Chapter 16

# An introduction to nonparametric analysis

CONSIDER THE FOLLOWING SITUATION. A researcher is interested in investigating a number of possible differences in behaviour between boys and girls in the classroom. One of the hypotheses the researcher wants to test is that girls are more attentive in class than boys. Whilst the researcher has access to a class of children that are suitable for testing it is not possible to video the classroom and analyse the recordings. Although aware of the problems, the researcher decides that the only solution in this specific case is to rely on the teacher's opinion. The teacher is asked to rate each of the children in the class in terms of their attentiveness on a scale of 0–100. The teacher is not, for obvious reasons, informed of the hypothesis of the test until after completing the task. In a class of ten children the following results are produced:

| *Child* | *Teacher's rating* |
|---|---|
| Susan | 67 |
| Linda | 55 |
| John | 26 |
| Mary | 70 |
| Peter | 36 |
| Ian | 57 |
| Trevor | 32 |
| Andrew | 65 |
| Helen | 59 |
| Christine | 24 |

I have plotted the results on the 0–100 scale below and indicated the teacher's rating of each child by their initial. It does look here as though there are more of the girls at the high end of the attentiveness scale and more of the boys at the lower end. And if these data were of the sort we have been considering up to now we could compare these results on a *t* test.

```
0              10        20        30        40        50         60       70        80        90        100
-----------------------------------------------------------------------------------------------------------
Girls                         C                           L H          S M
Boys                            J      T P                  I         A
```

The problem is that, in this case, we are making an assumption about the data which may not be valid. The problem has to do with using any form of rating scale. On the basis of the numbers there appears to be a small difference between Christine and John and a large difference between Mary and Linda. Also the difference between Christine and John, of 2, is the same as the difference between Andrew and Susan. The assumption that we are making is that the teacher is using the rating scale as an interval scale, where the numbers progress in equal intervals along the scale, with the difference between consecutive numbers always the same. (See Chapter 2 on different types of numbers.)

Why cannot we assume that the teacher is using the rating scale as an interval scale? There are two reasons. First, the teacher is not a clock or a thermometer or a tape measure. These are all measuring devices that have been deliberately designed to measure in equal intervals. Human beings may not be able to judge differences in the same formal way as other devices. Second, we cannot check the teacher in the same way as we can calibrate a clock to check that it is working properly.

In reality the teacher might see Christine and John as more similar than Andrew and Susan. Also the difference between Peter and Linda could be seen as the same as the difference between John and Trevor, even though the gap between Peter and Linda is numerically greater. It is quite possible that an interval scale is not being used. An interval scale is like a tape measure made out of rigid material, the intervals are always the same. Now consider a tape measure made out of an elastic material. The teacher's 'tape measure' (the rating scale) might be stretched at certain points and squashed at others, providing quite a different scale. The teacher's rating scale could in reality look like the scale below.

```
0            10      20             30              40     50      60              70     80     90       100
-----------------------------------------------------------------------------------------------------------
Girls                       C                                     LH        S     M
Boys                          J             T P                   I       A
```

When we have doubts about whether a scale is interval or not we should assume that it is not, otherwise we risk producing erroneous conclusions in our data analysis. Unfortunately, this produces another problem. All the statistical tests that we have examined so far in the book (*z*, *t* test, ANOVA) assume that the dependent variable has been measured on an interval scale. In fact they require it, in order that means, standard deviations and other statistics can be properly calculated. Without an interval scale these calculations are meaningless.

We can see the problem of calculating statistics in the above example. To the teacher the difference between Andrew and Susan is larger than the difference between Christine and John as the 'tape measure' is stretched more between 60 and 70 than between 20 and 30. Even though both differences are written as 2 the Andrew–Susan difference is a larger '2' than the Christine–John '2'. Calculating a mean, or a standard deviation, is clearly inappropriate as the numbers do not reflect the underlying scale being used.

We can refer to two kinds of data here: that which comes from an interval scale and we can perform statistics on, and that which comes from an ordinal scale. Interval data is usually obtained from experiments where the dependent variable is measured on a formal measuring device, such as reaction times, weight loss, certain test scores and so forth. We can perform parametric tests on these data, such as *t* tests or an ANOVA. Parametric tests require interval data. The other important feature of parametric tests is that they make parametric assumptions, assumptions concerning characteristics of the underlying populations that the samples come from. These include the assumptions that populations are normally distributed and that samples come from distributions with equal variance. All the tests attempt to estimate unknown population parameters by using the sample statistics and these parameters are constrained by the assumptions. If we believe that the assumptions of the parametric tests are not met then it is inappropriate to use them as they may not test the hypothesis properly. When we are concerned that our data is not interval or that the parametric assumptions might not be valid we employ a nonparametric test instead, one that does not make the interval assumption about the scale of measurement nor any assumptions about the underlying distributions.

How can we analyse data nonparametrically? The first point to note, for the reasons cited above, is that we cannot use the actual numbers in our analysis. We cannot perform calculations on the raw data or make assumptions about the underlying population distributions. What we can assume about the numbers produced in a rating scale, such as the one the teacher used, is that these numbers allow us to rank order the data. Whilst we are unable to decide what the difference between ratings of 24 and 26 means to the teacher, what we can say is that the teacher rates the person who scored 26 as more attentive than the one who is rated at 24. Ratings are therefore ordinal data, they place the subjects into a specific order. We can look at the teacher's ratings of the children and say, from the numbers, that Mary is rated as the most attentive and Christine the least. Indeed we are able to rank order the participants on the basis of the ratings. In the table below I have ranked the children from least attentive (rank 1) to most attentive (rank 10).

| *Child* | *Teacher's rating* | *Rank* |
|---|---|---|
| Susan | 67 | 9 |
| Linda | 55 | 5 |
| John | 26 | 2 |
| Mary | 70 | 10 |
| Peter | 36 | 4 |
| Ian | 57 | 6 |
| Trevor | 32 | 3 |
| Andrew | 65 | 8 |
| Helen | 59 | 7 |
| Christine | 24 | 1 |

We can be confident that the information we have extracted from the data, the ranks, is valid as long as the data is ordinal. In analysing the ranks we will not be making any assumptions about intervals or underlying distributions. Essentially, all nonparametric analyses compare the ranks obtained in the different conditions of the independent variable. We can compare the ranks of the girls to those of the boys. If the girls receive all the high ranks and the boys the low ones then this can be used in support of the experimental hypothesis. How and when we can decide that one set of ranks is significantly different from another set of ranks lies at the heart of the various nonparametric tests. In many cases statisticians have developed nonparametric tests that can be undertaken instead of a particular parametric test when its assumptions are not met. The following table gives the nonparametric equivalents of the most popular parametric tests.

| *Number of samples* | *Parametric test* | *Nonparametric test* |
|---|---|---|
| Two (independent) | Independent *t* test | Mann–Whitney *U* test |
| Two (related) | Related *t* test | Wilcoxon signed-ranks test |
| Two or more (independent measures) | One factor independent measures ANOVA | Kruskal–Wallis test |
| Two or more (repeated measures) | One factor repeated measures ANOVA | Friedman test |

## Calculating ranks

When working out ranks it is usual in statistical analysis to give the lowest score a rank of 1 and work up through the scores, giving the highest score the top rank. In a number of tests it does not matter whether the data are ranked from the top down or from the bottom up but when it does matter the bottom up ranking is required. It is therefore a good idea to get into the habit of ranking in this way.

It often occurs that more than one subject achieves the same score in a test. In this case it is sensible to give these subjects the same rank. The way to do this is to find out how many subjects have the same raw score. We will refer to this number as $s$, so if three subjects scored the same score then $s = 3$. The rank we are about to allocate is labelled $r$. If we had ranked the first five scores before we got to the tied scores then $r = 6$. The formula for calculating the rank to give to the tied subjects is as follows:

$$\text{rank} = \frac{r + (r + 1) + \ldots + (r + s - 1)}{s}$$

With $s = 3$ and $r = 6$ then: $\text{rank} = \dfrac{6 + 7 + 8}{3} = 7$. The three subjects are all given a rank of 7.

Looking at the example it is easy to see the reason for giving out these ranks. If the numbers had been different they would have been given the ranks 6, 7 and 8. As they are the same we give then an equal share of these three ranks. The next rank to be allocated is $r + s$. In our example, the next rank to be allocated is 9.

Sometimes the rank allocated to identical values will not be a whole number. If two subjects have identical scores and the next ranking to be allocated is 6 then both subjects would be given a rank of 6.5. It is only when scores are tied in this way that we obtain ranks that are not whole numbers.

## Calculations using ranks

There are a number of calculations that we can perform with ranks. These calculations can then be used in the construction of statistical tests. Calculations with ranks rather than scores are often simpler as, say, ten scores can be anything but ten ranks are always the numbers 1 to 10. With ranks we only need to know the number of scores and then we can work out a range of rank statistics. If the number of scores is $n$, and $R$ refers to a rank, then:

1 The sum of all the ranks ($\sum R$) is $\dfrac{n(n+1)}{2}$

If we have 10 ratings ($n = 10$) and rank them then $\sum R = \dfrac{10\,(10+1)}{2} = 55$

2 The sum of the top $n_1$ ranks, where $n_1 + n_2 = n$ is $n_1 n_2 + \dfrac{n_1(n_1+1)}{2}$

Again, with $n = 10$, if we wish to sum the top 3 ranks then $n_1 = 3$, $n_2 = 7$. The sum of the top three ranks $= (3 \times 7) + \dfrac{3\,(3+1)}{2} = 27$.

3 The mean of the ranks, which is $\left(\dfrac{\sum R}{n}\right) = \dfrac{n+1}{2}$

When $n = 10$ the mean of the ranks $= \dfrac{10+1}{2} = 5.5$

4 The sum of the squared ranks ($\sum R^2$) is $\dfrac{n(n+1)(2n+1)}{6}$ as long as there are no tied ranks. For this reason there are some statistics that become less valid the more tied ranks there are.

When $n = 10$ the sum of the squared ranks $= \dfrac{10(10 + 1)(20 + 1)}{6} = 385$

(as long as none of the ranks are tied).

In the following chapters we will use these calculations in the nonparametric analysis of data.

Chapter 17

# Two sample nonparametric analysis

chapter 17

A COMPARISON BETWEEN two samples, comparing two conditions of an independent variable on a dependent variable, would normally be analysed by a *t* test if we were able to make the assumptions that the *t* test requires about the data in our samples. When we cannot make those assumptions and can only assume that the data are ordinal, we have to build a nonparametric analysis based on the rank ordering of the data. In this chapter we will consider the nonparametric equivalents of the related and independent *t* tests, namely the Mann–Whitney *U* test and the Wilcoxon signed-ranks test.

## The Mann–Whitney *U* Test (for independent samples)

The teacher's ratings of pupils' attentiveness from the previous chapter provide us with a suitable example of a two sample case with independent samples. We cannot assume that the teacher's ratings are based on an interval scale, nor can we assume any underlying distributions concerning these ratings. The statistical analysis has to be based on the ranks. The rank ordering of the participants is shown below.

| *Pupil* | *Rank* |
|---|---|
| Mary | 10 |
| Susan | 9 |
| Andrew | 8 |
| Helen | 7 |
| Ian | 6 |
| Linda | 5 |
| Peter | 4 |
| Trevor | 3 |
| John | 2 |
| Christine | 1 |

The researcher's hypothesis was that the girls would be rated as more attentive. If this was the case then we would expect the girls' ranks to be higher than the boys' ranks. Alternatively, if the boys were more attentive then they should achieve the higher ranks. And if there was no difference between the groups on attentiveness then we would expect the boys and girls to be evenly spread amongst the ranks. One way of finding out whether the groups are clustered at the top or bottom of the ranks is to find out how many participants from one group have a higher rank than each member of the other group. If we look at the table below we can see that no boys are above Mary and Susan, one above Helen, two above Linda, and five above Christine. We can do this for the boys as well and this is also shown in the table.

| *Pupil* | *Rank* | *Boys above* | *Girls above* |
|---|---|---|---|
| Mary | 10 | 0 | |
| Susan | 9 | 0 | |
| Andrew | 8 | | 2 |
| Helen | 7 | 1 | |
| Ian | 6 | | 3 |
| Linda | 5 | 2 | |
| Peter | 4 | | 4 |
| Trevor | 3 | | 4 |
| John | 2 | | 4 |
| Christine | 1 | 5 | |
| Total | | 8 | 17 |

Now if all five girls had been at the top of the rankings their total would have been 5 × 0 = 0 (as they would have had no boys above any of them) and the boys would have scored 5 × 5 = 25 (as all five of them would have had five girls above them). If the boys had all been at the top then the totals would have been reversed. With the researcher's one-tailed test we are focusing on the girls' total being small, indicating their ranks are at the top. If the girls score 0 then it seems reasonable to conclude that there is a genuine difference between the girls' and boys' ratings. If the girls scored 25 then clearly they are not ranked higher than the boys. When the score is midway between the two (12 or 13) then the two groups are mixed in their ranking. Our total for the girls is 8: is this low enough to conclude that they are genuinely higher in the ranks as a group?

The analysis we are developing here is that of the Mann–Whitney *U* test (for two independent samples). It compares the actual ranks achieved with the 'best possible ranks', that is what the group would have scored if all its members had been at the top of the ranks.

To work out the calculations we shall label the girls as Sample 1 with a sample size of $n_1 = 5$ and the boys as Sample 2, with $n_2 = 5$. If the girls had occupied the top $n_1(5)$ ranks then they would have had a rank total of 10 + 9 + 8 + 7 + 6 = 40, or as a formula:

$$n_1 n_2 + \frac{n_1(n_1 + 1)}{2} = 5 \times 5 + \frac{5(5 + 1)}{2} = 40$$

How close did the girls get to this? If we add up the actual ranks of the girls we find they achieved:

$$\sum R_1 = 10 + 9 + 7 + 5 + 1 = 32$$

The top ranks minus the actual ranks for Sample 1 is $40 - 32 = 8$. We refer to this figure as $U_1$, $U_1 = 8$.

We can also find a *U* for the boys. If they had occupied the top $n_2$ ranks then they would have had a rank total of:

$$n_1 n_2 + \frac{n_2(n_2 + 1)}{2} = 5 \times 5 + \frac{5(5 + 1)}{2} = 40$$

The boys' actual rank total is: $\sum R_2 = 8 + 6 + 4 + 3 + 2 = 23$. For the boys $U_2 = 40 - 23 = 17$.

Notice that we have arrived at the same figures of 8 and 17 as in the table above. The is because the two analyses are the same. The Mann–Whitney *U* statistic is the difference between the sample's actual ranks and the maximum ranks they could have got, with a small value of *U* indicating a group is close to the top. It is calculated using the formulae:

$$U_1 = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - \sum R_1 \qquad U_2 = n_1 n_2 + \frac{n_2(n_2 + 1)}{2} - \sum R_2$$

As a check it is worth noting that $U_1 + U_2 = n_1 n_2$.

## The significance of *U*

To decide whether there is a significant difference between the samples we need the probability of obtaining the two values of *U* when there really is no difference between the populations the samples are drawn from. What range of values, and with what probabilities, would we expect for *U* when the null hypothesis is true?

Imagine for a moment that we had only tested two girls and two boys and we had obtained a *U* for the girls of 1 and a *U* for the boys of 3. What is the probability of getting this result by chance rather than as a result of a genuine difference in populations? We can see that there are six possible ways in which we can order two boys and two girls:

| *Rank* | *Order 1* | *Order 2* | *Order 3* | *Order 4* | *Order 5* | *Order 6* |
|---|---|---|---|---|---|---|
| 4 | Girl | Girl | Girl | Boy | Boy | Boy |
| 3 | Girl | Boy | Boy | Girl | Girl | Boy |
| 2 | Boy | Girl | Boy | Girl | Boy | Girl |
| 1 | Boy | Boy | Girl | Boy | Girl | Girl |
| *U*(Girls) | 0 | 1 | 2 | 2 | 3 | 4 |
| *U*(Boys) | 4 | 3 | 2 | 2 | 1 | 0 |

When the null hypothesis is true we would expect each of these possibilities to occur with equal probability. As there are six of them each one has a probability of 1/6 or 0.167. We can now work out the probability of getting a *U* value by chance. There is only one way for the girls to get a *U* of 0, 1, 3, or 4 so each has a probability of 0.167, but two ways of getting a *U* of 2, with a probability of 0.33. In hypothesis testing we are concerned with probabilities greater than or less than a certain value. In this example it is the girls' score of 1. The probability of getting 1 or less by chance is the probability of getting 1 (0.167) plus the probability of getting 0 (0.167), which equals 0.33. If we choose the $p = 0.05$ level of significance then we can say that the probability of getting 1 or less by chance is so large (0.33) that it is not significant at $p = 0.05$.

Returning to our example of 5 boys and 5 girls, we can do the same calculation of probabilities. It is more tedious to work out as there are 252 different ways of ordering these samples but the logic is the same. When the null hypothesis is true each possibility is equally likely and we are able to work out the probability of achieving a certain value. There is one way of the girls obtaining a $U$ of 0, so this has a probability of 1/252 or 0.004, one way of obtaining a $U$ of 1 (probability = 0.004), two ways of obtaining a $U$ of 2 (probability 0.008) and so on, as shown in the table below.

| *U* | *Number of ways of getting this value by chance* | *Probability of getting this value by chance* | *Probability of getting this value or lower by chance* | |
|---|---|---|---|---|
| 0 | 1 | 0.004 | 0.004 | |
| 1 | 1 | 0.004 | 0.008 | |
| 2 | 2 | 0.008 | 0.016 | |
| 3 | 3 | 0.012 | 0.028 | |
| 4 | 5 | 0.020 | 0.048 | ⇦ $p < 0.05$ |
| 5 | 7 | 0.028 | 0.076 | |

I stopped calculating $U$ at 5 for two reasons. One, it is getting rather hard work and two, if we look at the last column, we have found out which values of $U$ occur by chance with a probability less than 0.05 (our significance level). With five boys and five girls a value of $U$ of 4 or less can be taken as significant (at $p = 0.05$) as it is occurs by chance with a probability less than the significance level.

Fortunately we do not have to work the probability tables ourselves, they have been worked out and the critical value of $U$ is listed for the level of probability chosen (see Table A.5 in the Appendix). You will see that for small values of $n_1$ and $n_2$ no critical value is given, there is a dash instead. As we saw with two boys and two girls, it is not possible with these small sample sizes to obtain a value with a probability lower than the significance level of $p = 0.05$.

In looking up the values in the table we must decide whether we are testing a one- or two-tailed prediction. In this example we have a one-tailed prediction: we test the girls' value of $U(U_1)$ as we are not interested in the boys' value. If we specify a two-tailed test then we simply select the smaller of $U_1$ and $U_2$ to compare with the table value. When looking up the value in the table it is important to remember that we want the calculated value to be equal to or *smaller* than the table value to be significant for the reason given above.

We can now look up the table value (Table A.5) to compare with the calculated value of $U$ for the girls. For a one-tailed test, with $n_1 = 5$ and $n_2 = 5$, the critical value of $U$ is 4 at a significance level of $p = 0.05$. As the girls' value is larger (8) we cannot reject the null hypothesis at this level of significance. We have not found a difference in the girls and boys in the teacher's ratings of the attentiveness.

## The distribution of *U*

When the null hypothesis is true, any variation in the ranks between the two samples will have arisen from chance factors. Clearly we want to know what differences we would expect by chance in order to make a decision about our calculated value, so we need to know the distribution of $U$ when the null hypothesis is true. As we saw above a value of $U$ is calculated for each sample. The possible values of $U$ range from 0 up to $n_1n_2$ but when the null hypothesis is true we would not expect the extreme values very often and we would expect both values of $U$ to be similar, around $\frac{n_1n_2}{2}$, the midpoint of the distribution. As we saw above it is not too difficult to work out the distributions for small values of $n_1$ and $n_2$. These values are shown in the tables. However, when the sample sizes are large (both 20 or more) then the distribution of $U$ turns out to approximate a normal distribution with:

$$\mu = \frac{n_1n_2}{2} \text{ and } \sigma = \sqrt{\frac{n_1n_2(n_1 + n_2 + 1)}{12}}$$

With these large samples, we can work out a $z$ score for the calculated value of $U$ and look up the probability in the standard normal tables (Table A.1), where $z$ is calculated as follows:

$$z = \frac{U - \frac{n_1n_2}{2}}{\sqrt{\frac{n_1n_2(n_1 + n_2 + 1)}{12}}}$$

We have to be a little careful in our use of $U$. The more tied values we have the more inaccurate the test becomes. If we do get a lot of tied values then it is worth questioning the use of the dependent variable; is it too crude a measure to differentiate between the subjects and rank order them appropriately?

## Procedure for calculating the Mann–Whitney *U* statistic

1 Rank all the scores from lowest to highest.

2 Calculate a *U* value for each sample using the following formulae:

$$U_1 = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - \sum R_1 \qquad U_2 = n_1 n_2 + \frac{n_2(n_2 + 1)}{2} - \sum R_2$$

3 Compare the smaller value with the critical value in the table (Table A.5 in the Appendix). The calculated value must be equal to or smaller than the table value for significance. (In a one-tailed test, if the sample predicted to have the highest ranks does not produce the smallest of the two *U* values then it certainly will not be significant!)

## A worked example

Two social clubs, the Hilltop Social Club and the Valley Social Club, decide to join forces and hire a coach to take them to see a Shakespearian play in the nearby city. One of the club secretaries decides to find out how much the members enjoyed the play, so on the coach home asks everyone to rate their enjoyment of the play on a 0 to 100 scale. The members of Valley Social like to see themselves as very cultured people so the club secretary predicts that they will rate their enjoyment of the play higher than the members of Hilltop. Is the secretary's prediction supported by the following data?

| *Hilltop Social Club* | *Valley Social Club* |
|---|---|
| 23 | 46 |
| 54 | 45 |
| 35 | 62 |
| 42 | 62 |
| 14 | 75 |
| 24 | 50 |
| 38 | 80 |
| | 55 |
| | 33 |

We are not going to make any assumptions about the data (except that it is ordinal) or about the underlying distributions of the populations, so will perform a Mann–Whitney *U* test.

First we rank the rating values across all conditions, taking into account ties:

| *Sample 1* | | *Sample 2* | |
|---|---|---|---|
| *Hilltop* | *Rank* | *Valley* | *Rank* |
| 23 | 2 | 46 | 9 |
| 54 | 11 | 45 | 8 |
| 35 | 5 | 62 | 13.5 |
| 42 | 7 | 62 | 13.5 |
| 14 | 1 | 75 | 15 |
| 24 | 3 | 50 | 10 |
| 38 | 6 | 80 | 16 |
| | | 55 | 12 |
| | | 33 | 4 |
| $n_1 = 7$ | $\sum R_1 = 35$ | $n_2 = 9$ | $\sum R_2 = 101$ |

We work out the two values of *U*:

$$U_1 = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - \sum R_1 = 7 \times 9 + \frac{7(7+1)}{2} - 35 = 56$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2 + 1)}{2} - \sum R_2 = 7 \times 9 + \frac{9(9+1)}{2} - 101 = 7$$

The prediction is one-tailed so the Valley value is the *U* we choose. As this is the smaller value the data do follow the direction predicted. To decide if this is significant we look up the critical value using $n_1$ and $n_2$. From Table A.5, $U = 9$, $n_1 = 7$, $n_2 = 9$, $p = 0.01$ for a one-tailed test. As the calculated value of 7 is lower than the table value we can conclude that the members of Valley Social Club gave significantly higher ratings of their enjoyment of the play than the members of Hilltop Social Club.

## The Wilcoxon signed-ranks test (for related samples)

The nonparametric test for comparing two related samples is the Wilcoxon signed-ranks test. This will be explained by considering an example. A teacher wanted to test the effect of a new television programme designed to encourage children's interest in mathematics. A group of nine children ($n = 9$) were asked to rate their interest in mathematics on a 0 to 10 scale before and after the programme. The results are shown below.

| | *Interest in mathematics* | |
|---|---|---|
| *Child* | *Before* | *After* |
| 1 | 2 | 4 |
| 2 | 5 | 8 |
| 3 | 5 | 4 |
| 4 | 2 | 8 |
| 5 | 3 | 7 |
| 6 | 2 | 9 |
| 7 | 7 | 4 |
| 8 | 7 | 7 |
| 9 | 4 | 9 |

The Wilcoxon test often has the words matched pairs in its title. This is because each score is matched in one sample with a score in the second sample, in this example the children are matched with themselves. We match the pairs in order to produce a difference score. It is not unreasonable to assume that the scores of a matched pair can be compared despite any differences in the way in which the rating scale is being used between the children. If there really is a beneficial effect of the television programme (the one-tailed prediction is correct) then we would expect the interest ratings to be consistently higher after the programme than before. This consistency should show up as a set of negative differences when we subtract the rating after the programme from the rating before the programme.

A mixture of equal positive and negative differences would indicate a lack of consistency in the differences between the samples, with some children's interest going up and others' going down after the programme. This is what we would expect with the null hypothesis. So, for significance we are looking for a consistent pattern where most difference scores are of the same sign, either mostly positive or mostly negative.

The differences are shown in the table below. Notice that child 8 produces a difference score of zero. This cannot be used to support negative differences or positive differences so we reject this participant from the analysis as the data is unhelpful to our decision making. We reduce *n* by one to 8.

| *Child* | *Before*<br>*Sample 1* | *After*<br>*Sample 2* | *Sign of*<br>*difference* | *Size of*<br>*difference* | *Rank of*<br>*difference* |
|---|---|---|---|---|---|
| 1 | 2 | 4 | – | 2 | 2 |
| 2 | 5 | 8 | – | 3 | 3.5 |
| 3 | 5 | 4 | + | 1 | 1 |
| 4 | 2 | 8 | – | 6 | 7 |
| 5 | 3 | 7 | – | 4 | 5 |
| 6 | 2 | 9 | – | 7 | 8 |
| 7 | 7 | 4 | + | 3 | 3.5 |
| 8 | 7 | 7 |  | 0 |  |
| 9 | 4 | 9 | – | 5 | 6 |

The Wilcoxon test does not just compare the sign of the differences, it also compares the size of the differences. Clearly the inconsistent differences (in our example the positive ones) are more of a problem to the research hypothesis if they are large rather than if they are small, as they are harder to explain away. The Wilcoxon test considers this by ranking the size of the differences (their absolute values) by ignoring the sign of the differences and treating them all as positive for ranking purposes. The ranks are shown in column six of the above table.

The inconsistent differences, the two positive differences (+) have ranks of 1 and 3.5. Are these small enough for us to conclude that this result is very unlikely to have occurred by chance? What is the probability of getting such ranks by chance? What we do in the Wilcoxon test is to look at the sum of the inconsistent ranks, $1 + 3.5 = 4.5$, which we call *T*. What is the probability of getting a *T* as small as 4.5 when the null hypothesis is true? We are interested in *T* being small for significance as it indicates a high degree of consistency: when *T* is zero there is no inconsistency in the ranking and the higher of each pair of scores is always in the same sample.

By chance each rank could be positive (+) or negative (−), so we have two equal possibilities for each participant when the null hypothesis is true. With eight participants that gives us $2^8 = 256$ different possibilities in total. How many of these possibilities have positive rank totals as small as or smaller than 4.5? There is only one way of achieving a positive rank total of zero (every difference is negative), so the probability of getting zero by chance is 1/256 or 0.004. There is only one way of getting a positive rank total of 1 (the lowest difference is positive and the rest are negative) and one way of a positive rank total of 2 (the second lowest difference is positive and the rest are negative). We can get a positive rank total of 3 in two ways: either the third lowest rank is the only positive one or the lowest two ranks are positive and the rest negative. We can work out further values as in the table below.

| *T* | *Number of ways of getting this value by chance* | *Probability of getting this value by chance* | *Probability of getting this value or lower by chance* | |
|---|---|---|---|---|
| 0 | 1 | 0.004 | 0.004 | |
| 1 | 1 | 0.004 | 0.008 | |
| 2 | 1 | 0.004 | 0.012 | |
| 3 | 2 | 0.008 | 0.020 | |
| 4 | 2 | 0.008 | 0.027 | |
| 5 | 3 | 0.012 | 0.039 | ⇦ $p < 0.05$ |
| 6 | 3 | 0.012 | 0.051 | |

(Slight differences between the sums of the figures in columns 3 and 4 are due to rounding of the third decimal place.)

Notice that the probability gets larger than 0.05 with a *T* of 6, but the probability of obtaining a *T* of 5 and below is less than 0.05. In our example with a *T* of 4.5 we can reject the null hypothesis at the $p = 0.05$ level of significance and conclude that there is a significant increase in the ratings of mathematical interest after the programme.

Fortunately, we do not have to work out the probability values under the null hypothesis every time. Tables of these have been constructed (Table A.6 in the Appendix). Our example was a one-tailed prediction but if we had performed a two-tailed test we would have to consider both the sum of the negative ranks and the sum of the positive ranks and taken the smaller value as *T*. The critical value of *T* for significance would also have to take into account both tails of the distribution (i.e. the chances of getting a small *T* with positive values *or* negative values) and hence be more conservative than for the one-tailed test. We have to remember that when we look up *T* we need the calculated value to be equal to or *lower* than the table value for significance.

## The distribution of *T*

For small values of *n*, less than 25, we have the tables of the critical values of *T* when the null hypothesis is true. However, the distribution of *T* approximates a normal distribution as *n* (the number of subjects) gets larger with:

$$\mu = \frac{n(n+1)}{4} \text{ and } \sigma = \sqrt{\frac{n(n+1)(2n+1)}{24}}$$

Hence, when *n* is 25 or larger, we can test the significance of *T* by calculating a *z* score and comparing it to the standard normal distribution tables, where

$$z = \frac{T - \dfrac{n(n+1)}{4}}{\sqrt{\dfrac{n(n+1)(2n+1)}{24}}}$$

We must be cautious in the use of *T* when we are dealing with data that includes more than a few tied ranks as it is unlikely to be appropriate to use. In this case we should examine the measure of the dependent variable and see if we can make it more sensitive, to produce more distinction between the difference scores and hence fewer tied ranks.

## Procedure for calculating the Wilcoxon signed-ranks test

1 Calculate a difference score for each subject, the score in Sample 1 minus the score in Sample 2. When a subject has a zero difference score we remove the subject from the analysis and reduce the size of *n* by 1 in each case.

2 Rank the difference scores from lowest to highest, ignoring the sign.

3 Sum the ranks of the positive differences ($\sum R_+$) and sum the ranks of the negative differences ($\sum R_-$). The smaller of the positive and negative sums of ranks is the calculated value of $T$. (If a one-tailed prediction has been made the smaller of the two values should be consistent with the prediction. If it is not then it certainly is not significant.) It is worth checking that $\sum R_+ + \sum R_- = \frac{n(n+1)}{2}$, as both sides of the equation add up to the sum of the ranks.

4 Compare the calculated value of $T$ with the critical value in the table (Table A.6), using $n$ to find the correct value, at the chosen level of significance. The calculated value of $T$ must be equal to or smaller than the value in the table for significance.

## A worked example

An interview panel of ten interviewers were asked to rate the two final candidates on a scale of 1 to 20 in terms of their suitability for a vacant post. Is one candidate rated significantly higher than the other by the interviewers?

| *Interviewer* | *Candidate 1* | *Candidate 2* |
|---|---|---|
| 1 | 14 | 10 |
| 2 | 17 | 7 |
| 3 | 12 | 14 |
| 4 | 16 | 6 |
| 5 | 14 | 14 |
| 6 | 10 | 4 |
| 7 | 17 | 10 |
| 8 | 12 | 4 |
| 9 | 6 | 11 |
| 10 | 18 | 6 |

We shall make no assumptions about the data or the population distributions except that the data is ordinal and so perform a Wilcoxon signed-ranks test to examine the hypothesis. First we work out the difference scores (Candidate 1 – Candidate 2) for each participant (interviewer). Zero differences are excluded from the analysis and the differences are ranked on their size as in the table below.

| *Interviewer* | *Candidate 1* | *Candidate 2* | *Sign of difference* | *Size of difference* | *Rank* |
|---|---|---|---|---|---|
| 1 | 14 | 10 | + | 4 | 2 |
| 2 | 17 | 7 | + | 10 | 7.5 |
| 3 | 12 | 14 | − | 2 | 1 |
| 4 | 16 | 6 | + | 10 | 7.5 |
| 5 | 14 | 14 | | 0 | |
| 6 | 10 | 4 | + | 6 | 4 |
| 7 | 17 | 10 | + | 7 | 5 |
| 8 | 12 | 4 | + | 8 | 6 |
| 9 | 6 | 11 | − | 5 | 3 |
| 10 | 18 | 6 | + | 12 | 9 |

Interviewer 5 is rejected from the analysis as the difference score is zero, so the number of participants, *n*, is now 9. We next calculate the sum of ranks for the positive differences and the negative differences.

$$\sum R_{+} = 2 + 7.5 + 7.5 + 4 + 5 + 6 + 9 = 41$$

$$\sum R_{-} = 1 + 3 = 4$$

No specific prediction is being made so it is a two-tailed test. We take the smaller value for the calculated value of $T$, so $T = 4$. At the $p = 0.05$ level of significance, with $n = 9$, the table value of $T$ is 5 for a two-tailed test. As the calculated value of $T$ is smaller than the table value we can say that the interviewers significantly favour Candidate 1 in their ratings.

Chapter 18

# One factor ANOVA for ranked data

chapter 18

When the data for analysis is not from an interval scale or the assumptions of the ANOVA are not met, we have to perform a nonparametric test. With a one factor design where we are analysing more than two samples we perform either the Kruskal–Wallis test, if the samples are independent, or the Friedman test, if the samples are related. These tests are the nonparametric equivalents of the one factor independent measures ANOVA and the one factor repeated measures ANOVA.

## Kruskal–Wallis test (for independent measures)

The Kruskal–Wallis test performs an analysis that is very similar to an analysis of variance on the ranks. The test is performed when the assumptions of the parametric ANOVA cannot be made. An example of such data occurs in the following illustration. A researcher was interested in differences in attractiveness and the selection of candidates for jobs. As well as examining female attractiveness a number of experiments were undertaken on male attractiveness. One of the questions considered was whether different types of facial hair led to different judgements of male attractiveness by women. A female personnel officer in a large company agreed to rate photographs of men's faces on attractiveness on a 0 to 50 scale, with a high value indicating a high level of attractiveness. Out of a large pool of photographs of different men, five men with beards, five men with moustaches and five clean shaven men were randomly selected. (The photographs in the pool had been matched on age, hairstyle and tidiness.) If we examine the data below can we observe an effect of facial hair on the attractiveness judgements?

| *Facial hair* | | | | | |
|---|---|---|---|---|---|
| *Beard* | | *Moustache* | | *Clean shaven* | |
| *Rating* | *Rank* | *Rating* | *Rank* | *Rating* | *Rank* |
| 5 | 1 | 9 | 3 | 23 | 10 |
| 6 | 2 | 16 | 6 | 28 | 12 |
| 10 | 4 | 19 | 8 | 35 | 13 |
| 15 | 5 | 25 | 11 | 44 | 14 |
| 17 | 7 | 20 | 9 | 47 | 15 |
| | $T_1 = 19$ | | $T_2 = 37$ | | $T_3 = 64$ |

As we have independent measures on the factor *facial hair* we rank all the scores in the data, irrespective of condition. These ranks are shown above. If there was no difference between the conditions we would expect the ranks to be evenly scattered across them. If there is an effect of the independent variable we would expect there to be systematic differences between the conditions, such as all the high ranks in one condition. We need to find a way of measuring the clustering of similar ranks within specific conditions.

If we had been performing an ANOVA we would work out $F$, where $F = \dfrac{MS_{bet.conds}}{MS_{error}}$. However, in the Kruskal–Wallis test we calculate a slightly different statistic on the ranks, called $H$, where

$$H = \frac{SS_{bet.conds}}{MS_{total}}$$

We use the usual formulae for working out sums of squares and mean square but as we are dealing with ranks we can work out much simpler formulae for them in our calculation of $H$.

We know that $SS_{total} = \sum X^2 - \dfrac{(\sum X)^2}{N}$ but as we are dealing with ranks ($R$) rather than scores ($X$), with no tied ranks we can replace some of the terms in the formulae:

$$\sum X^2 = \sum R^2 = \frac{N(N+1)(2N+1)}{6} \text{ and}$$

$$\sum X = \sum R = \frac{N(N+1)}{2}$$

From this we have that $(\sum X)^2 = (\sum R)^2 = \dfrac{N^2(N+1)^2}{4}$

Substituting these formulae for ranks into the formula for the total sums of squares we get:

$$SS_{total} = \frac{N(N+1)(N-1)}{12}$$

As the total degrees of freedom in the data is $N - 1$, then:

$$MS_{total} = \frac{N(N+1)}{12}$$

This means that <u>whatever</u> data we collect, the $MS_{total}$ of the ranks will be a fixed value for $N$. We can see why $H$ is calculated rather than $F$ here. $MS_{total}$ provides us with a fixed value of 'average' variance that we get with $N$ ranks regardless of the effect of the independent variable. If we measure the between conditions variability against this fixed value we can see how much greater the variability between the conditions actually is. For example, with an $N$ of 15 the $MS_{total}$ will always be 20 (when there are no tied ranks).

From the usual formula for sums of squares:

$$SS_{bet.conds} = \frac{\sum T^2}{n} - \frac{(\sum X)^2}{N}$$

When we substitute the ranks formula for $(\sum X^2)$ we get:

$$SS_{bet.conds} = \frac{\sum T^2}{n} - \frac{N(N+1)^2}{4}$$

where $T$ is the total of the ranks in a condition and $\sum T^2 = T_1^2 + T_2^2 + \ldots + T_k^2$, $k$ being the number of conditions, and $n$ the number of scores in each condition.

From these calculations we can work out a relatively simple formula for $H$:

$$H = \frac{12}{N(N+1)} \times \frac{\sum T^2}{n} - 3(N+1)$$

$H$ is a formula which tells us how much variability there is between the conditions (the sums of squares) compared to the 'average' variance in the ranks. As $MS_{total}$ is always fixed for $N$ the important degrees of freedom is that between conditions, $df = df_{bet.conds} = k - 1$ as $H$ is influenced by the number of conditions under study.

In the *facial hair* example, $N = 15$, $n = 5$, $k = 3$, $T_1 = 19$, $T_2 = 37$, $T_3 = 64$ and

$$H = \frac{12}{15(15+1)} \times \frac{19^2 + 37^2 + 64^2}{5} - 3(15+1) = 10.26,\ df = 2$$

So the variability between the ranks of the conditions (the between conditions sums of squares) is 10.26 times larger than the 'average' variance (the total mean squares) in the ranks.

## Unequal sample sizes

Just like the independent measures ANOVA we can have a different number of subjects in each condition. If this is the case then the formula for $H$ is:

$$H = \frac{12}{N(N+1)} \times \sum \frac{T^2}{n} - 3(N+1)$$

where $\sum \frac{T^2}{n} = \frac{T_1^2}{n_1} + \ldots + \frac{T_k^2}{n_k}$, and $n_1$ to $n_k$ are the number of subjects in conditions 1 to $k$.

## The distribution of $H$

We can ask why we find $H$ rather than $F$ when we have ranks. There are a number of reasons. As noted above, $MS_{total}$ is a fixed value for $N$. In our example, with $N = 15$, it will always be 20 regardless of the number of conditions and the variability between them. We can therefore use $MS_{total}$ as a benchmark with which to compare the actual variability of the ranks between the conditions. If there is no variability between the conditions $SS_{bet.conds}$ will be zero as the total of ranks within each condition will be the same, and if there is lots of variability between the conditions then $SS_{bet.conds}$ will be large, as the similar ranks will cluster within specific conditions. But how large is large? This is why we compare it to $MS_{total}$. In our example, when we calculate them separately we find $SS_{bet.conds} = 205.2$ and $MS_{total} = 20$, so $SS_{bet.conds}$ is over 10 times larger than $MS_{total}$, implying that the variability between conditions is not random, and indicates an effect of *facial hair* on the judgements of attractiveness. We now need to find the distribution of $H$ under the null hypothesis to find the value of $H$ required for significance.

This is where we can see how useful $H$ is as a statistic. It turns out that the distribution of $H$ is known, as $H$ closely approximates a distribution called the chi-square ($\chi^2$) distribution, which is known. As long as we have at least 5 scores in each condition $H$ is accurate to two decimal places.[14] We shall be looking at the $\chi^2$ distribution in more detail in the next chapter but it is worth noting the following: $z$ is a deviation from a mean divided by a standard deviation. If we square $z$ then $z^2$ is a squared deviation divided by a variance. A distribution of $z^2$ is a $\chi^2$ distribution. A sum of $z^2$s is also a $\chi^2$ distribution, and a sum of $z^2$s is a sums of squares divided by a variance, which is what we have with $H$.

Clearly, the size of $H$ depends on the number of conditions and so we must look up the significance of $H$ using $df = df_{bet.conds} = k - 1$. Fortunately the $\chi^2$ distribution has been worked out for different degrees of freedom. In our example with $df = 2$ we can look up the appropriate value of $\chi^2$. From the tables of the $\chi^2$ distribution, Table A.7 in the Appendix, $\chi^2 = 9.21$, $p = 0.01$, $df = 2$. As our calculated value of $H$ is larger than the table value we can conclude that there is a significant difference (at $p = 0.01$) between the different conditions of facial hair in the judgements of attractiveness.

## Tied ranks

If we have tied ranks we really should use the original formulae on the ranks for $SS_{bet.conds}$ and $MS_{total}$. When we use the formula for $H$ with tied ranks the calculated value for $H$ will tend to be smaller than it really should be and we might miss a significant difference. To compensate we may wish to employ a correction $C$, where: $C = 1 - \frac{\sum t}{N^3 - N}$, and the corrected value of $H_c = \frac{H}{C}$. In the formula for $C$, $N$ is the total number of scores in the data (as above) but $\sum t = \sum(t_i^3 - t_i)$, which means that for each group of tied ranks $i$, $t_i$ is their number. Consider the following ranks: 1, 2.5, 2.5, 4, 5, 7, 7, 7, 9, 10. Here there are two sets of tied ranks: 2 at 2.5 and 3 at 7, so:

$\sum t = \sum(t_i^3 - t_i) = (2^3 - 2) + (3^3 - 3) = 6 + 24 = 30$, giving $C = 1 - \frac{30}{10^3 - 10} = 0.97$, so our calculated value of $H$ would be divided by 0.97 which would give us a slightly higher value for comparison with the table value for significance.

However, it is only when the calculated value is close to significance that this would arise and we should always pay attention to results that only just miss significance. In most cases we can work out the value of $H$ using the simpler formula without worrying about tied ranks, as long as there are not too many of them.

## Procedure for calculating the Kruskal–Wallis test

1 Rank all the scores in the experiment, irrespective of condition.

2 Add up the ranks for each condition to produce a rank total for each condition: $T_1, \ldots, T_k$ where $k$ is the number of conditions.

3 Calculate $H$ using the formula: $H = \frac{12}{N(N+1)} \times \sum \frac{T^2}{n} - 3(N+1)$, which allows for different numbers of subjects in each condition. $N$ is the total number of subjects and $n_1, \ldots, n_k$ are the number of subjects in the $k$ conditions.

4 The calculated value of $H$ must equal or exceed the table value of $\chi^2$ with $k - 1$ degrees of freedom at the chosen level of significance to reject the null hypothesis. Table A.7 in the Appendix gives the critical values of the $\chi^2$ distribution.

## A worked example

A group of 18 people who found it hard to relax agreed to take part in a test of three relaxation techniques, a pill to aid restfulness, hypnosis and exercise.

After a week employing the technique the participants were asked to rate their ability to relax on a 50 point scale (ranging from 0 much worse, 25 no change, through to 50 much better than before). Six people undertook the pill methods, five hypnosis and seven exercise. Is there an effect of *relaxation method* on their ratings?

The data are shown in the table below with their ranks.

| *Condition 1* | | *Condition 2* | | *Condition 3* | |
|---|---|---|---|---|---|
| *Pill* | *Rank* | *Hypnosis* | *Rank* | *Exercise* | *Rank* |
| 14 | 2.5 | 29 | 11 | 44 | 18 |
| 10 | 1 | 38 | 15 | 30 | 12 |
| 18 | 4 | 27 | 9 | 40 | 16 |
| 22 | 6 | 25 | 7 | 28 | 10 |
| 14 | 2.5 | 26 | 8 | 33 | 13 |
| 20 | 5 | | | 35 | 14 |
| | | | | 42 | 17 |
| $n_1 = 6$ | $T_1 = 21$ | $n_2 = 5$ | $T_2 = 50$ | $n_3 = 7$ | $T_3 = 100$ |

We now calculate *H*:

$$H = \frac{12}{N(N+1)} \times \sum \frac{T^2}{n} - 3(N+1)$$

$$= \frac{12}{18(18+1)}\left(\frac{21^2}{6} + \frac{50^2}{5} + \frac{100^2}{7}\right) - 3(18+1)$$

$$H = \frac{12}{342}(73.5 + 500 + 1428.57) - 57 = 13.25$$

Degrees of freedom, $df = k - 1 = 3 - 1 = 2$.

From the $\chi^2$ tables, at $p = 0.01$, $\chi^2 = 9.21$, $df = 2$. As the calculated value of 13.25 is greater than the table value (Table A.7 in the Appendix) we can conclude that there is a significant difference (at $p = 0.01$) between the relaxation methods on the participants' ratings.

## Post hoc multiple comparisons following the Kruskal–Wallis test

We can perform a post hoc multiple comparison test after a significant Kruskal–Wallis test in a similar manner to a Tukey test. From Chapter 12 we can write: Tukey's honestly significant difference = $q \times$ standard error, where $q$ is the Studentized range statistic. We use a variation of this called the Nemenyi test to compare pairs of samples following a Kruskal–Wallis test, where, instead of comparing the sample means, we compare the sample rank totals. Futhermore the standard error (SE) is now calculated as follows: $\mathrm{SE} = \sqrt{\frac{n(nk)(nk+1)}{12}}$, where $k$ is the number of conditions and $n$ the number of scores in each condition. We look up the value of $q$ in Table A.4 using the significance level (usually 0.05), the number of samples $k$ and, in this case, the infinity line of the degrees of freedom ($\infty$). If the difference between a pair of rank totals (e.g. $T_1$ and $T_2$) is greater than $q \times \mathrm{SE}$ then the difference between the conditions is significant at the chosen significance level.

The problem with the Nemenyi test is that it requires all samples to be of the same size ($n$). With unequal sample sizes we can use Dunn's test with

$$\mathrm{SE} = \sqrt{\frac{N(N+1)}{12}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

where $n_i$ and $n_j$ are the sample sizes of the two conditions.[15] We must compare the mean rank for our conditions rather than the rank totals (e.g. for condition 1 the mean rank will be $\frac{T_1}{n_1}$). A difference in mean ranks must be greater than $Q \times \mathrm{SE}$. $Q$ is the statistic for differences in mean ranks and the values of $Q$ are found in the table overleaf for the different values of $k$ at the significance levels of 0.05 and 0.01.[16]

| *Critical values of the Q statistic* | | |
|---|---|---|
| $k$ | $p = 0.05$ | $p = 0.01$ |
| 2 | 1.960 | 2.576 |
| 3 | 2.394 | 2.936 |
| 4 | 2.639 | 3.144 |
| 5 | 2.807 | 3.291 |
| 6 | 2.936 | 3.403 |
| 7 | 3.038 | 3.494 |
| 8 | 3.124 | 3.570 |
| 9 | 3.196 | 3.635 |
| 10 | 3.261 | 3.692 |

In the above worked example, the mean ranks are:

$$\bar{R}_1 = \frac{T_1}{n_1} = \frac{21}{6} = 3.50, \quad \bar{R}_2 = \frac{T_2}{n_2} = \frac{50}{5} = 10.00,$$

$$\bar{R}_3 = \frac{T_3}{n_3} = \frac{100}{7} = 14.29$$

For a significance level of $p = 0.05$, with three conditions ($k = 3$), $Q = 2.394$. For condition 1 versus condition 2: $\text{SE} = \sqrt{\frac{18(18+1)}{12}\left(\frac{1}{6} + \frac{1}{5}\right)} = 3.23$, so $Q \times \text{SE} = 7.73$. Hence the difference in mean ranks for conditions 1 and 2 of 6.5 is not significant at $p = 0.05$. For condition 1 versus condition 3, SE = 2.97 and $Q \times \text{SE} = 7.11$. The difference in mean ranks of 10.79 is significant at $p = 0.05$. Finally, for conditions 2 and 3, SE = 3.13, giving $Q \times \text{SE} = 7.49$. Hence their difference in mean ranks of 4.29 is not significant at $p = 0.05$.

## The Friedman test (for related samples)

The Friedman test is a nonparametric test that can be performed when we cannot make the assumptions necessary for the parametric one factor repeated measures ANOVA. In this test the analysis is performed on the ranks. As there are repeated measures the scores are ranked within each subject rather than across all the scores. In the example below six personnel officers were asked to rate, on a 0–10 scale, colours of business suits in terms of *professional image*. Three suit colours were chosen for the conditions: brown, black and blue.

| | *Suit colour* | | | | | | |
|---|---|---|---|---|---|---|---|
| | *Brown* | | *Black* | | *Blue* | | *Rank* |
| *Participant* | *Rating* | *Rank* | *Rating* | *Rank* | *Rating* | *Rank* | *total* |
| 1 | 5 | 1 | 8 | 2 | 9 | 3 | 6 |
| 2 | 4 | 1 | 6 | 3 | 5 | 2 | 6 |
| 3 | 3 | 1 | 4 | 2 | 9 | 3 | 6 |
| 4 | 5 | 2 | 4 | 1 | 8 | 3 | 6 |
| 5 | 4 | 1 | 5 | 2 | 6 | 3 | 6 |
| 6 | 5 | 2 | 3 | 1 | 7 | 3 | 6 |
| | | $T_1 = 8$ | | $T_2 = 11$ | | $T_3 = 17$ | |

If there was no difference in the samples we would expect the ranks to be evenly spread amongst the conditions. If there is an effect of the independent variable then we would expect similar ranks to cluster in specific conditions. In the above example most the Rank 1s are in the 'brown' condition, most of the Rank 2s in the 'black' condition and most of the Rank 3s in the 'blue' conditions so we would expect our statistic to indicate a significant difference between the conditions.

With the one way repeated measures ANOVA we work out *F* but in the Friedman test we work out $\chi_r^2$ which is a chi-square on the ranks, where

$$\chi_r^2 = \frac{SS_{bet.conds}}{MS_{with.subjs}}$$

Notice from the above table that when we rank the data for each participant there is no variation between the subjects ($SS_{bet.subjs} = 0$) as the rank total for each subject is always the same, in our case they all add up to 6. So all the variation in the ranks is within the subjects ($SS_{total} = SS_{with.subjs}$). We can see from this the similarity of the Kruskal–Wallis and the Friedman tests.

The formula for $SS_{with.subjs}$ is:

$$SS_{with.subjs} = \sum X^2 - \frac{\sum T_S^2}{k}$$

As we are dealing with ranks, if there are no tied ranks:

$$\sum X^2 = \sum R^2 = \frac{nk(k+1)(2k+1)}{6} \text{ and } \sum T_S^2 = \frac{nk^2(k+1)^2}{4}$$

These formulae for ranks are slightly different for those shown in Chapter 16 as we are ranking within each subject, not across all the scores in the experiment. We can now replace the ANOVA formula for scores with the replacement formulae for ranks.

$$SS_{with.subjs} = \frac{nk(k+1)(2k+1)}{6} - \frac{nk(k+1)^2}{4}$$

Simplifying the formula we get:

$$SS_{with.subjs} = \frac{nk(k+1)(k-1)}{12}$$

The degrees of freedom within the subjects is $n(k-1)$, so:

$$MS_{with.subjs} = \frac{k(k+1)}{12}$$

This is a fixed value for each value of $k$. With three conditions, as in our example, $MS_{with.subjs}$ will always be 1.

The sums of squares between the conditions can be worked out from the following formula:

$$SS_{bet.conds} = \frac{\sum T^2}{n} - \frac{(\sum X)^2}{nk}$$

where $nk = N$ the total number of scores and $T_1, \ldots, T_k$ are the totals of the scores in each condition.

As we have ranks, assuming no ties, we can replace $\sum X$ with $\dfrac{nk(k+1)}{2}$ in the formula and $T$ becomes the total of the ranks in a condition:

$$SS_{bet.conds} = \frac{\sum T^2}{n} - \frac{nk(k+1)^2}{4}$$

And finally,

$$\chi_r^2 = \frac{12}{nk(k+1)}\sum T^2 - 3n(k+1) \text{ with } k - 1 \text{ degrees of freedom}$$

In our *business suit colour* example, $n = 6$, $k = 3$, $T_1 = 8$, $T_2 = 11$, $T_3 = 17$:

$$\chi_r^2 = \frac{12}{6 \times 3(3+1)}(8^2 + 11^2 + 17^2) - 3 \times 6(3+1) = 7, \text{ with } df = 2$$

## The distribution of $\chi_r^2$

As with the Kruskal–Wallis $H$ statistic, $\chi_r^2$ compares the between conditions sums of squares to a fixed value, the 'average' variance in the ranks. If the null hypothesis is true we would expect the variability between conditions to be zero. When the null hypothesis is false we would expect the between conditions variability to be large. Our definition of large in this case is taken relative to the fixed value $MS_{with.subjs}$.

Again, as with the Kruskal–Wallis $H$ statistic, $\chi_r^2$ approximates the $\chi^2$ distribution, with the appropriate distribution found using the degrees of freedom between the conditions, $k - 1$. However, when there are few conditions and a small number of subjects ($k = 3$ and $n < 10$ or $k = 4$ and $n < 5$) then the $\chi^2$ distribution is not such a good fit for $\chi_r^2$.[17] In these cases we must work out the various probabilities for $\chi_r^2$ when the null hypothesis is true. Let us take, for example, the case where $k = 3$ and $n = 3$. For each subject there are six ways in which the ranks 1, 2, and 3 could be arranged across the three conditions, so for three subjects there are $6 \times 6 \times 6 = 216$ ways of arranging the ranks in total. The maximum value of $\chi_r^2$ is 6. This occurs when, for every subject, the same rank is in the same condition. This can occur in six ways. This gives us a probability of 6/216 or $p = 0.028$ of obtaining a value of $\chi^2_r = 6$ by chance. The next largest value of $\chi^2_r$ is 4.67 and the probability of obtaining this value is larger than 0.05. Thus, with $k = 3$ and $n = 3$ only $\chi^2_r = 6$ is significant at $p = 0.05$. The critical values of $\chi^2_r$ for small sample sizes are shown in Table A.8 in the Appendix.

The example of the business suits is a small sample case with $k = 3$ and $n = 6$. The table value for $p < 0.05$, is 7. As the calculated value of $\chi^2_r = 7$ is the same we can conclude that there is a significant effect (at $p = 0.05$) of business suit colour on the judgements of professional image.

We must be careful if there are a lot of tied ranks in the data as this might make the analysis inaccurate. Fortunately as we are ranking within each subject this is not likely to occur often. However, if there are more than a few tied ranks it is worth considering whether it is possible to make the dependent variable more sensitive and reduce the number of ties.

## Procedure for calculating the Friedman test

1 Set out the data with the subjects as rows and the conditions as columns.

2 Rank each of the $n$ subjects' scores separately, from lowest to highest.

3 Work out the rank total ($T$) for each condition: $T_1, \ldots, T_k$, where $k$ is the number of conditions.

4 Calculate $\chi^2_r$ using the following formula:

$$\chi^2_r = \frac{12}{nk(k+1)} \sum T^2 - 3n(k+1) \text{ with } k - 1 \text{ degrees of freedom.}$$

5 The calculated value of $\chi^2_r$ must be larger than or equal to the appropriate table value of $\chi^2$ (Table A.7 in the Appendix) or larger or equal to the value of $\chi^2_r$ in the small samples table (Table A.8).

## A worked example

Ten people stay at a hotel where they eat all their meals. On one day they are asked to rate the quality of food for the three meals, breakfast, lunch and dinner, on a scale of 0 to 100 (from bad to good). Is there a difference between the three meals in their rated quality?

The results of the ratings are shown in the table below. The data is assumed only to be ordinal and no assumptions are made about the underlying distributions.

| | *Breakfast* | | *Lunch* | | *Dinner* | |
|---|---|---|---|---|---|---|
| *Participant* | *Rating* | *Rank* | *Rating* | *Rank* | *Rating* | *Rank* |
| 1 | 50 | 1 | 58 | 3 | 54 | 2 |
| 2 | 32 | 2 | 37 | 3 | 25 | 1 |
| 3 | 60 | 1 | 70 | 3 | 63 | 2 |
| 4 | 41 | 1 | 66 | 3 | 59 | 2 |
| 5 | 72 | 1 | 73 | 2 | 75 | 3 |
| 6 | 37 | 3 | 34 | 2 | 31 | 1 |
| 7 | 39 | 1 | 48 | 3 | 44 | 2 |
| 8 | 25 | 2 | 29 | 3 | 18 | 1 |
| 9 | 49 | 2 | 54 | 3 | 42 | 1 |
| 10 | 51 | 1 | 63 | 2 | 68 | 3 |
| $n = 10$ | | | | | | |
| $k = 3$ | | $T_1 = 15$ | | $T_2 = 27$ | | $T_3 = 18$ |

The ratings are ranked for each participant as in the table above and the total of the ranks in each condition is calculated. We now calculate $\chi_r^2$:

$$\chi_r^2 = \frac{12}{nk(k+1)}\sum T^2 - 3n(k+1)$$

$$= \frac{12}{10 \times 3 \times 4}(15^2 + 27^2 + 18^2) - 3 \times 10 \times 4$$

$$= 0.1 \times 1278 - 120 = 7.8 \quad \text{with } df = k - 1 = 3 - 1 = 2$$

From Table A.7, $p = 0.05$, $df = 2$, $\chi^2 = 5.99$. As our calculated value of $\chi_r^2$ is larger than the table value of $\chi^2$ we can conclude that there is a significant difference between the meals in terms of the ratings of meal quality.

## Post hoc multiple comparisons following a Friedman test

We can employ a Nemenyi test, a variation of the Tukey test, to undertake pairwise comparisons of the conditions after a significant Friedman test. In this test we compute a standard error (SE) using the formula:

$$\text{SE} = \sqrt{\frac{nk(k+1)}{12}}$$

We then look up the appropriate value of the Studentized range statistic $q$ from Table A.4 using the chosen significance level (e.g. 0.05), the number of conditions $k$, and the infinity row for the degrees of freedom ($\infty$). If a difference in the rank totals of two conditions is larger than $q \times \text{SE}$ then we can claim a significant difference between the conditions.

In the above example, with $n = 10$ and $k = 3$, $\text{SE} = \sqrt{\frac{10 \times 3 \times (3+1)}{12}} = 3.16$ and $q = 3.31$ at $p = 0.05$. From these values we work out that $q \times \text{SE} = 10.46$. As $T_1 = 15$, $T_2 = 27$ and $T_3 = 18$ we can conclude the following. There is a significant difference between conditions 1 and 2 (as the rank total difference of 12 is greater than 10.46), but the differences between conditions 1 and 3 (rank total difference of 3) and between conditions 2 and 3 (rank total difference of 9) are not significant at $p = 0.05$.

Details on how to calculate the Kruskal–Wallis and Friedman tests using the SPSS computer statistical package can be found in Chapter 13 of Hinton *et al.* (2004).