Avoiding an inflated alpha

Summary of the first part

- We have spent the last lectures talking about the basics of null-hypothesis significance testing (NHST), including considerations of power and what a *p*-value actually means
- We have also discussed an alternative statistical approach that does not involve a null hypothesis (Bayesian Statistics) -In this last part, I want to introduce you to some approaches to diagnose (and perhaps remedy) an inflated alpha rate and questionable research practices in general

Checking p-values

- A surprising amount of published articles contains errors in reporting the p-value
  - For example, the p-value reported does not correspond to the test statistic and degrees of freedom
  - Nuijten, Hartgerink, van Assen, Epskamp, and Wicherts found that half of psychology papers published between 1985 and 2013 contained at least one incorrect p-value.
    - They now have a website that can analyse a manuscript automatically and spot problematic p-values

Analysing digits

- Benford's law:
  - The first digit of any number is more likely to be 1 than to be any other number
  - The distribution of numbers for the last digit should be uniform
  - If this is not true for the data of the experiment,something strange is going on
    - Not necessarily fraud, but maybe some weird rounding issue?
    - See this Datacolada (Simonsohn, Nelson and Simmon's blog) post for an example

Reporting requirements

- Have authors report the full design that was run, not just the subset that they find interesting
- Extreme (and artificial) example: Simmon, Nelson, and Simonsohn's False Positive Psychology paper
- You can get anything significant if you add enough participants, subconditions, etc. without correcting for multiple comparisons
- How can you make sure authors tell the truth about their design?

Preregistration

- Have authors pre-register their study *before* actually running it
- Either at an independent institution such as the Open Science Foundation (OSF)
  - Anyone can do it, and even if you can't get the manuscript published elsewhere, you can put it online there
- Or at a journal
  - Advantage: The journal commits to publishing the manuscript, even if the tests yield null results
  - Unfortunately, not all journals offer this option yet (although some big cognitive psychology ones have just started)

Open Science

- Require authors to share their data (that also enables the digit analysis described above)
- Ideally, data sharing becomes the norm voluntarily
  - but it may also be mandated by journals and research funders (e.g. UK Research and Innovation)
- Extra work necessary to prepare the data for publication
  - Need to safeguard participant privacy
- Opens authors up to greater scrutiny
  - But wouldn't you want to know if you had made an error?
    - Authors should be given opportunity to fix (if possible and error wasn't deliberate)

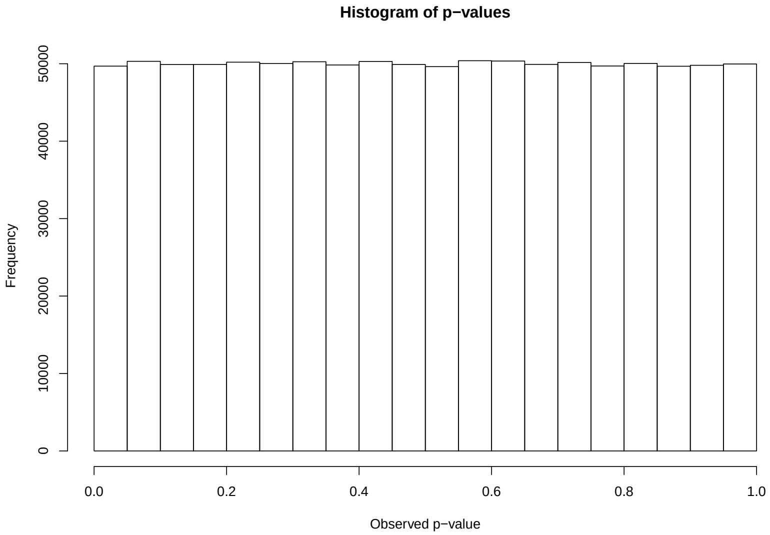
Meta-analysis

- Analyse many studies to get a more consistent picture of the research field
- Some studies are clearly outliers
- Bayesian meta-analysis gives posterior estimate of effect size – very useful
- However: what to do about the file-drawer problem?
  - What about all the experiments with null effects that were never published?

The p-curve

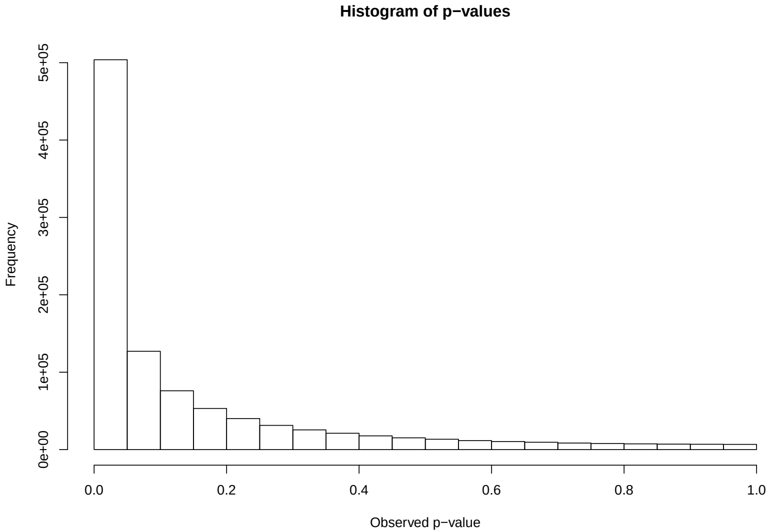
- What is the distribution of p-values given that the null-hypothesis is true?
- You have already seen a bit of this in the "Dance of the p-values" video that you watched in KTS.
- Try this visualisation by Kristoffer Magnusson.

The distribution of p-values when the null hypothesis is true



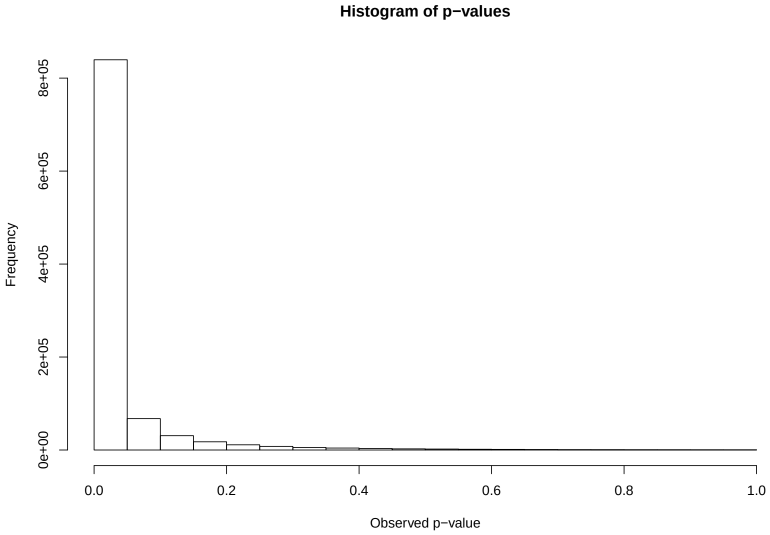
The distribution of p-values when the null hypothesis is false

- and we have 50% power



The distribution of *p*-values when the null hypothesis is false

- and we have 80% power



Using the *p*-curve as a diagnostic tool

- You need a large(r) number of significance tests (e.g. from all the studies on a particular phenomenon such as Power Posing)
- If there is no effect (or low power), the *p*-curve will be approximately flat.
- If there is a real effect (and at least medium power), the *p*-curve will be right-skewed, with low *p*-values more likely than high *p*-values

What if a research field systematically neglects to publish significant results?

- To the left of .05, the *p*-distribution is the same as it should be (all results with p<.05 are published)
- To the right of .05, *p*-values get a lot less frequent as they end up in the file drawer
- There will be a bump in the distribution just below .05

Try it in Felix Schönbrodt's *p*-hacker simulation

- *p*-hacker: Train your *p*-hacking skills!
  - You can run lots of studies without correcting for multiple comparisons
    - You can also add predictor variables that weren't in your original hypothesis
    - Eliminate outliers, test more participants, etc. while always checking the p-value after every change
- You can then send the *p*-values from your simulations to the *p*-checker app to draw a *p*-curve
- The *p*-checker app also has several other useful tests that are explained on the website

Lowering the *alpha*

- This is a fairly extreme proposal, but it has had a lot of support in recent years.
- Basic idea: since most studies are going to have an inflated false positive rate anyway, let's keep it acceptable as a whole by lowering the alpha level required for calling a result "significant"
- Is this a good idea? Lots of researchers think so. Lots of others don't.
- Your task for Assignment 2: Come to your own conclusion and describe the debate and your standpoint in your own words.