Unit 10

Nonparametric and goodness-of-fit tests

Introduction

This unit concerns two different topics, which do not have a great deal in common except that they both involve testing hypotheses.

Two tests that were discussed in Unit 9, the t-test and the test of the value of a proportion, concerned:

- null and alternative hypotheses about the value of a population parameter (the population mean μ and population proportion p, respectively)
- a particular assumption about the underlying distribution of the data involved (that they follow a normal distribution or a binomial distribution, respectively).

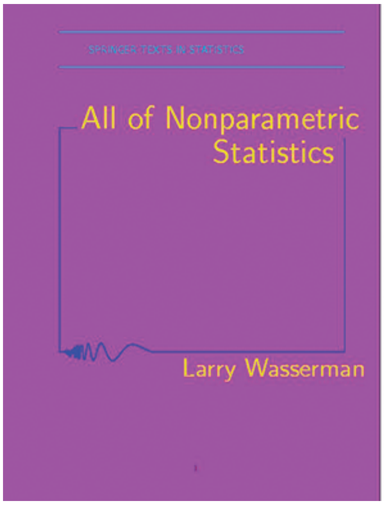
Since these tests involve probability distributions indexed by the value of a parameter, and specifically concern the value taken by that parameter, they are examples of *parametric* tests. What do we do if parametric assumptions like those above do not seem to be justified? One possibility, if we have a large sample and are interested in the value of the population mean, is to appeal directly to the Central Limit Theorem, as we did in developing the z-test in Unit 9. Alternatives which also make no assumption about the specific form of population distribution underlying the data are called *nonparametric* hypothesis tests. Hypothesis tests of this sort are the first topic of this unit. In particular, two nonparametric tests, one appropriate to a one-sample situation, the other to a two-sample situation, are investigated in Section 1.

This raises (once again!) the question of how we can actually tell whether a sample of data could plausibly have been drawn from a particular probability distribution. You have already seen certain graphical methods for investigating this, most explicitly in the form of normal probability plots in Unit 6. In Section 2, you will learn about a more formal, numerical, procedure for testing what is known as the *goodness-of-fit* of a probability model, in the particular context of models for discrete data.

1 Nonparametric tests

See Subsection 3.1 of Unit 9.

The t-test involves the assumption that the underlying distribution of the population is normal. Thus, as mentioned in the Introduction, this test is said to be **parametric**. The distributional properties of the test statistic depend on this parametric assumption of normality. If the underlying population distribution is not normal, the t-test statistic does not in general have a Student's t-distribution, and therefore any p-value that you calculate on the basis of this assumption might be incorrect. How serious this is in practice would depend on how different the underlying distribution is from the assumed normal form. In some cases the discrepancy might be small, and no practical problems might arise; but in other cases the discrepancy might be crucial.



Nonparametrics – statistics without the parametric assumptions – is another huge area of the subject that we do little but scratch the surface of here. This fine book on modern nonparametric statistics has an ambitious title that might be disputed on the grounds that nonparametric tests like those in this unit are not mentioned there!

In this section, techniques that do not require a specific probability model will be explored. Such techniques are called **nonparametric**. In a nonparametric test, there is no assumption that the underlying distribution comes from a specified family indexed by parameters (such as the normal distribution). Since no particular distributional form is assumed, the tests are also sometimes called distribution-free, though, as you will see, this does not mean that there are no distributions involved at all!

Nonparametric statistical tests pre-date parametric tests like the t-test. They can be traced back at least as far as 1710, to a study that was mentioned in Subsection 3.2 of Unit 3. There, we said that:

> John Arbuthnot, having examined parish records in London, noted that in each of the previous 82 years more boys than girls had been christened. He deduced that for 82 years more boys than girls had been born.

Arbuthnot's deduction involved a probability calculation that is nowadays considered to be the first recorded instance of the use of a nonparametric test that is called the *sign test*. You might well have come across the sign test in your earlier statistical studies. However, while the sign test is of historical (and basic pedagogical) interest, it is now rarely used in practice. The reason for this is that it simply ignores too much valuable information, and as a result is not a very powerful test. (What it ignores will be described in Examples 1 and 3.) That is, it is prone to failing to reject a null hypothesis except when the evidence against the null hypothesis is very clear indeed.

For this reason, in Subsection 1.1, as a replacement for the sign test, we will introduce and investigate an alternative, much more powerful, test. A second nonparametric test, for a two-sample rather than one-sample problem, is described in Subsection 1.2. A short discussion of the relative merits of nonparametric and parametric tests completes the section in Subsection 1.3.

1.1 The Wilcoxon signed rank test

The test we are about to investigate is a nonparametric test for a one-sample situation. The test will be introduced via consideration of an example which happens to comprise paired data, a type of data considered briefly in Subsection 3.1 of Unit 9. The test can also be used to test the null hypothesis that a single sample of data is drawn from a population with a specified value of its median. An extension to this case, and a general formulation of the test that covers both cases, will be made later in the subsection.

The Wilcoxon signed rank test for paired differences

Let us start with an example in which there are two, related, observations on each case and interest lies in the set of differences between the values observed on each case.

Activity 1 *Corneal thickness – the t-test?*

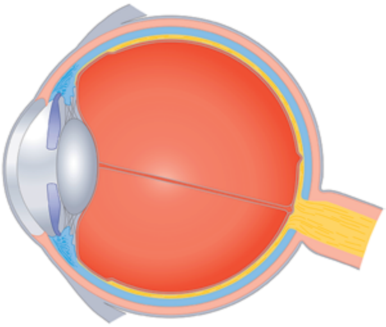
Table 1 gives the thickness, in microns, of the corneas of eight people, each of whom had one eye affected by glaucoma and one eye not.

Modern measurements yield values of corneal thickness somewhat larger than those with which we work here.

Table 1 Corneal thickness in patients with glaucoma (microns)

Patient	1	2	3	4	5	6	7	8
Glaucomatous eye	488	478	480	426	440	410	458	460
Normal eye	484	478	492	444	436	398	464	476

(Source: Ehlers, N. (1970) 'On corneal thickness and intraocular pressure. II. A clinical study on the thickness of the corneal stroma in glaucomatous eyes', *Acta Ophthalmologica*, vol. 48, no. 6, pp. 1107–12)



The structure of the eye; the cornea is the part bulging out to the front of the eye, that is, to the left of this picture

These data were collected to investigate whether, on average, there is a difference between corneal thickness in the eye affected by glaucoma and the other eye. A t-test for zero mean difference would involve calculating the differences between the thicknesses in the two eyes, and assuming that these differences are adequately modelled by a normal distribution.

The differences (glaucomatous eye − normal eye) are as follows.

4 0 − 12 − 18 4 12 − 6 − 16

(a) A normal probability plot of these differences is shown in Figure 1.

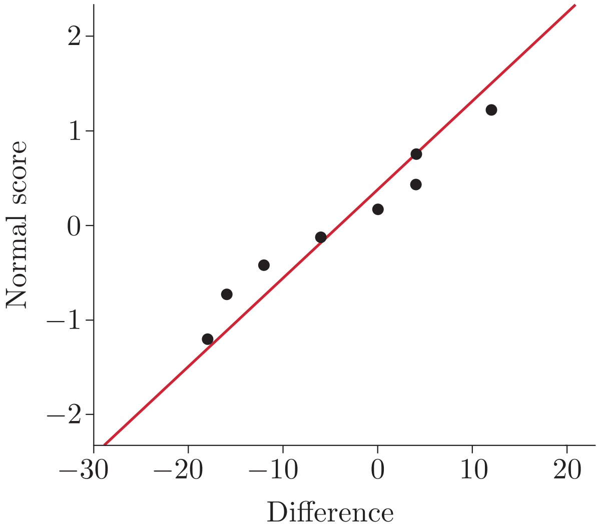


Figure 1 A normal probability plot of corneal thickness differences

Based on this plot, comment on the suitability or otherwise of the normal distribution as a model for the distribution of differences.

(b) The question of whether there is a difference between corneal thickness in the eye affected by glaucoma and the other eye can be framed in terms of the mean difference, μ_D.

(i) Write down the null and alternative hypotheses associated with a two-sided t-test of whether the mean difference is zero.

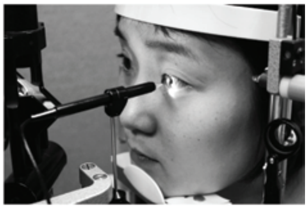
(ii) The t-test of the hypotheses set up in part (b)(i) turns out to give a p-value of 0.327. Interpret the result of this t-test, assuming the t-test to be appropriate.

(iii) Comment on the result in part (b)(ii) in the light of the considerations of part (a).

The normality assumption is a rather precise distributional assumption that, as you have just been reminded, may not always be an appropriate assumption to make. In order to avoid the normality assumption, one approach is to extract only the most important and relevant information from the numbers we have by replacing their exact values by quantities reflecting that information. In the context of Activity 1, relevant information means information that still allows us to investigate whether or not, on average, there is a non-zero difference between corneal thickness in the glaucomatous eye and the other eye. So, what relevant information do we have? Well, we know the signs of the data values, that is, whether they are positive or negative (or, indeed, zero). And we can order data values (indeed, this idea was used to produce the probability plot above). Let's see how these ideas might be taken forward in the context of the example on corneal thickness.

Example 1 *Corneal thickness – the sign test*

The least information that we can keep while still retaining some relevant information about the differences in corneal thicknesses is the signs of the data values (whenever those differences are non-zero). That is, we replace the difference 4 by its positive sign, +, the difference −12 by its negative sign, −, and so on. This has been done for the dataset on corneal thicknesses in Table 2.



Measuring corneal thickness

Table 2 Table 1 expanded to include differences and their signs

Patient	1	2	3	4	5	6	7	8
Glaucomatous eye	488	478	480	426	440	410	458	460
Normal eye	484	478	492	444	436	398	464	476
Difference	4	0	−12	−18	4	12	−6	−16
Sign of difference	+		−	−	+	+	−	−

These signs of differences are the basis of the sign test. This test proceeds by counting the number of differences with + signs and the number with − signs and, effectively, seeing if they are about the same. (It is common practice simply to ignore zeros in the sign test and to reduce the sample size accordingly.)

It turns out, however, that as we have already said, the ensuing sign test is not very powerful. This is because it throws out too much information in replacing exact data values by just their signs. We will therefore not pursue the sign test further here except for making one important observation. Suppose that the number of sample differences with + signs and the number of sample differences with − signs are exactly equal. For the underlying (continuous) population, this estimates the probability of a positive difference and the probability of a negative difference to be the same, and equal to one-half. These population probabilities correspond to the population *median* of the differences being zero. For this reason, the null hypothesis for the sign test is considered to be that the underlying median difference, m_D, is zero:

$$H_0 : m_D = 0.$$

Alternative hypotheses can then be two-sided,

$$H_1 : m_D \neq 0,$$

or one-sided,

$$H_1 : m_D < 0 \quad \text{or} \quad H_1 : m_D > 0,$$

as appropriate.

If the sign test is not especially effective, how else might we nonparametrically test $H_0 : m_D = 0$? The signs of the data values will still come into it, but so will ordering of values or, more accurately, the *ranks* of ordered values.

Suppose that you have n observations, $x_1, x_2, \ldots, x_n$. These observations can be put into numerical order. As in Section 4 of Unit 1 and Section 5 of Unit 6, we rearrange the data in order of increasing size: if the ith ordered observation is denoted by $x_{(i)}$, then

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}.$$

Suppose for a moment that all the x values are distinct so that all the $\leq$ signs above are $<$ signs:

$$x_{(1)} < x_{(2)} < \cdots < x_{(n)}.$$

Then the **ranks** of the x values are their positions in the list, that is, the bracketed subscripts to the ordered values. So the smallest observation has rank 1, the next smallest has rank 2, the next smallest has rank 3, right up to the largest having rank n.

Note that observations are ranked from lowest (rank 1) to highest (rank n), not the other way round . . . which is what it is for most sports and rankings in society!

What if some observations are equal? Well, if two or more observations are tied, then their *average* rank is given to each of the tied observations. This is used in the next example and activity, where we consider the ranks of corneal thickness for glaucomatous and normal eyes separately, for illustration purposes.

Example 2 *Ranks of glaucomatous corneal thicknesses*

The values of corneal thickness in the glaucomatous eyes of each of eight patients were given in Table 1. These values, unordered, are

488 478 480 426 440 410 458 460

Ordered in increasing size, these values are

410 426 440 458 460 478 480 488

and these observations have ranks $1, 2, \ldots, 8$. That is: patient 1, observed value 488, has rank 8; patient 2, observed value 478, has rank 6; and so on. The full set of patient ranks is given in Table 3.

Table 3 Ranks of corneal thickness of glaucomatous eyes

Patient	1	2	3	4	5	6	7	8
Glaucomatous eye	488	478	480	426	440	410	458	460
Rank	8	6	7	2	3	1	4	5

Suppose now, for illustration, that the corneal thickness values above were rounded to the nearest 10 microns. The rounded unordered values are

490 480 480 430 440 410 460 460

Ordered in increasing size, these values are

410 430 440 460 460 480 480 490

The observations now have ranks

$1 \quad 2 \quad 3 \quad 4\frac{1}{2} \quad 4\frac{1}{2} \quad 6\frac{1}{2} \quad 6\frac{1}{2} \quad 8$

Notice that the two values of 460 would have been ranked 4 and 5, so are given the same average rank of $(4+5)/2 = 4\frac{1}{2}$; similarly, the two values of 480 would have been ranked 6 and 7, so are given the same average rank of $(6+7)/2 = 6\frac{1}{2}$. So, for example, while patient 1, observed value 490, still has rank 8, patient 2, observed value 480, now has rank $6\frac{1}{2}$, and so on. The full set of patient ranks is given in Table 4.

Table 4 Ranks of rounded corneal thickness of glaucomatous eyes

Patient	1	2	3	4	5	6	7	8
Glaucomatous eye	490	480	480	430	440	410	460	460
Rank	8	$6\frac{1}{2}$	$6\frac{1}{2}$	2	3	1	$4\frac{1}{2}$	$4\frac{1}{2}$

Activity 2 *Ranks of normal corneal thicknesses*

The values of corneal thickness in the normal eyes of each of eight patients were also given in Table 1. These values, unordered, are

484 478 492 444 436 398 464 476

(a) Obtain the patients' ranks using these observations.

(b) Round each corneal thickness measurement to the nearest 10 microns. Obtain the patients' ranks using these rounded observations.

Returning, then, to the problem of nonparametrically testing the null hypothesis that the median paired difference is zero, ignoring the information on the *size* of the differences, as the sign test does, seems unsatisfactory. It fell to the American chemist and statistician Frank Wilcoxon in 1945 to propose a method of testing which takes into account the relative sizes (but not the exact values) of the differences.

Frank Wilcoxon (1892–1965)

An overview of Wilcoxon's method, in the context of a single sample of differences, is as follows. The null and alternative hypotheses for a two-sided test are

$$H_0: m_D = 0, \quad H_1: m_D \neq 0,$$

where m_D is the population median difference. The test then proceeds as follows. The signs of the differences are initially set to one side, that is, we consider the absolute values of the differences. Ranks are then allocated to the absolute values of the differences, the smallest being given a rank of 1, the next smallest a rank of 2, and so on. These ranks are then allocated to two groups: ranks for differences with a positive sign are allocated to one group; ranks for differences with a negative sign are allocated to the other group. The ranks are added up separately for the two sign groups. If the total for one of the sign groups is very small (which means that the total for the other sign group is very large, because they add up to a fixed total), then the null hypothesis of zero median difference is rejected. The test is called the **Wilcoxon signed rank test**. An example should make the method clear.

The hypotheses are the same as for the sign test at the end of Example 1. We concentrate for the moment on the two-sided version of the test for clarity.

Example 3 *Corneal thickness – the Wilcoxon signed rank test*

We will now develop the two-sided Wilcoxon signed rank test of the hypotheses

$$H_0: m_D = 0, \quad H_1: m_D \neq 0,$$

where m_D is the population median difference between the corneal thicknesses of glaucomatous and normal eyes.

Building on Table 2, Table 5 separates the absolute values of the differences from their associated signs, and shows the ranks of the absolute values.

Table 5 Table 2 expanded to include absolute differences and their ranks

Patient	1	2	3	4	5	6	7	8
Glaucomatous eye	488	478	480	426	440	410	458	460
Normal eye	484	478	492	444	436	398	464	476
Difference	4	0	-12	-18	4	12	-6	-16
Sign of difference	+		−	−	+	+	−	−
Absolute value of difference	4	0	12	18	4	12	6	16
Rank of absolute value of difference	$1\frac{1}{2}$		$4\frac{1}{2}$	7	$1\frac{1}{2}$	$4\frac{1}{2}$	3	6

As in our discussion of the sign test for these data in Example 1, the difference of zero for patient 2 has not been included in the ranking; it is ignored, and the sample size is taken as 7 instead of 8. Notice too that averaging of ranks has needed to be used here because two pairs of absolute differences have the same value.

There is an alternative procedure in which zeros are incorporated into the analysis. Except for very small samples, this alternative approach does not usually lead to substantially different conclusions.

Now the signs are taken into account. The ranks of the positive differences (of which there are three) are $1\frac{1}{2}$, $1\frac{1}{2}$ and $4\frac{1}{2}$ (corresponding to the differences 4, 4 and 12, respectively). The sum of the ranks of the positive differences is therefore

$$w_+ = 1\tfrac{1}{2} + 1\tfrac{1}{2} + 4\tfrac{1}{2} = 7\tfrac{1}{2}.$$

The ranks of the negative differences (of which there are four) are $4\frac{1}{2}$, 7, 3 and 6 (corresponding to the differences -12, -18, -6 and -16, respectively). The sum of the ranks of the negative differences is therefore

$$w_- = 4\tfrac{1}{2} + 7 + 3 + 6 = 20\tfrac{1}{2}.$$

If either of these sums were particularly large or particularly small, this would provide evidence against the null hypothesis of zero median difference.

In general, the sum w, say, where $w = w_+ + w_-$, is equal to $1 + 2 + \cdots + n = \frac{1}{2}n(n+1)$, where n is the sample size (after excluding zeros). In this case, the sample size n is 7, so $w = w_+ + w_-$ should be $\frac{1}{2} \times 7 \times 8 = 28$; this is indeed true for the values $w_+ = 7\frac{1}{2}$ and $w_- = 20\frac{1}{2}$ calculated above.

This result was used in Unit 4. The use of average ranks continues to make this true even where there are ties.

Thus, for a given sample size, w_+ is small exactly when w_- is large, and vice versa. Thus we can concentrate on just one of these quantities. The test statistic for the Wilcoxon signed rank test is taken to be w_+: under the null hypothesis of zero median difference, values of w_+ that are extremely small or extremely large will lead to rejection of the null hypothesis (for a two-sided test).

The null distribution of the test statistic w_+ is different for each value of n, as well as being rather complicated. So to obtain the p-value for a Wilcoxon signed rank test, a computer is generally used.

For the data being considered here, the p-value for a two-sided test with $w_+ = 7\frac{1}{2}$ turns out to be 0.344. On this analysis (as for the t-test in Activity 1 and, had we developed it that far, the sign test) there is little or no evidence of a non-zero difference in corneal thickness between glaucomatous and normal eyes. (Note that interpretation of the p-value for the Wilcoxon signed rank test follows the guidelines of Table 3 of Unit 9 in the same way as for z-tests or t-tests or any other hypothesis test.)

This is not the same as the p-value given by Minitab, which uses an approximation to calculate p-values for the Wilcoxon signed rank test.

Despite our result on this rather old and small dataset, modern ophthalmological thinking is that people with thinner corneas are at higher risk of developing glaucoma.

So far, we have explicitly discussed the two-sided Wilcoxon signed rank test, testing the hypotheses

$$H_0 : m_D = 0, \quad H_1 : m_D \neq 0,$$

where m_D is the population median difference.

Consider now the one-sided Wilcoxon signed rank test testing the hypotheses

$$H_0 : m_D = 0, \quad H_1 : m_D > 0.$$

The test statistic remains w_+, the sum of the ranks of the positive differences. If, as under this one-sided H_1, the median difference is positive, then you would expect rather more and, for many distributions, rather larger, positive differences than negative ones. Since w_+ is the sum of the ranks of the positive differences, large values of w_+ give evidence against H_0 and in favour of H_1. The p-value for the one-sided test is therefore $P(W_+ \geq w_+)$, where W_+ is the random variable version of the test statistic which follows the null distribution which holds when H_0 is true; this is because values greater than or equal to w_+ are at least as extreme as w_+ 'in the direction of' H_1.

This is illustrated in a particular case in Figure 2. There, a plot of the null distribution of the Wilcoxon signed rank test statistic is given for the situation where $n = 7$ and the simplified special case of no tied ranks has been assumed. (This is *not* usually the case of interest in practice, as you saw in Example 3.) In this case, $w = \frac{1}{2}n(n+1) = \frac{1}{2} \times 7 \times 8 = 28$ (so that the possible values of w_+ are $0, 1, 2, \ldots, 28$), and the observed value of W_+ has been assumed to be $w_+ = 24$. The one-sided p-value $P(W_+ \geq 24)$ is also shown on the figure. Notice that the distribution in Figure 2 is symmetric about the value 14; since $w = 28$, this is $w/2$.

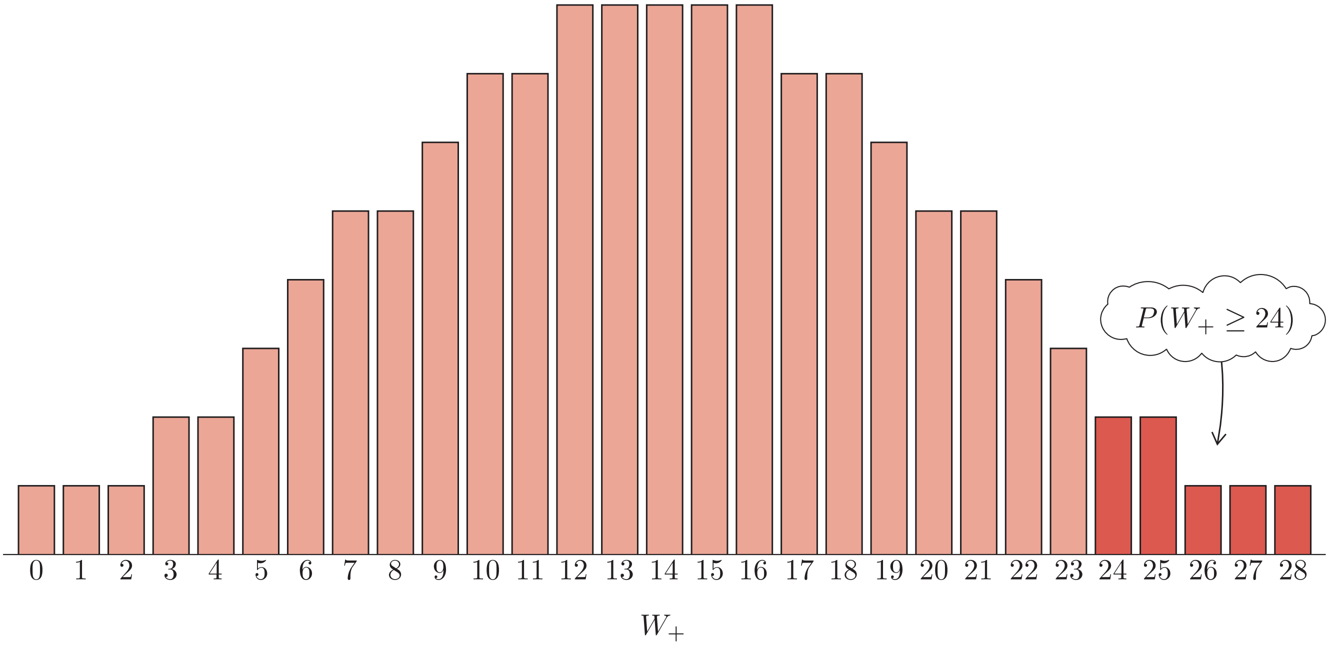


Figure 2 The null distribution of W_+ when $n = 7$, showing the p-value for a one-sided test

Returning attention to the two-sided test, evidence against H_0 is provided by values of w_+ that are extremely small or extremely large. Suppose that $w_+ > w_-$, so that the observed value w_+ is in the upper tail of the null distribution of W_+. The two-sided p-value is the sum of the probabilities that, if H_0 is true, the random variable W_+ is greater than or equal to its observed value w_+ and is less than or equal to its equally extreme value 'in the opposite direction'. But what do we mean by 'equally extreme in the opposite direction' in this case?

Figure 3 repeats the null distribution of the Wilcoxon signed rank test given in Figure 2. Here, $w = 28$ and we have observed $w_+ = 24$. From consideration of the distribution in Figure 3, it is clear that the value 'equally extreme in the opposite direction' to $w_+ = 24$ is 4. This comes from the symmetry of the distribution and the fact that $w - w_+ = 28 - 24 = 4$. The two-sided p-value is therefore $P(W_+ \geq 24) + P(W_+ \leq 4)$. There is also a rather strong indication from the figure that $P(W_+ \leq 4) = P(W_+ \geq 24)$, so the two-sided p-value appears to be

$$p = P(W_+ \geq 24) + P(W_+ \leq 4) = 2P(W_+ \geq 24),$$

that is, twice the one-sided p-value. Is this true in general?

Figure 3 also shows the two-sided p-value for this test if the observed value of w_+ had been 4.

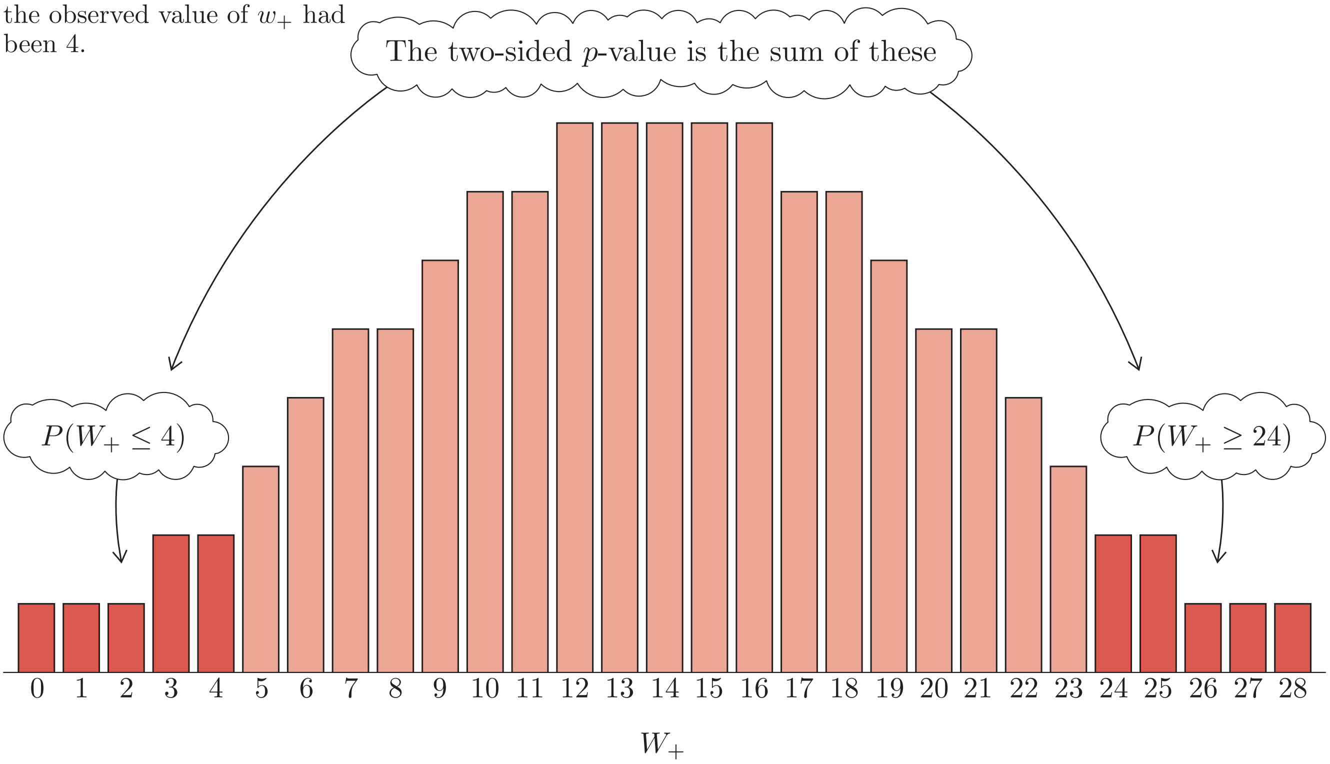


Figure 3 The null distribution of W_+ when $n = 7$, showing the p-value for a two-sided test

The argument is similar to that for z- and t-tests made in Section 4 of Unit 9, except for the definition of 'equally extreme in the opposite direction'. There, the null distributions are symmetric about zero, so x and $-x$ are 'equally extreme' values.

Well, it is true that, in general, even when tied values are accounted for, the null distribution of the Wilcoxon signed rank test statistic is symmetric. However, the symmetry here is not about the value zero (as it was in Unit 9 for the z- and t-tests) but, as you have seen in a special case, about the value $w/2$. Now, adding and subtracting $w/2$, we can write $w_+ = (w/2) + (w_+ - (w/2))$. Therefore, the value as far from $w/2$ as w_+ but in the opposite direction must be $(w/2) - (w_+ - (w/2))$. And, remembering that $w = w_+ + w_-$, this value is none other than w_- since

$$(w/2) - (w_+ - (w/2)) = w - w_+ = w_-.$$

In particular, if H_0 is true, then we have that

$$P(W_+ \geq w_+) = P(W_+ \leq w_-). \qquad (1)$$

So, in the two-sided test, the p-value is

$$p = P(W_+ \geq w_+) + P(W_+ \leq w_-).$$

But Equation (1) implies that

$$p = 2P(W_+ \geq w_+).$$

Therefore, the two-sided p-value is indeed twice the one-sided p-value or, equivalently, the one-sided p-value is one-half of the two-sided p-value.

A similar argument applies to the one-sided Wilcoxon signed rank test for the situation

Details are omitted.

$$H_0 : m_D = 0, \quad H_1 : m_D < 0.$$

The test statistic remains w_+, the sum of the ranks of the positive differences. But small values of w_+ give evidence against H_0 in the direction of H_1 in this case. The one-sided p-value is again one-half of the two-sided p-value.

The Wilcoxon signed rank test in general

The Wilcoxon signed rank test has been introduced above as a test of the median difference between paired values. Because the data are differences, the natural null hypothesis to consider has been that the median difference is zero. However, the Wilcoxon signed rank test can also be used to test the null hypothesis that a single sample of data is drawn from a population with a specified, and quite likely non-zero, median m_0. This can be achieved simply by subtracting the specified median from each data value and then testing in the same way as above for zero median of the data values minus m_0.

Activity 3 *Byzantine coins from the second coinage*

In Unit 6, data were introduced on the silver (Ag) content of coins from different coinages of the reign of the Byzantine king Manuel I Comnenus. The coins in question arose from four different coinages. In your Minitab work associated with that unit, you explored, using probability plots, the normality or otherwise of the distribution of the coins from each of the coinages. The samples of coins are all small, but the investigation suggested that while normality was an appropriate assumption for coins of the first and fourth coinages, normality is more questionable for coins of the second and third coinages.

A coin from the reign of Manuel I Comnenus. It contains silver, but that doesn't necessarily mean it's silver coloured!

The data associated with the *second* coinage – one of those of 'more questionable' normality – are given in Table 6 (overleaf).

Table 6 Silver content of coins: second coinage (% Ag)

6.9	9.0	6.6	8.1	9.3	9.2	8.6

Suppose that an archaeologist wishes to investigate whether it is plausible that the coins from the second coinage could come from a population where the median silver content is 7.5%.

(a) What are the null and alternative hypotheses associated with the two-sided version of the Wilcoxon signed rank test?

(b) Subtract the hypothesised median value, m_0, from each of the data values. What are the null and alternative hypotheses associated with the two-sided version of the Wilcoxon signed rank test based on the data values minus m_0?

So, the Wilcoxon signed rank test can be thought of as always involving looking at *differences* – either the differences between paired data, as in Example 3, or the differences between the data and the specified value of the median, as in Activity 3. The procedure for performing the Wilcoxon signed rank test is summarised in the following box.

The Wilcoxon signed rank test

The Wilcoxon signed rank test is a test on a single sample of data, $x_1, x_2, \ldots, x_{n_1}$. Let m denote the underlying population median and m_0 a specified value for it.

1 Determine the null and alternative hypotheses. The null hypothesis for the test takes the form

$$H_0 : m = m_0.$$

For a two-sided test, the alternative hypothesis is

$$H_1 : m \neq m_0.$$

For a one-sided test, the alternative hypothesis is

$$H_1 : m < m_0 \quad \text{or} \quad H_1 : m > m_0,$$

as appropriate.

2 (a) If the data are a set of paired differences, then we are testing for a zero median, so $m_0 = 0$; set $d_i = x_i$, $i = 1, 2, \ldots, n_1$.

(b) For a single sample for which we are testing for a non-zero median, $m_0 \neq 0$, form the differences from the specified value m_0: $d_i = x_i - m_0$, $i = 1, 2, \ldots, n_1$.

3 In either case of step 2, delete any zeros from the dataset of differences and let $n \leq n_1$ be the sample size of the dataset with zeros removed.

4 As the hypothesised median value is now 0, reframe the null hypothesis for the test based on differences as

$$H_0 : m = 0.$$

For a two-sided test, the alternative hypothesis is

$$H_1 : m \neq 0.$$

For a one-sided test, the alternative hypothesis is

$$H_1 : m < 0 \quad \text{or} \quad H_1 : m > 0,$$

as appropriate. These hypotheses based on the differences are equivalent to the hypotheses for the original data in step 1.

5 Without regard to their sign, order the absolute values of the differences from least to greatest, and allocate rank i to the ith absolute difference. In the event of ties, allocate the average rank to the tied differences.

6 Now consider again the signs of the original differences. Denote by w_+ the sum of the ranks of the positive differences. This is the Wilcoxon signed rank test statistic.

7 Obtain the p-value, p (usually by using computer software).

8 Interpret p and state your conclusions.

Activity 4 *Foetal movements*

Chorionic villus sampling is a test to detect genetic disorders in a foetus at an early stage of pregnancy. Such a test should have no effect on the amount of foetal movement before and after the test. Foetal movement can be measured by ultrasound and was checked for 10 women in a research study. The results are given in Table 7 in the form of differences between the percentage of time the foetus was moving pre-test and the percentage of time the foetus was moving post-test.

Table 7 Percentage of time moving pre-test minus percentage of time moving post-test

Patient	1	2	3	4	5	6	7	8	9	10
Difference	7	−3	3	−5	3	−1	−3	−2	1	8

(Source: Boogert, A., Mantingh, A. and Visser, G.H.A. (1987) 'The immediate effects of chorionic villus sampling on fetal movements', *American Journal of Obstetrics and Gynecology*, vol. 157, no. 1, pp. 137–9)

The small integer nature of the data suggests that a normal distribution is not a suitable model for these data (a fact borne out by a normal probability plot, not shown). You are therefore asked to analyse the data using the Wilcoxon signed rank test, which does not assume normality (so works just as well for discrete data like these). The hypotheses are

$$H_0 : m = 0, \quad H_1 : m \neq 0,$$

where m is the (population) median difference between pre-test and post-test movement. Notice that this is an occasion on which the 'no difference' nature of the null hypothesis is the one that researchers would prefer not to find evidence against.

(a) Calculate the Wilcoxon signed rank test statistic for the data.

(b) The p-value for this test is 0.709. What do you conclude?

Activity 5 *Byzantine coins – completing the test*

In Activity 3, you considered data on the silver content of coins from the second coinage of the reign of the Byzantine king Manuel I Comnenus. Interest lay in testing the hypotheses

$$H_0 : m = 7.5, \quad H_1 : m \neq 7.5,$$

where m is the median % Ag content of the coins from the second coinage. In part (b) of Activity 3, you produced a table (reproduced here as Table 8) of differences between the original data values and the hypothesised value of the median, $m_0 = 7.5$.

Table 8 Silver content minus 7.5 (% Ag)

−0.6	1.5	−0.9	0.6	1.8	1.7	1.1

In this activity, you will complete the Wilcoxon signed rank test of the above hypotheses.

(a) Calculate the Wilcoxon signed rank test statistic for these data.

(b) The p-value for this test is 0.125. What do you conclude?

In Activity 6, you are asked to use the Wilcoxon signed rank test to investigate the Shoshoni rectangle data that you first met in a computer activity associated with Section 5 of Unit 6. In fact, there is a snag in using the test with this dataset. However, ignore that for now and proceed with the test. (This snag is discussed later in the subsection.)

Activity 6 *Shoshoni rectangles*

In your computer work on Unit 6, data on the width-to-length ratios of twenty rectangles produced by Shoshoni native North Americans were considered. Interest lies in whether (in some average sense) the ratios matched the ancient Greek 'golden ratio' of $(\sqrt{5}+1)/2 \simeq 0.618$. The data are given in Table 9.

Many buildings have the golden ratio in them, the Parthenon in Athens being a prime example

Table 9 Width-to-length ratios of Shoshoni rectangles

0.693	0.662	0.690	0.606	0.570	0.749	0.672	0.628	0.609	0.844
0.654	0.615	0.668	0.601	0.576	0.670	0.606	0.611	0.553	0.933

In Unit 6, you produced a normal probability plot of these data in order to investigate whether a normal model might be appropriate; such a plot is provided in Figure 4. The normal probability plot is actually decidedly curved, so there must be some doubt over the appropriateness of an assumption of normality. In particular, a t-test might not be appropriate for testing the mean of this distribution.

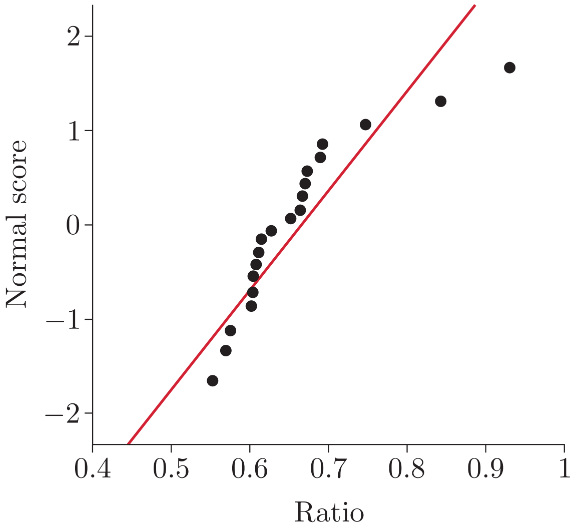


Figure 4 A normal probability plot of Shoshoni rectangle ratios

An analysis which avoids a normality assumption uses the Wilcoxon signed rank test to test the null hypothesis that the population median m of width-to-length ratios of Shoshoni rectangles is 0.618 against the alternative hypothesis that it is not 0.618:

$$H_0 : m = 0.618, \quad H_1 : m \neq 0.618.$$

(a) Obtain a table of differences for the data on width-to-length ratios of Shoshoni rectangles by subtracting the value in the null hypothesis, 0.618, from each of the data values in Table 9. Allocate ranks to the absolute values of the differences, and hence calculate the Wilcoxon signed rank test statistic, w_+.

(b) The Wilcoxon signed rank test was used to test the hypotheses above. The p-value for the test turns out to be 0.088. What can you conclude?

There is a sense in which the Central Limit Theorem operates with the Wilcoxon signed rank test statistic: provided that the number of differences is sufficiently large, a normal approximation to the null distribution of the test statistic may be used. This approximation is described in the following box.

Normal approximation to the null distribution of the Wilcoxon test statistic

Under the null hypothesis of zero median difference, for a sample of size n (excluding any zero differences), the random variable W_+, whose observed value is the Wilcoxon signed rank test statistic w_+, has mean and variance given by

$$E(W_+) = \frac{n(n+1)}{4}, \quad V(W_+) = \frac{n(n+1)(2n+1)}{24}.$$

The distribution of

$$Z = \frac{W_+ - E(W_+)}{\sqrt{V(W_+)}}$$

is approximately standard normal.

The approximation above is quite good, but should not be used for sample sizes that are very small. As a rule of thumb, the normal approximation is generally adequate as long as the sample size n is at least 16. The following example and activity illustrate inappropriate and appropriate use of the normal approximation, respectively!

Example 4 *Corneal thickness – normal approximation*

In Activity 1, Example 1 and Example 3, we looked at differences in corneal thickness in patients with one normal and one glaucomatous eye. There were seven such differences (excluding one with zero difference), so

$$E(W_+) = \frac{n(n+1)}{4} = \frac{7 \times 8}{4} = 14,$$

$$V(W_+) = \frac{n(n+1)(2n+1)}{24} = \frac{7 \times 8 \times 15}{24} = 35.$$

In this case, the observed sum of ranks for the positive differences is $w_+ = 7\frac{1}{2}$, so the corresponding observed value of Z is

$$z = \frac{w_+ - 14}{\sqrt{35}} = \frac{7\frac{1}{2} - 14}{\sqrt{35}} \simeq -1.10.$$

Using the table of probabilities of the standard normal distribution in the Handbook gives

$$P(Z \leq -1.10) = 1 - \Phi(1.10) = 0.1357 \simeq 0.136.$$

So, according to the approximation, the probability of obtaining a Wilcoxon signed rank test statistic of $7\frac{1}{2}$ or less is approximately 0.136. For a two-sided test, the p-value is double this:

$$P(Z \leq -1.10) + P(Z \geq 1.10) = 2P(Z \leq -1.10) \simeq 0.271.$$

Here, however, the (effective) sample size is only $n = 7$, and the approximate p-value of 0.271 is noticeably different from the exact p-value of 0.344 given in Example 3. But 7 is a lot less than the minimum sample size of 16 given in the rule of thumb for adequacy of the normal approximation, so it is not very surprising that the approximation is quite poor.

Activity 7 *Shoshoni rectangles – normal approximation*

In Activity 6, you found that the value of the Wilcoxon signed rank test statistic is 151 for the data on width-to-length ratios of Shoshoni rectangles. Use a normal approximation to test the null hypothesis that the Shoshoni rectangles conform (on average) to the Greek golden ratio standard. Compare the p-value you obtain using the approximation with the exact p-value (0.088) given in Activity 6.

We should next spend a little time further considering the assumptions behind the Wilcoxon signed rank test. The test is indeed nonparametric, in that it does not involve an assumption that the data can be modelled by a particular parametric family of distributions. But that does not mean that it does not involve any assumptions at all about the population distribution. Its advantage over the sign test is that it makes some use of the sizes of the differences, instead of just using their signs. For example, in the Shoshoni rectangles example of Activities 6 and 7, the weak evidence against the null hypothesis arises because 8 out of the 9 largest absolute differences between width-to-length ratios and their hypothesised value of 0.618 are positive. The sign test would ignore the information about the relative size of the absolute differences, and indeed fails to find evidence against H_0 for these data ($p = 0.824$).

This increased power, however, comes at a price. The null distribution of the Wilcoxon signed rank test statistic is found by assuming that an absolute difference with a particular rank is just as likely to be associated with a positive difference as with a negative one. Suppose, however, that the null hypothesis of the test (zero population median difference) is true, but that the differences have a distribution that is not symmetric about zero. For definiteness, suppose that the true distribution of differences is right-skew, that is, that its upper tail (positive values) is more spread out than its lower tail. Then, because we have assumed that the median difference is zero, we would expect the number of positive differences to be about the same as the number of negative differences, but on the whole the positive differences would tend to be larger in absolute value than the negative differences. This is illustrated in Figure 5 (overleaf). In other words, a difference whose absolute value had a large rank would be more likely to be positive than to be negative. If this were indeed the case, the null distribution of the Wilcoxon test statistic would be wrong.

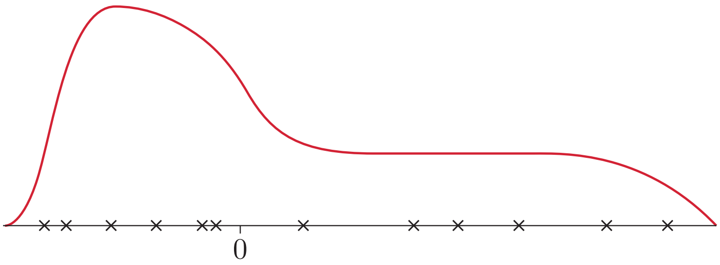


Figure 5 A right-skew distribution with zero median, and some sample values therefrom

Therefore, in order to use information about the relative size of the differences and use the Wilcoxon signed rank test, we must make an assumption that the differences can reasonably be modelled by a *symmetric* distribution, at least under the null hypothesis. The particular shape of the distribution does not matter at all, *as long as it is symmetric.* If this is not the case, the test is not valid.

In many circumstances, particularly for differences in paired data, such an assumption of symmetry is perfectly reasonable. However, this is not always so.

Example 5 *Shoshoni rectangles – assumptions*

The argument regarding the symmetry of the distribution might be 'taken with a pinch of salt' because the sample size is quite small, $n = 20$

Judging from the sample values, the normal probability plot of the data given in Figure 4, and especially the histogram of the data given in Figure 6, it looks as if it may well be inappropriate to model the Shoshoni rectangles data by a symmetric distribution because the sample is quite heavily right-skew. Thus it may be inappropriate to use the Wilcoxon signed rank test with these data. (We already doubted the reasonableness of a normality assumption, and hence of a t-test, in Activity 6, based on Figure 4.)

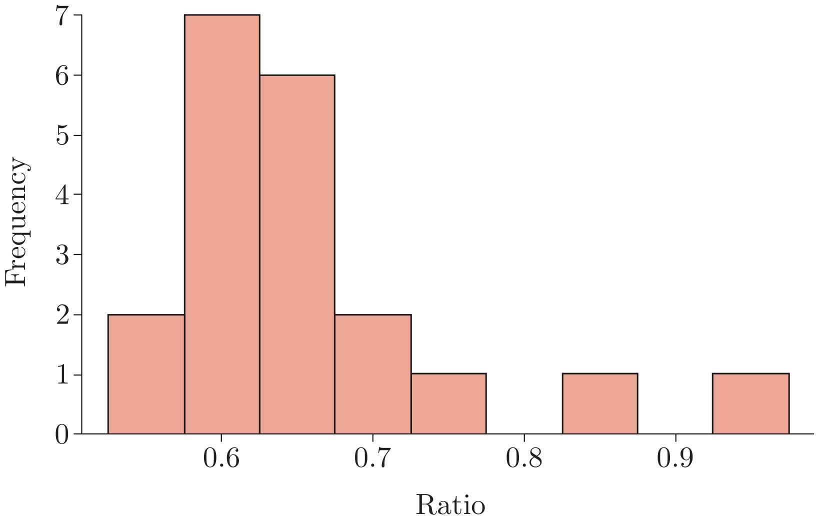


Figure 6 A histogram of the Shoshoni rectangles data

Now, had a two-sided t-test of whether the population mean width-to-length ratio is 0.618 been conducted anyway, the p-value would have been 0.054. One plausible explanation for the fact that both the t-test and the Wilcoxon signed rank test (with a p-value of 0.088, as given in Activity 6) provide some evidence against the null hypothesis may be simply that both reflect the skewness of the data, rather than the possibility that their mean or median is truly different from 0.618. In fact, further investigation, using yet other methods, indicates that the Wilcoxon signed rank test and the t-test do *not* provide misleading information in this case, even though the sample data are skew. That is, the data do really provide some weak evidence that the underlying mean or median is not 0.618.

Note that since the Wilcoxon signed rank test involves an assumption of symmetry, the care that was taken to express its null hypothesis as

$$H_0 : m = 0$$

(where m is the population median) was actually misplaced. If the underlying distribution is symmetric, then its median is equal to its mean, so we could have written the null hypothesis as

$$H_0 : \mu = 0$$

(with corresponding forms for the alternative hypothesis), just as for the one-sample t-test.

The calculations for the Wilcoxon signed rank test statistic are tedious except for very small samples, and we have not told you how to calculate the exact p-value for such a test at all. These calculations are, in general, best done on a computer. To complete this subsection, you will learn how to use Minitab to carry out this test.

Refer to Chapter 10 of Computer Book B for the rest of the work in this subsection.

1.2 The Mann–Whitney test

The idea of using ranks instead of the original data values, as in the Wilcoxon signed rank test, is a logical and appealing one. Furthermore, it has an extension to testing for a difference between two samples of independent, unpaired, data. This two-sample test was first proposed by H.B. Mann and D.R. Whitney in 1947, and later (in a different form that was shown to be equivalent) by Wilcoxon. We will follow the most usual convention and call it the Mann–Whitney test although it is sometimes called the Mann–Whitney–Wilcoxon test.

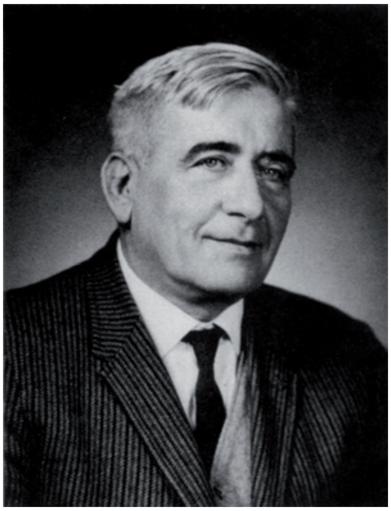
A recent article in the *Electronic Journal for History of Probability and Statistics* claims that 'the test was independently developed by at least six researchers in the late 1940s and early 1950s'!

The two-sample t-test was not covered in Unit 9, but a brief mention of its hypotheses and assumptions is useful here so that we can appreciate the differences between it – a parametric two-sample test – and the Mann–Whitney test – a nonparametric two-sample test. The two-sample

These are the same assumptions that underlie the two-sample t confidence interval for the difference between two population means investigated in Subsection 4.4 of Unit 8.

t-test is a test of the null hypothesis that the population means underlying two independent samples of data are equal. Moreover, the (strong) assumptions underlying the two-sample t-test are that each sample is normally distributed and that the two population variances are equal.

The Mann–Whitney test, being based on ranks, does not test the same null hypothesis as the t-test. It can perhaps best be thought of as a test of the null hypothesis that both samples are drawn from the same population distribution, with no assumptions being made about the form of this distribution (so that, for instance, no assumption even of symmetry of the underlying distributions is needed). The null hypothesis in the two-sample t-test is that the two population means are equal. But since the assumptions of the t-test are that both populations have normal distributions and equal variances, this means that, under the null hypothesis, the two samples are drawn from the same *normal* distribution.



Henry Berthold Mann (1905–2000) won the prestigious Cole Prize in Number Theory in 1946. (The next winner of the Cole prize was the very famous mathematician Paul Erdös.)

Under the alternative hypothesis of the two-sample t-test, the assumptions of normality and equal variance continue to hold, but the population means differ; that is, the population distributions are the same shape and have the same spread, but differ in location. Theoretically, for the Mann–Whitney test, the alternative hypothesis can include situations in which the two population distributions differ in shape and/or spread as well as location. However, in most situations, it makes sense to think of the alternative hypothesis in terms of a difference in location between the two populations. In this sense, the Mann–Whitney test is a valid alternative to the two-sample t-test, without the necessity of assuming normal population distributions.

Suppose, then, that we have two independent samples. The Mann–Whitney test may be used to test the null hypothesis that the samples arise from the same population, in particular, there being no difference in location between them, using the procedure given in the box below. Many of the ingredients of the Mann–Whitney test are similar to the Wilcoxon signed rank test, with which you have just become familiar.

The Mann–Whitney test

The Mann–Whitney test is a test on two independent samples of data. Let ℓ denote the underlying difference in location between the populations from which the samples were drawn. (No specific measure of location (e.g. mean, median or other) need be specified.)

1 Determine the null and alternative hypotheses. The null hypothesis for the test takes the form

$$H_0 : \ell = 0.$$

For a two-sided test, the alternative hypothesis is

$$H_1 : \ell \neq 0.$$

For a one-sided test, the alternative hypothesis is

$$H_1 : \ell < 0 \quad \text{or} \quad H_1 : \ell > 0,$$

as appropriate.

2 Pool (that is, combine) the two samples, keeping track of the sample to which each data value belongs.

3 Order the pooled data values from least to greatest, and allocate rank i to the ith pooled value. In the event of ties, allocate the average rank to the tied values.

4 Call one of the samples sample A (with sample size n_A) and the other sample B (with sample size n_B). Add up the ranks for sample A and write

$$u_A = \text{the sum of the ranks for sample A.}$$

5 The Mann–Whitney test statistic is u_A. Very small and/or very large observed values provide evidence against the null hypothesis, suggesting respectively that the values in sample A are 'too frequently' smaller than or larger than the values in sample B.

For a two-sided test, choosing sample A to be the smaller sample saves a little on calculations.

Let us delay consideration of the distribution of the Mann–Whitney test statistic under the null hypothesis until you have first had some practice at calculating the test statistic itself.

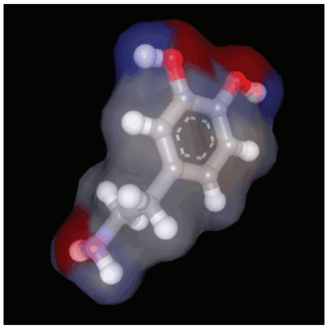
Example 6 *Dopamine activity – the Mann–Whitney test statistic*

In a study into the causes of schizophrenia, 25 hospitalised patients with schizophrenia were treated with antipsychotic medication, and after a period of time were classified as psychotic or non-psychotic by hospital staff. A sample of cerebro-spinal fluid was taken from each patient and assayed for dopamine β-hydroxylase enzyme activity. (Dopamine is a chemical messenger.) The data are given in Table 10; the units are nmol/(ml)(h)/mg of protein. The sample sizes are $n_A = 10$ and $n_B = 15$; notice that we have assigned the label A to the smaller of the two samples. The data are also graphed in a comparative boxplot in Figure 7 (overleaf).

Table 10 Dopamine β-hydroxylase activity (nmol/(ml)(h)/mg)

(A) Judged psychotic							
0.0150	0.0204	0.0208	0.0222	0.0226	0.0245	0.0270	0.0275
0.0306	0.0320						
(B) Judged non-psychotic							
0.0104	0.0105	0.0112	0.0116	0.0130	0.0145	0.0154	0.0156
0.0170	0.0180	0.0200	0.0200	0.0210	0.0230	0.0252	

(Source: Sternberg, D.E. et al. (1982) 'Schizophrenia: dopamine β-hydroxylase activity and treatment response', *Science*, vol. 216, no. 4553, pp. 1423–5)



A model of a dopamine molecule

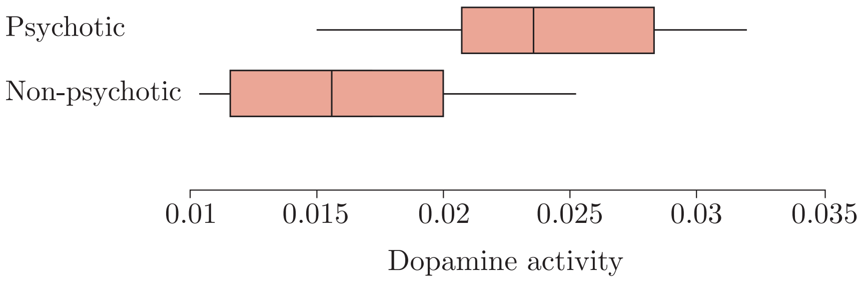


Figure 7 A comparative boxplot of the dopamine activity data

The Mann–Whitney test may be used to test the hypothesis that the distribution of dopamine β-hydroxylase activity is the same for patients judged non-psychotic as for patients judged psychotic. The first step is to pool the two samples, that is, to combine them into one. Then the combined sample is ordered so that the ranks of the data values *in the combined sample* can be obtained.

When the data have already been ordered *within each sample*, as they have been in Table 10, it is easier to assign pooled ranks using a different table format. This is illustrated for the dopamine data in Table 11. First, list the two samples of data vertically. Then, starting from the top, cast your eye down the next values in each column, allocating ranks from 1 onwards as you go. So, for example, 0.0104 is smaller than 0.0150, so 0.0104 gets rank 1. Now compare 0.0105 and 0.0150, the smallest remaining unranked values in each column: 0.0105 is smaller, so it gets rank 2. Next, it's 0.0112 versus 0.0150, 0.0112 being smaller, so getting rank 3. And so on, taking care when you come to tied values.

Table 11 Pooled and ranked data

(A) Psychotic	Rank	(B) Non-psychotic	Rank
0.0150	7	0.0104	1
0.0204	14	0.0105	2
0.0208	15	0.0112	3
0.0222	17	0.0116	4
0.0226	18	0.0130	5
0.0245	20	0.0145	6
0.0270	22	0.0154	8
0.0275	23	0.0156	9
0.0306	24	0.0170	10
0.0320	25	0.0180	11
		0.0200	$12\frac{1}{2}$
		0.0200	$12\frac{1}{2}$
		0.0210	16
		0.0230	19
		0.0252	21

It is fairly clear from the table that the values in sample A on the whole have larger ranks than the values in sample B, though it is not entirely straightforward to make this comparison because of the different sample sizes. Summing the ranks for each member of sample A gives

$$u_A = 7 + 14 + 15 + 17 + 18 + 20 + 22 + 23 + 24 + 25 = 185.$$

This is the value of the Mann–Whitney test statistic in this case.

Had you chosen the other sample as sample A, the value of the Mann–Whitney test statistic would have been the sum of the ranks in the other sample, namely,

$$1 + 2 + 3 + \cdots + 19 + 21 = 140.$$

Using this test statistic instead of the one designated u_A above would still work because the null distribution would be adjusted accordingly. A possibly useful check on the arithmetic is that if

$$u_B = \text{the sum of the ranks for sample B},$$

then the sum of u_A and u_B is

$$u_A + u_B = 1 + 2 + \cdots + (n_A + n_B) = \tfrac{1}{2}\,(n_A + n_B)\,(n_A + n_B + 1)\,.$$

For the current dataset, $u_A + u_B = 185 + 140 = 325$ and

$$\tfrac{1}{2}\,(n_A + n_B)\,(n_A + n_B + 1) = \tfrac{1}{2} \times 25 \times 26 = 325.$$

The Wilcoxon signed rank test statistic of Subsection 1.1 is based on obtaining the absolute values of differences, ranking them and adding up the ranks associated with positive differences. The Mann–Whitney test statistic is based on obtaining a combined sample of values, ranking them and adding up the ranks associated with sample A. The two differ mainly in how the values to be ranked are obtained from the raw data: as absolute differences in the Wilcoxon case, as a combined sample in the Mann–Whitney case.

Activity 8 *Aboriginal peoples of Alaska and California – the Mann–Whitney test statistic*

The lives of hunter-gatherer tribes on the west coast of North America were based on aquatic resources. They therefore had seasonal permanent villages. There is interest in comparing the sizes of the groups associated with these villages along the Californian coast (sample A; $n_A = 12$) and along the Alaskan coast (sample B; $n_B = 13$). The data (they are estimated averages, hence occasionally half-people!) are given in Table 12.

Tlingit totem pole and community house, Alaska

Table 12 Village group size

(A) California												
23	26	30	33	42	45	45	50	50.5	96	113	557	
(B) Alaska												
39	48	53.5	55	57	66	77	79	108	121	162	197	309

(Source for Table 12: Binford, L.R. (2002) *Constructing Frames of Reference: an Analytical Method for Archaeological Theory Building Using Hunter-Gatherer and Environmental Data Sets*, University of California Press)

Construct a table of pooled and ranked data from the data in Table 12 (in the style of Table 11). Hence calculate the value of the Mann–Whitney test statistic u_A.

Considering the test statistic u_A as the observed value of a random variable U_A, it may then be compared with the null distribution of U_A to yield a p-value for the test. As for the Wilcoxon signed rank test, the null distribution is complicated: its calculation would normally require the use of a computer. However, at least when there are no ties (values that are the same) in the data, the null distribution is symmetric, so the p-value for a two-sided test is exactly double that for the corresponding one-sided test.

In Subsection 1.1, you saw that there is a normal approximation for the null distribution of the Wilcoxon signed rank test statistic. There is also a normal approximation for the null distribution of the Mann–Whitney test statistic, which may be used to calculate approximate p-values. This approximation is given in the following box, followed by a rough rule for when the approximation is adequate.

Normal approximation to the null distribution of the Mann–Whitney test statistic

For independent samples of sizes n_A and n_B, the null distribution of the random variable U_A, whose observed value is the Mann–Whitney test statistic u_A, has mean and variance given by

$$E(U_A) = \frac{n_A\,(n_A + n_B + 1)}{2}, \quad V(U_A) = \frac{n_A n_B\,(n_A + n_B + 1)}{12}.$$

The distribution of

$$Z = \frac{U_A - E(U_A)}{\sqrt{V(U_A)}}$$

is approximately standard normal.

This normal approximation can be used for quite modest values of n_A and n_B; say, each of size 8 or more. (Also, it is valid as long as the number of tied values in the pooled dataset is not too great.) The normal approximation to the null distribution of the Mann–Whitney statistic is illustrated as we complete our analysis of the dopamine dataset in Example 7.

Example 7 *Dopamine activity – completing the test*

In Example 6, we calculated the value of the Mann–Whitney test statistic for comparing the distribution of dopamine β-hydroxylase enzyme activity in patients judged to be psychotic (sample A) and non-psychotic (sample B). We found that $u_A = 185$; also, $n_A = 10$ and $n_B = 15$.

Let ℓ denote the population difference in location between the distributions for the psychotic and non-psychotic patients. Let us continue from the results of Example 6 to complete the two-sided Mann–Whitney test of

$$H_0 : \ell = 0$$

against

$$H_1 : \ell \neq 0.$$

The expected value of U_A under the null hypothesis that $\ell = 0$ and hence that the two samples are from identical populations is

$$E(U_A) = \frac{n_A\,(n_A + n_B + 1)}{2} = \frac{10(10 + 15 + 1)}{2} = 130.$$

The observed value $u_A = 185$ is substantially larger than this (in accord with our observation in Example 6 that the values in sample A tend to be larger than the values in sample B), but is it significantly larger? When there are ties in the data (there is one tie here), the null distribution of U_A can have a very complicated shape with many modes. Exact computation of p-values in such a context is quite difficult. A computer gives the (two-sided) p-value as $p = 0.0015$.

Alternatively, the variance of U_A under the null hypothesis is

$$V(U_A) = \frac{n_A n_B\,(n_A + n_B + 1)}{12} = \frac{10 \times 15 \times 26}{12} = 325.$$

For the observed value $u_A = 185$, the corresponding z value is

$$z = \frac{u_A - 130}{\sqrt{325}} = \frac{185 - 130}{\sqrt{325}} \simeq 3.05.$$

So the approximate p-value for the two-sided test based on the normal approximation is

$$\begin{aligned} p &= P(Z \geq 3.05) + P(Z \leq -3.05) = 2P(Z \geq 3.05) \\ &= 2(1 - P(Z < 3.05)) = 2(1 - \Phi(3.05)) \\ &= 2(1 - 0.9989) = 0.0022. \end{aligned}$$

This is close to the exact p-value.

The p-value, however calculated, is very small, so there is strong evidence that the difference in locations of the distributions of dopamine activity in the two groups is non-zero. The dopamine activity in psychotic and non-psychotic patients appears to differ and, looking at the data in Table 10 and/or the comparative boxplot in Figure 7, it would seem that those judged psychotic (the population corresponding to sample A) have higher dopamine activity, on average. Here we use the loose term ‘on average’ as shorthand for ‘in terms of the locations of their distributions’.

Activity 9 *Aboriginal peoples of Alaska and California – completing the test*

In Activity 8, you calculated the value of the Mann–Whitney test statistic for comparing the distributions of average aboriginal village group size in

California (sample A) and Alaska (sample B). You found that $u_A = 115$; also, $n_A = 12$ and $n_B = 13$.

Complete the test of the null hypothesis that there was no difference between the locations of the distributions of average village group size in California and Alaska against the two-sided alternative hypothesis that there was a non-zero difference in location, using the normal approximation to the distribution of U_A under H_0. (Both sample sizes are greater than 8 in this case.)

A further activity will allow you to perform a complete Mann–Whitney test from the start.

Activity 10 *Recall of pleasant and unpleasant memories*

We now revisit a dataset that you first met (in pictorial form) in Subsection 5.2 of Unit 1. Let us repeat its description from there:

> In a study of memory recall times, a series of stimulus words was shown to a subject on a computer screen. For each word, the subject was instructed to recall either a pleasant or an unpleasant memory associated with that word. Successful recall of a memory was indicated by the subject pressing a bar on a computer keyboard. Of key interest in this study was whether pleasant memories could be recalled more easily and quickly than unpleasant ones.

The data on recall times (in seconds) for twenty pleasant and twenty unpleasant memories are given in Table 13. In Unit 1, you were presented with unit-area histograms of these data, which showed that the distributions of recall times were both right-skew and suggested a greater spread of recall times for the unpleasant memories than for pleasant ones. Please do not mistake the data in Table 13 for paired data (in which each row would correspond to measurements on the same subject): the data come from independent groups of subjects which happen to have the same number of individuals in each.

Now carry out a two-sided nonparametric test of the null hypothesis that there is no difference in location between the distributions of recall times for pleasant and unpleasant memories.

Table 13 Memory recall times (seconds)

Pleasant memory	Unpleasant memory
1.07	1.45
1.17	1.67
1.22	1.90
1.42	2.02
1.63	2.32
1.98	2.35
2.12	2.43
2.32	2.47
2.56	2.57
2.70	3.33
2.93	3.87
2.97	4.33
3.03	5.35
3.15	5.72
3.22	6.48
3.42	6.90
4.63	8.68
4.70	9.47
5.55	10.00
6.17	10.93

(Source: Dunn, G. and Master, D. (1982) 'Latency models: the statistical analysis of response times', *Psychological Medicine*, vol. 12, pp. 659–65)

As is the Wilcoxon signed rank test, so the Mann–Whitney test is usually carried out by computer. To complete this subsection, you will learn how to use Minitab to carry out the Mann–Whitney test.

Refer to Chapter 11 of Computer Book B for the rest of the work in this subsection.

1.3 Nonparametric tests and parametric tests

So far in this section, you have seen that, by the simple ploy of replacing data values by ranks, it is possible to carry out tests of statistical hypotheses without making detailed distributional assumptions. Of course, some of the information in the data has been discarded by using the ranks rather than the original values: information on how far apart the data values are can no longer be used. But in most circumstances this loss is not very important. Even in the situation where the data really do come from normal distributions, the Wilcoxon signed rank test and the Mann–Whitney test are almost as powerful as their t-test analogues, and they can be applied in situations where t-tests cannot.

In view of this, it may occur to you to wonder why statisticians do not simply use such nonparametric tests routinely, rather than putting themselves at risk of making false assumptions of normality by using parametric methods like the t-test. There are several reasons:

- because there are no explicit parameters in use to describe the data, it can be more difficult to make quantitative statements about the actual difference(s) between the underlying populations
- in some cases, the nonparametric test does have some sort of distributional assumptions attached (for example, symmetry for the Wilcoxon signed rank test), so that we cannot forget about distributions entirely
- some departures from the usual assumptions behind tests – such as the assumption that the data are sampled randomly and are thus independent of one another – interfere with nonparametric tests just as much as with standard parametric tests
- for many of the more complicated statistical testing procedures, there is no straightforward nonparametric alternative to the parametric procedure
- and, as mentioned above, there is a small loss of power of the test if parametric assumptions are met.

Thus, while nonparametric tests can be extremely useful, they certainly do not completely solve all the awkward problems of hypothesis testing!

Exercises on Section 1

Exercise 1 *Decontamination using poplar trees*

Plants that take up high concentrations of pollutants can be used to help decontaminate polluted soils. In this activity, you will look at a small part of a study into pollutant uptake by poplar trees. Aluminium content (Al, in units of micrograms per gram, μg/g) was measured in August and again in November in poplar coppices (a coppice is the sampling unit) planted on an old household waste disposal site near Antwerp, Belgium. If the trees are helping to decontaminate the site, their aluminium content should have increased. The differences in Al content, that is, November Al content

A poplar plantation

minus August Al content, are given for each of 13 poplar coppices in Table 14.

Table 14 November Al content minus August Al content

3.1	6.3	−1.2	2.0	1.0	7.2	−5.6	−2.2	12.0	12.3	5.3	0.1	23.4

(Source: Laureysens, I. et al. (2004) 'Clonal variation in heavy metal accumulations and biomass production in a poplar coppice culture: I. Seasonal variation in leaf, wood and bark concentrations', *Environmental Pollution*, vol. 131, no. 3, pp. 485–94)

A normal probability plot of these differences is shown in Figure 8.

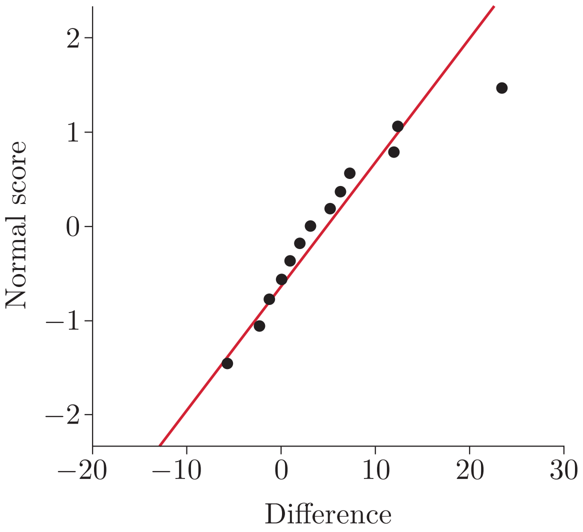


Figure 8 A normal probability plot of differences in aluminium content in poplar coppices

The points do not lie particularly close to a straight line, so the evidence is not compelling that a normal distribution is appropriate for modelling these data. You are therefore asked to analyse the data using the Wilcoxon signed rank test. The hypotheses are

$$H_0 : m = 0, \quad H_1 : m > 0,$$

where m is the (population) median Al difference.

(a) Calculate the Wilcoxon signed rank test statistic for the data.

(b) The p-value for the Wilcoxon signed rank test is 0.020. What do you conclude?

Exercise 2 *Byzantine coins – nonparametric tests*

In Activities 3 and 5, you considered data on the silver (Ag) content of coins from the reign of the Byzantine king Manuel I Comnenus. In particular, you tested whether the silver content of coins from the second of four coinages had a specified population median value. In this exercise, you are going to test whether coins from the first and second coinages differ in their silver content, on average. You will do so nonparametrically because the distribution of silver contents of coins from the second coinage is, arguably, non-normal. The data associated with the first two coinages are given, ordered within each sample, in Table 15.

Table 15 Silver content of coins: first and second coinages (% Ag)

First coinage	5.9	6.2	6.4	6.6	6.8	6.9	7.0	7.2	7.7
Second coinage	6.6	6.9	8.1	8.6	9.0	9.2	9.3		

Use the Mann–Whitney test to investigate whether it is plausible that the silver contents of both of these sets of coins come from the same distribution. Use a two-sided test, the alternative hypothesis being that the distributions of the silver contents in the coinages differ in location. Use the normal approximation to compute the p-value. State your conclusion.

2 A test for goodness-of-fit

The statistical tests that you have encountered so far are generally used to investigate scientific or other hypotheses of interest. This applies just as much to the tests you learned about in Unit 9 as to the nonparametric tests described in Section 1.

In this section, a different type of test is described, which you can apply to test the validity of a statistical model. In the module so far, a variety of graphical techniques have been described which you can use to explore the validity of the distributional assumptions you may have made. For example, you can investigate whether the shape of a bar chart or histogram of the data mirrors that of your chosen distribution. A little more formally, you can use a normal probability plot to look at whether or not a normal model is reasonable.

These graphical techniques, although extremely useful, are necessarily exploratory and generally involve a degree of subjective judgement. Hence, developing a test of the null hypothesis that the model 'fits' the data would be attractive.

In this section, a hypothesis test is introduced that can be used with discrete data to decide how reasonable it is to assume that a particular probability model generated the data. The test provides a test of the null hypothesis that the model 'fits' the data, allowing for random variation consistent with the model, and is thus referred to as a test of 'goodness-of-fit'. Note that this somewhat informal sounding phrase is the standard technical terminology for what we are about to do. The particular test that we explore can be adapted to test for goodness-of-fit of continuous distributions. However, that adaptation involves a degree of arbitrary choice and will not be covered here.

In Subsection 2.1, a test statistic for testing the goodness-of-fit of a discrete distribution is developed; the distribution of the test statistic is described in Subsection 2.2; and the test is further illustrated in Subsection 2.3. The test in question is called the *chi-squared goodness-of-fit test*. You might, in previous statistical studies, have come across a chi-squared test in the context of a topic called 'contingency tables'. If you have, you will recognise a number of the elements of the two chi-squared tests as being in common. The two chi-squared tests are not unrelated, but the chi-squared goodness-of-fit test developed here is not the same as the chi-squared test for contingency tables. Indeed, it can be argued that the latter is a special case of the former. The introduction of both chi-squared tests is credited to a paper in 1900 by the great early statistician Karl Pearson.

'chi' is pronounced 'kye'.

2.1 Goodness-of-fit of discrete distributions

The assessment of goodness-of-fit is based on quantifying the discrepancy between the data observed and the values that are expected under the model. The method is demonstrated through an example.

Example 8 *Testing the fit of a Poisson model*

Let us revisit some data that we previously considered in Example 13 of Unit 5. In order to elucidate the phenomenon of radioactivity, the scientists Rutherford and Geiger counted the number of alpha particles emitted from a radioactive source during 2612 different intervals of 7.5 seconds duration; the data are reproduced in Table 16 (this is essentially Table 7 of Unit 5). In 57 of these intervals there were zero emissions, in 203 there was a single emission, and so on. These observed frequencies, which we will now denote by O_i, constitute the data. In Unit 5, we made an informal check of Rutherford and Geiger's observation that a Poisson model seemed to provide a good fit to these data. But how can this claim be tested more formally?

Table 16 Emissions of alpha particles: counts and observed frequencies

Count i	Observed frequency O_i
0	57
1	203
2	383
3	525
4	532
5	408
6	273
7	139
8	49
9	27
10	10
11	4
12	2
≥ 13	0

The idea is to assume that the Poisson model is indeed correct – this phrase is convenient shorthand for 'the data were randomly sampled from a Poisson distribution' – and to calculate the corresponding expected frequencies E_i of intervals containing $0, 1, 2, \ldots$ emissions. If the model fits the data well, then the differences between the observed and expected frequencies, $O_i - E_i$, should be small in magnitude. This idea is precisely the same one that we employed informally in Example 13 of Unit 5. There, we just said that 'because the observed frequencies and the expected frequencies are close, a Poisson(3.877) distribution seems to fit the data very well'. (This was backed up by a graphical comparison of relative frequencies with the Poisson p.m.f.) Here, we wish to quantify what we mean by the frequencies being close or, equivalently, by the differences between observed and expected frequencies being small.

Before we get to that quantification, let us make sure we know what we mean by 'expected frequencies', and where they come from. The first step in obtaining the expected frequencies is to estimate the parameter of the

Poisson distribution by calculating the mean number of emissions in a 7.5-second interval. The sample mean turns out to be 3.877. To obtain the expected frequencies, we are therefore going to make use of the 'fitted' Poisson(3.877) distribution. The idea is that if the Poisson distribution is correct, the fitted Poisson distribution is the appropriate version of the Poisson distribution to compare with the data.

Recall from Subsection 4.1 of Unit 7 that the sample mean is the maximum likelihood estimate of the parameter of the Poisson distribution.

If the Poisson model is correct, then the probability that i emissions occur in a 7.5-second interval is given by the corresponding Poisson probability:

$$P(X = i) = \frac{e^{-3.877} 3.877^i}{i!}.$$

A total of 2612 intervals were observed during the experiment. If the Poisson model is correct, the expected number of those 2612 intervals in which i emissions occur is

$$E_i = 2612 \times P(X = i) = 2612 \times \frac{e^{-3.877} 3.877^i}{i!}.$$

For example, the expected frequency for one emission is

$$E_1 = 2612 \times \frac{e^{-3.877} 3.877^1}{1!} \simeq 209.75,$$

and hence the difference between the observed and expected frequencies for one emission is

$$O_1 - E_1 \simeq 203 - 209.75 = -6.75.$$

Yes, Geiger counters are named after Hans Geiger who performed this experiment with Ernest Rutherford

Activity 11 *Some observed and expected frequencies*

For the data and model in Example 8, calculate each of E_0, $O_0 - E_0$, E_2 and $O_2 - E_2$. (Give your answers correct to two decimal places.)

Example 9 *Testing the fit of a Poisson model, continued*

When calculations similar to those made at the end of Example 8 and in Activity 11 are made for every emission frequency in Example 8, the results are as shown in the third and fourth columns of Table 17 (overleaf). (The expected frequencies were also given, but to the nearest whole number, in Table 8 of Unit 5.)

The differences $O_i - E_i$ between the observed frequencies and the expected frequencies appear to be quite small in magnitude, at least in relation to the total sample size of 2612. However, to make this kind of statement precise, a test statistic is required with which to quantify the overall agreement, or lack of agreement, between the observed and expected frequencies.

Table 17 Emissions of alpha particles: observed and expected frequencies and differences between them, assuming a Poisson model

Count i	Observed frequency O_i	Expected frequency E_i	Difference $O_i - E_i$
0	57	54.10	2.90
1	203	209.75	−6.75
2	383	406.61	−23.61
3	525	525.47	−0.47
4	532	509.31	22.69
5	408	394.92	13.08
6	273	255.19	17.81
7	139	141.34	−2.34
8	49	68.50	−19.50
9	27	29.51	−2.51
10	10	11.44	−1.44
11	4	4.03	−0.03
12	2	1.30	0.70
≥ 13	0	0.53	−0.53

Since we are interested not in whether the differences are positive or negative, but only in their magnitudes, it makes sense to use squared differences in an overall assessment. However, a problem arises when this is done: even though most of the differences are small, when squared a few become inordinately large. For example, the difference of 2.90 in an expected frequency of 54.10 (count = 0) is virtually the same percentage difference as the difference of −23.61 in an expected count of 406.61 (count = 2): the former is $2.90 \times 100/54.10 \simeq 5.4\%$, the latter $23.61 \times 100/406.61 \simeq 5.8\%$. Yet when the differences are squared, one ($-23.61^2 \simeq 557.43$) is much larger than the other ($2.90^2 = 8.41$) – the squared difference associated with count = 2 is about $557.43/8.41 \simeq 66.3$ times the squared difference associated with count = 0 – misrepresenting the discrepancy between the data and the model.

The (standard) solution to the problem identified in Example 9 is to scale the squared differences by dividing by the expected frequency. Thus the scaled squared differences $(O_i - E_i)^2/E_i$ are used. These are then added up to give the overall measure of goodness-of-fit:

χ is the Greek lower-case letter chi.

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}.$$

This quantity is the **chi-squared goodness-of-fit** statistic.

Activity 12 *Some scaled squared differences between observed and expected frequencies*

(a) For the data and model in Examples 8 and 9, calculate the scaled squared difference terms $(O_i - E_i)^2/E_i$ for $i = 0, 1, 2$. (Give your answers correct to three decimal places.)

(b) How does the ratio of scaled squared differences for counts 2 and 0 compare with the ratio of unscaled squared differences for counts 2 and 0 mentioned at the end of Example 9?

Since each element in the statistic χ^2 is a squared term divided by an expected frequency, which must be positive, we must have $\chi^2 \geq 0$. Note that χ^2 is zero when $O_i = E_i$ for all categories, that is, when the observed and expected frequencies are equal. Clearly, due to random variation, we would not expect this to occur even when the model is correct. In order to assess what values of χ^2 are consistent or inconsistent with the model, the sampling distribution of the statistic χ^2 is required when the model is correct. In fact, the goodness-of-fit test is based on the *approximate* distribution of χ^2 under the null hypothesis that the model is correct. This distribution, known as the chi-squared distribution, is described in Subsection 2.2.

2.2 The chi-squared distribution

Define a random variable W as the sum of the squares of r independent random variables $Z_1, Z_2, \ldots, Z_r$, each following the standard normal distribution, $N(0,1)$:

$$W = Z_1^2 + Z_2^2 + \cdots + Z_r^2.$$

Then the distribution followed by W is called the **chi-squared distribution with r degrees of freedom**. This is written

$$W \sim \chi^2(r).$$

This defines a family of distributions which, like the family of t-distributions, is indexed by a parameter called the **degrees of freedom**.

Tai chi in a square. OK, so it's pronounced 'tie chee' not 'tie kye'!

Chi-squared random variables are sums of squared values so are defined on the set of positive values; that is, the range of the chi-squared distribution is $(0, \infty)$. The p.d.f.s of chi-squared distributions with 1, 2, 3 and 8 degrees of freedom are shown in Figure 9 (overleaf).

Notice that for smaller values of the degrees of freedom, the distribution is very right-skew. Indeed, for $r = 1$ and $r = 2$, the shape of the p.d.f. is quite different from what it is for larger values of r: for $r = 1$ and $r = 2$, the p.d.f. is a decreasing function for all $w > 0$, while for larger values of r it is more like normal and t p.d.f.s, being an increasing-then-decreasing function with a single maximum. Unlike normal and t p.d.f.s, however, a χ^2 p.d.f. is *not* symmetric.

For $r = 3, 4, \ldots$, the maximum of the $\chi^2(r)$ p.d.f. is actually at $w = r - 2$.

For larger values of r, however, the χ^2 distribution does become more symmetric. In fact, as the number of degrees of freedom increases, the distribution approaches a normal distribution. This should not be too surprising: for large values of r, W may be regarded as the sum of a large number of independent random variables, each with the same distribution, and hence by the Central Limit Theorem the distribution of W is approximately normal.

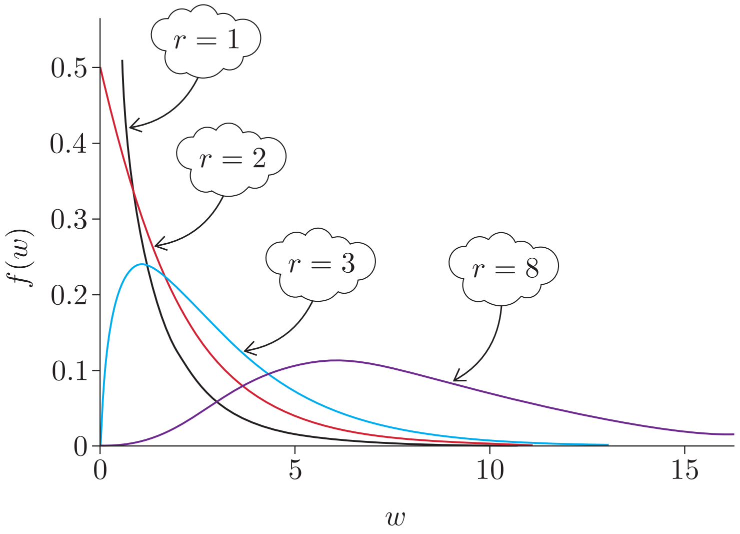


Figure 9 The p.d.f.s of four chi-squared distributions $\chi^2(r)$

Example 10 *The mean and variance of $\chi^2(1)$*

Suppose that W is a chi-squared random variable with one degree of freedom. The mean of W is quite easy to obtain. Since $W = Z^2$, where $Z \sim N(0,1)$,

$$\mu_W = E(W) = E\big(Z^2\big).$$

But $V(Z) = E\big(Z^2\big) - (E(Z))^2$, so

$$\mu_W = V(Z) + (E(Z))^2 = \sigma_Z^2 + \mu_Z^2.$$

But $Z \sim N(0,1)$, so $\mu_Z = 0$ and $\sigma_Z^2 = 1$. Therefore

$$\mu_W = 1 + 0^2 = 1.$$

It is not quite so easy to obtain the variance of W, and the details will not be given. In fact,

$$\sigma_W^2 = 2.$$

Activity 13 *The mean and variance of a chi-squared random variable*

The chi-squared random variable W with r degrees of freedom is defined to be

$$W = Z_1^2 + Z_2^2 + \cdots + Z_r^2,$$

where the Z_i, $i = 1, 2, \ldots, r$, are independent standard normal random variables. Show that the mean and variance of W are given by

$$\mu_W = r, \quad \sigma_W^2 = 2r.$$

The definition of a chi-squared distribution and the results of Activity 13 are summarised in the following box.

The chi-squared distribution

The continuous random variable W given by

$$W = Z_1^2 + Z_2^2 + \cdots + Z_r^2,$$

which is the sum of r independent squared standard normal random variables, is said to have a **chi-squared distribution with *r* degrees of freedom**. This is written

$$W \sim \chi^2(r).$$

The mean and variance of W are given by

$$\mu_W = r, \quad \sigma_W^2 = 2r.$$

As for the normal and t-distributions, straightforward formulas for the c.d.f. and quantiles of the distribution of the random variable $W \sim \chi^2(r)$ are not available in general; such quantities need to be computed numerically, or obtained from tables. Some quantiles of $\chi^2(r)$ for values of $r = 1, 2, \ldots, 10$ are shown in Table 18; the values of r are in the column headed 'df' (for degrees of freedom). This table is part of a much larger table of quantiles of the χ^2 distribution that is given in the Handbook.

Table 18 Selected quantiles of chi-squared distributions

df	0.01	0.05	0.90	0.95	0.99
1	0.0001	0.0039	2.71	3.84	6.63
2	0.020	0.103	4.61	5.99	9.21
3	0.115	0.352	6.25	7.81	11.34
4	0.297	0.711	7.78	9.49	13.28
5	0.554	1.14	9.24	11.07	15.09
6	0.872	1.64	10.64	12.59	16.81
7	1.24	2.17	12.02	14.07	18.48
8	1.65	2.73	13.36	15.51	20.09
9	2.09	3.33	14.68	16.92	21.67
10	2.56	3.94	15.99	18.31	23.21

Something to note is that Table 18 and the corresponding table in the Handbook include quantiles for small as well as large values of α, where $0 < \alpha < 1$ denotes the probability associated with the α-quantile. This is because the former cannot be deduced from the latter as they could in the case of the normal and t-distributions. This is a consequence of the lack of symmetry of the family of χ^2 distributions.

Example 11 *Using the table*

The 0.95-quantile of $\chi^2(5)$ is the number in the row labelled 5 (df $= 5$) and in the column headed 0.95, which is 11.07. Similarly, the 0.01-quantile of $\chi^2(7)$ is 1.24, and the 0.99-quantile of $\chi^2(2)$ is 9.21.

Activity 14 *Tail probabilities for chi-squared distributions*

Use the table of quantiles for chi-squared distributions in the Handbook to answer the following.

(a) Find the 0.01-quantile of W, where $W \sim \chi^2(18)$.

(b) Find the value w such that $P(W > w) = 0.05$, where $W \sim \chi^2(12)$.

(c) Find a good lower bound and a good upper bound on $P(W > 12.03)$, where $W \sim \chi^2(4)$.

More links: seaside golf courses are often called links

The χ^2 distribution actually has close links to two of the other continuous distributions that you have already met in this module. The closest link of all is explored in Activity 15.

Activity 15 *The $\chi^2(2)$ distribution*

The p.d.f. of the $\chi^2(r)$ distribution is of the form

$$g_r(w) = k_r w^{(r/2)-1} e^{-w/2}, \quad w > 0,$$

where k_r is the normalising constant. You need do nothing with this p.d.f. in general in this module, just use it in this activity to identify the $\chi^2(r)$ distribution in the case $r = 2$.

(a) When $r = 2$, it turns out that $k_r = k_2 = \frac{1}{2}$. Hence write down and identify the p.d.f. of the $\chi^2(2)$ distribution. Hint: if you don't recognise the p.d.f., take a look back at Unit 5.

(b) Expressions have been given for the mean and variance of the $\chi^2(r)$ distribution above. Check that when $r = 2$, they equate to the mean and variance of the distribution that you identified in part (a).

The other link is with the t-distribution. It turns out to be the case that if $Z \sim N(0, 1)$, $W \sim \chi^2(r)$ and Z and W are independent, then the random variable T can be defined such that

$$T = \frac{\sqrt{r}Z}{\sqrt{W}} \sim t(r).$$

The derivation of this result – although it surreptitiously underlies the introduction of the t-distribution in Unit 8 – is far beyond the scope of this module. It is worth mentioning, however, to show that use of the terminology 'degrees of freedom' is, at least, consistent across distributions. Although we have given them different symbols (r in the χ^2 case, ν in the t case), the two degrees of freedom parameters are actually one and the same.

2.3 The chi-squared goodness-of-fit test

In Subsection 2.1, the chi-squared goodness-of-fit test statistic was defined to be

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}.$$

This statistic can be calculated for any discrete model. If the model is correct, the distribution of the test statistic is approximately chi-squared, with degrees of freedom that depend on the number of categories and the assumed model. The chi-squared goodness-of-fit test is described in the following box.

The chi-squared goodness-of-fit test

Suppose that in a random sample of size n, each observation can be classified into one of k distinct classes or categories, and that the number of observations out of a total of n that fall into category i is denoted by O_i. Suppose that a model is set up, including p parameters whose values are estimated from the data, and that, according to the model, the probability of an observation falling into category i is θ_i. Then the expected number of observations falling into category i is denoted by E_i, and $E_i = n\theta_i$.

The chi-squared goodness-of-fit test statistic for the model is

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}.$$

If the model is correct, then for large n, the distribution of the test statistic is approximately chi-squared with $k - p - 1$ degrees of freedom:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \approx \chi^2(k - p - 1). \qquad (2)$$

Since the chi-squared test statistic measures the extent to which observed frequencies differ from those expected under the assumed model, the higher the value of χ^2, the greater the discrepancy between the data and the model. Thus the appropriate test is one-sided: only high values of χ^2 indicate that the model does not fit the data well.

Note that k is the number of categories in the data, and p is the number of model parameters that have been estimated.

As an aside, it is possible to argue that low values of χ^2 suggest a fit so good that the data are suspect, showing less variation than might be expected, and hence that a two-sided test is required. However, this is not the approach adopted in M248.

Distributional Result (2) is presented without proof. You should merely note that it is yet another consequence of the Central Limit Theorem.

Activity 16 *Obtaining the degrees of freedom*

Suppose that a discrete uniform distribution is proposed as a model for a large dataset taking values $0, 1, \ldots, m$. The fit of the model is to be checked using a chi-squared goodness-of-fit test. What would be the degrees of freedom associated with the approximate chi-squared null distribution of the test statistic in this case?

But how good is the chi-squared approximation to the null distribution of the chi-squared statistic when n is not large? Here is a simple rule to follow.

The validity of the chi-squared approximation

The chi-squared approximation to the null distribution of the chi-squared goodness-of-fit test statistic is adequate if no expected frequency E_i is less than 5. Otherwise, the approximation may not be adequate.

As categories are combined, both observed and expected frequencies simply add up.

Happily, if one or more expected frequencies are less than 5 so that the approximate chi-squared null distribution is not adequate, we can do something about it: combine categories together until all expected frequencies are 5 or more, remembering to reduce the number of categories, k, in a corresponding fashion. We often need to do this, as in the ongoing example of testing the Poisson model for the data on emissions of alpha particles.

Example 12 *Calculations for the data on emissions of alpha particles*

The calculations in Examples 8 and 9 for the data on emissions of alpha particles fit into the framework of the chi-squared goodness-of-fit test as follows. The total number of observations is 2612 (so $n = 2612$), and there are 14 categories, $0, 1, \ldots, 12$ and ≥ 13. Fitting the Poisson model requires the estimation of the sample mean $\widehat{\mu} = \overline{x} = 3.877$; hence the number of parameters estimated is 1 (so $p = 1$).

Note that, for convenience, i has been taken to range from 0 to 13 rather than from 1 to 14.

The expected frequencies were given in Table 17 ($E_i = 2612\,\theta_i$ where $\theta_i = P(X = i)$, $i = 0, 1, \ldots, 12$, and $\theta_{13} = P(X \geq 13)$). The values of E_i corresponding to counts of 11, 12 and ≥ 13 are all less than 5, thus violating the rule that each E_i must be at least 5 for the chi-squared approximation to be valid. This problem is overcome by combining categories until all values of E_i are greater than or equal to 5. First, combine categories 12 and ≥ 13. This has expected frequency $1.30 + 0.53 = 1.83 < 5$. This is still too small, so combine the already combined categories with category 11 also. Now, replacing categories 11, 12 and ≥ 13 with a single ≥ 11 category results in expected frequency $4.03 + 1.30 + 0.53 = 5.86$, which is greater than 5, so we can stop. The resulting frequencies and the corresponding values of $(O_i - E_i)^2/E_i$ are shown in Table 19.

Table 19 Calculating χ^2 for the data on emissions of alpha particles

i	O_i	E_i	$O_i - E_i$	$(O_i - E_i)^2/E_i$
0	57	54.10	2.90	0.155
1	203	209.75	−6.75	0.217
2	383	406.61	−23.61	1.371
3	525	525.47	−0.47	0.000
4	532	509.31	22.69	1.011
5	408	394.92	13.08	0.433
6	273	255.19	17.81	1.243
7	139	141.34	−2.34	0.039
8	49	68.50	−19.50	5.551
9	27	29.51	−2.51	0.213
10	10	11.44	−1.44	0.181
≥ 11	6	5.86	0.14	0.003

There are now 12 categories (so $k = 12$), and the value of the goodness-of-fit test statistic is

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} = 0.155 + 0.217 + \cdots + 0.003 = 10.417 \simeq 10.42.$$

Under the null hypothesis that the Poisson model is correct, the distribution of χ^2 is approximately a chi-squared distribution with $k - p - 1 = 12 - 1 - 1 = 10$ degrees of freedom.

Remember that the chi-squared test is one-sided; only high values of χ^2 give evidence against the fit of the model. Comparing the observed value of 10.42 with quantiles in the df = 10 row of the table of χ^2 quantiles in the Handbook, we see that it lies somewhere between the 0.5-quantile (which is 9.34) and the 0.9-quantile (which is 15.99). In fact, a computer calculation shows that $P(X \geq 10.42) \simeq 0.404$ when $X \sim \chi^2(10)$; that is, the p-value is 0.404. This is illustrated in Figure 10. There is therefore little or no evidence against the null hypothesis that the assumed Poisson model is correct. On the basis of this, we can be quite confident in using the Poisson distribution as a good model for these data.

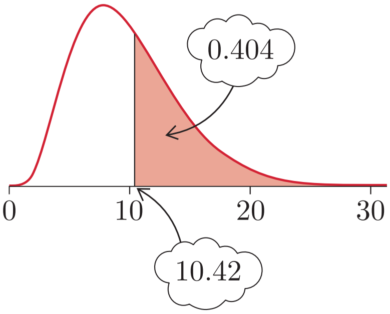


Figure 10 The null distribution and the p-value for the data on emissions of alpha particles, assuming a Poisson model

To finish this section, you can practise testing the goodness-of-fit of models that have been proposed previously in this module for certain datasets. Perhaps surprisingly, Minitab does not offer a facility for performing chi-squared goodness-of-fit tests, so you will perform such tests only ‘by hand’.

Checking the goodness-of-fit

Activity 17 *Testing the fit of a geometric model*

In Subsection 3.2 of Unit 3, a large genealogical study was reported in which the stage at which the first daughter was born in each of $n = 7745$ families where there was at least one daughter was recorded. The data are given in Table 20 (overleaf).

Table 20 Position of first girl in family

Position of girl	1	2	3	4	≥ 5
Number of families	3684	1964	1011	549	537

In Unit 3, the geometric model with parameter $p = 17/35$ was suggested for these data. Notice that the value of the parameter is given in this case (by Nicolaus Bernoulli) and is *not* a parameter value estimated from the data. In Unit 3, you also considered the fit of this model informally, and suggested that the fit of the model was reasonable. As part of that consideration, the probabilities under the presumed geometric model were obtained. They are given below in fractional (rather than approximate decimal) form to help retain accuracy in computations. If $X \sim G\left(\frac{17}{35}\right)$, then

$$P(X=1) = \frac{17}{35}, \quad P(X=2) = \frac{18}{35} \times \frac{17}{35}, \quad P(X=3) = \left(\frac{18}{35}\right)^2 \times \frac{17}{35},$$
$$P(X=4) = \left(\frac{18}{35}\right)^3 \times \frac{17}{35}, \quad P(X \geq 5) = P(X > 4) = \left(\frac{18}{35}\right)^4.$$

(a) Produce a table like Table 19 giving the positions of first girls i, observed frequencies O_i, expected frequencies E_i, differences $O_i - E_i$, and components of the χ^2 goodness-of-fit test statistic $(O_i - E_i)^2/E_i$ for this model and data.

(b) Hence obtain the value of the χ^2 test statistic and the number of degrees of freedom of the approximate χ^2 null distribution. Between which quantiles of the appropriate χ^2 distribution does the value of χ^2 lie?

(c) What conclusion can you come to?

Activity 18 *Leaves of Indian creeper plants*

The leaves of the Indian creeper plant *Pharbitis nil* can be variegated or unvariegated and, at the same time, faded or unfaded. In an experiment that you first considered in Unit 7, plants were crossed. Of 290 offspring plants observed, the four types of leaf, 0 for unvariegated and unfaded, v for variegated and unfaded, f for unvariegated and faded, and vf for variegated and faded, occurred with frequencies 187, 37, 35 and 31, respectively.

(a) According to one genetic theory, the four types should have occurred with probabilities $\frac{9}{16}$, $\frac{3}{16}$, $\frac{3}{16}$ and $\frac{1}{16}$, respectively. Use a chi-squared test of goodness-of-fit to show that the data offer strong evidence against this theory.

(b) A more sophisticated theory, which was the focus of your work with this dataset in Unit 7, allows for so-called 'genetic linkage'. The corresponding model includes a parameter, θ, that you estimated to take a value a bit above 0.05. Using the exact MLE, $\widehat{\theta}$, the hypothesised proportions turn out to be 0.6209, 0.1291, 0.1291, 0.1209, respectively. Use a chi-squared goodness-of-fit test to investigate the validity of this theory.

Exercises on Section 2

Exercise 3 *Testing the fit of another Poisson model*

In Units 2 and 3, data were considered on the numbers of yeast cells in each of $n = 400$ squares on a microscope slide. The data are given once again in Table 21.

Table 21 Yeast cells on a microscope slide

Cells in a square	0	1	2	3	4	5	≥ 6
Frequency	213	128	37	18	3	1	0

In Unit 3, the Poisson model was suggested for these data. The sample mean of the data is $\overline{x} = 0.6825$; this is used as an estimate of the Poisson parameter. You already worked out the probabilities associated with each number of cells in a square under the Poisson(0.6825) model in Unit 3; they are given in Table 22.

Table 22 Yeast cells on a microscope slide

Cells in a square	0	1	2	3	4	5	≥ 6
Probability	0.505	0.345	0.118	0.027	0.005	0.001	0

Informal comparisons in Section 4 of Unit 3 suggested that the fit of this Poisson model to these data seemed reasonable. Now make a formal test of the goodness-of-fit of the Poisson model to these data, and report your conclusion.

Exercise 4 *Testing the fit of another geometric model*

The final exercise of Unit 7 concerned the pattern of healthy and diseased trees in a plantation of Douglas firs. In particular, the data, repeated in Table 23, include the lengths of unbroken runs of diseased trees up to and including the first healthy tree. Observations were made on a total of $n = 109$ runs of trees.

Table 23 Run lengths of diseased trees

Run length	1	2	3	4	5	6	≥ 7
Number of runs	71	28	5	2	2	1	0

In the original work by E.C. Pielou and in Unit 7, it was proposed that the geometric distribution might be a good model for these data. Also, using these data, the geometric parameter p was estimated by maximum likelihood to be 0.657. (Here, p is the proportion of healthy trees in the plantation.)

Investigate the goodness-of-fit of the geometric model, and report your conclusion.

Summary

In this unit, you have learned about certain nonparametric tests, and about a method for testing the goodness-of-fit of a probability model for discrete data.

A nonparametric or distribution-free hypothesis test is a statistical testing procedure that does not involve making specific assumptions about the form of the distribution of the population(s) involved. This means, for example, that such procedures can be used in place of t-tests when the populations involved cannot be assumed to have a normal distribution. You have met a test whose null hypothesis is that a single sample of data, which might actually arise as a set of differences between pairs of data, comes from a population whose median has a specified value. The Wilcoxon signed rank test generally has reasonable power, but it involves the assumption that the distribution from which the data were drawn is symmetric. You have also met the Mann–Whitney test, which is used to compare the locations of the distributions of the populations from which two independent samples of data were drawn.

A family of distributions, called chi-squared distributions, has been introduced. These are indexed by a parameter r, known as the degrees of freedom.

The chi-squared goodness-of-fit test for discrete probability models involves calculating expected frequencies for each possible value of the random variable involved, and producing a summary measure of how these differ from the frequencies that were actually observed. Under the null hypothesis that the proposed model fits the data, the distribution of this summary measure is approximately a chi-squared distribution. This approximate result for the null distribution of the chi-squared goodness-of-fit test statistic is valid only if the expected frequencies are not too small: if any of the expected frequencies are less than 5, then some of the categories should be combined before the value of the test statistic is calculated.

Learning outcomes

After you have worked through this unit, you should be able to:

- be aware that there are statistical tests that do not involve making specific distributional assumptions about the data
- be aware that nonparametric tests may still involve certain broad distributional assumptions, such as an assumption of symmetry
- perform the Wilcoxon signed rank test for testing the location of a distribution, both by hand and using Minitab
- perform the Mann–Whitney test for testing whether two populations are equal, both by hand and using Minitab
- be aware that the validity of a particular model for a given set of data can be tested using a goodness-of-fit test
- perform a chi-squared goodness-of-fit test for discrete data
- understand some basic properties of the family of chi-squared distributions
- use the table of quantiles for chi-squared distributions.

Solutions to activities

Solution to Activity 1

(a) The points do not lie particularly close to a straight line, so the evidence is not compelling that a normal distribution is appropriate for modelling these data. However, it must be borne in mind that the sample size is small, and therefore that the evidence *against* a normal distribution is not compelling either.

(b) (i) The null and alternative hypotheses are

$$H_0: \mu_D = 0, \quad H_1: \mu_D \neq 0.$$

(ii) A p-value of 0.327 implies little or no evidence against the null hypothesis of a zero mean difference.

(iii) The t-test is based on assuming normality of the differences. But the conclusion of the t-test may be in doubt because of the possible non-normality of the underlying distribution.

Solution to Activity 2

(a) **Table 24** Ranks of corneal thickness of normal eyes

Patient	1	2	3	4	5	6	7	8
Normal eye	484	478	492	444	436	398	464	476
Rank	7	6	8	3	2	1	4	5

(b) **Table 25** Ranks of rounded corneal thickness of normal eyes

Patient	1	2	3	4	5	6	7	8
Normal eye	480	480	490	440	440	400	460	480
Rank	6	6	8	$2\frac{1}{2}$	$2\frac{1}{2}$	1	4	6

Note that the ranks of patients 1, 2 and 8, who all have rounded value 480, are 6 because it is the average of the ranks 5, 6 and 7.

Solution to Activity 3

(a) The hypotheses are

$$H_0: m = 7.5, \quad H_1: m \neq 7.5,$$

where m is the median % Ag content of the coins from the second coinage.

(b) The data minus $m_0 = 7.5$ are given in Table 26.

Table 26 Silver content minus 7.5 (% Ag)

−0.6	1.5	−0.9	0.6	1.8	1.7	1.1

The hypotheses are now

$$H_0: m_{7.5} = 0, \quad H_1: m_{7.5} \neq 0,$$

where $m_{7.5}$ is the median % Ag content of the coins from the second coinage minus 7.5.

Solution to Activity 4

(a) The Wilcoxon signed rank test statistic can be calculated using Table 27.

Table 27

Patient	1	2	3	4	5	6	7	8	9	10
Difference	7	-3	3	-5	3	-1	-3	-2	1	8
Sign	$+$	$-$	$+$	$-$	$+$	$-$	$-$	$-$	$+$	$+$
Absolute difference	7	3	3	5	3	1	3	2	1	8
Rank	9	$5\frac{1}{2}$	$5\frac{1}{2}$	8	$5\frac{1}{2}$	$1\frac{1}{2}$	$5\frac{1}{2}$	3	$1\frac{1}{2}$	10

There are no 0s but there are a number of tied observations.

The test statistic w_+ is the sum of the ranks associated with the positive differences, so

$$w_+ = 9 + 5\tfrac{1}{2} + 5\tfrac{1}{2} + 1\tfrac{1}{2} + 10 = 31\tfrac{1}{2}.$$

(b) The p-value is much greater than 0.1, indicating little or no evidence against the null hypothesis. We conclude that these data give no cause to fear that chorionic villus sampling has any effect on foetal movement.

Solution to Activity 5

(a) The Wilcoxon signed rank test statistic can be calculated using Table 28.

Table 28

Difference	-0.6	1.5	-0.9	0.6	1.8	1.7	1.1
Sign	$-$	$+$	$-$	$+$	$+$	$+$	$+$
Absolute difference	0.6	1.5	0.9	0.6	1.8	1.7	1.1
Rank	$1\frac{1}{2}$	5	3	$1\frac{1}{2}$	7	6	4

There are no 0s but there are two tied observations.

The test statistic w_+ is the sum of the ranks associated with the positive differences, so

$$w_+ = 5 + 1\tfrac{1}{2} + 7 + 6 + 4 = 23\tfrac{1}{2}.$$

(b) The p-value is greater than 0.1, indicating little or no evidence against the null hypothesis. We conclude that these data give little or no evidence against the notion that the median silver content in the second coinage of Manuel I Comnenus is 7.5%.

Solution to Activity 6

(a) Table 29 shows the results of subtracting 0.618 from each entry in Table 9 and then allocating ranks to the absolute values of the resulting differences.

Table 29

Value	Difference	Sign	Absolute difference	Rank
0.693	0.075	+	0.075	17
0.662	0.044	+	0.044	10
0.690	0.072	+	0.072	16
0.606	−0.012	−	0.012	$5\frac{1}{2}$
0.570	−0.048	−	0.048	11
0.749	0.131	+	0.131	18
0.672	0.054	+	0.054	14
0.628	0.010	+	0.010	4
0.609	−0.009	−	0.009	3
0.844	0.226	+	0.226	19
0.654	0.036	+	0.036	8
0.615	−0.003	−	0.003	1
0.668	0.050	+	0.050	12
0.601	−0.017	−	0.017	7
0.576	−0.042	−	0.042	9
0.670	0.052	+	0.052	13
0.606	−0.012	−	0.012	$5\frac{1}{2}$
0.611	−0.007	−	0.007	2
0.553	−0.065	−	0.065	15
0.933	0.315	+	0.315	20

There are no 0s and there are only two tied differences.

The value of the test statistic w_+ is the sum of the ranks associated with positive differences. Thus

$$w_+ = 17 + 10 + 16 + 18 + 14 + 4 + 19 + 8 + 12 + 13 + 20 = 151.$$

(b) The p-value is between 0.05 and 0.1, indicating only weak evidence against the null hypothesis that the median width-to-length ratio is 0.618. Looking at the data, it would seem that the median width-to-length ratio *may* be greater than 0.618.

Solution to Activity 7

The sample size is 20 and there are no zero differences, so $n = 20$. Therefore

$$E(W_+) = \frac{n(n+1)}{4} = \frac{20 \times 21}{4} = 105,$$

$$V(W_+) = \frac{n(n+1)(2n+1)}{24} = \frac{20 \times 21 \times 41}{24} = 717.5.$$

The observed value of the test statistic is $w_+ = 151$, so

$$z = \frac{w_+ - 105}{\sqrt{717.5}} = \frac{151 - 105}{\sqrt{717.5}} \simeq 1.72.$$

The table of probabilities for the standard normal distribution in the Handbook gives

$$P(Z \geq 1.72) = 1 - \Phi(1.72) = 1 - 0.9573 \simeq 0.043.$$

So there is a probability of 0.043 of being at least this far out into the (right-hand) tail of the standard normal distribution. Since you are performing a two-sided test, you need to consider the other tail as well. Thus the approximate p-value is $2 \times 0.043 = 0.086$. This is very close to the value given by the exact test. The p-value of 0.086 provides weak evidence that Shoshoni rectangles do not conform to the Greek golden ratio standard (and the data suggest that the median width-to-length ratio may be greater than 0.618).

Solution to Activity 8

The table of pooled and ranked data is as follows.

Table 30

(A) California	Rank	(B) Alaska	Rank
23	1	39	5
26	2	48	9
30	3	53.5	12
33	4	55	13
42	6	57	14
45	$7\frac{1}{2}$	66	15
45	$7\frac{1}{2}$	77	16
50	10	79	17
50.5	11	108	19
96	18	121	21
113	20	162	22
557	25	197	23
		309	24

The Mann–Whitney test statistic is

$$u_A = 1 + 2 + 3 + 4 + 6 + 7\tfrac{1}{2} + 7\tfrac{1}{2} + 10 + 11 + 18 + 20 + 25 = 115.$$

(To check,

$$u_B = 5 + 9 + 12 + \cdots + 23 + 24 = 210$$

and

$$u_A + u_B = \tfrac{1}{2}(n_A + n_B)(n_A + n_B + 1) = 325.)$$

Solution to Activity 9

Under the null hypothesis that there is no difference in population location between the distributions underlying the two samples, we have

$$E(U_A) = \frac{n_A\,(n_A + n_B + 1)}{2} = \frac{12(12 + 13 + 1)}{2} = 156$$

and

$$V(U_A) = \frac{n_A n_B\,(n_A + n_B + 1)}{12} = \frac{12 \times 13 \times 26}{12} = 338.$$

For the observed value $u_A = 115$, the corresponding z value is

$$z = \frac{u_A - 156}{\sqrt{338}} = \frac{115 - 156}{\sqrt{338}} \simeq -2.23.$$

So the approximate p-value based on the normal approximation, against the two-sided alternative that the difference in population location is non-zero, is

$$\begin{aligned} p &= P(Z \leq -2.23) + P(Z \geq 2.23) = 2P(Z \geq 2.23) \\ &= 2(1 - \Phi(2.23)) = 2(1 - 0.9871) \simeq 0.026. \end{aligned}$$

The (approximate) p-value is such that $0.01 < p \leq 0.05$, so there is moderate evidence that the distribution of village group sizes differed in (statistical) location between California and Alaska. Looking at the data in Tables 12 or 30, there is moderate evidence that village groups were larger in Alaska than they were in California, on average.

Solution to Activity 10

The appropriate test is the Mann–Whitney test. The ranks are given in Table 31.

Table 31

Pleasant memory	Rank	Unpleasant memory	Rank
1.07	1	1.45	5
1.17	2	1.67	7
1.22	3	1.90	8
1.42	4	2.02	10
1.63	6	2.32	$12\frac{1}{2}$
1.98	9	2.35	14
2.12	11	2.43	15
2.32	$12\frac{1}{2}$	2.47	16
2.56	17	2.57	18
2.70	19	3.33	25
2.93	20	3.87	27
2.97	21	4.33	28
3.03	22	5.35	31
3.15	23	5.72	33
3.22	24	6.48	35
3.42	26	6.90	36
4.63	29	8.68	37
4.70	30	9.47	38
5.55	32	10.00	39
6.17	34	10.93	40
Sums of ranks	$345\frac{1}{2}$		$474\frac{1}{2}$

If the group of pleasant memory recall times is labelled A, then the test statistic u_A is $345\frac{1}{2}$. There are only two tied values, and the samples are not particularly small, so the normal approximation should be adequate for calculating the p-value:

$$E(U_A) = \frac{n_A\,(n_A + n_B + 1)}{2} = \frac{20 \times 41}{2} = 410,$$

$$V(U_A) = \frac{n_A n_B\,(n_A + n_B + 1)}{12} = \frac{20 \times 20 \times 41}{12} \simeq 1366.667.$$

The z value is

$$z = \frac{345.5 - 410}{\sqrt{1366.667}} \simeq -1.74.$$

So, since the test is two-sided, the approximate p-value is

$$\begin{aligned} p &= P(Z \leq -1.74) + P(Z \geq 1.74) = 2P(Z \geq 1.74) \\ &= 2(1 - \Phi(1.74)) = 2(1 - 0.9591) \simeq 0.082. \end{aligned}$$

This p-value is between 0.05 and 0.1, so there is only weak evidence against the null hypothesis that the distribution of recall times has the same location for pleasant and unpleasant memories. Looking at the data, it would seem that there is a suggestion that pleasant memories might have shorter recall times, on average.

Had you chosen to label the group of unpleasant memory recall times by A (and both sample sizes are the same in this example, so either sample could just as reasonably have been chosen), the test statistic u_A would be $474\frac{1}{2}$. The population mean and variance of U_A are the same as above, and

$$z = \frac{474.5 - 410}{\sqrt{1366.667}} \simeq 1.74.$$

The approximate p-value is now

$$\begin{aligned} p &= P(Z \geq 1.74) + P(Z \leq -1.74) \\ &= 2P(Z \geq 1.74) = 2(1 - 0.9591) \simeq 0.082. \end{aligned}$$

The p-value is the same as above, as are the conclusions that follow from it. (This should not be surprising since it was said in the box defining the Mann–Whitney test near the start of this subsection that it does not matter which sample is which.)

Solution to Activity 11

The expected frequency for no emissions is

$$E_0 = 2612 \times e^{-3.877} \simeq 54.10.$$

From Table 16, the observed frequency for no emissions is $O_0 = 57$, so the difference between the observed and expected frequencies for no emissions is

$$O_0 - E_0 \simeq 57 - 54.10 = 2.90.$$

The expected frequency for two emissions is

$$E_2 = 2612 \times \frac{e^{-3.877} 3.877^2}{2!} \simeq 406.61.$$

From Table 16, the observed frequency for two emissions is $O_2 = 383$, so the difference between the observed and expected frequencies for two emissions is

$$O_2 - E_2 \simeq 383 - 406.61 = -23.61.$$

Solution to Activity 12

(a) Taking values from Table 17, we have: $O_0 - E_0 \simeq 2.90$, $E_0 \simeq 54.10$ and

$$(O_0 - E_0)^2/E_0 \simeq 2.90^2/54.10 \simeq 0.155;$$

$O_1 - E_1 \simeq -6.75$, $E_1 \simeq 209.75$ and

$$(O_1 - E_1)^2/E_1 \simeq (-6.75)^2/209.75 \simeq 0.217;$$

$O_2 - E_2 \simeq -23.61$, $E_2 \simeq 406.61$ and

$$(O_2 - E_2)^2/E_2 \simeq (-23.61)^2/406.61 \simeq 1.371.$$

(b) The scaled squared difference corresponding to count $= 2$ is $1.371/0.155 \simeq 8.8$ times the scaled squared difference corresponding to count $= 0$. This is still a considerable discrepancy, showing that the model does not fit the observed frequency O_2 as well as it does O_0.

However, it is not so great as the factor of 66.3 that is the ratio of the unscaled squared differences mentioned at the end of Example 9.

Solution to Activity 13

Since W is the sum of r independent observations $Z_1^2, Z_2^2, \dots, Z_r^2$, we can use the following results introduced in Unit 4:

$$E(W) = E\left(Z_1^2\right) + E\left(Z_2^2\right) + \cdots + E\left(Z_r^2\right),$$
$$V(W) = V\left(Z_1^2\right) + V\left(Z_2^2\right) + \cdots + V\left(Z_r^2\right).$$

So, since each Z_i^2 has mean 1,

$$E(W) = \underbrace{1 + 1 + \cdots + 1}_{r \text{ terms}} = r;$$

and since each Z_i^2 has variance 2,

$$V(W) = \underbrace{2 + 2 + \cdots + 2}_{r \text{ terms}} = 2r.$$

Solution to Activity 14

The following values were obtained using the table of quantiles of chi-squared distributions in the Handbook.

(a) For $r = 18$, $q_{0.01} = 7.01$.

(b) The 0.95-quantile of $\chi^2(12)$ is required, so $w = 21.03$.

(c) The value 12.03 lies between the 0.975-quantile and the 0.99-quantile of $\chi^2(4)$. (These quantiles are 11.14 and 13.28, respectively.) Thus

$$P(W > 12.03) < 0.025$$

and

$$P(W > 12.03) > 0.01.$$

That is,

$$0.01 < P(W > 12.03) < 0.025.$$

(In fact, $P(W > 12.03) = 0.0171$.)

Solution to Activity 15

(a) When $r = 2$, the p.d.f. is

$$g_2(w) = k_2 w^{(2/2)-1} e^{-w/2} = \tfrac{1}{2} e^{-\frac{1}{2}w}.$$

From Subsection 2.2 of Unit 5, this is the p.d.f. of the exponential distribution with parameter $\lambda = \frac{1}{2}$ (i.e. $M(\frac{1}{2})$).

(b) The $\chi^2(r)$ distribution has mean r and variance $2r$, so when $r = 2$, the mean is 2 and the variance is 4. From Subsection 2.2 of Unit 5, the $M(\lambda)$ distribution has mean $1/\lambda$ and variance $1/\lambda^2$, so when $\lambda = 1/2$, the mean is indeed $1/(1/2) = 2$ and the variance is indeed $1/(1/2)^2 = 4$.

Solution to Activity 16

There are $k = m+1$ 'classes' or 'categories', namely the values $0, 1, \ldots, m$. No parameters are estimated when fitting the discrete uniform distribution to data (from Subsection 5.1 of Unit 3, the discrete uniform distribution has p.m.f.

$$p(x) = \frac{1}{m+1}, \quad x = 0, 1, \ldots, m,$$

whatever the data may be); so $p = 0$. Therefore the appropriate degrees of freedom in this situation are $k - p - 1 = (m+1) - 0 - 1 = m$.

Solution to Activity 17

(a) To obtain the values of E_i, multiply the corresponding probabilities given in the question by 7745. The required values are shown in Table 32. (Note that no categories are combined because no E_i is less than 5.)

Table 32

Position i	O_i	E_i	$O_i - E_i$	$(O_i - E_i)^2/E_i$
1	3684	3761.86	−77.86	1.611
2	1964	1934.67	29.33	0.445
3	1011	994.97	16.03	0.258
4	549	511.70	37.30	2.719
≥ 5	537	541.80	−4.80	0.043

(b) The value of the chi-squared goodness-of-fit test statistic is

$$\begin{aligned}\chi^2 &= \sum \frac{(O_i - E_i)^2}{E_i} \\ &= 1.611 + 0.445 + 0.258 + 2.719 + 0.043 = 5.076 \simeq 5.08.\end{aligned}$$

There are 5 categories, so $k = 5$. No parameter has been estimated, since the hypothesised geometric model was fully specified, including the value of the geometric parameter, 17/35. Thus p, the number of estimated parameters, is 0. The null distribution of χ^2 is therefore approximately chi-squared with $k - p - 1 = 5 - 0 - 1 = 4$ degrees of freedom. For the $\chi^2(4)$ distribution, the value of $\chi^2 = 5.08$ lies between the 0.5-quantile (which is 3.36) and the 0.9-quantile (which is 7.78).

(c) The p-value is greater than 0.1 (since 5.08 is smaller than the 0.9-quantile of $\chi^2(4)$). There is therefore little or no evidence against the hypothesis that the observations are from the geometric distribution $G(17/35)$. (The exact p-value happens to be 0.279.)

Solution to Activity 18

(a) The categories have expected frequencies given by

$$E_i = n\theta_i = 290\,\theta_i, \quad i = 0, v, f, vf,$$

where

$$\theta_0 = \tfrac{9}{16}, \quad \theta_v = \tfrac{3}{16}, \quad \theta_f = \tfrac{3}{16}, \quad \theta_{vf} = \tfrac{1}{16}.$$

This leads to the values in Table 33.

Table 33

i	O_i	E_i	$O_i - E_i$	$(O_i - E_i)^2/E_i$
0	187	163.125	23.875	3.494
v	37	54.375	−17.375	5.552
f	35	54.375	−19.375	6.904
vf	31	18.125	12.875	9.146

The value of the chi-squared test statistic is

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$
$$= 3.494 + 5.552 + 6.904 + 9.146 = 25.096 \simeq 25.10.$$

There are four categories and no model parameters were estimated, so the null distribution of the test statistic has $4 - 0 - 1 = 3$ degrees of freedom. The value 25.10 is greater than the 0.995-quantile of $\chi^2(3)$, which is 12.84, so the p-value is less than 0.005. (The actual p-value is about 0.000015.) This is very small, so there is strong evidence against the null hypothesis that the model arising from this genetic theory is appropriate; it appears that the simple theory may be flawed.

(b) Allowing for genetic linkage, the expected frequencies are given by

$$E_i = n\theta_i = 290\,\theta_i, \quad i = 0, v, f, vf,$$

where

$$\theta_0 = 0.6209, \quad \theta_v = 0.1291, \quad \theta_f = 0.1291, \quad \theta_{vf} = 0.1209.$$

This leads to the values in Table 34.

Table 34

i	O_i	E_i	$O_i - E_i$	$(O_i - E_i)^2/E_i$
0	187	180.061	6.939	0.267
v	37	37.439	−0.439	0.005
f	35	37.439	−2.439	0.159
vf	31	35.061	−4.061	0.470

The value of the chi-squared test statistic is

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$
$$= 0.267 + 0.005 + 0.159 + 0.470 = 0.901 \simeq 0.90.$$

The number of categories is again 4. However, one parameter has been estimated, so the null distribution of the test statistic has $4 - 1 - 1 = 2$ degrees of freedom. The 0.1-quantile of $\chi^2(2)$ is 0.211, and the 0.5-quantile is 1.39. The observed value 0.91 lies between these quantiles, so the p-value is between 0.5 and 0.9. (The actual p-value is 0.638.) Hence there is little or no evidence against the theory. The model appears to fit the data well.

Solutions to exercises

Solution to Exercise 1

(a) The Wilcoxon signed rank test statistic can be calculated using Table 35.

Table 35

Coppice	1	2	3	4	5	6	7	8	9	10	11	12	13
Difference	3.1	6.3	−1.2	2.0	1.0	7.2	−5.6	−2.2	12.0	12.3	5.3	0.1	23.4
Sign	+	+	−	+	+	+	−	−	+	+	+	+	+
Absolute difference	3.1	6.3	1.2	2.0	1.0	7.2	5.6	2.2	12.0	12.3	5.3	0.1	23.4
Rank	6	9	3	4	2	10	8	5	11	12	7	1	13

The test statistic w_+ is the sum of the ranks associated with the positive differences, so

$$w_+ = 6 + 9 + 4 + 2 + 10 + 11 + 12 + 7 + 1 + 13 = 75.$$

Note that, in this case, it may be slightly quicker to work out the sum of the ranks for the negative differences ($w_- = 3 + 8 + 5 = 16$) and to use the fact that the sum of all the ranks is $\frac{1}{2}n(n+1) = \frac{1}{2} \times 13 \times 14 = 91$, to give

$$w_+ = 91 - 16 = 75.$$

(b) The p-value is between 0.01 and 0.05, indicating moderate evidence against the null hypothesis. We conclude that there is moderate evidence that the median difference in aluminium content is positive for poplar coppices planted on this kind of polluted ground.

Solution to Exercise 2

(a) The ranks are given in Table 36.

Table 36

First coinage	Rank	Second coinage	Rank
5.9	1	6.6	$4\frac{1}{2}$
6.2	2	6.9	$7\frac{1}{2}$
6.4	3	8.1	12
6.6	$4\frac{1}{2}$	8.6	13
6.8	6	9.0	14
6.9	$7\frac{1}{2}$	9.2	15
7.0	9	9.3	16
7.2	10		
7.7	11		

If the group of silver contents of coins from the second coinage is labelled A, then the test statistic u_A is

$$4\tfrac{1}{2} + 7\tfrac{1}{2} + 12 + 13 + 14 + 15 + 16 = 82.$$

With $n_A = 7$, $n_B = 9$,

$$\begin{aligned} E(U_A) &= \frac{n_A\,(n_A + n_B + 1)}{2} = \frac{7 \times 17}{2} = 59.5, \\ V(U_A) &= \frac{n_A n_B\,(n_A + n_B + 1)}{12} = \frac{7 \times 9 \times 17}{12} = 89.25. \end{aligned}$$

The z value is

$$z = \frac{82 - 59.5}{\sqrt{89.25}} \simeq 2.38.$$

So the approximate p-value for the two-sided test is

$$p = 2P(Z \geq 2.38) = 2(1 - \Phi(2.38)) \simeq 0.017.$$

Since $0.01 < p \leq 0.05$, there is moderate evidence against the null hypothesis. The silver contents appear to differ, and looking at the data, it would seem that the second coinage contains more silver, on average.

Solution to Exercise 3

To obtain the values of E_i, multiply the probabilities given in Table 22 by 400; these are shown in Table 37.

Table 37

Cells in a square	O_i	E_i
0	213	202.0
1	128	138.0
2	37	47.2
3	18	10.8
4	3	2.0
5	1	0.4
≥ 6	0	0.0

To ensure that all expected values are at least 5, squares with 3 or more yeast cells on them are combined. The further calculations are set out in Table 38.

Table 38

Cells in a square	O_i	E_i	$O_i - E_i$	$(O_i - E_i)^2/E_i$
0	213	202.0	11.0	0.599
1	128	138.0	−10.0	0.725
2	37	47.2	−10.2	2.204
≥ 3	22	13.2	8.8	5.867

The value of the chi-squared goodness-of-fit test statistic is

$$\chi^2 = 0.599 + 0.725 + 2.204 + 5.867 = 9.395 \simeq 9.40.$$

After combination, there are four categories, so $k = 4$, and one parameter, the Poisson mean, has been estimated, so $p = 1$. The null distribution of χ^2 is therefore approximately chi-squared with $k - p - 1 = 4 - 1 - 1 = 2$ degrees of freedom. For the $\chi^2(2)$ distribution, the value of $\chi^2 = 9.40$ lies between the 0.99-quantile (which is 9.21) and the 0.995-quantile (which is 10.60). The p-value therefore lies below 0.01 (since 9.40 is greater than the 0.99-quantile of $\chi^2(2)$). There is therefore strong evidence against the hypothesis that the observations are from the Poisson distribution. It seems that our earlier modelling of these data by the Poisson distribution was somewhat over-confident and inappropriate.

Solution to Exercise 4

For the geometric model, the probability of a run of length i ($i = 1, 2, \ldots, 6$) is

$$\theta_i = (1 - p)^{i-1} p.$$

That no run was greater than 6 must be accounted for by

$$\theta_7 = P(X \geq 7) = P(X > 6) = (1 - p)^6.$$

The geometric parameter p has been estimated from the data to be 0.657. The expected frequencies under the geometric model are $109\,\theta_i$; these are shown in Table 39.

Table 39

Run length	O_i	E_i
1	71	71.61
2	28	24.56
3	5	8.43
4	2	2.89
5	2	0.99
6	1	0.34
≥ 7	0	0.18

To ensure that all expected values are at least 5, runs of length 3 or more are combined. The calculations are set out in Table 40.

Table 40

Run length	O_i	E_i	$O_i - E_i$	$(O_i - E_i)^2/E_i$
1	71	71.61	-0.61	0.005
2	28	24.56	3.44	0.482
≥ 3	10	12.83	-2.83	0.624

The value of the chi-squared test statistic is

$$\chi^2 = 0.005 + 0.482 + 0.624 = 1.111 \simeq 1.11.$$

There are three categories and one parameter has been estimated from the data (the geometric parameter, $\widehat{p} = 0.657$), so the null distribution of the test statistic has $3 - 1 - 1 = 1$ degree of freedom. The observed value 1.11 lies between the 0.5-quantile and the 0.9-quantile of $\chi^2(1)$ (which are 0.455 and 2.71, respectively), so the p-value is greater than 0.1. (The actual p-value turns out to be 0.292.) There is therefore little or no evidence against the geometric model. This confirms that Pielou's assumptions were reasonable.

Acknowledgements

Grateful acknowledgement is made to the following sources:

Page 276: Taken from: http://www.stat.cmu.edu/ larry/all-of-nonpar/

Page 277: © Peter Hermes Furian/Dreamstime.com

Page 278: The National Library of Medicine

Page 279: Leonard Zhukovsky/www.123rf.com

Page 281: Taken from:
http://www.swlearning.com/quant/kohler/stat/biographical_sketches/bio21.1.html

Page 285: www.vcoins.com. This file is licensed under the Creative Commons Attribution-Share Alike Licence
http://creativecommons.org/licenses/by-sa/3.0/

Page 288: Sergio Bertino

Page 290: National Institutes of Health

Page 292: Taken from:
http://www.lifelesshurried.com/2014_03_01_archive.html

Page 294: Taken from:
https://math.osu.edu/about-us/history/henry-berthold-mann

Page 295: Taken from:
http://thebrain.mcgill.ca/flash/d/d_03/d_03_m/d_03_m_que/d_03_m_que.html

Page 297: Bob and Ira Spring / https://www.britannica.com/topic/Tlingit

Page 301: © 4photos.net

Page 305: Bullet308. This file is licensed under the Creative Commons Attribution-Share Alike 3.0 Unported license

Page 307: Craig Nagy. This file is licensed under the Creative Commons Attribution-Share Alike Licence
http://creativecommons.org/licenses/by-sa/3.0/

Page 310: Oleksandr Sokolov / www.123rf.com

Page 313: © kadmy/www.istockphoto.com

Every effort has been made to contact copyright holders. If any have been inadvertently overlooked, the publishers will be pleased to make the necessary arrangements at the first opportunity.