Tutorial 5: Hypothesis Testing

Rob Nicholls – nicholls@mrc-lmb.cam.ac.uk

MRC LMB Statistics Course 2014

Contents

1 Introduction

It is often the case that we want to infer information using collected data, such as whether two samples can be considered to be from the same population, whether one sample has systematically larger values than another, or whether samples can be considered to be correlated. Such hypotheses may be formally tested using inferential statistics, allowing conclusions to be drawn, allowing the potential for objective decision making in the presence of a stochastic element.

The general idea is to predict the likelihood of an event associated with a given statement (i.e. the hypothesis) occurring by chance, given the observed data and available information. If it is determined that the event is highly unlikely to randomly occur, then the hypothesis may be rejected, concluding that it is unlikely for the hypothesis to be correct. Conversely, if it is determined that there is a reasonable chance that the event may randomly occur, then it is concluded that it is not possible to prove nor disprove the hypothesis, using the particular test performed, given the observed data and available information. Conceptually, this is similar to saying that the hypothesis is 'innocent until proven guilty'. Such hypothesis testing is at the core of applied statistics and data analysis.

The ability to draw valid conclusions from such testing is subject to certain assumptions, the most simple/universal of which being the base assumptions that the observed data are ordinal, and are typical of the populations they represent. However, assumptions are often also made about the underlying distribution of the data. Different statistical tests require different assumptions to be satisfied in order to be validly used. Tests that make fewer assumptions about the nature of the data are inherently applicable to wider classes of problems, whilst often suffering from reduced *Statistical Power* (i.e. reduced ability to correctly detect thus reject the hypothesis in cases when the hypothesis is truly incorrect).

Statistical tests may be separated into two classes: parametric tests and non-parametric tests. Parametric tests make assumptions about the data's underlying probability distribution, essentially making assumptions about the parameters that define the distribution (e.g. assuming that the data are Normally distributed). In contrast, non-parametric tests make no assumptions about the specific functional form of the data's distribution. As such, non-parametric tests generally have reduced power but wider applicability in comparison with parametric tests.

In practice, we consider the *Null Hypothesis* – this term is used to refer to any hypothesis that we are interested in disproving (or, more correctly, accumulating evidence for rejection). The converse is referred to as the *Alternative Hypothesis.* These are written:

$$H_0 : \text{statement is true}$$
$$H_1 : \text{statement is not true}$$

By convention, H_0 denotes the null hypothesis and H_1 denotes the alternative hypothesis.

For example, if we wanted to test the hypothesis that a coin is a fair (i.e. the coin lands on heads or tails with equal probability) then we could consider:

$$H_0 : p = 0.5$$
$$H_1 : p \neq 0.5$$

where p is the probability of the coin landing on heads. This could be tested by repeatedly tossing the coin, recording the number of times that the coin landed on heads, and testing H_0 using the Binomial distribution. This would be a one-sample test.

For comparison, a two-sample test might be used if we wanted to test the hypothesis that a two coins are equally fair/unfair (i.e. both coins land on heads with equal probability), in which case we could consider:

$$H_0 : p_1 = p_2$$
$$H_1 : p_1 \neq p_2$$

where p_1 and p_2 are the probabilities of coins 1 and 2 landing on heads, respectively. In this case, we could test the null hypothesis by repeatedly tossing both coins, recording the number of times that each coin landed on heads, and obtaining the probability that both values come from Binomial distributions with equal success probability $p = p_1 = p_2$, given the numbers of trials n_1 and n_2, respectively.

Whether or not the null hypothesis is rejected depends on the *statistical significance* of the test, given by $P(H_0)$, commonly referred to as a *p-value.* A result is considered *significant* if it has been predicted to be highly unlikely to have occurred randomly by chance, given some threshold level. This threshold, often denoted α, is called the *significance level.* The significance level is commonly set to $\alpha = 0.05$, which represents the threshold at which there is only 5% probability that the null hypothesis is correct. If a p-value is found to be less than this value, then the result would be considered statistically significant. However, note that different significance levels may be selected depending on the nature of the application (e.g. a lower α-level may be selected if the incorrect rejection of a hypothesis would lead to human fatalities).

The significance level α is equal to the rate of false positive (type I) errors, called the *size* of the test. The rate of false negative (type II) errors is denoted β.

$$\alpha = P(\text{reject } H_0 | H_0 \text{ is correct})$$

$$\beta = P(\text{do not reject } H_0 | H_0 \text{ is incorrect})$$

The *size* of a test (α) may be controlled by adjusting the significance level. The *power* of a test is equal to $1 - \beta$, and is determined by the nature of the particular statistical test used to test the null hypothesis. Given that non-parametric tests tend to have lower power than parametric tests, non-parametric tests will have a greater tendency to fail to reject the null hypothesis in cases where the null hypothesis is actually incorrect.

	H_0 is correct	H_0 is incorrect
Reject null hypothesis	false positive type I error (α)	true positive
Fail to reject null hypothesis	true negative	false negative type II error (β)

It should be noted that there are two types of tests – one-tailed and two-tailed tests – which correspond to two different ways of computing the significance level (p-value). A two-tailed test considers any values that are extremes of the distribution to be of interest for the purposes of testing the hypothesis, irrespective of whether those values are particularly large or small. In contrast, a one-tailed test is directed, being interested in detecting extreme outliers that are either particularly large or particularly small, but not both.

For example, suppose there are two classes of students that sit a particular exam. A random selection of n students is selected from each of the two classes – these samples are to be used to test hypotheses regarding differences in the performance of each class. A two-tailed test might be used to test the hypothesis that both classes performed equally well in the exam. However, a one-tailed test might be used to test the hypothesis that class 1 performed better than class 2 in the exam.

Acknowledging whether a test is one-tailed or two-tailed is important in determining the probability required to achieve a given level of significance. For example, suppose the outcome of a statistical test is $P(X \leq x) = 0.04$. If the hypothesis test is one-sided (i.e. testing whether the random variable X is no greater than x) then the p-value is 0.04, thus the null hypothesis is rejected. However, if the test is two-sided (i.e. testing whether the random variable X is no more extreme than the value x) then the p-value is 0.08, thus the null hypothesis is not rejected (assuming $\alpha = 0.05$).

When performing a statistical test, a *confidence interval* is often reported. This is the interval in which a test statistic could potentially lie without resulting in the null hypothesis being rejected, given a particular significance level α, and is referred to as the $100 \times (1-\alpha)\%$ confidence interval. For example, if $\alpha = 0.05$, then the 95% confidence interval would be of interest, which would be the interval with boundaries at the 2.5% and 97.5% levels, for a two-tailed test.

There are many different statistical tests, designed for various purposes. For example, when testing protein expression in different conditions, it may be of interest to test whether one condition results in a systematically greater yield than another. In such circumstances, it may be appropriate to test whether the average value of the underlying distribution corresponding to one particular sample is systematically larger/smaller than that of another. Such tests, which are designed to compare measures of centrality, are very commonly used. There are various such tests, intended for use with different types of data, e.g. a single data sample, two independent samples, or two dependent samples (paired, with known correspondences), and different tests depending on what assumptions can be made (e.g. ability to reasonably assume Normality). The following table highlights various statistical tests that may be used in various circumstances, assuming that the objective is to test for differences in the average values of distributions:

	1-sample	2-sample independent	2-sample dependent (paired)
Parametric	t-test	t-test Welch's t-test	paired t-test
Non-parametric	sign test Wilcoxon signed-rank test	median test Mann-Whitney U-test	sign test Wilcoxon signed-rank test

The remainder of this tutorial will provide an introduction to some of the most common statistical tests, which may be used to test various types of hypotheses, with various types of data.

2 Testing distributional assumptions

Testing for Normality

Since some statistical tests require certain assumptions to be satisfied, e.g. the t-test requires the sample to be (approximately) Normally distributed, it is useful to be able to test such distributional assumptions.

The Shapiro-Wilk test tests the null hypothesis that a particular sample can be considered to be Normally distributed, and can be performed in R using the command:

```
shapiro.test(rnorm(10))
```

Here, we test whether a random sample of 10 variates from the $\mathcal{N}(0,1)$ distribution can be considered to be Normally distributed. Since the data were generated from a Normal distribution, the p-value should be large, thus the null hypothesis is not rejected. Now consider performing the test on the squares of standard Normal variates (i.e. the data now follow a χ^2_1 distribution):

```
shapiro.test(rnorm(10)^2)
```

In this case, the p-value should be small, thus allowing the null hypothesis to be rejected.

Remember that in Tutorial 2 we considered the use of Q-Q plots to visually explore relationships between distributions. In particular, the `qqnorm` function was used to compare a sample against the Normal distribution. Such representations provide a visual indication of the nature of the data (i.e. the degree of Normality in this case), allowing insight to be gained, whilst tests such as the Shapiro-Wilk test allow such hypotheses to be tested in a more objective manner, providing quantitative (i.e. test statistic and p-value) and qualitative (significant / not significant) results. Nevertheless, manual visual exploration of the data is always useful, especially for identifying peculiarities in the dataset that would not be automatically detected during the standard course of statistical hypothesis testing.

Testing for equality of distributions

Whilst the Shapiro-Wilk test specifically tests whether a given sample can be considered to be Normally distributed, the Kolmogorov-Smirnov test tests for the equality of two arbitrary distributions (empirical or theoretical). Specifically, it considers the maximum difference between corresponding values of the cumulative distribution functions of the compared distributions, and tests whether such a difference could have arisen by chance.

The Kolmogorov-Smirnov test can be performed in R using the command:

```
ks.test(rnorm(10),rnorm(10))
```

which tests whether two samples of 10 random variates from the $\mathcal{N}(0,1)$ distribution could be from the same distribution. Since the data were both generated from the same distribution, the p-value should be large, thus the null hypothesis is not rejected. Comparing two distributions that are truly different should result in the null hypothesis being rejected, e.g.:

```
ks.test(rnorm(10),rnorm(10,2))
```

Note that the test statistic, the Kolmogorov-Smirnov distance, can be used to quantity the distance between distributions.

Testing for outliers

Grubb's test for outliers tests the null hypothesis that there are no outliers in a given sample, making the assumption:

- The data can reasonably be considered to follow a Normal distribution.

The test considers the maximum absolute difference between observations and the mean, normalised with respect to the sample standard deviation:

$$G = \max_i \left| \frac{x_i - \bar{x}}{s} \right|$$

and considers the chance of such an extreme value occurring given the number of observations, given that the data are Normally distributed.

Grubb's test is available for R in the package *outliers*. To install and load this package, type:

```
install.packages("outliers")
library(outliers)
```

Grubb's test can now be executed using the command:

```
grubbs.test(rnorm(10))
```

In this case, since the data are generated from a standard Normal distribution, the null hypothesis that there are no outliers will not be rejected. However, if we manually insert a value that should be considered an outlier:

```
grubbs.test(c(rnorm(10),5))
```

then Grubb's test will detect the outlier, and the null hypothesis will be rejected.

Note that Dixon's Q-test is a non-parametric outlier detection test also available in the *outliers* R package (function: `dixon.test`), which may be used if the data are not Normal.

Testing for equality of variances between two independent samples

The F-test is the most common test used to test for equality of variance between two independent samples. The F-test tests the null hypothesis that the variances of two samples are equal using the test statistic:

$$F = \frac{s_1^2}{s_2^2} \qquad\qquad F \sim \mathcal{F}_{n_1-1,n_2-1}$$

where s_1 and s_2 are the sample standard deviations, and n_1 and n_2 the number of observations in the two samples, respectively. Consequently, this test is also referred to as the variance ratio test.

The F-test for equality of variances makes the following assumptions:

- Within each sample, the observations are independent and identically distributed (i.i.d.);
- Both data samples are Normally distributed.

Since the F-test is sensitive to the assumption of Normality, it is important for this assumption to be tested prior to application.

An F-test can be performed in R using the `var.test` function. For example, typing:

```
var.test(rnorm(10),rnorm(10))
```

will perform an F-test on two independent samples of 10 random numbers taken from $\mathcal{N}(0,1)$ – the standard Normal distribution – testing the null hypothesis that the variances of the two samples are equal. In fact, it actually tests the hypothesis that the ratio of the two variances is equal to one. Consequently, since the null hypothesis is not rejected, the value '1' is contained within the reported 95% confidence interval.

Note that the F-test is independent of the location of the distributions – it does not test equality of means. Note that changing the mean value of the distribution does not affect the F-test:

```
var.test(rnorm(10),rnorm(10,5))
```

However, altering the variance of one of the compared samples has a dramatic affect on the result of the F-test:

```
var.test(rnorm(10),rnorm(10,0,3))
```

The F-test is useful for assessing equality of variances, assuming Normality has already been ascertained. Determining equality of variances is useful when attempting to perform other tests that assume equal variances, such as the two-sample t-test.

In cases where the testing for equality of variances is required, but data cannot be considered to be Normal, other tests less sensitive to the assumption of Normality could be considered (e.g. Levene's test, Bartlett's test, or the Brown-Forsythe test).

Task 1:

In R, the inbuilt dataset 'CO2' contains data from an experiment on the cold tolerance of the grass species *Echinochloa crus-galli.* The dataset records the carbon dioxide uptake rates (response), ambient carbon dioxide concentration (independent variable), and three factors (Plant, Type and Treatment).

1. Consider the three Plants that are both Chilled and from Mississippi – these are labelled 'Mc1', 'Mc2' and 'Mc3'. Extract the data corresponding to these Plants, and, separately for each of the three plants, perform statistical tests to test whether they can be considered to be Normally distributed.

2. For each of the three plants, perform statistical tests to detect any outliers, ensuring that assumptions are satisfied for any statistical tests performed. Which of the plants have corresponding distributions that exhibit at least one outlier?

Caution – in this task we have simultaneously performed multiple hypothesis tests. This is dangerous, as it increases the chances of randomly observing a significant result (i.e. type I error). For example, suppose that 20 hypothesis tests are performed, then clearly it is quite possible that at least one of the tests is significant at the 95% level, purely by chance. In order to account for this effect, we would usually use the *Bonferroni correction*, which essentially involves using a higher α-level (significance threshold) in order to account for the fact that we are performing multiple hypothesis tests. Specifically, α is divided by the number of tests being performed. For instance, in the above example three tests are simultaneously performed. Consequently, the significance level α would be reduced from 0.05 to 0.0167.

3 One-sample tests

One-sample t-test

The one-sample t-test tests the null hypothesis that the population mean is equal to some value μ_0.

The test statistic is:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \qquad\qquad t \sim \mathcal{T}_{n-1}$$

where $\bar{x}$ is the sample mean, s is the sample standard deviation, and n is the number of observations.

Note the similarity between the formula for the test statistic and that of a z-score – in computing the t-test statistic the data are normalised with respect to the hypothesised mean (not the sample mean!) and the sample variance. Note also that, according to the Central Limit Theorem, if μ_0 is the true population mean then the distribution of the test statistic converges to the standard Normal distribution as $n \to \infty$.

The one-sample t-test essentially makes the assumptions:

- The observations $x_1, \ldots, x_n$ are independent and identically distributed (i.i.d.);
- The data are approximately Normally distributed (or large sample).

The latter condition is flexible to some degree; t-tests can be used on non-Normal data providing the data are not too non-Normal, e.g. the distribution is unimodal and symmetric. For non-Normal data, the reliability of the test increases as the number of observations increases (which causes the sample mean to become more Normal). However, if the data are non-Normal then a non-parameteric test may be preferred.

A t-test can be performed in R using the `t.test` function. For example, typing:

```
t.test(rnorm(10))
```

will perform a one-sample t-test on a sample of 10 random numbers taken from $\mathcal{N}(0,1)$ – the standard Normal distribution – testing the null hypothesis that the mean of the sample is equal to zero. Of course, in this case we know that the 'true' mean is zero, as the sample has been taken from the standard Normal distribution. Consequently, we would expect a p-value greater than 0.05, indicating no evidence with which to reject the null hypothesis. Further to providing a t-test statistic and associated p-value, note that the R output from the `t.test` function call also includes the 95% confidence interval, and the sample mean. Note that the 'true' mean (0) is indeed contained within the 95% confidence interval.

For comparison, now consider a t-test performed on a sample of 10 random numbers taken from $\mathcal{N}(1,1)$, again testing the hypothesis that the population mean is zero, which can be performed using the command:

```
t.test(rnorm(10,1))
```

This time, the mean should not be equal to 0, given that we have artificially generated numbers from a distribution whose mean is 1. Indeed, inspecting the output of the command should indicate that the null hypothesis is rejected, with a p-value less than the $\alpha = 0.05$ threshold, noting that the value '0' is outside the confidence interval. Note also that increasing the mean of the sample higher than 1 would result in smaller p-values (i.e. higher significance levels), and also that inflating the variance higher than 1 would result in larger p-values (due to the increased uncertainty).

One-sample sign test

An alternative to the one-sample t-test is the one-sample sign test, which is a simple non-parametric test that makes very few assumptions about the nature of the data, namely:

- The observations $x_1, \ldots, x_n$ are independent and identically distributed (i.i.d.).

Whilst the t-test allows the testing of a hypothesis regarding the value of the mean, the sign test tests a hypothesis regarding the value of the median (a more robust statistic).

The sign test counts the number of observations greater than the hypothesised median m_0, and calculates the probability that this value would result from a

$\text{Bin}(n, 0.5)$ distribution, where n is the number of observations not equal to m_0. As such, the one-sample sign test is also referred to as the Binomial test.

For example, suppose we want to use the sign test to test the null hypothesis that the median of a random sample of 10 variates from the $\mathcal{N}(0,1)$ distribution is equal to zero. This can be done in R using the commands:

```
x=rnorm(10))
binom.test(sum(x>0),length(x))
```

Here, we use the `binom.test` function to test whether the median of `x` could be equal to 0. As expected, the test should result in the null hypothesis not being rejected, with a p-value much larger than the 0.05 threshold.

Repeating the test, this time testing whether the median could be equal to 1, will most often result in a significant result:

```
x=rnorm(10))
binom.test(sum(x>1),length(x))
```

However, if more than one out of the ten standard Normal variates randomly have a value greater than 1 then the null hypothesis will not be rejected (indicating a type II error) due to the low power of the test.

One-sample Wilcoxon signed-rank test

Another non-parametric alternative to the one-sample t-test is the one-sample Wilcoxon signed-rank test, which makes more assumptions regarding the nature of the data than the sign test, thus has increased power. Specifically:

- The observations $x_1, \ldots, x_n$ are independent and identically distributed (i.i.d.);
- The distribution of the data is symmetric.

The one-sample Wilcoxon signed-rank test ranks the observations according to their absolute differences from the hypothesised median $|x_i - m_0|$, and uses the sums of the ranks corresponding to the positive and negative differences as test statistics.

The one-sample Wilcoxon signed-rank test may be performed in R using the command:

```
wilcox.test(rnorm(10))
```

which again tests the null hypothesis that the median of a random sample of 10 variates from the $\mathcal{N}(0,1)$ distribution is equal to zero. The null hypothesis should not be rejected in the majority of cases.

Now perform a one-sample Wilcoxon signed-rank test on a sample of 10 random numbers taken from the $\mathcal{N}(1,1)$ distribution, again testing the hypothesis that the population mean is zero:

```
wilcox.test(rnorm(10,1))
```

This should result in a significant result, thus rejecting the null hypothesis, in the majority of cases.

Task 2:

In R, the inbuilt dataset ‘LakeHuron’ contains data corresponding to annual measurements of the level of Lake Huron (in feet).

1. Visually inspect the data by creating a histogram and Normal Q-Q plot, in order to gain insight regarding the nature of the data.

2. Test whether the data can be considered to be Normally distributed.

3. Use an appropriate statistical test to test the null hypothesis:

   $$H_0 : \mu = 578$$

   where μ is the population mean corresponding to the data. Can this hypothesis be rejected?

4. According to the test used in the previous step, what is the 95% confidence interval corresponding to estimated distribution of μ?

5. List all integer values that could reasonably be equal to μ, according to your results from the previous steps.

Task 3:

In R, the inbuilt dataset ‘Nile’ contains data corresponding to annual flow of the river Nile at Ashwan.

1. Visually inspect the data by creating a histogram and Normal Q-Q plot, in order to gain insight regarding the nature of the data.

2. Test whether the data can be considered to be Normally distributed.

3. Use an appropriate statistical tests to test the null hypotheses:

   $$H_0 : \mu = 850$$

   and

   $$H_0 : \mu = 950$$

   where μ is the population mean corresponding to the data. Can these hypotheses be rejected?

4. Use (1) the one-sample t-test, and (2) the one-sample Wilcoxon signed-rank test, to test the hypothesis:

   $$H_0 : \mu = 880$$

   where μ is the population mean corresponding to the data. Is this hypothesis rejected by none, one, or both of the tests?

5. Which of the two tests would you use to test the hypothesis considered in the previous step? Discuss the pros and cons associated with using each test to draw conclusions in this particular case.

4 Testing for differences between two independent samples

Independent two-sample t-test

The two-sample version of the t-test, designed for use with two independent samples, tests the null hypothesis that the population means of the two groups are equal.

The test statistic is:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_{12}\sqrt{n_1^{-1} + n_2^{-1}}} \qquad\qquad t \sim \mathcal{T}_{n_1+n_2-2}$$

where $\bar{x}_1$ and $\bar{x}_2$ are the means of the two samples, s_{12} is an estimator of the common standard deviation of the two samples, and n_1 and n_2 are the numbers of observations in the two samples, respectively.

The independent two-sample t-test essentially makes the assumptions:

- Within each sample, the observations are independent and identically distributed (i.i.d.);
- Both data samples are approximately Normally distributed;
- The data samples have the same variance.

Consequently, further to requiring Normality, the independent two-sample t-test also requires that the compared samples can be considered to have the same variance. This assumption can be tested, e.g. using an F-test.

Indeed, it is always important to (1) test for Normality, and (2) test for equal variances, before performing a two-sample t-test. If assumptions are violated, then other statistical tests should be used to test the hypothesis. For example, if the Normality assumption is violated then a non-parametric test could be used, and if the equal variance assumption is violated then Welch's t-test (see below) could be used instead of the standard variant.

A t-test can be performed in R using the `t.test` function. For example, typing:

```
t.test(rnorm(10),rnorm(10),var.equal=TRUE)
```

will perform a two-sample t-test on two independent samples of 10 random numbers taken from $\mathcal{N}(0, 1)$ – the standard Normal distribution – testing the null hypothesis that the means of the two samples are equal (the fact that the means happen to be zero in this case is irrelevant). The `var.equal=TRUE` argument specifies to assume that the variances of the compared sample can be considered to be equal.

Since the data are generated from the same distribution, the t-test should not reject the null hypothesis that the means are equal (i.e. the p-value should be greater than 0.05). Note that the reported 95% confidence interval corresponds to the difference between means, in contrast with the one-sample t-test.

Note that this is the same function as used for a one-sample t-test – the `t.test` function is context-dependent, performing a one-sample t-test if one vector (data sample) is provided, and performing a two-sample t-test if two vectors are provided as input arguments.

For comparison, now consider a t-test performed on two samples of 10 random numbers taken from $\mathcal{N}(0,1)$ and $\mathcal{N}(1,1)$, respectively. The hypothesis that the population means are equal can be tested using the command:

```
t.test(rnorm(10),rnorm(10,1),var.equal=TRUE)
```

This time, the means should not be equal, given that we have artificially generated numbers from distributions with different means. Indeed, inspecting the output of the command should generally indicate that the null hypothesis is rejected, with a p-value less than the $\alpha = 0.05$ threshold, noting that the value '0' is outside the confidence interval (i.e. the difference between the population means is unlikely to be equal to zero).

Welch's t-test

Welch's t-test is a generalisation of the independent two-sample t-test that doesn't assume that the variances of the two data samples are equal. Consequently, the test statistic is:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

where $\bar{x}_1$ and $\bar{x}_2$ are the means, s_1 and s_2 the standard deviations, and n_1 and n_2 the numbers of observations in the two samples, respectively.

Welch's two-sample t-test makes the assumptions:

- Within each sample, the observations are independent and identically distributed (i.i.d.);
- Both data samples are approximately Normally distributed.

This version of the t-test can be performed in R by omitting the '`var.equal=TRUE`' argument, e.g.:

```
t.test(rnorm(10),rnorm(10))
```

Relative to the independent two-sample t-test, relaxation of the equal variance criterion in Welch's t-test results in reduced power, but wider applicability.

Mann-Whitney U-test

Similar to how the Wilcoxon signed-rank test is a non-parametric analogue of the one-sample t-test, the Mann-Whitney U-test test is a non-parametric analogue of the independent two-sample t-test that tests whether the medians of the compared samples can be considered to be equal.

The test makes the assumptions:

- Within each sample, the observations are independent and identically distributed (i.i.d.);
- The distributions of both data samples are symmetric.

The Mann-Whitney U-test, which is also referred to as the two-sample Wilcoxon signed-rank test, may be performed in R using the command:

```
wilcox.test(rnorm(10),rnorm(10))
```

which tests the null hypothesis that the median of two random samples of 10 variates from the $\mathcal{N}(0, 1)$ distribution are equal. In this case, since the data are generated from identical distributions, the null hypothesis should not be rejected.

Task 4:

Recall R's inbuilt dataset 'CO2', which contains data from an experiment on the cold tolerance of the grass species *Echinochloa crus-galli.* The dataset records the carbon dioxide uptake rates (response), ambient carbon dioxide concentration (independent variable), and three factors (Plant, Type and Treatment).

1. Visually inspect the CO2 uptake data (i.e. the `CO2$uptake` vector) by creating a histogram and Normal Q-Q plot, in order to gain insight regarding the nature of the data.

2. Test whether CO2 uptake can be considered to be Normally distributed.

3. Now consider the data corresponding only to concentrations (`conc`) less than 300 mL/L.

   (a) Create side-by-side box plots of the CO2 uptake corresponding to the two levels of Treatment (i.e. 'chilled' and 'nonchilled'), for observations for which concentration is less than 300 mL/L.

   (b) Can the two samples corresponding to 'chilled' and 'nonchilled' plants (i.e. the two sets of data displayed as box plots in the previous step) be considered to be Normally distributed?

   (c) Can the variances of these two samples be considered to be equal?

   (d) Perform an appropriate test to determine whether the average CO2 uptake for nonchilled plants is significantly greater than for chilled plants, for concentrations less than 300mL/L.

4. Now consider the data corresponding only to concentrations greater than 300 mL/L.

   (a) Create side-by-side box plots of the CO2 uptake corresponding to the two levels of Treatment (i.e. 'chilled' and 'nonchilled'), for observations for which concentration is greater than 300 mL/L.

   (b) Test whether these two samples corresponding to 'chilled' and 'nonchilled' plants can be considered to be Normally distributed.

   (c) Perform an appropriate test to determine whether the average CO2 uptake for nonchilled plants is significantly greater than for chilled plants, for concentrations greater than 300mL/L.

5. Now consider the data corresponding only to concentrations greater than 400 mL/L.

   (a) Create side-by-side box plots of the CO2 uptake corresponding to the two levels of Type (i.e. 'Quebec' and 'Mississippi'), for observations for which concentration is greater than 400 mL/L.

(b) Test whether these two samples corresponding to 'Quebec' and 'Mississippi' plants can be considered to be Normally distributed.

(c) Perform an appropriate test to determine whether the average CO2 uptake is significantly different in the plants from Quebec and Mississippi, for concentrations greater than 400mL/L.

6. Reflect on your results from parts 3, 4 and 5 of this task – i.e. consider in which of the different groups the average CO2 uptake was found to be significantly different, and for which groups no differences were detected. Which results do you "believe"? Were all results conclusive? Use your observations from visual inspection of the box plots in order to support your conclusions.

5 Testing for differences between two dependent (paired) samples

Paired two-sample *t*-test

In cases where there is a known correspondence between the two compared samples, it is necessary to account for the fact that the observations between the samples are not independent when performing statistical tests. Such correspondences may exist because the observations correspond to the same individuals (i.e. repeated measurements) or simply because the samples have been matched in some way. In such circumstances, it is possible to test for differences between the samples accounting for such dependencies. In such cases, the quantities of interest are the differences between the paired observations.

The paired two-sample *t*-test, designed for use with two independent samples, tests the null hypothesis that the population means of the two groups are equal.

The test statistic is:

$$t = \frac{\bar{x}_\Delta - \mu_0}{s_\Delta/\sqrt{n}} \qquad\qquad t \sim \mathcal{T}_{n-1}$$

where $\bar{x}_\Delta$ and s_Δ are the mean and standard deviation of the differences between the two samples, respectively (compare with the one-sample *t*-test).

The paired two-sample *t*-test makes the assumptions:

- Within each sample, the observations are independent and identically distributed (i.i.d.);
- The distribution of the paired differences is approximately Normally distributed;

The paired two-sample *t*-test can be performed in R by supplying the `paired=TRUE` argument to the `t.test` function, e.g.:

```
x=rnorm(10)
y=rnorm(10)
t.test(x,y,paired=TRUE)
```

noting that this command is equivalent to:

```
t.test(x-y)
```

since the paired two-sample t-test is effectively a one-sample t-test performed on the distribution of differences between paired observations.

Two-sample sign test

A non-parametric alternative to the paired two-sample t-test is the two-sample sign test. Similarly to how the paired two-sample t-test is effectively a one-sample t-test performed on the distribution of differences between paired observations, the two-sample sign test is effectively a one-sample sign test performed on the distribution of differences between paired observations.

Being a non-parametric test, the two-sample sign test makes very few assumptions about the nature of the data:

- Within each sample, the observations are independent and identically distributed (i.i.d.).

The two-sample sign test counts the number of differences between paired observations that are greater than zero, and calculates the probability that this value would result from a $\text{Bin}(n, 0.5)$ distribution, where n is the number of differences not equal to zero.

For example, suppose we want to use the sign test to test the null hypothesis that the medians of two paired samples of 10 random variates from the $\mathcal{N}(0, 1)$ distribution are equal. This can be done in R using the commands:

```
x=rnorm(10))
y=rnorm(10))
binom.test(sum(x>y),length(x))
```

Here, we use the `binom.test` function to test whether the median of `x` could be equal to the median of `y`. As expected, the test should result in the null hypothesis not being rejected, with a p-value much larger than the 0.05 threshold.

Paired Wilcoxon signed-rank test

Another non-parametric alternative to the paired two-sample t-test is the paired Wilcoxon signed-rank test. Similarly to with the two-sample sign test, the paired Wilcoxon signed-rank test is effectively a one-sample Wilcoxon signed-rank test performed on the distribution of differences between paired observations. The test makes the assumptions:

- Within each sample, the observations are independent and identically distributed (i.i.d.);
- The distribution of the paired differences is symmetric.

The paired two-sample Wilcoxon signed-rank test may be performed in R using the command:

```
x=rnorm(10)
y=rnorm(10)
wilcox.test(x,y,paired=TRUE)
```

which again tests the null hypothesis that the medians of two samples of 10 random variates from the $\mathcal{N}(0,1)$ distribution are equal, noting that this command is equivalent to:

```
wilcox.test(x-y)
```

which is a one-sample Wilcoxon signed-rank test performed on the distribution of differences between paired observations.

Task 5:

Recall R's inbuilt dataset 'CO2', which contains data from an experiment on the cold tolerance of the grass species *Echinochloa crus-galli.* The dataset records the carbon dioxide uptake rates (response), ambient carbon dioxide concentration (independent variable), and three factors (Plant, Type and Treatment).

Suppose we want to test whether the CO2 uptake is significantly larger for concentrations of 1000 mL/L than for 675 mL/L.

1. Create box plots to display the distributions of CO2 uptake for the data corresponding to concentrations of 1000 mL/L and 675 mL/L (i.e. create two side-by-side box plots). Does it appear that the uptake is substantially larger for concentrations of 1000 mL/L than for 675 mL/L?

2. Test whether these two distributions can be considered to be Normally distributed.

3. Test whether these two distributions can be considered to have equal variances.

4. Perform an independent two-sample *t*-test to compare the means of these distributions. Can they be considered to be significantly different? From this test, can we conclude that CO2 uptake is substantially larger for concentrations of 1000 mL/L than for 675 mL/L?

5. Note that each observation in each of the two samples corresponds to a different Plant (i.e. a different individual). Note also that there is a direct correspondence between observations in the 1000 mL/L sample and the 675 mL/L sample, and that the corresponding observations have the same indices in their respective vectors. Perform an appropriate statistical test to test whether the CO2 uptake is significantly larger for concentrations of 1000 mL/L than for 675 mL/L. Is it possible to conclude that CO2 uptake is significantly larger for concentrations of 1000 mL/L than for 675 mL/L?