Many tests, many strata:

How should we use hypothesis tests to guide the next experiment?

Jake Bowers
University of Illinois @ Urbana-Champaign
Departments of Political Science & Statistics
Fellow, Office of Evaluation Sciences, GSA
Methods Director, EGAP
Fellow, CASBS
http://jakebowers.org

Nuole Chen
University of Illinois @ Urbana-Champaign
Department of Political Science
Associate Fellow, Office of Evaluation Sciences, GSA
https://publish.illinois.edu/nchen3

25 January 2019

The Plan

Motivation and Application:
An experiment with many blocks raises questions about where to focus the next experiment

Background: Hypothesis testing and error for many tests

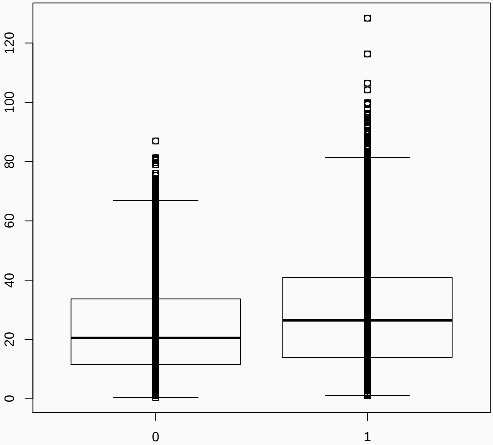
Adjustment based methods for controlling error rate.

Limitations and Advantages of Different Approaches

Motivation and Application: An experiment with many blocks raises questions about where to focus the next experiment

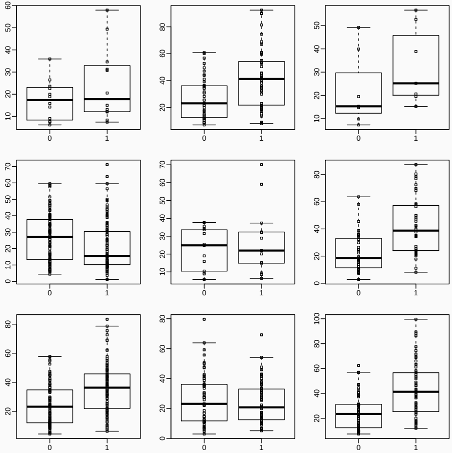
A (Fake) Field Experiment

”...We successfully randomized the college application communication to people within each of the 100 HUD buildings. The estimated effect of the new policy is 5 ($p < .001$).”



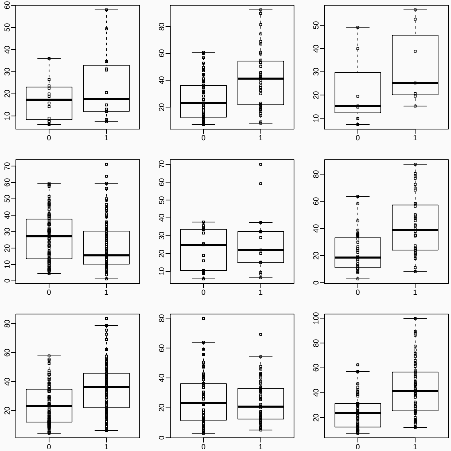
A (Fake) Field Experiment

Policy Expert: …Great! But this seems small. Could it be mainly effecting people in just a few buildings?



A (Fake) Field Experiment

Jake: This is a natural and important question both for the development of new theory and better policy. Let me get back to you...



One response: use large-scale hypothesis testing

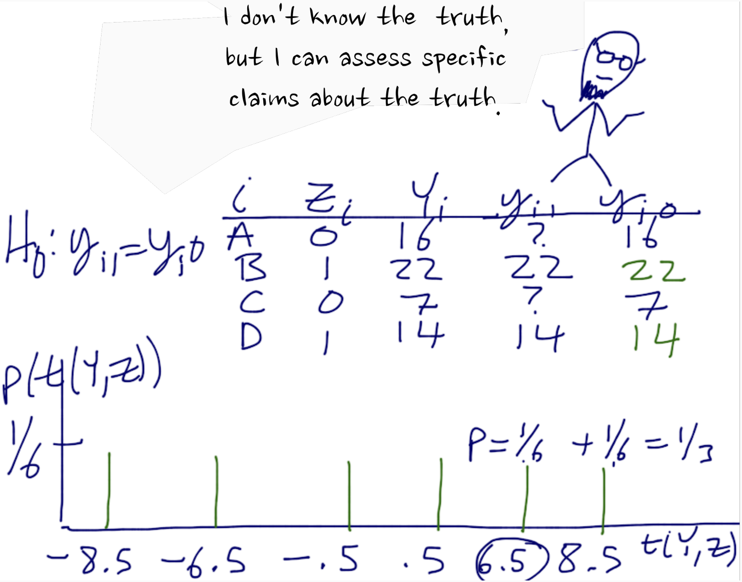
Given a block-randomized experiment, if effects are concentrated in a few blocks (but zero is most or many blocks), how can we detect them?

- Report 100 simple average treatment effects?
  But reporting $\tau_b = \{5, 0, -2, 100, \ldots\}$ will require p-values (or CIs, or SEs, or posterior dists).
- Report 100 simple p-values?
  But we know that the false rejection rate / false positive rate for 100 tests is $1 - (1 - .05)^{100} = .99$.
- Report 100 adjusted p-values?
  But we know that Bonferroni-style adjustments will reduce power to detect effects (i.e. would only reject if $p < .05/100 = .0005$.)

Our proposal to increase power: Test many hypotheses but control the false positive rate via **testing hypotheses in order** (novel in this context) OR by controlling the **false discovery rate** (common in genetics and machine learning).

Background: Hypothesis testing and error for many tests

A hypothesis test summarizes design information against a claim or model.



I don't know the truth, but I can assess specific claims about the truth.

$H_0: y_{i,1} = y_{i,0}$

i	Z_i	Y_i	$y_{i,1}$	$y_{i,0}$
A	0	16	?	16
B	1	22	22	22
C	0	7	?	7
D	1	14	14	14

$P = 1/6 + 1/6 = 1/3$

What is a good hypothesis test?

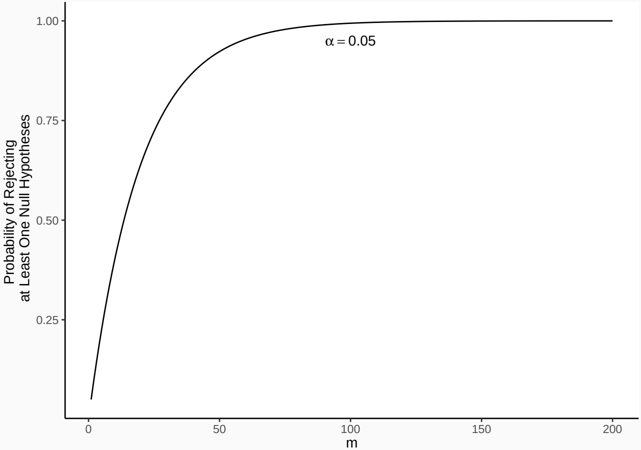
A good test casts doubt on the truth rarely (few false positive errors) and detects falsehoods often (high statistical power).

Rarely and often refer to repeated use in a given research design.

To assess false positive rate, for example:

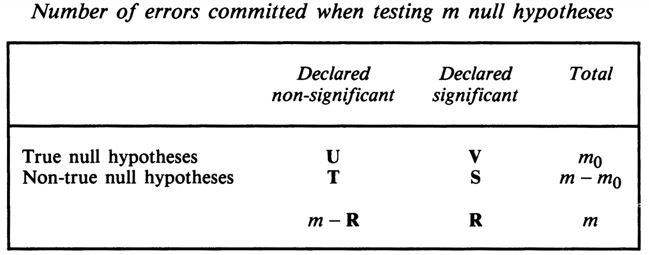
1. Shuffle the treatment assignment to make the true relationship zero but keeping everything else about the data fixed.
2. Test the true null hypothesis, $H_0 : \tau_i = 0$, and record the p-value.
3. Repeat steps 1 and 2 B times.
4. The false positive rate (aka Type I error rate aka size of the test) is $\sum_b I(p \leq \alpha)/B$.

The false positive error rate increases with the number of tests.



This figure shows the relationship between the number of hypotheses m and the probability of rejecting at least one null hypothesis $P(1-(1-\alpha)^{m})$, when $\alpha = 0.05$ Bretz, Hothorn, and Westfall, 2011.

Error rates of multiple testing procedures



Number of errors committed when testing m null hypotheses

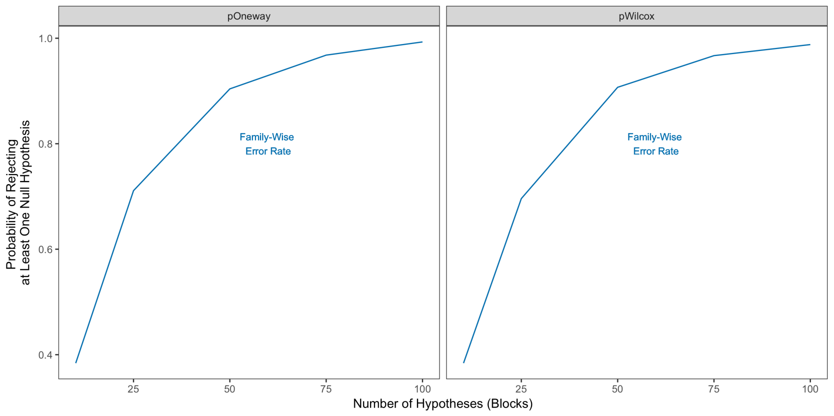
	Declared non-significant	*Declared significant*	*Total*
True null hypotheses	**U**	**V**	m_0
Non-true null hypotheses	**T**	**S**	$m - m_0$
	$m - \mathbf{R}$	**R**	m

(Benjamini and Hochberg, 1995)'s Table 1. Cells are numbers of tests. For example, R are number of ”discoveries” and V are false discoveries.

Two main error rates to control when testing many hypotheses:

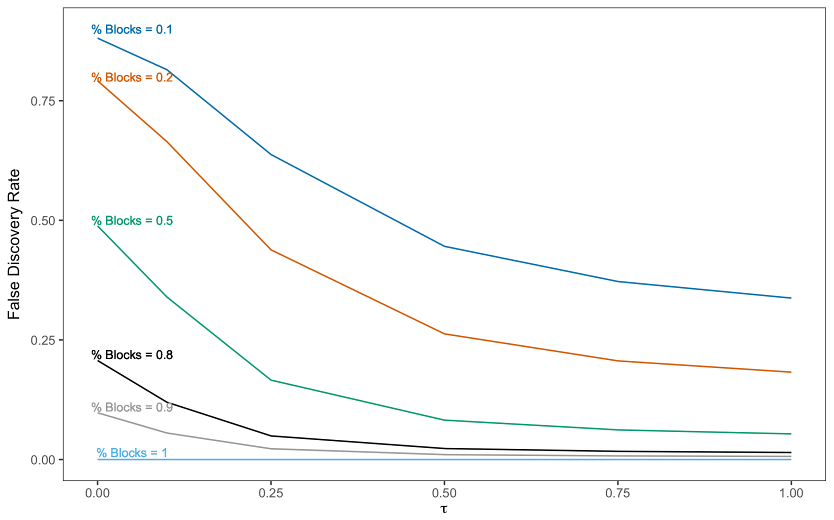
- **Family wise error rate (FWER)** is $P(V > 0)$ (Probability of any false positive error)
- **False Discovery Rate (FDR)** is $E(V/R|R > 0)$ (Average proportion of false positive errors given some rejections.)

Family Wise Error Rate (FWER)



Simulations to show FWER with two different tests, ranging from 10 to 100 hypotheses (blocks), where each hypothesis (block) is one test. Same data as shown earlier but sampled to show effects in datasets with 10 ...100 blocks. ”pOneway” is a t-test using means. ”pWilcox” is a Wilcox-Mann-Whitney rank based test.

False Discovery Rate



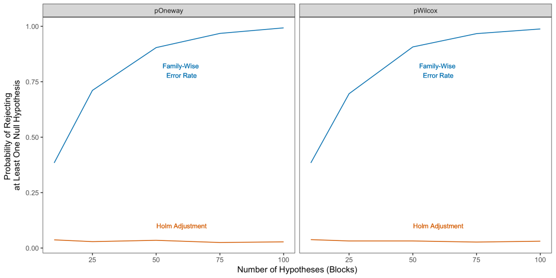
This simulation of 100 hypotheses tests shows that the false discovery rate decreases when τ increases and/or when more blocks (hypotheses) have true effects.

Adjustment based methods for controlling error rate.

Some ways to control error rates: Holm-Bonferroni-style adjustment

Family-Wise Error Rate (FWER) Control via Bonferroni-style adjustment

1. Set α
2. Test hypotheses and record p-values
3. Order p-values of tests from smallest to largest
4. Reject hypothesis i if $p_i \leq \alpha/(m - i + 1)$, where m is the number of hypothesis



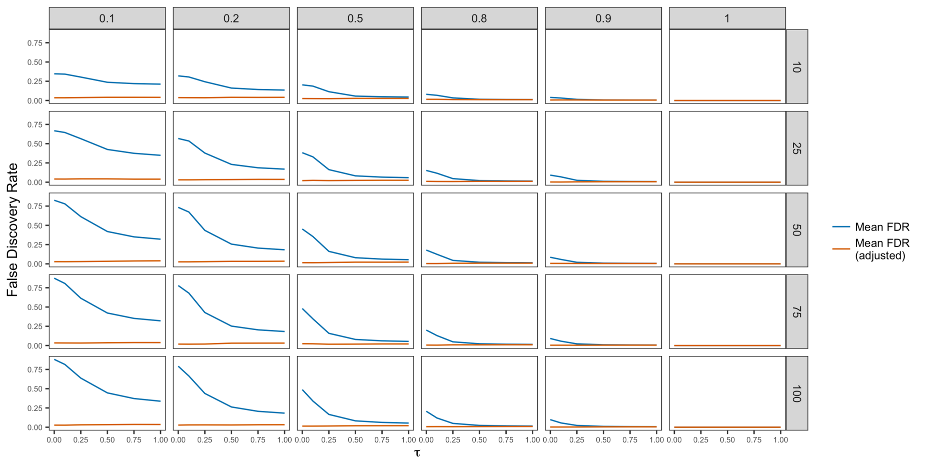
Simulations show FWER with Holm-Bonferroni adjustments.

Some ways to control error rates: Benjamini-Hochberg FDR control

- Set α
- Test hypotheses and record p-values
- Order p-values of tests from smallest to largest
- Reject hypothesis i if the corresponding p-value $p_i \leq (i/m)\alpha$, where m is the number of hypotheses
- Reject all hypotheses with p-values that are smaller than p_i

Some ways to control error rates: Benjamini-Hochberg FDR control

Controlling False Discovery Rate (FDR)



The false discovery rates and false discovery rates after the Benjamini-Hochberg adjustment from our simulations.

Some ways to control error rates: Structured hypothesis testing.

Our approach: (1) creating a hierarchy of hypotheses based on a priori power (e.g. sample size of block); (2) test hypotheses on smaller and smaller subsets of blocks in a predetermined order, stopping testing when $p > .05$.

1. Test overall $H_0 : \tau_i = 0$. If $p_0 > \alpha$ stop testing.
2. If $p_0 \leq \alpha$, split the sample.
3. Test H_0 in each set of blocks. Record maximum p-value for the set, p_s, so far (overall versus test in this split). If $p_s > \alpha$ or split contains only 1 block, stop.
4. Repeat steps 2 and 3 until stop.

Why should this work? Maximum power test is p_0. If we cannot reject then we should not be able to reject on any other subset.

Splitting algorithms

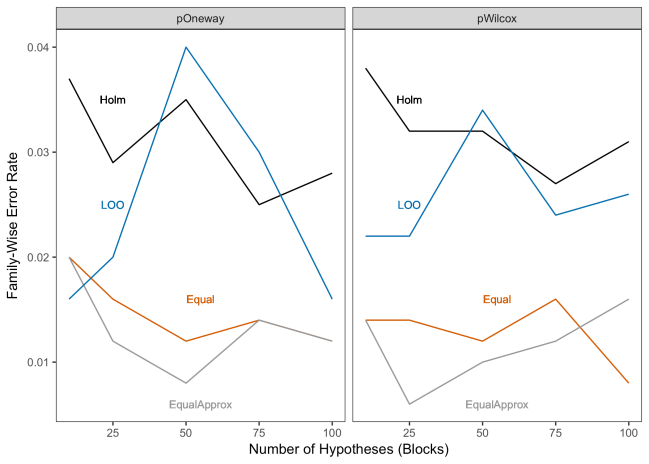
How should we choose to split? We need the statistical power of each split to be less than the preceding splits.

Idea 1: Leave one out splitting Quickly focus down on individual blocks. Algorithm: Rank blocks in order of sample size (or harmonic mean weight (w)). Split the data into (1) block with largest w and (2) the rest of the blocks.

Idea 2: Equal power splitting Ensure that each test has more or less the same statistical power (ex. sample size). Algorithm: Rank blocks in order of sample size (or harmonic mean weight (w)). Split the data into two groups of equal $\sum_b w_b$ (equal sums of weights).

SIU and FWER control

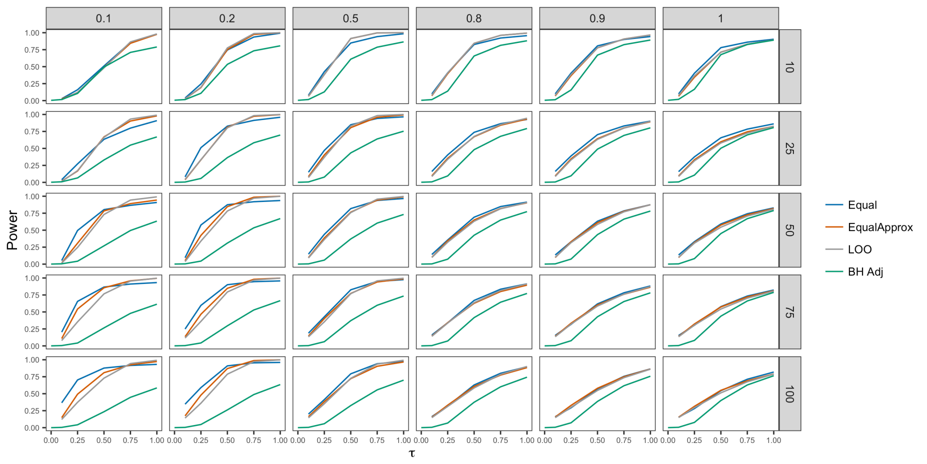
How well do the splitting algorithms control FWER?



Comparison of FWER control between splitting algorithms and Holm-Bonferroni Procedure. Average number of splits when all $\tau_i = 0$ is 1.5

SIU and power

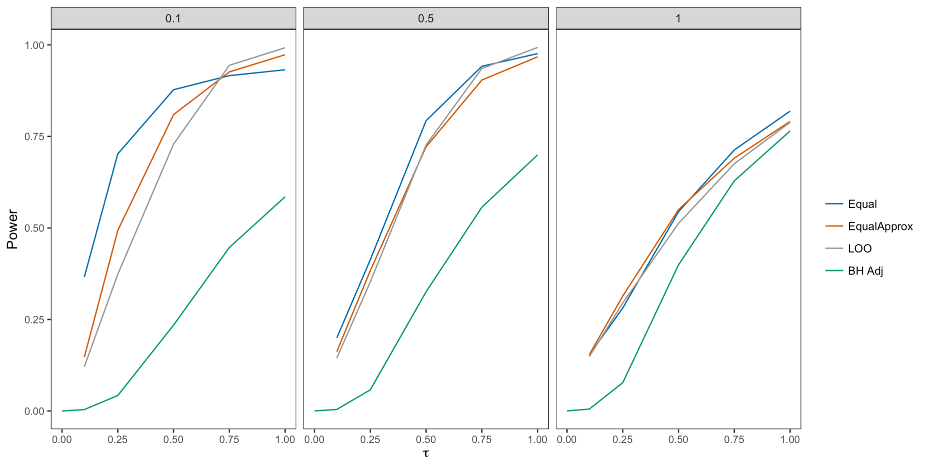
How powerful are the splitting algorithms?



Comparison of power between splitting algorithms and Benjamini-Hochberg Procedure.

SIU and power

How powerful are the splitting algorithms?



Comparison of power between splitting algorithms and Benjamini-Hochberg Procedure, for 100 hypotheses (blocks).

Applying the procedures

In our fake data of 100 blocks and 12,500 units we made 51 blocks with some treatment effect (between .5 and 1.4 sds) and 49 with no effects.

- SIU found 32/51 correctly and made no false positive errors (indicating a true effect where non-existed) (same in both splitting algorithms)
- FDR found 32/51 correctly and 1 false positive.
- HOLM found 27/51 correctly with no false positives.

Applying the procedures

We can now go back to our policy expert with a list of blocks (hypotheses) for:

- further exploration about mechanisms (ex. interviews, observation, databases, discussions, design-sprints,...)
- to help direct attention toward blocks where no effects could be detected
- to design studies powered towards specific blocks.

Limitations and Advantages of Different Approaches

Advantages and Limitations

Limitations:

- Limitation of testing versus estimation: Each block stands alone. No (partial) pooling of information across blocks. No model of differences between blocks to enhance precision and enable subgroup estimation (cf. Feller and Miratrix on multi-site trials)
- Limitation of the SIU approach versus FDR: The SIU approach requires choice of splitting algorithm and ranking of subgroups by a fixed quantity — little guidance on this compared to big literature on FDR.
- Limitation of the SIU approach versus FDR: SIU controls FWER and so should be conservative compared to FDR when it comes to FDR. (Probably could get more power from SIUP by focusing it on FDR control rather than FWER).

Advantages

Advantages and Limitations

- Advantage of testing over estimation of average treatment effects: Tests summarize information — of direct policy planning relevance. Flexibility with test statistics can increase power (ex. means are low powered when we have outliers compared to ranks or robust means.).
- Advantage of testing over estimation of average treatment effects: Randomization inference within small or skewed blocks ensures correct false positive rates for individual tests.

Advantages of Sequential, Structured Testing:

- Advantage of SIU over FDR: Fewer tests required (ex. when effect is zero everywhere, it often would report one single test rather than *m* tests; when effect is not zero it seems to do roughly 10% of the tests of the other procedures.)
- Advantage of SIU over FDR: More power with same FWER control.

References

Benjamini, Yoav and Yosef Hochberg (1995). “Controlling the false discovery rate: a practical and powerful approach to multiple testing”. In: Journal of the Royal Statistical Society 57.1, pp. 289–300. ISSN: 00359246. DOI: 10.2307/2346101. arXiv: 95/57289 [0035-9246]. URL: http://www.jstor.org/stable/2346101%7B%5C%%7D5Cnhttp://about.jstor.org/terms.

Bretz, Frank, Torsten Hothorn, and Peter Westfall (2011). Multiple comparisons using R. Boca Raton: CRC Press, p. 182. ISBN: 9781584885740. DOI: 10.2307/1266041. arXiv: arXiv:1011.1669v3. URL: http://www.jstor.org/stable/1266041?origin=crossref.