University of Glasgow
School of Mathematics & Statistics

Statistical Inference: Level M

Chapter 2

Hypothesis Testing and Nonparametric Inference

January 2021

Dr Benn Macdonald
Room 225, Maths & Stats Building
Benn.Macdonald@glasgow.ac.uk

This chapter will start by introducing the idea of **hypothesis testing**. This approach to statistical inference is used in all three of the areas that will be introduced in this course: nonparametric, parametric and Bayesian inference. After introducing the general principles of hypothesis testing this chapter will use and implement these ideas through the introduction of techniques in nonparametric inference. This type of inference requires very little assumptions to be made about the distribution of the data and hence can be introduced with only a little knowledge of probability. The other two types of inference rely heavily on probability theory and hence will be introduced later in the course once the required ideas have been introduced in Probability (M).

2.1 Hypothesis Testing

Well-conducted research studies often begin with a question about a population. Here are some examples.

- What is the average height of children in the West of Scotland when they start Primary School?
- Is the average level of nitrate in groundwater less than the environmental trigger of 50mg/l in areas which are not draining from substantial farmland?
- In 1972, the proportion of adults in the UK earning over £20K per annum was 15%, this figure had risen to 45% by 1992. Has this figure risen further since 1992?

We might say that we wish to investigate the plausibility of a hypothesis (or statement) about the population. Here are some possible hypotheses.

- When children in the West of Scotland start Primary School, their median height is different from that of children of the same age in the South of England (which is 1.06 metres).
- The mean level of nitrate in groundwater is less than the environmental trigger of 50mg/l in areas which are not draining from substantial farmland.
- The proportion of UK adults earning over £20K is now greater than 45%.

These are all statements about what researchers expect to find out. They are usually saying that something has changed, or is different, or can be made different. These are study hypotheses, which are often denoted H_1. Hypotheses are usually written in terms of population values (i.e. parameters). We will use η as a symbol for the population median, μ as a symbol for the population mean and θ as the symbol for a population proportion or probability.

- H_1 :
  where η is the population median height (metres) of West of Scotland children when they start Primary School.

- H_1 :
  where μ is the population mean level of nitrate (mg/l) in groundwaters that are not draining from substantial farmland.

- H_1 :
  where θ is the population proportion of UK adults who earn more than £20K.

We use the data collected from a sample to assess the evidence in favour of the study hypothesis. It is necessary also to consider what would be true if the study hypothesis were not true. We can frame this statement about the population as a hypothesis, called the **null hypothesis** (H_0). H_0 usually includes the possibility that nothing has changed, or two things are equal.

- H_0 :

- H_0 :

- H_0 :

On the basis of the evidence provided by the sample data, we must decide which of the null and study hypotheses is the more plausible. For this reason, the study hypothesis (H_1) is often called the **alternative hypothesis**.

It might seem as though we ought to treat the null and alternative hypotheses equally, but we do not. Null hypotheses usually, by their nature, cannot be proven to be true unless we have access to data from the whole population.

When assessing the evidence from a sample we have just two possibilities. Either we reject H_0 in favour of H_1 or we do not reject H_0. We can never

prove H_0 from a sample, although often we do not have sufficient evidence to reject it.

Generally hypotheses fall into two main classes: one-sided and two sided alternative hypotheses.

One-sided

- H_0: H_1:
- H_0: H_1:

Two-sided

- H_0: H_1:

2.1.1 Assessing the evidence against H_0

For each of the questions on Page 1 we are interested in calculating information about a 'population' i.e.

- The median height of (early primary school) children in the west of Scotland.
- The mean level of nitrate in groundwaters that do not drain from substantial farmland.
- The proportion of working adults in the UK that earn over £20K.

In each case we can summarise the evidence from our sample e.g. we can compute the median height of children in a representative sample of children starting West of Scotland primary schools. These summaries can be referred to as our **test statistics** for each question that we are interested in. We saw in Chapter 1 that a statistic is a function of the data in the sample and that this value will change from sample to sample. The observed value for each of the examples above (i.e. median height, proportion of adults) in our samples of data are referred to as our **observed test statistics**.

A test statistic enables us to carry out a **hypothesis test**, which is a formal procedure to decide between the null and alternative hypotheses. The strength of the evidence against H_0 (and for H_1) is judged using a **p-value**, which is determined on the basis of the observed value of the test statistic.

> The p**-value** is the probability of obtaining a value for your test statistic that is at least as extreme as the observed value of the test statistic (*assuming* H_0 *is true*).

The p-value is an attempt to measure the consistency of the data with the null hypothesis.

1. Low p-values imply that the data are improbable if H_0 is true. Since the data actually happened, this suggests rejecting H_0 in favour of H_1.

2. High p-values suggest that the data are quite probable under H_0. Since the data and hypothesis appear consistent, H_0 is not rejected.

How low should a p-value be in order to cause rejection of H_0? There is no 'correct' answer to this, and many statisticians prefer not to make one up, simply quoting the p-value instead. If a value, α, is chosen, such that H_0 will be rejected if the p-value is $\leq \alpha$ then α is known as the **significance level** of the test. Traditional choices for α are 0.05 (5%) and 0.01 (1%), with 0.05 being the most common.

For a two-sided alternative hypothesis we will typically use $\alpha = 0.05$ (and $\alpha = 0.025$ for a one-sided alternative hypothesis).

The usefulness of p-values as a measure of the evidence against H_0 stems from their general applicability: a p-value of 0.04 has the same meaning whatever hypothesis is being tested and whatever method is being used for the test, and that meaning is well defined.

2.2 Nonparametric Inference

In this course we will mainly focus on parametric inference, which depends on distributional assumptions. However, we will start with a brief look at nonparametric tests which require relatively weak distributional assumptions for their validity.

Nonparametric, or distribution free tests require few, if any, assumptions about the shapes of the underlying population distributions. For this reason, they are often used in place of parametric tests if/when one feels that the assumptions of the parametric test have been too extremely violated (e.g. if the sample sizes are small and the data do not follow a defined probability distribution).

The parameter of interest for such tests is generally the **median**, since it is a robust measure of location that is not heavily affected by outliers.

In this course we will focus mainly on methods for one or two populations.

2.2.1 Wilcoxon's Signed Ranks Test (One population or paired data)

When we have only one sample of data from a population, the main question of interest is often about estimating the 'average' value of a specific variable within the population.

The Wilcoxon Signed Ranks test investigates the null hypothesis that the median of the distribution of the data has a specified value. The idea behind the test is based on differences between each of the data points in the sample and a specified value for the median. If it is true that the population median value is equal to the specified value then we would expect half of the differences to be positive and half of the differences to be negative. If this is not the case we have evidence to reject the null hypothesis.

Similarly, the test is also commonly used for paired data to test the null hypothesis that the median difference between paired data points is zero. The difference between each pair of points is computed; and under the null hypothesis that the median difference is zero then we would expect half of the differences to be positive and half to be negative. If this is not the case we have evidence to reject the null hypothesis.

However, we do not wish to deal with the actual numerical values of the differences since these can be affected by outliers in the dataset. The test therefore gives ranks to the absolute value of the differences and computes the sum of the positive and negative ranks.

The test has a small number of **assumptions**:

- The observations come from a symmetric distribution.
- The observations are random and independent.

The format of the test is as follows:

1. Explore the data - initial impression;
2. Graphically check assumptions;
3. State the null and alternative hypotheses;
4. State the test statistic;
5. Compute the difference of the data from H_0;
6. Order and rank the differences;
7. Compute the observed test statistic;
8. Find the rejection region (the set of values of the test statistic for which H_0 is rejected);
9. State the conclusion.

Example 1 - House price

The following values are the house price (in £1,000's) of 8 houses in an extremely affluent area of England:

2491 2485 3433 2575 2521 2451 2550 2540

Investigate the hypothesis that the median house price of the area is £2,500,000.

Initial Impression

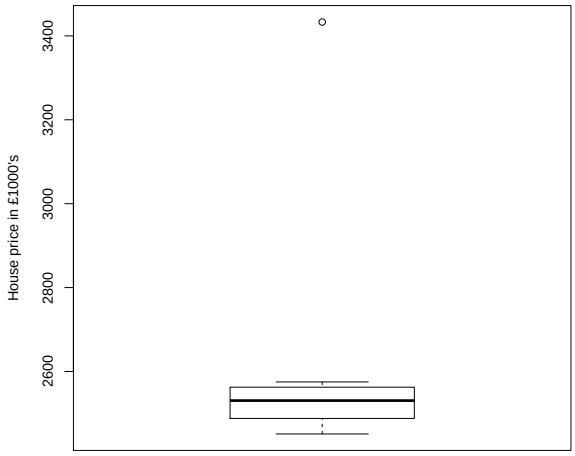

Figure 1: House price

A stem and leaf plot of house price data:

```
The decimal point is 2 digit(s) to the right of the |

 24 | 599
 25 | 2458
 26 |
 27 |
 28 |
 29 |
 30 |
 31 |
 32 |
 33 |
 34 | 3            34|3 = 3430 to 3434
```

The plots highlight that there is one extreme data point at £3,433,000.

The observed values (with the exception of the outlier) are consistent with an average house price of £2,500,000. There is some concern about the validity of the Wilcoxon Signed Ranks Test, since the outlier gives the sample data such a strong impression of asymmetry. Each house price can be considered to be independent.

Hypotheses

Let η denote the median house price of the houses in the area. Then,

H_0 :

H_1 :

Test Statistic (TS)

Observed Value of TS

Compute the sum of the positive and negative ranks, denoted $W+$ and $W-$ respectively. The test statistic is whichever of these has the smallest value. In this case, the test statistic is $W-$, the sum of the ranks of the $|d_i|$ corresponding to negative d_i values in the order statistic. (This is chosen because it is smaller than W+).

$W- =$

Rejection Region (RR)

An exact p-value can be computed by hand here and this will be described later. However, it is cumbersome to do this and so an alternative method is to use critical values from statistical tables to determine whether or not to reject H_0.

The statistical tables provide a critical value for the test to assess the evidence against H_0. The theoretical details of this are described fully in Section 2.2.1.1. Only the sample size and the significance level are required to obtain the critical value.

Since this is a two-sided test we will choose our significance level to be $\alpha = 0.05$. Using Statistical Tables Section 7 for $n = 8$,

RR =

Conclusion

We may wish to repeat the test after removing the extreme observation to see if the results of the analyses change. Try this as an example.

Analysis in R

The following R command can be used to perform this test, x here are the original data, the house prices..

```
wilcox.test(x, mu=2500)
```

Results in R:

```
Wilcoxon signed rank test

data:  x
V = 28, p-value = 0.1953
alternative hypothesis: true location is not equal to 2500
```

Statistical packages will return a p-value for the test as illustrated above. Note, the test statistic here is the larger of W+ and W-, R is using a slightly different formulation of the test.

Conclusion

Since the p-value > 0.05 at 0.1953, we do not reject H_0 and conclude that there is insufficient evidence of a difference from £2,500,000.

2.2.1.1 Exact Distribution

Under H_0, each rank has equal probability of being assigned a positive or negative sign and the value of the test statistic is evaluated for each of these 2^n possibilities, where n is the sample size. An assumption of symmetry is needed.

Example 2 - Artificial Example with $n = 3$

Table 1 shows all the possible combinations of ranks that may be assigned to positive and negative values of d_i, ignoring the possibility of ties.

Since each rank may be assigned to a + or to a − value, it follows that there are $2^3 = 8$ different possibilities.

Rank 1	Rank 2	Rank 3	W+	W-
-	-	-	0	6
+	-	-	1	5
-	+	-	2	4
-	-	+	3	3
+	+	-	3	3
-	+	+	5	1
+	-	+	4	2
+	+	+	6	0

Table 1: All possible combinations of ranks that may be assigned

In general, there are 2^n different possibilities (with no ties). Under H_0 because the data have a symmetric distribution each of the possibilities is equally likely. Under H_0 the distribution of all of the d_i is symmetric around the median value of 0. So given that an observed d_i has rank j in the order statistic of the d_i's there is no reason why d_i should be more likely to be a $+$ than a $-$ and vice versa. In other words,

$$\text{P(rank } j \text{ is assigned to a } + \mid H_0) = \text{P(rank } j \text{ is assigned to a } - \mid H_0) = \tfrac{1}{2}$$

By the multiplication principle, then all 2^n different possibilities are equally likely under H_0. We can use the results from Table 2 to obtain the null distribution of W+ or W-. Here we will look at W+ (for illustration purposes):

w	Combinations of + ranks giving W+=w	Number of combinations	P(W+=w $\mid$ H_0)
0	all ranks negative	1	2^{-3}
1	(1)	1	2^{-3}
2	(2)	1	2^{-3}
3	(3), (1,2)	2	2×2^{-3}
4	(1,3)	1	2^{-3}
5	(2,3)	1	2^{-3}
6	(1,2,3)	1	2^{-3}

Table 2: Probability mass function of W+

The p-value for the Signed Ranks Test of H_0 against H_1 can be found from the cumulative distribution function $P(W+ \leq w)$.

For example, if our value for W+ is 2.

$$\mathrm{P}(W+ \leq 2|H_0) = (1+1+1) \times 2^{-3} = 0.375$$

So, the p-value of the test would be found as follows:

$p = \mathrm{P}(\mathrm{W}+ \text{ has at least as "extreme" a value as that observed } | \ \mathrm{H}_0)$

$p = P(W+ \leq 2) + P(W+ \geq \frac{1}{2}n(n+1) - 2)$

$p = P(W+ \leq 2) + P(W+ \geq 4)$

Since we have a symmetric distribution for $W+$ then,

$p = 2 \times P(W+ \leq 2) = 2 \times 0.375 = 0.75$

Note: For a one-sided alternative hypothesis the p-value would not be multiplied by 2.

It is clearly cumbersome to calculate the p-value for this test and hence critical values can be obtained from Statistical Tables relating to the significance level (α) of the test. This is usually set at $\alpha = 0.05$ (or $\alpha = 0.01$), for a two-sided test.

The **critical value** w_α is such that

$\alpha = \mathrm{P}(\mathrm{W}+ \text{ has at least as "extreme" a value as } \mathrm{w}_\alpha \ | \ \mathrm{H}_0)$

$\alpha = \mathrm{P}(\mathrm{W}+ \leq w_\alpha \ ) + P(W+ \geq \frac{1}{2}n(n+1) - w_\alpha)$

$\alpha = 2 \text{ x } \mathrm{P}(\mathrm{W}+ \leq w_\alpha)$ (By symmetry)

This is the largest value of W+ for which we would be willing to reject H_0. In other words, we would only reject H_0 when $W+ \leq w_\alpha$.

Critical values (w_α) for the Wilcoxon Signed Ranks Test, for various sample sizes, are given in the Statistical Tables Section 7.

2.2.1.2 Notes

- We only use + or - values of differences in this test. Zeros are ignored. If there are zeros then these are removed and the sample size is reduced.

- One-sided tests may be conducted. However, these tests have significance levels of $\alpha = 0.025$ (or $\alpha = 0.005$). These significance levels are exactly half those of the corresponding two-sided tests, since a test with a one-sided alternative hypothesis has a p-value that is exactly half the p-value of the corresponding test with a two-sided alternative hypothesis.

- For one-sided tests, the null hypothesis is either

  $H_0 : \eta \leq 0$ or $H_0 : \eta \geq 0$

  p or α is then calculated for the 'worst case' which is $\eta = 0$. The test statistic depends on the direction of the hypotheses, which here would be W- and W+ respectively. For example, if $H_0 : \eta \leq 0$, then reject H_0 for values of W- which are not larger than the critical value from the table (and for small p-values).

- In principle, exact p-values may always be calculated by tabulating all the possible arrangements of the ranks, as shown above.

- Approximate p-values for the Wilcoxon Signed Ranks test can be easily obtained from R.

- The Wilcoxon Signed Ranks test is commonly used with paired data and examples of this will be illustrated in the tutorial sheet.

2.2.2 Mann-Whitney U Test (Wilcoxon Rank Sum Test)

A similar approach can be used to derive a test that compares the medians of two independent populations.

Although this test is often attributed to Mann and Whitney, and called the Mann-Whitney U Test, Wilcoxon himself independently realised how to extend his earlier work on one-sample problems to the two-sample context and obtained the test in an equivalent form.

The idea of the test is that if there is no difference between the two populations, in terms of the values of a specific variable, then we should be able to collect all of the data together, ignoring which population the data come from, and rank the magnitude of the values.

By using ranks we avoid working with the numerical values where outliers may influence the result.

If we then aggregate (add up) the ranks for each population then we should get roughly the same answer for each population. If the sum of the ranks is very different for each population then we have evidence that the values of the two populations are not the same.

Assumptions

- All the recorded values are independent observations;
- The variable of interest has the same distribution in the two populations, except possibly for a difference in the medians. (In other words, the distribution of the variable of interest has the same shape and spread in the two populations.)

Example 3 - Preferred Room Temperatures

In a controlled environment laboratory, 10 men and 10 women were tested individually to determine the room temperature (^{0}F) they found to be most comfortable. The following results were obtained:

Men	74	72	77	76	76	73	75	73	74	75
Women	75	77	78	79	77	73	78	79	78	80

Assuming that these values represent a random sample from the respective populations, is the average comfortable temperature the same for men and women?

Initial Impression

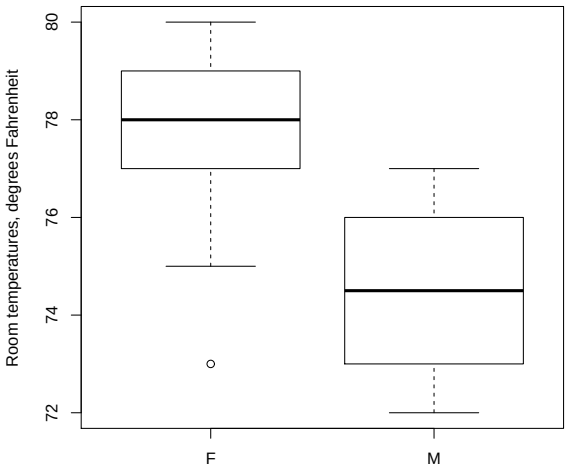

Figure 2: Preferred room temperatures for males and females

From Figure 2 it can be seen that the preferred room temperatures for females are quite a bit higher than those for males.

Assumptions

If X is the preferred room temperature of a randomly-selected individual, then we assume:

1. all the recorded values of X are independent observations;
2. X has the same distribution in the populations of men and women, except possibly for a difference in the medians. (In other words, the distribution of X has the same shape and spread in the two populations.)

The assumptions appear plausible in this example.

Hypotheses

If η_A, η_B are the population median preferred temperatures for men and women, then
H_0 :

H_1 :

Test Statistic

The Mann-Whitney U Test proceeds as follows. Form the combined order statistic (c.o.s.) of the data (i.e. the order statistic of the pooled data from both samples), and rank the values in the c.o.s.

Consider now the two equivalent statistics:

R_A = sum of ranks (in the c.o.s.) of observations from group A.
R_B = sum of ranks (in the c.o.s.) of observations from group B.

In this example,

$R_A =$

$R_B =$

In general, if m and n are the sample sizes ($m = n = 10$), then

$$R_A + R_B = \tfrac{1}{2}(m+n)(m+n+1)$$

In this example, $m = n = 10$, so $\frac{1}{2}(m+n)(m+n+1) = \frac{1}{2} \times 20 \times 21 = 210$. (It is not a requirement that the sample sizes are equal, it is just coincidence in this example).

Consider R_A. Since there are m observations in group A, the smallest value that R_A can take is for the ranks 1, ..., 10 which would give $R_A = 55$ i.e. $\frac{1}{2}m(m+1)$, which occurs when A values take the ranks 1, 2, ..., m in the c.o.s. So, it is conventional to work with

$$U_A = R_A - \tfrac{1}{2}m(m+1) \text{ and } U_B = R_B - \tfrac{1}{2}n(n+1)$$

both of which can take integer values starting at 0. This way differences in sample size do not influence the result.

In this example,

$U_A =$

$U_B =$

The conventional test statistic is:

$U = \min(U_A, U_B) =$

Null Distribution of U

Ignoring the possibility of ties, there are different ways of assigning m ranks to the observation from Group A. Under the null hypothesis, each of these is equally likely. This allows us (in principle) to establish the exact null distribution of U, and hence calculate an exact p-value for the test.

In practice, this is very tedious (as for the Wilcoxon Signed Ranks Test). So, we will generally be content to choose a Significance Level (α) and draw a conclusion with the help of the critical values tabulated in Statistical Tables.

Significance Level

Chosen to be $\alpha = 0.05$ (or 5%), since the hypotheses are two-sided.

Critical Value

Using Statistical Tables Section 7, the critical value of the test, when $m = n = 10$, is $u_{0.05} = 23$. This means that we would reject H_0 in favour of H_1 only if $U \leq 23$. Alternatively, we can say that a REJECTION REGION for the test is given by the interval,

Rejection Region

RR =

Conclusion

Analysis in R[1]

The following R command can be used to perform this test:

```
wilcox.test(data~grp)
```

Result in R

```
Wilcoxon rank sum test with continuity correction

data:  data by grp
W = 13, p-value = 0.005452
alternative hypothesis: true location shift is not equal to 0

Warning message:
In wilcox.test.default(x = c(74, 72, 77, 76, 76, 73, 75, 73, 74,  :
  cannot compute exact p-value with ties
```

Conclusion

The results in R provided a p-value of 0.005 and since this is < 0.05 we reject H_0. Therefore, there is evidence that the preferred mean temperatures are different for each population.

2.2.2.1 Notes

- How can we be sure that the second assumption holds? As well as exploring the data graphically, we might also compare the sample Inter-Quartile Ranges. If the ratio of the larger IQR to the smaller one is no greater than 2, then we are usually willing to proceed as though the observations come from populations with equal spreads.
- One-sided hypothesis tests can also be conducted. Since the p-values (and significance levels) of such tests are exactly one half of their corresponding values under a two-sided alternative, we work at significance levels of $\alpha = 0.025$ and 0.005 for one-sided tests.
- If the null hypothesis is $H_0 : \eta_A \leq \eta_B$, then the test statistic is U_B i.e. we reject H_0 if U_B is not larger than the critical value in the table, and vice versa for the opposite hypotheses.

[1] R code and datasets for Ch2 are available on Moodle for you to try when you have a chance.

Normal Approximations

For both the Wilcoxon Signed Ranks test and the Mann-Whitney U test, if the sample sizes are large enough and there are not too many ties then normal approximations can be used to estimate p-values. These will sometimes be automatically implemented in `R` packages if the sample size is large enough.

We will not go into the detail of these at the moment since normal approximations will not be considered until later in the Probability (Level M) course. For more information on these see Rice (1995) and Garthwaite (2006).

2.2.3 Permutation and Randomization Tests

The Wilcoxon Signed Ranks test and the Mann-Whitney test are classical statistical nonparametric approaches for making inferences about the 'average' values of one and two populations, respectively. They are straight-forward in the sense that they can be computed by hand. However, the approaches in this section are more modern and have been introduced as a result of fast computing power.

In several fields of mathematics, the term permutation is used with different but closely related meanings. They all relate to the notion of (re-)arranging elements from a given finite set into a sequence. The basic idea of a permutation test is to find a family of permutations of the data, such that the probability of each permutation is known under the null hypothesis, H_0. Typically, each permutation has equal probability. The probability distribution (under H_0) of a specified test statistic may then be evaluated by determining the statistic's value for each permutation.

A permutation test is a type of statistical significance test in which the labels on the observed data points are rearranged to provide a 'permutation of the data'. All possible rearrangements are found and for each rearrangement the test statistic is computed. This provides the distribution of the test statistic under the null hypothesis. We can use a permutation test only when we can see how to resample in a way that is consistent with the study design and with the null hypothesis.

Since such methods require no particular assumptions concerning statistical distributions (with the exception of the important assumption of independent observations), permutation tests are increasingly applied even in the context of traditional statistical tests.

The procedure for a permutation test is:

1. Devise a test statistic.
2. Define your null hypothesis.
3. Create a new data set consisting of your data, randomly rearranged.
4. Calculate your test statistic for this data set, and compare it to your observed value.
5. Repeat steps 3 and 4 for all permutations of the data.

Example 4 - Company profits
A permutation test will be used here as an alternative to the Mann-Whitney test. The idea behind the test is that if there is no difference between the 'average' values of two populations then if we re-arrange the values between the two populations the 'average' should remain unchanged.

An analyst was interested in comparing the energy and retail sectors in terms of annual profit. The profits of 3 companies in the energy sector and 3 companies in the retail sector are recorded in Table 3 below. (The sample sizes are very small here and would ideally be higher. However, the data can be used to illustrate the technique of a permutation test.)

Sector	Profit (in millions)
Energy	47 45 41
Retail	31 33 39

Table 3: Profits for retail and energy companies

Is there evidence that the profits are systematically different between the energy and retail sectors?

The most obvious feature of this experiment is that it is extremely small scale. Looking at the little data we have, it appears as though the profits might be a bit higher for the energy sector than for the retail sector.

Hypotheses

What is a suitable **null hypothesis** for this experiment?

Letting η_A and η_B denote the population median profits under the two sectors, we might settle for testing the following hypotheses:

Assumptions

We would also assume that the shape and spread of the distributions of profits are the same for the two sectors.

Test Statistic

A suitable test statistic is

$$\text{TS} = \text{sample median for energy } (\eta_A) \text{ - sample median for retail } (\eta_B)$$

Under the null hypothesis, this test statistic should be about 0. Values of the test statistic that are either 'too large' or 'too small' provide evidence in favour of the alternative hypothesis.

The **observed value of the test statistic** was

We face the problem of obtaining information about the null distribution of the test statistic so that we can decide whether the observed value of 12 is extreme compared to what we would expect under H_0.

A permutation test is constructed by noting that, under the null hypothesis, the two sets of data are observations from the same distribution. In fact, all our knowledge of the null distribution must be inferred from these 6 (= 3 + 3) data values.

There are $\binom{6}{3} = 20$ different ways of permuting the observed data values so that 3 are associated with the Energy sector and the other 3 are associated with the Retail sector. These possibilities are all listed in Table 4. Under the null hypothesis, all of these configurations are equally-likely to have occurred as a result of the experiment. So, these 20 possible outcomes allow us to determine how likely or unlikely the observed value of the test statistic was, given the null hypothesis.

DataE	MedianE	DataR	MedianR	MedE-MedR
31 33 39	33	41 45 47	45	-12
31 33 41	33	39 45 47	45	-12
31 33 45	33	39 41 47	41	-8
31 33 47	33	39 41 45	41	-8
31 39 41	39	33 45 47	45	-6
31 39 45	39	33 41 47	41	-2
31 39 47	39	33 41 45	41	-2
31 41 45	41	33 39 47	39	2
31 41 47	41	33 39 45	39	2
31 45 47	45	33 39 41	39	6
33 39 41	39	31 45 47	45	-6
33 39 45	39	31 41 47	41	-2
33 39 47	39	31 41 45	41	-2
33 41 45	41	31 39 47	39	2
33 41 47	41	31 39 45	39	2
33 45 47	45	31 39 41	39	6
39 41 45	41	31 33 47	33	8
39 41 47	41	31 33 45	33	8
39 45 47	45	31 33 41	33	12
41 45 47	45	31 33 39	33	12

Table 4: All permutations for energy and retail data

4 out of 20 of the possible permutations give a sample median difference whose absolute value is greater than or equal to the observed value, 12. So, the p-value of the permutation test is

$$p = P(|TS| \geq 12|H_0) = \frac{4}{20} = 0.2$$

Based on our knowledge of the (assumed common) distribution of profits, and the knowledge of the null distribution of the test statistic that is derived from that, a difference of 12 or more between the sample medians is quite likely to occur.

Therefore, we do not reject H_0. There is no evidence of a significant difference between the population median profits.

It is clear that as the sample sizes increase it would become very cumbersome to compute all permutations by hand. The availability of fast computers has

made permutation tests increasingly feasible, even for large data sets. If the number of permutations is very large (i.e. the sample size is very large) then the null distribution is estimated using a random sample of the permutations. This is commonly referred to as a randomization test. This will be illustrated below.

The following `R` packages and functions can be used to perform one and two sample permutation tests.

`R` functions:

```
perm.test (library(exactRankTests))

oneway_test (library(coin))
```

Example 3 - Cont...

Ideally, we might like to use a permutation test to analyse the data from the experiment involving preferred temperatures of 10 men and 10 women. However, in order to do so, we would have to be able to deal with all equally-likely possible combinations of data under the null hypothesis.

As a more practical alternative, we will carry out a randomization test. Instead of investigating every possible permutation of the data, we choose a simple random sample of (say) 10,000 of them (usually with replacement), calculate the value of the test statistic for each of these possibilities, and estimate the p-value by comparing these results against the observed value.

Hypotheses

If η_A and η_B are the population median preferred temperatures for men and women, then,

$$H_0 : \eta_A = \eta_B \text{ vs. } H_1 : \eta_A \neq \eta_B$$

Test Statistic

The test is based on the difference between the sample medians. The observed value of this statistic is:

sample median for males - sample median for females = 74.5 - 78.0 = -3.5

p-Value

All 20 observations were pooled. At random, 10 were selected as the 'male' sample. The remaining 10 were the 'female' sample. The difference between the medians of the two groups was then calculated. This was repeated 10,000 times, and the values shown in Table 5 were recorded for the median differences.

In total, a value at least as extreme as the observed value (-3.5) was recorded on 93 occasions (4+37+46+6) and so our estimate of the p-value is 93/10000 = 0.0093.

Conclusion

As the p-value is < 0.05 we reject H_0 in favour of H_1 and conclude that there is evidence of a difference between the two populations in terms of preferred mean temperature.

Median Difference	Frequency
-4.0	4
-3.5	37
-3.0	168
-2.5	488
-2.0	860
-1.5	999
-1.0	1584
-0.5	76
0.0	1533
0.5	65
1.0	1583
1.5	977
2.0	901
2.5	503
3.0	170
3.5	46
4.0	6

Table 5: Frequency of median differences for 10,000 samples

2.2.4 Kruskal-Wallis Test (for k independent samples)

The Kruskal-Wallis H-test goes by various names, including Kruskal-Wallis one-way analysis of variance by ranks. It is for use with k independent groups, where k is equal to or greater than 3, and the measurements are at least ordinal. (When $k = 2$, you would use the Mann-Whitney U test instead. The Kruskal-Wallis test is a generalisation of the Mann-Whitney test). Note that because the samples are independent, they can be of different sizes.

Example 5 - Children's Comprehension

A student was interested in comparing the effects of four kinds of reinforcement on children's performance on a test of reading comprehension. The four reinforcements used were: (a) praise for correct responses; (b) a jelly bean for each correct response; (c) reproof for incorrect responses; and (d) silence. Four independent groups of children were tested, and each group received only one kind of reinforcement. The measure of performance given in Table 6 is the number of errors made during the course of testing following each type of reinforcement.

a	b	c	d
68	78	94	54
63	69	82	51
58	58	73	32
51	57	67	74
41	53	66	65
61	80		

Table 6: Number of errors made during the course of testing

The null hypothesis is that the k samples come from the same population, or from populations with identical medians. The alternative hypothesis states that not all population medians are equal i.e. the population median no. of errors made differs between at least two types of reinforcement. It is assumed that the underlying distributions are continuous; but only ordinal measurements are required.

This test is especially useful in small-sample situations and since the data are replaced by their ranks, outliers will have less influence on the test. The test can be performed in `R` using the function: `kruskal.test()`. We will not go into the detail of this test in this course. For more information on the test see Rice (1995).