a sample array x can be calculated in MATLAB using the following calculation:

```
3*(mean(x) - median(x))/std(x)
```

***z*-values**:

```
z = zscore(x)
```

Compute the *z*-values of the sample data x.

Shapiro–Wilk Test:

There is no built-in function in MATLAB to perform the Shapiro–Wilk test. The basic implementation that we have provided below will perform the test at 0.05 significance for a maximum sample size of 10. See Section 7.8 for details of how to download a more flexible implementation.

shapiro_wilk_test.m:

```
function h = shapiro_wilk_test(d)
% perform Shapiro-Wilk test at 0.05 significance
% usage:
%   h = shapiro_wilk_test(d)
%     d: sample data
%     h: result: zero means do not reject null hypothesis
%                that data are normal

% make sure it's a column vector
d = d(:);

% check sample size is valid
if length(d) > 10
   error('shapiro_wilk_test:  sample size must be <= 10');
end

% Shapiro-Wilk coefficients a for n<=10
swcoeff=[0.0000 0.7071 0.7071 0.6872 0.6646 ...
          0.6431 0.6233 0.6052 0.5888 0.5739; ...
         0.0000 0.0000 0.0000 0.1677 0.2413 ...
          0.2806 0.3031 0.3164 0.3244 0.3291; ...
         0.0000 0.0000 0.0000 0.0000 0.0000 ...
          0.0875 0.1401 0.1743 0.1976 0.2141; ...
         0.0000 0.0000 0.0000 0.0000 0.0000 ...
          0.0000 0.0000 0.0561 0.0947 0.1224; ...
         0.0000 0.0000 0.0000 0.0000 0.0000 ...
          0.0000 0.0000 0.0000 0.0000 0.0399];

% Shapiro-Wilk tabulated values for alpha=0.05 and n<=10
swtab=[0 0 0.767 0.748 0.762 0.788 0.803 0.818 0.829 0.842];

% sort data
ds = sort(d);

% compute b
n = length(ds);
s = std(ds);
b = 0;
for i=1:n/2
    b = b + (ds(n-i+1)-ds(i))*swcoeff(i,n);
end

% compute test statistic w
calcw = (b*b) / ((n-1)*s*s);
```

```
% find tabulated value w
tabw = swtab(n);

% compare calculated and tabulated w
% null hypothesis: data are from normal distribution
% alternative hypothesis: it's not
if (calcw > tabw)
    h = 0; % do not reject null hypothesis
else
    h = 1; % reject null hypthesis — data are not normal
end

end
```

Chi-Square Test for Normality:

```
[h, p] = chi2gof(x)
```

Performs a χ^2 goodness-of-fit test at 95% confidence for data sample `x` against a normal distribution. A return value of `h` $= 1$ means that the null hypothesis that the sample is normally distributed is rejected, that is, the sample is not normally distributed. A return value of `h` $= 0$ means that the null hypothesis cannot be rejected. The return value `p` is the p-value of the hypothesis test.

7.8 FURTHER RESOURCES

- The MATLAB Statistics and Machine Learning Toolbox documentation features a full list of functions available for performing other types of hypothesis test: https://mathworks.com/help/stats/hypothesis-tests-1.html
- A MATLAB script for performing the Shapiro–Wilk test is available for download from the Mathworks File Exchange: https://mathworks.com/matlabcentral/fileexchange/13964-shapiro-wilk-and-shapiro-francia-normality-tests

7.9 EXERCISES

Perform the following tasks, either by hand or using MATLAB, as specified.

Exercise 7.1

The contaminant level data for Professor A's new "Wonder 2" drug are contained in the file "profa_wonder2.txt", which is available from the book's web site. Use MATLAB to produce a Q–Q plot of these data against a standard normal distribution. ■ *O7.A*

■ Exercise 7.2

O7.B Use MATLAB to verify the PPCC figure of 0.9205 for Professor A's purity data given in Section 7.3.1. ■

■ Exercise 7.3

O7.A The MATLAB `qqplot` function can be used to produce a Q–Q plot of one sample against another or (if only one argument is provided) of one sample against a standard normal distribution. Write a new function that produces a Q–Q plot of a sample against a standard t-distribution with specified number of degrees of freedom. The function should take two arguments: the data sample (an array) and an integer representing the degrees of freedom. It does not need to return any value. The basic steps that your code should implement are:

- Generate a (large) random sample from a standard t-distribution (i.e. with $\mu = 0$ and $\sigma = 1$) with the specified number of degrees of freedom.
- Produce a Q–Q plot of the t-distribution sample against the sample provided as an argument.
- Annotate the plot appropriately.

Once you have written your function, use it to produce a Q–Q plot against an appropriate t-distribution for Professor A's "Wonder 2" drug contaminant level data. ■

■ Exercise 7.4

O7.C Use MATLAB to compute the skew of the "Wonder 2" drug contaminant level data. ■

■ Exercise 7.5

O7.D Use MATLAB to verify the computation of the z-values of the "Wonder 2" drug contaminant level data. You can check your results against the values given in Section 7.3.3. ■

■ Exercise 7.6

O7.E As part of a study on Alzheimer's disease patients, 8 patients were given a cognitive impairment test. The scores out of 20 were (high meaning impairment): (18.44 14.18 19.79 15.73 15.36 16.17 13.91 15.35). Perform a Shapiro–Wilk test to determine if the sample comes from a normal distribution. You can use MATLAB to do the calculations if you want, but you should follow the steps to apply the test by hand as outlined in Section 7.3.4. ■

■ Exercise 7.7

In Activity 7.4, you performed a χ^2 goodness-of-fit test on volunteer heart rate data. The heart rate data are available to you through the book's web site as the file "hr_data.mat". Use MATLAB to perform the same χ^2 goodness-of-fit test to verify the result that you got in Activity 7.4. ■ *O7.F*

■ Exercise 7.8

Image *segmentation* refers to the process of delineating the boundaries of regions of interest in an image. In medical imaging, segmentation can be useful for delineating organs or tumors to derive clinical measurements and assess disease. Segmentation can be performed manually using an interactive software tool. Manual segmentations are normally very accurate but can be very time-consuming to produce. Therefore, there is significant interest in developing automatic segmentation algorithms that can delineate regions of interest with no human interaction at all. *O7.G*

A research team has developed a new technique for automatic liver segmentation from magnetic resonance (MR) images and wishes to compare the new approach to a comparative technique, which is considered to be the current state-of-the-art.

One way of assessing the similarity between segmentations is by computing a *Dice coefficient*. A Dice coefficient measures the degree of overlap between two segmentations and is a number between 0 and 1, with 0 representing no overlap and 1 representing full overlap.

To assess the accuracy of the liver segmentations produced by the new and comparative techniques, a junior doctor has manually segmented the livers from 35 MR scans. These manual segmentations are considered to be the "gold standard" segmentations. Dice coefficients have been computed between the segmentations produced by the new/comparative techniques and the gold standard.

These data are available to you in the file "segmentation.mat" from the book's web site, which contains the following variables:

- `dice_new`: The Dice coefficients between the segmentations produced by the new technique and the gold standard segmentations.
- `dice_comparative`: The Dice coefficients between the segmentations produced by the comparative technique and the gold standard segmentations.

Based on these data, use MATLAB to determine if the new segmentation technique is more accurate than the comparative technique. Be as thorough as possible in your data analysis and make sure that you choose and apply an appropriate hypothesis test to answer the question. Use a 95% degree of confidence. ■

FAMOUS STATISTICIAN: CARL FRIEDRICH GAUSS

This chapter's Famous Statistician is more of a general mathematician/scientist, but he made undoubted contributions that have greatly influenced the field of statistics. Carl Friedrich Gauss is sometimes referred to as the "Prince of Mathematicians". He had a remarkable influence in many fields of mathematics and science and is ranked as one of history's most influential mathematicians.

Gauss was born in 1777 to poor parents in Braunschweig, now part of Lower Saxony, Germany. He was a child prodigy and completed a major (and still influential today) work on number theory at the age of 21. Among his many achievements is his work on the normal (or Gaussian) distribution. This has proved very influential on the field of statistics and is central to the problems of estimating confidence intervals and testing hypotheses. Gauss was notoriously committed to his work, and one story goes that he was once interrupted in the middle of a problem and told that his wife was dying. He is reported to have said "Tell her to wait a moment till I'm done." Gauss himself died in Germany in 1855, aged 77.