BMI 541/699 Lecture 13

Where we are:

1. Introduction and Experimental Design
2. Exploratory Data Analysis
3. Probability
4. Distribution of the sample mean
5. Testing hypotheses about the sample mean(s): t-based methods
6. **Power and sample size for t-based methods**

Sample Size Calculations

A major contribution of biostatistics to biomedical research is in the design of studies that can achieve scientific goals (e.g., answer questions) with high reliability and efficiency.

Statistical design is more than just a good idea.

- It is essential in order to obtain unequivocal results that answer scientific questions.
- It minimizes costs.
- In research involving animals or humans, ethical considerations require detailed study design justification, including sample size and statistical power.

The **process** of sample size calculation can substantially improve study design. It requires one to think through:

- definition of the scientific issue
- how the scientific issue is being formulated as an empirical question
- sampling plan
- variables to be collected
- statistical analysis plan
- expected results

In general, if the details of implementation has been glossed over, this will become obvious during sample size calculation.

Recall that the p-value from a hypothesis test can be used to

1. decide whether to reject the null hypothesis (reject if p-value less than α)
2. summarize the evidence against the null

For the purposes of designing a study we use the first method. Typically $\alpha = 0.05$.

When we run the study we can also interpret the p-value as evidence against the null.

Definitions:

- A **type I error** occurs if the null hypothesis is true and *it is* rejected.
- A **type II error** occurs if the null hyp. is false and *it is not* rejected.

Table of the possible outcomes of a hypothesis test

	H_0 true	H_0 false
Do not Reject H_0	correct decision	**Type II error**
Reject H_0	**Type I error**	correct decision

What is probability of a type I error occurring?

Recall that the cutoff p-value α is the probability of rejecting the null hypothesis when the null hypothesis is true.

$$\alpha = \Pr(\text{reject } H_0 | H_0 \text{ is true})$$

also

$$\Pr(\text{Type I error}) = \Pr(\text{reject } H_0 | H_0 \text{ is true})$$

so

$$\alpha = \Pr(\text{Type I error})$$

Usually α is set to 0.05.

We Define:

$$\beta = \Pr(\text{Type II error}) = \Pr(\text{fail to reject } H_0 | H_0 \text{ is false})$$

When designing a study we must chose both α and β.

We want both to be small but typically we set α smaller than β

- A type II error (claiming no difference when there is a difference)

  is not considered as bad as

- a type I error (claiming that there is a difference when there isn't).

While α is usually set to 0.05, β is usually set to 0.20 or 0.10

From before:

	H_0 true	H_0 false
Do not Reject H_0	correct decision	**Type II error**
Reject H_0	**Type I error**	correct decision

The **power** of a test is the probability of the correct decision when the null hypothesis is false.

$$\text{Power } = \Pr(\text{reject } H_0 | H_0 \text{ is false})$$

That is, the power is the probability of finding an effect when an effect exists.

$$\text{Power } = \Pr(\text{reject } H_0 | H_0 \text{ is false})$$

$$= 1 - \Pr(\text{fail to reject } H_0 | H_0 \text{ is false}) = 1 - \beta$$

We want β to be small and the power to be large

In most computer programs you are asked to specify the power rather than specifying β.

The probabilities α and β refer to what *could happen* in the study

Once the study has been done and the data analyzed, these probabilities are *no longer relevant* - e.g., we calculate the actual p-value, and since we only see the sample, we don't know if a type I or type II error has occurred.

By designing the study with respect to these parameters, we minimize the probability of incorrect conclusions.

That is, for a given null and alternative hypothesis of interest, we design the study that has adequate statistical power to lead to a correct decision.

Power typically should be .8 or above.

Problems with Over- and Under-Powered Studies

Over powered: If the sample size is too large the study will be able to detect very small differences.

This is a waste of money and time if the difference is so small it is scientifically or clinically unimportant.

If the intervention is risky you have put too many individuals at risk.

Under powered: If the sample size is too small the study will be unable to detect differences that are scientifically or clinically important.

The risk taken by the individuals in the study was unnecessary because the study was unlikely to detect clinically important effects.

Also a waste of money and time.

Components of a sample size calculation: two sample t-test

We wish to test $H_0 : \mu_1 - \mu_2 = 0$ vs. $H_A : \mu_1 - \mu_2 \neq 0$.

The process of designing the study involves:

- Specify $\boldsymbol{\alpha}$: this is (usually) not difficult
- Specify the power $(1 - \boldsymbol{\beta})$
- Estimate/guess $\boldsymbol{\sigma}$ the population standard deviation
- Specify a *specific* alternative — Ideally the smallest difference $\boldsymbol{\delta} = \mu_1 - \mu_2$ that has scientific or clinical importance.

Given α, $(1 - \beta)$, σ, and δ we can calculate $\boldsymbol{n}_g$ the sample size in each group.

Approximate sample size formulas

These are approximate because they are based on the normal distribution rather than the t distribution.

Since σ and δ are not known exactly the additional error added by using the normal distribution is not important unless the sample sizes are very small.

Two sample t-test: $H_0 : \mu_1 = \mu_2$

$$n_g \doteq (z_{\alpha/2} + z_\beta)^2 \frac{\sigma_1^2 + \sigma_2^2}{(\mu_1 - \mu_2)^2}$$

where n_g = sample size per group.

Assuming equal variances, which we usually do in study planning, the formula is

$$n_g \doteq 2(z_{\alpha/2} + z_\beta)^2 \left(\frac{\sigma}{\mu_1 - \mu_2} \right)^2$$

For all sample size calculations, round the result up to the nearest integer. The total sample size is $n = 2n_g$

Typically we set

- $\alpha = 0.05$ so $Z_{\alpha/2} = 1.960$
- $\beta = 0.20$ so $Z_{\beta} = 0.8416$

which gives

$$\begin{aligned} n_g &\doteq 2(z_{\alpha/2} + z_{\beta})^2 \left(\frac{\sigma}{\mu_1 - \mu_2} \right)^2 \\ &\doteq 2 \times (1.960 + 0.8416)^2 \left(\frac{\sigma}{\mu_1 - \mu_2} \right)^2 \\ &= 2 \times 7.849 \left(\frac{\sigma}{\mu_1 - \mu_2} \right)^2 \end{aligned}$$

Rounding 7.849 up to 8 provides a quick formula for the number of observations per group for a 2 sample t-test

$$n_g \doteq 16 \left(\frac{\sigma}{\mu_1 - \mu_2} \right)^2$$

Example:

- Suppose we are designing a new study to compare the mean LDL cholesterol levels on two diets of oats: low oat consumption and high oat consumption
- Our plan is to collect a sample of $n_g + n_g$ subjects randomly assigned to either of the low oats and the high oats diets so that we have two groups, each with n_g subjects
- We think that the two groups will have will have different mean LDL
- Let

$$\mu_1 = \text{population mean LDL on high oats}$$

$$\mu_2 = \text{population mean LDL on low oats}$$

We will test:

$$H_0 : \mu_1 - \mu_2 = 0$$

versus

$$H_A : \mu_1 - \mu_2 \neq 0$$

- We set $\alpha = 0.05$ and $\beta = 0.2$ (80% power).
- From previous studies, our best guess for the standard deviation in the two groups is 1.0 ($\sigma = 1$).
- We consider differences $\mu_1 - \mu_2$ of 0.7 mmol/l or greater to be biologically important ($\delta = 0.7$).

From the equation for a two sample t-test:

$$n_g \doteq 2(z_{\alpha/2} + z_\beta)^2 \left(\frac{\sigma}{\mu_1 - \mu_2}\right)^2$$

$$= 2(1.960 + 0.8416)^2 \left(\frac{1}{0.7}\right)^2 = 32.036$$

We round up to a whole number, so for this experiment we need 33 subjects per group or 66 total.

One sample and paired t-tests

We need just one group and the sample size for that group is 1/2 the size of the sample size per group for the 2-sample t-test.

Null hypothesis: $H_0 : \mu = \mu_0$

$$n \doteq (z_{\alpha/2} + z_\beta)^2 \left(\frac{\sigma}{\mu - \mu_0} \right)^2$$

If $\alpha = 0.05$ and $\beta = 0.20$ we have $n \doteq 8 \left(\frac{\sigma}{\mu - \mu_0} \right)^2$

- One-sample t-test
  - $\mu =$ the population mean
  - $n =$ the number of observations
  - $\sigma =$ the population standard deviation
- Paired t-test
  - $\mu = \mu_d =$ the population mean difference.
  - $n = n_d =$ the number of pairs
  - $\sigma = \sigma_d =$ the population standard deviation of the paired differences.

What if we don't know σ

Typically we don't know the population standard deviation σ.

We can

1. Use an estimate from the literature. (May not be applicable to our situation.)
2. Run a pilot study to obtain enough data to estimate σ. (Expensive)
3. State $\mu_1 - \mu_2$ as a percentage of σ.

Recall the formula for sample size for the two sample t-test.

$$n_g \doteq 2(z_{\alpha/2} + z_\beta)^2 \left(\frac{\sigma}{\mu_1 - \mu_2}\right)^2$$

If we let

$$\delta' = \frac{\mu_1 - \mu_2}{\sigma}$$

Then δ' is the size of the difference between the means in standard deviation units and

$$n_g \doteq 2(z_{\alpha/2} + z_\beta)^2 \left(\frac{1}{\delta'}\right)^2$$

If we put $\sigma = 1$ into our power calculation then $\mu_1 - \mu_2$ can be entered in standard deviation units.

For example $\sigma = 1$ and $\mu_1 - \mu_2 = .7$ means we expect to see a difference between means that 70% the size of the standard deviation.

Sample size calculations using a computer

R Commander plugin EZR does do sample size calculations but I don't really trust it.

R does exact calculations using the t-distribution.
In R:

```
> power.t.test(sig.level=0.05, sd = 1, delta = .7, power = 0.8)

Two-sample t test power calculation

n = 33.02467
delta = 0.7
sd = 1
sig.level = 0.05
power = 0.8
alternative = two.sided

NOTE: n is number in *each* group
```

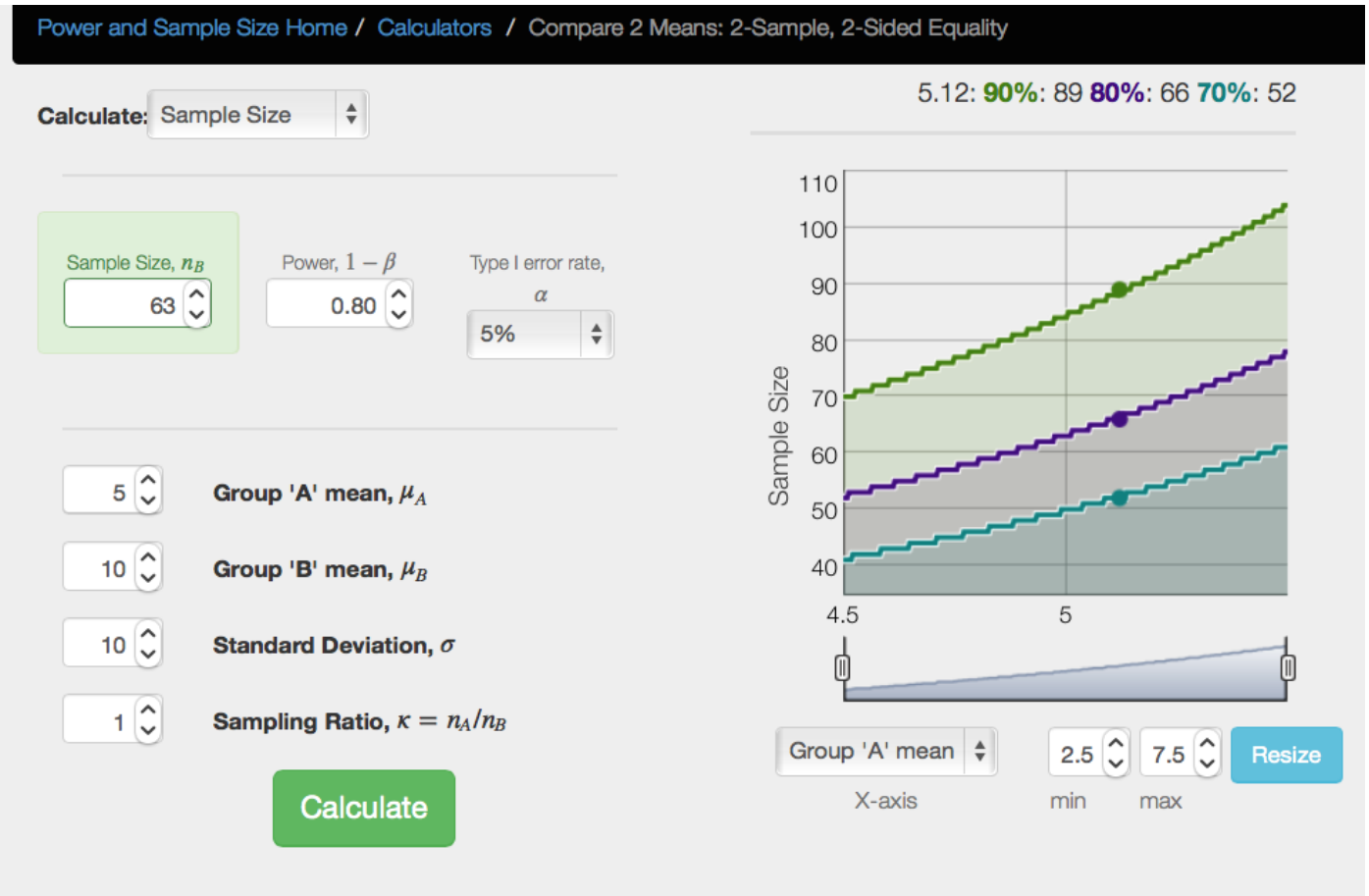
Rounding up we need 34 subjects in each group to obtain 80% power to detect a difference of 0.7

The web site:

http://powerandsamplesize.com/Calculators/

provides a nice interface to many sample size calculations.

On the left hand side under "Compare 2 Means" you see "2-sample, 2-sided Equality". Click on it.



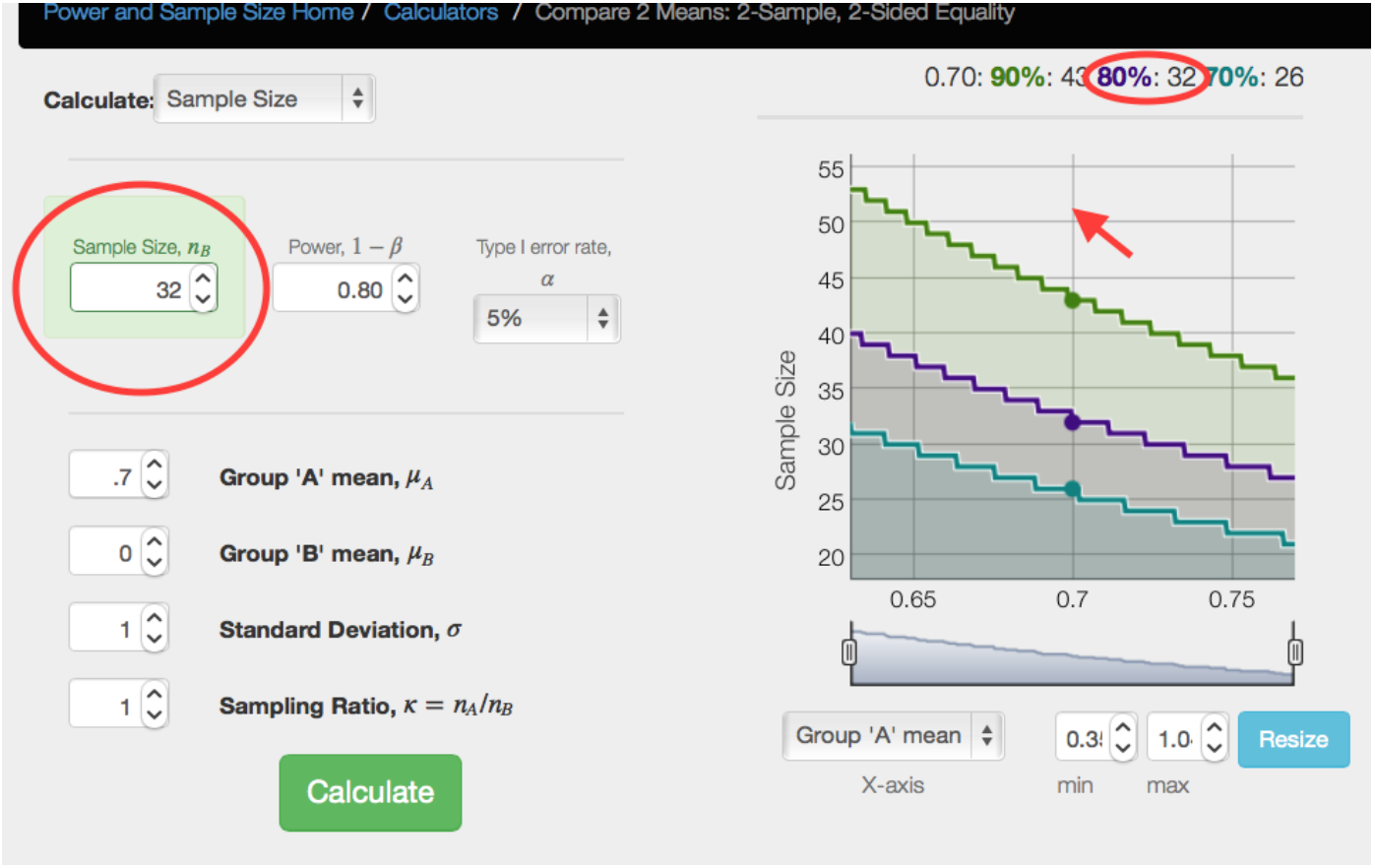
The sample size in group B (n_B) is highlighted in green and is what will be calculated (don't change that number)

The defaults are power $= 0.80$ and $\alpha = 0.05$. These can be changed.

Just below that you enter

- Group 'A' mean, μ_A
- Group 'B' mean, μ_B
- Standard Deviation, σ (the population standard deviation)
- Sampling Ratio, $\kappa = n_A/n_B$ (if this is 1 then the groups are the same size)

When you hit "Calculate" The sample size for group B appears in the green highlighted box.



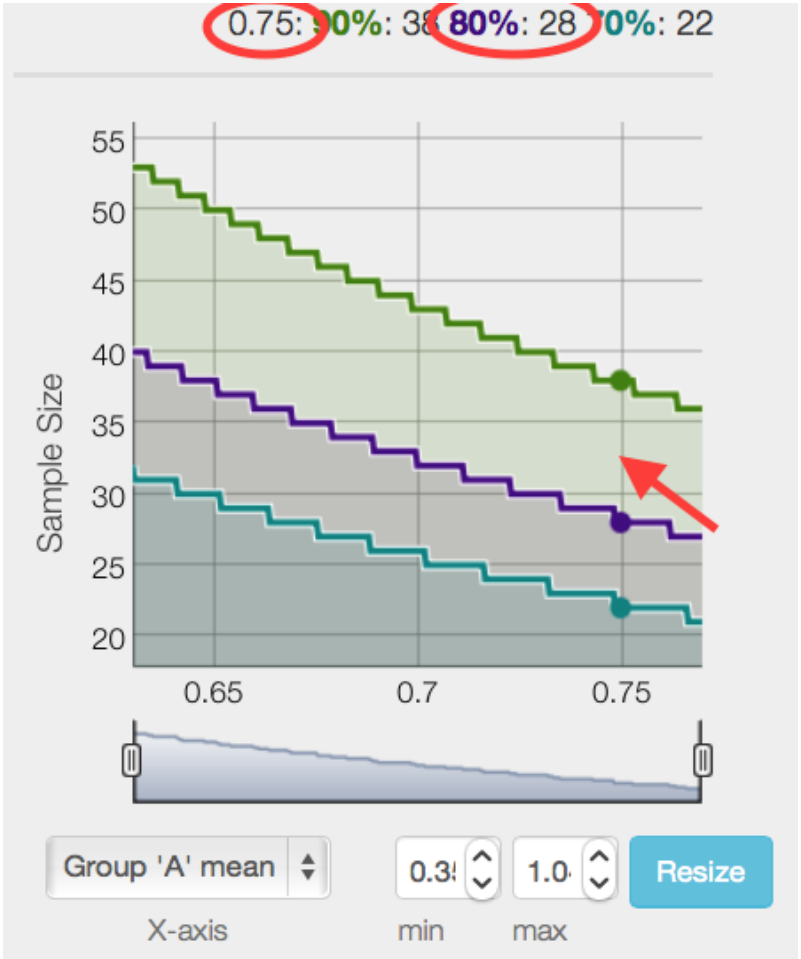
For our example we had $\sigma = 1$. We specified that the difference in the means is 0.7. We will set the mean of group B equal to 0 and the mean of group A $= .7$ so that the difference between the two is 0.7.

The calculated sample size is 32 per group. The actual value calculated with no rounding is 32.036. Not sure why they didn't round up.

If you move your cursor (red arrow) over the graph at a particular value for the mean of group A, the sample sizes needed obtain 90%, 80% and 70% power are shown at the top of the plot.

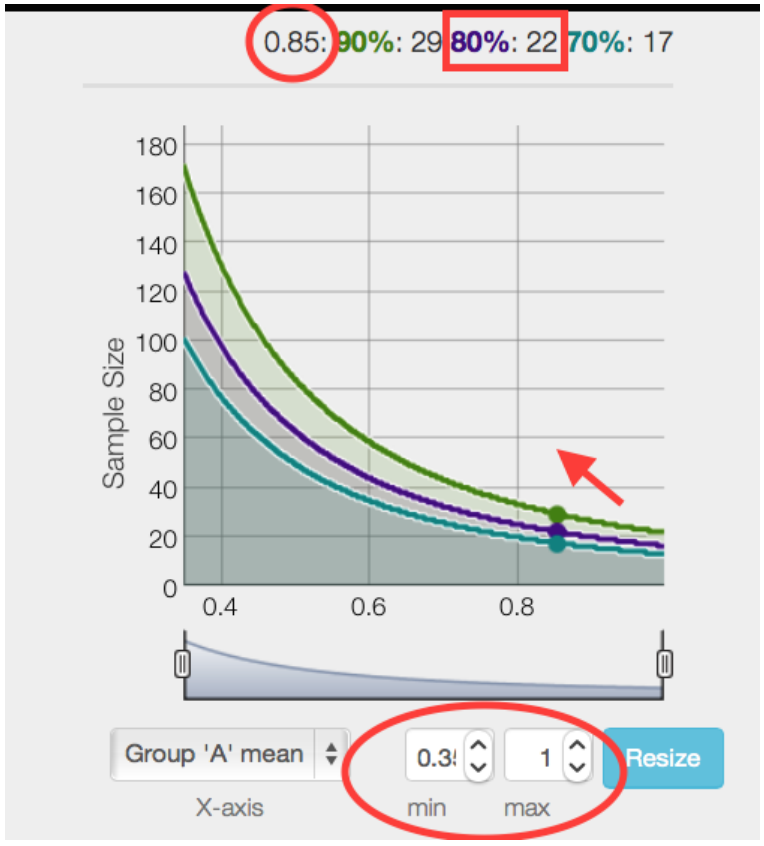
As the group A mean gets further away from the group B mean (delta gets larger) the sample size decreases.

As it gets closer to the group B mean (δ gets smaller) the sample size increases.



Say we decided that 32 per group was too many and we can only afford 22 per group.

Moving the cursor (red arrow) over the graph allows us to see that if we change to 0.75 for mean of group A we only need 28 per group to obtain 80% power.



The options below the graph allow us to change the range of the x-axis and change what is plotted. To see a larger range of means for group A set "max" equal to 1 and click the resize button.

Moving the cursor (red arrow) over the graph and watching the sample size listed next to 80% above the graph shows us that a difference in the means of 0.85 (red circle) requires 22 observations (red rectangle) for 80% power.

The other ways to decrease the sample size needed are to increase α, decrease σ, and decrease the power.

- Typically α is set at 0.05.
- We will talk about σ but it can't be adjusted at will.
- Power can be changed but it shouldn't be lower than 0.08.

Usually the only option to decrease the sample size is to change the alternative hypothesis so that the size of the difference that the experiment can detect is increased.

How design parameters affect sample size / power

Changing the parameter in the left column will cause the indicated changes to sample size or power.

If we Increase	**Direction of resulting change in** needed sample size	power
power	$\uparrow$	–
σ	$\uparrow$	$\downarrow$
sample size per group (n_g)	–	$\uparrow$
δ	$\downarrow$	$\uparrow$
α (sig.level)	$\downarrow$	$\uparrow$

Sample Size Determination for confidence intervals

Sometimes we do not plan to do a hypothesis test but we wish to estimating a mean

We can choose n so that the confidence interval for the mean is a certain length.

The normal based confidence interval is: $\bar{x} \pm z_{\alpha/2}\sigma/\sqrt{n}$

Define L to be the length of the confidence

$$L = 2\ z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

We can rearrange this equation to get:

$$\sqrt{n} \doteq \frac{2\ z_{\alpha/2}\sigma}{L}$$

and

$$n \doteq \left(\frac{2\sigma z_{\alpha/2}}{L}\right)^2$$

For a 95% CI $\alpha = 0.05$ and $z_{\alpha/2} = 1.96$ and we get:

$$n \doteq 15.37 \left(\frac{\sigma}{L}\right)^2$$

or as a rough approximation

$$n \doteq 16 \left(\frac{\sigma}{L}\right)^2$$

Example:

Suppose we wish to estimate the population mean and a 95% CI.

Our best guess at the population standard deviation σ is 6.

We would like to have a confidence interval of length 3.

We calculate that

$$n \doteq 15.37 \left(\frac{\sigma}{L}\right)^2 = 15.37 \times (6/3)^2 = 15.37 \times 4 = 61.48$$

Summary: Why do sample size calculations?

Prospective study design with sample size calculation helps to avoid studies that are:

- Too small: leads to equivocal results. An under powered study may dismiss a potentially beneficial treatment, or may fail to detect an important relationship.
- Too large: wastes resources.

Both sample size errors create ethical issues when using humans or animals.

- Too small: you have exposed them to harm with little likelihood of learning anything.
- Too big: you have exposed more of them to harm than was necessary.

Secondary benefit: Makes for better studies. Before you can do a sample size calculation, you will have to:

- Define the scientific issue you are addressing.
- Translate the issue into research questions or hypotheses.
- Determine what data are needed.
- Formulate the questions or hypotheses in terms of parameters describing the distribution of the data to be collected.
- Map out the statistical analysis plans