(mis)interpration of p-values

Inspired by the work of Andrew Gelman et al.

Purpose of this talk

- Make you worry about things
- Trust p-values less
- NOT to shame you for not understanding statistics

What is a p-value?

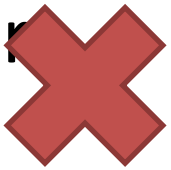
- Result of most statistical packages
- An answer to a question we are rarely asking
- The probability that the data would be „more extreme“ if the null hypothesis is true
- Formally:
  - Statistic T
  - $p(Y_{obs}) = Prob(T(Y_{obs}) > T(Y_{null}) \mid null\ is\ true)$
  - $p \sim Uniform(0, 1)$ If null is true

Example: Two sample T-test

- Null hypothesis: treatment and control have the same mean
- Alternative: treatment and control have different means
- Difference $E(treatment) - E(control)$ is called *effect size*
- Significance: $p < 0.05$

p-value quiz

- If $p < 0.05$ the control is more likely to differ from treatment than not
- If $p > 0.6$ the control is more likely to be the same as treatment than not
- Within a set of results with $p < 0.05$ I expect to get about 5% of false positives



p-value quiz 2

- $p = 0.0001$ means that null is LESS likely to be true than if $p = 0.01$
- $p = 0.01$ with $N = 100$ means that null is LESS likely to be true than if $p = 0.01$ with $N = 20$
- $p = 0.8$ means that null is MORE likely to be true than if $p = 0.2$
- $p = 0.2$ with $N = 100$ means that null is MORE likely to be true than if $p = 0.2$ with $N = 20$

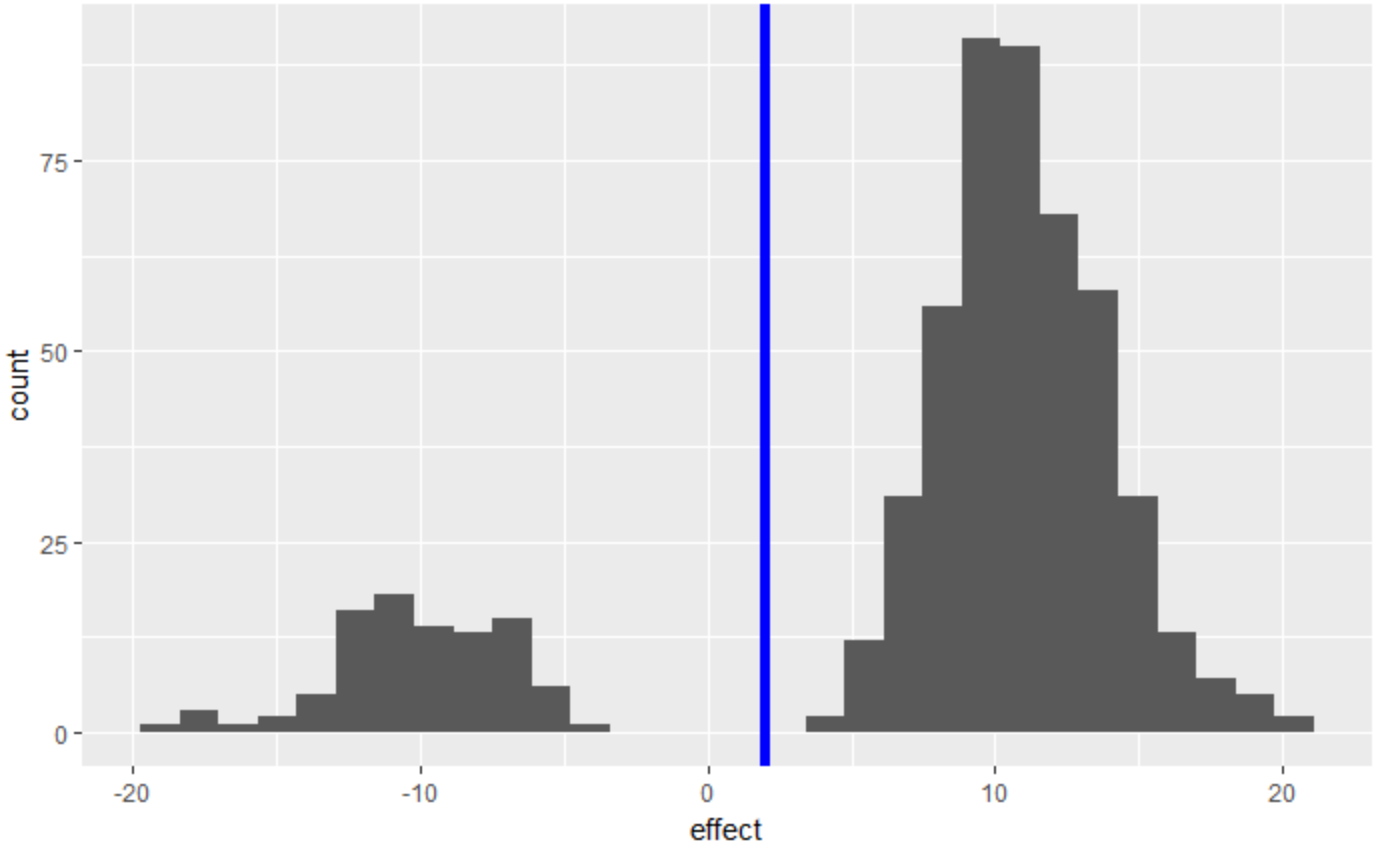
?

p-values are unintuitive

- Heuristic:
  - Strength of evidence against null.
- Interpret as a percentile:
  - $p = 0.04$: "Our data has scored in the 4th percentile"

Type M and Type S errors

- Is obtaining statistical significance always a win?
- Totally made up example:
  - Amount of product produced with or w/o a chemical
  - The chemical helps a little (+2µg on average)
  - Biological variance is large (sd = 8µg)
  - 5 replicates
- Mostly $p > 0.05$ (94% of time)
- But if $p < 0.05$, we are good, right?
  - Right?



- 17% significant effect in the opposite direction
  - Type S(ign) error
- 61% effect > 10 (5x exaggerated)
  - Type M(agnitude) error

Artwork by Viktor Beekman
viktor.beekman@gmail.com

Published effects are exaggerated

- Consequence of filtering published results
  - Not specific to p-values
- How much?
  - Hard to say, depends on amount of noise
- Way out?
  - Publish everything
  - Preregistration

Adjusted p-values

- Running many tests is risky
- Mostly inflated (more stringent)
- Assuming false discoveries are bigger problems than non-discoveries
- False Discovery Rate (FDR)
  - Proportion of discoveries (e.g. $p < 0.05$) that are incorrect

Intro to DESeq2

- The simplest case: two groups of samples
- Assumes:
  - avg. expression in control = 2^{base}
  - avg. expression in treatment = $2^{base + lfc}$
- $lfc = log\ fold\text{-}change$ <- the effect
- Null hypothesis: $|lfc| < threshold$
- We work triplicate experiments, 1000 genes each, 200 genes are DE

DESeq2 guesswork

- DESeq2 controls for $FDR < 0.05$
  - Will the observed FDR be < 0.05 on average?
- Assume true $LFC = 4$ (treatment expression is 16x control), $threshold = 2$
  - how frequently will we get a discovery? (p-adjusted < 0.05)
    - < 33%
    - 33-66%
    - > 66%

DESeq2 guesswork II

- When replicating the experiment, the estimated LFC of DE genes will be
  - Mostly larger
  - Mostly equal
  - Mostly smaller
- Assume true LFC = 4 (treatment expression is 16x control), threshold = 2
  - how frequently will discoveries replicate?
    - < 33%
    - 33-66%
    - > 66%

DESeq2

True LFC	Threshold	True positive	False positives	FDR	P-adjusted for false negatives
0	0	0	3	100%	NA
4	0	129	15	10%	~0.18
4	2	20	<1	2%	~0.28
6	0	168	19	10%	~0.18
6	2	90	<1	1%	~0.23

- DESeq2 has very low Type S and Type M errors!

Replication & DESeq2

Remember 200 genes are truly DE

True LFC	Threshold	Avg. significant	Avg. replicated	Replication smaller lfc
4	0	141	91 (65%)	~50%
4	2	19	5 (27%)	84%
6	0	187	152 (81%)	~50%
6	2	90	53 (58%)	~50%

Other tools then DESeq2?

- You can do these tests yourself:
  - Just need to write code to simulate data
  - No fancy maths needed
- Code to produce today's examples available at BJC web (check your inbox)

Things we ignored today

- What if the model assumptions do not hold?
  - They almost never do
- Hidden assumptions:
  - No systematic bias
  - The effect (LFC) can be 0

Take home messages

- p-values are unintuitive. Don't use intuition when interpreting them.
  - „Strength of evidence against null"
- Worry about Type S and Type M errors
- Worry about variance and measurement error
- Worry about false non-discovery
- Run tests with simulated data to understand your tools