Power and Multiple Hypothesis Testing

Garret Christensen[1]

[1]UC Berkeley:
Berkeley Initiative for Transparency in the Social Sciences
Berkeley Institute for Data Science

IDB, March 2018
Slides available online at
http://www.github.com/BITSS/IDBMarch2018

Outline

BITSS

BERKELEY INITIATIVE FOR TRANSPARENCY
IN THE SOCIAL SCIENCES

What is Statistical Power?

The power of a statistical hypothesis test is the probability that the test correctly rejects the null hypothesis when it is false.

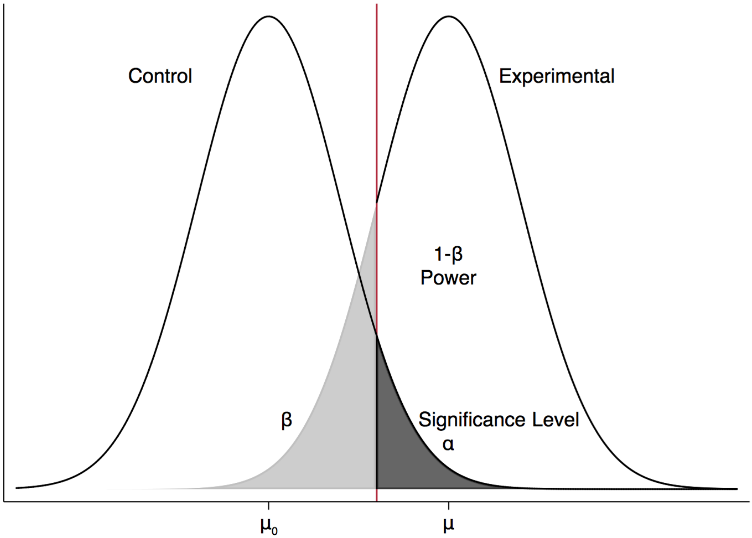
That is, if there's a real effect, what's the likelihood you'll detect it? 80% is the standard.

What is Statistical Power?

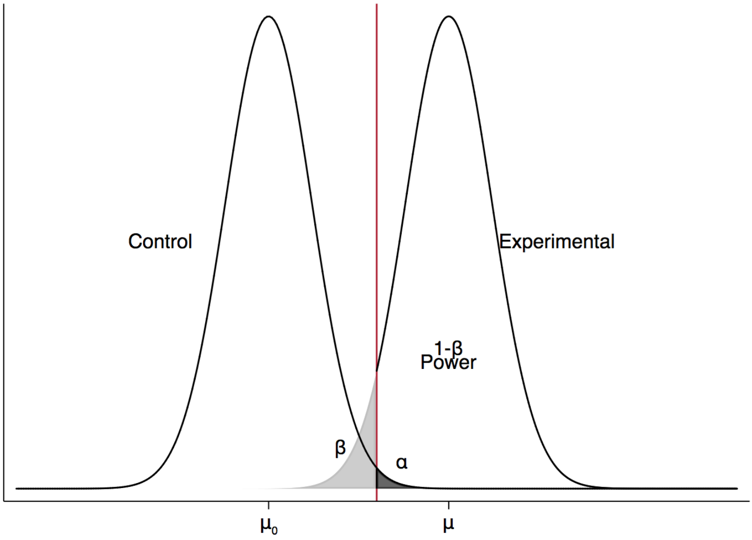
In terms of Type I (false positive) and Type II (false negative) errors:

- Type I error rate is α
- Type II error rate is β
- Power is $1 - \beta$.

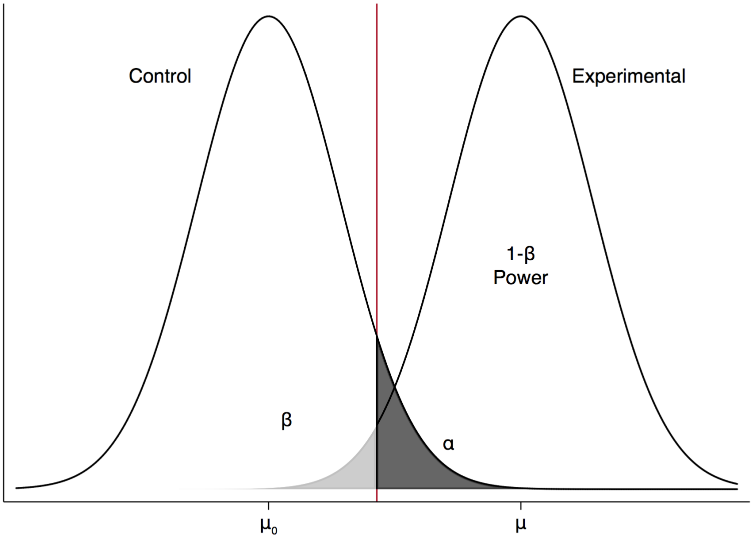
$Power = 1 - \beta$



Less noise, more power



Larger true effect, more power



Derivation

$$Power = 1 - \beta = Pr(Y \geq \ \mu_0 + z_{1-\alpha}\sigma/\sqrt{n}|H_1 \ : \ \mu > \mu_0)$$

$$= 1 - Pr(Y < \ \mu_0 + z_{1-\alpha}\sigma/\sqrt{n}|H_1)$$

$$= 1 - Pr(\frac{Y-\mu}{\frac{\sigma}{\sqrt{n}}} < \ \frac{\mu_0 + \frac{z_{1-\alpha}\sigma}{\sqrt{n}} - \mu}{\frac{\sigma}{\sqrt{n}}}|H_1)$$

$$= 1 - Pr(\frac{Y-\mu}{\frac{\sigma}{\sqrt{n}}} < \ \frac{\mu_0 - \mu}{\frac{\sigma}{\sqrt{n}}} + z_{1-\alpha}|H_1)$$

$$= 1 - \Phi(\ \frac{\mu_0 - \mu}{\frac{\sigma}{\sqrt{n}}} + z_{1-\alpha}|H_1)$$

$$= \Phi(\ \frac{\mu_0 - \mu}{\frac{\sigma}{\sqrt{n}}} - z_{1-\alpha}|H_1)$$

Increasing Power

$$= \Phi\left(\frac{\mu_0 - \mu}{\frac{\sigma}{\sqrt{n}}} - z_{1-\alpha} | H_1 \right)$$

Hopefully the equation makes clear that:

- larger n
- lower σ
- larger true effect size $(\mu_0 - \mu)$
- and a larger α, though that's kind of cheating

all increase power.

Rearranging

Rather than solving for power, you may want to solve for the minimum detectable effect (MDE).

$$MDE = (t_\beta + t_\alpha) * \sqrt{\frac{1}{P(1-P)}}\sqrt{\frac{\sigma^2}{n}}$$

Or, if you've got unlimited funds, pick the minimum biologically or practically meaningful effect, (or your estimate from previous literature of how big the effect will be) and solve for n.

Design Effect

We've so far assumed independent observations, which isn't the case if we cluster treatment. Multiply MDE by the Design Effect:

$$\sqrt{1 + (n-1)\rho}$$

Where n is households per sampling unit, and ρ is the intracluster correlation–variance between clusters divided by sum of within and between.

Complications

Clusters not equal sized? Use the coefficient of variation, but it may not matter much. (Eldridge, Ashby, Kerry 2006)

You get the most power with equal proportions of treated/control. If treatment is very expensive, maximize power subject to your budget constraint. (Randomization Toolkit: Duflo, Glennerster, and Kremer 2007)

Panel with serial correlation? (Burlig, Preonas, Woerman 2017)

Complicated? Simulate it. (Arnold et al. 2011)

Problem of Low Power

So what happens if we have low power?

- More false negatives (Type II error, just β).
- More false positives! More precisely, the likelihood that a reported effect represents a true finding decreases.

Ioannidis 2005

"Why most published research findings are false" (Ioannidis 2005), cited 5600 times.

$$PPV = Pr(True | T > t_\alpha)$$

$$= \frac{(1-\beta)\cdot R}{(1-\beta)R + \alpha}$$

- R is ratio of true relationships to non-relationships tested in a literature.

Derivation

How Bad in Economics?

*"It's **bad**! It's REALLY **bad**."*

–Tom Stanley [Emphasis original]

Source

THE POWER OF BIAS IN ECONOMICS RESEARCH*

John P. A. Ioannidis, T. D. Stanley and Hristos Doucouliagos

We investigate two critical dimensions of the credibility of empirical economics research: statistical power and bias. We survey 159 empirical economics literatures that draw upon 64,076 estimates of economic parameters reported in more than 6,700 empirical studies. Half of the research areas have nearly 90% of their results under-powered. The median statistical power is 18%, or less. A simple weighted average of those reported results that are adequately powered (power ≥ 80%) reveals that nearly 80% of the reported effects in these empirical economics literatures are exaggerated; typically, by a factor of two and with one-third inflated by a factor of four or more.

Statisticians routinely advise examining the power function, but economists do not follow the advice.

McCloskey (1985, p. 204)

Ioannidis, Stanley, Doucouliagos 2017

If we adopt the conventional 5% level of statistical significance and 80% power level, as well, then the 'true effect' will need to be 2.8 standard errors from zero to discriminate it from zero. The value of 2.8 is the sum of the usual 1.96 for a significance level of 5% and 0.84 that is the standard normal value that makes a 20/80% split in its cumulative distribution. Hence, for a study to have adequate power, its standard error needs to be smaller than the absolute value of the underlying effect divided by 2.8. We make use of this relationship to survey adequate power in economics.

Ioannidis, Stanley, Doucouliagos 2017

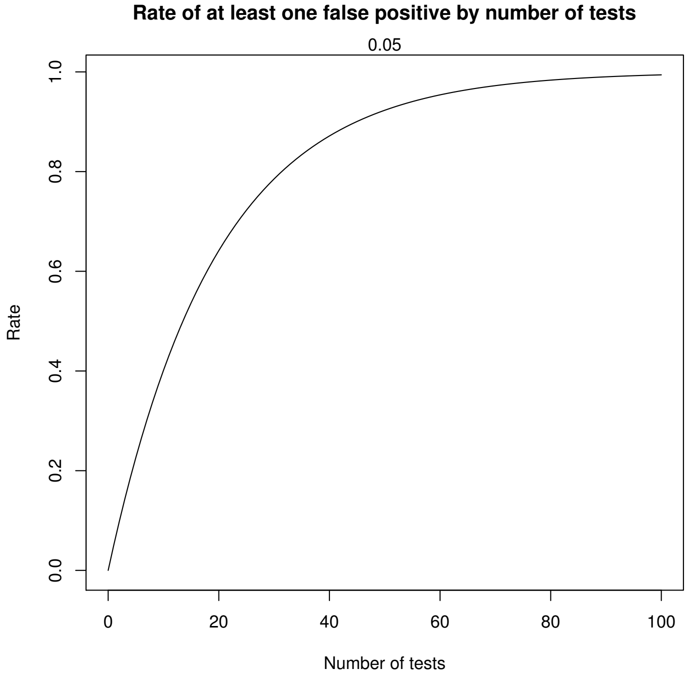
But you still have to find the 'true effect.'
How? Meta-Analysis.

- simple weighted average of all estimates ('fixed effect')
- same for top 10% (smallest s.e.) estimates
- single smallest s.e. estimate (Ioannidis 2013)
- meta-regression estimate (regress estimate on s.e., Stanley 2008)

Speaking of False Positives

It's not just journals and researchers collectively creating publication bias, you can create the same problem all by yourself by testing multiple hypotheses and not adjusting for this. Especially if you only report the significant tests, but also if you report everything.

P(false positive)$= \alpha$
P(no false positives)$= 1 - \alpha$
P(no false positives in m tests)$= (1 - \alpha)^m$
P(at least one false positive in m tests)$= 1 - (1 - \alpha)^m$


Rate of at least one false positive by number of tests
0.05
Rate
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100
Number of tests

Reduce Tests: Summary Index Tests

Reduce number of tests conducted by grouping outcomes into indexes.

- Started with O'Brien (1984)
- Economists know from MTO: Kling, Liebman, Katz (2007).

How:

- Group outcomes into families
- Align direction
- Normalize and sum
- Could also weight for more efficiency (unlikely to matter in practice)
- Interpret as standard deviation unit

Control the Type I Error Rate

Primary methods:

- Family-wise error rate (FWER): the probability of at least one Type I error
- False discovery rate (FDR): the expected proportion of Type I errors among rejected hypotheses

FWER Methods

Bonferroni: divide your cutoff by the number of tests (or multiply p-value by number of tests, same thing)

- Not suggesting you do this
- In fact, I am suggesting you *not* do this (Pernerger 1998)
- It's just easy to understand

Better FWER Methods

Westfall & Young (1993): Free stepdown method

1. Sort by increasing p-value
2. Simulate null data with resampling
3. Calculate simulated p-values, $p_1^*, \ldots, p_M^*$
4. Enforce original monotonicity $p_r^{**} = \min\{p_r^* \ldots p_M^*\}$ where r is original rank
5. $L \geq 10,000$ repetitions, S_r is number of times $p_r^{**} < p_r$
6. $p_r^{fwer*} = S_r/L$
7. original monotonicity one more time: $p_r^{fwer} = \max\{p_1^{fwer*} \ldots p_r^{fwer*}\}$

Free Stepdown

- Dependence between outcomes preserved by resampling
- Larger unadjusted p's correspond to larger adjusted p's

Example

FDR Methods

Maybe one false positive isn't the end of the world, and you want more power. Control the expected proportion of false positives instead (FDR).

Let $V =$ false rejections, $U =$ correct rejections, $t =$ total rejections.

FWER is $P(V > 0)$, FDR is $E[Q = V/t]$

Benjamin Hochberg 1995

1. Sort p-values $1 \ldots M$
2. $q \in (0, 1)$
3. Let c be be largest r for which $p_r < qr/M$
4. Reject all hypotheses $1 \ldots c$ to control FDR at q
5. That is, starting with p_M, check if $p_r < qr/M$. If true, reject it and all smaller p

Benjamini, Krieger, Yekutieli 2006

Sharpen the procedure by estimating number of true null hypotheses

1. Apply BH procedure at level $q' = q/(1+q)$. Let c be number of hypotheses rejected. Continue if $c \neq 0$
2. Let $\hat{m}_0 = M - c$
3. Apply BH at $q^* = q'M/\hat{m}_0$

Note: Procedures describe test for a given q, so test every value from $1, .999, .998\ldots$ and find q when hypothesis stops being rejected.

Software

Stata

- Michael Anderson
  - Wonderfully written JASA paper ▸ Link
  - Stata for Benjamini & Hochberg 1995 ▸ Link
  - Stata for Benjamini, Krieger, & Yekuteli 2006 ▸ Link
- Roger Newson
  - `ssc install qqvalue` ▸ Stata Journal
  - `ssc install smileplot` ▸ Old Stata Journal

R

- p.adjust ▸ Link

Active Area of Research

List, Shaikh, Xu

- Useful in experimental economics:
  - jointly identifying treatment effects for a set of outcomes
  - estimating heterogeneous treatment effects through subgroup analysis
  - conducting hypothesis testing for multiple treatment conditions
- Builds on Romano, Wolf (2010)
- NBER WP
- Github (Stata, Matlab)
- `ssc install mhtexp`

Conclusion

1. Design studies with adequate power
2. Adjust for multiple tests
   - Summary Indexes
   - FWER
   - FDR

Questions?

Thank you!

Derivation of PPV I

$$PPV = Pr(True | T > t_\alpha)$$

Prior to the study, the quantities involved are as follows:

- Probability of a relationship being true: $\frac{R}{R+1}$
- Probability of a relationship being false: $1 - \frac{R}{R+1} = \frac{1}{R+1}$
- Probability of finding a positive statistical association given that the relationship is false: α
- Probability of finding a positive statistical association given that the relationship is true (i.e., power): $1 - \beta$

Derivation of PPV II

Bayes' law says that $Pr(A|B) = \frac{Pr(B|A)Pr(A)}{Pr(B)}$, though it is almost always the case that the denominator is more useful when written out with the law of total probability, as follows: $Pr(A|B) = \frac{Pr(B|A)Pr(A)}{Pr(B|A)Pr(A)+Pr(B|\neg A)Pr(\neg A)}$.
By using Bayes' law, we know that:

$$Pr(True|T > t_\alpha) = \frac{Pr(T > t_\alpha|True) \cdot Pr(True)}{Pr(T > t_\alpha|True) \cdot Pr(True) + Pr(T > t_\alpha|False \cdot Pr\,(\mathrm{False}))}$$

Derivation of PPV III

Substituting, we find:

$$Pr(True|T > t_\alpha) = \frac{(1-\beta)\frac{R}{R+1}}{(1-\beta)\frac{R}{R+1} + \alpha \cdot \frac{1}{R+1}}$$

$$Pr(True|T > t_\alpha) = \frac{\frac{(1-\beta)\cdot R}{R+1}}{\frac{(1-\beta)R+\alpha}{R+1}}$$

Simplifying:

$$Pr(True|T > t_\alpha) = \frac{(1-\beta)\cdot R}{(1-\beta)R+\alpha} = \frac{(1-\beta)R}{R-\beta R+\alpha}$$

This is the same as the formula in Ioannidis (2005) and equation 1 above. Back

Free Stepdown Example

Michael Anderson JASA 2008! Link

"An example may aid interpretation of FWER-adjusted *p*-values. In this research, M = 9 summary indexes were tested for each gender. Consider the smallest summary index *p*-value of the nine male summary indexes, which occurs for adult Early Training males (Table 3). The unadjusted *p*-value is approximately .011. The corresponding adjusted *p*-value, calculated by the free step-down resampling method for the entire family of male summary tests, is $p^{fwer} = .090$. Suppose that we simulate the male data 100,000 times under the null hypothesis of no treatment effect.

Free Stepdown Example II

If we compute an entire set of nine summary effect p-values for each simulation, then the minimum p-value of that set will be less than or equal to the unadjusted p-value of .011 approximately 9% of the time. Thus a minimum observed p-value of .011 is not unlikely under the null given the number of tests conducted?a fact that helps explain why this particular effect goes in the "wrong" (negative) direction. For unadjusted p-values above the family's minimum p-value, the number of tests in the family effectively decreases, making the adjustment less severe."

Back