Uncommon Hypothesis Tests to Debunk Common Misconceptions

Kellie Ottoboni
September 28, 2017

University of California, Berkeley
Department of Statistics

Berkeley Institute for Data Science

Road map

- With great *p*-values comes great responsibility
- Examples
    1. Pseudo-random number generators
    2. Student evaluations of teaching
    3. Risk-limiting election audits

Monkey Cage

Does social science have a replication crisis?

Cancer Research Is Broken

There's a replication crisis in biomedicine—and no one even knows how deep it runs.

By Daniel Engber

SundayReview

Why Do So Many Studies Fail to Replicate?

Gray Matter

By JAY VAN BAVEL MAY 27, 2016

RESEARCH ARTICLE

Estimating the reproducibility of psychological science

Open Science Collaboration[*,†]

NATURE | EDITORIAL

Reality check on reproducibility

POLICY & ETHICS

Is There a Reproducibility Crisis in Science?

PLOS | MEDICINE

OPEN ACCESS

ESSAY

Why Most Published Research Findings Are False

John P. A. Ioannidis

Published: August 30, 2005 • http://dx.doi.org/10.1371/journal.pmed.0020124

About 40% of economics experiments fail replication survey

By John Bohannon | Mar. 3, 2016 , 2:00 PM

NATURE | NEWS

Over half of psychology studies fail reproducibility test

Largest replication study to date casts doubt on many published positive results.

Monya Baker

27 August 2015

The ASA's Statement on p-values (2016)

Informally, a *p*-value is **the probability** under a specified statistical model that a statistical summary of the data would be equal to or **more extreme than its observed value**.

A *p*-value is **not**

- the probability that the model is true
- evidence for the model
- a measure of effect size
- a measure of importance
- valid after trying out many different models

The ASA's Statement on p-values (2016)

Informally, a *p*-value is the probability **under a specified statistical model** that a statistical summary of the data would be equal to or more extreme than its observed value.

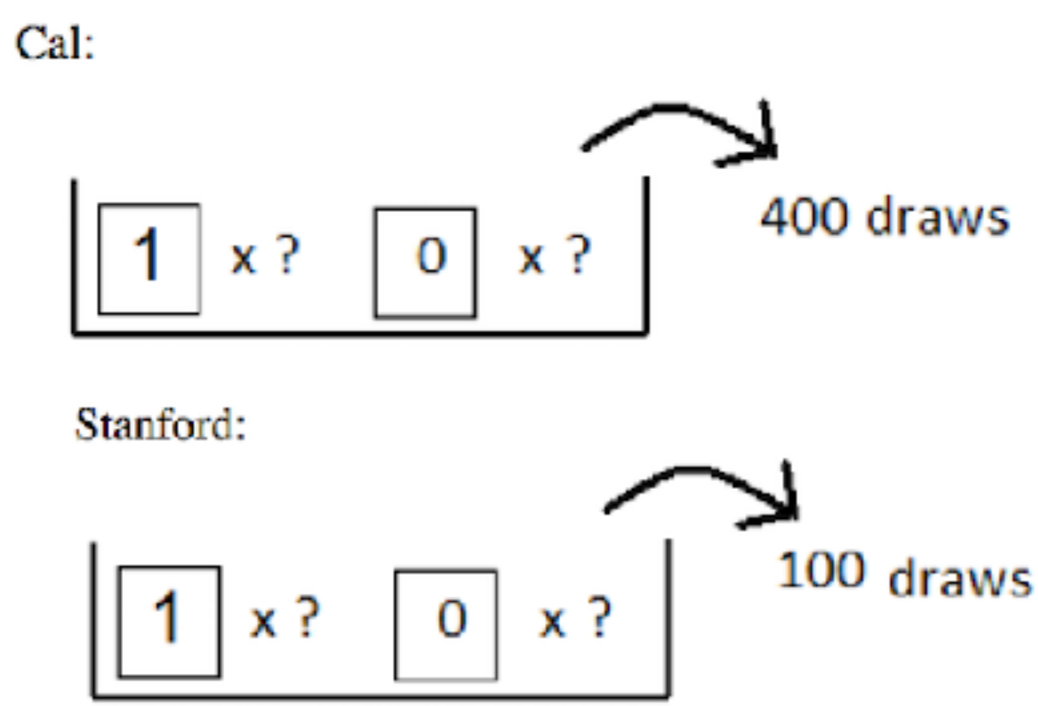
Freedman's Rabbit-Hat Theorem

To pull a rabbit out of a hat, at least one rabbit must first be placed in the hat.

Compare assumptions…

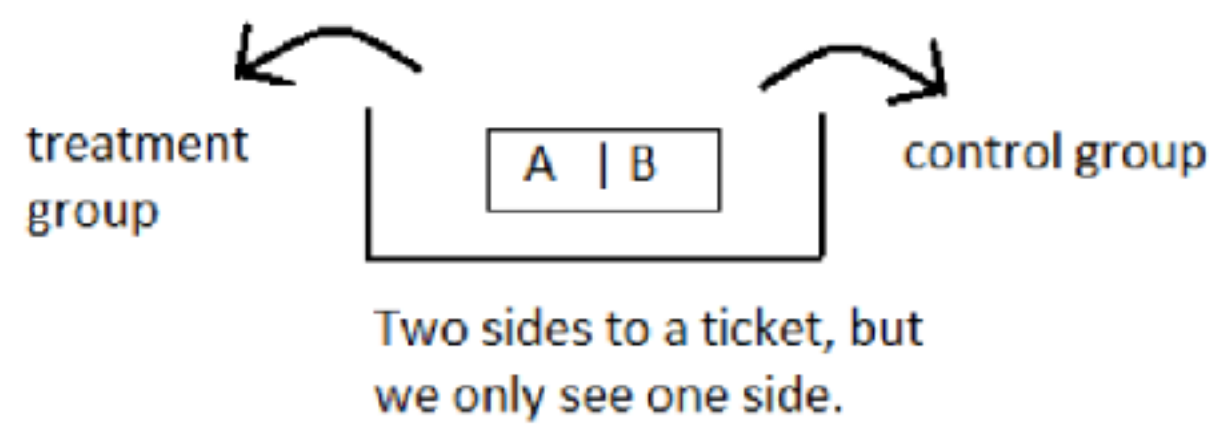
Two sample *t* test

- Samples are from two independent populations
- Data are normally distributed



Permutation test

- Samples are from a single population
- Group membership is randomly assigned



Permutation tests and confidence sets

build passing coverage 99%

Permutation tests and confidence sets for a variety of nonparametric testing and estimation problems, for a variety of randomization designs.

- **Website (including documentation):** http://statlab.github.io/permute
- **Mailing list:** http://groups.google.com/group/permute
- **Source:** https://github.com/statlab/permute
- **Bug reports:** https://github.com/statlab/permute/issues

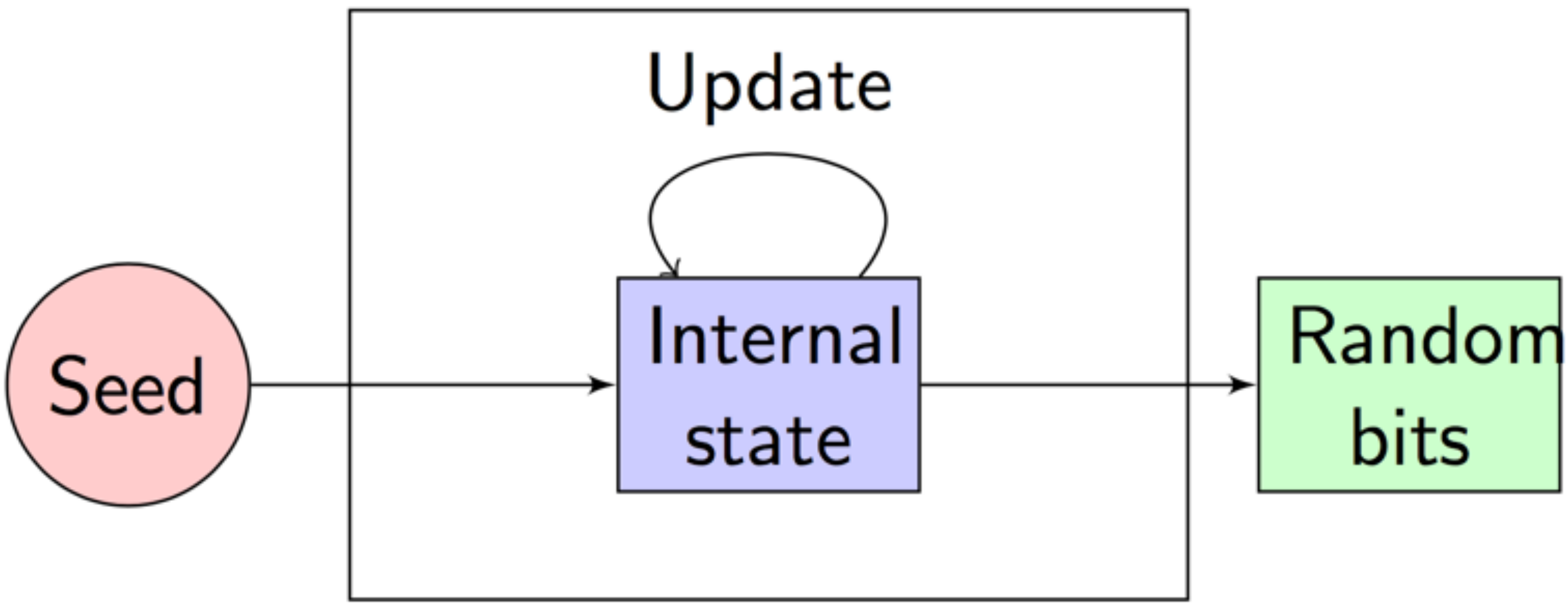
Installation from binaries

```
$ pip install permute
```

Myth #1:
The default pseudo-random number generator works well enough.

http://dilbert.com/strip/2001-10-25

Pseudo-random number generators



The Reverse Pigeonhole Principle

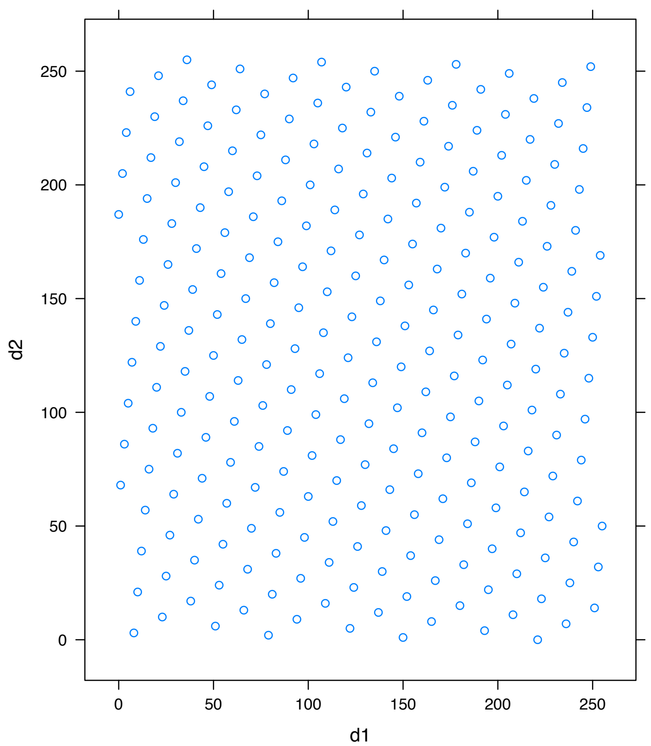
If the number of possible random samples is larger than the size of a PRNG's state space, then the PRNG cannot possibly generate all samples.

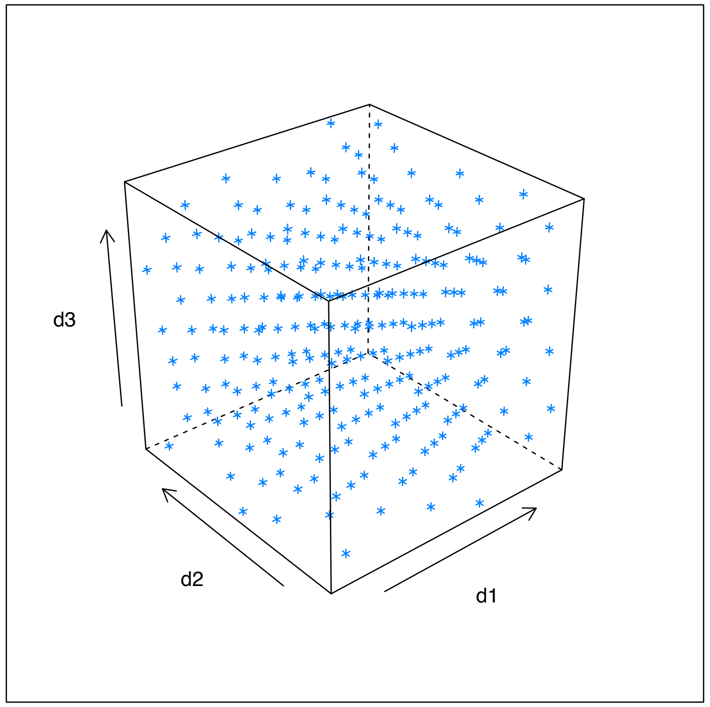
Does it matter in practice?

PRNG	# Internal States	# Possibilities	Proportion attainable
32-bit linear congruential generator	4 billion	Samples of 10 items out of 50: ~ 10 billion	0.4
Mersenne Twister	~ 2 x 10^{6010}	Permutations of 2084 items: ~ 3 x 10^{6013}	0.0001

Testing PRNGs

- Uniformity: samples should occur with equal frequency
- Independence: there should be no serial correlation in outputs
- Compressibility: a predictable sequence contains less information than a random one





Test for serial correlation

1 2 3 4 5

1 3 4 5 2

Test for serial correlation

1 2 3 4 5

1 3 4 5 2 $\longrightarrow$ **1**

Test for serial correlation

1 2 3 4 5

1 3 4 5 2 $\longrightarrow$ **1**

2 4 5 3 1

.

.

.

Test for serial correlation

1 2 3 4 5

1 3 4 5 2 $\longrightarrow$ **1**

2 4 5 3 1 $\longrightarrow$ **0**

. **.**

. **.**

. **.**

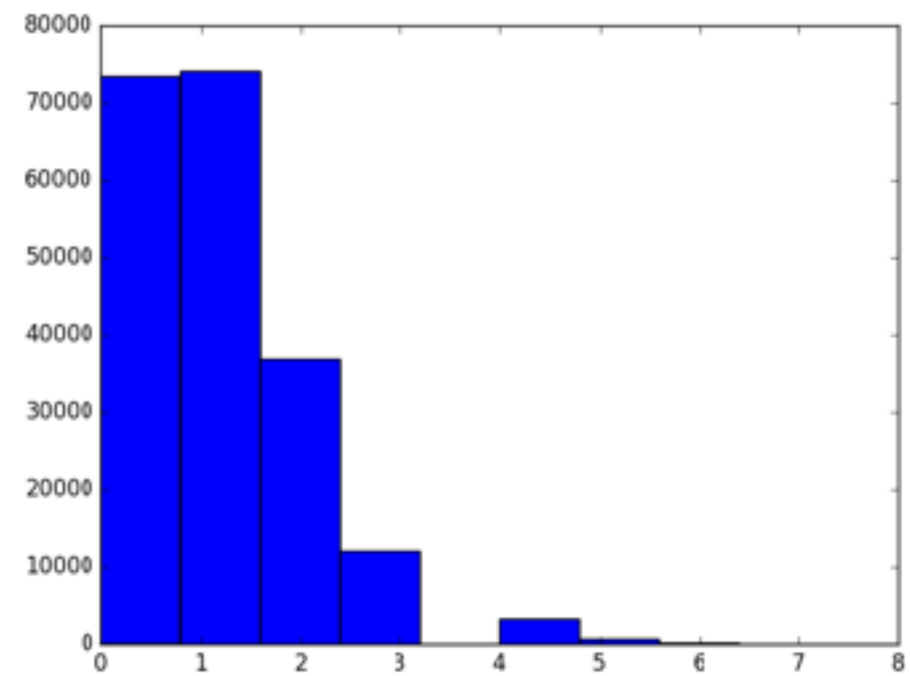
Test for serial correlation

1 2 3 4 5

Poisson(1)?

1 3 4 5 2 → 1

2 4 5 3 1 → 0 →



.

.

.

.

.

.

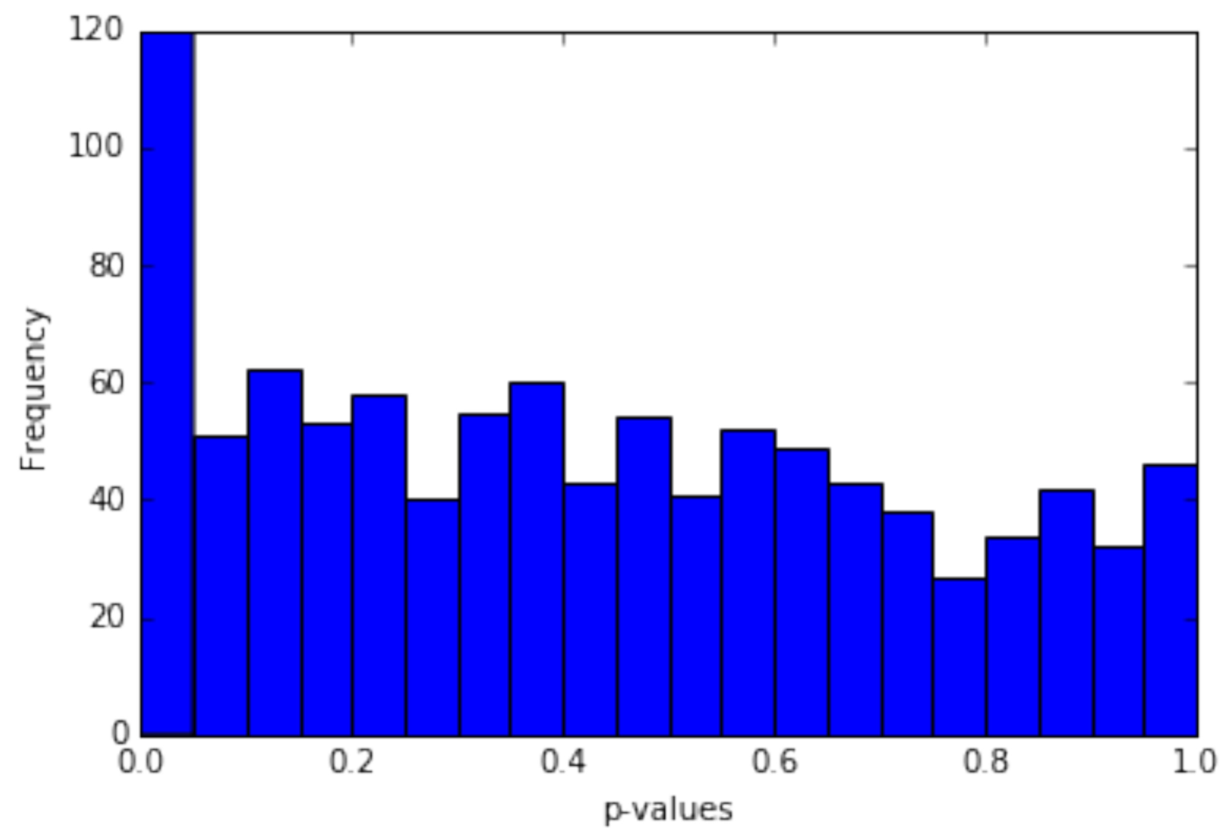
Test for serial correlation

1. Generate 1000 random seeds
2. For each seed:
    1. Generate many permutations
    2. Compute the *p*-value
3. Test whether these 1000 *p*-values are uniform between 0 and 1

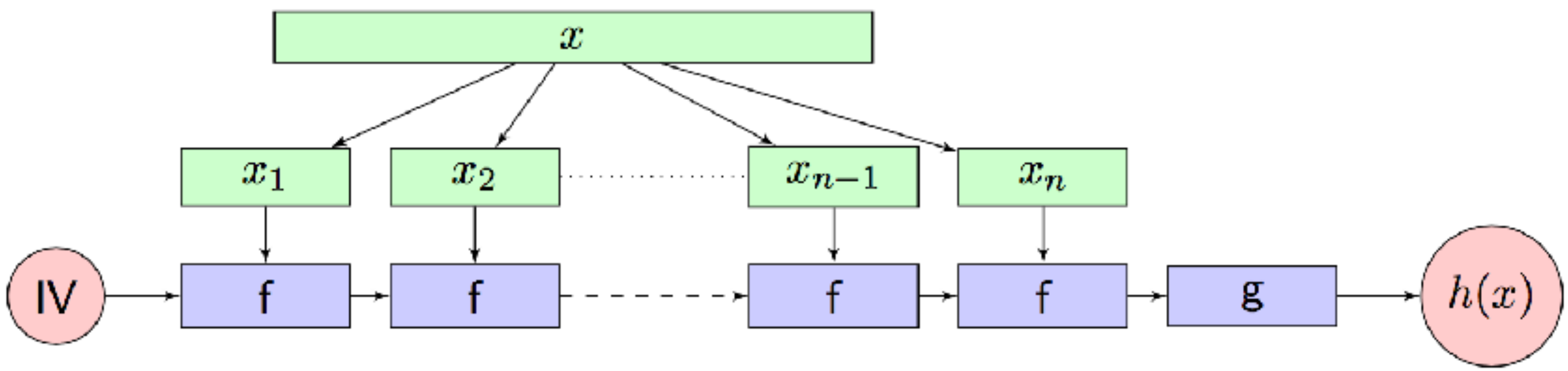
Test for serial correlation

1. Generate 1000 random seeds
2. For each seed:
    1. Generate many permutations
    2. Compute the *p*-value
3. Test whether these 1000 *p*-values are uniform between 0 and 1

For Mersenne Twister

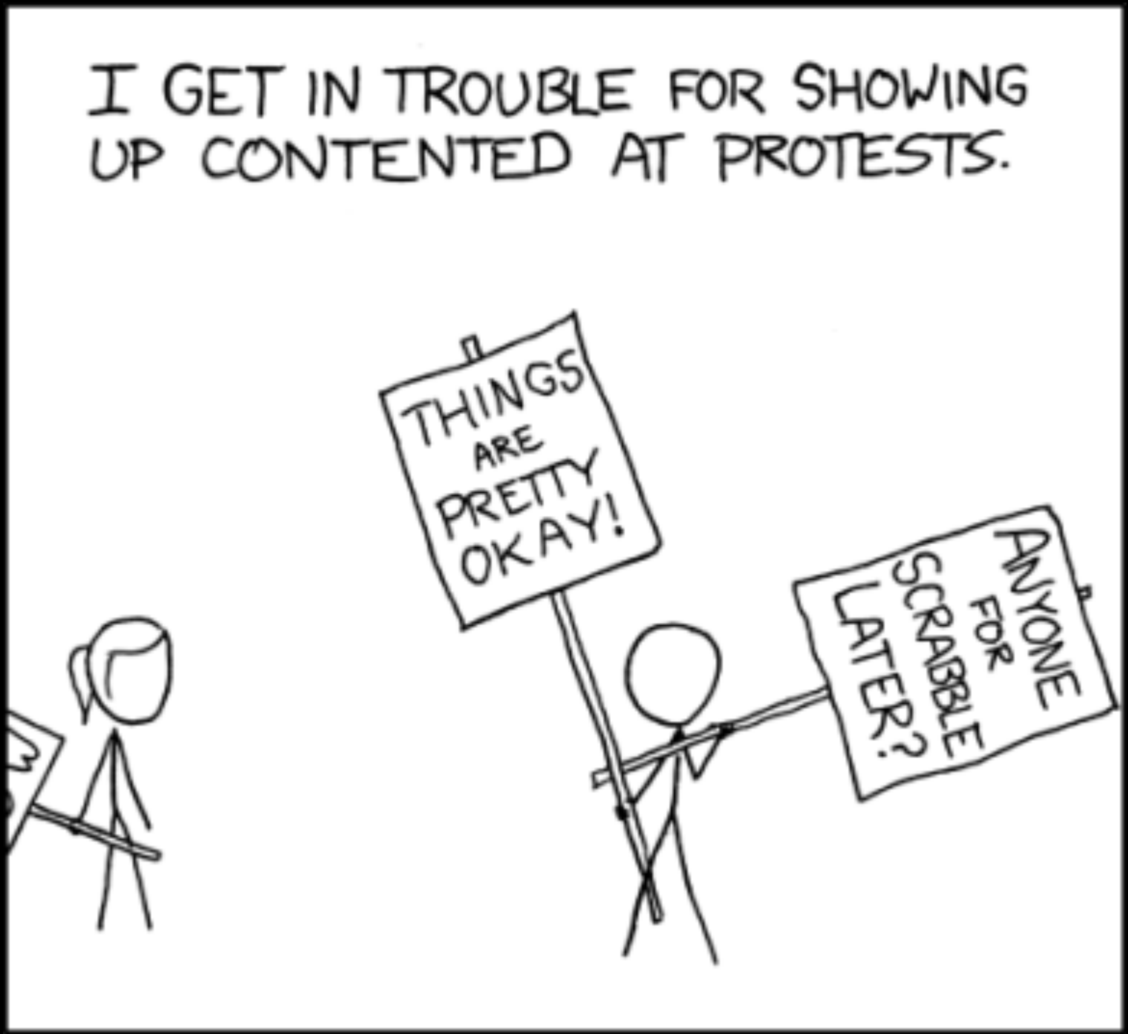


One solution: cryptographic hash functions.



- Designed to have good pseudorandom behavior
- Infinite state space when used in “counter mode”
- Some are designed for speed or even built into hardware

Myth #2: Student evaluations of teaching measure teaching effectiveness.



https://xkcd.com/470

The experiment:

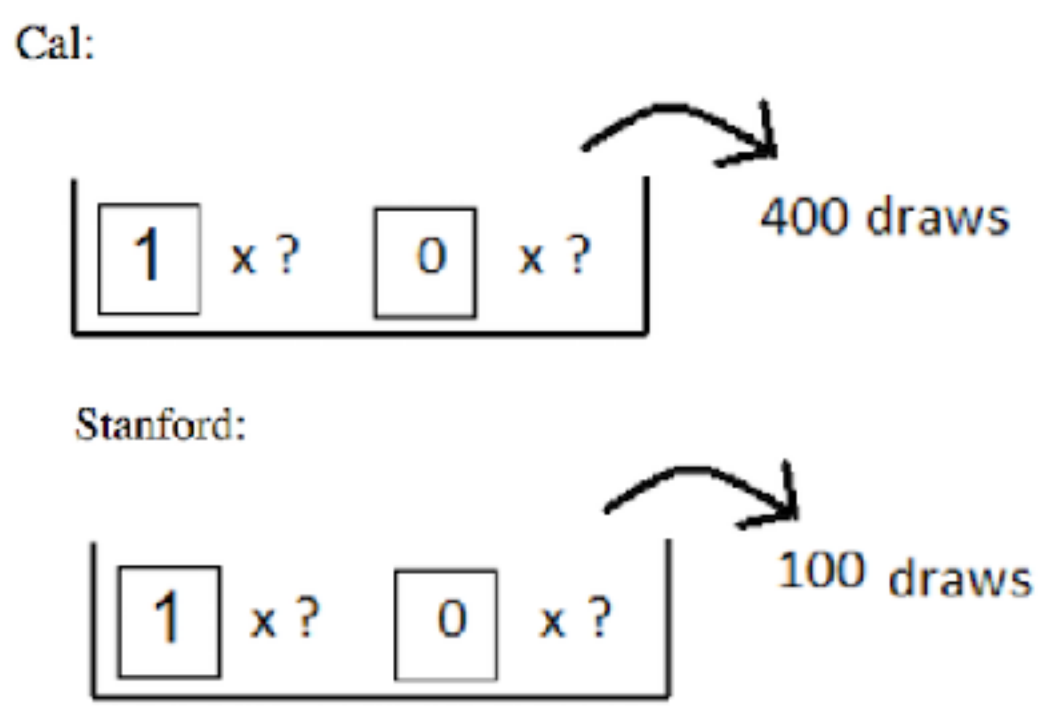
MacNell et al. (2015)

- Students were randomized into 4 online sections of a course
- In two sections, the instructors swapped identities
- Was the instructor who identified as female rated lower on average?

Compare assumptions…

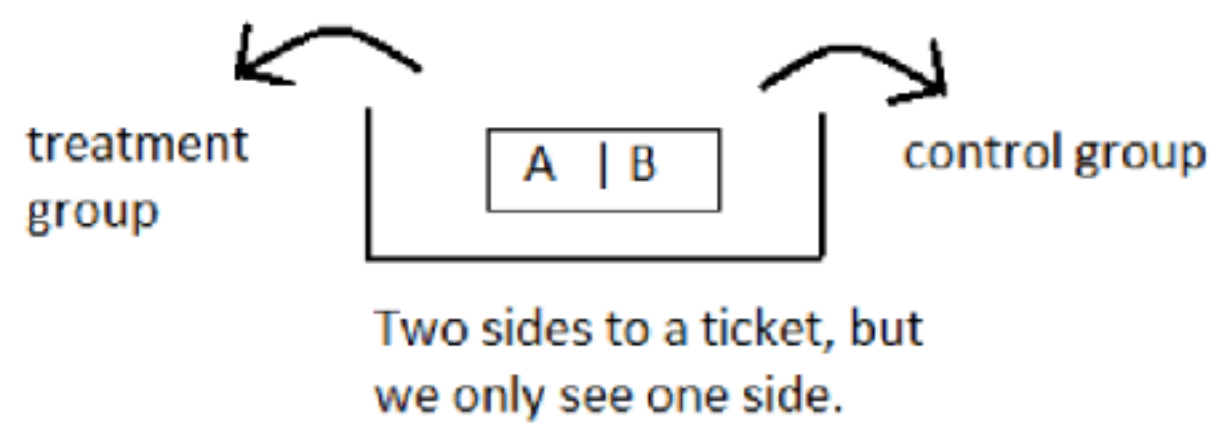
Two sample *t* test

- Samples are from two independent populations
- Data are normally distributed

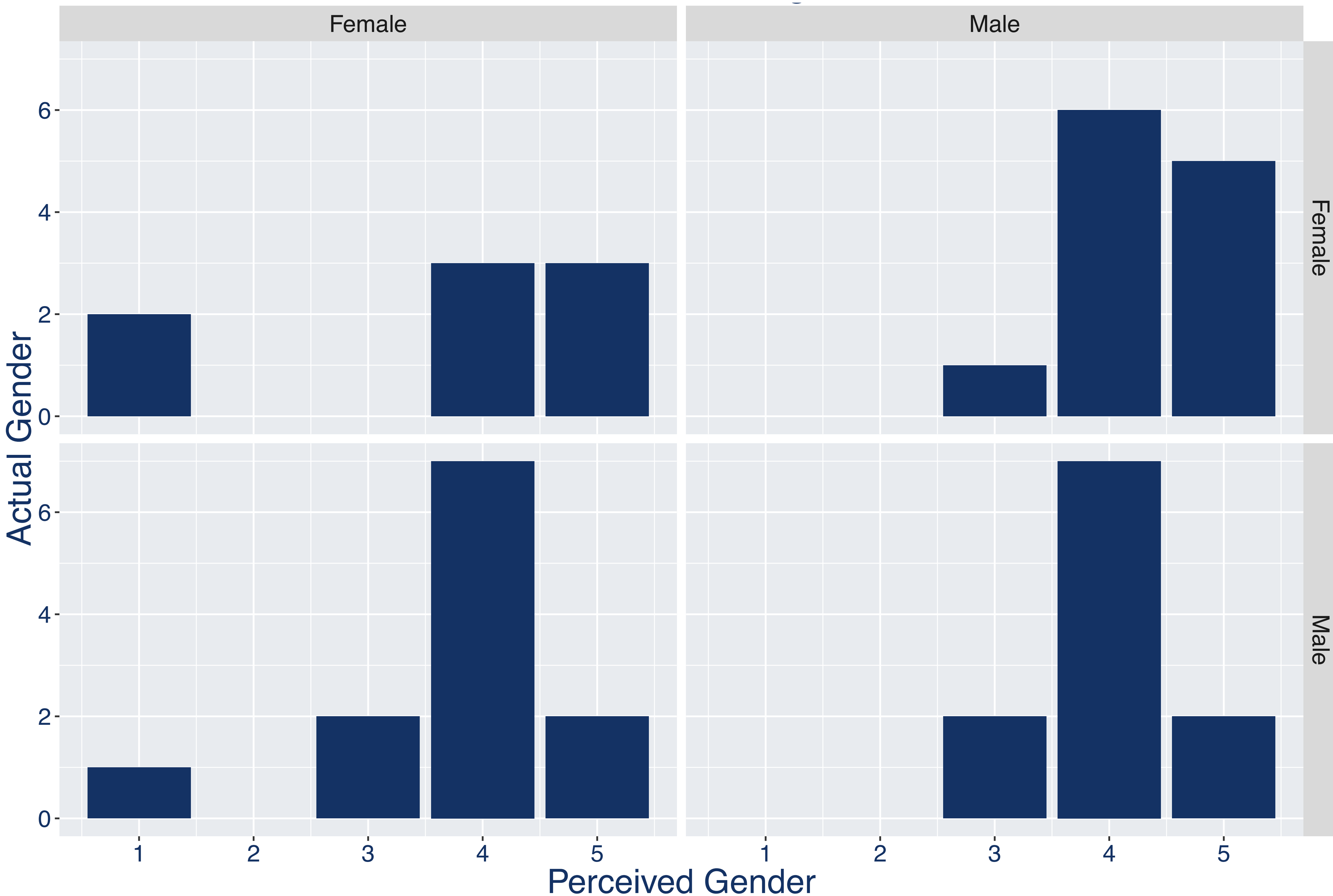


Permutation test

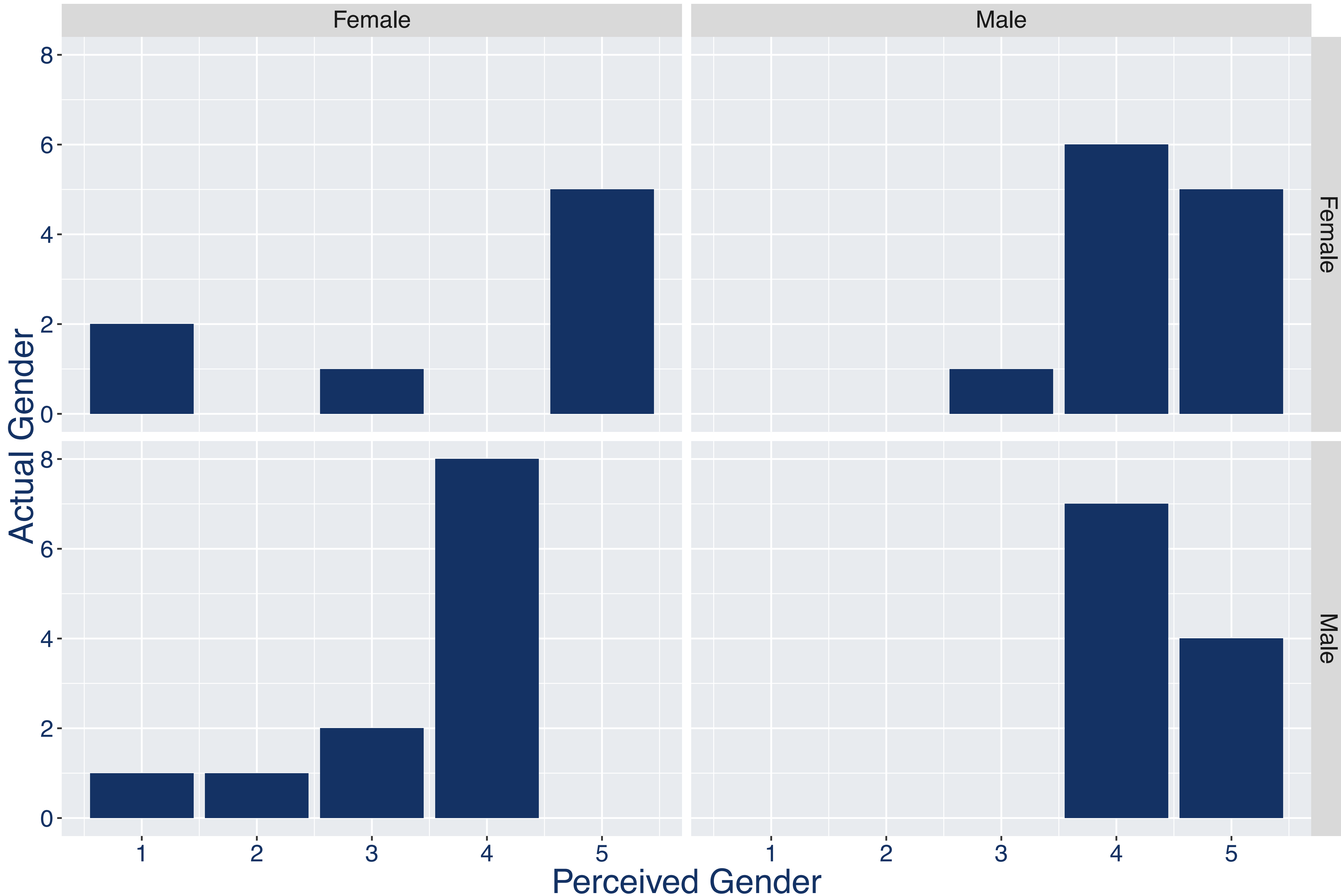
- Samples are from a single population
- Group membership is randomly assigned



The overall, main rating



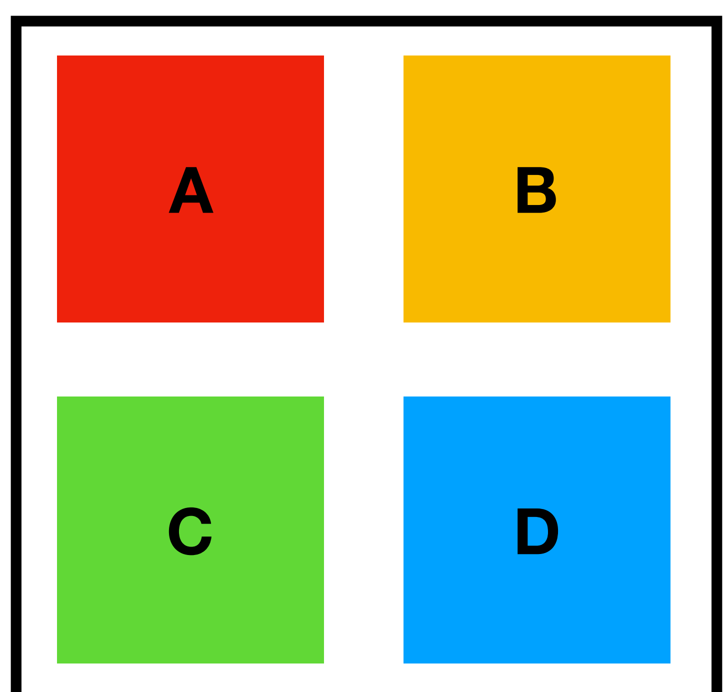
Objective things, like how promptly assignments were graded



Generalized model of potential outcomes

Numbers are fixed; randomization reveals one of the numbers.

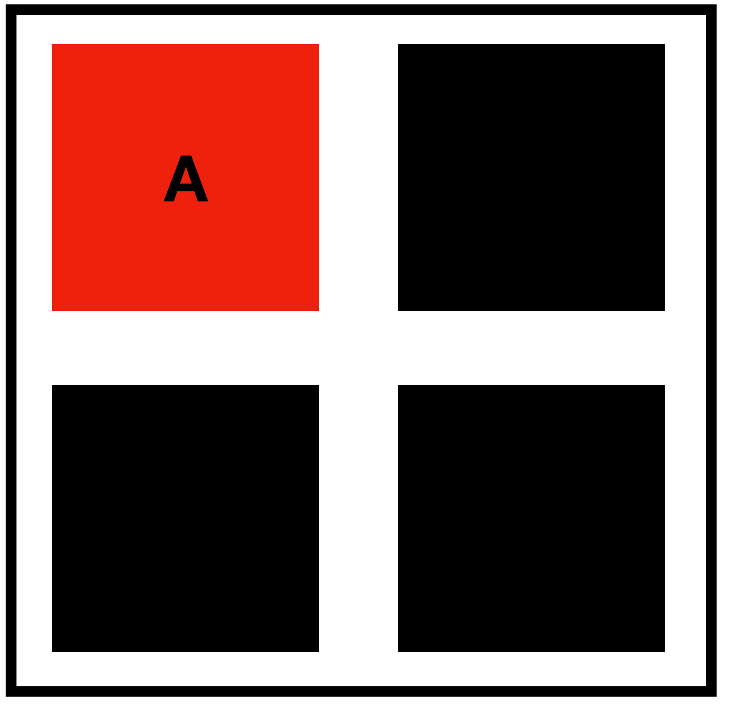
Assume non-interference: each student's response depends only on that student’s treatment.



Generalized model of potential outcomes

Numbers are fixed; randomization reveals one of the numbers.

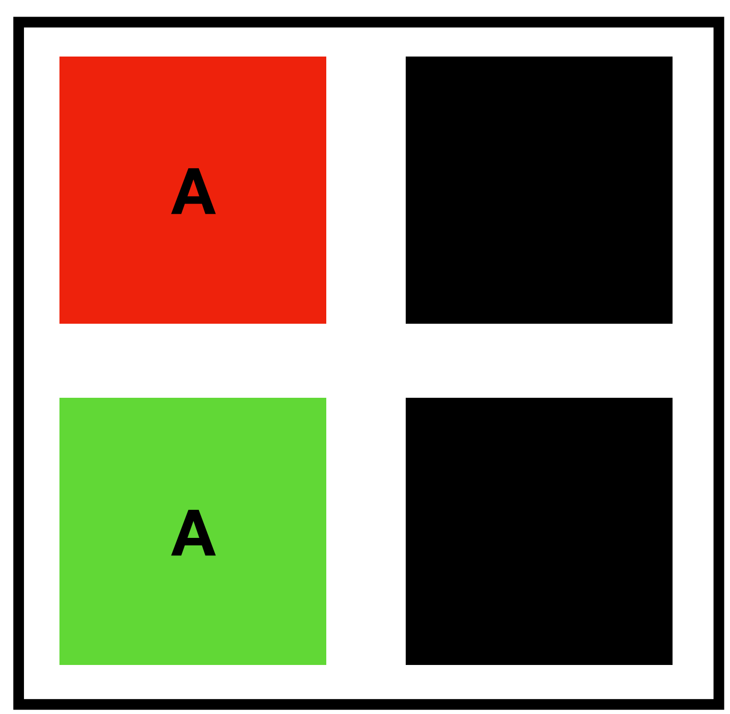
Assume non-interference: each student's response depends only on that student’s treatment.



Generalized model of potential outcomes

Numbers are fixed; randomization reveals one of the numbers.

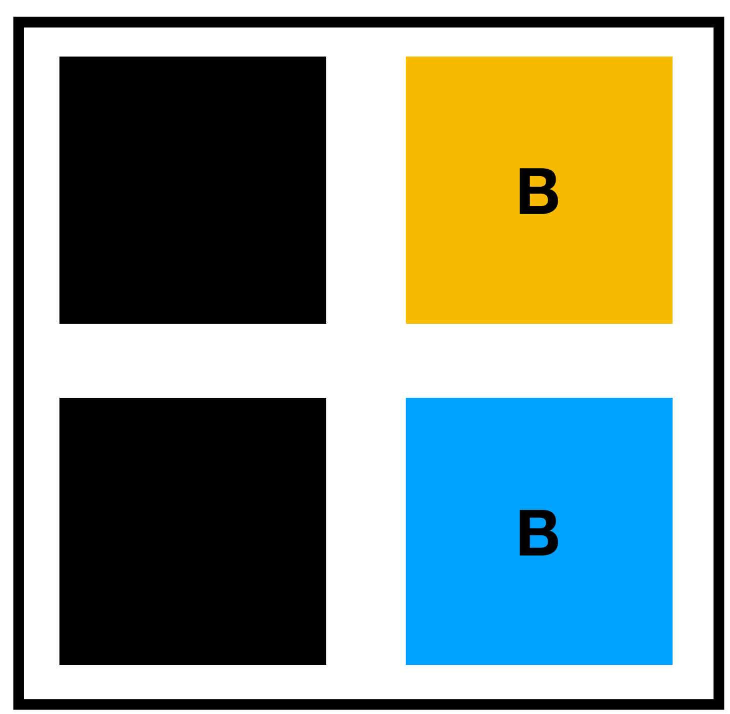
Assume non-interference: each student's response depends only on that student’s treatment.



Generalized model of potential outcomes

Numbers are fixed; randomization reveals one of the numbers.

Assume non-interference: each student's response depends only on that student’s treatment.



Permutation test

- Two conceptual levels of randomization: first instructor, then reported instructor gender

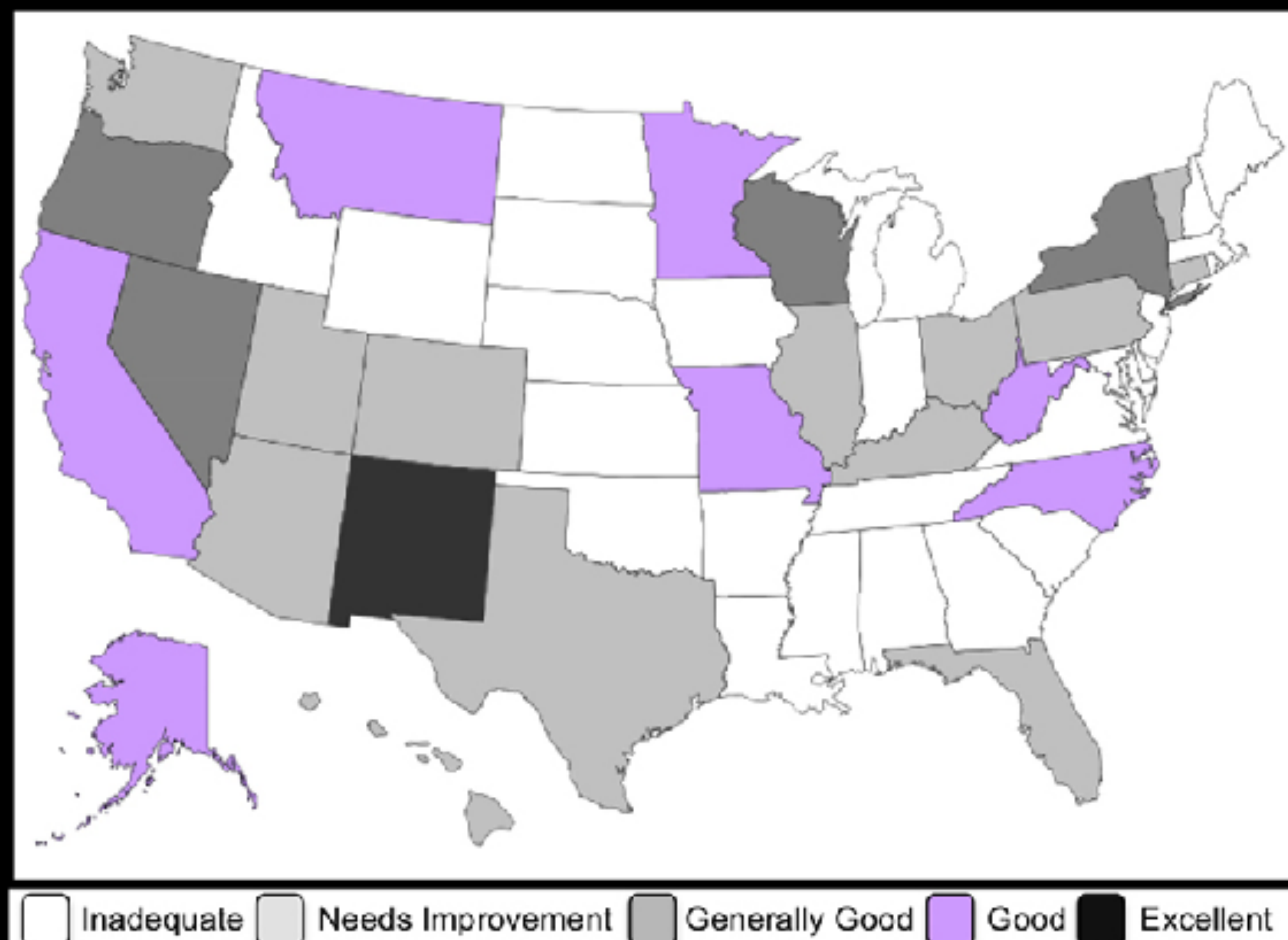
- All assignments of students that fix the number in each section are equally likely

- Stratified two-sample test: permute *perceived gender assignments* and measure difference in mean ratings by perceived gender

In all categories, the male-identified instructor was rated higher.

Characteristic	M - F	perm *P*-value	t-test *P*-value
Overall	0.47	0.12	0.128
Caring	0.52	0.10	0.071
Consistent	0.47	0.21	0.045
Enthusiastic	0.57	0.06	0.112
Fair	0.76	0.01	0.188
Feedback	0.47	0.16	0.054
Helpful	0.46	0.17	0.049
Knowledgeable	0.35	0.29	0.038
Praise	0.67	0.01	0.153
Professional	0.61	0.07	0.124
Prompt	0.80	0.01	0.191
Respectful	0.61	0.06	0.124
Responsive	0.22	0.48	0.013

Myth #3: election integrity



https://www.verifiedvoting.org/resources/post-election-audits/

What’s an audit?

- Convince everyone that the outcome was decided correctly
- **Compliance audit:** were election procedures followed properly?
- **Materiality audit:** were errors introduced despite compliance?

Auditing framed as a hypothesis test

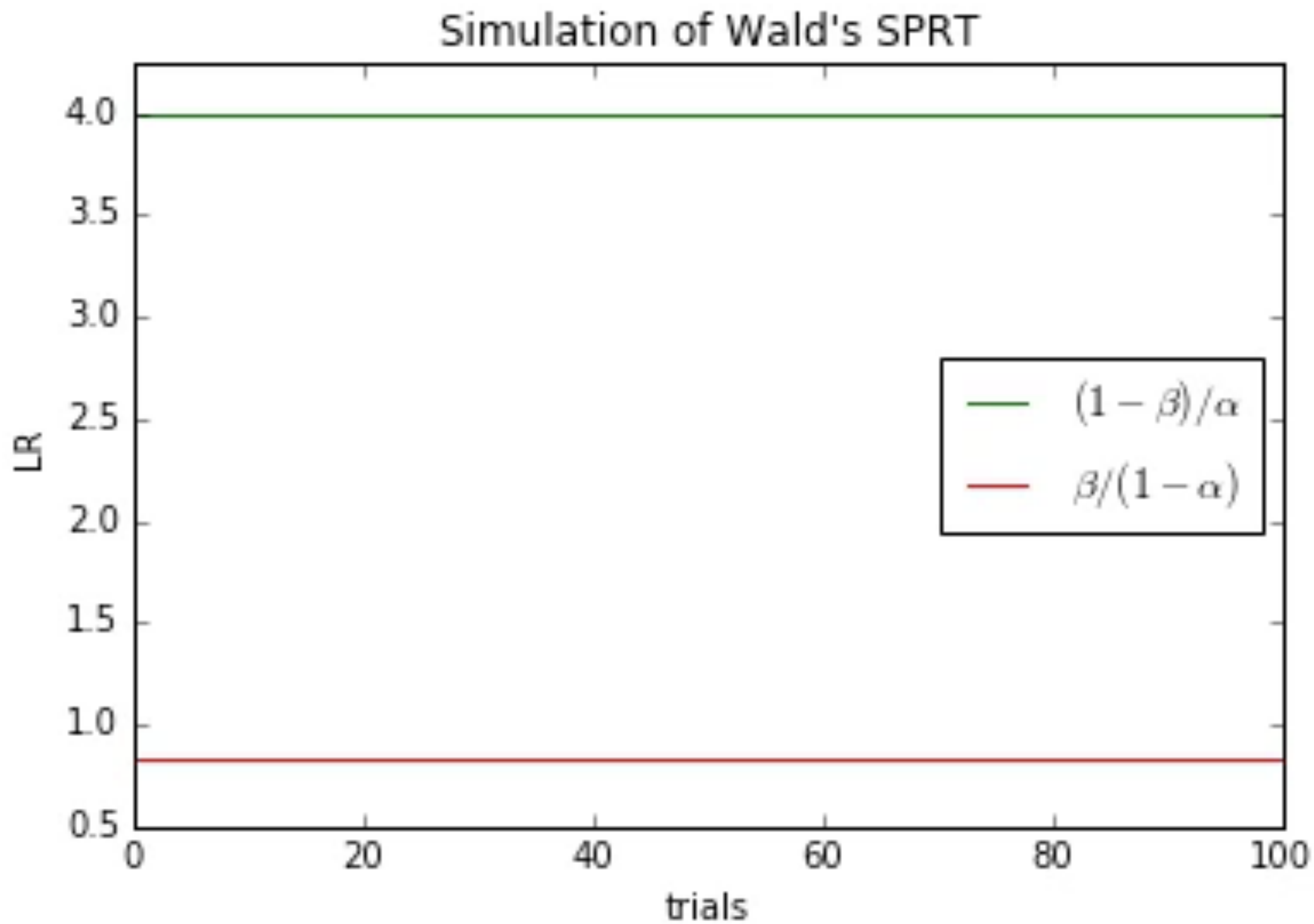
Null hypothesis: the reported winner received **fewer votes** than the runner-up

Alternative hypothesis: the reported winner received **more votes** than the runner-up

We **certify** the election when the null is rejected. Either:

- the correct winner was named, or
- something very unlikely happened

Sequential testing



The hypothesis test depends on the type of voting machine.

- **Ballot polling:** check that vote shares in a sample of paper ballots match the reported outcomes
- **Ballot comparison:** check for errors at the ballot level by comparing paper ballot to electronic record

Take-aways

- *P*-values can be a useful tool when interpreted correctly
- The null model must match the way the data were generated