

4 HYPOTHESIS TESTING: TWO SAMPLE TESTS

Objectives

After studying this chapter you should

- appreciate the need for two sample tests;
- be able to carry out a test for the equality of two normal population variances;
- understand when, and be able, to carry out normal and t -tests for the equality of two normal population means using information from two independent samples;
- understand when, and be able, to carry out a t -test for the equality of two population means using information from paired samples;
- be able to carry out a sign test for a paired samples design;
- be able to analyse the results of a paired samples design using the Wilcoxon signed-rank test.

4.0 Introduction

In the previous chapter, tests were described for a parameter of a single population. In this chapter, some of these tests will be developed, and new ones introduced, to test for the equality of a parameter in two populations.

It is a well-known and frequently-stated fact that, on average, men are significantly taller than women. Less well-known is the fact that the variability in men's heights is greater than that in women's heights. However, is this latter difference significant?

In this chapter you will see how similar facts are justified, and similar questions are answered.

Activity 1 Hand span 1

Using a 30 cm ruler, measure and record the span (thumb to fourth finger) of the dominant hand of each of a random sample of between 10 and 15 males. Repeat for a random sample of females; the two sample sizes need not be equal.

- (a) Calculate appropriate summary statistics for each of your two samples.
 - (b) Comment upon any differences or similarities between males and females as regards dominant hand span.
 - (c) Name a probability distribution that may be appropriate for hand span.
 - (d) State, in words, two possible hypotheses regarding the difference, if any, between male and female dominant hand spans.
-

Activity 2 Hand span 2

Using a 30 cm ruler, measure and record the span (thumb to fourth finger) of both the dominant and non-dominant hands of each of a random sample of between 10 and 15 people.

- (a) For each person, calculate the difference (dominant hand span) minus (non-dominant hand span).
 - (b) Calculate summary statistics for your differences.
 - (c) Formulate hypotheses in terms of the variable 'difference' to investigate the claim that, on average, the span of a person's dominant hand is greater than that of their non-dominant hand.
-

What is the key difference in experimental design between Activity 1 and Activity 2?

4.1 Two normal population variances

One possible question arising from the data collected in Activity 1 might be:

"Is the variability in dominant hand span the same for males and females?"

This question may be formulated as the following two hypotheses.

$$H_0: \sigma_1^2 = \sigma_2^2 \quad (1 = \text{male}, 2 = \text{female})$$

$$H_1: \sigma_1^2 \neq \sigma_2^2 \quad (\text{where } \sigma^2 \text{ denotes population variance})$$

It is important to note that, although an assumption of normal populations is required for such tests, no assumption is required as to the equality or otherwise of the population means μ_1 and μ_2 .

In Section 3.2, a statistic based upon the sample variance, $\hat{\sigma}^2$ was used in testing hypotheses concerning a single normal population variance (or standard deviation). Here, when testing the equality of two normal population variances (or standard deviations), the statistic used is the ratio of the two sample variances.

In fact

$$F = \frac{\hat{\sigma}_1^2 / \sigma_1^2}{\hat{\sigma}_2^2 / \sigma_2^2}$$

has an F distribution with degrees of freedom

$$v_1 = n_1 - 1 \quad \text{and} \quad v_2 = n_2 - 1, \quad \text{written } F_{(n_1-1, n_2-1)}.$$

The F distribution was developed by the American statistician, *G. W. Snedecor*, and so named in honour of *R. A. Fisher* (1890-1962) an eminent British statistician, who originally discovered the distribution in a slightly different form.

The distribution depends on the two parameters v_1 and v_2 for its shape and, except for large values of both, is positively skewed (like the χ^2 distribution). The distribution function for F exists for the range zero to infinity and is very complicated. The mean of

$F_{(v_1, v_2)}$ exists only for $v_2 > 2$ and is given by $\frac{v_2}{(v_2 - 2)}$. The

variance exists only for $v_2 > 4$ and is a complicated function of v_1 and v_2 .

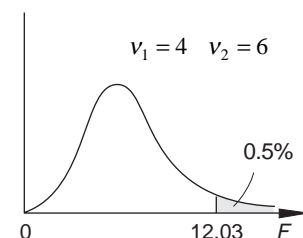
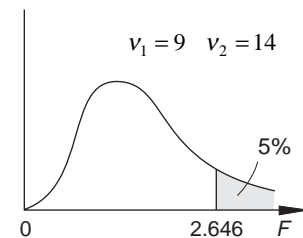
If $H_0: \sigma_1^2 = \sigma_2^2$ is true, then $F = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} \sim F_{(n_1-1, n_2-1)}$

and, adopting the convention of always putting the larger $\hat{\sigma}^2$ in the numerator, will result in H_0 being rejected when F becomes significantly large.

This significance can be assessed by making reference to tables of Upper Percentage Points of the F Distribution in which $v_1 = \text{column}$ and $v_2 = \text{row}$. For example, using Table 10:

the upper 5.0% point of $F_{(9,14)}$ is 2.646,

the upper 0.5% point of $F_{(4,6)}$ is 12.03.



Note that, for larger values of v_1 and / or v_2 , linear interpolation may be necessary to obtain 'accurate' percentage points.

What is the upper 2.5% point for $F_{(14,9)}$?

Example

A random sample of 10 hot drinks from Dispenser A had a mean volume of 203 ml and a standard deviation (divisor $(n-1)$) of 3 ml. A random sample of 15 hot drinks from Dispenser B gave corresponding values of 206 ml and 5 ml. The amount dispensed by each machine may be assumed to be normally distributed. Test, at the 5% significance level, the hypothesis that there is no difference in the variability of the volume dispensed by the two machines.

Solution

$$H_0: \sigma_A^2 = \sigma_B^2$$

$$H_1: \sigma_A^2 \neq \sigma_B^2 \quad (\text{two-tailed})$$

Significance level, $\alpha = 0.05$

Since $\hat{\sigma}_A < \hat{\sigma}_B$ and the larger value is always placed in the numerator,

$$v_1 = n_B - 1 = 14 \text{ and } v_2 = n_A - 1 = 9$$

Using interpolation, the upper 2.5% point for

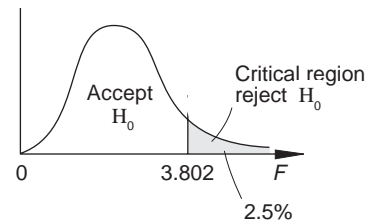
$$\begin{aligned} F_{(14,9)} &= F_{(12,9)} - \frac{2}{3}(F_{(12,9)} - F_{(15,9)}) \\ &= 3.868 - \frac{2}{3}(3.868 - 3.769) \\ &= 3.802 \end{aligned}$$

Thus critical region is $F > 3.802$

Test statistic is

$$F = \frac{\hat{\sigma}_B^2}{\hat{\sigma}_A^2} = \frac{5^2}{3^2} = 2.78$$

This value does not lie in the critical region. Thus there is no evidence, at the 5% level of significance, of a difference in the variability of the volume dispensed by the two machines.



Activity 3 Hand span 1 revisited

Test the hypothesis that the variability in dominant hand span for males is the same as that for females.

What assumption regarding hand span distributions did you make?

Explain why you consider the assumption to be reasonable.

Exercise 4A

1. An investigation was conducted into the dust content in the flue gases of two types of solid-fuel boilers. Thirteen boilers of type A and nine boilers of type B were used under identical fuelling and extraction conditions. Over a similar period, the following quantities, in grams, of dust were deposited in similar traps inserted in each of the twenty-two flues.

Type A	73.1	56.4	82.1	67.2
	78.7	75.1	48.0	
	53.3	55.5	61.5	
	60.6	55.2	63.1	
Type B	53.0	39.3	55.8	
	58.8	41.2	66.6	
	46.0	56.4	58.9	

Assuming that these independent samples come from normal populations, test for an equality of population variances. (AEB)

2. Korn Krispies are a type of breakfast cereal, and they are packed in boxes with a nominal net mass of 296 grams. Owing to overwhelming demand, the manufacturers have installed a new and faster machine to fill the boxes with cereal. However, to meet government regulations, amongst other things the variability in the packed masses of these boxes should not increase over present levels. The table below gives the masses of a random sample of 10 boxes of cereal from the original packing machine, and the masses of a random sample of 12 boxes of cereal from the new machine.

Original machine		New machine	
301.0	292.4	295.3	320.4
293.6	298.7	289.4	312.2
291.1	285.1	288.5	292.9
305.1	290.0	299.8	300.2
297.0	302.2	293.6	276.3
		308.9	280.3

Assuming that these independent samples came from underlying normal populations, use a 1% level of significance to determine whether an increase in variance has occurred.

(AEB)

3. In a research study aimed at improving the design of bus cabs it was necessary to measure the functional arm reach of bus drivers. In a pilot study a research worker made this measurement on a random sample of ten bus drivers from a large depot, and next day her assistant made this measurement on a random sample of eight bus drivers from the same depot. The results, in millimetres, were as follows.

Research worker	730	698	712	686	724
	711	679	762	683	673
Assistant	701	642	651	700	672
	674	656	649		

Assuming a normal distribution for functional arm reach, test, at the 10% significance level, whether the samples could have come from populations with the same variance. (AEB)

4. As part of a research study into pattern recognition, subjects were asked to examine a picture and see if they could distinguish a word. The picture contained the word 'technology' written backwards and camouflaged by an elaborate pattern. Of the 23 librarians who took part 11 succeeded in recognising the word whilst of 19 designers, 13 succeeded. The times, in seconds, for the successful subjects to recognise the word were as follows.

Librarians	55	18	99	54	87	11
	62	68	27	90	57	
Designers	23	69	34	27	51	29
	42	48	74	31	30	31

Stating any necessary assumptions, investigate the hypothesis that the variability of times for librarians significantly exceeds that for designers. (AEB)

5. The light attenuation of trees may be measured by photometric methods, which are very time consuming, or by photographic techniques which are much quicker. The light attenuation of an oak tree was repeatedly measured by both methods independently. The following results, expressed as percentages, were obtained.

Photometric	85.6	86.1	86.5	85.1	86.8	87.3
Photographic	82.4	84.7	86.1	87.2	82.4	85.8

Assuming normal distributions, test at the 5% level of significance whether there is a difference in the variability of the two methods.
(AEB)

6. A firm is to buy a fleet of cars for use by its salesmen and wishes to choose between two alternative models, A and B. It places an advertisement in a local paper offering 20 litres of petrol free to anyone who has bought a new car of either model in the last year. The offer is conditional on being willing to answer a questionnaire and to note how far the car goes, under typical driving conditions, on the free petrol supplied. The following data were obtained.

	Km driven on 20 litres of petrol			
Model A	187	218	173	235
Model B	157	198	154	184
	202	174	146	173

Assuming these data to be random samples from two normal populations, test whether the population variances may be assumed equal.

List good and bad features of the experimental design and suggest how you think it could be improved.

(AEB)

4.2 Two normal population means – case 1

Independent samples and known population variances – normal test

It is claimed that Brand A size D alkaline batteries last longer than those of Brand B.

Is this claim likely to be true for all Brand A size D alkaline batteries?

All mass-produced articles are liable to random variation which should be monitored and controlled, but cannot be eliminated entirely. Such variation is generally assumed, with good cause, to be approximately normally distributed. Thus it is quite possible that, whilst the claim may be true **on average**, it is not the case for every Brand A size D alkaline battery.

Which other population parameter may influence the validity of the claim?

Investigating claims of a population mean difference generally requires a comparison of two sample means; in the illustration above, measuring the lifetimes of all Brand A (and B) battery lifetimes would leave none for sale!

As noted earlier, the variance of a sample mean depends upon the sample size and the variance of the population from which the sample is selected. Consequently the sizes of the two samples and the variances of the two populations will influence the comparison of sample means.

From earlier work you have seen that:

$$(a) \text{ if } X \sim N(\mu, \sigma^2), \text{ then } \bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

$$(b) \text{ if } X_1 \sim N(\mu_1, \sigma_1^2) \text{ independent of } X_2 \sim N(\mu_2, \sigma_2^2), \\ \text{then } X_1 - X_2 \sim N(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2)$$

Combining these two results gives

$$\boxed{\bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)}$$

Hence

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

is a standardised normal statistic and may be used to test the equality of two normal population means, μ_1 and μ_2 , based upon independent random samples.

It is perhaps worth noting here that for $n_1 > 30$ and $n_2 > 30$, the sample variances, $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ may be used as estimates of σ_1^2 and σ_2^2 , respectively, so providing an approximate z statistic.

In this case, why can the requirement of normal populations be relaxed?

Example

The alkalinity, in milligrams per litre, of water in the upper reaches of rivers in a particular region is known to be normally distributed with a standard deviation of 10 mg/l. Alkalinity readings in the lower reaches of rivers in the same region are also known to be normally distributed, but with a standard deviation of 25 mg/l.

Ten alkalinity readings are made in the upper reaches of a river in the region and fifteen in the lower reaches of the same river with the following results.

Upper reaches	91	75	91	88	94	63	86	77	71	69
Lower reaches	86	95	135	121	68	64	113	108	79	62
	143	108	121	85	97					

Investigate, at the 1% level of significance, the claim that the true mean alkalinity of water in the lower reaches of this river is greater than that in the upper reaches.

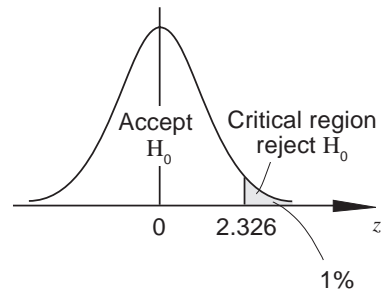
Solution

$$H_0: \mu_1 = \mu_2 \quad (1 = \text{lower}, 2 = \text{upper})$$

$$H_1: \mu_1 > \mu_2 \quad (\text{one-tailed})$$

Significance level, $\alpha = 0.01$

Critical region is $z > 2.326$



Under H_0 , the test statistic is

$$z = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Calculation gives $\bar{x}_1 = \frac{1485}{15} = 99.0$ and $\bar{x}_2 = \frac{805}{10} = 80.5$,

so

$$z = \frac{(99.0 - 80.5)}{\sqrt{\frac{25^2}{15} + \frac{10^2}{10}}} = 2.57$$

This value does lie in the critical region so H_0 is rejected. Thus there is evidence, at the 1% level of significance, to suggest that the true mean alkalinity of water in the lower reaches of the river is greater than that in the upper reaches.

Activity 4 Random numbers

Random numbers generated on calculators are claimed to be rectangularly distributed within the range 0 to 1.

- (a) State the mean and variance of this distribution.
- (b) Explain why the average of 10 such random numbers may be assumed to be approximately normally distributed.

Using a calculator (Model A) with a random number key, generate and calculate $n_A = 12$ such averages together with their mean \bar{x}_A . Repeat, using a different make of calculator (Model B), to obtain \bar{x}_B from $n_B = 8$ averages each of 10 random numbers.

- (c) Investigate the claim that $\mu_A = \mu_B$, assuming that

$$\sigma_A^2 = \sigma_B^2 = \frac{1}{120}.$$

- (d) Explain how the assumption in (c) was obtained.

Exercise 4B

- 1. The mass of crisps delivered into bags by a machine is known to be normally distributed with a standard deviation of 0.5 g.

Prior to a minor overhaul of the machine, the contents, in grams, of a random sample of six bags are as follows.

151.7 152.6 150.8 151.9 152.3 151.5

After the overhaul, which from past experience is known not to affect the standard deviation, the contents of a random sample of twelve bags were measured with the results below.

151.1 150.7 149.0 150.3 151.3 151.4
150.8 149.5 150.2 150.6 150.9 151.3

Test, at the 5% significance level, the hypothesis that the minor overhaul has had no effect on the mean mass of crisps delivered by the machine.

- 2. A firm obtains its supply of steel wire of a particular gauge from each of two manufacturers A and B. The firm suspects that the mean breaking strength, in newtons(N), of wire from manufacturer A differs from that supplied by manufacturer B.

The table below shows the breaking strengths of random samples of wire from each of the two manufacturers.

A	80.5	83.1	73.6	70.4	68.9	71.6	82.3	78.6	73.4
B	71.4	86.2	81.4	72.3	78.9	80.3	81.4	78.0	

Assuming all such breaking strengths to be normally distributed with a standard deviation of 5 N, investigate the firm's suspicion.

- 3. The manager of a lemonade bottling plant is interested in comparing the performance of two production lines, one of which has only recently been installed. For each line she selects 10 one-hour periods at random and records the number of crates completed in each hour. The table below gives the results.

Production line	Number of crates completed per hour
1 (new)	78 87 79 82 87 81 85 80 82 83
2 (old)	74 77 78 70 87 83 76 78 81 76

From past experience with this kind of equipment it is known that the variance in these figures will be 10 for Line 1 and 25 for Line 2. Assuming that these samples came from normal populations with these variances, test the hypothesis that the two populations have the same mean. (AEB)

4. Rice Pops (RP) are a type of breakfast cereal which are packed into boxes with a quoted net mass of 296 g by one of two different filling machines. The mass of RP delivered by filling machine A, an old machine, is known to be normally distributed with a standard deviation of 5 g. The mass of RP delivered by machine B, a new machine, is also known to be normally distributed but with a standard deviation of 3 g. The table below shows the net masses, in grams, of a random sample of 12 boxes filled by machine A and of a random sample of 15 boxes filled by machine B

Boxes filled by machine A

296	296	302
304	300	306
305	297	307
303	299	306

Boxes filled by machine B

296	297	299
301	297	299
298	302	304
298	299	303
297	296	299

Test the hypothesis that there is no significant difference in the mean mass of RP delivered by the two filling machines. (AEB)

5. During the first three months of 1993 a technician was timed for the repair of an electronic instrument on 12 separate occasions. In the same period a trainee technician was timed for the repair of a similar instrument on 14 occasions. These times, in minutes, are given in the table below.

Technician	344	278	267	234	212	271	
	341	391	176	164	214	399	
Trainee	279	351	282	280	258	267	312
	357	322	249	228	315	311	341

- (a) Assuming that these observations may be regarded as independent random samples from normal populations with known standard deviations of 80 minutes (for the technician) and 40 minutes (for the trainee technician), test the hypothesis that there is no difference in the mean times.
- (b) Subsequently it was learned that the times for the trainee were incorrectly recorded and that each of the values above is 30 minutes too small. What, if any, difference does this make to the result of the test you have just completed? (AEB)

6. James and his sister, Alison, each deliver 30 papers on their evening paper rounds and each are paid the same amount. One evening Alison claims this system of equal payment to be unfair as her round takes on average longer than that of her brother, James. To test her claim their father, unknown to them, records their delivery times for 12 consecutive days. One of Alison's times had to be discounted as the front tyre of her bicycle was punctured. The recorded times, in minutes, were as shown below.

	Delivery times (minutes)											
James	45	30	39	32	34	43	38	35	43	39	32	34
Alison	49	42	39	45	38	49	43	36	33	41	36	

- (a) Assuming each child's delivery times are normally distributed with the same known standard deviation of 5 minutes, test whether Alison's claim is justified, using a 5% level of significance.
- (b) If Alison had in fact claimed that the system of equal payment was unfair because the average delivery times for the two rounds were different, what changes would you make to your test procedure in (a)? (AEB)

4.3 Two normal population means – case 2

Independent samples and unknown but equal population variances – *t*-test

In Section 4.2, the test statistic z required that the two populations are normal with known variances, σ_1^2 and σ_2^2 . If however both sample sizes are greater than 30, it was stated that sample variances may be used as estimates to provide an approximate z statistic.

In most practical situations, the population variances are unknown and the sample sizes are less than 30. Refer to your data collected in Activity 1.

If it may be assumed, or has been confirmed using the *F*-test of Section 4.1, that the population variances are equal, but unknown, then a test is available for all sample sizes providing the two populations are normal. (For small samples from normal populations with unknown and unequal variances, an involved approximate *t*-test is available, but it is outside the scope of this text.)

Returning to the test statistic of Section 4.1, defined by

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

If $\sigma_1^2 = \sigma_2^2 = \sigma^2$, then

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Now both sample variances, $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$, are estimates of σ^2 , so this information can be combined to form a pooled (weighted) estimate of variance defined by

$$\hat{\sigma}_p^2 = \frac{(n_1 - 1)\hat{\sigma}_1^2 + (n_2 - 1)\hat{\sigma}_2^2}{n_1 + n_2 - 2}$$

From Section 2.1 you saw that when σ^2 is replaced by $\hat{\sigma}^2$ in a z statistic the result is a *t* statistic.

Hence

$$\frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\hat{\sigma}_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

is a t statistic with degrees of freedom given by $\nu = n_1 + n_2 - 2$.

What interpretation has the pooled estimate of variance if the two samples are of the same size?

Example

Mr Brown is the owner of a small bakery in a large town. He believes that the smell of fresh baking will encourage customers to purchase goods from his bakery. To investigate this belief, he records the daily sales for 10 days when all the bakery's windows are open, and the daily sales for another 10 days when all the windows are closed. The following sales, in £, are recorded.

Windows open	202.0	204.5	207.0	215.5	190.8
	215.6	208.8	187.8	204.1	185.7
Windows closed	193.5	192.2	199.4	177.6	205.4
	200.6	181.8	169.2	172.2	192.8

Assuming that these data may be deemed to be random samples from normal populations with the same variance, investigate the baker's belief.

Solution

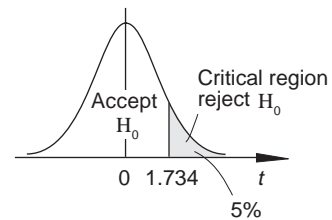
$$H_0: \mu_1 = \mu_2 \quad (1 = \text{open}, 2 = \text{closed})$$

$$H_1: \mu_1 > \mu_2 \quad (\text{one-tailed})$$

Significance level, $\alpha = 0.05$ (say)

Degrees of freedom, $\nu = 10 + 10 - 2 = 18$

Critical region is $t > 1.734$



Under H_0 , the test statistic is

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\hat{\sigma}_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Calculation gives

$$\bar{x}_1 = 202.18, \quad \hat{\sigma}_1^2 = 115.7284$$

and $\bar{x}_2 = 188.47, \quad \hat{\sigma}_2^2 = 156.6534$

$$\begin{aligned} \text{Hence } \hat{\sigma}_p^2 &= \frac{9 \times 115.7284 + 9 \times 156.6534}{10 + 10 - 2} \\ &= \frac{115.7284 + 156.6534}{2} \quad (\text{mean when } n_1 = n_2) \end{aligned}$$

$$\text{so } \hat{\sigma}_p = 11.67$$

$$\text{Thus } t = \frac{202.18 - 188.47}{11.67 \sqrt{\frac{1}{10} + \frac{1}{10}}} = 2.63$$

This value does lie in the critical region so H_0 is rejected. Thus there is evidence, at the 5% level of significance, to suggest that the smell of fresh baking will encourage customers to purchase goods from Mr Brown's bakery.

Example

Referring back to the Example in Section 4.1 concerning the two drink dispensers, test, at the 5% level of significance, the hypothesis that there is no difference in the mean volume dispensed by the two machines.

Solution

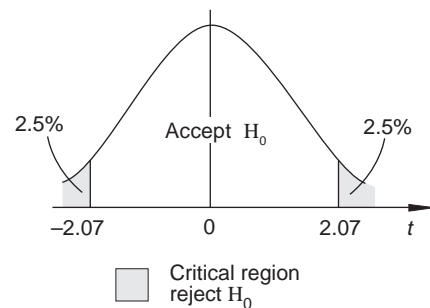
$$H_0: \mu_A = \mu_B$$

$$H_1: \mu_A \neq \mu_B \quad (\text{two-tailed})$$

Significance level, $\alpha = 0.05$

Degrees of freedom, $\nu = 10 + 15 - 2 = 23$

Using interpolation, critical region is
 $t < -2.07$ or $t > 2.07$



Under H_0 , the test statistic is

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\hat{\sigma}_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$\bar{x}_A = 203, \hat{\sigma}_A = 3 \text{ and } \bar{x}_B = 206, \hat{\sigma}_B = 5$$

Hence

$$\hat{\sigma}_p^2 = \frac{9 \times 3^2 + 14 \times 5^2}{10 + 15 - 2} = 18.7391 = 4.33^2$$

Thus
$$t = \frac{203 - 206}{4.33 \sqrt{\frac{1}{10} + \frac{1}{15}}} = -1.70$$

This value does not lie in the critical region so H_0 is not rejected. Thus there is no evidence, at the 5% level of significance, to suggest that there is a difference in the mean volume dispensed by the two machines.

Activity 5 *Hand span 1 revisited again, perhaps!*

If in Activity 3, you did NOT reject the hypothesis that the variability in dominant hand span for males is the same as that for females, now test the hypothesis that the true mean hand span for males is greater than that for females.

Activity 6 *Random numbers revisited*

Re-investigate the claim made in Part (c) of Activity 4, assuming only that $\sigma_A^2 = \sigma_B^2$. Compare your two conclusions.

Exercise 4C

1. A microbiologist wishes to determine whether there is any difference in the time it takes to make yoghurt from two different starters; lactobacillus acidophilus (A) and bulgarius (B). Seven batches of yoghurt were made with each of the starters. The table below shows the time taken, in hours, to make each batch.

Starter A	6.8	6.3	7.4	6.1	8.2	7.3	6.9
Starter B	6.1	6.4	5.7	5.5	6.9	6.3	6.7

Assuming that both sets of times may be considered to be random samples from normal populations with the same variance, test the hypothesis that the mean time taken to make yoghurt is the same for both starters.

2. Referring to Question 1 of Exercise 4A, test for an equality of population means. (AEB)
3. Referring to Question 2 of Exercise 4A, test the hypothesis that the original and new machines deliver the same mean mass of Korn Krispies. (Use a 1% significance level.) (AEB)
4. Referring to Question 3 of Exercise 4A, show that, at the 1% significance level, the hypothesis that the samples are from populations with equal means is rejected. (AEB)
5. Referring to Question 5 of Exercise 4A, test, at the 5% significance level, whether there is a difference in the mean measurement by the two methods. (AEB)
6. A new chemical process is developed for the manufacture of nickel-cadmium batteries. The company believes that this new process will increase the mean lifetime of a battery by 5 hours as compared to that of batteries produced by the old process. Sixteen batteries produced by the old process were randomly selected and the mean and the standard deviation of the lifetimes of these batteries were 105.2 hours and 9.1 hours, respectively. Fifteen batteries produced by the new process were also randomly selected and calculations gave corresponding values of 112.4 and 8.3 hours.
Assuming all battery lifetimes to be normally distributed, test at the 5% significance level whether there is
 - (a) a difference in the variability of the two processes,
 - (b) an increase of 5 hours in the mean lifetime of batteries produced by the new process as compared to that of batteries produced by the old process.

4.4 Two normal population means – case 3

Paired samples – *t*-test

To assess the claims of a new weight-reducing diet programme, a researcher weighs each of a random sample of 10 people about to enrol on the programme. The researcher then calculates correctly their mean and standard deviation to be 72.1 kg and 6.8 kg, respectively. Later, the researcher weighs each of a random sample of 12 people who have adhered strictly to the programme for three months. Correct calculations give their mean weight as 80.4 kg with a standard deviation of 5.1 kg. The researcher therefore concludes that, rather than reduce people's weights, the diet programme actually appears to cause people to gain weight!

Why would the researcher's data and conclusion be challenged, even ridiculed, by the diet programme's sponsors?

In some investigations, the inherent variation of the subjects or items used in the study can negate, mask or enhance the actual differences of interest. However in many of these cases, with sensible planning of the investigation, the differences of interest can be assessed separately from the inherent variation.

Thus in the previous illustration regarding a diet programme, the researcher should have weighed the same 10 people after three months as were weighed before enrolment. The measure of weight loss for each person would then be 'weight before minus weight after 3 months'. The mean and standard deviation of the 10 differences so calculated could then be validly used to assess the programme's claim. Note that variations in weight between the 10 people are no longer confused with weight loss. On the other hand, the two samples are no longer independent since the same 10 people form each sample, with the results thereby occurring in pairs; hence the name, paired samples.

Assuming that the two populations from which the paired samples of size n are selected are distributed with means μ_1 and μ_2 , respectively, then from earlier work the differences between pairs will be distributed with mean $\mu_1 - \mu_2 = \mu_d$ and variance σ_d^2 , say. Thus a test of $H_0: \mu_1 = \mu_2$ is equivalent to a test of $H_0: \mu_d = 0$.

Although the two populations may well be normally distributed, the key distributional assumption for the test is that the differences between pairs of values are approximately normally distributed.

Let \bar{d} and $\hat{\sigma}_d^2$ denote the mean and variance of the sample of n differences.

Then
$$\bar{d} \sim N\left(\mu_d, \frac{\sigma_d^2}{n}\right)$$

or
$$\frac{\bar{d} - \mu_d}{\frac{\hat{\sigma}_d}{\sqrt{n}}} \sim N(0, 1)$$

Thus, from Section 3.1,

$$\boxed{\frac{\bar{d} - \mu_d}{\frac{\hat{\sigma}_d}{\sqrt{n}}}}$$

is a t statistic with degrees of freedom, $\nu = n - 1$.

Example

A school mathematics teacher decides to test the effect of using an educational computer package, consisting of geometric designs and illustrations, to teach geometry. Since the package is expensive, the teacher wishes to determine whether using the package will result in an improvement in the pupils' understanding of the topic. The teacher randomly assigns pupils to two groups; a control group receiving standard lessons and an experimental group using the new package. The pupils are selected in pairs of equal mathematical ability, with one from each pair assigned at random to the control group and the other to the experimental group. On completion of the topic the pupils are given a test to measure their understanding. The results, percentage marks, are shown in the table.

Pair	1	2	3	4	5	6	7	8	9	10
Control	72	82	93	65	76	89	81	58	95	91
Experimental	75	79	84	71	82	91	85	68	90	92

Assuming percentage marks to be normally distributed, investigate the claim that the educational computer package produces an improvement in pupils' understanding of geometry.

Solution

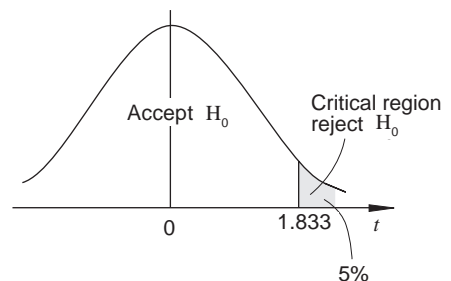
$H_0: \mu_d = 0$ Difference = Experimental – Control

$H_1: \mu_d > 0$ (one-tailed)

Significance level, $\alpha = 0.05$ (say)

Degrees of freedom, $\nu = 10 - 1 = 9$

Critical region is $t > 1.833$



Under H_0 , the test statistic is

$$t = \frac{\bar{d}}{\frac{\hat{\sigma}_d}{\sqrt{n}}}$$

The 10 differences (Experimental - Control) are

$$d: 3 \quad -3 \quad -9 \quad 6 \quad 6 \quad 2 \quad 4 \quad 10 \quad -5 \quad 1$$

Hence $\sum d = 15$ and $\sum d^2 = 317$

so $\bar{d} = 1.5$ and $\hat{\sigma}_d = 5.72$

Thus $t = \frac{1.5}{\frac{5.72}{\sqrt{10}}} = 0.83$

This value does not lie in the critical region so H_0 is not rejected. Thus there is no evidence, at the 5% level of significance, to suggest that the educational computer package produces an improvement in pupils' understanding of geometry.

Activity 7 Hand span 2 revisited

Assuming the difference between dominant and non-dominant hand spans to be normally distributed, investigate the claim made in Part (c) of Activity 2.

Exercise 4D

1. A random sample of eleven students sat a Chemistry examination consisting of one theory paper and one practical paper. Their marks out of 100 are given in the table below.

Student	A	B	C	D	E	F	G	H	I	J	K
Theory mark	30	42	49	50	63	38	43	36	54	42	26
Practical mark	52	58	42	67	94	68	22	34	55	48	17

Assuming differences in pairs to be normally distributed, test, at the 5% level of significance, the hypothesis of no difference in mean mark on the two papers. (AEB)

2. A convenience food, known as 'Quicknosh', was introduced into the British market in January 1992. After a poor year for sales the manufacturers initiated an intensive advertising campaign during January 1993. The table below records the sales, in thousands of pounds, for a one-month period before and a one-month period after the advertising campaign, for each of eleven regions.

Region	A	B	C	D	E	F	G	H	I	J	K
Sales before campaign	2.4	2.6	3.9	2.0	3.2	2.2	3.3	2.1	3.1	2.2	2.8
Sales after campaign	3.0	2.5	4.0	4.1	4.8	2.0	3.4	4.0	3.3	4.2	3.9

Determine, at the 5% significance level, whether an increase in mean sales has occurred by using the t -test for paired values. (AEB)

3. In an investigation to compare the accuracy of Crackshot and Fastfire 12-bore shotguns in clay pigeon shooting, ten competitors each fired 100 shots with each make of gun. Their scores are shown in the table below.

Competitor	A	B	C	D	E	F	G	H	I	J
Crackshot	93	99	90	86	85	94	87	91	96	79
Fastfire	87	91	86	87	78	95	89	84	88	74

It may be assumed that the differences between pairs of scores are approximately normally distributed. Examine the claim that the Crackshot shotgun is the more accurate for clay pigeon shooting. (AEB)

4. The following data are the third and fourth round scores of a random sample of five competitors in an open golf tournament.

Competitor	A	B	C	D	E
3rd round	76	75	72	75	79
4th round	70	73	71	68	76

Use a paired t -test and a 5% significance level to test whether there is a difference in the mean score of all competitors in the two rounds.

(AEB)

5. In a study of memory recall, 12 students were given ten minutes to try to memorise a list of 20 nonsense words. Each student was then asked to list as many of the words as he or she could remember both one hour and twenty-four hours later. The numbers of words recalled correctly for each student are shown below.

Student	A	B	C	D	E	F	G	H	I	J	K	L
1 hr later	14	9	18	12	13	17	16	16	19	8	15	7
24 hrs later	10	6	14	6	8	10	12	10	14	5	10	5

Stating any necessary assumptions, use a paired t -test to determine whether there is evidence, at the 5% level of significance, that for all such students, the mean number of words recalled after one hour exceeds that recalled after twenty-four hours by 5 words. (AEB)

6. The temperature of the earth may be measured either by thermometers on the ground (x), which is an accurate but tedious method, or by sensors mounted in space satellites (y), which is a less accurate method and may be biased. The following table gives readings ($^{\circ}\text{C}$) taken by both methods at eleven sites.

Site	Ground therm, x	Satellite sensors, y
1	4.6	4.7
2	17.3	19.5
3	12.2	12.5
4	3.6	4.2
5	6.2	6.0
6	14.8	15.4
7	11.4	14.9
8	14.9	17.8
9	9.3	9.7
10	10.4	10.5
11	7.2	7.4

Given that all readings are normally distributed, investigate the hypothesis that satellite sensors give, on average, significantly higher readings than the ground thermometers. (AEB)

4.5 Two population medians – case 1

Paired samples – sign test

The table below shows the grades obtained by each of a random sample of ten students in two pieces of Statistics coursework.

Student	1	2	3	4	5	6	7	8	9	10
Coursework 1 grade	A	B	B	C	D	C	C	A	B	C
Coursework 2 grade	B	C	B	D	C	D	C	B	C	D

What distinguishes these data from those considered so far in this Chapter?

The above grades are the results of a paired samples study since the same ten students constituted both samples. However the paired t -test of the previous Section requires the differences in pairs to be normally distributed. This is certainly not the case here so a non-parametric test is required. In fact since the grades are letters, rather than numbers, differences can reasonably be listed only as:

+ + 0 + - + 0 + + +

where

+ denotes (coursework 1 grade > coursework 2 grade)

0 denotes (coursework 1 grade = coursework 2 grade)

- denotes (coursework 1 grade < coursework 2 grade)

Now if the distribution of grades in the two courseworks is the same, then the two population median grades, η_1 and η_2 , will be the same.

Hence, under the null hypothesis of no difference between the grades in the two courseworks,

$$\begin{aligned} P(\text{difference is positive}) &= P(\text{difference is negative}) \\ &= 0.5. \end{aligned}$$

This hypothesis can therefore be tested by reference to a binomial distribution with n = number of non-zero differences and $p = 0.5$.

What probability is required in the above illustration?

In the above illustration, the number of non-zero differences is $10 - 2 = 8$, so X = number of +ve signs $\sim B(8, 0.5)$ under H_0 , with the observed value of X being 7.

Since for $X \sim B(8, 0.5)$

$$\begin{aligned} P(X \geq 7) &= 1 - P(X \leq 6) \\ &= 1 - 0.9648 \\ &= 0.0352 > 0.025 \quad (5\% \text{ two-tailed}) \end{aligned}$$

there is no significant evidence to suggest that there is a difference between the grades achieved in the two statistics courseworks.

What probability would have been evaluated if the observed value of X had been 1?

Example

On a particular day the incoming mail in each of twelve selected towns was randomly divided into two similar batches prior to sorting. In each town one batch was then sorted by the traditional hand sorting method, the other by a new Electronic Post Code Sensor Device (EPCSD). The times taken, in hours, to complete the sorting of the batches are recorded below.

Town	A	B	C	D	E	F	G	H	I	J	K	L
Hand sort time	4.3	4.1	5.6	4.0	5.9	4.9	4.3	5.4	5.6	5.2	6.1	4.7
EPCSD sort time	3.7	5.3	4.5	3.1	4.8	4.9	3.5	4.9	4.6	4.1	5.7	4.7

Use the sign test and a 1% level of significance to investigate the claim that the EPCSD method is quicker.

Solution

Let difference = (Hand sort time) – (EPCSD sort time)

$$H_0: \text{No difference in the times } (\eta_1 = \eta_2)$$

$$H_1: \text{EPCSD method is quicker } (\eta_1 > \eta_2; \text{ one-tailed})$$

Significance level, $\alpha = 0.01$

Signs of differences are

$$+ \quad - \quad + \quad + \quad + \quad 0 \quad + \quad + \quad + \quad + \quad + \quad 0$$

Let X denote the number of + signs.

Then, ignoring the two 0's in this case, under H_0 ,

$$X \sim B(10, 0.50) \text{ with observed value of } X = 9$$

$$\text{From tables, } P(X \geq 9) = 1 - P(X \leq 8)$$

$$= 1 - 0.9893$$

$$= 0.0107 > 0.01 \quad (\text{one-tailed})$$

Thus there is no evidence, at the 1% level of significance, that the EPSCD method is quicker than the traditional hand sorting method.

Activity 8 Brand preference

Obtain from your local supermarket two large bottles of cola (or orange or lemonade); one bottle having a well-known brand label, the other having the supermarket's own brand label.

(You may alternatively, or additionally, investigate crisps, biscuits, chocolate, baked beans, etc, providing that they are comparable in taste and cannot be identified readily by observation.)

Arrange for between 10 and 20 tasters to sample a small amount of each drink in (plastic) beakers labelled simply I or II with only you knowing which beaker contains the well-known brand and which contains the supermarket's own brand.

Ask the tasters independently to grade the quality of each of their samples as either

A: excellent, B: good, C: acceptable or D: unacceptable.

Using the sign test, investigate the claim that people cannot taste the difference between a well-known brand of soft drink and a supermarket's own brand.

Exercise 4E

1. Fifteen girls were each given an oral examination and a written examination in French. Their grades (highest = A, lowest = F) in the two examinations were as follows.

Girl	1	2	3	4	5	6	7	8
Oral exam	A	B	C	D	E	C	B	E
Written exam	B	D	D	C	E	D	C	D

Girl	9	10	11	12	13	14	15
Oral exam	E	C	D	C	E	C	B
Written exam	C	C	E	E	F	D	C

Using the sign test, investigate the hypothesis that one examination produces significantly different grades from the other. (AEB)

2. Apply the sign test to the data in Question 1 of Exercise 4D. Compare your conclusion with that obtained previously. (AEB)
3. Apply the sign test to the data in Question 2 of Exercise 4D. What assumption, made when using the paired *t*-test, is not needed for the sign test? (AEB)
4. Apply the sign test to the data in Question 3 of Exercise 4D. Comment on your conclusions to the two analyses. (AEB)

5. To measure the effectiveness of a drug for asthmatic relief, twelve subjects, all susceptible to asthma, were each randomly administered either the drug or a placebo during two separate asthma attacks. After one hour an asthmatic index was obtained on each subject with the following results.

Subject	1	2	3	4	5	6	7	8	9	10	11	12
Drug	28	31	17	18	31	12	33	24	18	25	19	17
Placebo	32	33	23	26	34	17	30	24	19	23	21	24

Making no distributional assumptions, investigate the claim that the drug significantly reduces the asthmatic index. (AEB)

6. Two methods for measuring the level of vitamin B12 in red blood cells were compared. Blood samples were taken from ten healthy adults, and, for each blood sample, the B12 level was determined using both methods. The resultant data are given below.

Adult	1	2	3	4	5	6	7	8	9	10
Method 1	204	238	209	277	197	226	203	131	282	76
Method 2	199	230	198	253	180	209	213	137	250	82

Use the signs of the differences to test the hypothesis that there is no difference between the two methods as regards the measurement of the B12 level in red blood cells. (AEB)

4.6 Two population medians – case 2

Paired samples – Wilcoxon signed-rank test

As stated earlier, the paired t -test of Section 4.4 requires the differences in pairs to be normally distributed. In addition, the relative magnitudes of the differences are taken into account by the calculations of \bar{d} and $\hat{\sigma}_d^2$.

The sign test of the previous section requires no such assumption of normality but, on the other hand, it makes no account for the relative magnitudes of the differences.

The alternative test for paired samples now to be considered, called the Wilcoxon (signed-rank) test, may be considered to be a compromise between the two previous methods in that it

- (i) does not require any distributional assumptions of normality, so is non-parametric;
- (ii) does take into account the relative magnitudes of the differences.

It is therefore to be preferred to the sign test when the sizes of the differences can be determined readily.

What do you understand by the term 'rank'?

The Wilcoxon test procedure is best developed using a specific example.

In a comparison of two computerised methods, A and B, for measuring physical fitness, a random sample of eight people was assessed by both methods. Their scores (maximum 20) were recorded as follows.

Subject	1	2	3	4	5	6	7	8
Method A	11.2	8.6	6.5	17.3	14.3	10.7	9.8	13.3
Method B	10.4	12.1	9.1	15.6	16.7	10.7	12.8	15.5
Difference (A - B)	+0.8	-3.5	-2.6	+1.7	-2.4	0.0	-3.0	-2.2

Since previous investigations of each method have concluded that scores are not normally distributed, a paired t -test is not appropriate.

A sign test could be used, but this would only use the signs of the differences (2 positive, 5 negative, with 1 zero).

A closer examination of these differences reveals that the 2 positive ones are the smallest in absolute value. This suggests that they are less important than the five negative values, which are all greater in absolute value. This idea is the basis of the **Wilcoxon test**.

Having determined the differences, they are ranked ignoring their signs, or, in other words, the absolute differences are ranked. Next the signs of the differences are attached to the ranks, hence the term 'signed-rank'.

The test statistic, T , is then determined as

$$T = \text{maximum of } T_+ \text{ and } T_-,$$

where T_+ = sum of ranks with positive sign

and T_- = sum of ranks with negative sign.

What will be sum of T_+ and T_- for n ranked differences?

Thus for the current example,

Difference (A - B)	+0.8	-3.5	-2.6	+1.7	-2.4	0.0	-3.0	-2.2
Absolute difference A - B	0.8	3.5	2.6	1.7	2.4	0.0	3.0	2.2
Rank of A - B	1	7	5	2	4		6	3
Signed -rank of difference	+1	-7	-5	+2	-4		-6	-3

and hence,

$$T_+ = 1 + 2 = 3 \quad T_- = 3 + 4 + 5 + 6 + 7 = 25$$

giving $T = 25$

Note that here, $n = 8 - 1$ (zero difference) = 7,

$$(T_+ + T_-) = 28 = (7 \times 8) \div 2,$$

and in general, $(T_+ + T_-) = n(n + 1) \div 2$.

Under the null hypothesis of no real difference (i.e. $\eta_A = \eta_B$), each rank is equally likely to be associated with a positive or negative sign. Thus in this example, the list of signed ranks consists of all the possible combinations of

$$\pm 1 \quad \pm 2 \quad \pm 3 \quad \pm 4 \quad \pm 5 \quad \pm 6 \quad \pm 7$$

There are $2^7 = 128$ possible combinations, all equally likely under H_0 , and this enables probabilities for each tail to be calculated as follows.

Value of T	Possible arrangements (ranks of same sign)	Probability	Cumulative probability
28	all	$(\frac{1}{2})^7$	0.0078125
27	rank 1	$(\frac{1}{2})^7$	0.0156250
26	rank 2	$(\frac{1}{2})^7$	0.0234375
25	rank 3 or ranks 1 & 2	$2 \times (\frac{1}{2})^7$	0.0390625
24	rank 4 or ranks 1 & 3	$2 \times (\frac{1}{2})^7$	0.0546875

Thus under the null hypothesis, H_0 , $P(T \geq 25) = 0.0391$

For a two-tailed 5% test, this probability is greater than 0.025, so the null hypothesis of no difference between the two methods A and B would not be rejected.

What is the smallest value of T for which H_0 would have been rejected?

When two or more differences have the same absolute value they are termed ties. In such cases the general rule is to replace the ties by the average rank of all the observations involved in the tie.

e.g. The ranks for 2.5 2.7 2.8 2.8 2.8 3.1 3.2 3.2
 would be 1 2 4 4 4 6 7.5 7.5

Probability calculations as above for values of T are not in general necessary since tables of (approximate) critical values are available (see Appendix, Table 11).

Example

An athletics coach wishes to test the value to his athletes of an intensive period of weight training and so he selects twelve 400-metre runners from his region and records their times, in seconds, to complete this distance. They then undergo his programme of weight training and have their times, in seconds, for 400 metres measured again. The table below summarises the results.

Athlete	A	B	C	D	E	F	G	H	I	J	K	L
Before	51.0	49.8	49.5	50.1	51.6	48.9	52.4	50.6	53.1	48.6	52.9	53.4
After	50.6	50.4	48.9	49.1	51.6	47.6	53.5	49.9	51.0	48.5	50.6	51.7

Use the Wilcoxon signed-rank test to investigate the hypothesis that the training programme will significantly improve athletes' times for the 400 metres.

Solution

Let difference = (Time before)– (Time after)

H_0 : Training programme has no effect ($\eta_B = \eta_A$)

H_1 : Training programme improves time ($\eta_B > \eta_A$)
(one-tailed)

Significance level, $\alpha = 0.05$, say

$n = 12 - 1 = 11$ (1 zero difference)

From tables, critical region is $T > 52$

d	+0.4	-0.6	+0.6	+1.0	0.0	+1.3	-1.1	+0.7	+2.1	+0.1	+2.3	+1.7
$ d $	0.4	0.6	0.6	1.0	0.0	1.3	1.1	0.7	2.1	0.1	2.3	1.7
rank	2	3.5	3.5	6		8	7	5	10	1	11	9
s-rank	+2	-3.5	+3.5	+6		+8	-7	+5	+10	+1	+11	+9

Hence $T_+ = 55.5$ and $T_- = 10.5$

(Check: $T_+ + T_- = 66 = (11 \times 12) \div 2$)

Thus $T = 55.5$

This value does lie in the critical region so H_0 is rejected. There is evidence, at the 5% level of significance, that the weight training programme does improve athletes' times for the 400 metres.

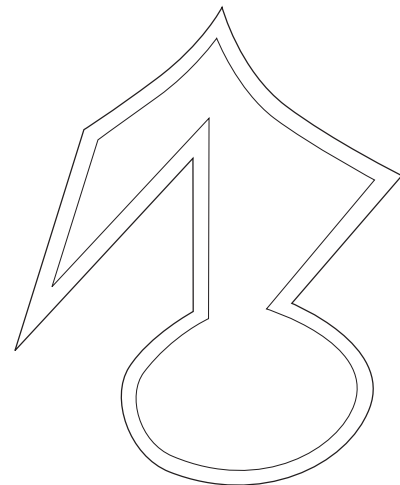
What alternative tests are available to you for these data, and what, if any, assumptions are necessary?

Activity 9 *Faster with practice*

Construct a simple random shape such as that printed here. Arrange for it to be printed at least 5 times on a plain sheet of A4 white paper.

Ask between 10 and 20 subjects to draw round each shape using their non-dominant hand. Time each subjects first and last tracings.

Investigate the hypothesis that tracing times improve with practice.



Exercise 4F

- Apply the Wilcoxon signed-rank test to the data in Question 2 of Exercise 4D. Compare your conclusion with those obtained previously. (AEB)
- Apply the Wilcoxon test to the data in Question 3 of Exercise 4D. List and discuss your three conclusions to this set of data. (AEB)
- Apply the Wilcoxon test to the data in Question 5 of Exercise 4E. Compare your conclusion with that from the sign test. (AEB)
- As part of her research into the behaviour of the human memory, a psychologist asked 15 schoolgirls to talk for five minutes on 'my day at school'. Each girl was then asked to record how many times she thought that she had used the word nice during this period. The table below gives their replies together with the true values.
- Re-analyse the data in Question 6 of Exercise 4E using Wilcoxon's signed-rank test. Compare your conclusion with that obtained from the sign test. What other test could be applied to the data, and what assumption would be necessary? (AEB)
- The Ministry of Defence is considering which of two shoe leathers it should adopt for its new Army boot. They are particularly interested in how boots made from these leathers wear and so 15 soldiers are selected at random and each man wears one boot of each type. After six months the wear, in millimetres, for each boot is recorded as follows.

Girl	A	B	C	D	E	F	G	H
True value	12	20	1	8	0	12	12	17
Recorded value	9	19	3	14	4	12	16	14

Girl	I	J	K	L	M	N	O
True value	6	5	24	23	10	18	16
Recorded value	5	9	20	16	11	17	19

Use Wilcoxon's test to investigate whether schoolgirls can remember accurately the frequency with which they use a particular word in a verbal description. (AEB)

Soldier	1	2	3	4	5	6	7	8
Leather A	5.4	2.6	4.3	1.1	3.3	6.6	4.4	3.5
Leather B	4.7	3.2	3.8	2.3	3.6	7.2	4.4	3.9

Soldier	9	10	11	12	13	14	15
Leather A	1.2	1.3	4.8	1.2	2.8	2.0	6.1
Leather B	1.9	1.2	5.8	2.0	3.7	1.8	6.1

Use the Wilcoxon signed-rank test to investigate the hypothesis that the wear in the two leathers is the same. Why may this test be considered a better approach to this problem than the sign test? (AEB)

4.7 Miscellaneous Exercises

- The vitamin content of the flesh of each of a random sample of eight oranges and of a random sample of five lemons was measured. The results are given in milligrams per 10 grams.

Oranges	1.14	1.59	1.57	1.33	1.08	1.27	1.43	1.36
Lemons	1.04	0.95	0.63	1.62	1.11			

Assuming vitamin content to be normally distributed, test the hypothesis that both samples come from populations with the same variance. (AEB)
- Over a certain period of time, a random sample of 15 private subscribers connected to telephone exchange X used a total of 7980 units. Over the same period of time, a random sample of 20 subscribers connected to exchange Y used a total of 10 220. It is known that, for both exchanges, the numbers of units used by private subscribers during the period are normally distributed with standard deviation 100.

Test the hypothesis that there is no difference between the mean number of units used by private subscribers at exchanges X and Y .
- An economist believes that a typical basket of weekly provisions, purchased by a family of four, costs more in Southville than it does in Nortown. Six stores were randomly selected in each of these two cities and the following costs, for identical baskets of provisions, were observed.

Southville	12.32	13.10	12.11	12.84	12.52	12.71
Nortown	11.95	11.84	12.22	12.67	11.53	12.03

 - Explain why a paired test would not be appropriate here.
 - Assuming costs in both towns to be normally distributed with the same variance, test the economist's belief. (AEB)
- A high nitrate intake in food consumption is suspected of retarding the growth of some animals. The following data are the results of an experiment to measure the percentage gain in mass of young laboratory mice given either a standard diet (A) or an extra 200 parts per million of nitrate in their diet (B).

A	18.2	25.8	16.8	14.9	19.6	26.5	17.5
B	13.4	18.8	20.5	6.5	22.2	15.0	12.2
	14.3	18.0	15.1				

Assuming that both percentages are normally distributed with a standard deviation of 4.5, test, at the 1% level of significance, the hypothesis that a high nitrate intake retards the mean percentage gain in mass of mice.

After the experiment was performed it was discovered that the laboratory mice used were not a homogeneous population. In fact, most of the mice in the control group were appreciably heavier than those in the experimental group. Discuss briefly the possible effect of this information on the validity of your analysis.

(AEB)

5. As part of an investigation into the effects of alcohol on the human body at high altitude, ten male subjects were taken to a simulated altitude of 8000 m and given several tasks to perform. Each subject was carefully observed for deterioration in performance due to lack of oxygen, and the time, in seconds, at which useful consciousness ended was recorded. Three days later, the experiment was repeated one hour after the same ten subjects had unknowingly consumed 1 ml of 100%-proof alcohol per 5 kilograms of body mass. The time, in seconds, of useful consciousness was again recorded. The resulting data are given below.

Subject	1	2	3	4	5
No alcohol	260	565	900	630	280
Alcohol	185	375	310	240	215

Subject	6	7	8	9	10
No alcohol	365	400	735	430	900
Alcohol	420	405	205	255	900

Using an appropriate parametric test, determine whether or not these data support the hypothesis that the consumption of the stated amount of alcohol reduces the mean time of useful consciousness at high altitudes.

Name an alternative non-parametric test, and indicate the assumption that is then no longer required.

(AEB)

6. An on-line catalogue of books is being introduced into a college library. Formerly the catalogue was held on microfiche. To test the new system, students were selected at random and asked to obtain some specified information from the microfiche catalogue and a further sample of students was asked to obtain the same information from the on-line catalogue. The times, in seconds, were as follows.

Microfiche	68	91	71	96	97	75		
On-line	85	69	93	79	117	79	78	102

- (a) Assuming a normal distribution, test, at the 5% significance level, whether there is a difference in
- the standard deviations of times for the two methods,
 - the means of the times for the two methods.
- (b) One student had taken 297 seconds to obtain the information using the on-line catalogue due to an initial misunderstanding of how to use the equipment. It had been decided to exclude this result from the data above. Comment on this decision and on the effect the inclusion of this result would have had on the assumptions you made in carrying out the test in (a)(ii).
- (AEB)
7. Two analysers are used in a hospital laboratory to measure blood creatinine levels. These are used as a measure of kidney function.

- (a) To compare the performance of the two machines a technician took eight specimens of blood and measured the creatinine level (in micromoles per litre) of each specimen using each machine. The results were as follows.

Specimen	1	2	3	4	5	6	7	8
Analyser A	119	173	100	99	77	121	84	73
Analyser B	106	153	83	95	69	123	84	67

The technician carried out a paired t -test and reported that there was a difference between analysers at the 5% significance level.

Verify that this is in fact the case, assuming a normal distribution.

- (b) A statistician requested that each analyser should be used repeatedly to measure a standard solution which should give a creatinine level of 90.
- Analyser A was used 7 times and gave a mean result of 93.5 and a standard deviation ($\hat{\sigma}$) of 4.7. Test at the 5% significance level whether these results could have come from a population with mean 90, assuming a normal distribution.
 - Results were also obtained for Analyser B. Explain why the results requested by the statistician are more useful in comparing the performance of the two analysers than the results in (a). What further analysis would you carry out if all the results were available to you? Justify your answer.
- (AEB)

8. It is claimed that Examiner V is more severe than Examiner W. This claim is based upon an analysis of the marks awarded by each examiner to independent random samples of scripts from a particular examination which had been marked by the two examiners. Some details of the marks awarded are as follows.

	Sample size	Sum of marks
Examiner V	25	1060
Examiner W	15	819

Investigate the claim that Examiner W awards, on average, more marks than Examiner V, assuming all marks are normally distributed with a standard deviation of 15.

Give **two** reasons why the outcome of the test may not necessarily imply that Examiner V is really more severe than Examiner W.

Suggest a more effectively designed study to investigate the claim.

9. Students on a statistics course are assessed on coursework and by a written examination. The marks obtained by a sample of 14 students were as follows (3 of the students failed to hand in any coursework).

Student	A	B	C	D	E	F	G
Coursework (%)	68	66	0	65	0	66	69
Examination (%)	53	45	67	52	43	71	37

Student	H	I	J	K	L	M	N
Coursework (%)	68	70	67	0	67	69	68
Examination (%)	43	68	27	34	79	57	54

- (a) Use the sign test on all these results to examine whether coursework marks are higher than examination marks.
- (b) Use the Wilcoxon signed-rank test, ignoring the 3 students who failed to hand in any coursework, to examine whether coursework marks are higher than examination marks.
- (c) Compare your conclusions to (a) and (b) and indicate with reasons which analysis you consider to be the more appropriate. (AEB)

10. A large consignment of similarly graded apples arrived at a company's warehouse for distribution to retail outlets. Two varieties were chosen and a random sample of each had their masses, in grams, measured. The results are tabulated below.

Variety I	110.5	89.6	89.1	85.6	115.0	98.2
	113.1	92.0	104.3	100.7	97.5	106.1
Variety II	125.6	118.3	118.0	110.8	116.5	108.7
	108.2	104.4	114.4	98.4	111.2	

Assuming that these independent samples came from underlying normal populations, use a 5% significance level to test the hypothesis that the population variances are the same.

Further, use a 5% level of significance to test the hypothesis that the population means are the same.

Later it transpired that the measuring device used to determine the above masses was inaccurate. The true masses of the 23 apples considered were all 10 grams more than the results given above. What effect do you think this information will have on the above test results and why? (Further tests are not required.) (AEB)

11. A large food processing firm is considering introducing a new recipe for its ice cream. In a preliminary trial, a panel of 11 tasters were asked to score ice cream made from both the existing and the new recipe for sweetness. The results, on a scale from 0 to 100 with the sweeter ice cream being given the higher score, were as follows.

Taster	A	B	C	D	E	F
Existing recipe	88	35	67	17	24	32
New recipe	94	49	66	82	25	96

Taster	G	H	I	J	K
Existing recipe	8	44	73	47	25
New recipe	14	56	27	44	79

Use the sign test, at the 5% significance level, to test whether the new recipe is sweeter than the existing one.

Because of the erratic nature of the scores obtained, it was decided to repeat the trial with a new panel of 10 tasters, this time giving some guidance as to the scores to allocate. Two other ice creams were tasted first. One was very sweet and the tasters were told that it had a score of 90. The other was not sweet and had a score of 10. The new trial gave the following results.

Taster	L	M	N	O	P	Q	R	S	T	U
Existing recipe	52	44	57	49	61	55	49	69	64	46
New recipe	74	65	66	47	71	55	62	66	73	59

Use a paired *t*-test, stating any necessary assumptions, to test the hypothesis that there is no difference in sweetness between the two recipes at the 1% significance level.

Discuss briefly the suitability of the choice of the sign test for the first set of data and the paired *t*-test for the second. (AEB)

12. Industrial waste dumped in rivers reduces the amount of dissolved oxygen in the water. A factory was suspected of illegally dumping waste in the river. Samples of water were taken from the river, six above the factory and eight below the factory and the dissolved oxygen content in parts per million (ppm) were as follows.

Above factory	4.9	5.1	4.7	5.0	5.3	4.6		
Below factory	3.8	4.9	4.0	3.6	5.0	3.4	3.5	3.9

Making any necessary assumptions, test, at the 5% significance level, whether

- (a) the variability of the dissolved oxygen content is the same above and below the factory,
- (b) the mean of the dissolved oxygen content is less below than above the factory. (AEB)
13. Two trainee estate agents, A and B, each valued independently a random sample of eight small properties. Their valuations, in £000s, are shown below.

Property	A	B	C	D
Trainee A	83.7	58.8	77.7	85.1
Trainee B	79.6	59.2	75.8	84.3

Property	E	F	G	H
Trainee A	91.9	66.4	69.8	48.5
Trainee B	90.1	65.2	66.9	53.8

- (a) Stating any assumptions necessary, use a paired t -test to investigate whether there is evidence that the two trainees differ in their valuations.
- (b) Repeat the test in (a) using an appropriate non-parametric test.
14. The manager of a road haulage firm records the time, in minutes, taken on six occasions for a lorry to travel from the depot to a particular customer's factory. Roadworks are due to start on the usual route so the manager decides to try an alternative route and records the times, in minutes, of eight journeys on this new route.

Old route	34	45	36	48	49	38		
Alternative route	43	35	47	39	58	40	39	51

Test, at the 5% significance level, whether there is a difference in

- (a) the variances of the times taken on the two routes,
- (b) the means of the times taken on the two routes.

One driver had taken 99 minutes on the alternative route. Investigation showed that this was due to losing his way and it was decided to exclude this result from the above tests. Comment on this decision and state what assumptions may have been violated if the result had been included in the analysis. (AEB)

15. Celebrity endorsement of a product is a common advertising technique. In one study, a randomly selected group of 125 people was shown a TV commercial involving a celebrity endorsement. A second randomly selected group of 75 people was shown the same TV commercial, but involving an unknown actress rather than the celebrity. Each of the 200 people was asked to rate on a scale from 0 (not persuaded) to 20 (totally persuaded) the effect on them of the commercial. A summary of the scores is shown below.

	With celebrity	Without celebrity
Sum of scores	1275	705
Sum of squares of scores	16485	9705

Explain why the sample variances may be used as accurate estimates of the corresponding population variances.

Hence investigate the claim that the celebrity endorsement of this particular TV commercial increases its mean persuasiveness score.

Why were no distributional assumptions necessary in carrying out your test?

16. A biologist weighs each individual mouse in a random sample consisting of ten mice and records each weight to the nearest gram. The mice are then fed on a special diet and after 15 days each mouse is weighed again and the weight to the nearest gram is recorded. The results are as follows.

Initial weight (x)	50	49	48	52	40	43	51	46	41	42
Weight after 15 days (y)	52	50	50	55	42	45	52	48	42	44

(You may assume that $\sum x^2 = 21520$ and $\sum y^2 = 23226$.)

- (a) Assuming that the results are given in random order on both occasions,
- (i) test the hypothesis that $\sigma_x^2 = \sigma_y^2$, where σ_x^2 and σ_y^2 are the variances of the populations from which these data are taken,
- (ii) examine the possibility that there has been a significant increase in mean weight over the 15 days.
- (b) If the results are given in the same order on both occasions, explain (without calculation) how this fact would alter your analysis of these data. (AEB)

17. Trace metals in drinking water affect the flavour of the water and high concentrations can pose a health hazard. The following table shows the zinc concentrations, in milligrams per 1000 litres, of water on the surface and on the river bed at each of twelve locations on a river.

Location	1	2	3	4	5	6
Surface	387	515	721	341	689	599
Bed	435	532	817	366	827	735
Location	7	8	9	10	11	12
Surface	734	541	717	523	524	445
Bed	812	669	808	622	476	387

- (a) Using a Wilcoxon signed-rank test, examine the claim that zinc concentration of water in this river is higher on the river bed than on the surface. Explain why this test is preferred here to the sign test.
- (b) If differences in zinc concentrations of water in this river may be assumed to be normally distributed, re-examine the claim in (a) using an appropriate alternative test.
18. Samples are taken from two batches of paint and the viscosity, x poise, measured. The information is summarised below.

Paint	Mean (\bar{x})	Standard deviation ($\hat{\sigma}$)	Size (n)
A	114.44	0.62	4
B	114.93	0.94	6

- Assuming normal distributions, test, at the 5% significance level, whether
- (a) the mean viscosity of Paint A is more than 114,
- (b) the standard deviations of the viscosities of the two paints are equal,
- (c) the mean viscosities of the two paints are equal. (AEB)
19. Five joints of meat were each cut in half. One half was frozen and wrapped using a standard process and the other half using a new process. The ten halves were placed in a freezer and the number of days to spoilage (which can be detected by the colour of the package) was noted for each pack.

Joint number	1	2	3	4	5
Standard process	96	194	149	185	212
New process	117	190	186	776	263

A statistician queried the observation on the new process for joint 4. The experimenter agreed that an error must have been made but said that he was certain that, for this joint, the half frozen by the new process had lasted longer than the other half.

He had used the sign test on the five joints and had accepted, at the 5% significance level, that there was no difference in the number of days to spoilage.

- (a) Confirm, by making any necessary calculations, that the sign test applied to these data does lead to the experimenter's conclusion.
- (b) Use a paired t -test on joints 1, 2, 3 and 5 to test whether there is a difference, at the 5% significance level, in the mean number of days to spoilage.
- Comment on the validity and advisability of using each of these tests on these data. A larger trial is to be carried out and, before the data are collected, you are asked to advise on which test should be used. List the advantages of each. (AEB)
20. The development engineer of a company making razors records the time it takes him to shave on seven mornings using a standard razor made by the company. The times, in seconds, were
- 217, 210, 254, 237, 232, 228, 243.

He wishes to compare the time taken by different designs of razor. He decides that rather than test all the designs himself it would be quicker to find other employees who would be willing to test one design each. As a preliminary step his assistant agrees to test the standard razor and produces the following times.

186, 219, 168, 202, 191, 184.

- Regarding the samples as coming from normal distributions,
- (a) show that there is no significant evidence of a difference between variances,
- (b) test whether the mean shaving times of the engineer and his assistant are the same.
- Advise the engineer how to proceed with his investigation. (AEB)