SECTION 3 : Hypothesis Tests

The goodness of fit tests which we introduced in the previous section using the χ^2 statistic were an example of a **hypothesis test**. In this section we now consider hypothesis tests more generally.

3.1 : Simple Hypothesis Tests

A **simple hypothesis test** is one where we test a **null hypothesis**, denoted by H_1 (say), against an **alternative hypothesis**, denoted by H_2 – i.e. the test consists of only **two** competing hypotheses. We construct a **test statistic**, t, and based on the value of t observed for our real data we make one of the following two decisions:-

1. accept H_1, and reject H_2
2. accept H_2, and reject H_1

As an example of a simple hypothesis test, let X be a RV drawn from a normal pdf with variance equal to unity and mean value equal to μ, where it is known that either $\mu = 2$ or $\mu = -2$. Let our test statistic be simply $t = x$, the observed value of X in a random sample of size one. Let our null and alternative hypotheses be:-

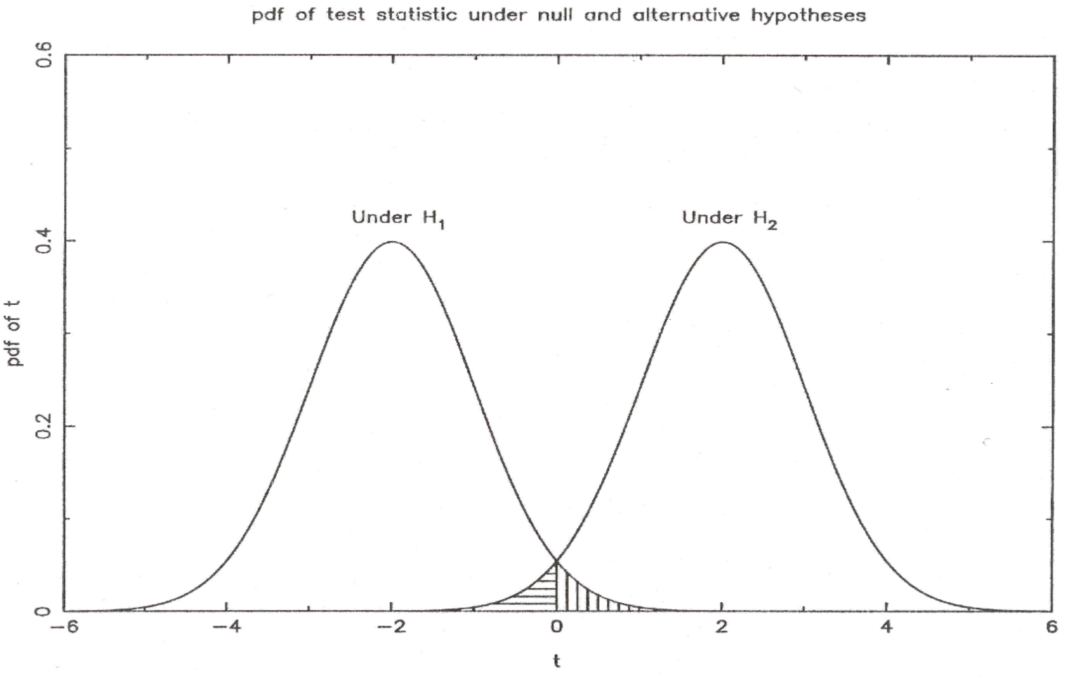
$$H_1 \; : \; \mu \; = \; -2 \qquad H_2 \; : \; \mu \; = \; 2$$

(Note that we could equal have chosen the null hypothesis to be $H_1 \; : \; \mu \; = \; 2$. The choice of which is the null and which is the alternative hypothesis – abbreviated as NH and AH – is basically up to the experimenter). Figure 15 shows the distribution of the test statistic, t, under the NH and AH specified above.

To carry out the hypothesis test we choose the **critical region** for the test statistic, t. This is the set of values of t for which we will choose to **reject** the null hypothesis and accept the alternative hypothesis. The region for which we accept the null hypothesis is known as the **acceptance region**. Note that we must choose the critical region and acceptance region ourselves. For example we might choose the critical region as the set of values of t for which $t > 0$. In other words, if our sampled value of x is found to be positive then we accept the alternative hypothesis that x was sampled from a normal pdf with mean $\mu = 2$, whereas if our sampled value of x is found to be negative, or equal to zero, then we accept the null hypothesis that x was sampled from a normal pdf with mean $\mu = -2$. In Figure 15, this particular choice of acceptance region and critical region is shown as the horizontally and vertically striped area respectively under the normal pdfs.

Note that our decision about whether the accept or reject the NH depends on the critical region, which has to be chosen by the observer. A different choice of critical region might lead to a different decision. This might seem to make the business of hypothesis testing a little subjective, but in some ways this subjectivity is inevitable. Statistical theory can never **absolutely** determine which of two competing hypotheses is correct – all it can do is tell us, provided certain assumptions are valid, how **probable** the two competing hypotheses are. Whether one (or indeed both!) of the hypotheses is then deemed to be *too* improbable to be accepted is – in the final analysis – up to the observer to decide. Very often this decision will depend on whether one is trying to prove one's own theory or model (i.e. by finding observational evidence to back it up), or disprove someone else's theory! We will return to this point shortly when we discuss significance.

Figure 15: PDF of test statistic under NH and AH for a simple hypothesis test.



While the choice of critical region may be subjective, once we have specified our choice of critical region we can objectively quantify what is the probability of making an **incorrect decision**.

3.2 : Incorrect Decisions

We can make an incorrect decision in one of two ways

3.2.1 : Type I error

A **type I error** occurs when we **reject** the null hypothesis when it is **TRUE** – i.e. when we should have accepted it. P(I) often denotes the probability of incurring a type I error.

3.2.2 : Type II error

A **type II error** occurs when we **accept** the null hypothesis when it is **FALSE** – i.e. when we should have rejected it. P(II) often denotes the probability of incurring a type II error.

We can calculate P(I) and P(II) in the simple example introduced above (see lectures), where the areas under the appropriate normal pdf can be found by consulting the tables provided.

A good hypothesis test should have small P(I) and P(II). Broadly speaking, this means that the distributions of the test statistic under H_1 and H_2 should have little overlap. We can always reduce P(I) by suitable choice of critical region, but this is inevitably at the cost of increasing P(II). It is often useful to choose the critical region which minimises some weighted combination of P(I) and P(II), but there is no general strategy suitable for all situations.

One frequently adopted criterion is the **power** of a hypothesis test, defined as the probability of rejecting H_1 when it is false, i.e. power = 1 - P(II). Choosing a critical region which maximises the power for a given alternative hypothesis is generally a useful strategy for defining a good hypothesis test.

3.3 : Level of Significance

The **level of significance** of a hypothesis test is the maximum probability of incurring a type I error which we are willing to risk when making our decision. In practice a level of significance of 5% or 1% is common. If a level of significance of 5% is adopted, for example, then we choose our critical region so that the probability of rejecting the null hypothesis when it is *true* is no more than 0.05

If the test statistic *is* found to lie in the critical region then we say that the null hypothesis is rejected at the 5% level, or equivalently that our rejection of the null hypothesis is **significant** at the 5% level. This means that, **if** the null hypothesis **is** true, and we were to repeat our experiment or observation a large number of times, then we would expect to obtain – by chance – a value of the test statistic which lies in the critical region (thus leading us to reject the NH) in no more than 5% of the repeated trials. In other words, we expect our rejection of the null hypothesis to be the *wrong* decision in no more than 5 times out of every 100 experiments. Yet another way to express this is to say that we are '95 % confident' that we have made the correct decision in rejecting the null hypothesis.

As mentioned above, the choice of significance level is somewhat subjective. Suppose, for example, that one is comparing the model prediction of another astronomer's favourite theory (here the NH) to the prediction of one's own pet theory (here the AH). In this case one might regard rejection at the 10% significance level to be sufficient grounds for ruling out the other astronomer's theory. Why? Because **if** the other astronomer's theory is true, there is at most a one in ten chance of the test statistic falling in the critical region (i.e. a one in ten chance of obtaining data similar to – as 'bad' as, if you like – the **actual** data which we do obtain). If, on the other hand, one were seeking support for one's own theory as the NH, then rejection at the 10% significance level might not be sufficient grounds to give up on one's theory, since one can argue that the **actual** data obtained happens to be one of those one in ten data sets which, by chance (or 'bad luck', if you like), yield a test statistic lying in the critical region – even when the NH is true.

How can we get around this? As remarked in section 3.1, we can always choose a more stringent critical region. For example, if we could reject the NH at, say, the 1% or 0.1% level, then we can be much more sure that the test statistic obtained for our real data does not lie in our critical region by chance, even though the NH is true. In other words, we reduce the probability of a type I error. But recall from section 3.1 that this will inevitably increase the chances of accepting the alternative hypothesis when it is false – i.e. making a type II error. Again, the key here is for the distribution of the test statistic to be so nearly disjoint under the null and alternative hypotheses (i.e. having so little overlap) that we can afford to adopt a 'tough' critical region without increasing P(II) too much. Clearly one effective way to reduce the overlap between the pdf of test statistics is to acquire more, and better, data, but in astronomy this is often a painful – and expensive – solution!

3.4 : Two Tailed Tests

It is common for the critical region to be defined as both the upper *and* lower tails of the distribution of the test statistic under H_1. For example, consider the random variable $X \sim N(\mu, 1)$ and the test statistic $t = x$. Consider the null and alternative hypotheses

$$H_1 \quad : \quad \mu = 0 \qquad\qquad H_2 \quad : \quad \mu \neq 0$$

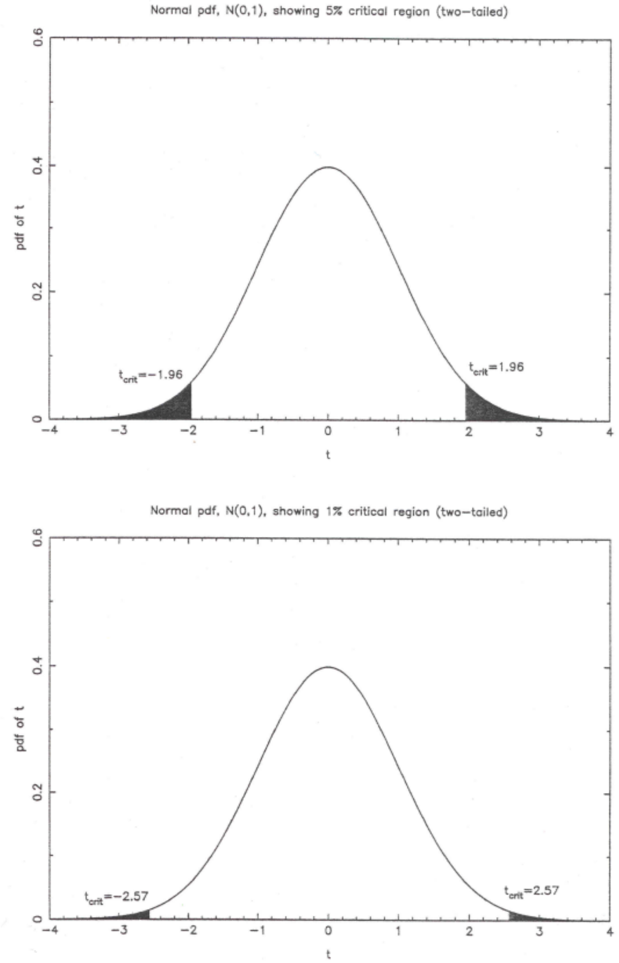
Then a value of t either much larger or smaller than zero might lead us to reject H_1 and accept H_2, since H_2 only states that the mean value is *different* from zero. In this example, adopting a 5% level of significance with a two tailed test would give as the critical region for t

$$\{t : |t| \geq 1.96\}$$

while for a 1% level of significance with a two tailed test, the critical region for t would be

$$\{t : |t| \geq 2.57\}$$

Figure 16: Two-tailed critical regions



In these examples $t = \pm 1.96$ and $t = \pm 2.57$ are the **critical values** for the test statistic; i.e. they indicate the boundary between the critical region and acceptance region. These two-tailed critical regions are shown in Figure 16.