INFERENCE, REGRESSION, & STORIES

MPA 630: Data Science for Public Management

December 6, 2018

Fill out your reading report on Learning Suite

PLAN FOR TODAY

Why all this simulation?

Inference and regression

Telling stories with data

WHY ALL THIS SIMULATION?

Find δ

The sample statistic: diff in means, mean, diff in props, etc.

Invent world where δ is null

Simulate what the world would look like if there was no effect.

Look at δ in the null world

Is it big and extraordinary, or is it a normal thing?

Calculate probability that δ could exist in the null world

This is your p-value!

Decide!

GOAL OF STATISTICS

Make inferences about a whole population using only a sample of the population

MAKING INFERENCES

We have to know how confident we are that the thing we find in the sample reflects the whole population

Confidence intervals

p-values

P-VALUES

The probability of observing an difference or effect at least that large when no difference or effect exists

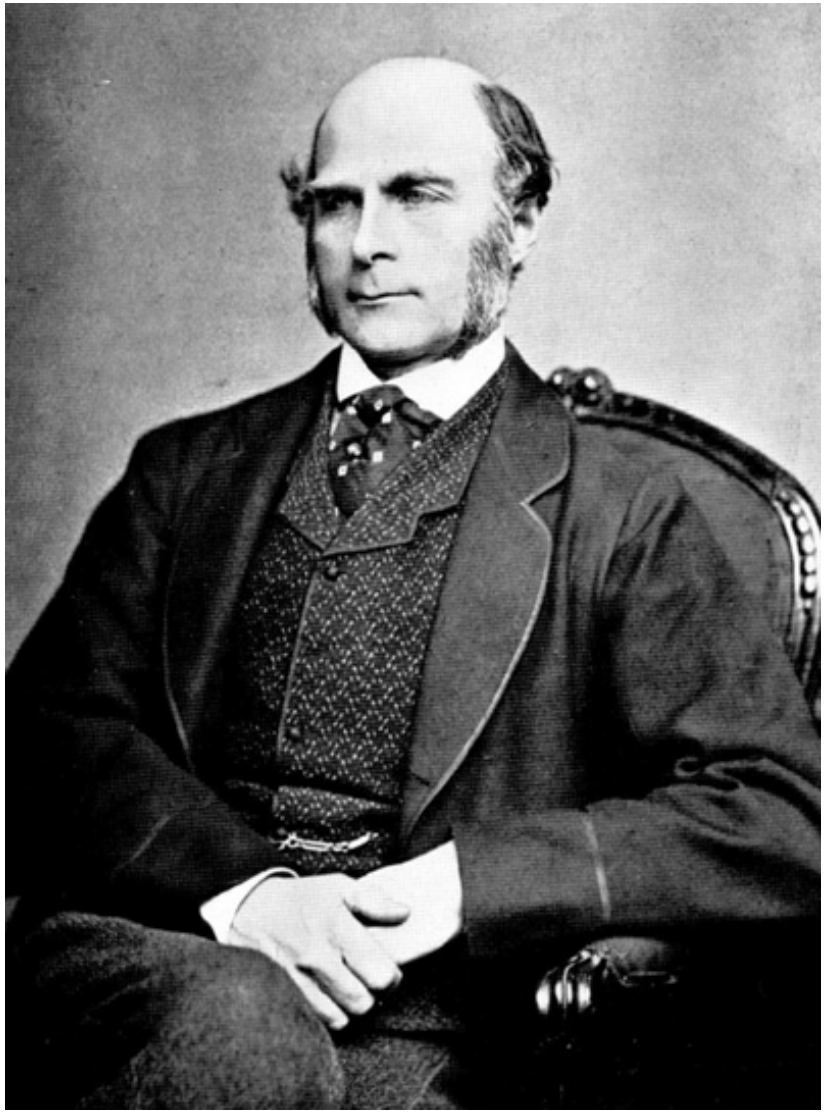
NULL WORLDS

But how do you determine what δ would look like in a null world?

Option 1: Use math and calculus and formulas to determine probabilities

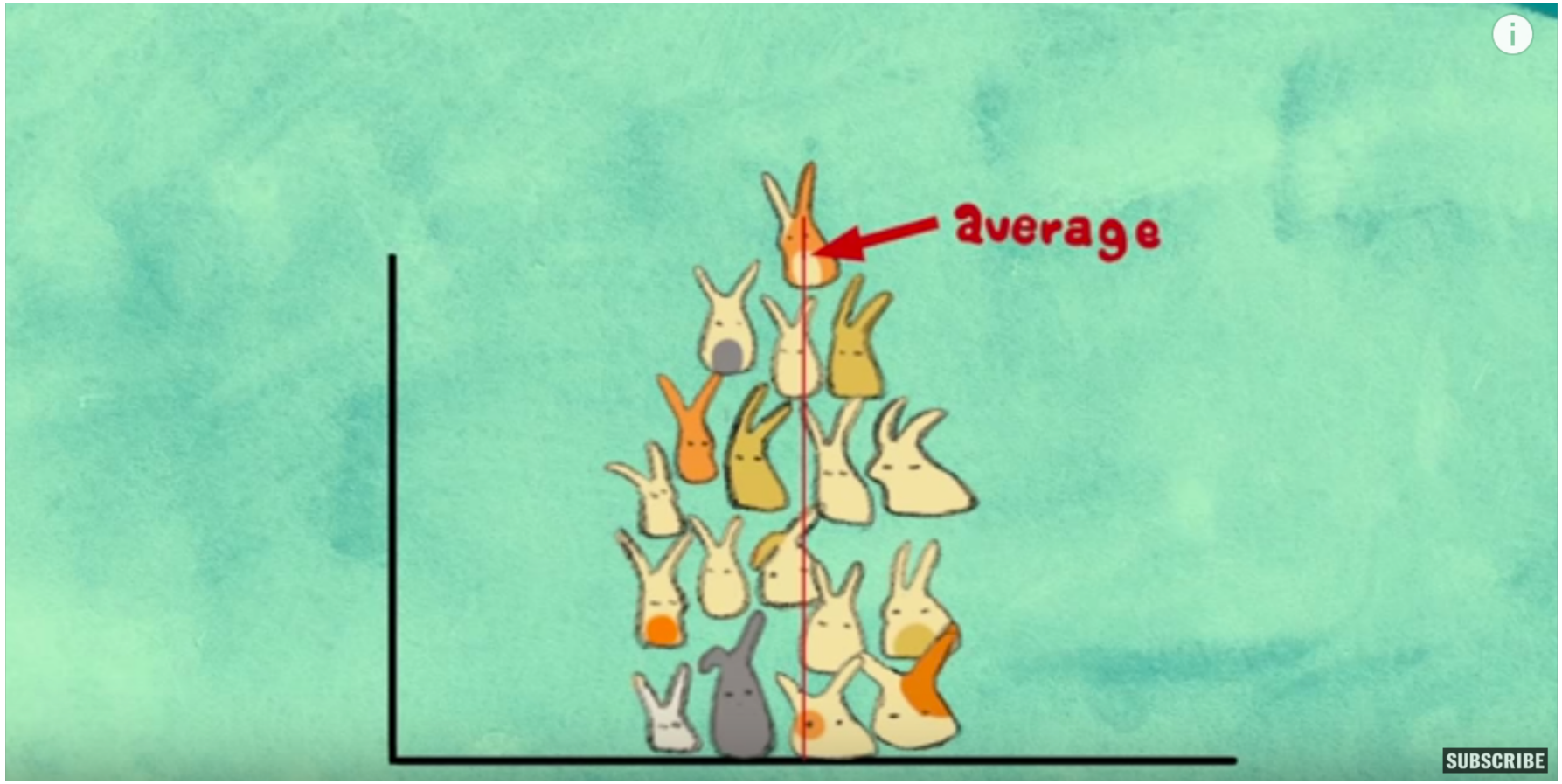
Carl Friedrich Gauss

Francis Galton



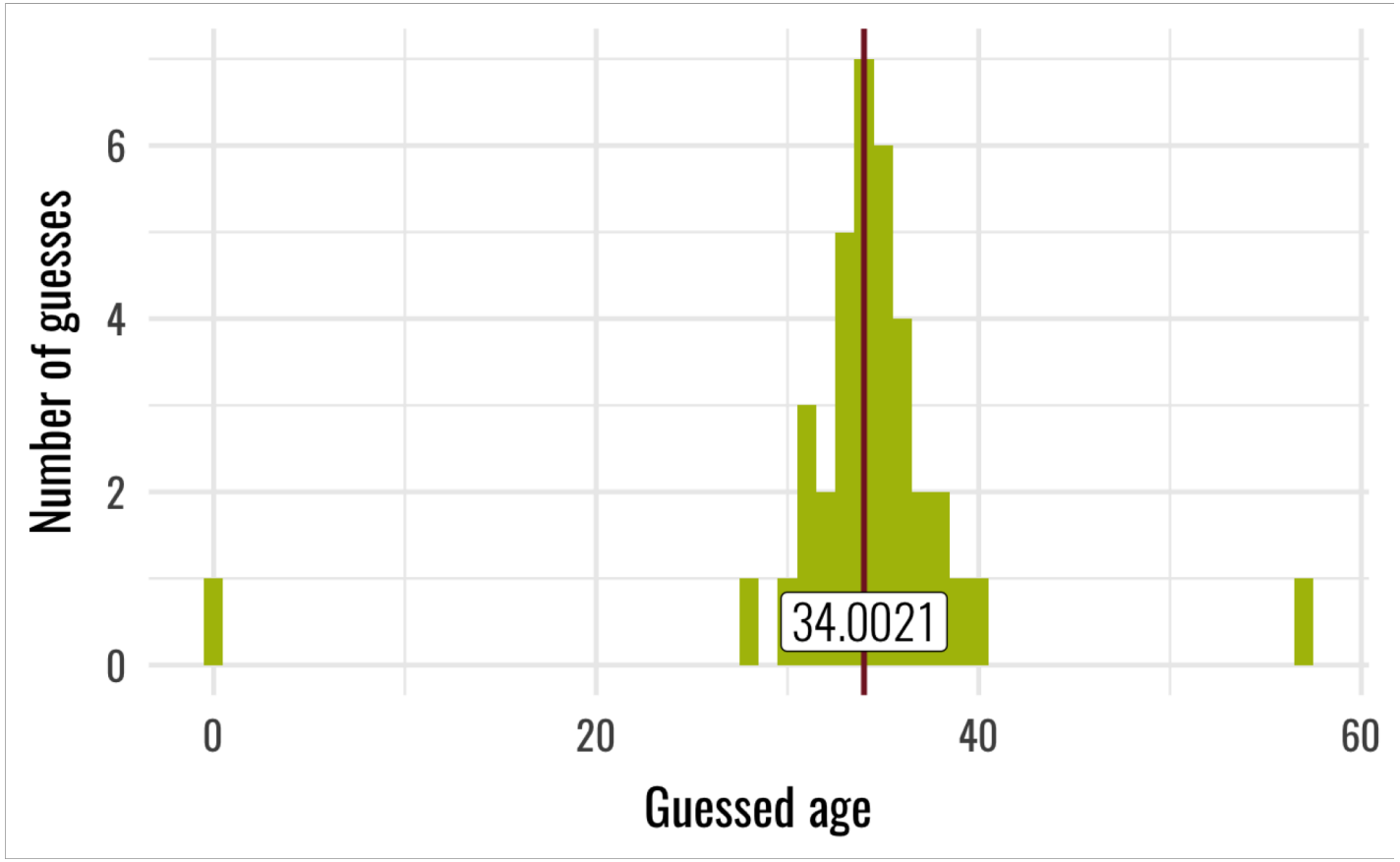
CENTRAL LIMIT THEOREM

https://www.youtube.com/watch?v=jvoxEYmQHNM



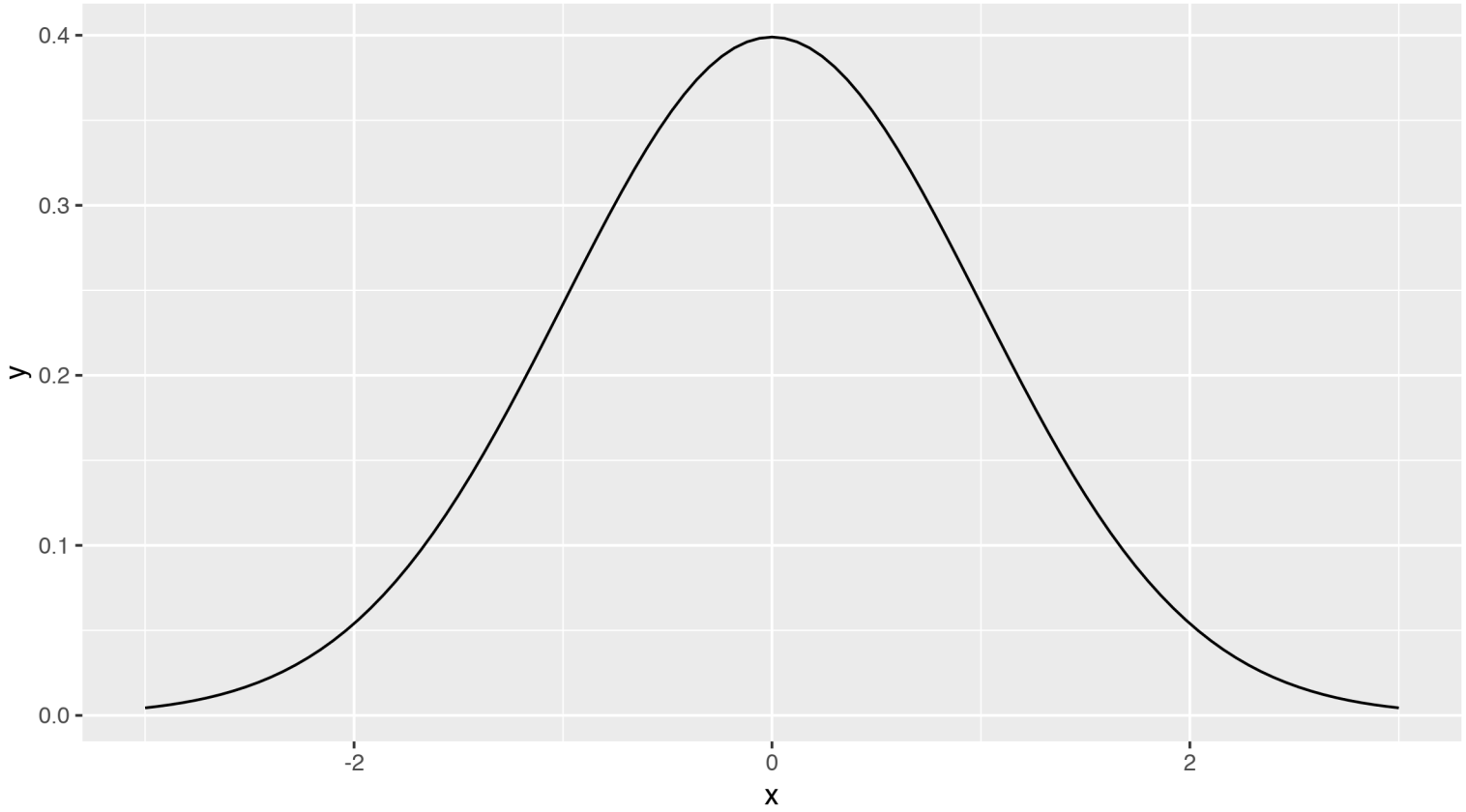
WISDOM OF THE CROWD

	guess
1	35.000
2	35.476
3	36.600
4	35.000
5	34.000
6	36.000
7	32.000
8	38.000
9	36.000
10	34.000
11	35.000
12	38.000
13	31.000



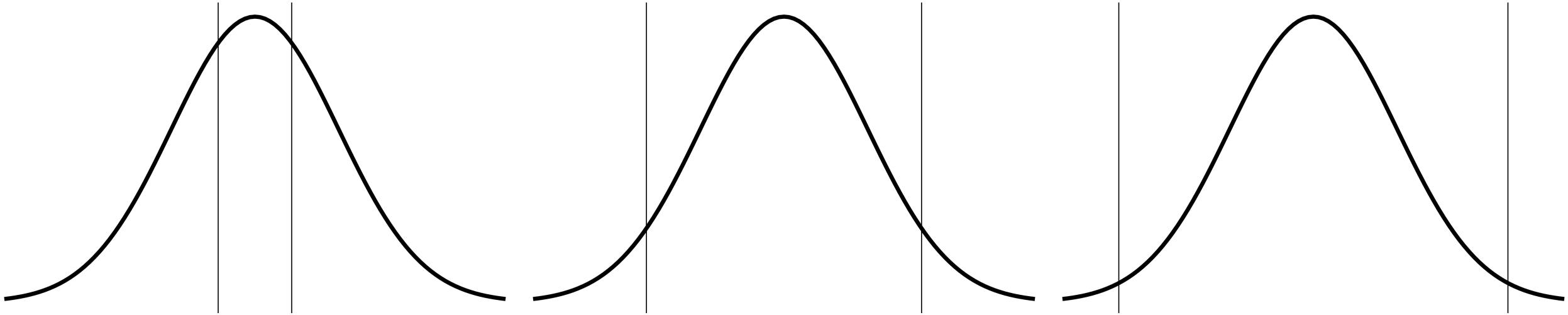
$$\frac{12,405}{365.25} \parallel 33.963$$

NORMAL DISTRIBUTION



$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

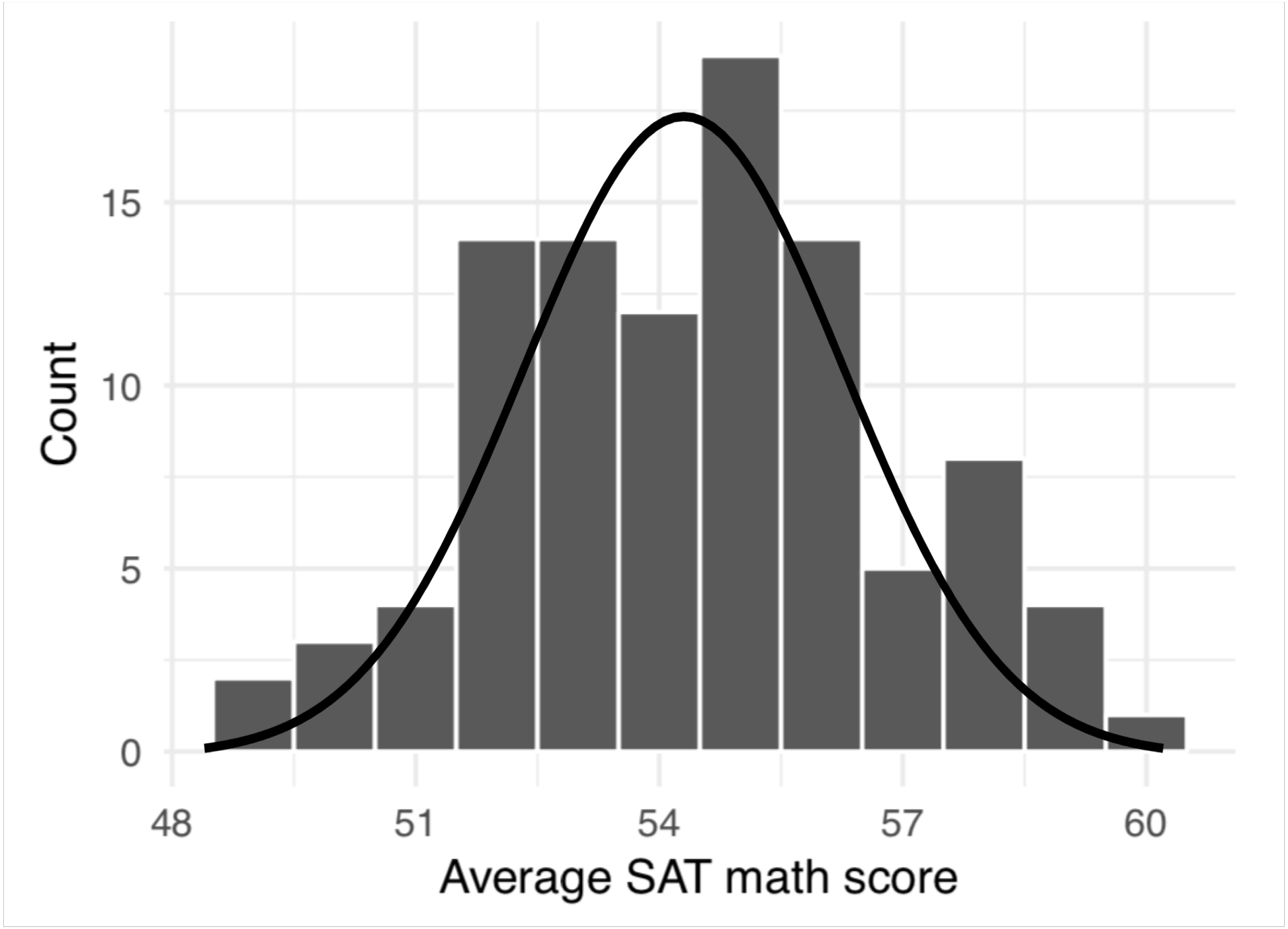
Z-SCORES AND AREAS



1 SD
67%

2 SD
95%

3 SD
99%


Count
15
10
5
0
48
51
54
57
60
Average SAT math score

IDEAL SAMPLE SIZES

How many apartments should we sample to generate a 95% confidence interval where we're ±$50 around the mean?

(Assume the population standard deviation is $200)

$$\text{Margin of error} = Z \times \frac{\sigma}{\sqrt{n}}$$

$$50 = 1.96 \times \frac{200}{\sqrt{n}}$$

$$n = 61.4655$$

SAMPLE SIZES IN REAL LIFE

Search Google for “sample size calculator”

https://www.qualtrics.com/blog/calculating-sample-size/

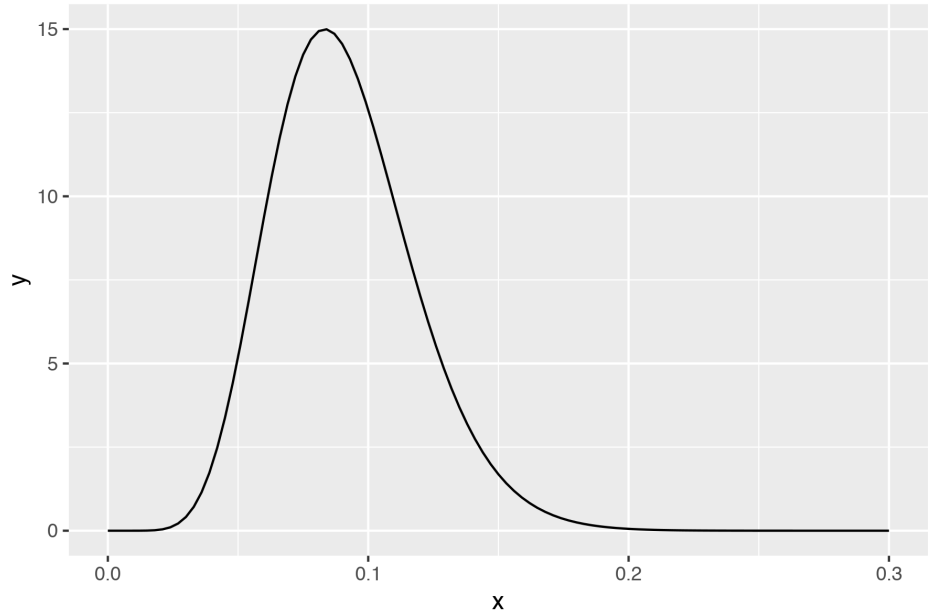
NOT EVERYTHING IS NORMAL

What if the data doesn’t fit a normal distribution?

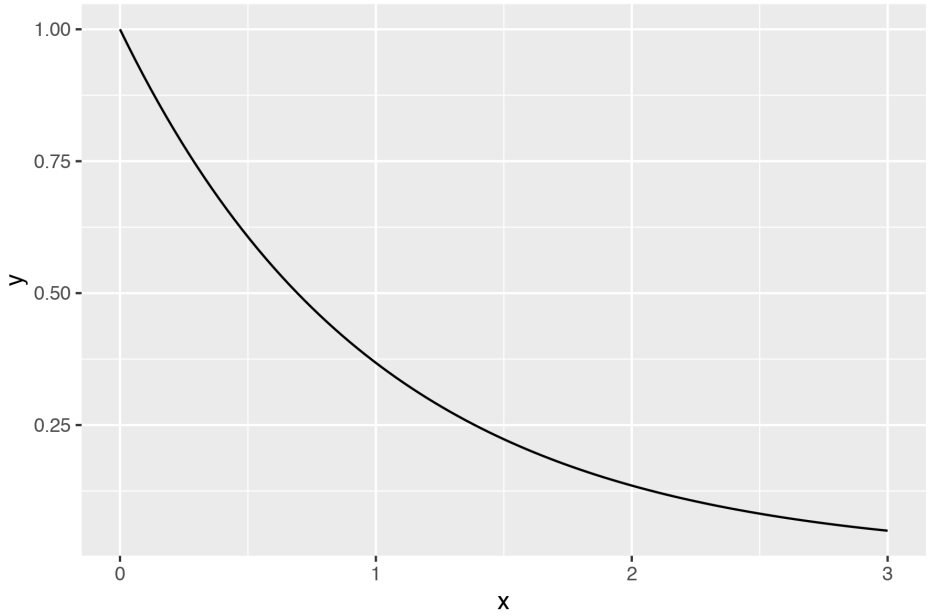
Use a different distribution!

Beta distribution

shape1 = 10; shape2 = 100

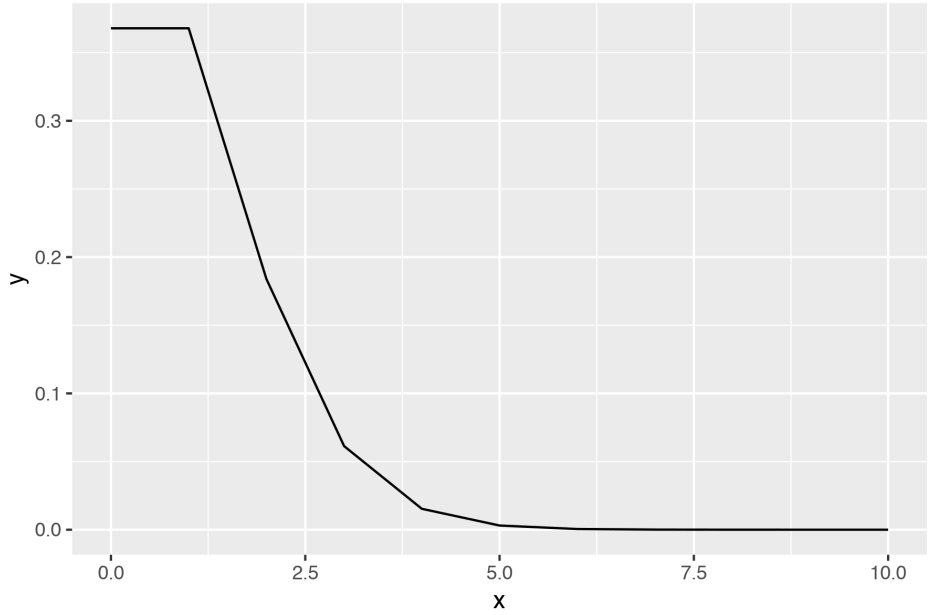


Expontential distribution



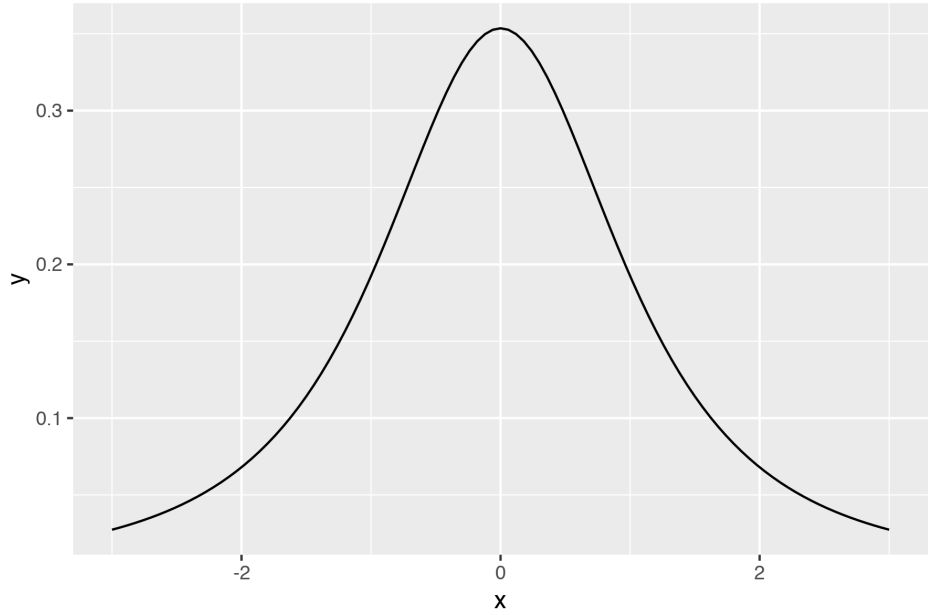
Poisson distribution

lambda = 1



t distribution

df = 2

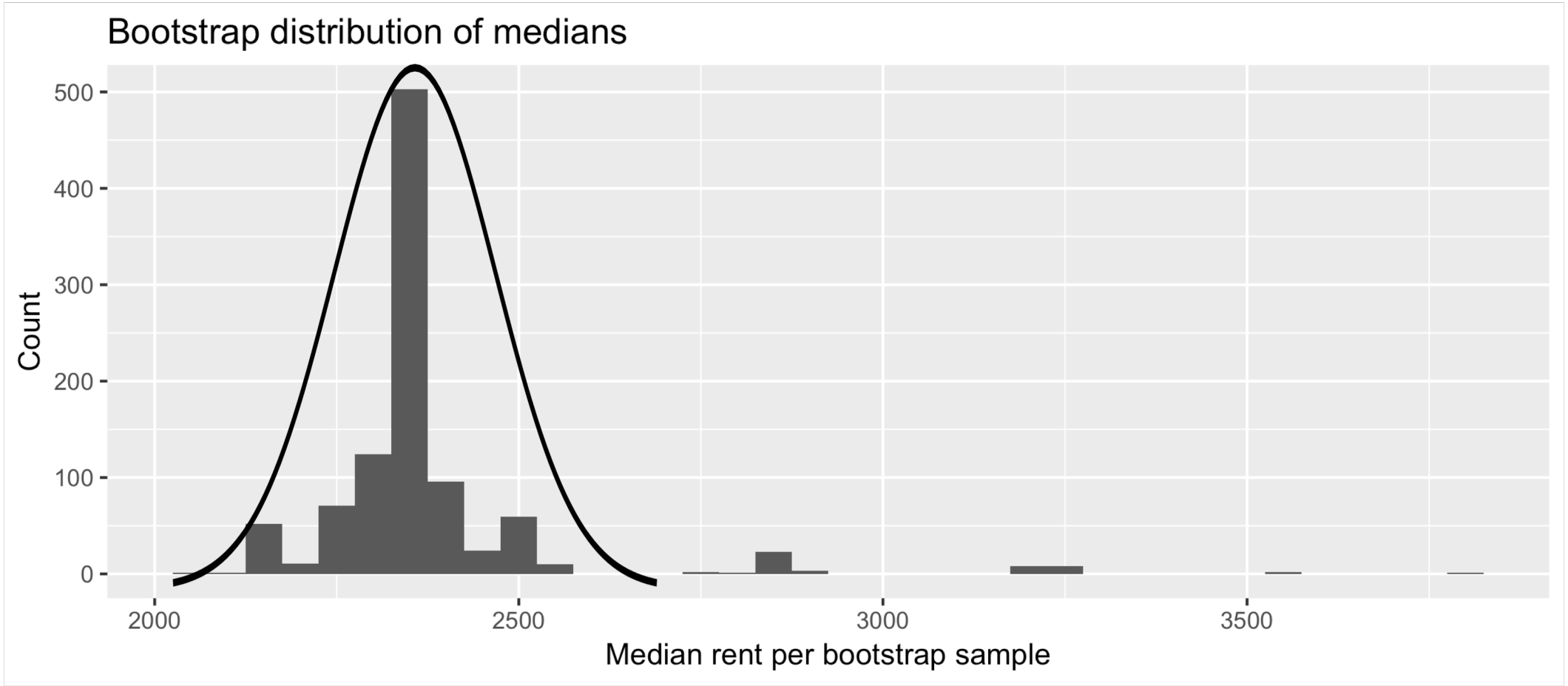


NOT EVERYTHING IS MATHY

What if the data doesn't fit any of those distributions?

Give up or use one that's close enough

NO GOOD MATHY FIT



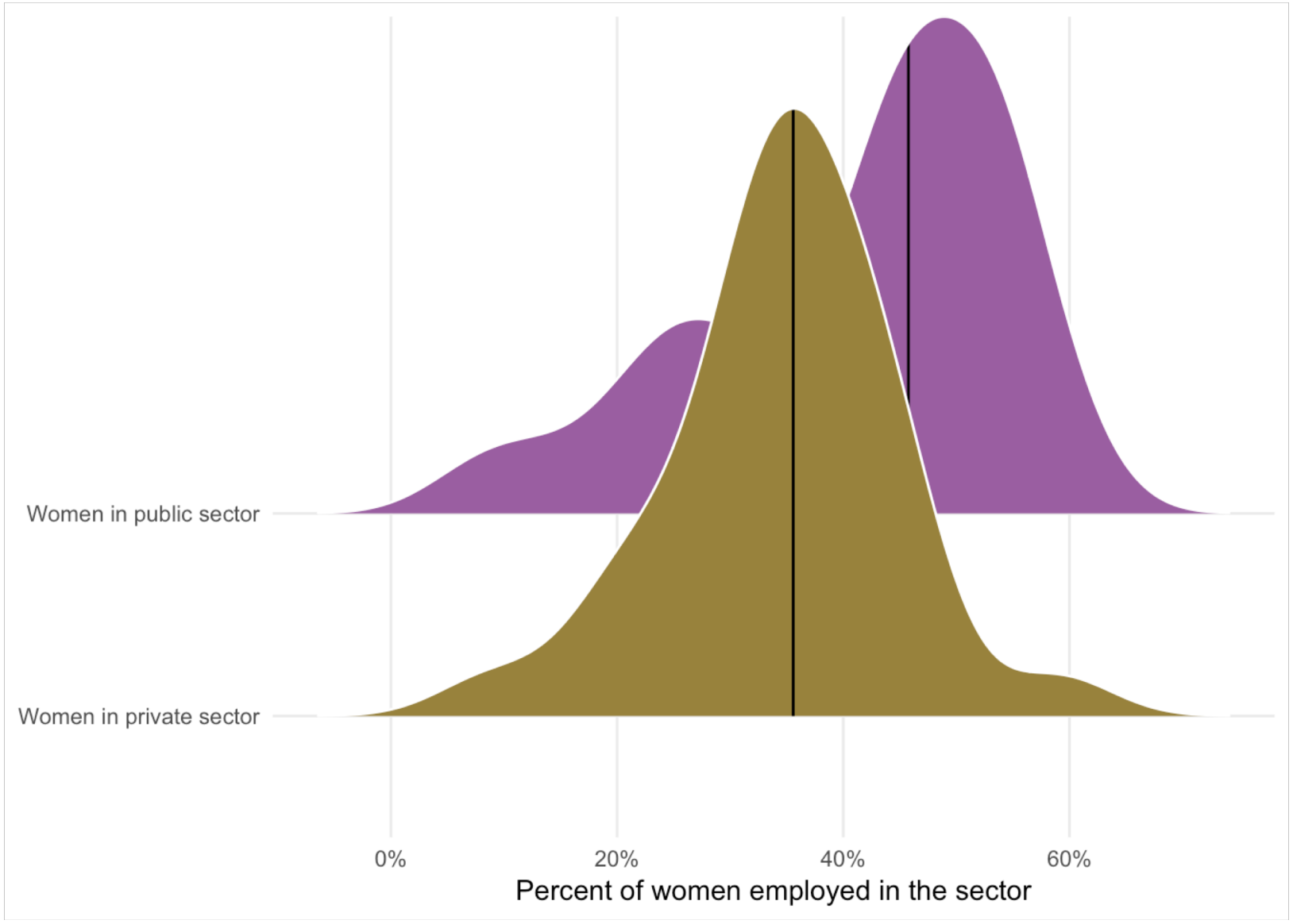
NULL HYPOTHESES TOO

In classical statistics,
null worlds are just some math formula

We have to hope they fit our data

(hence the complicated flowcharts)

WOMEN AND SECTORS



WOMEN AND SECTORS

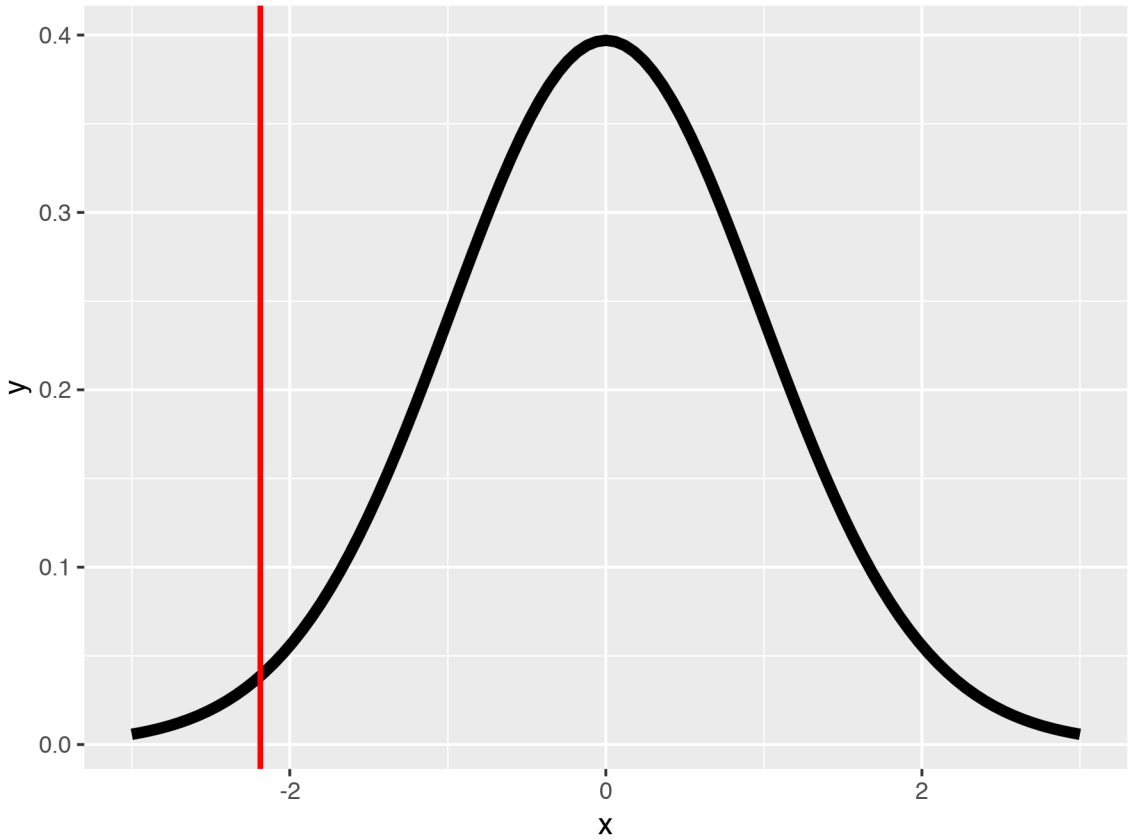
```
> t.test(proportion ~ sector, data = wwbi_2012)

        Welch Two Sample t-test

data:  proportion by sector
t = -2.186, df = 52.604, p-value = 0.03329
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.128090056 -0.005496389
sample estimates:
mean in group Women in private sector  mean in group Women in public sector
                           0.3463885                           0.4131817
```

WOMEN AND SECTORS

```
data:  proportion by sector
t = -2.186, df = 52.604, p-value = 0.03329
```



NULL WORLDS

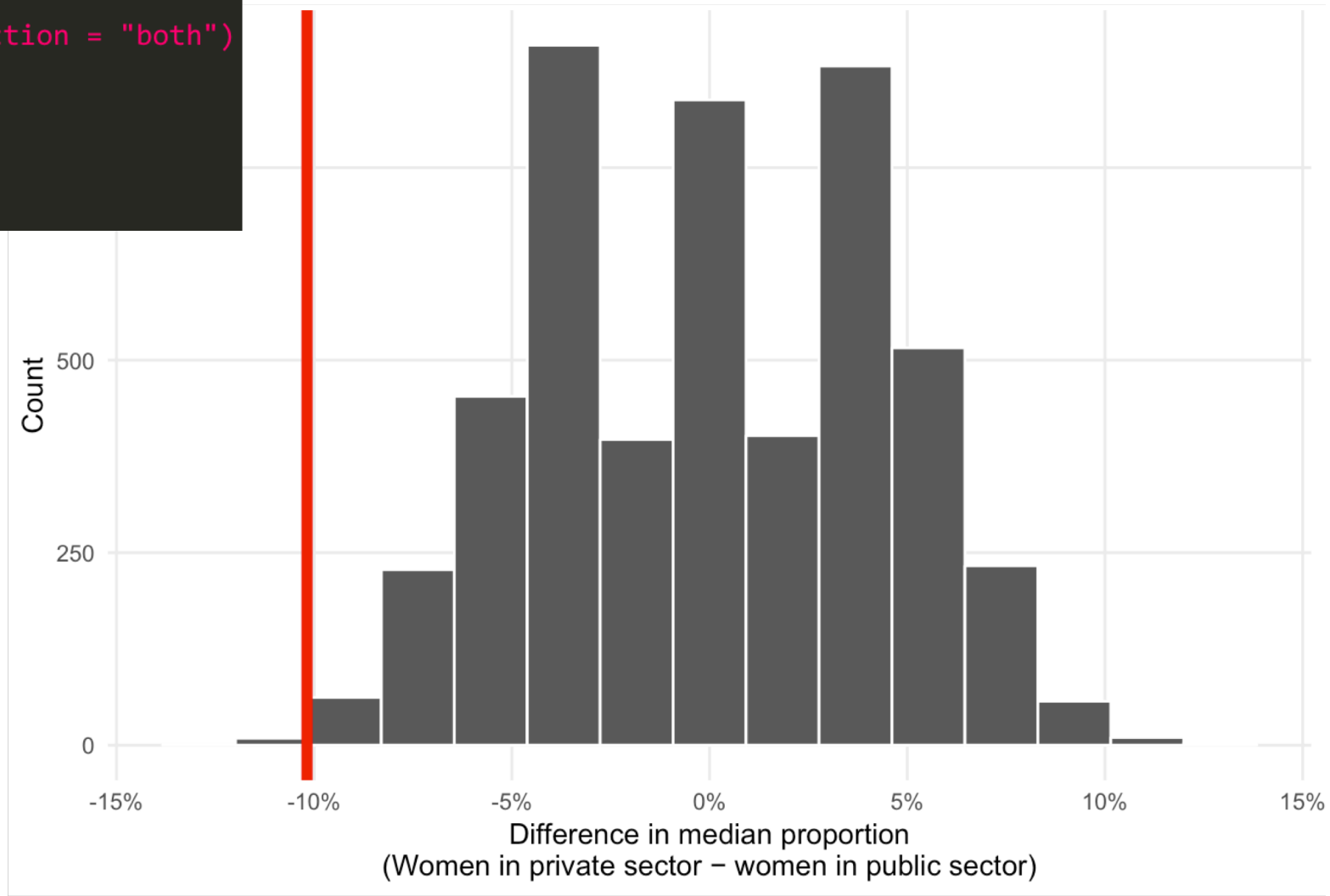
But how do you determine what δ would look like in a null world?

Option 1: Use math and calculus and formulas to determine probabilities

Option 2: Use brute force with simulation

WOMEN AND SECTORS

```
> pub_private_null %>%
...   get_pvalue(obs_stat = diff_prop, direction = "both")
# A tibble: 1 x 1
  p_value
    <dbl>
1 0.00440
```



WHY SIMULATE?

You can test any hypothesis without math and calculus

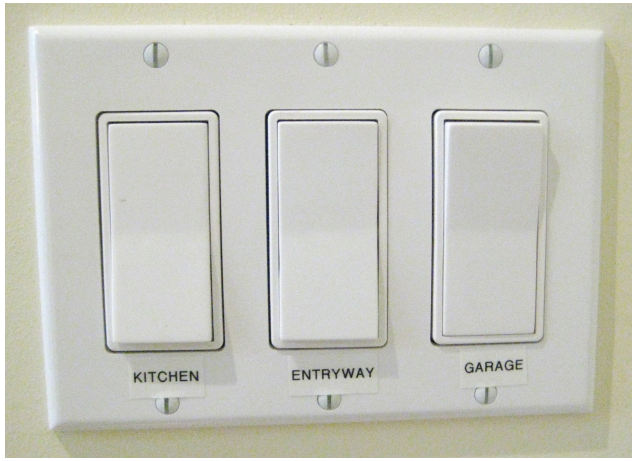
INFERENCE AND REGRESSION

COEFFICIENTS AND UNCERTAINTY

Every regression coefficient is a δ

Something that reflects the population and might be zero or might not be zero

SLOPES AND ZEROES



If a β is 0, Y doesn't change if X changes

If a β is not 0, Y changes if X changes

```r
lm(temperatureHigh ~ humidity + windSpeed + cloudCover +
     precipProbability + visibility + month,
   data = winter_spring)
```

```
# A tibble: 10 x 7
   term              estimate std_error statistic p_value lower_ci upper_ci
   <chr>                <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
 1 intercept            88.2      8.50      10.4    0        71.4    105.
 2 humidity             -0.764    0.081     -9.48   0        -0.924   -0.605
 3 windSpeed            -0.039    0.429     -0.09   0.928    -0.887    0.809
 4 cloudCover            0.019    0.031      0.61   0.543    -0.043    0.081
 5 precipProbability     0.112    0.041      2.77   0.006     0.032    0.192
 6 visibility           -0.8      0.74      -1.08   0.282    -2.26     0.664
 7 monthFebruary         5.90     1.93       3.06   0.003     2.08     9.71
 8 monthMarch            9.62     2.20       4.36   0         5.26    14.0
 9 monthApril            9.32     2.30       4.06   0         4.78    13.9
10 monthMay             19.4      2.38       8.17   0        14.7     24.1
```

term	estimate	std_error	statistic	p_value
intercept	-412.5	118.1	-3.493	0.001
median_home_value	0.004	0	21.99	0
prop_houses_with_kids	14.09	2.853	4.941	0
stateCalifornia	123.3	88.22	1.397	0.164
stateIdaho	9.526	82.74	0.115	0.908
stateNevada	102.5	98.25	1.043	0.299
stateUtah	-213.2	91.21	-2.337	0.021

$$\widehat{\text{property taxes}} = \beta_0 + \beta_1 \text{home values} + \beta_2 \% \text{ houses with kids} +$$
$$\beta_3 \text{California} + \beta_4 \text{Idaho} + \beta_4 \text{Nevada} + \beta_6 \text{Utah} + \epsilon$$

term	estimate	std_error	statistic	p_value
intercept	-2.821	1.355	-2.083	0.04
life_expectancy	0.102	0.017	5.894	0
school_enrollment	0.008	0.01	0.785	0.435
regionEurope & Central Asia	0.031	0.255	0.123	0.902
regionLatin America & Caribbean	0.732	0.294	2.489	0.015
regionMiddle East & North Africa	0.189	0.317	0.597	0.552
regionNorth America	1.114	0.581	1.917	0.058
regionSouth Asia	-0.249	0.45	-0.553	0.582
regionSub-Saharan Africa	0.326	0.407	0.802	0.425

$$\hat{\text{happiness}} = \beta_0 + \beta_1 \text{life expectancy} + \beta_2 \text{school enrollment} +$$
$$\beta_3 \text{Europe} + \beta_4 \text{Latin America} + \beta_5 \text{MENA} +$$
$$\beta_6 \text{North America} + \beta_7 \text{South Asia} + \beta_8 \text{SSA} + \epsilon$$

Y = test scores

Not small vs. small

Class doesn't have aide vs. class has aide

Student not white/Asian vs. yes

Student is boy vs. girl

Student does not receive FRL vs. yes

Teacher not white/Asian vs. yes

Years (actual number)

Teacher does not have MA vs. yes

Table 2: OLS models for four standardized tests

VARIABLES	(1) Reading	(2) Math	(3) Listening	(4) Words
Small class	6.47***	8.84***	3.24**	6.99***
	(1.45)	(2.32)	(1.42)	(1.60)
Regular + aide class	1.00	0.42	-0.58	1.27
	(1.26)	(2.14)	(1.32)	(1.42)
White or Asian	7.85***	16.91***	17.98***	7.08***
	(1.61)	(2.40)	(1.70)	(1.91)
Girl	5.39***	6.46***	2.67***	5.03***
	(0.78)	(1.12)	(0.74)	(0.94)
Free/reduced lunch	-14.69***	-20.08***	-15.23***	-15.97***
	(0.91)	(1.33)	(0.90)	(1.07)
Teacher white or Asian	-0.56	-1.01	-3.68	0.46
	(2.66)	(3.80)	(2.59)	(3.07)
Years of teacher experience	0.30**	0.42**	0.25*	0.30**
	(0.12)	(0.20)	(0.15)	(0.14)
Teacher has MA	-0.75	-2.20	0.50	0.24
	(1.25)	(2.08)	(1.24)	(1.46)
School fixed effects	X	X	X	X
Constant				
	(3.12)	(4.49)	(2.84)	(3.59)

***** p < 0.001, ** p < 0.01, * p < 0.05**

TELLING STORIES WITH DATA

Jennifer Thompson
@jent103

Following

Ratio of time spent "doing analysis" to time spent interpreting & communicating the meaning of said analysis in a way that makes sense to the audience: Currently at 1:100.

4:29 PM - 25 Aug 2018

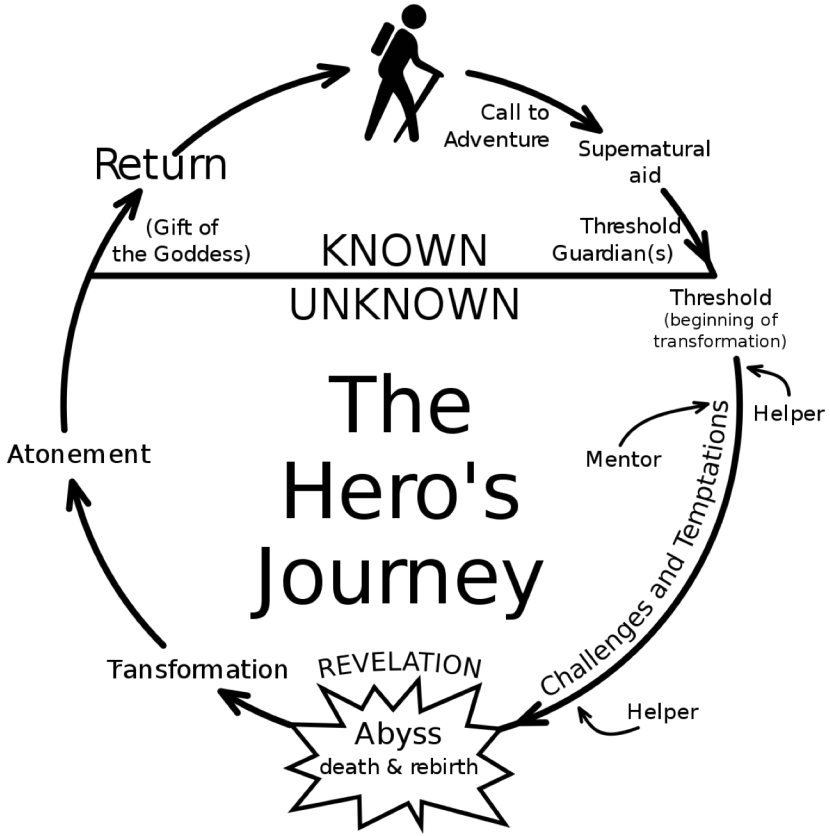
36 Retweets **157** Likes

36 157

Stories are how we translate core, essential content to different forms for specific audiences

EVERY STORY IS THE SAME

https://www.youtube.com/watch?v=LuD2Aa0zFiA


Return
(Gift of the Goddess)
Call to Adventure
Supernatural aid
Threshold Guardian(s)
KNOWN
UNKNOWN
Threshold
(beginning of transformation)
Helper
Mentor
Challenges and Temptations
Helper
Abyss
death & rebirth
REVELATION
Tansformation
Atonement
The Hero's Journey


1. YOU
2. NEED
3. GO
4. SEARCH
5. FIND
6. TAKE
7. RETURN
8. CHANGE

What does this all have to do with statistics and data analysis?

Who is the hero?

You are not the hero.

- About us
  - Company history
  - Market cap
  - # employees and # locations
- About our product and service
  - What it is
  - How it works
  - Why it's better than the alternative
- Call to action (ideally)

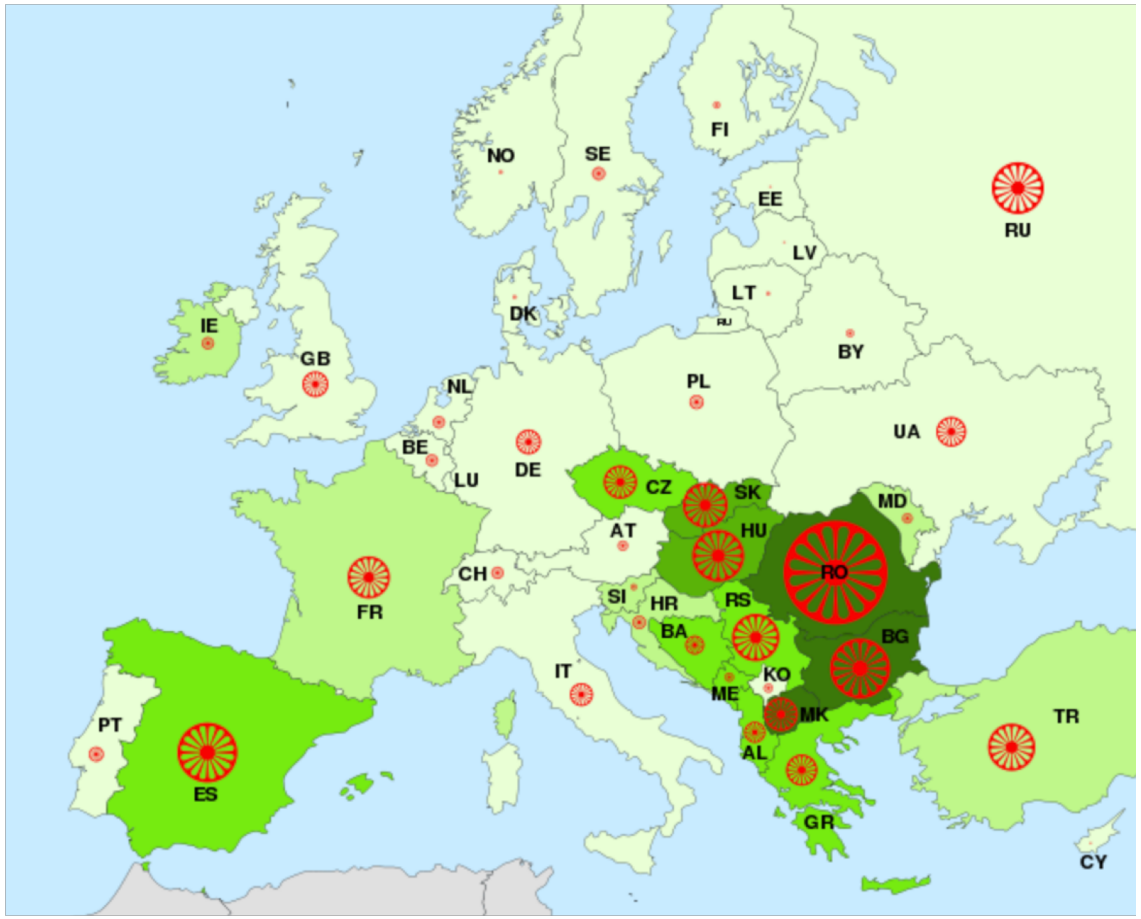
XYZ Co. Equity Partners, LLC

- Founded in 1988 in Anchorage, Alaska
- Invest in companies who:
  - Provide professional IT services
  - Offer exceptional technical and project management expertise
  - Deliver complex data and information management solutions as systems and/or applications integrators
- Average annual revenue: $51.5M

XYZ Co. Software

- Established in 1984
- Headquarters: San Francisco, CA
- Integrated P&C Insurance software and services
- Focused on Alternative Risk & Self-Insured markets
- Recognized leader in risk management solutions
- Over 100 customers in U.S. and Canada


1
A likable audience is unaware that they have a problem or opportunity.
2
They are shown a unique idea that brings their world into an imbalance.
3
They are skeptical, afraid, and resistant to adopt it because it will require change, and change is hard.
4
But a presenter with experience, valuable insights, and magical tools will help on the journey.
5
So they decide to jump in and commit to the idea.
6
Now the real work begins, but it's hard. People and things oppose the effort to change.
7
They are determined to push the idea forward and begin to work on new skills to be successful.
8
They take a major step toward your idea, and it doesn't quite work out as they'd thought.
9
They get discouraged and consider giving up on the idea, but they begin to see some benefit from their efforts.
10
They decide to continue on with a renewed excitement, even though the resistance around them is chronic.
11
Utilizing their new tools, they try one final time to push the idea forward and are victorious.
12
The idea is widely adopted and the galaxy is a better place.
Ordinary World
Call to Adventure
Refusal of the Call
Meeting with the Mentor
Crossing the Threshold
Tests, Allies, and Enemies
Approach the Inmost Cave
Ordeal
Reward (seizing the sword)
Road Back
Resurrection
Return with the Elixir
Ordinary World
Special World


FI
NO
SE
EE
RU
LV
LT
DK
IE
GB
BY
NL
PL
UA
BE
DE
LU
CZ
SK
MD
AT
HU
RO
CH
FR
SI
HR
RS
BA
BG
IT
KO
ME
MK
PT
TR
AL
ES
GR
CY

doi:10.1017/S0003055417000478

Seeing the World Through the Other's Eye: An Online Intervention Reducing Ethnic Prejudice

GÁBOR SIMONOVITS *New York University*
GÁBOR KÉZDI *University of Michigan*
PÉTER KARDOS *Bloomfield College*

We report the results of an intervention that targeted anti-Roma sentiment in Hungary using an online perspective-taking game. We evaluated the impact of this intervention using a randomized experiment in which a sample of young adults played this perspective-taking game, or an unrelated online game. Participation in the perspective-taking game markedly reduced prejudice, with an effect-size equivalent to half the difference between voters of the far-right and the center-right party. The effects persisted for at least a month, and, as a byproduct, the intervention also reduced antipathy toward refugees, another stigmatized group in Hungary, and decreased vote intentions for Hungary's overtly racist, far-right party by 10%. Our study offers a proof-of-concept for a general class of interventions that could be adapted to different settings and implemented at low costs.

Intergroup prejudice has been recognized as one of the most important social problems, leading to discrimination, inequality, and violence in countries across the world. Understanding the mechanisms behind and reducing prejudice are thus of eminent interests for scientific as well as political reasons. Decades of research have accumulated a vast array of knowledge hind (Paluck and Green 2009) because the conditions needed for intergroup contact to reduce prejudice are often extremely difficult, costly, or time-consuming to achieve under realistic circumstances. In this article, we report the results of an intervention that was both effective in reducing prejudice and appears to be easily implementable in a broad class of settings.

You were born eighteen years ago, in June in a Gypsy settlement near Szekszard. Your mother, as you know, long-dead, no news about your father. You might have some siblings somewhere in the country, but you never met them… Entangled dreams and stories swirl in your head. You are scared, but you know that now you can achieve anything. Your net worth of two hundred and ten thousand forints is all in your pocket, that is what you received from the state after all the deductions. You nervously feel your dress as the train rolls in between the platforms.

You are browsing through some websites offering housing and feel the notes in your pocket. You pick the four sublets that look the best among those you could possibly afford. One might be a little overpriced, but all of them offer immediate move in. You call each of the landlords, and agree to check each of them out.

Choose which one you visit first:

- You go to the one on Nemet street
- You go to the one on Rozsa street
- You go to the one on Dob street
- You go to the slightly expensive one on Jokai street

Prejudice toward Roma ↓

Side effects

Prejudice toward refugees ↓

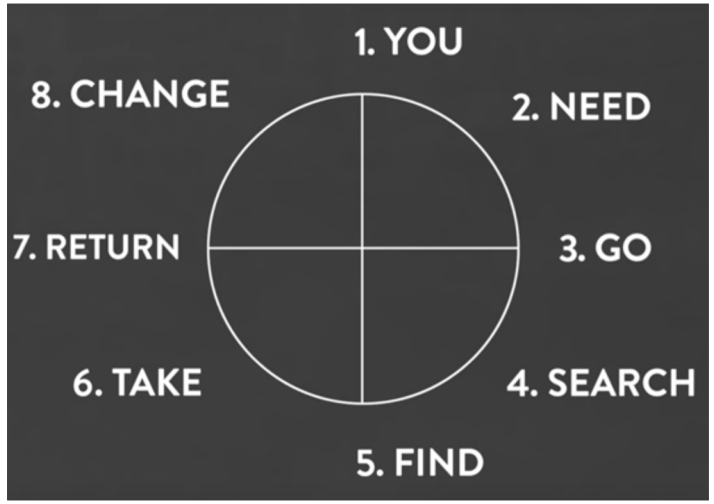
Support for far right party ↓

Provo City
SCHOOL DISTRICT

Requirements

In your report, you need to include at least one of **each** of the following elements (i.e. at least one plot, but more is fine; at least one regression model, but more is fine):

- A plot of a single variable (like a histogram; see ModernDive 3)
- A plot of multiple variables (like a scatterplot; see ModernDive 3)
- 2-3 hypotheses that you will test
- A comparison of proportions or means (see ModernDive 9 and 10)
- A multiple regression model (see ModernDive 6, 7, and 11)



Outline

Here is a suggested outline for your final report:

1. **Executive summary**: one-page summary of your questions, methods, findings, and recommendations
2. **Introduction and description of research questions**: describe the motivation for this study, outline and define what questions you are exploring and why
3. **Data and methods**: explain how the data was collected, provide basic summary statistics (tables and figures) of the main variables you're interested in, and describe what statistical tools you will use to answer your questions (i.e. regression, bootstrapped comparisons of means, etc.)
4. **Results**: answer each of your questions using statistical tools and interpret the results of the different statistical tests you use
5. **Limitations of the study**: provide caveats for your analysis and explain how confident you are in your results
6. **Recommendations and conclusion**: discuss the implications of these findings and make recommendations based on the results
7. **Appendices**: if you want to include tables of summary statistics or tables showing alternative models, you can include them in an appendix instead of in the body of the report itself.

Check the storyness of your final project

Does it take the audience on a journey?