Contents 45

Non-parametric Statistics

Learning outcomes

You will learn about some significance tests which may be used when we are not willing to assume that the data come from a probability distribution of a particular type. In the first Section you will learn about the one-sample case and, in the second Section, you will learn about the two-sample case.

Non-parametric Tests for a Single Sample

45.1

Introduction

In earlier Workbooks we have looked at a number of significance tests, such as the t-test, the F-test and the χ^2 test. All of these depend on the assumption that the data are drawn from normal distributions. Although the normal distribution is very common, and this is what gave it its name, there are clearly cases when the data are not drawn from normal distributions and there are other cases when we might simply be unwilling to make that assumption. It is possible to make tests for cases where the data are drawn from some other specified distribution but sometimes we are unable or unwilling to say what kind of distribution it is. In such cases we can use tests which are designed to do without an assumption of a specific distribution. Sometimes these tests are called *distribution-free* tests, which seems like a very sensible name, but usually they are called *non-parametric* tests because they do not refer to the parameters which distinguish members of a particular family of distributions. For example, a t-test is used to consider questions concerning the statistic μ, the mean of a normal distribution, which distinguishes one normal distribution from another. In a non-parametric test we do not have a parametric formula for the form of the underlying probability distribution.

Prerequisites

Before starting this Section you should ...

- be familiar with the general ideas and terms of significance tests
- be familiar with t-tests
- understand and be able to apply the binomial distribution

Learning Outcomes

On completion you should be able to ...

- explain what is meant by a nonparametric test and decide when such a test should be used
- use a sign test
- use and interpret the results of a Wilcoxon signed rank test

1. Non-parametric tests

Sometimes it is possible to measure a quantity and express the measurements numerically in such a way that meaningful arithmetic can be done. For example, if you measure three spacers and determine that they are 1 mm 2 mm and 3 mm spacers you can certainly assert that $1 + 2 = 3$ in the sense that the combination of the 1 mm and 2 mm spacers are effectively the same as the 3 mm spacer. There are occasions when data may be expressed numerically but doing arithmetic leads to nonsensical conclusions. Suppose, for example that as a manager, you are asked to assess the work of three colleagues, John, Tony and George. You might come to the conclusion that overall George is the "best" worker, followed in order by John and the Tony. You may present the results as follows:

Name	**Rating**
George	1
John	2
Tony	3

In this case, if you assert that $1 + 2 = 3$ you may be interpreted as saying that the combined work of George and John is equivalent to the work of Tony. This, of course, is in complete contradiction to the way you have rated the work of your colleagues! Remember that the appearance of numbers does not imply that you can do meaningful arithmetic. In fact, meaningless arithmetic, while giving a piece of work the appearance of careful analysis can (and almost certainly will) be totally misleading in any conclusions reached. In other statistical problems, the variable measured may allow meaningful arithmetic but we might not feel able to assume that it follows a probability distribution of any particular type. In particular, we might not be willing to assume that it has a normal distribution. In cases such as these we use tests which do not depend on the assumption of a particular distribution, unlike t-tests, F-tests etc., where a normal distribution is assumed. Tests which do not require such distributional assumptions are called non-parametric tests.

Very often, the non-parametric procedure described in this Workbook may be thought of as direct competitors of the t-test and F-test when normality can be assumed and we will compare the performance of parametric and non-parametric methods under conditions of normality and non-normality. In general terms, you will find the non-parametric methods fail to use all of the information that is available in a sample and as a consequence they may be though to as less efficient than parametric methods. Essentially, you should remember that in cases where it is difficult or impossible to justify normality but it is known that the underlying distribution is continuous, non-parametric methods remain valid while parametric methods may not. You should also bear in mind that in terms of practical application it may be difficult to decide whether to use parametric or non-parametric tests since both the t-test (and the F-test) are relatively insensitive to small departures from normality.

Our work concerning non-parametric tests begins with the sign test.

2. The sign test

The sign test is used to test hypotheses concerning the **median** of a continuous distribution. Some authors use the symbol θ to represent to median of the distribution - remember that μ is used to represent the mean of a distribution. We will use the θ notation for the median throughout this Workbook. Remember that in the case of a normal distribution the mean is equal to the median and so the sign test can be used to test hypotheses concerning the mean of a normal distribution. The test procedure is straightforward to describe. The usual null hypothesis is

$$H_0 \; : \; \theta = \theta_0$$

As you might expect, the alternative hypothesis can take one of three forms

$$H_1 \; : \; \theta \neq \theta_0 \qquad H_1 \; : \; \theta > \theta_0 \qquad H_1 \; : \; \theta < \theta_0$$

Now suppose the sample taken from a population is $X_1, X_2, X_3, \ldots, X_n$. We form the differences

$$X_i - \theta_0 \qquad i = 1 \ldots n$$

Assuming that the null hypothesis is true, each difference $X_i - \theta_0$ is equally likely to be positive or negative and in order to test a particular pair of hypotheses we need only test the number of plus signs (say). Under the null hypothesis this is a value of the binomial distribution with parameter $p = \dfrac{1}{2}$. In order to decide whether we should reject a null hypothesis, we can calculate probabilities directly from the binomial distribution (see HELM 37) using the formula

$$P(X = r) = \binom{n}{r} q^{n-r} p^r = \binom{n}{r} (1-p)^{n-r} p^r$$

or by using the normal approximation to the binomial distribution.

The following Examples and Tasks illustrate the test procedure.

Example 1

The compressive strength of insulating blocks used in the construction of new houses is tested by a civil engineer.

The engineer needs to be certain at the 5% level of significance that the median compressive strength is at least 1000 psi. Twenty randomly selected blocks give the following results:

Observation	Compressive Strength	Observation	Compressive Strength	Observation	Compressive Strength	Observation	Compressive Strength
1	1128.7	6	718.4	11	1167.1	16	1153.6
2	679.1	7	787.4	12	1387.5	17	1423.3
3	1317.2	8	1562.3	13	679.9	18	1122.6
4	1001.3	9	1356.9	14	1323.2	19	1644.3
5	1107.6	10	1153.2	15	788.4	20	737.4

Test (at the 5% level of significance) the null hypothesis that the median compressive strength of the insulting blocks is 1000 psi against the alternative that it is greater.

Solution

The hypotheses are

$$H_0 : \theta \quad = 1000$$

$$H_1 : \theta \quad > 1000$$

Comp. Strength	Sign	Comp. Strength	Sign	Comp. Strength	Sign	Comp. Strength	Sign
1128.7	$+$	718.4	$-$	1167.1	$+$	1153.6	$+$
679.1	$-$	787.4	$-$	1387.5	$+$	1423.3	$+$
1317.2	$+$	1562.3	$+$	679.9	$-$	1122.6	$+$
1001.3	$+$	1356.9	$+$	1323.2	$+$	1644.3	$+$
1107.6	$+$	1153.2	$+$	788.4	$-$	737.4	$-$

We have 14 plus signs and the required probability value is calculated directly from the binomial formula as

$$\begin{aligned}
P(X \geq 14) &= \sum_{r=14}^{20} \binom{20}{r} \left(\frac{1}{2}\right)^{20-r} \left(\frac{1}{2}\right)^{r} \\
&= \frac{20.19.18.17.16.15}{1.2.3.4.5.6}\left(\frac{1}{2}\right)^{20} + \frac{20.19.18.17.16}{1.2.3.4.5}\left(\frac{1}{2}\right)^{20} + \frac{20.19.18.17}{1.2.3.4}\left(\frac{1}{2}\right)^{20} \\
&\quad + \frac{20.19.18}{1.2.3}\left(\frac{1}{2}\right)^{20} + \frac{20.19}{1.2}\left(\frac{1}{2}\right)^{20} + \frac{20}{1}\left(\frac{1}{2}\right)^{20} + \left(\frac{1}{20}\right)^{20} \\
&= \left(\frac{1}{2}\right)^{20}(38760 + 15504 + 4845 + 1140 + 190 + 20 + 1) \\
&= 0.05766
\end{aligned}$$

Since we are performing a one-tailed test, we must compare the calculated value with the value 0.05.

Since $0.05 < 0.05766$ we conclude that we cannot reject the null hypothesis and that on the basis of the available evidence, we cannot conclude that the median compressive strength of the insulating blocks is greater than 1000 psi.

Example 2

A certain type of solid rocket fuel is manufactured by bonding an igniter with a propellant. In order that the fuel burns smoothly and does not suffer either "flame-out" or become unstable it is essential that the material bonding the two components of the fuel has a shear strength of 2000 psi. The results arising from tests performed on 20 randomly selected samples of fuel are as follows:

Observation	Shear Strength	Observation	Shear Strength	Observation	Shear Strength	Observation	Shear Strength
1	2128.7	6	1718.4	11	2167.1	16	2153.6
2	1679.1	7	1787.4	12	2387.5	17	2423.3
3	2317.2	8	2562.3	13	1679.9	18	2122.6
4	2001.3	9	2356.9	14	2323.2	19	2644.3
5	2107.6	10	2153.2	15	1788.4	20	1737.4

Using the 5% level of significance, test the null hypothesis that the median shear strength is 2000 psi.

Solution

The hypotheses are $\qquad H_0 : \theta = 2000 \qquad H_1 : \theta \neq 2000$

We determine the signs associated with each observation as shown below and perform a two-tailed test.

Shear Strength	Sign	Shear Strength	Sign	Shear Strength	Sign	Shear Strength	Sign
2128.7	$+$	1718.4	$-$	2167.1	$+$	2153.6	$+$
1679.1	$-$	1787.4	$-$	2387.5	$+$	2423.3	$+$
2317.2	$+$	2562.3	$+$	1679.9	$-$	2122.6	$+$
2001.3	$+$	2356.9	$+$	2323.2	$+$	2644.3	$+$
2107.6	$+$	2153.2	$+$	1788.4	$-$	1737.4	$-$

We have 14 plus signs and the required probability value is calculated directly from the binomial formula:

$$\begin{aligned}
P(X \geq 14) &= \sum_{r=14}^{20} \binom{20}{r} \left(\frac{1}{2}\right)^{20-r} \left(\frac{1}{2}\right)^{r} \\
&= \frac{20.19.18.17.16.15}{1.2.3.4.5.6}\left(\frac{1}{2}\right)^{20} + \frac{20.19.18.17.16}{1.2.3.4.5}\left(\frac{1}{2}\right)^{20} + \frac{20.19.18.17}{1.2.3.4}\left(\frac{1}{2}\right)^{20} \\
&\quad + \frac{20.19.18}{1.2.3}\left(\frac{1}{2}\right)^{20} + \frac{20.19}{1.2}\left(\frac{1}{2}\right)^{20} + \frac{20}{1}\left(\frac{1}{2}\right)^{20} + \left(\frac{1}{2}\right)^{20} \\
&= \left(\frac{1}{2}\right)^{20}(38760 + 15504 + 4845 + 1140 + 190 + 20 + 1) = 0.05766
\end{aligned}$$

Since we are performing a two-tailed test, we must compare the calculated value with 0.025.

Since $0.025 < 0.05766$ we cannot reject the null hypothesis on the basis of the evidence and conclude that the median shear strength is not significantly different from 2000 psi.

Now do the following Task.

Task

A certain type of solid rocket fuel is manufactured by binding an igniter with a propellant. In order that the fuel burns smoothly and does not suffer either "flame-out" or become unstable it is essential that the material bonding the two components of the fuel has a shear strength of 2000 psi. The results arising from tests performed on 10 randomly selected samples of fuel are as follows.

Observation	Shear Strength	Observation	Shear Strength
1	2128.7	6	1718.4
2	1679.1	7	1787.4
3	2317.2	8	2562.3
4	2001.3	9	2356.9
5	2107.6	10	2153.2

Using the 5% level of significance, test the null hypothesis that the median shear strength is 2000 psi.

Your solution

Answer
The hypotheses are

$$H_0 : \theta = 2000 \qquad H_1 : \theta \neq 2000$$

We determine the signs associated with each observation as shown below and perform a two-tailed test.

Shear Strength	Sign	Shear Strength	Sign
2128.7	$+$	1718.4	$-$
1679.1	$-$	1787.4	$-$
2317.2	$+$	2562.3	$+$
2001.3	$+$	2356.9	$+$
2107.6	$+$	2153.2	$+$

We have 7 plus signs and the required probability value is calculated directly from the binomial formula as

$$\begin{aligned} P(X \geq 7) &= \sum_{r=7}^{10} \binom{10}{r} \left(\frac{1}{2}\right)^{10-r} \left(\frac{1}{2}\right)^{r} \\ &= \frac{10.9.8}{1.2.3}\left(\frac{1}{2}\right)^{10} + \frac{10.9}{1.2}\left(\frac{1}{2}\right)^{10} + \frac{10}{1}\left(\frac{1}{2}\right)^{10} + \left(\frac{1}{2}\right)^{10} \\ &= \left(\frac{1}{2}\right)^{10}(120 + 45 + 10 + 1) \simeq 0.172 \end{aligned}$$

Since we are performing a two-tailed test, we must compare the calculate value with the value 0.025. Since $0.025 < 0.172$ we cannot reject the null hypothesis on the basis of the available evidence and we cannot conclude that the median shear strength is different to 2000 psi.

3. The sign test for paired data

Very often, experiments are designed so that the results occur in matched pairs. In these cases the sign test can often be applied to decide between two hypotheses concerning the data. Performing a sign test involves counting the number of times when, say, the first score is higher then the second – designated by a "$+$" sign and the number of times that the first score is lower than the second – designated by a "$-$" sign.

Ties

It is, of course, possible that in some cases, the scores will be equal, that is, they are said to be **tied**. There are two ways in which tied scores are dealt with.

Method 1

Ties may be counted as minus signs so that they count *for* the null hypothesis. The logic of this is that equal scores cannot be used as agents for change.

Method 2

Ties may be discounted completely and not used in any analysis performed. The logic of this is that ties can sometimes occur because of the way in which the data are collected. Throughout this Workbook, any ties occurring will be discounted and ignored in any subsequent analysis.

Essentially, we take paired observations, say $(X_{1i}, X_{2i}),\ i = 1 \ldots n$, from a continuous population and proceed as illustrated below.

Example 3

In an experiment concerning gas cutting of steel for use in off-shore structures, 48 test plates were prepared. Each plate was cut using both oxy-propane cutting and oxy-natural gas cutting and, in each case, the maximum Vickers hardness near the cut edge was measured. The results were as follows.

Plate	Propane	Nat. gas	Plate	Propane	Nat. gas	Plate	Propane	Nat. gas
1	291	296	17	295	272	33	325	313
2	315	281	18	327	300	34	312	323
3	318	310	19	329	309	35	318	317
4	319	312	20	319	291	36	314	317
5	312	320	21	327	317	37	324	334
6	296	297	22	317	279	38	319	293
7	331	319	23	289	282	39	305	294
8	316	290	24	321	301	40	305	332
9	321	301	25	299	259	41	306	330
10	283	259	26	325	302	42	303	296
11	316	327	27	307	337	43	321	311
12	342	306	28	291	320	44	328	338
13	302	259	29	312	300	45	302	292
14	312	314	30	335	330	46	324	278
15	293	268	31	319	307	47	327	352
16	346	300	32	310	307	48	329	295

Use a sign test to test the null hypothesis that the mean difference between the hardnesses produced by the two methods is zero against the alternative that it is not zero. Use the 1% level of significance.

Solution

We are testing to see whether there is evidence that the media difference between the hardnesses produced by the two methods is zero. The null and alternative hypotheses are:

$$H_0 \ : \ \theta_{\text{differences}} = 0 \qquad\qquad H_1 \ : \ \theta_{\text{differences}} \neq 0$$

We perform a two-tailed test. The signs of the differences (propane minus natural gas) are shown in the table below.

Plate	Prop.	N.gas		Plate	Prop	N.gas		Plate	Prop	N.gas	
1	291	296	$-$	17	295	272	$+$	33	325	313	$+$
2	315	281	$+$	18	327	300	$+$	34	312	323	$-$
3	318	310	$+$	19	329	309	$+$	35	318	317	$+$
4	319	312	$+$	20	319	291	$+$	36	314	317	$-$
5	312	320	$-$	21	327	317	$+$	37	324	334	$-$
6	296	297	$-$	22	317	279	$+$	38	319	293	$+$
7	331	319	$+$	23	289	282	$+$	39	305	294	$+$
8	316	290	$+$	24	321	301	$+$	40	305	332	$-$
9	321	301	$+$	25	299	259	$+$	41	306	330	$-$
10	283	259	$+$	26	325	302	$+$	42	303	296	$+$
11	316	327	$-$	27	307	337	$-$	43	321	311	$+$
12	342	306	$+$	28	291	320	$-$	44	328	338	$-$
13	302	259	$+$	29	312	300	$+$	45	302	292	$+$
14	312	314	$-$	30	335	330	$+$	46	324	278	$+$
15	293	268	$+$	31	319	307	$+$	47	327	352	$-$
16	346	300	$+$	32	310	307	$+$	48	329	295	$+$

There are 34 positive differences and 14 negative differences.The probability of getting 14 or fewer negative differences, if the probability that a difference is negative is 0.5, is

$$\begin{aligned} P(X \leq 14) &= \sum_{r=0}^{14} \binom{48}{r} \left(\frac{1}{2}\right)^r \left(\frac{1}{2}\right)^{48-r} = \sum_{r=0}^{14} \binom{48}{r} \left(\frac{1}{2}\right)^{48} \\ &= 0.0027576 \end{aligned}$$

We can find this value approximately by using the normal approximation. The required mean and variance are $48 \times 0.5 = 24$ and $48 \times 0.5 \times 0.5 = 12$ repectively. So we calculate the probability that a normal random variable with mean 24 and variance 12 is less than 14.5.

$$\begin{aligned} P(X \leq 14) \approx P(Y < 14.5) &= P\left(\frac{Y-24}{\sqrt{12}} < \frac{14.5-24}{\sqrt{12}}\right) = \Phi\left(\frac{14.5-24}{\sqrt{12}}\right) \\ &= \Phi(-2.742) = 1 - \Phi(2.742) \\ &= 1 - 0.9969 = 0.0031 \end{aligned}$$

For a two-sided test at the 1% level we must compare this probability with 0.5%, that is 0.005. We see that, even using the larger approximate value, our probability is less than 0.005 so our test statistic is significant at the 1% level. We therefore reject the null hypothesis and conclude that the evidence suggests strongly that the median of the differences is not zero but is, in fact, positive. Use of propane tends to result in greater hardness.

Example 4

Automotive development engineers are testing the properties of two anti-lock braking systems in order to determine whether they exhibit any significant difference in the stopping distance achieved by different cars.

The systems are fitted to 10 cars and a test is run ensuring that each system is used on each car under conditions which are as uniform as possible.

The stopping distances (in yards) obtained are given in the table below.

	Anti-lock Braking System	
Car	1	2
1	27.7	26.3
2	32.1	31.0
3	29.6	28.1
4	29.2	28.1
5	27.8	27.9
6	26.9	25.8
7	29.7	28.2
8	28.9	27.6
9	27.3	26.5
10	29.9	28.3

Solution

We are testing to find any differences in the median stopping distance figures for each braking system. The null and alternative hypotheses are:

$H_0 : \theta_1 = \theta_2$ or $H_0 : \theta_{\text{differences}} = 0$

$H_1 : \theta_1 \neq \theta_2$ or $H_1 : \theta_{\text{differences}} \neq 0$

We perform a two-tailed test.

The signed differences shown by the two systems are shown in the table below:

	Anti-lock Braking System		
Car	1	2	Sign
1	27.7	26.3	$+$
2	32.1	31.0	$+$
3	29.6	28.1	$+$
4	29.2	28.1	$+$
5	27.8	27.9	$-$
6	26.9	25.8	$+$
7	29.7	28.2	$+$
8	28.9	27.6	$+$
9	27.3	26.5	$+$
10	29.9	28.3	$+$

Solution (contd.)

We have 9 plus signs and the required probability value is calculated directly from the binomial formula as

$$\begin{aligned} P(X \geq 9) &= \sum_{r=9}^{10} \binom{10}{r} \left(\frac{1}{2}\right)^{10-r} \left(\frac{1}{2}\right)^{r} \\ &= \frac{10}{1}\left(\frac{1}{2}\right)^{10} + \left(\frac{1}{2}\right)^{10} = 11 \times \left(\frac{1}{2}\right)^{10} \simeq 0.011 \end{aligned}$$

Since we are performing a two-tailed test, we must compare the calculated value with the value 0.025. Since $0.011 < 0.025$ we reject the null hypothesis on the basis of the available evidence and conclude the the differences in the median stopping distances recorded is significant at the 5% level.

General comments about the sign test

1. Before the sign test can be applied we must be sure that the underlying distribution is continuous. Usually, the second score being higher than the first score counts as a plus sign. The null hypothesis H_0 is that the probability of obtaining each sign is the same, that is $p = \frac{1}{2}$. The alternative hypothesis H_1 may be that $p \neq \frac{1}{2}$ which gives a two-tailed test or $p > \frac{1}{2}$ or $p < \frac{1}{2}$ each of which gives a one-tailed test.

2. If H_0 is correct, the test involves the $B(n, 0.5)$ distribution which, if n is "large" and the conditions for the normal approximation hold, can be approximated by the $N\left(n \times \frac{1}{2}, \sqrt{n \times \frac{1}{2} \times \frac{1}{2}}\right)$ distribution. This approximation can save much tedious arithmetic and time.

3. The sign test may not be as reliable as an equivalent parametric test since it relies only on the sign of the difference of each pair and not on the size of the difference. If it is possible it is suggested that an equivalent parametric test is used.

4. If the underlying distribution is normal, either the sign test or the t-test may be used to test the null hypothesis $H_0 : \theta = \theta_0$ against the usual alternative, but the t-test will not give valid results when the data are non-normal. It can be shown that the t-test produces a smaller Type II error probability for one-sided tests and also for two-sided tests where the critical regions are symmetric. Hence we may claim that the t-test is superior to the sign test when the underlying distribution is normal.

4. The Wilcoxon signed-rank test

As you will now appreciate, the sign test only makes use of the signs of the differences between observed data and the median θ or pairs of differences between observed data in the case of a paired sample. In either case, no account is taken of the size of the differences arising. The statistician Frank Wilcoxon developed a procedure which takes into account both the sign and the magnitude of the differences arising. The resulting test is now widely known as the Wilcoxon signed-rank test. You should note that the test applies to *symmetric* continuous distributions and it is important that you justify this assumption before applying the procedure to a set of data. Note that under this condition, the mean and the median of a distribution are equal and we can use this fact to test the null hypothesis.

$$H_0 : \mu = \mu_0$$

against the alternatives

$$H_1 : \mu \neq \mu_0$$

$$H_1 : \mu > \mu_0$$

$$H_1 : \mu < \mu_0$$

While the theory underpinning this test is complex and is not considered here, the actual test procedure is straightforward and involves the use of special tables. A copy of the Wilcoxon signed-rank test table is given at the end of this Workbook (Table 1). The test procedure is as follows.

1. On the assumption that $x_1, x_2, x_3, \ldots, x_n$ is a random sample taken from a continuous symmetric distribution with mean and median $\mu = \theta$ we test the null hypothesis $H_0 : \mu = \mu_0$ against one of the alternatives given above.
2. Calculate the differences $x - \mu_0,\ i = 1, \ldots, n$.
3. Rank the absolute differences $|x_i - \mu_0|,\ i = 1, \ldots, n$ in ascending order.
4. Label the ranks with the signs of their corresponding differences.
5. Sum the ranks corresponding to positive differences to obtain the value S_P.
6. Sum the ranks corresponding to negative differences to obtain the value S_N.
7. Let $S = \min(S_P, S_N)$.
8. Use Table 1 at the end of this Workbook to reject (if appropriate) the null hypothesis as follows:

Case 1	$H_0 : \mu = \mu_0$ $H_1 : \mu \neq \mu_0$	Reject H_0 if $S \leq$ tabulated value
Case 2	$H_0 : \mu = \mu_0$ $H_1 : \mu > \mu_0$	Reject H_0 if $S_N \leq$ tabulated value
Case 3	$H_0 : \mu = \mu_0$ $H_1 : \mu < \mu_0$	Reject H_0 if $S_P \leq$ tabulated value

Note

It is possible that calculation will result in data with equal rankings. Ties in ranking are dealt with in the usual way. The short example below reminds you how to deal with equal ranking.

Data	Incorrect ranks	Correct ranks
3.1	1	1
4.2	2	2.5
4.2	3	2.5
5.7	4	4.5
5.7	5	4.5
7	6	6
8.1	7	7

To illustrate the application of the Wilcoxon signed-rank test, we will use one of the examples used previously when considering the sign test. The example is repeated here for convenience.

Example 5

The compressive strength of insulating blocks used in the construction of new houses is tested by a civil engineer. The engineer needs to be certain at the 5% level of significance that the median compressive strength is at least 1000 psi. Twenty randomly selected blocks give the following results:

Observation	Compressive Strength
1	1128.7
2	679.1
3	1317.2
4	1001.3
5	1107.6
6	718.4
7	787.4
8	1562.3
9	1356.9
10	1153.2
11	1167.1
12	1387.5
13	679.9
14	1323.2
15	788.4
16	1153.6
17	1423.3
18	1122.6
19	1644.3
20	737.4

Use the Wilcoxon signed-rank test to decide (at the 5% level of significance) whether the hypothesis that the median compressive strength of the insulating blocks is at least 1000 psi is acceptable.

Solution

Assume that the data are taken from a symmetric continuous distribution, so the mean and median are identical. The hypotheses may be stated as:

$$H_0 : \mu = 1000$$

$$H_1 : \mu > 1000$$

The differences are:

Observation	Compressive Strength	$x_i - 1000$	$\lvert x_i - 1000\rvert$	Ascending Order	Signed Rank
1	1128.7	128.7	128.7	1.3	$+1$
2	679.1	-320.9	320.9	107.6	$+2$
3	1317.2	317.2	317.2	122.6	$+3$
4	1001.3	1.3	1.3	128.7	$+4$
5	1107.6	107.6	107.6	153.2	$+5$
6	718.4	-281.6	281.6	153.6	$+6$
7	787.4	-212.6	212.6	167.1	$+7$
8	1562.3	562.3	562.3	211.6	-8
9	1356.9	356.9	356.9	212.6	-9
10	1153.2	153.2	153.2	262.6	-10
11	1167.1	167.1	167.1	281.6	-11
12	1387.5	387.5	387.5	317.5	$+12$
13	679.9	-320.1	320.1	320.1	-13
14	1323.2	323.2	323.2	320.9	-14
15	788.4	-211.6	211.6	323.2	$+15$
16	1153.6	153.6	153.6	356.9	$+16$
17	1423.3	423.3	423.3	387.5	$+17$
18	1122.6	122.6	122.6	423.3	$+18$
19	1644.3	644.3	644.3	562.3	$+19$
20	737.4	-262.6	262.6	644.3	$+20$

We now calculate the sum S_N in order to decide whether to reject the null hypothesis. Note that the form of the null hypothesis dictates that we only need to calculate S_N,

$$S_N = \lvert -8 - 9 - 10 - 11 - 13 - 14\rvert = 65$$

From Table 1, the critical value at the 5% level of significance for a one-tailed test performed with a sample of 20 values is 60. Since $60 < 65$ we conclude that we cannot reject the null hypothesis and that on the basis of the available evidence, the median compressive strength of the insulating blocks is not significantly different to 1000 psi.

Now do the following Tasks.

Again you have seen this problem previously (Task on page 7). This time you are required to use the Wilcoxon signed-rank test to decide whether to reject the null hypothesis.

A certain type of solid rocket fuel is manufactured by bonding an igniter with a propellant. in order that the fuel burns smoothly and does not suffer either "flame-out" or become unstable it is essential that the shear strength of the material bonding the two components of the fuel has a shear strength of 2000 psi. The results arising from tests performed on 10 randomly selected sample of fuel are as follows.

Observation	Shear Strength	Observation	Shear Strength
1	2128.7	6	1718.4
2	1679.1	7	1787.4
3	2317.2	8	2562.3
4	2001.3	9	2356.9
5	2107.6	10	2153.2

Using the Wilcoxon signed-rank test and the 5% level of significance, test the null hypothesis that the median shear strength is 2000 psi.

Your solution

Answer

Assume that the data are taken from a symmetric continuous distribution. The hypotheses are

$$H_0 : \mu = 2000$$

$$H_1 : \mu \neq 2000$$

The Wilcoxon calculations are as shown below. We perform a two-tailed test.

Shear Strength	$x_1 - 2000$	Sorted $\|x_i - 2000\|$	Signed Rank
2128.7	128.7	1.3	+1
1679.1	−320.9	107.6	+2
2317.2	317.2	128.7	+3
2001.3	1.3	153.2	+4
2107.6	107.6	212.6	−5
1718.4	−281.6	281.6	−6
1787.4	−212.6	317.2	+7
2562.3	562.3	320.9	−8
2356.9	356.9	356.9	+9
2153.2	153.2	562.3	+10

We now calculate the sums S_N, S_P and S in order to decide whether to reject the null hypothesis.

$$S_N = |-5-6-8| = 19$$

$$S_p = |1+2+3+4+7+9+10| = 36$$

$$S = \min(S_p, S_N) = \min(36, 19) = 19$$

From Table 1, the critical value at the 5% level of significance for a two-tailed test performed with a sample of 10 values is 8. Since $8 < 19$ we conclude that we cannot reject the null hypothesis and that, on the basis of the available evidence, the median compressive strength of the insulating blocks is not significantly different to 2000 psi.

An automotive development engineer is investigating the properties of two fuel injection systems in order to determine whether they exhibit any significant difference in the level of fuel economy measured on different cars. The systems are fitted to 12 cars and a test is run ensuring that each injection system is used on each car under conditions which are as uniform as possible. The fuel consumption figures (in miles per gallon) obtained are given in the table below. Use the Wilcoxon signed-rank test applied to the differences in the paired data to decide whether the median fuel consumption figures are significantly different at the 5% level of significance.

	Fuel Injection System	
Car	1	2
1	27.6	26.3
2	29.4	31.0
3	29.5	28.2
4	27.2	26.1
5	25.8	27.6
6	26.9	25.8
7	26.7	28.2
8	28.9	27.6
9	27.3	26.9
10	29.2	30.3
11	27.8	26.9
12	29.2	28.3

Your solution

Answer

We assume that each data set is taken from separate continuous distributions. It can be shown that this ensures that the distribution of differences is then symmetric and continuous. In this case the median and mean are identical. We are testing to find any differences in the median miles per gallon figures for each injection system. The null and alternative hypotheses are:

$H_0 : \mu_1 = \mu_1$ or $H_0 : \mu_{\text{differences}} = 0$

$H_1 : \mu_1 \neq \mu_2$ or $H_1 : \mu_{\text{differences}} \neq 0$

We perform a two-tailed test.

The signed ranks are obtained as shown in the table below:

	Fuel Injection System				
Car	1	2	Differences	Sorted Abs	Signed
1	27.6	26.3	1.3	0.4	$+1$
2	29.4	31.0	-1.6	0.9	$+2.5$
3	29.5	28.2	1.3	0.9	$+2.5$
4	27.2	26.1	1.1	1.1	$+5$
5	25.8	27.6	-1.8	1.1	$+5$
6	26.9	25.8	1.1	1.1	-5
7	26.7	28.2	-1.5	1.3	$+8$
8	28.9	27.6	1.3	1.3	$+8$
9	27.3	26.9	0.4	1.3	$+8$
10	29.2	30.3	-1.1	1.5	-10
11	27.8	26.9	0.9	1.6	-11
12	29.2	28.3	0.9	1.8	-12

We now calculate the sums S_N, S_P and S in order to decide whether to reject the null hypothesis.

$$S_N = |-5 - 10 - 11 - 12| = 38$$

$$S_P = |1 + 2.5 + 2.5 + 5 + 5 + 8 + 8 + 8| = 40$$

$$S = \min(S_P, S_N) = \min(40, 38) = 38$$

From Table 1, the critical value at the 5% level of significance for a two-tailed test performed with a sample of 12 values is 13.

Since $13 < 38$ we conclude that we cannot reject the null hypothesis and that on the basis of the available evidence, the two injection systems do not differ significantly in respect of the fuel economy they offer.

General comments about the Wilcoxon signed-rank test

1. For underlying normal populations, either the t-test or the Wilcoxon signed-rank test may be used to test the null hypothesis, say $H_0 : \mu = \mu_0$, concerning the mean of the distribution against the usual alternative. Comparisons between the two tests are difficult since it is hard to obtain the Type II error for the Wilcoxon signed-rank test and hard to obtain the Type II error for the t-test in the case of non-normal populations. For the t-test, the Type I error rate is wrong in non-normal populations.

2. Investigations have shown that the Wilcoxon signed-rank test is never much worse than the t-test and in the case of non-normal populations it may be rather better. The Wilcoxon signed-rank test may be seen as a useful alternative to the t-test, especially when doubt is cast on the normality of the underlying distribution.

Exercises

1. Springs used in the lids of portable CD players are subjected to testing by repeated flexing until they fail. The times, in hours, to failure of forty springs are given below. Those times marked * indicate cases where the experiment was stopped before the spring failed.

*48.0	41.2	1.2	*48.0	*48.0	0.7	0.2	12.2
0.7	19.0	1.9	0.0	42.6	*48.0	15.7	*48.0
4.3	24.2	*48.0	47.5	33.3	17.8	15.9	8.2
4.6	2.7	25.3	3.2	15.7	10.5	2.4	37.1
4.1	30.0	*48.0	19.9	39.3	*48.0	17.5	*48.0

Use a sign test to test the null hypothesis that the median time to failure is 15 hours against the alternative that it is greater than 15 hours. Use the 5% level of significance.

2. In dual-pivot bicycle brakes the control cable enters on one side and there is potential for greater wear in the brake pads on one side than the other. Thirty trials were conducted with a test rig in which a brake was fitted to a wheel connected to a flywheel which was repeatedly set in motion and then brought to rest by the brake with a fixed force applied. The abrasion loss of each brake pad was measured (mg).

Run	Left	Right	Run	Left	Right
1	114	105	16	150	132
2	149	141	17	160	161
3	116	144	18	50	56
4	69	130	19	128	192
5	134	185	20	147	121
6	117	108	21	72	74
7	78	111	22	120	131
8	146	170	23	103	92
9	88	107	24	145	120
10	105	96	25	96	112
11	117	139	26	63	73
12	102	140	27	85	103
13	68	137	28	137	133
14	105	111	29	107	141
15	65	123	30	67	83

Use a sign test to test the null hypothesis that the median difference between left-pad wear and right-pad wear is zero against the two-sided alternative. Use the 5% level of significance.

3. Loaded lorries leaving a quarry are weighed on a weigh bridge. To test the weigh bridge, each of a sample of twelve lorries is driven to a second weigh bridge and weighed again. The differences (kg) between the two weights (first − second) are given below.

$$38 \quad 14 \quad 16 \quad 54 \quad 36 \quad -19 \quad -24 \quad 1 \quad -18 \quad 5 \quad -14 \quad -28$$

Use a Wilcoxon signed-rank test to test the null hypothesis that there is no systematic difference in the weights given by the two weigh bridges. Use the 5% level of significance. Comment on any assumptions which you need to make.

4. Apply a Wilcoxon signed-rank test to test to the data in Exercise 2 to test the null hypothesis that the mean difference in abrasion loss between the left and right pads is zero. Use the 5% level of significance. Comment on any assumptions which you need to make.

Answers

1. Under the null hypothesis the probability that the failure time is greater than 15 hours is 0.5 and the distribution of the number with failure times greater than 15 hours in binomial $(40, 0.5)$. Of the forty test springs, 25 had failure times greater than 15 hours. The probability under the null hypothesis of observing at least 25 can be found approximately using the normal distribution $N(20, 10)$. Now

$$\frac{24.5 - 20}{\sqrt{10}} = 1.423$$

and the probability that a standard normal random variable is greater than 1.423 is $1 - \Phi(1.423) = 0.077$. Since $0.077 > 0.05$, the result is not significant at the 5% level and we do not reject the null hypothesis that the median failure time is 15 hours.

2. In 9 cases the left-pad wear is greater than the right-pad wear. Let X be the number of cases where left-pad wear is greater than right-pad wear. Under the null hypothesis X has a binomial $(30, 0.5)$ distribution. The probability of observing a value less than or equal to 9 from this distribution is 0.0214. Because we are testing against the two-sided alternative we double this to 0.0428 and, because $0.0428 < 0.05$, the result is significant at the 5% level. We reject the null hypothesis and conclude that left-pad wear tends to be less than right-pad wear.

3. The observations and their signed ranks are as follows.

Observation	38	14	16	54	36	-19
Signed rank	11.0	3.5	5.0	12.0	10.0	-7.0
Observation	-24	1	-18	5	-14	-28
Signed rank	-8.0	1.0	-6.0	2.0	-3.5	-9.0

The sum of the positive ranks is 44.5 and the sum of the negative ranks is 33.5. For a two-tailed test at the 5% level of significance, the critical value is 13 and we compare the smaller rank sum with this. We see that the rank sum is not less than 13 so the result is not significant and we do not reject the null hypothesis. There is no significant evidence of a systematic difference between the weigh bridges.

Comment: We are assuming that, under the null hypothesis, the distribution of the differences is symmetric. This may well be valid in this case since, if the weigh bridges are really the same then the differences between values given by them should be distributed symmetrically about zero. (We also have to assume that the weight does not change systematically on the journey between the weigh bridges, for example by spillage.)

Answers

4. The thirty differences (left $-$ right) and their signed ranks are as follows.

Run	Difference	Signed rank	Run	Difference	Signed rank
1	9	8.0	16	18	15.5
2	8	6.0	17	-1	-1.0
3	-28	-22.0	18	-6	-4.5
4	-61	-28.0	19	-64	-29.0
5	-51	-26.0	20	26	21.0
6	9	8.0	21	-2	-2.0
7	-33	-23.0	22	-11	-11.5
8	-24	-19.0	23	11	11.5
9	-19	-17.0	24	25	20.0
10	9	8.0	25	-16	-13.5
11	-22	-18.0	26	-10	-10.0
12	-38	-25.0	27	-18	-15.5
13	-69	-30.0	28	4	3.0
14	-6	-4.5	29	-34	-24.0
15	-58	-27.0	30	-16	-13.5

The sum of the positive ranks is 101. The sum of the negative ranks is 364. (The total of the ranks is $0.5 \times 30 \times 31 = 465$.) With $n = 30$ the distribution of the rank sum under the null hypothesis is approximately normal with mean $M = n(n+1)/4 = 30 \times 31/4 = 232.3$ and standard deviation $S = \sqrt{n(n+1)(2n+1)/24} = \sqrt{30 \times 31 \times 61/24} = 48.62$. For a two-sided test at the 5% level we reject the null hypothesis if either rank sum is outside the range $M \pm 1.96S$, which is 232.3 ± 95.3 or 137.0 to 327.6. We see that the rank sums are indeed outside of this range so we reject the null hypothesis at the 5% level and conclude that left-pad wear tends to be less than right-pad wear.

Comment: We are assuming that, under the null hypothesis, the distribution of the differences is symmetric. This seems reasonable since the assumption that there is no systematic difference bewteen left and right would imply that the distribution of differences in observed wear should be symmetric.