Lecture 7

Power and sample size in *t*-tests and proportion tests

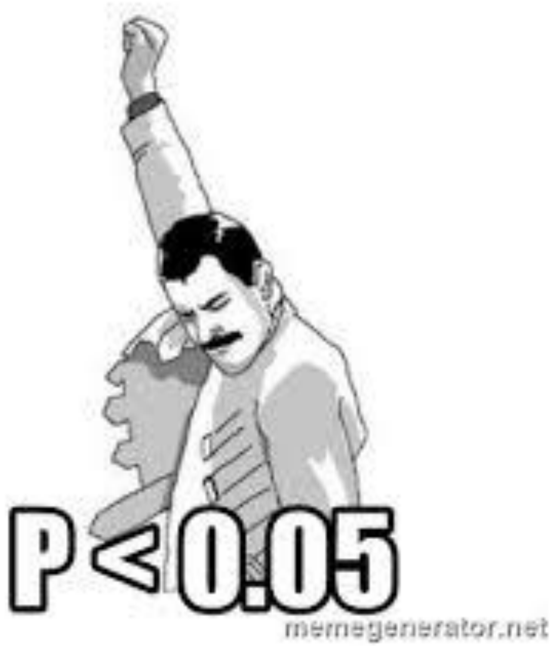
Detecting true differences

- Suppose that two groups *truly* differ in a variable mean; or that some proportion significantly differs between two populations
  - elephants and mice truly differ in mean weight
  - proportion of vegetarians truly differs between Indian and Argentina

- Question is: how large should my sample size be if…
  - the two species differ by 1 kg? Or 5,000 kg?
  - the proportion of vegetarians differs by at least 5%? Or 50%
- Calculating right sample size avoids two problems:
  - collecting less data than needed to test a hypothesis
  - collecting more data than needed!

Types of error: Type I

- Type I error: when you incorrectly reject a ‘true’ null hypothesis
  - =test says that groups are different (=$P<0.05$), but in fact they are similar
  - difference is just due to sampling from a single distribution
  - probability of type I error is the significance level
    - = probability of obtaining a P-value below the significance level purely by chance due to sampling
- So whenever we reject a null hypothesis and accept the alternative due to $P < 0.05$, we are accepting a risk of 5% of being wrong



*****
Type II error

- Type II error is the opposite: it is when you incorrectly accept a wrong null hypothesis
  - test says that groups are similar (P > 0.05), but if fact they are different
  - probability of type II error is the probability of randomly obtaining a *P*-value above the chosen significance level (P > 0.05)

- Type II error occurs when your test is not *powerful* enough to identify a true difference, i.e. to reject the null hypothesis (of no difference)
  - test is 'myopic', or does not have enough resolution to detect that level of difference (or *effect size*) between the groups
    - The effect size: the true effect of species on weight (the effect of being an elephant vs. a mouse), or the true effect of country on the probability of being a vegetarian (the effect of being Indian or Argentinian)

Statistical power = test resolution

- = power to identify a true difference
- = power to obtain P < 0.05
- = power to reject a wrong null hypothesis
- = power to avoid Type II error

- So what makes a test 'short-sighted'?
  - Small sample size!
    - makes it more difficult to obtain a P <0.05
  - Small effect size!
    - more difficult to detect true difference of 1 than a true difference of 10
  - Large standard deviations
    - larger overlap between groups reduces

$$t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{N}}}$$

- Calculations of statistical power must take all those factors into account

Statistical power

The larger the power, the lower the type II error

- Statistical power = 1 - (probability of type II error)
  - If power to detect true difference is β=0.9, the chance of a type II error (and not detecting it) is 0.1=10%

- Parameters determining power:
  - n = sample size
  - α = significance level (=0.05 by default; modified with "sig.level=0.01")
  - σ = standard deviation
  - δ = delta or the real difference between sample means that

- You should design tests with power of at least 80%; ideally, 90%
  - Power of 80%: you have a chance of 4 in 5 of detecting a true difference between groups

Statistical power detects the significance level (how confident it could be), but the significance level should be or above 95%

- Notice that confidence levels and statistical power are different things

Noncentral *t*-distribution

- It is possible to adapt the *t*-distribution to calculate probability of type II error through a noncentral parameter *v* (noo)

- *v* is similar to *t*-statistic and varies according to the specific test (one-sample *t*-test, two-sample *t*-test, paired *t*-test, two-proportions test)
  - = it is the test statistic to estimate the probability of an effect size under a given sample size, confidence level (P value) and standard variation

Power of one-sample *t*-test

- Question is: what is the probability of identifying a true difference of δ between a group mean and a test value?
  - i.e. what is the power of the test?

- In one-sample t-tests, noncentral parameter ν is

$$\nu = \frac{\delta}{\frac{\sigma}{\sqrt{n-1}}}$$

- Noncentral parameter ν is the difference divided by sem (standard error of mean)

Power of one-sample *t*-test

- In R, we use function *power.t.test*

```
>power.t.test(delta, sd, n, power)
```

- If you enter any 3 parameters (in any order), the 4th is calculated

Example:

- We have height for 20 Agta women from the Philippines
- What is the power of a one-sample test to demonstrate that their height truly differs by (at least) 5 cm from the mean height of neighbouring farmers? Assume sd of height is 7.

Parameters:

- $n=20$
- $\delta=5$ cm
- $\sigma=7$

Power of one-sample *t*-test

- For one-sample t-test, add ***type="one.sample"***

```
>power.t.test(n=20, delta=5, sd=7, type="one.sample")

One-sample t test power calculation
   n = 20
   delta = 5
   sd = 7
   sig.level = 0.05
   power = 0.8575538
   alternative = two.sided
```

- Conclusion: for a true difference of 5cm, a sample of 20 Agta women would provide a t-test with power β=0.86 (which is good enough)

(for a one-tailed test, add ***alt="one.sided"***)

Calculating sample size

- If we want a power of 90%, what sample size do we need?

```
> power.t.test(delta=5, sd=7, power=0.9, type="one.sample")

     One-sample t test power calculation

              n = 22.60315
          delta = 5
             sd = 7
      sig.level = 0.05
          power = 0.9
    alternative = two.sided
```

It defines the lowest level

- So what's the sample size needed? 22? 23?
  - If the estimate is 22.6 (minimum), then you need n= 23

- Calculating power shows that the answer to the question 'what is a large sample' depends on what we want to detect

Two-sample *t*-tests

- In the case of a two-sample test, noncentral parameter is

$$\nu = \frac{\frac{\delta}{\sigma}}{\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

- where n_1 and n_2 are sizes of the two samples
  - test assumes n1= n2 (to maximise power)
    - =the minimum sample size for each group
  - variance also assumed similar in the two groups, i.e. we enter only one standard deviation

Sample size, two-sample *t*-test

- We want to test for a true difference between two groups; for a power of 80%, a true difference of 2, and sd= 5, what is the required sample size?
  - (two-sample *t*-test is default; no need to specify type now

Two sample test is default in all tests, no need to point out like type="one.sample" or "paired"

```
> power.t.test(power=0.8, delta=2, sd=5)
        Two-sample t test power calculation
     n = 99.08057
     delta = 2
     sd = 5
     sig.level = 0.05
     power = 0.8
     alternative = two.sided
   NOTE: n is number in *each* group
```

- Sample size = 100
  - = minimum size **in each group**; so total is **200**

Sample size, two-sample t-test

- Now what if the true difference is 7?

```
> power.t.test(power=0.8, delta=7, sd=5)
   Two-sample t test power calculation
    n = 9.07768
    delta = 7
    sd = 5
    sig.level = 0.05
    power = 0.8
    alternative = two.sided
 NOTE: n is number in *each* group
```

- Now you need a much smaller sample (10 per group)
  - a larger difference is easier to spot; required ‘resolution’ to identify a true difference of 7 is much lower

Paired t-tests

- For a paired t-test, add *type="paired"*

```
> power.t.test(power = 0.8, delta=2, sd=5, type="paired")
     Paired t test power calculation
              n       = 51.00957
              delta   = 2
              sd      = 5
          sig.level   = 0.05
              power   = 0.8
        alternative   = two.sided

NOTE: n is number of *pairs*, sd is std.dev. of *differences* within pairs
```

- Sample size
  - ~ half the size in two-sample t-test
  - Same size as a one-sample t-test

Minimal relevant difference vs. statistical significance

- What if you have *no idea* about the expected difference between two groups?
- In this case, select the minimum value that you may consider relevant to report!
  - = 'minimal relevant difference' or 'smallest meaningful difference'

- For example, identifying a significant difference ($P < 0.05$) of *1 second* in life expectancy between people who eat some vegetable and people who don't will not get you a medical award!
- So if you want to test whether eating some vegetable or anything else affects lifespan, design the test around a relevant 'effect size'
  - you may decide that a relevant effect on life expectancy should be at least one year
- In summary: a difference may be statistically significant, and yet irrelevant!

Power of two-proportion tests

- To calculate power and sample sizes in two-proportion tests, we use function *power.prop.test*
  - based on a binomial approximation to a normal distribution

- limitations:
  - only works for independent proportions – *this is not a power test for one-sample proportion tests*
  - it cannot be used when sample size is smaller than 5 (a limitation of model distribution)

- only two parameters needed:
  - δ is replaced by p_1=proportion 1 and and p_2=proportion 2.
  - standard deviation not required

Power of two-proportion tests

- Example: What sample size do we need to detect a difference of 15% in preference for hybrid cars between Swedish and American people (let's say between 25% and 10%)?

```
> power.prop.test(power=0.9, p1=0.1, p2=0.25)
 Two-sample comparison of proportions power calculation
   n = 132.7557
   p1 = 0.1
   p2 = 0.25
   sig.level = 0.05
   power = 0.9
   alternative = two.sided

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

- One-tailed option is available

Other tests

- Package *pwr* has a series of functions to estimate power, sample sizes etc. of
  - t-tests
  - Proportion (chi-square) tests, one and two independent proportions
  - ANOVAs
  - Correlations

Summary

- Your test should have a power of at least 80%; if possible, try 90%
- When designing experiments, collecting or analysing data, run power and sample size tests first
- When you design a test, aim at a difference between groups that is worth reporting or considering a ‘result’
- A ‘result’ is not defined exclusively by statistical significance; relevance of finding (or effect size) is as important, and this can be determined by statistical power

Exercises

- Suppose that the average baby girl is born weighing 3300 g.
  - Which sample size do you need to show (with a probability of 90%) that newborn size in boys is at least 5% different from that? (assume sd=200 g)

Group means, t test

Real difference: 3300*0.05

power.t.test(delta=165,sd=200,power=0.9,type="one.sample")

- Proportion tests failed to identified a difference in proportion of boys among all births in rural gypsies and Hungarians
  - Calculate the sample size required if a two-independent proportions test is to have an 80% chance of detecting the observed difference in proportions of boys in gypsies and non-gypsies

power.prop.test (power=0.8, p1=0.47 p2=0.53)

1089

Table 2. *Sex ratios at birth for each population*

	number of sons per 100 daughters			
	rural populations		urban populations	
	Gypsy	Hungarian	Gypsy	Hungarian
A. all children				
sample size	254	216	239	224
males/100 females	89.3	111.8	89.7	113.3
B. first-born children only				
sample size	87	85	77	102
males/100 females	81.3	157.6	94.3	131.8