Sample Size Calculations

Analyses rely on means rather than individual values

Means are more precise

Precision measured by $\frac{\sigma}{\sqrt{n}}$

So precision depends on *n*

This can be used, directly or indirectly, as a basis for setting a sample size

Why set a sample size?

Should aim for a degree of precision appropriate to purposes of the study

Too much precision is wasteful and possibly unethical

Too little precision will yield a study that is uninformative (and so possibly unethical)

Methods usually based on either:

Specifying width of confidence interval

or

Power of a hypothesis test

Principal Difficulties

Methods usually use formulae, tables or a computer program

All require values of unknown parameters

Study is a method of estimating these parameters

So do sample size calculations put the cart before the horse?

To an extent, yes

Must use imprecise estimates, so

- results must not be taken too ‘numerically’
- can rule out silly studies

Method based on confidence intervals

Confidence interval is: $\bar{x} \pm ts / \sqrt{n}$

Use *z* in place of *t*, OK for *n*>30

Width of confidence interval is $2zs / \sqrt{n} \approx 4s / \sqrt{n}$

Limit on confidence interval is equivalent to limit on SE

SE is less than *L* if $n > s^2 / L^2$ - use literature to get estimate of *s* and/or σ.

For two-groups, SE=$s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$, leading to $n \geq \frac{2s^2}{L^2}$

Method based on hypothesis tests

Principles but not details apply whatever the type of outcome

Start with Normal outcome, two groups

Slightly different emphasis to hypothesis test:

Type I error if a true null hypothesis is rejected

Type II error if a false null hypothesis is not rejected.

Constraining error rates

Test null hypothesis by referring $z = \frac{\bar{x}_1 - \bar{x}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ to SNV

If you reject null hypothesis if z>1.96 or z<-1.96
then Type I error rate is 5% (independent of sample size)

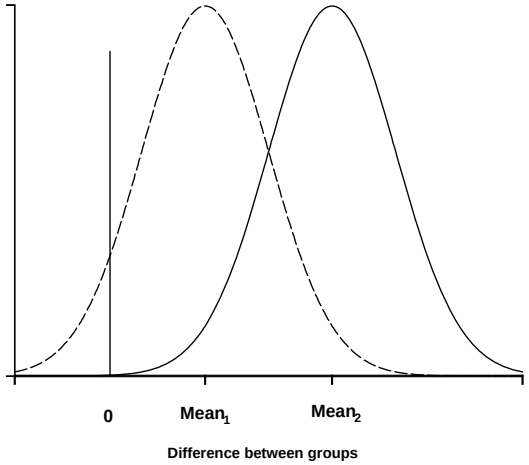
Type II error rate depends on sample size, so Type II error rate is the thing we seek to control through sample size.

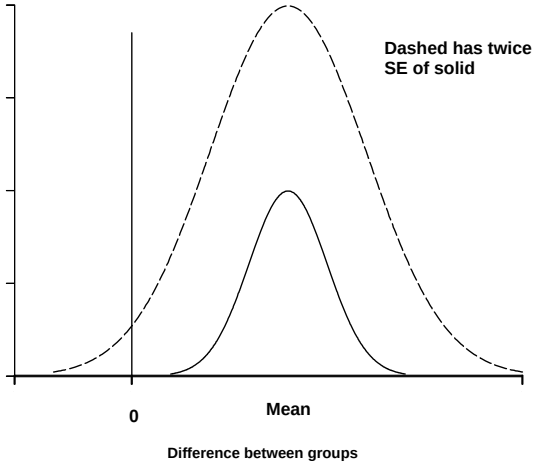
Note POWER = 1 -Type II error rate
= Probability reject a false null hypothesis

Heuristic explanation of process can be given

Basic Ideas

Two cases: different means, same se and same means different ses





Will be able to tell $\mu_1 - \mu_2 \neq 0$ more easily is $\mu_1 - \mu_2$ larger

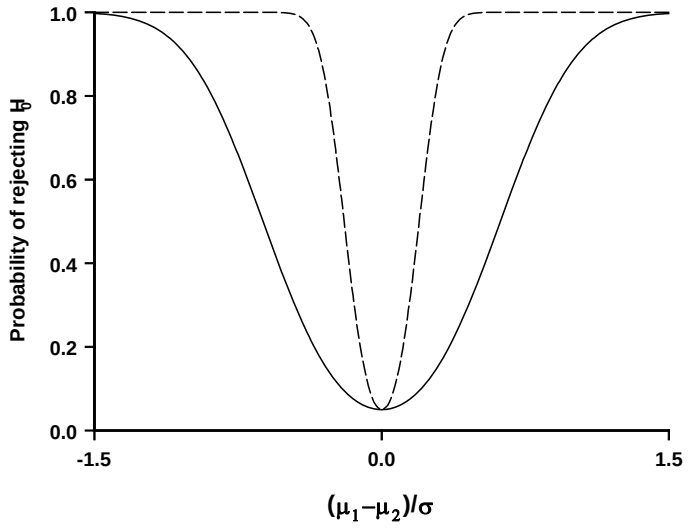
OR

if se is smaller, i.e. sample size larger

Power as a function

Null hypothesis specifies $\mu_1 - \mu_2 (= 0)$ whereas the null hypothesis can be false in an infinite number of ways

It follows that power is not a single number, but depends on $\mu_1 - \mu_2$



All curves must go through (0,0.05) [or (0,α)] and tend to 1 as $|\mu_1 - \mu_2|$ increases.

Sample size formula for Normal outcomes

Two groups with common SD, testing they have a common mean.

Number in each group is *n*. Power to detect difference $|\mu_1 - \mu_2|$ is β at two-sided level α if:

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

z_β cuts off top proportion β from a SNV etc.

Common values

Formula can be written

$$n = A \times \frac{\sigma^2}{(\mu_1 - \mu_2)^2} = A \times \left(\frac{\sigma}{(\mu_1 - \mu_2)} \right)^2$$

Power	80%	90%	95%
A	15.7	21.0	26.0

(Type I error 0.05)

σ needs to be estimated from literature/pilot study (latter can be very variable)

Specifying $\mu_1 - \mu_2$

Power to detect some values of $\mu_1 - \mu_2$ will always be small

Huge studies needed to detect small $\mu_1 - \mu_2$

But small values of $\mu_1 - \mu_2$ will not be of interest

Need to specify $\mu_1 - \mu_2$ as smallest difference that is of clinical interest.

Essentially set $\mu_1 - \mu_2$ as limit so that investigators are prepared to accept that their study may miss differences less than $\mu_1 - \mu_2$

Example

Placebo controlled trial of zanamivir, a new treatment for influenza (MIST study group, Lancet 352, 1877-81)

Outcome is number of days to alleviation of symptoms. Common SD is taken as 2.75 days

Decided that a mean improvement of 1 day is minimum that is clinically valuable

For 90% power at 5% level A=21.0, so number required in each group is

$$21.0 \times \frac{2.75^2}{1^2} = 159$$: for σ=2, 3.5 we get 84 or 258

Binary outcome

α, β and $\pi_1 - \pi_2$ specified as before

No analogue of σ: remember SE(*r*/*n*) = $\sqrt{\frac{\pi(1-\pi)}{n}}$, i.e. information about σ comes through π itself.

(sample size for change 0.1 to 0.3 different from 0.4 to 0.6)

Actually need to specify one of π_1, π_2 as well as $\pi_1 - \pi_2$

Table available, in terms of $\pi_{smaller}$ and $\pi_{larger} - \pi_{smaller}$

Note change from 0.6 to 0.8 is same problem as change from 0.2 to 0.4

Concluding remarks

Sample size calculation is a bit of a black art: depends on unknown parameters that need to be estimated by whatever means you can - still leaves you with very variable estimates

Is a sample size estimate that is plausibly between 400 and 800 any use? Yes, if all you can manage is 100.

Can use formula to get power from a given *n* - need to be cautious

Anticipate dropouts, but remember that loss of power may not be your main concern about dropouts

Post hoc power calculations inappropriate - use confidence intervals.