# Stat 215A
# Hypothesis Testing in the Wild

Kellie Ottoboni

Department of Statistics, UC Berkeley
Berkeley Institute for Data Science

October 11, 2016

# Outline

# Recommendations

- Look at your data – plots, summary stats, etc.
- What's the alternative? Think about the science
- Make sure the null distribution matches how the data actually arose

# Outline

# Teaching Evaluations

Student evaluations of teachers (SET) are used to

- Quantify teaching effectiveness
- Compare instructors across courses
- Make hiring, firing, and promotion decisions

Are SET a valid measure of teaching effectiveness?

In Boring et al. (2016), we reanalyzed two datasets:

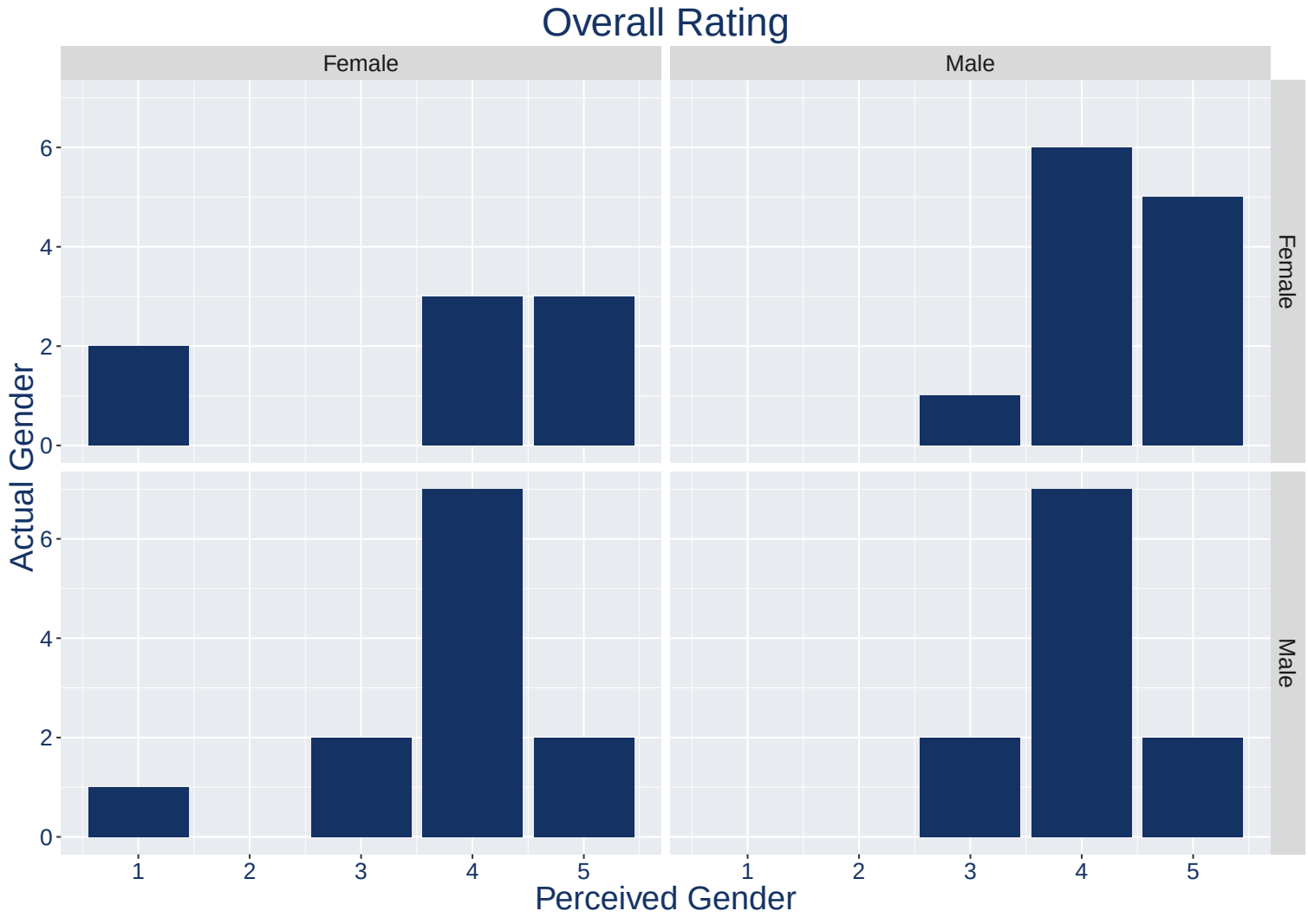
- an online, randomized experiment from North Carolina State University
- a natural experiment at Sciences Po, Paris

# Teaching evaluations

The US data: MacNell et al. (2015)

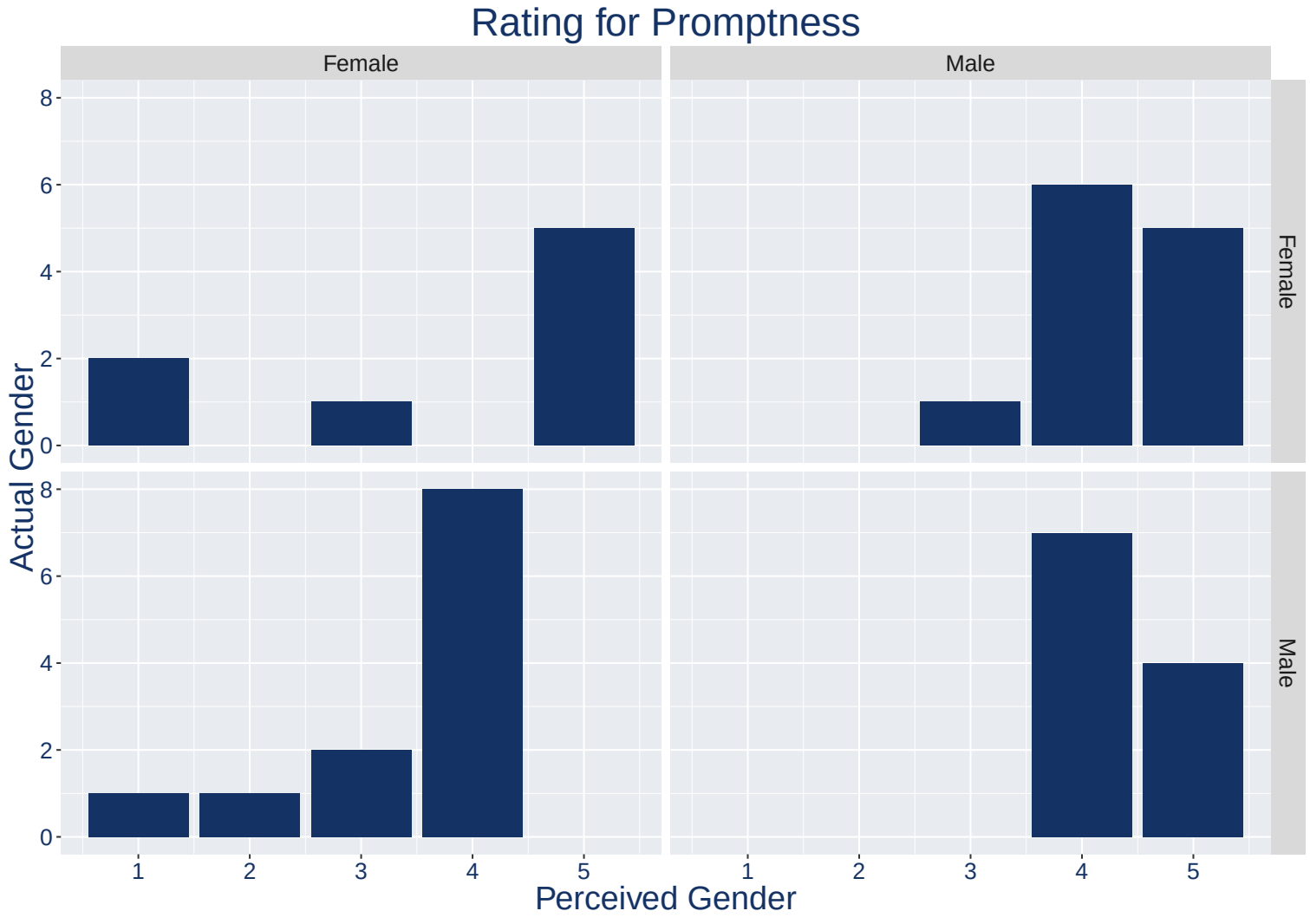
- Students were randomized to 4 online sections of a course.
- In two sections, the instructors swapped identities.
- Was the instructor who identified as female rated lower on average?

# Teaching evaluations




The overall, main rating

# Teaching evaluations




Objective things like how promptly assignments were graded...

# Neyman-Rubin model, generalized

Student $i$ is represented by a ticket with $4$ numbers, their response to each "treatment."

$$r_{ijk} = \text{ SET given by student } i \text{ to instructor } j \text{ when they appear to have gender } k$$

$$i = 1, \ldots, N; \qquad j = 1, 2; \qquad k \in \{\text{male, female}\}$$

Numbers are fixed; randomization reveals one of the numbers.

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

If gender doesn't matter,

$$r_{ij\text{male}} = r_{ij\text{female}}.$$

# Randomization

Conceptually, there are two levels of randomization:

1. $N_m$ students are randomly assigned to the male instructor, and the remaining $N_f$ get the female instructor.
2. Of the $N_j$ assigned to instructor $j$, $N_{jm}$ are told that the instructor is male, for $j = 1, 2$.

All $\binom{N_m}{N_{mm}} \times \binom{N_f}{N_{fm}}$ assignments of students to sections are equally likely.

## Stratified two-sample test

- For each instructor, permute perceived gender assignments
- Use difference in mean ratings for female-identified minus male-identified

# Results

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

| **Characteristic** | **M-F** | **perm** $P$ | **t-test** $P$ |
|---:|---|---:|---:|
| 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 |

# Teaching evaluations

The French data: Boring (2015)

- 23,001 SET of 379 instructors by 4,423 students between 2008 and 2013
- All first-year students take the same six courses
- Students enroll in "triads," so they cannot select individual instructors
- Section instructors assign interim grades during the semester – a proxy for their expected grade
- All students take the same final exam, written and graded by another professor – a proxy for amount learned
- SET are mandatory: response rates are nearly 100%

# Randomization

- The hypothetical randomization holds triads fixed, to allow for cohort effects
- The test is conditional on which students signed up for which triad
- The unconditional level of the test is controlled at level $\alpha$:

$$\begin{aligned}
\mathbb{P}\{\text{Type I error}\} &= \sum_{\text{triads}} \mathbb{P}\{\text{Type I error}|\text{triads}\}\mathbb{P}\{\text{triads}\} \\
&\leq \sum_{\text{triads}} \alpha\mathbb{P}\{\text{triads}\} \\
&= \alpha \sum_{\text{triads}} \mathbb{P}\{\text{triads}\} \\
&= \alpha
\end{aligned}$$

# Randomization

Test the null hypothesis that there is no correlation between gender and average SET from a triad

- It's as though instructors are assigned at random. Gender is arbitrary.
- Things are *not* exchangeable across years or across courses!

## Stratified test of correlation

- Independently for each course and each year, permute average SET
- For each permutation, compute the Pearson correlation between average SET and instructor gender*
- Average the correlations across years (and courses, if considering overall)

* This is permutationally equivalent to difference in means

# Results

Table : Average correlation between SET and instructor gender

| | $\bar{\rho}$ | $p$-value |
|---|---|---|
| Overall | 0.09 | 0.00 |
| History | 0.11 | 0.08 |
| Political institutions | 0.11 | 0.10 |
| Macroeconomics | 0.10 | 0.16 |
| Microeconomics | 0.09 | 0.16 |
| Political science | 0.04 | 0.63 |
| Sociology | 0.08 | 0.34 |

*Note: $p$-values are two-sided. Positive values of $\bar{\rho}$ indicate that students of male instructors received higher average SET than female instructors.*

# Results

But maybe female instructors didn't teach as well?

Table : Average correlation between final exam scores and instructor gender

| | $\bar{\rho}$ | $p$-value |
|---|---|---|
| Overall | -0.06 | 0.07 |
| History | -0.08 | 0.22 |
| Macroeconomics | -0.06 | 0.37 |
| Microeconomics | -0.06 | 0.37 |
| Political science | -0.03 | 0.70 |
| Sociology | -0.05 | 0.55 |

*Note: $p$-values are two-sided. Negative values of $\bar{\rho}$ indicate that students of female instructors did better on average than students of male instructors.*

# Effect sizes and confidence intervals

From one of the reviewers:

> *"The article does not discuss effect sizes and considerations around practical significance. This is especially important for the French dataset which has a large number of observations. Indeed in Section 6 the authors discuss the importance of finding small p-values with the small US dataset, however a similar discussion is not included for the French dataset."*

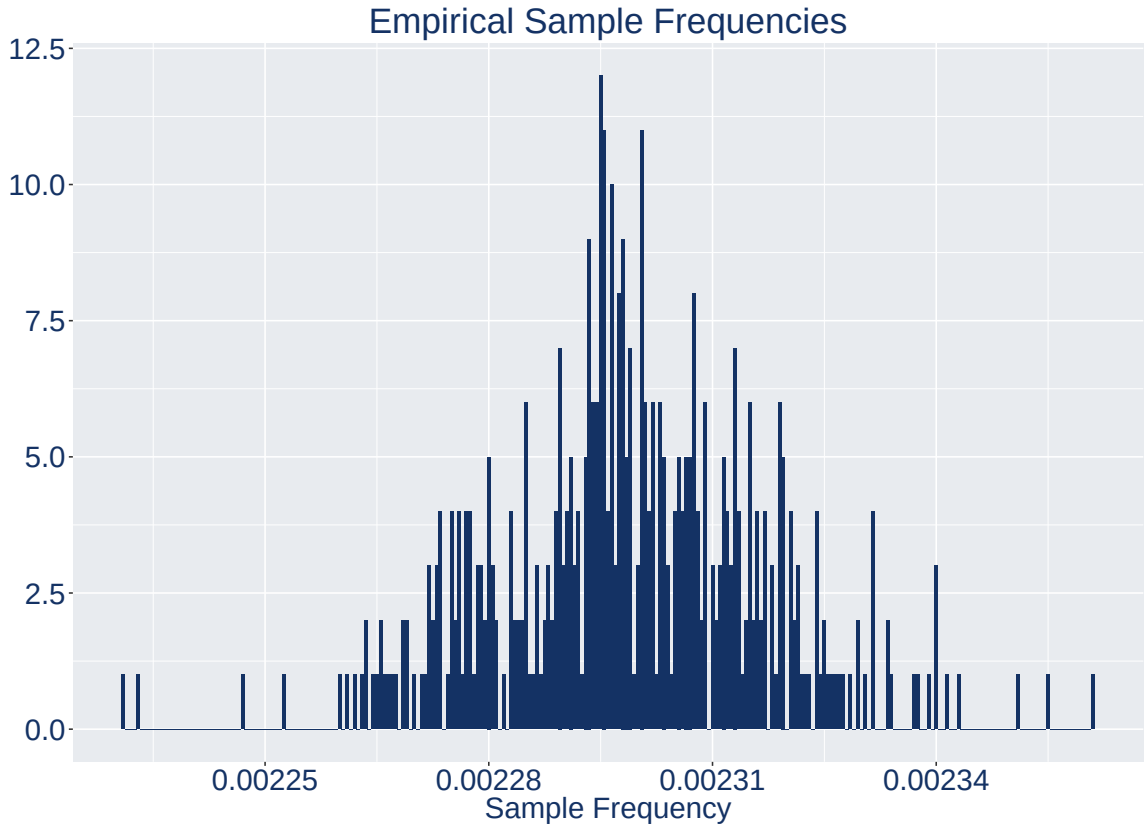
- Inverting hypothesis tests depends on the alternative
- For permutation tests, it requires strong parametric assumptions, e.g. constant additive treatment effect

# Outline

# PRNGs

- For moderate sample and population sizes, common PRNGs cannot generate all possible simple random samples (SRSs): e.g. RANDU can't do $\binom{40}{10}$
- In principle, Mersenne Twister and other PRNGs with a large state space can generate all SRSs
- Do samples have equal probability?

# PRNGs




- Take $10^7$ SRSs of size 2 from a population of 30, using RANDU with PIKK
- Departure from uniformity is less obvious with better PRNGs – do a hypothesis test

# Set-up

$$H_0 : \mathbb{P}(\text{sample}_i) = \frac{1}{N} \text{ for all samples } i = 1, \ldots, N$$

$$H_1 : \mathbb{P}(\text{sample}_i) \neq \frac{1}{N} \text{ for some sample}$$

- Test by generating a large "sample" of $B$ SRSs
- Under $H_0$, the number of times each SRS is observed follows a multinomial distribution with $B$ trials and equal selection probabilities $1/N$

# Proposals

1. Chi-squared test:

$$X^2 = \sum_{i=1}^{N} \frac{(O_i - E_i)^2}{E_i}$$

where $E_i = \frac{B}{N}$ is the expected number and $O_i$ is the observed number of sample $i$. Under $H_0$, $X^2 \sim \chi^2_{N-1}$.

2. Range statistic test:

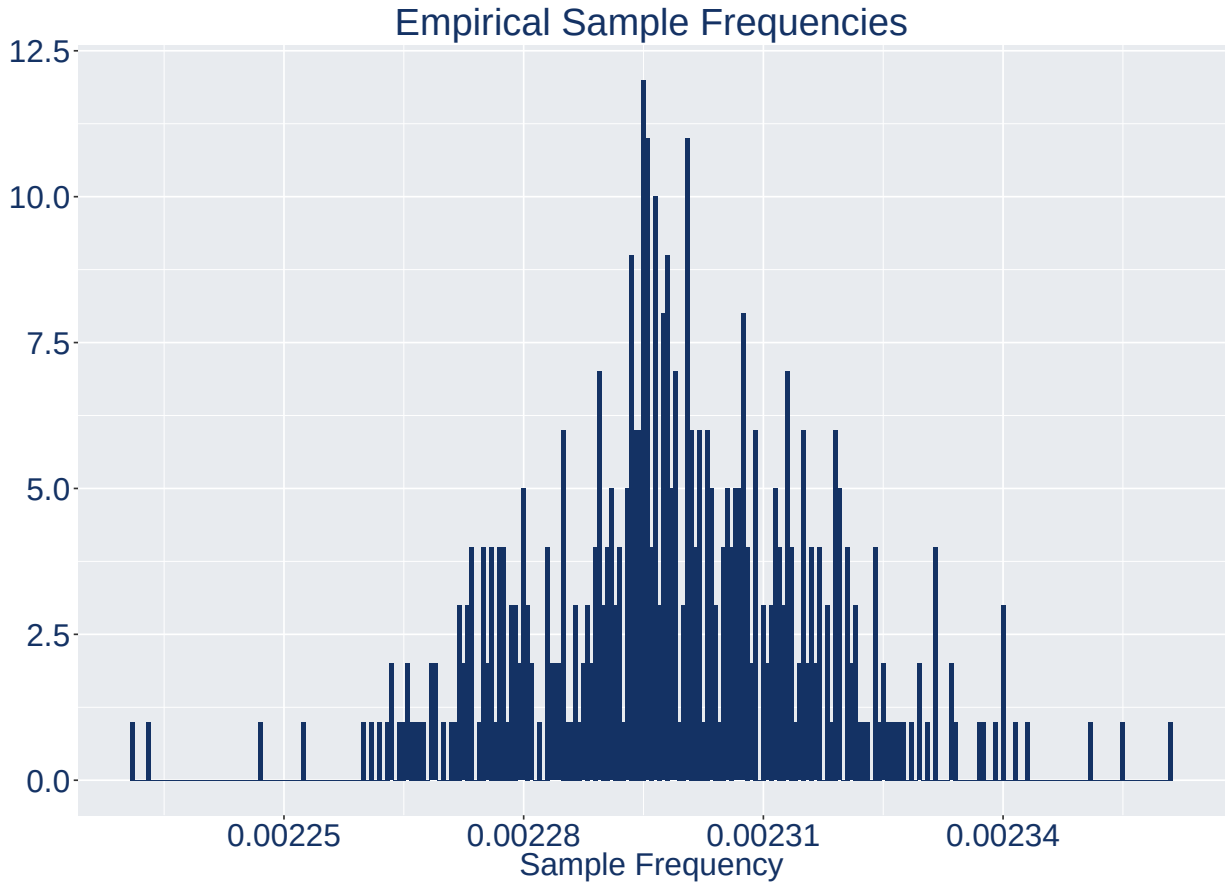
$$R = \max_i O_i - \min_i O_i$$

$R$ has a complicated distribution... use asymptotic approximation from Young (1962):

$$\mathbb{P}(R \leq r) \approx P(W_N \leq (r - (2B)^{-1})(N/B)^{1/2})$$

where $W_N$ denotes the sample range of $N$ independent standard normal random variables.

# What's the alternative?




We want to be sensitive to a few large deviations from $1/N$.

# Power

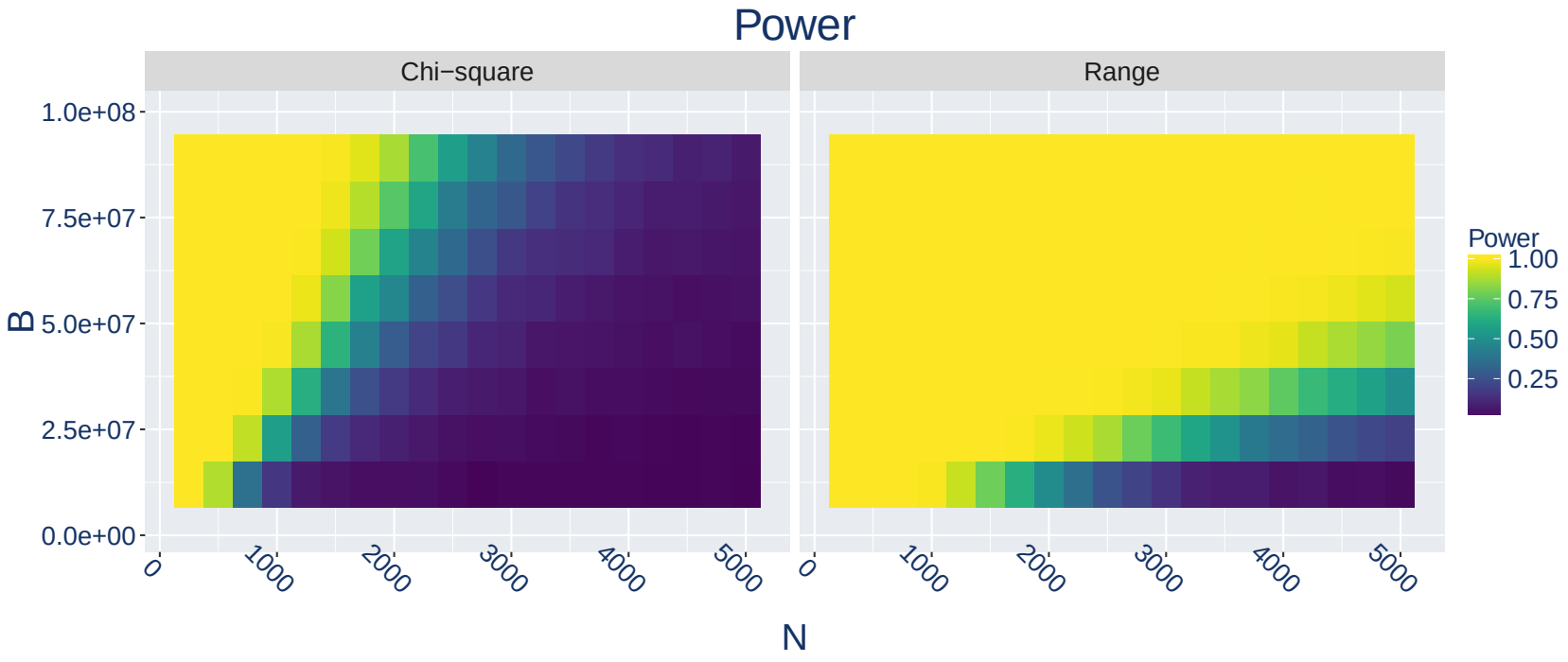
Estimate the power for different $B$ and $N$ under the following alternative hypothesis:

- All samples but two have probability $1/N$
- One sample has probability $0.95/N$, the other has probability $1.05/N$
- Reject the null hypothesis at level $1\%$
- Power $= \mathbb{P}(\text{Reject null at level } 1\% \text{ under this alternative})$

Two ways to do power calculations:

1. Simulation
2. Analytically

# Simulations




For large values of $N$, you need fewer samples (smaller $B$) to achieve high power.

Choose your statistics wisely!