Foundations of Medical Statistics
Statistical Inference 6: Construction of a hypothesis test

Aims

The aim of this session is to introduce the concepts underlying hypothesis testing, 'best' hypothesis tests; and to revisit the definition of p-values.

Objectives
At the end of this session you should:

- understand the ideas of type I and type II error;
- understand the idea of power;
- understand what makes a 'best' test;
- understand the difference between simple and composite hypothesis tests, and how we justify doubling the p-value for two-sided tests;
- understand how to quantify the evidence against simple hypotheses in novel situations.

6.1 Hypothesis testing

In general, a **hypothesis** is a statement about the probability distribution of the population from which the data we observe are drawn. Usually, in the context of **parametric** tests, this takes the form of statements about the population value of parameters. Typically, two complementary hypotheses are stated, called the **null** hypothesis (denoted H_0) and the **alternative** hypothesis (denoted H_1or H_A).

For example, suppose X is a random variable with a binomial distribution $Bin(5, \theta)$, and consider the null hypothesis H_0: $\theta = ½$, and an alternative hypothesis H_1: $\theta = \frac{2}{3}$. Since, in this example, we completely specify the distribution of the data under both H_0 and H_1, these are termed a **simple hypotheses**. In the early part of this session we will restrict ourselves to this somewhat unnatural situation; later we shall consider more complex **composite hypotheses**, for example of the form $\theta > \frac{2}{3}$.

Having specified our null and alternative hypotheses, a **hypothesis test** is a rule that specifies:

1. for which sample values H_0 is *rejected* (and hence H_1 is *accepted* as true)

2. for which sample values H_0 is *not rejected* (but H_0 is not necessarily 'accepted' as true).

   Note: In a context where H_0 and H_1 are both simple hypotheses, there may be no alternative but to accept the null when it is not rejected. But it is important to note that generally in practice there is an asymmetry, with H_0 simple and H_1 composite. In such a context we do not necessarily accept the null hypothesis when it is not rejected. *Absence of evidence is not evidence of absence*: in other words, absence of evidence against the null hypothesis is not of itself evidence that the null hypothesis is true, since generally in practice the null hypothesis is not the only hypothesis 'compatible' with the sample values, when it is not rejected; on the other hand, when the null hypothesis is rejected, generally the only hypothesis compatible with the data is the alternative, so then evidence against the null is also evidence for the alternative hypothesis.

The subset of the sample space for which H_0 will be rejected is called the **rejection region** or **critical region**, and is denoted by $\mathfrak{R}$.

6.2 Error probabilities and the power function

SAMPLE

		$\underline{x} \notin \mathfrak{R} \Rightarrow$ **Do not reject** H_0	$\underline{x} \in \mathfrak{R} \Rightarrow$ **Reject** H_0
TRUTH	**H_0 is true**	✓	Type I error
	H_1 is true	Type II error	✓

Suppose the hypotheses are about a parameter θ:

H_0: $\theta = \theta_0$ vs H_1: $\theta = \theta_1$,

Then the **power** of the test is defined as the probability of rejecting the null hypothesis when the alternative hypothesis is true:

power = Prob($\underline{x} \in \mathfrak{R}$ | H_1 is true) = 1 – Prob(Type II error)

(since the data is either in the rejection region or not). The probability of type II error is conventionally denoted by $\boldsymbol{\beta}$, so power = 1 – β.

Conversely, the **statistical size** of a test (*not to be confused with sample size*!) is defined as

Prob($\underline{x} \in \mathfrak{R}$ | H_0 is true) = Prob(Type I error)

The size of the test is conventionally denoted by $\boldsymbol{\alpha}$. Thus α is the probability of the test giving a 'false positive' (wrongly stating that H_1 is true) while β is the probability of failing to detect a true H_1.

Example *6.2.1 Binomial distribution*
As above, $X \sim Bin(5, \theta)$, and consider testing H_0 : $\theta = \frac{1}{2}$ vs H_1: $\theta = \frac{2}{3}$.
Consider two tests:

Test A: reject H_0 if and only if 5 'successes' are observed i.e. the observed value of X is 5). Thus $\mathfrak{R}$ is $X = 5$. The power of Test A is Prob($X = 5|H_1$ true), which is $(\frac{2}{3})^5 = 0.1317$. The size is Prob($X = 5|H_0$ true) $= (\frac{1}{2})^5 = 0.03125$.

> *Test B*: reject H_0 if the observed value of X is either 3, 4 or 5. Thus $\mathfrak{R}$ is $X = 3$, 4 or 5. The power of Test B is Prob(X = 3, 4 or 5|H_1 true), which is
>
> $$\sum_{i=3}^{5} \left(\frac{2}{3}\right)^{i} \left(\frac{1}{3}\right)^{5-i} \binom{5}{i} \approx 0.7901$$
>
> while the size is Prob(X = 3, 4 or 5|H_0 true) = 0.5.

Comparing Test A and B, we are more likely to wrongly reject H_0 using test B, but also more likely to correctly reject H_0 (hence accept H_1) using test B.

This is an example of a general phenomenon: as we increase the power to detect H_1, we also increase the size, i.e. the probability of a type I error.

To resolve this, in practice the type I error probability is fixed, typically at $\alpha = 0.05$, and then the test is chosen to make the power as high as possible given this fixed size.

6.3 Choice of a test statistic

This leads to the next question: what should our test statistic be? In the binomial distribution above, it was X, the number of successes. But it could have been the number of repeats before the first success, or indeed, any statistic which is related to the hypotheses. Similarly, if the data are continuous and we have several observations, we could use the mean, median, mode, geometric mean etc..

Fortunately, for testing *simple* hypotheses, where H_0 and H_1 are each completely specified, then the *Neyman-Pearson lemma* (proof Rice p. 332; Cox and Hinkley, p. 92) tells us that the best, i.e. **most powerful**, test for a given size α rejects H_0 for small values of the likelihood ratio

$$\frac{L_{H_0}}{L_{H_1}}$$

This is intuitive, as then we reject H_0 if the data are much more likely under H_1 than H_0. Typically, the value of the likelihood ratio only depends on a particular statistic: this then is the best test statistic.

Since the likelihood ratio is only small if its logarithm is small, we usually work in terms of the log of the likelihood ratio, $l_{H_0} - l_{H_1}$.

> **Note**: do not confuse this likelihood ratio and its logarithm with the log-likelihood ratio as defined in 4.1. In the latter, the term is specialised so that the denominator is always specifically the likelihood at the maximum, while here we are talking simply about a ratio of two likelihoods.

How small is small? The threshold beyond which we reject is determined by the sampling distribution of the test statistic, and is chosen to give the required value of α.

Example *6.3.1 Mean of normal distribution with known variance*

Suppose $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$ and σ^2 is known. Let H0: $\mu = 5$ and H1: $\mu = 10$. We will now find the 'best' statistic. Recall from 4.2 that

$$l(\mu|\underline{x}) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2.$$

Then, as σ^2 is known and the same under both hypotheses,

$$l_{H_0} - l_{H_1} = -\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}(x_i - 5)^2 - \sum_{i=1}^{n}(x_i - 10)^2\right).$$

However, we only need to consider quantities which vary with the data, so we can ignore fixed constants. We will therefore reject H0 for small values of

$$= -\left(\sum_{i=1}^{n}(x_i - 5)^2 - \sum_{i=1}^{n}(x_i - 10)^2\right) = 75n - 2 \times (10 - 5)\sum_{i-1}^{n} x_i.$$

And again, ignoring the fixed constants, we reject H0 for *large* values of $\sum_{\mathrm{i=1}}^{\mathrm{n}} x_i$, or of any constant multiple of this. It is convenient to use the constant multiple $\frac{1}{n}\sum_{\mathrm{i=1}}^{\mathrm{n}} x_i$, the sample mean: thus the best test of H0: $\mu = 5$ vs H1: $\mu = 10$ rejects H0 for large values of the sample mean. It also follows that tests based on the mode, median or geometric mean, say, would all be less powerful and waste sample information.

Furthermore, we know the sampling distribution of the sample mean in this context: this will allow us to construct a suitable rejection region (see Example 6.5.1).

Note: Remember that $l_{\mathrm{H}_0} - l_{\mathrm{H}_1}$ is a random variable: the data varies each time we sample, and we will get a different amount of relative support for the hypotheses. We are therefore only interested in that part of $l_{\mathrm{H}_0} - l_{\mathrm{H}_1}$ which depends on quantities that vary with each sample. The constant part provides no information on the relative support the data give to the hypotheses, so we ignore it.

6.4 Composite hypotheses

The situation we have described so far is too artificial to be of much practical use. In practice, we are usually interested in hypotheses such as

1) H0 : $\theta = \theta_0$ vs H1: $\theta > \theta_0$ [**one-sided** alternative hypothesis]
2) H0 : $\theta = \theta_0$ vs H1: $\theta \neq \theta_0$ [**two-sided** alternative hypothesis]

In both cases we are faced with a problem: it is not clear that the same test statistic will be most powerful for all alternative values of the parameter.

We consider cases 1) and 2) separately.

6.4.1 Case 1: the one-sided alternative hypothesis

Often there is a single test statistic that is most powerful for each $\theta > \theta_0$. Such a test statistic is called **uniformly most powerful**.

Example *6.4.1* *Mean of normal distribution with known variance*

In the example above (*6.3.1*), we saw that large values of the sample mean imply small values of $l_{H_0} - l_{H_1}$, so that the most powerful test rejects H$_0$ for large values of the sample mean for all values of μ under H$_1$ that are greater than $\mu_0 = 5$. In other words, instead of H$_1$: $\mu = 10$, any value of μ greater than 5 would still result in an expression obtained from the log likelihoods showing that larger values of the sample mean provide less support for H$_0$.

Thus, in this example, there is a uniformly most powerful test of H$_0$: $\mu = 5$ vs H$_1$: $\mu > 5$: the test that rejects H$_0$ for large values of the sample mean.

If there is no uniformly most powerful test, we would usually use our scientific knowledge of a problem to identify a particular $\theta > \theta_0$ and choose a test that is most powerful for that particular value.

6.4.2 Case 2: the two-sided alternative hypothesis

No uniformly most powerful test can exist in this situation.

To see this, we return to the example of normal distribution. We have seen that for H$_0$: $\mu = 5$ vs H$_1$: $\mu > 5$, the most powerful test rejects H$_0$ for large values of the sample mean.

An analogous argument shows that if we wish to test H$_0$: $\mu = 5$ vs H$_1$: $\mu < 5$, the most powerful test rejects H$_0$ for *small* values of the sample mean.

Here lies the problem: the most powerful test is different when we test H$_1$: $\mu > 5$ from when we test H$_1$: $\mu < 5$. We return to this issue in Section 6.6 below. First, though, we consider how to quantify evidence against H$_0$.

6.5 Quantifying evidence against H$_0$: the one-sided p-value

Consider testing the composite hypothesis H$_0$: $\theta = \theta_0$ vs H$_1$: $\theta > \theta_0$, and suppose a uniformly most powerful test exists, using test statistic T (more formally, a function $T(x)$ of the sample data), large values of which reject H$_0$. Recall that we need to define the rejection region for a fixed Type I error α:

$$\text{Prob}(\underline{x} \in \Re \mid \text{H}_0) = \alpha.$$

If we know the sampling distribution for T, we can readily determine the rejection region using a threshold c such that

Prob($T \geq c$| H$_0$) = α.
(More formally, we define $\Re$ as {$\underline{x}$:Prob($T(\underline{x}) \geq c$| H$_0$) = α)}.)

In other words, we reject H$_0$ if $T > c$. However, if we are not required to make a binary decision, "reject or not reject", we can use a continuous measure to quantify the evidence against the null hypothesis. Consider the observed value of the test statistic, its realisation t. Note that larger values of t are more 'extreme' with reference to H$_0$. We can then use the sampling distribution of T to calculate the probability of observing data at least as extreme, in terms of probability, as that observed:

p = Prob($T \geq t$| H$_0$)

This is known as the **one-sided p-value**. The smaller the p-value, the smaller the proportion of the sample space of the statistic which is less favourable to the null hypothesis. In other words, you cannot get observed data much more unfavourable to H$_0$ than that which yields a very small p-value: the smaller the p-value, the more evidence provided by the data against the null hypothesis.

We can also use the p-value in the context of formal hypothesis testing to reject or not reject H$_0$, since

$p < \alpha \quad \Leftrightarrow \quad t > c.$

For example, if we set $\alpha = 0.05$ (following convention), then if $p < 0.05$ the test rejects H$_0$ at the 5% level.

Note: Why is the p-value a tail area of the distribution, rather than just the value of the distribution at the relevant point of the sample space? Here are some reasons which contribute to the usefulness of the tail area. For a continuous distribution, the value at any point is a density, not a probability, which would be a problem: it could take values greater than 1, for example; so we need an area. But why not an arbitrary small region around the point on the sample space, giving us the probability of a result being in that small region, a 'local' probability? But even if we used a local probability, there would be no calibration: different shaped distributions would give us different local probabilities without a standardising element which was meaningful in all, or most distributions. Moreover, we don't just want the value of the local probability, we want to know the relative value, how small or large is it relative to local probabilities under that distribution for other values in the sample space; neighbouring points on the sample space could provide similar local probabilities, or these could be changing rapidly; and we are generally interested in the probability not of obtain a particular result, but a similar result falling in that region. By looking at the tail area beyond the result observed, we are able to calibrate the meaning to most distributions and compare the result with the rest of the sample space in terms of how much evidence it provides against the null hypothesis.

***Example** 6.5.1 Mean of normal distribution with known variance (continued)*

Consider $X_1, \dots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, where we know $\sigma^2 = 10$. The aim is to test

H$_0$: $\mu = 5$ vs H$_1$: $\mu > 5$

1. First we need to identify our test statistic. We have seen above that a uniformly most powerful test statistic exists if σ^2 is known: the statistic is the sample mean.
2. We now need the sampling distribution of the sample mean, which we know is:

$$\bar{X} \sim N(\mu, \sigma^2/n)$$

$$\Rightarrow \quad Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

so under H$_0$: $Z = \frac{\bar{X}-5}{\sqrt{10}/\sqrt{n}} \sim N(0,1)$

3. For a test with size α = 0.05, we choose the critical region $\bar{X} \geq c$ such that the type I error probability is 0.05:

   Prob($\bar{X} \geq c$| H$_0$) = 0.05.

Now, Prob(Z ≥ 1.64) = 0.05 = Prob$\left(\frac{\bar{X}-5}{\sqrt{10}/\sqrt{n}} \geq 1.64\right)$

Suppose to simplify that n = 10: then the critical region $\mathfrak{R}$ is $\bar{X} \geq 6.64$.

4. Suppose we observe $\bar{X} = 7.76$. Since our test statistic is in the rejection region, the most powerful test rejects H$_0$ in favour of H$_1$.

5. We quantify the evidence against H$_0$ with the one-sided p-value, using the known distribution of Z (here $Z = \bar{X} - 5$):

   p = Prob($\bar{X} \geq 7.76$|H$_0$) = Prob($5 + Z \geq 7.76$) = Prob($Z \geq 2.76$) = 0.003

   Thus 0.3% of possible results would give at least as much evidence as the observed result against H$_0$ and in favour of H$_1$.

6.6 Quantifying evidence against H$_0$: the two-sided p-value

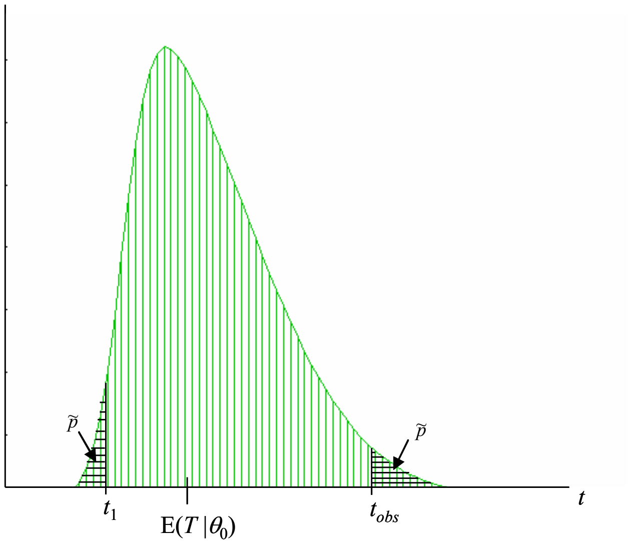

Figure 6.1

We here deliberately use an asymmetrical distribution to highlight the issues.

Suppose now we wish to test the composite alternative hypothesis:

H$_0$: $\theta = \theta_0$ vs H$_1$: $\theta \neq \theta_0$

Formally, this is analysed by splitting into two one-sided tests, with alternatives

i) H$_1$: $\theta > \theta_0$ and ii) H$_1$: $\theta < \theta_0$, each of which may have a uniformly most powerful test. Here we give a less formal account.

Let T be the test statistic uniformly most powerful for testing i), with a sampling distribution for example as in Figure 6.1 above.

There are two approaches; **the first approach is to double the observed one-sided p-value**, which can be justified informally as follows.

Suppose we observe t_{obs}. Then a result probabilistically at least as unfavourable as this to H$_0$ can occur in two ways:

i) $T \geq t_{\text{obs}}$, where Prob($T \geq t_{\text{obs}}$|H$_0$) = $\tilde{p}$.

Recall that $\tilde{p}$ quantifies the evidence provided by our observed result against H$_0$. Hence the second way we could see something as unfavourable is if:

ii) $T \leq t'$, where we choose t' such that Prob($T \leq t'$ |H$_0$) = $\tilde{p}$ (see Figure 6.1).

Therefore we calculate the **two-sided p-value**, being the total probability of observing a result at least as unfavourable to H$_0$ as t_{obs}, as $p = 2\tilde{p}$. In effect, we reject for large values of $|T|$.

Note:

1. t' is chosen *not* to give equal 'distance': $| t' - \text{E}(\text{T}|\theta_0)| = | t_{\text{obs}} - \text{E}(\text{T}|\theta_0)|$ (which would only be the case in *symmetrical* sampling distributions), but in probability terms, to give the same size of tail as that defined by t_{obs}.

2 In discrete distributions, such an opposite tail may not exist. In this case the observed tail is the only such unfavourable region, and then $p = \tilde{p}$. Also, in discrete distributions, there may be problems obtaining 'exact' p-values (see Analytical Techniques 3). A related issue is that for a hypothesis test with a discrete distribution, it may not be possible to obtain the nominated Type I error probability exactly when defining the rejection region. In this case a Type I error is chosen that is as large as possible without exceeding the nominated level.

The second approach, which is the one used in Analytical Techniques 3, is to construct the one-sided tail as in (i) above, obtaining $\tilde{p}$; then identify in the other tail a probability density (or probability in a discrete distribution, which may not always exist) equal to that of the observed result, and add the resulting tail to that in (i). In other words, find t'' in the other tail (i.e. < t_{obs}) such that Prob function($T = t''$ |H$_0$) = Prob function($T = t_{\text{obs}}$|H$_0$), and add to $\tilde{p}$ the tail Prob($T \leq t''$ |H$_0$) to obtain the two-sided p-value. This second approach will only yield different results for asymmetrical distributions. The two approaches in general give very similar p-values.

Example *6.6.1 Mean of normal distribution with known variance (continued)*

Suppose now we are testing H$_0$: $\mu = 5$ vs H$_1$: $\mu \neq 5$ and again we know that $\sigma^2 = 10$; $\bar{X}$ is our test statistic, and suppose that we observe $\bar{X} = 7.76$. Clearly results more favourable to H$_1$ and less favourable to H$_0$ will be in the region $\bar{X} > 7.76$.

Recall from Example 6.5.1 that

$$\text{Prob}(\bar{X} \geq 7.76 | \text{H}_0) = 0.003 = \tilde{p}.$$

We therefore report the two-sided $p = 2\tilde{p} = 0.006$.

Of course, in practice we do not go through this whole argument, we just report twice the one-side p-value. Nevertheless, it is useful to see this procedure is not ad-hoc but has a justification.

If you look at Inference 1.5 you will recall there is a link between confidence intervals and hypothesis tests: briefly, the test of H$_0$: $\mu = \mu_0$ vs H$_1$: $\mu \neq \mu_0$ rejects at the α level if and only if μ_0 lies outside the 100(1-α)% confidence interval for μ; this is further discussed in Analytical Techniques 3.

6.7 Type I error, type II error and statistical power revisited

Consider a simple test of the null hypothesis that the difference δ between the means of two groups in a Normally distributed population, with common variance, is zero; versus the alternative that it takes a value δ_1:

$$\text{H}_0: \delta = 0 \text{ vs } \text{H}_1: \delta = \delta_1$$

Consider the test statistic *D*, the random variable for the difference in sample means of the two groups, realised by *d*. Figure 6.2 shows the expected distributions of the test statistic under the null and alternative hypotheses.

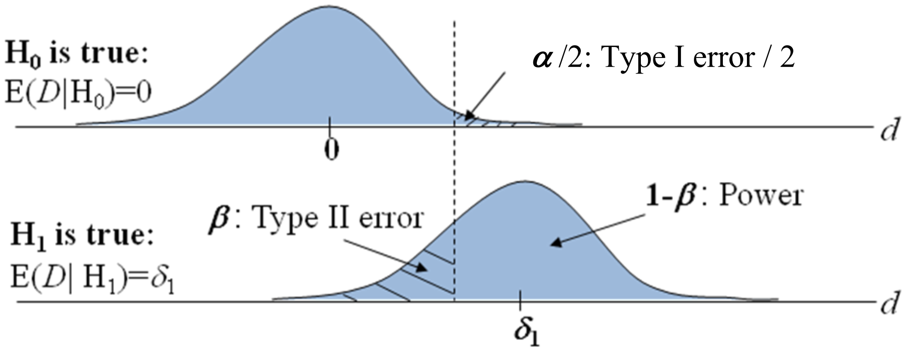

Figure 6.2

Figure 6.2 also shows the type I and II errors, along with the power of the test to detect a population difference of size δ_1. Note that having fixed the type I error at α, the power is then affected by either i) the size of the difference to be detected, which affects the separation between the expectations of the two distributions; or ii) the standard error of the test statistic, which affects the dispersion around the expectations. The standard error in turn is affected by the population standard deviation and the sample size.

Note: When do we use a one-sided p-value?
Rarely, and with care and special justification. One-sided p-values tend to be used in one of the following situations:

i) Where we know a priori that a result in one direction *has* to due to chance, there being no other *possible* scientific explanation for a result in that direction. We are rarely in a position to know that there can be no mechanism (other than chance) which could yield a result in the unexpected (or unwanted!) direction. There is a danger for researchers to perform a one-sided test in the direction they want to observe, to make a significant result more likely: this increases the chance of a false positive, a Type I error.

ii) Where the consequences of a Type I are not as bad as of a Type II error. For example, with safety data, where a significant result may classify a drug as unsafe, it may be preferable to classify a safe drug as unsafe rather than classifying an unsafe drug as safe: so a one-sided test may be preferred, increasing the sensitivity to detecting an unsafe drug, at the expense of false 'alarms'.

Note on interpretation of two-sided test:
There may be a temptation, when reporting and interpreting a two-sided test, to avoid giving a 'direction' of rejection. For example, when testing H_0: $\mu = 5$ vs H_1: $\mu \neq 5$, and observing a sample mean 7.76 with two-sided p = 0.006, there may be a temptation to report merely that there is evidence to reject the null hypothesis, without explicitly stating the direction of this evidence: without stating whether the evidence is for $\mu > 5$ or $\mu < 5$. **This temptation is mistaken and should be resisted**. The two-sided test allows the theoretical prior *possibility* of rejection in either direction, but in actuality the test can only be rejected in one direction! Observing a sample mean > 5 cannot in this context be evidence that $\mu < 5$! So to interpret fully and properly we would conclude that we can reject the null hypothesis at the 1% level **and that there is evidence, at this significance level, that $\mu > 5$**.

6.8 Construction of a test in a novel situation

We summarise here how the ideas we have discussed are involved in the construction of a hypothesis test in a novel situation. First we set up null and alternative hypotheses, which may be simple or composite. Then the following are the steps involved:

1. Define an appropriate test statistic (a function of the random variable(s) which the data is assumed to realise). You might use the Neyman-Pearson lemma to assist with this.
2. Obtain the sampling distribution for that test statistic under the null hypothesis. This is often difficult, but is sometimes easier for a test which is not the optimal one: this may justify not using the most powerful test statistic suggested by the Neyman-Pearson lemma.

   Note that when the test cannot be defined in terms of a parameter in a statistical model, the sampling distribution may have to be obtained from first principles by considering the probability distribution of a chosen statistic across the sample space of observed results. This is typical of a non-parametric test.
3. Define a rejection region which gives a pre-specified type I error probability, usually $\alpha = 0.05$.
4. Calculate the value of the test statistic for the observed sample of data.
5. If the observed value of the test statistic is in the rejection region, conclude that the data reject the null hypothesis in favour of the alternative hypothesis. In practice, because of the asymmetry in specifying null and composite alternative hypotheses, if the observed value of the statistic is not in the rejection region we conclude merely that the data fail to reject the null hypothesis (which is not the same as accepting the null hypothesis!).

6. We report the p-value quantifying the evidence against the null hypothesis; and very often this is all we are doing, as statisticians: assessing the weight of evidence provided by the data, rather than making a decision to reject or not reject a hypothesis.