Hypothesis Testing and Confidence Intervals I: One Sided Tests and Intervals.

Evaluating the weight of evidence supporting an idea and assessing the amount of confidence one can attribute to an idea laying in particular region are two intimately related problems that arise in the practice of applying statistics. Indeed they turn out to be the opposite sides of the same coin and we shall study the latter as the obverse of the former. Hypothesis testing is all about making choices under uncertainty, the general problem can be characterised as follows.

Organizing The Test.

Suppose the true state of the world can be one of two things denoted H_0 (the null hypothesis) and H_1 (the alternative hypothesis), however it is not known which state of the world prevails. Within the context of this uncertainty based upon some information (the data) a choice or decision has to be made, is the truth H_0 or H_1? The situation is best described in the following table:

Table 1.		**The True State of the World**	
		H_0	**H_1**
The Decision	**H_0**	✔	**Type II Error**
	H_1	**Type I Error**	✔

fundamentally two types of mistake can be made, concluding the world is H_1 when in fact it is H_0 (a Type I Error) and concluding that the world is H_0 when in fact it is H_1 (a Type II Error). The problem is how to organize the decision process and data in such a way as to minimize the impact of the errors.

In minimizing the impact of the errors the first important thing to note is that the errors need not be equally critical or important, the following example will suffice as an illustration. Suppose you manufacture brake pads for cars and your customers, the car manufacturers require pads that have at least 10000 lbs per square inch breaking power. To examine the breaking power of a set of pads you have to test them to destruction so you manufacture a batch, test a sub-sample and, based upon the test, deduce the breaking power properties of those remaining in the batch. If your test suggests the batch has adequate breaking power you will sell the rest to the manufacturers otherwise you will scrap them. Reproducing the above table for this problem yields:

Table 2. (One Sided Lower Tailed Test)		**The True State of the World**	
		H_0 : Braking power ≥ 10000 lbs p.s.i.	**H_1 : Braking power < 10000 lbs p.s.i.**
The Decision	**H_0 : Braking power ≥ 10000 lbs p.s.i.**	✔	**Type II Error**
	H_1 : Braking power < 10000 lbs p.s.i.	**Type I Error**	✔

Concluding H_1 when in fact H_0 is true (a **Type I Error**) results in scrapping a batch of acceptable brake pads. Concluding H_0 when in fact H_1 is true (a **Type II Error**) involves selling a bunch of defective brakes. Aside from the obvious moral issue, in a litigious world such as ours it is pretty obvious that a Type II error has far more serious consequences than a Type I error! The impact of an error is minimized by controlling the probability of it occurring at some preassigned low value chosen by the investigator. The usual practice in statistics is to control the probability of a type one error (**P(Type I Error)** is usually denoted **α** and called the **size of the test**) and take whatever Type II probabilities arise (**P(Type II Error)** is usually denoted by **β** and **$1-\beta$**, the probability of correctly rejecting a false null hypothesis, is called the **power of the test**). Consequently hypotheses are organised in such a way that the most serious error is a Type I Error and the probability it occurs is set at α. For example in the above mentioned problem the set up would be re-arranged in the following fashion:

Table 3. (One Sided Upper Tailed Test)		**The True State of the World**	
		H_0 : Braking power ≤ 10000 lbs p.s.i.	**H_1 : Braking power > 10000 lbs p.s.i.**
The Decision	**H_0 : Braking power ≤ 10000 lbs p.s.i.**	✔	**Type II Error**
	H_1 : Braking power > 10000 lbs p.s.i.	**Type I Error**	✔

Tests in the form of Table 2 are called Lower Tailed Tests (because the alternative is in the lower tail of the null distribution) and for a similar reason tests in the form of Table 3 are referred to as Upper Tailed Tests.

Constructing The Test.

Suppose the test has been structured as in Table 3, and we have to hand a sample of maximum breaking power values X_i, i = 1,.., n (the results from n independent experiments) the average value of which is $\bar{X}$. Employing ideas from the sampling theory chapter we work with the notion that the sample mean $\bar{X}$ is distributed N(μ, σ^2/n) and assume for simplicity that σ^2 is known (usually it is not known but we shall discover how to account for this later). If the null hypothesis were true, the highest possible value that the true mean breaking power could take on would be 10000 lbs p.s.i (in general denote this as μ_0). This is the boundary of the null - right next to the alternative without being in it - and in this case $\bar{X}$ is distributed N(10000, σ^2/n). We assume this to be the case. Intuitively large values of $\bar{X}$ provide evidence in favour of the alternative hypothesis whereas small values favour the null, we need to decide upon some criterion, C, (usually referred to as the **Critical Value**) for determining what constitutes large so that, if $\bar{X}$ > C we shall choose H_1: $\mu > \mu_0$ and if $\bar{X} \leq$ C we shall choose H_0 : $\mu \leq \mu_0$.

C is determined by noting three things:

1) That α has been chosen as the probability of a type 1 error.
2) That such an error would be committed if H_1 were chosen when H_0 was true.
3) H_1 will be chosen when $\bar{X}$ > C.

Bringing these together note that:

$$\alpha = P(Type\ I) = P(\bar{x} > C|_{H_0\ is\ true})$$

$$= P(\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} > \frac{C-\mu_0}{\sigma/\sqrt{n}}) = P(Z > \frac{C-\mu_0}{\sigma/\sqrt{n}})$$

where α, μ_0, σ and n are all known. The last equality comes from the fact that since $\bar{X}$ is distributed $N(\mu_0, \sigma^2/n)$, $(\bar{X} - \mu_0)/(\sigma/\sqrt{n})$ is distributed as Z, a standard normal (N(0,1)) random variable. It follows that the equality will hold when $(C - \mu_0)/(\sigma/\sqrt{n}) = Z_{1-\alpha}$ (the value of a standard normal random variable that yields an upper tail area of α. which knowing α can be found from the standard normal tables) and as a consequence:

$$\mathbf{C = \mu_0 + (\sigma/\sqrt{n})Z_{1-\alpha}.}$$

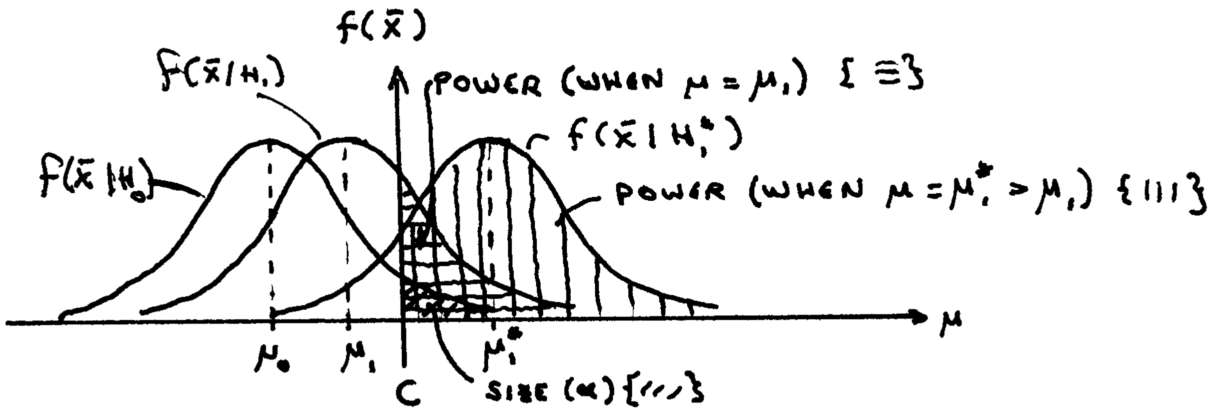
Having derived C we simply compare it with $\bar{X}$ and if $\bar{X} > C$ reject H_0, otherwise we accept H_0. Notice that C will be bigger than μ_0 since for α small $Z_{1-\alpha}$ will be positive as will $(\sigma/\sqrt{n})$ (thus in our example C will be bigger than 10000), furthermore it gets bigger as n gets smaller. This is because we are protecting ourselves against making a type one error, the average value in our sample has to be that much bigger than μ_0 to protect against pure sampling variability causing the rejection and sampling variability decreases as n increases. Notice also what would happen if the type one error probability α were set at 0, $Z_{1-\alpha} = \infty = C$ ensuring that the null would never be rejected and thus a type one error would never be made!

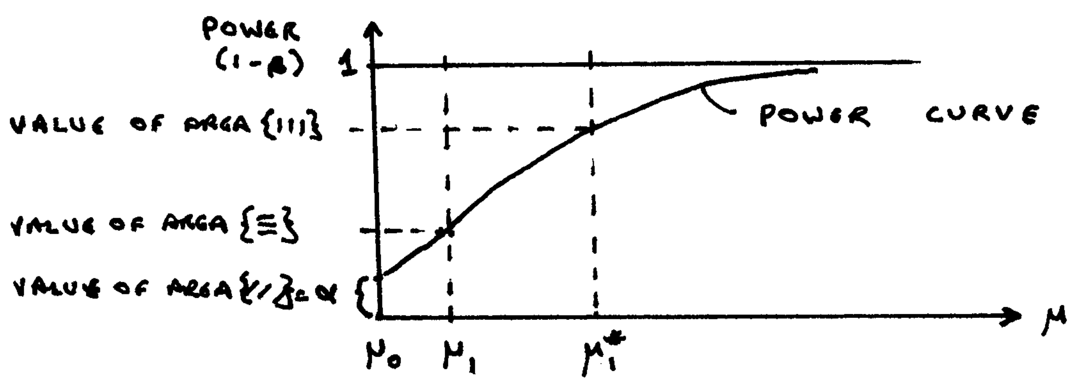
The Power of the Test.

The power of the test (1- P(Type II error)) has already been remarked upon, it is the probability of correctly rejecting the false null hypothesis and will depend upon how close to the null the true alternative hypothesis (H_1) is. Suppose that in our example H_1 is true and the true population mean is $\mu_1 > \mu_0$, then $\bar{X}$ will be distributed $N(\mu_1, \sigma^2/n)$ and the power of the test is given by the probability that the sample mean will exceed C under this alternative distribution which we write succinctly as $P(\bar{X} > C|\bar{X}\sim N(\mu_1, \sigma^2/n))$. Notice that:

1) The further away μ_1 is from μ_0, the closer to one will the power be.
2) The closer they are together the closer the power will be to α, the size of the test.

Indeed the plot of the power of the test as μ_1 moves away from μ_0 is known as the **Power Function** or **Power Curve**. These are best illustrated by the following diagrams.





One-Sided Confidence Intervals.

C, the critical value in a one sided upper tailed hypothesis test, is really an answer to the question

"How high does the mean of the sample have to be in order to decide on the falsity of the null hypothesis with a probability of being wrong no greater than α?."

One could equally well ask the question:

"Given a sample mean and given α, what an the estimate of the largest value μ^* for the unknown population mean μ that would yield a probability of $1-\alpha$ of the true mean being no bigger than that value?."

In this case we could write:

$$1-\alpha = P(\mu < \mu^*) = P\left(\frac{\mu-\bar{x}}{\sigma/\sqrt{n}} < \frac{\mu^*-\bar{x}}{\sigma/\sqrt{n}}\right)$$

$$= P\left(-Z < \frac{\mu^*-\bar{x}}{\sigma/\sqrt{n}}\right) = P\left(Z > \frac{\bar{x}-\mu^*}{\sigma/\sqrt{n}}\right)$$

This will true when $(\bar{x} - \mu^*)/(\sigma/\sqrt{n}) = Z_\alpha$ (the value of a standard normal random variable that yields an upper tail area of $1-\alpha$. which knowing α can be found from the standard normal tables) and as a consequence $\mu^* = \bar{x} - (\sigma/\sqrt{n})Z_\alpha$ which since $Z_{1-\alpha} = -Z_\alpha$ we may write as:

$$\mu^* = \bar{x} + (\sigma/\sqrt{n})Z_{1-\alpha}.$$

Notice the similarity between this upper bound of the $1-\alpha$ one sided confidence interval and the formula for the critical value C of an α size upper tailed hypothesis test, the former simply has $\bar{x}$ in the place of μ_0 in the latter. This is no accident, the hypothesis test bases the upper bound for the sample mean on the null hypothesized value of the population mean, the confidence interval bases the location of the upper bound for the population mean on the sample mean. They are simply opposite sides of the same coin.

Deciding upon a sample size.

Generally investigators accept whatever type two error probabilities arise in a particular situation but it is possible to specify an acceptable type two error probability for a given parameter value in the alternative and deduce thereby the sample size necessary to ensure both type one and type two errors. Consider a one sided upper tailed test $H_0\ \mu \le \mu_0$ versus $H_1\ \mu > \mu_0$ where the population variance σ^2 is known and where the size of the test is set to α and suppose that the investigator is prepared to accept a type II error probability of β for a specified value $\mu_1 > \mu_0$. The formula for C, the critical value for the test is given by:

$$C = \mu_0 + \frac{\sigma}{\sqrt{n}} Z_{1-\alpha}$$

as was derived earlier. Now, since the type II error probability (β) has been specified for a particular value in the alternative (μ_1), a second equation for the critical value can be determined. Since under the alternative $\bar{x}$ will be distributed $N(\mu_1, \sigma^2)$ and P(type II error) = P(accepting a false H_0) we have:

$$P(\textit{rejecting false null}) = P(\bar{X} < C|H_1) = P(\frac{\bar{X} - \mu_1}{\sigma/\sqrt{n}} < \frac{C - \mu_1}{\sigma/\sqrt{n}}) = \beta$$

which will be true when:

$$\frac{C - \mu_1}{\sigma/\sqrt{n}} = Z_\beta$$

or:

$$C = \mu_1 + \frac{\sigma}{\sqrt{n}} Z_\beta$$

Setting the two formulae for C to be equal and solving for n yields:

$$n = \frac{\sigma^2(Z_{1-\alpha} + Z_{1-\beta})^2}{(\mu_1 - \mu_2)^2}$$

Appendix. The χ^2, t and F distributions.

The standard normal distribution has already been introduced in Chapter 3, because of the Central Limit Theorem introduced in Chapter 4 it and three closely related distributions will be of major importance in the rest of the course. The distributions and their relationships will be briefly outlined here.

The χ^2 distribution

Given k independently distributed standard normal variates Z_i, i = 1, .., k, the random variable X defined by:

$$X = \sum_{i=1}^{k} Z_i^2$$

is distributed as a $\chi^2(k)$ random variable with k degrees of freedom. Noting that $E(Z_i^2) = 1$ it is readily seen that E(X) = k, it may also be shown that V(X) = 2k.Obviously X takes on only non negative values and is right skewed.

The t distribution

Given X, a $\chi 2(k)$ random variable and Z, a standard normal (N(0,1)) random variable, where the independently distributed standard normal variates Z_i, i = 1, .., k, defining the random variable X are also independent of Z the random variable Y defined by:

$$Y = \frac{Z}{\sqrt{\frac{X}{k}}}$$

is distributed as a t(k) random variable with k degrees of freedom. Y takes on values on the whole real line, E(X) = 0 and is symmetric around 0. As k becomes very large Y tends towards a standard normal random variable. For k small the t distribution looks very much like a standard normal distribution with a slightly lower peak and slightly fatter tails.

The F distribution

Given two χ^2 random variables X_1 (with k_1 degrees of freedom) and X_2 (with k_2 degrees of freedom) where the independently distributed standard normal variates z_i, $i = 1, .., k_1$, and the independently distributed standard normal variates z_j, $j = 1, .., k_2$, making up the respective variates are all mutually independent of one another, the random variable W defined by:

$$W = \frac{\frac{X_1}{k_1}}{\frac{X_2}{k_2}}$$

is distributed as an $F(k_1,k_2)$ distribution with k_1 and k_2 degrees of freedom.

Relationships.

$t(k) \rightarrow N(0,1)$ as $k \rightarrow \infty$

Y^2 is distributed as $F(1,k)$.

$1/W$ is distributed as $F(k_2,k_1)$.