

Chapter 13 The Wilcoxon signed rank test

If a man will begin with certainties, he shall end in doubts; but if he will be content to begin with doubts, he shall end in certainties.

Sir Francis Bacon, The Advancement of Learning

The Avonford Star

Public votes new shopping centre the tops – at last!

It's official – the public have voted the newly opened Regent Shopping Centre a definite asset to the town, despite the considerable and prolonged disruption the building works have caused. Town Centre Manager, Paul Clay, expressed relief at the outcome, which further showed that 25% of those asked voted it “the tops.”

Details of the survey may be found on the Avonford Star website.

Up to now the statistical tests you have used require the assumption that the underlying population is Normally distributed, or that you had large samples. Life is not always quite so simple, and there are many situations where data simply do not behave like that.

Take the poll on the shopping centre as an example; the data here are a collection of opinions. In fact the journalist asked a question that required more than just a “Yes” or “No” answer. People were asked to comment on the statement ‘The new Regent Shopping Centre is an asset to the town.’ and given a choice of answers from “Definitely an asset, it's the tops.” to “No definitely not an asset, it's awful.”

While it is undoubtedly impressive that 25% of those asked went so far as to say it was the tops, you should be asking yourself “I wonder what percentage thought it awful”. Darell Huff, who wrote a famous and humorous book called “How to Lie with Statistics”, might well have said of the newspaper report, “*It’s the figures that aren’t there that are important.*”

Here are all the responses, recorded on a data collection sheet.

| Person number | Definitely an asset, it's the tops. | Yes, it's a good asset. | Yes it is an asset, it's OK | I have no opinion | Not really an asset. | I don't think it an asset at all, is not very good | No definitely not an asset, it's awful. |
|---------------|-------------------------------------|-------------------------|-----------------------------|-------------------|----------------------|--|---|
| 1 | | | | | X | | |
| 2 | | X | | | | | |
| 3 | X | | | | | | |
| 4 | | | | X | | | |
| 5 | | | | | X | | |
| 6 | | | | X | | | |
| 7 | X | | | | | | |
| 8 | | | | | | X | |
| 9 | | | X | | | | |
| 10 | | X | | | | | |
| 11 | | | X | | | | |
| 12 | X | | | | | | |

A statistical hypothesis test would help here. It would tell you how likely it is that the result has arisen by pure chance, if it is not very likely you can sit up and take notice of the outcome.

The hypothesis you would want to test would be: “The public considers the new shopping centre an asset.”

The first problem is what figures to use in this situation? The responses are words not numbers, but you can solve this by allocating a number to each response, with 1 for “Definitely an asset, it’s the tops” to 7 for “No definitely not an asset, it’s awful”.

This gives you a *rating scale* of responses, which substitutes a number for each opinion.

? **What kind of scale is this?**



Always be cautious with rating scales as people are not like rulers, watches, or thermometers. The difference between opinions cannot be neatly measured in the same way as the differences between two lengths, times or temperatures.

Is the difference between “Yes, it’s a good asset” (2) and “Yes it is an asset, it’s OK” (3) the same amount of difference of opinion as between “I have no opinion” (4) and “Not really as asset.” (5)? Is it sensible to measure the difference between two opinions numerically?

Is the opinion “Yes it is an asset, it’s OK”, rated as 3, three times weaker than the opinion rated as 1, “Definitely an asset, it’s the tops.” ?

There is nevertheless an order to the different opinions. Think of other situations where there is an order, but doing arithmetic with the numbering does not make sense.

The journalist chose 12 people, at random, and from the data sheet their responses were

5 2 1 4 5 4 1 6 3 2 3 1

At first glance, as you have a small sample, this looks like a candidate for the t test, with which you are familiar, where you might find whether or not the mean is 4, representing the “no opinion” response.

? **What assumption underlies the t-test?**

Why would the t-test not be appropriate here?

In this situation an alternative test is the Wilcoxon Rank Sum Test. This is an example of a non-parametric test. It requires no assumption about an underlying Normal distribution.



The term non-parametric tests has generally come to mean tests where there is no assumption that the underlying population has a Normal distribution population. A common mistake is to think such tests require no assumptions, but this is not true. The assumption made for the Wilcoxon test is that the variable being tested is symmetrically distributed about the median, which would also be the mean. Remember too that it is still vitally important that your sample has been randomly chosen from the population.

Frank Wilcoxon's revolutionary idea was not to use the data themselves for a test, but to use the ranks. There are several Wilcoxon tests.

- The Wilcoxon signed-rank test can be used for a single sample, such as you have with the shopping centre survey.
- The Wilcoxon signed-rank test can also be used for paired data (see Chapter **)
- Where you have two unrelated samples, which you wish to compare, the Wilcoxon rank sum test is used (see Chapter **); this test is very similar to the Mann Whitney test.

? Is asking only 12 people too small a sample?

Do you think that asking 12 people about the new shopping centre gave too small sample in the circumstances?

The Wilcoxon signed rank test for a single sample.

The responses people gave about the shopping centre are represented by numerical ratings:

5 2 1 4 5 4 1 6 3 2 3 1

where having no opinion equates to 4, the median value. A quick 'eyeball' test shows that none of those questioned thought it was awful and only one person thought it not very good, so a first impression is that people generally approve.

If you start by assuming that in the population there is no opinion one way or the other, and that people's responses are symmetrically distributed about 'no opinion', you can test the hypothesis that people think the shopping centre is an asset, with the null hypothesis that people have no opinion about it, their response being the median value 4.

SETTING UP THE HYPOTHESIS TEST

As people who think it an asset will give a rating of less than 4, the null and alternative hypotheses can be stated as follows.

H_0 : the median response is 4

H_1 : the median response is less than 4

1 tail test

Significance level 5%

? Why is this a one tail test?

CALCULATING THE TEST STATISTIC

There are 4 steps to the test. Some people find the mnemonic DRASTIC helps them to remember it.

It's not so **DRAS**tiC after all:
Difference **R**ank **S**um **C**ompare

1. Find the difference between each value and the median.

Using a table is the easiest way to organise the figures:

| rating | rating – median (4) |
|--------|---------------------|
| 5 | 5 – 4 = 1 |
| 2 | 2 – 4 = -2 |
| 1 | 1 – 4 = -3 |
| 4 | 4 – 4 = 0 |
| 5 | 5 – 4 = 1 |
| 2 | 2 – 4 = -2 |
| 4 | 4 – 4 = 0 |
| 1 | 1 – 4 = -3 |
| 6 | 6 – 4 = 2 |
| 3 | 3 – 4 = -1 |
| 2 | 2 – 4 = -2 |
| 1 | 1 – 4 = -3 |

2. Ignore the zeros and rank the absolute values of the remaining scores.

Ignore the signs, start with the **smallest** difference and give it rank 1.

Where two or more differences have the same value find their mean rank, and use this.

| rating | rating - 4 | absolute value | ranking | + | - |
|--------|------------|----------------|---------|-----|------|
| 5 | 1 | 1 | 2 | 2 | |
| 2 | -2 | 2 | 5.5 | | 5.5 |
| 1 | -3 | 3 | 9 | | 9 |
| 4 | 0 | 0 | ignore | | |
| 5 | 1 | 1 | 2 | 2 | |
| 2 | -2 | 2 | 5.5 | | 5.5 |
| 4 | 0 | 0 | ignore | | |
| 1 | -3 | 3 | 9 | | 9 |
| 6 | 2 | 2 | 5.5 | 5.5 | |
| 3 | -1 | 1 | 2 | | 2 |
| 2 | -2 | 2 | 5.5 | | 5.5 |
| 1 | -3 | 3 | 9 | | 9 |
| | | | Total | 9.5 | 45.5 |

Ignoring the zeros leaves you with 10 values.

In calculating the ranks, you have four values of 1, three of 2, and three of 3.

The three values of 1 occupy the ranks 1, 2, and 3, which has a mean of 2, so they are all given the rank of 2..

The four values of 2 occupy the ranks 4,5, 6, and 7 which has a mean of 5.5, and so they are ranked 5.5.

The three values of 3 have ranks 8, 9, 10 which has a mean of 9; they are all ranked 9.

3. Sum the ranks of the positive differences, W_+ , and sum the ranks of the negative differences W_- .

W_+ , the sum of the ranks of the positive differences is $2 + 2 + 5.5 = 9.5$

W_- , the sum of the ranks of the negative differences is $5.5 + 9 + 5.5 + 9 + 2 + 5.5 + 9 = 45.5$

Now check that $W_+ + W_-$ are the same as $\frac{1}{2} n(n+1)$, where n is the number in the sample (having ignored the zeros). In this case $n = 10$.

$$\frac{1}{2} n(n+1) = \frac{1}{2} \times 10 \times 11 = 55$$

$$W_+ + W_- = 9.5 + 45.5 = 55$$

INTERPRETING THE TEST STATISTIC

4. Compare the test statistic with the critical value in the tables.

If the null hypothesis were true, and the median is 4, you would expect W_+ and W_- to have roughly the same value.

There are two possible test statistics here, $W_+ = 9.5$ and $W_- = 45.5$, and you have to decide which one to use.

You are interested in W_+ , the sum of the ranks of ratings greater than 4. W_+ is much less than W_- which suggests that more people felt the shopping centre was an asset.

It could also suggest that those who expressed a negative view expressed a very strong one, with lots of high numbers in the ratings. Now you need to compare the value of W_+ , the test statistic, with the critical value from the table.

| 1-tail 2-tail | 5% | 2½% | 1% | ½% |
|------------------|-----------|-----|----|----|
| n | | | | |
| 2 | - | - | - | - |
| 3 | - | - | - | - |
| 4 | - | - | - | - |
| 5 | 0 | - | - | - |
| 6 | 2 | 0 | - | - |
| 7 | 3 | 2 | 0 | - |
| 8 | 5 | 3 | 1 | 0 |
| 9 | 8 | 5 | 3 | 1 |
| 10 | 10 | 8 | 5 | 3 |
| 11 | 13 | 10 | 7 | 5 |
| 12 | 17 | 13 | 9 | 7 |
| 13 | 21 | 17 | 12 | 9 |
| 14 | 25 | 21 | 15 | 12 |
| 15 | 30 | 25 | 19 | 15 |

Figure.1 critical values for the Wilcoxon single sample and paired sample tests

Given that W_+ is small the key question becomes “Is W_+ significantly smaller than would happen by chance?” The table helps you decide this by supplying the critical value. For a sample of 10, at the 5% significance level for a 1 tailed test, the value is 10. As W_+ is 9.5, which is less than this, the evidence suggests that we can reject the null hypothesis.

Your conclusion is that the evidence shows, at the 5% significance level, that the public thinks the new shopping centre is an asset to the town.

Choosing the test statistic.

You need to be careful to choose the appropriate test statistic for the problem you are tackling.

For a **two** tailed test the test statistic is the smaller of W_+ and W_- .

For a **one** tailed test, where the alternative hypothesis is that the median is **greater** than a given value, the test statistic is W_- .

For a **one** tailed test, where the alternative hypothesis is that the median is **less** than a given value, the test statistic is W_+ .

Historical note: Frank Wilcoxon (1892 – 1965)

Frank Wilcoxon was an outstanding chemist, whose interest in statistics first started while studying fungicides, when he and colleagues studied Fisher's newly published Statistical Methods for Research Workers. In 1945 he published his paper setting out the rank-sum and signed-rank tests which are still named after him.



Roger – the picture is at http://www.swlearning.com/quant/kohler/stat/biographical_sketches/bio21.1.html

His background was colourful. A keen cyclist and motor cyclist, he and his twin sister were born in an Irish castle, to wealthy American parents. He grew up in the States, ran away to sea, worked as an oil worker and tree surgeon, and attended a military academy before finally entering college, aged 26, to read chemistry.

e The sign test

A very simple test of opinion about the shopping centre would assume that people are equally likely to approve or not, with probability 0.5 . You could count up the number of responses greater than 3 and work out, using the binomial distribution, how likely you are to have this number. This is known as the sign test, it is very useful, but it doesn't make full use of all the information you have available as it ignores the magnitudes of the differences from the median. The Wilcoxon test, on the other hand, makes use of this information, and as a result is more powerful.

Example 13.1

Student satisfaction surveys ask students to rate a particular course, on a scale of 1 (poor) to 10 (excellent). In previous years the replies have been symmetrically distributed about a median of 4. This year there has been a much greater on-line element to the course, and staff want to know how the rating of this version of the course compares with the previous one.

14 students, randomly selected, were asked to rate the new version of the course and their ratings were as follows:

1 3 6 4 8 2 3 6 5 2 3 4 1 2

Is there any evidence at the 5% level that students rate this version any differently?

SOLUTION

SETTING UP THE HYPOTHESIS TEST

The null hypothesis is that there is no change in the rating given by the students, and so the median is still 4. The alternative hypothesis is that the median is not 4.

H_0 : The median is 4.

H_1 : The median is not 4.

The alternative hypothesis is not that the median is greater than 4, nor is it that it is less than 4, just that it is not 4, so it is a two tailed test.

2 tail test

Significance level 5%

Assumption: the ratings are still symmetrically distributed.

CALCULATING THE TEST STATISTIC

Taking your 'DRAS*t*iC steps', you find the difference, rank, sum and compare our test statistic with the critical value. Much of the working is best done in a table.

| rating | frequency | rating - 4 | absolute value | ranking | + | - |
|--------|-----------|------------|----------------|---------|-------------------|----------------------|
| 1 | 2 | -3 | 3 | 11 | | $2 \times 11 = 22$ |
| 2 | 3 | -2 | 2 | 7 | | $3 \times 7 = 21$ |
| 3 | 3 | -1 | 1 | 2.5 | | $3 \times 2.5 = 7.5$ |
| 4 | 2 | 0 | ignore | | | |
| 5 | 1 | 1 | 1 | 2.5 | 2.5 | |
| 6 | 2 | 2 | 2 | 7 | $2 \times 7 = 14$ | |
| 8 | 1 | 3 | 3 | 11 | 11 | |
| | | | | Total | | |

To find the ranks:

Notice there are four with value 1 so these have the ranks 1, 2, 3, 4, with mean 2.5.

The five with value 2, have ranks 5,6,7,8 and 9, mean 7.

The three 3's have ranks 10, 11, 12, with mean 11

$$W_- = (2 \times 11) + (3 \times 7) + (3 \times 2.5) = 22 + 21 + 7.5 = 50.5$$

$$W_+ = 2.5 + (2 \times 7) + 11 = 2.5 + 14 + 11 = 27.5$$

Check that $W_+ + W_- = \frac{1}{2} n(n+1)$ $n = 12$ since we have ignored the two zeros.

$$W_+ + W_- = 27.5 + 50.5 = 78$$

$$\frac{1}{2} n(n+1) = \frac{1}{2} \times 12 \times 13 = 78$$



Do not be tempted to omit this check. It will tell you whether or not you have calculated W_+ and W_- correctly.

INTERPRETING THE TEST STATISTIC

Now you need to choose your test statistic.

For a two tailed test it is the smaller of W_+ and W_- .

Here it is $W_+ = 27.5$.

From the tables, the critical value for a 2 tailed test at the 5% significance level, for a sample of 12 is 13.

The test statistic, $W_+ = 27.5$ which is > 13 , you accept the null hypothesis at the 5% significance level.

You can write your conclusion as:

The evidence shows that the introduction of a greater on-line element to the course has had no significant effect on the ratings, at the 5% level.

? When you write “accept the null hypothesis at the 5% significance level” what do you mean by “at the 5% significance level”? What does it tell you?

Exercise 4.1

1. New recruits to a call centre are given initial training in answering customer calls. Following this training they are independently assessed on their competence, and are rated on a score of 1 to 10, 1 representing 'totally incompetent' to 10 'totally competent'. It is usual for the trainees' scores to be symmetrically distributed about a median of 6. A new trainer has been appointed and the scores of her first 19 trainees are:

6 5 6 9 7 3 4 6 7 2 9 8 7 4 5 6 9 5 7

Is there evidence at the 5% level that the new trainer has made any difference?

2. It is recommended that women should not consume more than 70g of fat per day. A random sample of 13 student nurses at St Clares were asked to estimate as carefully as possible how much fat they ate on one particular day. The results were (measured in grams)

85 120 45 95 100 50 65 85 105 125 65 49

Is there evidence at the 5% level that student nurses at St Clares are consuming more fat than they should?

3. The numbers of waders feeding in Sampford creek is being logged. At the same time last year, the median number seen each day at a certain time was 8. Records for the last 16 days show the following numbers of birds:

7 9 12 14 7 7 15 12 10 7 7 12 9 7 15 15

Is there evidence at the 5% level that the number of waders feeding at the creek at this time of year has increased?

4. A dental nurse has carried out a large survey of patient stress levels, asking patients to rate their stress level while having treatment on a score of 1 to 10; 1 being 'stress free' to 10 being 'unbearably stressful'. The responses were symmetrically distributed with a median of 4.

The dentist plays Radio 1 while treating patients as he thinks this relaxes them. The dental nurse now suggests changing to Classic fm for a week, and asks 17 patients at random to rate their stress levels. The results are:

9 1 5 3 4 2 8 3 6 6 4 4 3 1 10 7 2

Is there evidence at the 1% level that listening to Classic fm has reduced patients' stress levels?

5. It has been suggested that British office workers are not taking their full lunch breaks, but spend part of them working at their desks, the median lunch 'hour' now being 34 minutes. An office supervisor in a large purchasing department, intrigued by this, noted the time spent away from their desks at lunchtime by 10 randomly chosen staff, without their knowledge. The data were (minutes):

55 20 31 12 18 35 28 16 14 32

She was shocked when she looked at the data. Does it suggest, at the 5% level, that her staff are taking even less than 34 minutes away from their desks at lunchtime?

6. The local paper reckons the rate of pay for baby sitters is now £6 an hour. Teenagers dispute this. A random sample of teenagers who babysit are asked what their pay rates are per hour; this required some calculation as most were paid per night.

The data are (£ per hour) 5 5 4 3 7 3 5 3 8 2

Is there any evidence at the 5% level that the median is not £6?

Is there any evidence at the 5% level that the median rate of pay of the teenagers is less than £6 an hour?

Answers to Discussion points.

? ***What kind of scale is this?***

The scale of opinions represented by numbers, a rating scale, is an ordered scale where the order has meaning, but the differences between each opinion are not measurable.

There are other types of scale, where the differences do have meaning. Although it is not sensible to describe a day of 20° C as being twice as hot as one of 10° C the difference in the two temperatures can be measured, and has a meaning. This scale is called an interval scale.

If you are measuring snakes you can not only say that one snake is 1cm longer than another but also describe one snake as being twice, or half, or any ratio of, the length of another. Such a scale is called a ratio scale.

Other examples of ordered scales are where items are judged on merit and placed in order, or where items are grouped, eg companies placed into one of 3 groups according to turnover, the 3 groups being Large, Medium and Small.

? **What assumption underlies the t-test?**

The t test needs the assumption that the underlying population is Normally distributed. That's clearly not the case here with a set of responses.

? Is asking only 12 people too small a sample?

Not necessarily, provided it is representative of the population being sampled. It is surprising how many pieces of research are carried out with small samples, which is what makes the Wilcoxon test so useful. Make a habit of looking for the sample size, and even more importantly how the sample was chosen, when reading any report.

? Why is this a one tail test?

Because you are only interested in the median being less than 4. If you were interested in it being different from 4, either more or less than 4, it would have been a two tailed test.

? When you write “accept the null hypothesis at the 5% significance level”

Statisticians don't like egg on their face, and there is still a possibility, albeit small, that the null hypothesis is not correct. There is only a 5%, or 1 in 20, chance that it is wrong, but by always stating your conclusion with a significance level you are leaving that possibility open, and your reputation intact. You do not have to stick with a 5% significance level. In some circumstances you might want to be more certain and might go for a 1% significance level.

Key Points.

The Wilcoxon signed rank test (single sample)

1. This is used for testing the null hypothesis that the population median of a random variable is equal to a given value M . It is assumed that the variable is symmetrically distributed about its median.
2. To calculate the test statistic:
 - Find the difference between each value and the median;

- Ignore the zeros. Rank the remaining scores. Ignoring the signs, give the lowest rank to the smallest difference. Where two or more differences have the same value find their mean rank, and use this.
- Sum the ranks of the positive differences, W_+ , and sum the ranks of the negative differences W_- . Check that $W_+ + W_- = \frac{1}{2} n(n+1)$, where n is the number in the sample having ignored the zeros.
- For a two tailed test the test statistic is the smaller of W_+ and W_- .
For a one tailed test where the alternative hypothesis is that the median is greater than a given value, the test statistic is W_- .
For a one tailed test where the alternative hypothesis is that the median is less than a given value, the test statistic is W_+ .
- Compare the test statistic, W , with the critical value in the tables; the null hypothesis is rejected if W is less than or equal to the critical value.

Answers to Exercises.

Exercise 4.1

1. Two tailed test. $n = 15$ $W = W_- = 57.5$, critical value 25 Do not reject H_0 , no difference.
2. One tailed test. $n = 12$ $W = W_- = 21.5$. Critical value 17. Do not reject H_0 , no difference.
3. One tailed test. $n = 16$ $W = W_- = 27$ critical value 35. Reject H_0 , more waders.
4. One tailed test. $n = 14$ $W = W_+ = 64.5$ critical value 25. Do not reject H_0 , no difference.

5. One tailed test. $n=10$ $W = W_+ = 10$ critical value 10. Reject H_0 ; Less lunchtime.

6. Two tailed test. $n = 10$ $W = W_+ = 8$, critical value 8. Reject H_0 ; Different rate.

One tailed test. $n=10$ $W = W_+ = 8$ critical value 10. Reject H_0 ; Less pay.