Lecture 11

Confidence Intervals. Power.

The development of a methodology for testing whether the mean value is equal to μ_0 leads to providing a plausible range for the true value of the mean.

A so called *interval estimate* or *Confidence Interval*

Suppose I estimate the Modulus of Rupture (MOR) for planks of a certain origin to be 15N/mm^2. I have calculated $\overline{X}$ to be 15.

But was this the average MOR of 3 planks, 300 or 3mln? What is the variance (SD) ? Were the instruments used accurate to within 0.1, 1 or 10 or 100N?

It is reasonable to argue that without some statement of precision the estimate is useless.

If I take another sample am I likely to get a value close to 15 eg 14.5 or 16 or might I get 3 or 37?

A *confidence interval* gives a plausible range for the mean. The extent of the plausibility is defined by the confidence level 1-α.

It is stated in this way because it is precisely the set of values μ_0 for which we would not reject the hypothesis $\mu = \mu_0$ vs the not equal alternative at significance level α

If we test:

$H_0\, \mu = \mu_0$ vs $\mu \neq \mu_0$ and do not reject this at 5%.

There are other μ_0 for which we would not reject H_0, on the basis of this data. The set of all those μ_0 is the 95% confidence interval.

Formally the confidence interval covers (includes) the mean with probability 1-α.

The only point is that the interval is random not the value of the mean

Frequentists make a strong point about any parameter being an unknown constant not a random variable. Because another school called Bayesian's make a very strong point about it being a random variable!

So $P(L \leq \mu \leq U)$=0.95 is a valid statement, but it is U and L that are random.

Construction of a confidence interval for a mean

α =5% , assume known variance.

We do not reject H_0 if:

$$-1.96 \le T \le 1.96$$

$$-1.96 \le \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \le 1.96$$ where μ is the hypothesised mean.

$$-1.96\frac{\sigma}{\sqrt{n}} \le \bar{x} - \mu \le 1.96\frac{\sigma}{\sqrt{n}}$$

$$1.96\frac{\sigma}{\sqrt{n}} \ge \mu - \bar{x} \ge -1.96\frac{\sigma}{\sqrt{n}}$$

$$\bar{x} + 1.96\frac{\sigma}{\sqrt{n}} \ge \mu \ge \bar{x} - 1.96\frac{\sigma}{\sqrt{n}}$$

In practice, I would use $\bar{x} \pm 2\frac{\sigma}{\sqrt{n}}$.

For different α we have a number different than 1.96, for 90% confidence use 1.64. etc. If variance is unknown use the appropriate value from TINV(α,df).

Example

Suppose that the to sample of planks with $\overline{X} = 15$ we add the information that the sample size n = 49 and that the standard deviation was 8.4.

95% confidence interval then is:
With n = 49 it will not make much difference whether we use the normal or the T distribution.

The *standard error* = $s.d.(\overline{X})$ = $\dfrac{\sigma}{\sqrt{n}}$ = 1.2

The “quick” 95% confidence interval is :

$$\overline{X} \pm 2 * s.e. \;= 15 \pm 2.4$$

We are 95% confident that the mean is in the range :

12.6 to 17.4

Precisely: the interval (12.6,17.4) includes the mean with probability 0.95. (near enough 0.95)

For the “exact” normal confidence interval we use 1.96 rather than 2.

12.648 to 17.352 but I suggest the accuracy is spurious.

Finally for the proper T based interval we use:

TINV(0.05,49) =2.009574

The confidence interval not only tells us what we think the mean is but also gives a measure of the precision of the estimate.

<u>Comments</u>

Confidence intervals are probabilistically centred. If the parameter is outside the interval, the chances of it being either side of the interval are equal $(=\frac{\alpha}{2})$.

For the mean (based on a normal) they are also numerically centred – estimate in the middle.

Power of a test.

When testing a hypothesis we control for the probability of type I error by deciding on a significance level.

$$P(\text{Reject } H_0 \mid H_0 \text{ true}) = \alpha .$$

Again:
The reason we do this is so that we only have to compute the distribution of the test statistic only for the case H_0 true. This allows the alternative, H_1 , to be a general hypothesis such as $>$,$<$ or $\neq$ and makes the approach usable in practice.

We chose the test statistics in such a way so that probability of type II error is minimised.

$$P(\text{fail to reject } H_0 \mid H_0 \text{ is false}).$$

In the special (artificial) case where H_1 specifies a single value for the parameter we can compute P(Type II error).

The power of a test is:

$$\beta = 1 - P(\mathit{type}\ \text{II}\ \mathit{error}).$$

Lets go back to our wire example.

We had a sample of 25 observations and the standard deviation is known to be 150.

We observed $\bar{X} = 1312$ and were able to reject H_0 vs H_1 $\mu > 1250$ at a significance level of 2%. So all worked out OK. Had we failed to reject we would have learned nothing – the conclusion would have been insufficient evidence to reject H_0.

It is obviously desirable to avoid such situations. Can we get some assurance that we will reject H_0 when it is false?

Yes if we opt for a *simple* alternative. A simple hypothesis is one that specifies the values of the parameter(s). H_0 is always simple, H_1 usually isn't, it is a *compound* hypothesis. It consists of an infinity of simple hypotheses.

H_1: $\mu > 1250$ is equivalent to the set of hypotheses

$$H_1: \mu = \mu_1 \text{ for all } \mu_1 > 1250.$$

For each simple hypothesis we can compute the power.

For example:

n=25, $\sigma = 150$, $\alpha = 0.05$.

H$_0$: $\mu = 1250$ vs H$_1$: $\mu = 1300$

The power is the probability of rejecting H$_0$ when the mean is 1300.

We reject H$_0$ if the T statistic is big enough. (one tailed test because small T would not suggest mean = 1300)

With the significance level at 5%, the rule is reject H$_0$ if :

$$T > 1.64$$

$$\frac{\bar{X} - 1250}{s.d.(\bar{X})} > 1.64$$ the only variable here is $\bar{X}$.

If H$_1$ is true, the mean of X is 1300, the mean of $\bar{X}$ is 1300 and the standard deviation hasn't changed = $\frac{150}{\sqrt{25}} = 30$.

So

P(reject) $= P(\frac{\bar{X} - 1250}{30} > 1.64)$ where $\bar{X} \sim N(1300, 30^2)$

$$= P(\bar{X} > 1250 + 30 * 1.64)$$

$$= P\left(\frac{\bar{X} - 1300}{30} > \frac{1250 + 30 * 1.64 - 1300}{30}\right)$$

$$= P(N(0,1) > \frac{-50}{30} + 1.64) \qquad ***$$

$$= P(N(0,1) > -0.027) \ = 0.5106$$

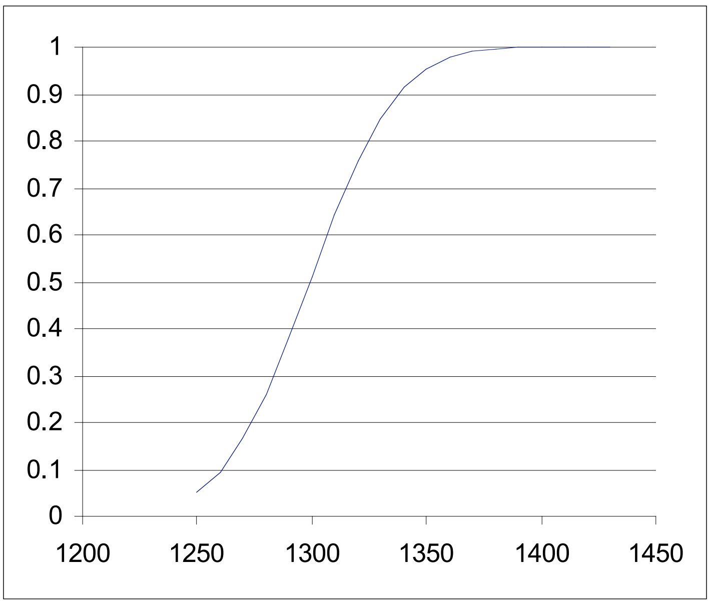
So if the true mean were 1300 we would need to be lucky (toss of a coin) to reject at 5% with 25 observations.

We can do this for other H_1 1260, 1275 , 1400 etc. the general form of the *** equation is:

$$P(N(0,1) > 1.64 - \frac{\Delta\mu}{sd(\bar{X})})$$

where: $\Delta\mu = \mu_{H_1} - \mu_{H_0}$

Power as a function of true mean



Designing the experiment.

Before we collect the data we want to ensure that we have reasonable power.

The power depends on:

$\Delta\mu$ - the difference in means the bigger this is the easier it is to reject.
We can chose this to be sensible. Not demand too much from the data.

α - The significance level. How sure we want to be that H_0 is false before we reject it?
The numbers are:

10%	$Z_\alpha = 1.28$
5%	$Z_\alpha = 1.64$
1%	$Z_\alpha = 2.33$

The bigger the Z_α, the lower the power.

$sd(\bar{X}) = \dfrac{sd(X)}{\sqrt{n}}$. Probably not much you can do about sd(X) but more accurate measuring might help.
What you can do is increase n =get more data.

What sample size (n) do you need to guarantee an acceptable level of power?

Suppose we pre-assign the power to be at least β.

The we have:

$$P(N(0,1) > Z_\alpha - \frac{\Delta\mu}{\sigma/\sqrt{n}}) \geq \beta$$

If we want, $P(N(0,1) > x) \geq \beta$ then we have to have

$x \leq -W_\beta$. where $W_\beta = NORMSINV(\beta)$

So:

$$Z_\alpha - \frac{\Delta\mu}{\sigma/\sqrt{n}} \leq -W_\beta$$

$$Z_\alpha + W_\beta \leq \frac{\Delta\mu}{\sigma}\sqrt{n}$$

$$(Z_\alpha + W_\beta)\frac{\sigma}{\Delta\mu} \leq \sqrt{n}$$

Note: $W_\beta = Z_{1-\beta}$

Example:

To be 90% sure that we will reject H_0 (significance =5%) if the mean is 10 higher than the hypothesised mean, sd(X)=150 we need:

Z_α=1.64, $W_\beta = 1.28$, $\Delta\mu = 10$, $\sigma = 150$

$\sqrt{n} \geq 31.488 \quad \Rightarrow n \geq 991.5$ i.e n=992 observations.

Don't demand to much of the poor data.

If we are considering alternatives less than H_0 everything reflects about 0. The formula is the same as long as $\Delta\mu$ measures the distance (positive) from the H_0 mean. Z_α is also treated as a positive value. (for this case reject H_0 if if T<-Z_α)

With a two sided test, ($\neq$), Z_α will be different because of the two tailed test (5% -> 1.96) . There is also the possibility of rejecting H_0 on the "wrong side" , but as the probability of this is tiny it can be ignored.

Comments.

We need to know σ. If we don't know σ the test should be a t-test, unless n>30 . Try as Z-test if n comes out less than 30, use n=30 or if the data are very expensive use the proper method. Requires non-central t distribution but is all done for you in MINITAB.

But you still need an estimate of σ. This is got from a pilot sample (small just to get σ). This can be used as part of the sample later.

However as estimates of σ are lousy unless you have a lot of observations you would usually allow some safety factor using the σ bigger than it really is will give you a larger n and hence a better β than you require. That is not as bad as concluding nothing.