Lecture 4: Multiple comparisons, power, and limitations of NHST

Bernhard Angele

28 February 2018

Type I and Type II error

- We already mentioned this in the previous lecture

	H_0 is true	H_1 is true
H_0 not rejected	correct	β, Type II error
H_0 rejected	α, Type I error	Power

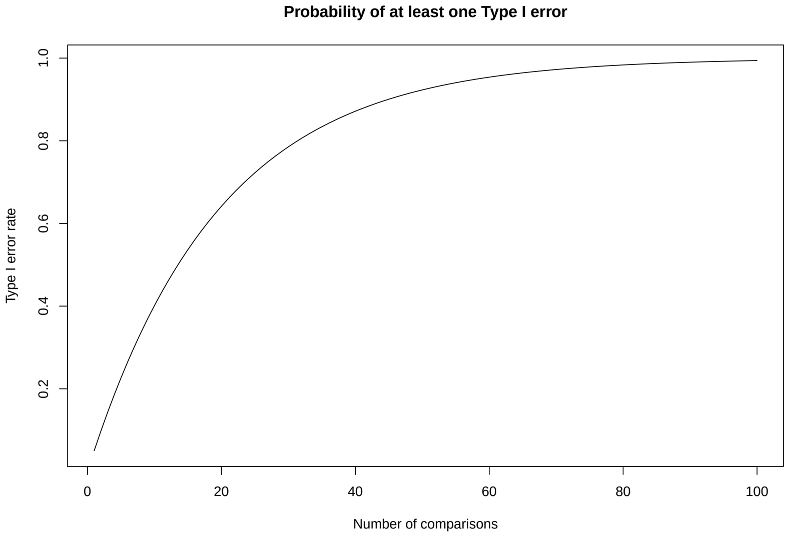
Controlling the Type I error rate

- Seems trivial: just set α to an appropriate level (0.05)
- However, there are many factors that can increase the Type I error rate:
- Multiple comparisons
  - See Week 3 In-Class activity Question 1
  - Each test has an individual α of .05.
  - But if you do a lot of tests, these individual alphas add up

Type I error rate for multiple comparisons

This is not particularly complicated, it's just the function

$$y = 1 - .95^x$$



The family-wise Type I error rate

- A family of hypothesis tests is a number of tests that would each, on their own, lead you to reject the null hypothesis if they came out significant
- Example: You are testing an intervention on depression. You want to see if the treatment group differs from the control group in terms of mood, activity, and sleep quality and conduct one t-test for each of these measures
    - You will reject the null hypothesis that the intervention has no effect if any one of the t-tests shows a significant effect.
    - What is the family-wise Type I error rate for the three tests (assuming they are two-tailed with a 5% α?
        - It's $1 - .95^3 = 0.142625$

Controlling the family-wise Type I error rate

- The simplest approach is applying a Bonferroni correction
- Simply divide the α by the number of tests: $\alpha_{corr} = \frac{\alpha}{n_{Tests}}$
- In our example, that would mean using an α of $\frac{.05}{3} = 0.0166667$
- Let's see what this does to our Type I error rate: $1 - 0.9833333^3 = 0.0491713$
- The correction is effective! Now our Type I error rate is below .05, just like we wanted.
- Of course, lowering the α increases the β, i.e. the Type II error rate, and lowers power.
- Other corrections such as Bonferroni-Holm are slightly more sophisticated and can help you maximise power for the comparison you care most about

Optional stopping rules

- As we've seen last week, multiple comparisons are not the only way to increase the Type I error rate
- A similar problem applies if we "peek" at the data after a number of participants, run a hypothesis test, and only continue collecting data if the test is not significant
- Given enough time and patience, you can get *any* test significant with such a strategy.
- Firm stopping rules (e.g. stop after 40 participant data sets no matter what happens) don't have this issue.
- A lot of researchers can't resist the temptation, especially if the test narrowly missed significance
- "$p = .07$? Surely if I just collect five more participants, this will get significant!"
  - "Still $p = .06$? But we're moving in the right direction, I'll just get 5 more participants"
- You must correct for each test that you do on the data, even if you were "just peeking to make sure everything is OK with the data".

The Type II error rate (beta)

- The probability that you fail to reject the null hypothesis even though it is actually false
- There is an effect, but your experiment failed to detect it.
- The opposite of β is power, the probability that you correctly reject the null hypothesis given that it is actually false
- How to estimate β/power?
  - Step 1: Determine the critical test statistic for rejecting the null hypothesis
    - This uses your assumption of how results will be distributed given that the null hypothesis is true and will depend on your chosen α
  - Step 2: Propose an alternative distribution given that the null hypothesis is actually false
  - This will require you to make some additional assumptions about the alternative hypothesis
  - How would observations be distributed given the alternative hypothesis?
  - Most commonly, you would choose the same distribution as for the null hypothesis, just with a different mean

Example

- Let's assume that you want to test the effect of a new depression treatment on mood
- Unrealistically, let's assume that the population of depressed patients you are studying score a mean of 0 on a standardised mood scale. Even more unrealistically, let's assume that the mood scale values are normally distributed with an SD of 4. You would be satisfied with your treatment if it increased mood by at least four units on the scale. In other words, the effect size you want to detect has a Cohen's d of $d = \frac{4-0}{4} = 1$.

Example (continued)

- Let's say you are testing 16 participants, and the null hypothesis is that this group of 16 is a sample from the general population (with a mean of 0).
- The null hypothesis predicts that the sample means come from a normal distribution with a mean $\mu = 0$ and a standard error $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}} = \frac{4}{\sqrt{16}} = \frac{4}{4} = 1$. (See what I did there to make things easier?)
- This means that, for a two-tailed test, the critical z-value for rejecting the H_0 is $z_{crit} = 1.96$ (or -1.96, but this doesn't matter for this analysis since we will be looking at the right side of the distribution).

Example (continued, 2)

- If the alternative hypothesis is true, we assume that the sample for the 16 participants will come from a normal distribution with a mean of $\mu = 4$ and a standard error $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}} = \frac{4}{\sqrt{16}} = \frac{4}{4} = 1$ instead.
- Now what is the probability that given this alternative distribution we get a sample with a mean that is lower than $z_{crit} = 1.96$? We can ask Excel: `=NORM.DIST(1.96,4,1,TRUE)`, which tells us that the area under the left tail of the distribution is $\beta = 0.0206752$. Therefore, our power is $1 - 0.0206752 = 0.9793248$ or 98%.

Example (continued, 3)

- How about if we have only 8 participants? We could calculate this again, but we can just try using the interactive visualisation by Kristoffer Magnusson at http://rpsychologist.com/d3/NHST/
- First, try to replicate our analysis here
- Then, try different values.
  - What happens if we increase/decrease sample size?
  - What sample size do we need for a power of .8?
  - What happens if we reduce α to .025 (e.g. because we need to correct for multiple comparisons)?

A more realistic example

- Let's say we don't know the population SD for the scale, so we need to use a t-test instead. Let's also say we have a control group that doesn't receive the treatment, so we will have an independent samples t-test.
- Based on previous experiments, we still think that the SD of mood scores in both groups will be around 4, so a change in scale by 4 corresponds to a Cohen's d of 1.
- I could show you how to do this by hand using different t-distributions now, but the principle is the same as for the z-test
- Instead, I will show you how to use GPower (http://www.gpower.hhu.de/en.html) to calculate β and the power

What does a significant test actually tell us?

- Let's look back at the table at the beginning of the lecture

	H_0 is true	H_1 is true
H_0 not rejected	correct	β, Type II error
H_0 rejected	α, Type I error	Power

- What proportion of significant p-values is actually due to a false positive?
  - If you want to say .05, that is not necessarily true!
  - Let's fill the table with some numbers to illustrate this. We need to make some assumptions:
  - Let's assume we have an α of .05 and a power of .8 and we are looking at 200 studies
  - Most importantly, let's assume that 50% of all hypotheses that are tested are actually true

False positive rate

	H_0 is true (50%)	H_1 is true (50%)
H_0 not rejected	95 studies (47.5%)	20 studies (10%)
H_0 rejected	5 studies (2.5%)	80 studies (40%)

- How realistic is a proportion of 50% true hypotheses? We don't really know!
    - But we would hope that our hypotheses aren't so bad that only a small fraction of them is true...
- We have 85 significant tests. Out of these 85, 5 are false positives. That is 5.88%. This is called the false positive rate.
- On the other hand, 80 of the 85 significant tests are true positives. That is 94.11% and is called the positive predictive values.
- These terms were coined by Ioannidis in a research paper that got a lot of attention: http://doi.org/10.1371/journal.pmed.0020124

Why this is a problem

- The example above doesn't look so bad. But let's look at a more realistic situation
- Felix Schönbrodt developed this interactive visualisation of the false positive rate: http://shinyapps.org/apps/PPV/
- Try using it to take a look at the situation in the table, then explore other situations
- The visualisation also has an option called % of p-hacked studies. This takes a portion of studies that should be non-significant and changes them to significant (e.g. in our example, setting 10% of studies as p-hacked would cause 11 of the 110 non-significant studies to be counted as significant)

How does p-hacking happen?

- Tests with an inflated Type I error rate are reported as having been performed with $\alpha = .05$
- This can be due to
  - experimenter degrees of freedom (“multiverse” of possible data analyses)
  - HARKing (hypothesising after the results are known)
  - Multiple comparisons without controlling
  - Assumption violations (the least likely by far)

Why this is a problem (2)

- Let's make some relatively safe assumptions
  1. 40% of hypotheses in Psychology are true (the Replication Project was able to replicate about this many results)
  2. The average power is .5 (I fear it is probably lower)
  3. 10% of studies are p-hacked (again, quite a conservative estimate)
- This results in a false discovery rate of 28.3%
  - According to this estimate, between 1 in 4 and 1 in 3 Psychology studies with "significant" results are actually false
  - More pessimistic assumptions paint a worse picture
- Try playing with the α slider. What happens if α is reduced below .05?

Fisher's interpretation of p-values

- Fisher did not like Neyman and Pearson's NHST approach
- His conception of probability was somewhere between subjective and objective
- In Fisher's opinion, a p-value of .01 provides more evidence against the null hypothesis than a p-value of .05
- This is where the common practice of labeling p-values as "highly significant" or "marginally significant" comes from
- The problem is that this goes against the strict logic of NHST
- The rejection rules are strict for a reason
- In general, using the p-value as a measure of evidence is a bad idea. We will see how Bayesian methods are much better suited for this

Criticism of NHST

- It is inflexible and offers only black and white decisions
- Two significant studies, e.g. one with 10 participants and one with 250, can provide drastically different amounts of evidence for and against the null hypothesis. NHST makes us ignore this difference
- The focus on null hypotheses limits the amount of thinking researchers do about the alternative hypotheses. Vague alternative hypotheses are the norm.
- Most null hypotheses are certain to be wrong – it is very unlikely for the true effect to be exactly 0 (but it may well be close to 0)

Criticism of NHST continued

- What is the reference class for the family-wise error rate?
  - Imagine you have conducted a t-test with p = .04
  - If this is the only t-test that you had planned on doing, it is significant
  - If you were going to do another t-test it is not, because you have to correct for multiple comparisons
  - Do you even still remember what you were going to do when you planned on doing the experiment?
  - Even more extreme: Stopping rules. Scientist X and Scientist Y both tested 30 participants and have test with p = .04, but Scientist X checked the data at 15 participants and would have stopped if the test had been significant then. Scientist X has a non-significant result, while Scientist Y, who didn't have a stopping rule, has a significant result.