# Quantifying the Gender Pay Gap in the United States: Two Different Paths to the Same Inference

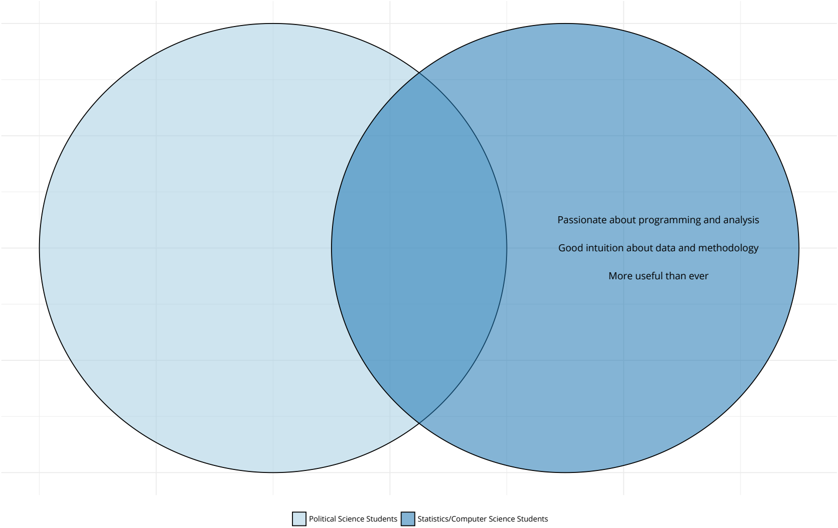
Steven V. Miller

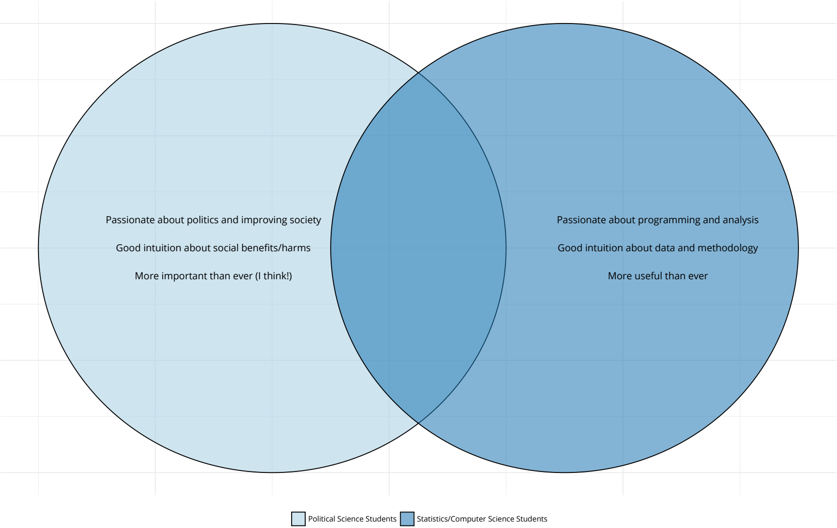
Department of Political Science

CLEMSON UNIVERSITY

# Goals for Today

1. Learn more about the gender pay gap (in the U.S.)
2. Discuss two paths to statistical inference
3. Tell you a bit more about myself


Passionate about programming and analysis
Good intuition about data and methodology
More useful than ever
Political Science Students
Statistics/Computer Science Students



Passionate about politics and improving society

Good intuition about social benefits/harms

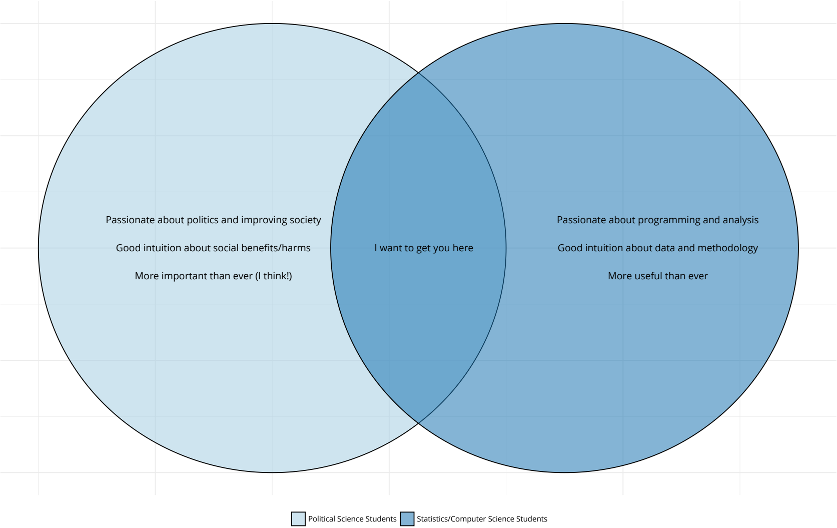
More important than ever (I think!)

Passionate about programming and analysis

Good intuition about data and methodology

More useful than ever

Political Science Students
Statistics/Computer Science Students



Passionate about politics and improving society

Good intuition about social benefits/harms

More important than ever (I think!)

I want to get you here

Passionate about programming and analysis

Good intuition about data and methodology

More useful than ever

□ Political Science Students ▨ Statistics/Computer Science Students

M=F̅ d cos a
m1+m2 / r2
PV= nRT
ω=2* pi*f
Newton's second law of motion :
F = net force = thrust = ṁ
Veq = equivalent engine exhaust
du = -Veq dM/M
Δu=-Veq ln(m)|me mf
Σ i = n(n+1)/2 = 1000+1001/2 =
Assume we move with rocket → u=0
i=1
instantaneous mass of rocket
velocity of rocket
time
net force = thrust = ṁ Veq
equivalent engine exhaust velocity = Isp go
Δu = Veq ln(mf/me) = Veq ln MR = Isp g

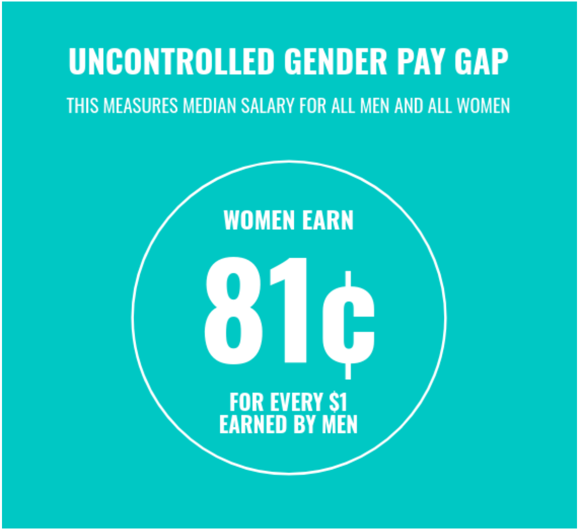
# My Perspective

1. *If you can learn about programming a computer to do something, you have direct access to the deepest, most fundamental ideas in statistics.*
2. Statistics aren't *that* mystifying (but the notation can be unintuitive).

You just need a problem you want to solve.

- Your computer will solve it for you.

# Motivating Issue: The Gender Pay Gap



## UNCONTROLLED GENDER PAY GAP

THIS MEASURES MEDIAN SALARY FOR ALL MEN AND ALL WOMEN

WOMEN EARN

**81¢**

FOR EVERY $1
EARNED BY MEN

# The Data

Let's bring individual-level survey data to bear on this topic.

- *Data:* General Social Survey (2010-2018). *N*: 914.
- *Outcome:* respondent's income (in 2019 USD)
- *Treatment*: respondent's gender (male, female)

Other notes:

- Data subset to those single/never married, with no children, and working full time.
- Data processed/matched to be identical in expectation for age, occupational prestige, college education.
- Raw data available in `gss_wages` in my `{stevedata}` R package.

Table 1: Income Averages for Men and Women in the General Social Survey (2010-2018)

| **Gender** | **Average Income** | **Std. Dev.** | **N** |
|---|---|---|---|
| Female | 51086 | 55360 | 457 |
| Male | 61756 | 76753 | 457 |

On average, women earn $10,670 less than men (or 82% of the average man’s income).

**Skeptic’s argument**: A $10,670 difference in income between men and women could have been observed just by random chance.

**Advocate’s argument**: A $10,670 difference in income between men and women is an important difference and is unlikely to have been observed by random chance.

1. The STAT 101 (analytical) method
2. The computational method

# The STAT 101 (Analytical) Method

To do a (two-sample [Welch’s]) *t*-test, you’ll need to calculate:

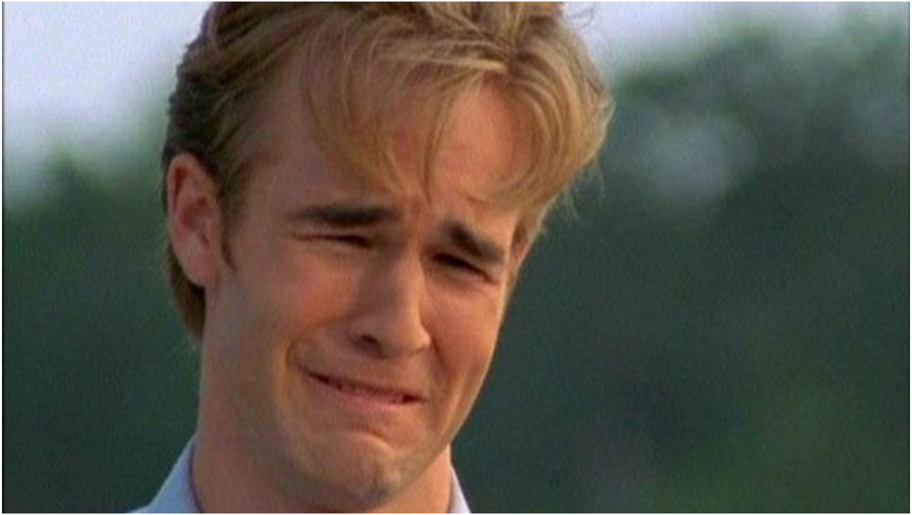
- *t*-statistic
- degrees of freedom
- critical value for rejecting skeptic’s argument.

# Calculating a $t$-statistic

$$
\begin{aligned}
t &= \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}} \\
t &= \frac{51086 - 61756}{\sqrt{\frac{55360^2}{457} + \frac{76753^2}{457}}} \\
t &\approx -2.41
\end{aligned}
$$

# Obtaining Degrees of Freedom

$$\nu \quad \approx \quad \frac{\left( \frac{s_1^2}{N_1} + \frac{s_2^2}{N_2} \right)^2}{\frac{s_1^4}{N_1^2 \nu_1} + \frac{s_2^4}{N_2^2 \nu_2}}$$



# Obtaining Degrees of Freedom

$$\nu \quad \approx \quad \frac{\left( \frac{s_1^2}{N_1} + \frac{s_2^2}{N_2} \right)^2}{\frac{s_1^4}{N_1^2 \nu_1} + \frac{s_2^4}{N_2^2 \nu_2}}$$

$$\nu \quad \approx \quad \frac{\left( \frac{55360^2}{457} + \frac{76753^2}{457} \right)^2}{\frac{55360^4}{457^2 456} + \frac{76753^4}{457^2 456}}$$

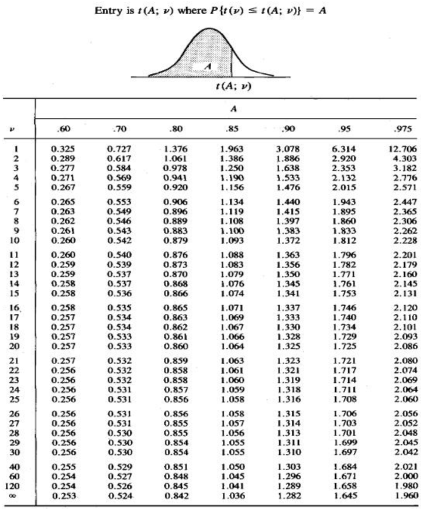
$$\nu \quad \approx \quad 829.39$$

# The *t*-distribution

$$\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

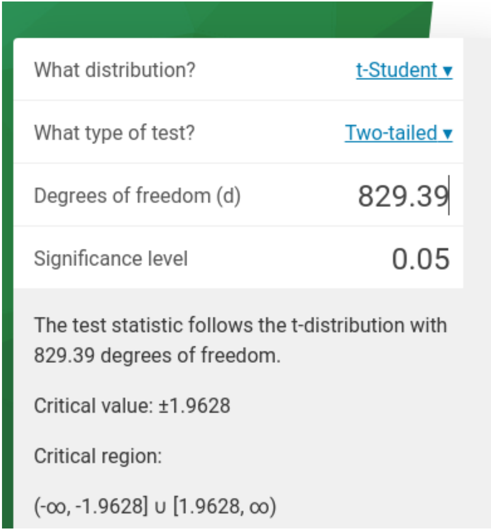
where

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}\,dx$$



Entry is $t(A;\,\nu)$ where $P\{t(\nu) \le t(A;\,\nu)\} = A$

| | A | | | | | | |
|---|---|---|---|---|---|---|---|
| ν | .60 | .70 | .80 | .85 | .90 | .95 | .975 |
| 1 | 0.325 | 0.727 | 1.376 | 1.963 | 3.078 | 6.314 | 12.706 |
| 2 | 0.289 | 0.617 | 1.061 | 1.386 | 1.886 | 2.920 | 4.303 |
| 3 | 0.277 | 0.584 | 0.978 | 1.250 | 1.638 | 2.353 | 3.182 |
| 4 | 0.271 | 0.569 | 0.941 | 1.190 | 1.533 | 2.132 | 2.776 |
| 5 | 0.267 | 0.559 | 0.920 | 1.156 | 1.476 | 2.015 | 2.571 |
| 6 | 0.265 | 0.553 | 0.906 | 1.134 | 1.440 | 1.943 | 2.447 |
| 7 | 0.263 | 0.549 | 0.896 | 1.119 | 1.415 | 1.895 | 2.365 |
| 8 | 0.262 | 0.546 | 0.889 | 1.108 | 1.397 | 1.860 | 2.306 |
| 9 | 0.261 | 0.543 | 0.883 | 1.100 | 1.383 | 1.833 | 2.262 |
| 10 | 0.260 | 0.542 | 0.879 | 1.093 | 1.372 | 1.812 | 2.228 |
| 11 | 0.260 | 0.540 | 0.876 | 1.088 | 1.363 | 1.796 | 2.201 |
| 12 | 0.259 | 0.539 | 0.873 | 1.083 | 1.356 | 1.782 | 2.179 |
| 13 | 0.259 | 0.537 | 0.870 | 1.079 | 1.350 | 1.771 | 2.160 |
| 14 | 0.258 | 0.537 | 0.868 | 1.076 | 1.345 | 1.761 | 2.145 |
| 15 | 0.258 | 0.536 | 0.866 | 1.074 | 1.341 | 1.753 | 2.131 |
| 16 | 0.258 | 0.535 | 0.865 | 1.071 | 1.337 | 1.746 | 2.120 |
| 17 | 0.257 | 0.534 | 0.863 | 1.069 | 1.333 | 1.740 | 2.110 |
| 18 | 0.257 | 0.534 | 0.862 | 1.067 | 1.330 | 1.734 | 2.101 |
| 19 | 0.257 | 0.533 | 0.861 | 1.066 | 1.328 | 1.729 | 2.093 |
| 20 | 0.257 | 0.533 | 0.860 | 1.064 | 1.325 | 1.725 | 2.086 |
| 21 | 0.257 | 0.532 | 0.859 | 1.063 | 1.323 | 1.721 | 2.080 |
| 22 | 0.256 | 0.532 | 0.858 | 1.061 | 1.321 | 1.717 | 2.074 |
| 23 | 0.256 | 0.532 | 0.858 | 1.060 | 1.319 | 1.714 | 2.069 |
| 24 | 0.256 | 0.531 | 0.857 | 1.059 | 1.318 | 1.711 | 2.064 |
| 25 | 0.256 | 0.531 | 0.856 | 1.058 | 1.316 | 1.708 | 2.060 |
| 26 | 0.256 | 0.531 | 0.856 | 1.058 | 1.315 | 1.706 | 2.056 |
| 27 | 0.256 | 0.531 | 0.855 | 1.057 | 1.314 | 1.703 | 2.052 |
| 28 | 0.256 | 0.530 | 0.855 | 1.056 | 1.313 | 1.701 | 2.048 |
| 29 | 0.256 | 0.530 | 0.854 | 1.055 | 1.311 | 1.699 | 2.045 |
| 30 | 0.256 | 0.530 | 0.854 | 1.055 | 1.310 | 1.697 | 2.042 |
| 40 | 0.255 | 0.529 | 0.851 | 1.050 | 1.303 | 1.684 | 2.021 |
| 60 | 0.254 | 0.527 | 0.848 | 1.045 | 1.296 | 1.671 | 2.000 |
| 120 | 0.254 | 0.526 | 0.845 | 1.041 | 1.289 | 1.658 | 1.980 |
| ∞ | 0.253 | 0.524 | 0.842 | 1.036 | 1.282 | 1.645 | 1.960 |



What distribution? t-Student ▾

What type of test? Two-tailed ▾

Degrees of freedom (d) 829.39

Significance level 0.05

The test statistic follows the t-distribution with 829.39 degrees of freedom.

Critical value: ±1.9628

Critical region:

(-∞, -1.9628] ∪ [1.9628, ∞)

# The STAT 101 (Analytical) Method

Since $|-2.41| > 1.9628$, we can reject the skeptic’s argument.

- The observed test statistic exceedingly rare, far from typical.

But there’s got to be a better way.

# In R

For one, make the computer do it for you.

```r
broom::tidy(t.test(realrinc20 ~ gender,
                   data = wages10_matched)) %>%
  # estimate1 = mean for women. estimate2 = mean for men
  rename(diff = estimate,
         t_stat = statistic,
         df = parameter) %>%
  select(diff:df)
```

```
## # A tibble: 1 x 6
##       diff estimate1 estimate2 t_stat p.value    df
##      <dbl>     <dbl>     <dbl>  <dbl>   <dbl> <dbl>
## 1 -10670.    51086.    61756.  -2.41  0.0162  829.
```

# The Computational Method

A computational alternative, called “permutations”, may be more accessible.

- Permutations randomly shuffle the outcome variable and recalculate statistics of interest.

Recall the skeptic’s argument: there are no meaningful differences by gender; the difference is due to chance.

- Permutation allows us to regenerate data to test the skeptic’s argument.

Table 2: Ten Select Gender-Income Pairings

| Gender | Income |
|---|---|
| Female | 45674 |
| Female | 82214 |
| Female | 33494 |
| Female | 23633 |
| Female | 48037 |
| Male | 100484 |
| Male | 28927 |
| Male | 66989 |
| Male | 61169 |
| Male | 72056 |

Table 3: Ten Select Gender-Income Pairings, with a Permutation

| Gender | Income | Perm. 1 |
|---|---|---|
| Female | 45674 | 23633 |
| Female | 82214 | 33494 |
| Female | 33494 | 45674 |
| Female | 23633 | 100484 |
| Female | 48037 | 48037 |
| Male | 100484 | 61169 |
| Male | 28927 | 28927 |
| Male | 66989 | 66989 |
| Male | 61169 | 82214 |
| Male | 72056 | 72056 |

Table 4: Ten Select Gender-Income Pairings, with Two Permutations

| Gender | Income | Perm. 1 | Perm. 2 |
|---|---|---|---|
| Female | 45674 | 23633 | 82214 |
| Female | 82214 | 33494 | 33494 |
| Female | 33494 | 45674 | 100484 |
| Female | 23633 | 100484 | 23633 |
| Female | 48037 | 48037 | 66989 |
| Male | 100484 | 61169 | 61169 |
| Male | 28927 | 28927 | 28927 |
| Male | 66989 | 66989 | 45674 |
| Male | 61169 | 82214 | 72056 |
| Male | 72056 | 72056 | 48037 |

Table 5: Ten Select Gender-Income Pairings, with Three Permutations

| Gender | Income | Perm. 1 | Perm. 2 | Perm. 3 |
|---|---|---|---|---|
| Female | 45674 | 23633 | 82214 | 61169 |
| Female | 82214 | 33494 | 33494 | 33494 |
| Female | 33494 | 45674 | 100484 | 28927 |
| Female | 23633 | 100484 | 23633 | 100484 |
| Female | 48037 | 48037 | 66989 | 23633 |
| Male | 100484 | 61169 | 61169 | 45674 |
| Male | 28927 | 28927 | 28927 | 82214 |
| Male | 66989 | 66989 | 45674 | 72056 |
| Male | 61169 | 82214 | 72056 | 48037 |
| Male | 72056 | 72056 | 48037 | 66989 |

Table 6: Ten Select Gender-Income Pairings, with Four Permutations

| Gender | Income | Perm. 1 | Perm. 2 | Perm. 3 | Perm. 4 |
|---|---|---|---|---|---|
| Female | 45674 | 23633 | 82214 | 61169 | 66989 |
| Female | 82214 | 33494 | 33494 | 33494 | 61169 |
| Female | 33494 | 45674 | 100484 | 28927 | 72056 |
| Female | 23633 | 100484 | 23633 | 100484 | 23633 |
| Female | 48037 | 48037 | 66989 | 23633 | 82214 |
| Male | 100484 | 61169 | 61169 | 45674 | 48037 |
| Male | 28927 | 28927 | 28927 | 82214 | 100484 |
| Male | 66989 | 66989 | 45674 | 72056 | 33494 |
| Male | 61169 | 82214 | 72056 | 48037 | 28927 |
| Male | 72056 | 72056 | 48037 | 66989 | 45674 |

Table 7: Ten Select Gender-Income Pairings, with Five Permutations

| Gender | Income | Perm. 1 | Perm. 2 | Perm. 3 | Perm. 4 | Perm. 5 |
|---|---|---|---|---|---|---|
| Female | 45674 | 23633 | 82214 | 61169 | 66989 | 82214 |
| Female | 82214 | 33494 | 33494 | 33494 | 61169 | 28927 |
| Female | 33494 | 45674 | 100484 | 28927 | 72056 | 33494 |
| Female | 23633 | 100484 | 23633 | 100484 | 23633 | 100484 |
| Female | 48037 | 48037 | 66989 | 23633 | 82214 | 48037 |
| Male | 100484 | 61169 | 61169 | 45674 | 48037 | 66989 |
| Male | 28927 | 28927 | 28927 | 82214 | 100484 | 45674 |
| Male | 66989 | 66989 | 45674 | 72056 | 33494 | 72056 |
| Male | 61169 | 82214 | 72056 | 48037 | 28927 | 23633 |
| Male | 72056 | 72056 | 48037 | 66989 | 45674 | 61169 |

## The Distribution of Possible Average Income Differences Between Men and Women

This effectively blank (for now) plot has a single dashed vertical line representing the actual difference (-$10,670).

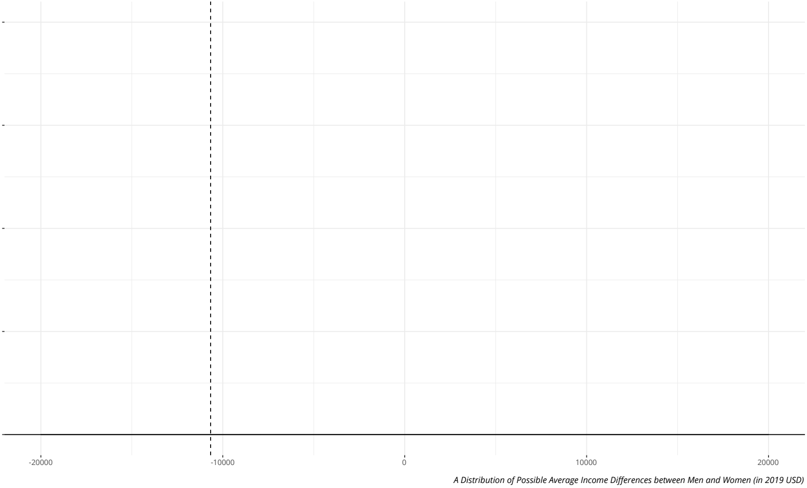


Table 8: Average Income Differences Between Men and Women Across Permutations

| **Perm. No.** | **Income Diff.** | **Mean Income (Women)** | **Mean Income (Men)** |
|---|---|---|---|
| 1 | 6920.9 | 59881.69 | 52960.79 |

Table 9: Average Income Differences Between Men and Women Across Permutations

| Perm. No. | Income Diff. | Mean Income (Women) | Mean Income (Men) |
|---|---|---|---|
| 1 | 6920.90 | 59881.69 | 52960.79 |
| 2 | -9448.46 | 51697.01 | 61145.47 |
| 3 | 1239.81 | 57041.14 | 55801.33 |
| 4 | 238.46 | 56540.47 | 56302.01 |
| 5 | 5834.58 | 59338.53 | 53503.95 |

## The Distribution of Possible Average Income Differences Between Men and Women

This dot plot has five dots for five different permutation means and a vertical line representing the actual difference (-$10,670).

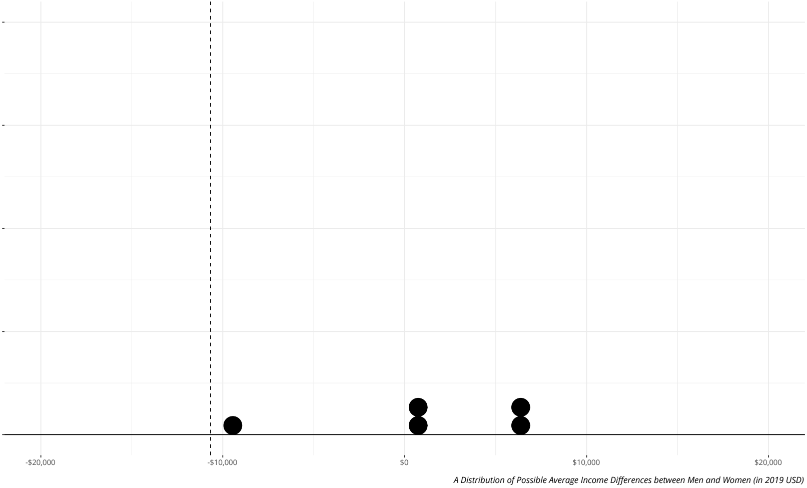
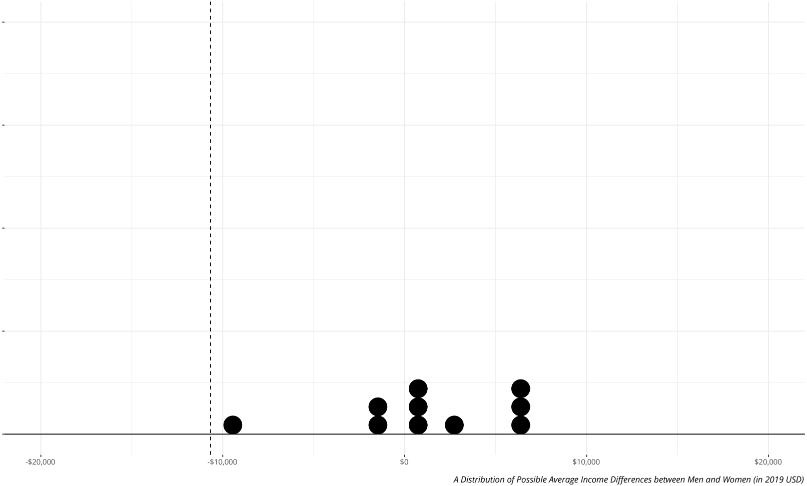


Table 10: Average Income Differences Between Men and Women Across Permutations

| Perm. No. | Income Diff. | Mean Income (Women) | Mean Income (Men) |
|---|---|---|---|
| 1 | 6920.90 | 59881.69 | 52960.79 |
| 2 | -9448.46 | 51697.01 | 61145.47 |
| 3 | 1239.81 | 57041.14 | 55801.33 |
| 4 | 238.46 | 56540.47 | 56302.01 |
| 5 | 5834.58 | 59338.53 | 53503.95 |
| 6 | -1313.68 | 55764.40 | 57078.08 |
| 7 | 1229.56 | 57036.02 | 55806.46 |
| 8 | -1630.85 | 55605.81 | 57236.66 |
| 9 | 2726.30 | 57784.39 | 55058.09 |
| 10 | 6921.70 | 59882.09 | 52960.39 |

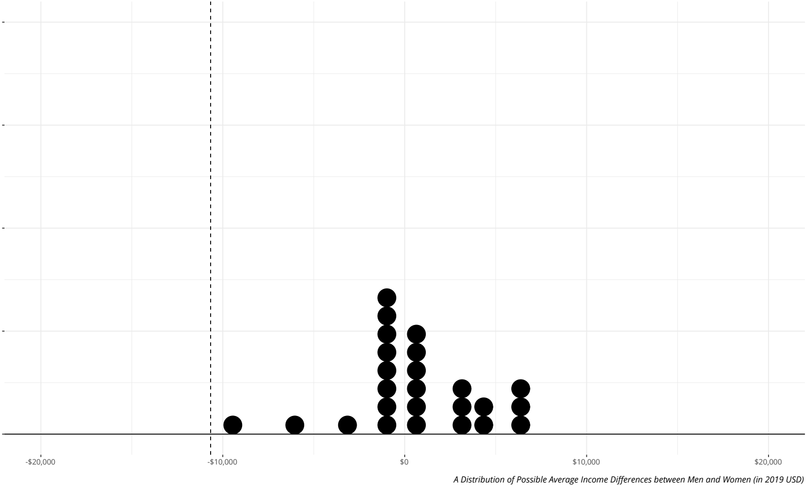
## The Distribution of Possible Average Income Differences Between Men and Women

This dot plot has 10 dots for 10 different permutation means and a vertical line representing the actual difference (-$10,670).



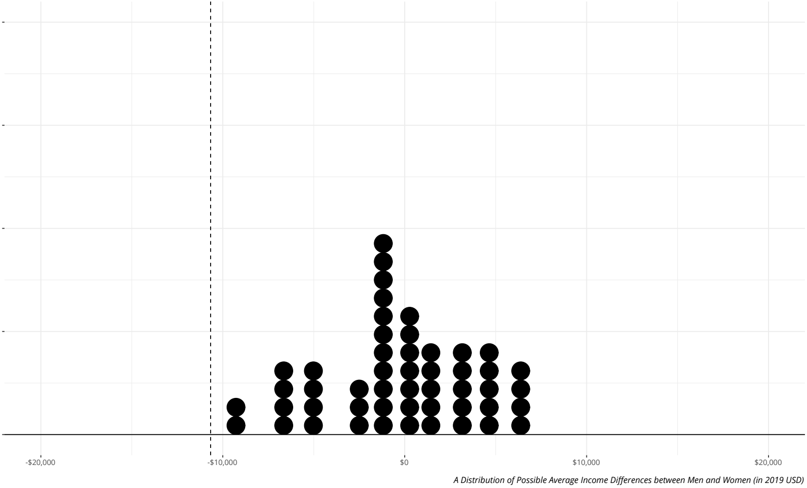
## The Distribution of Possible Average Income Differences Between Men and Women

This dot plot has 25 dots for 25 different permutation means and a vertical line representing the actual difference (-$10,670).



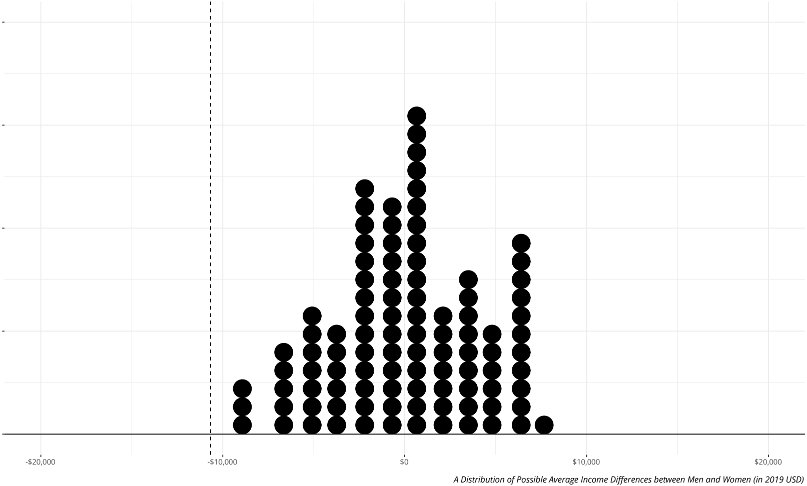
## The Distribution of Possible Average Income Differences Between Men and Women

This dot plot has 50 dots for 50 different permutation means and a vertical line representing the actual difference (-$10,670).



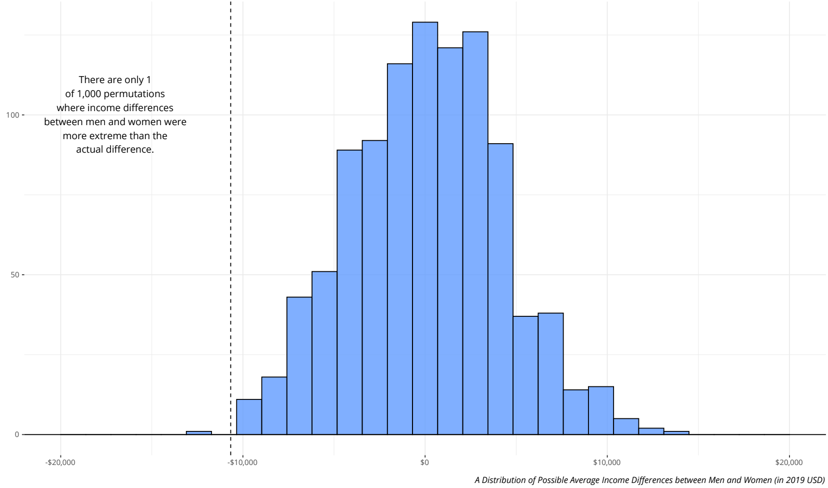
## The Distribution of Possible Average Income Differences Between Men and Women

This dot plot has 100 dots for 100 different permutation means and a vertical line representing the actual difference (-$10,670).



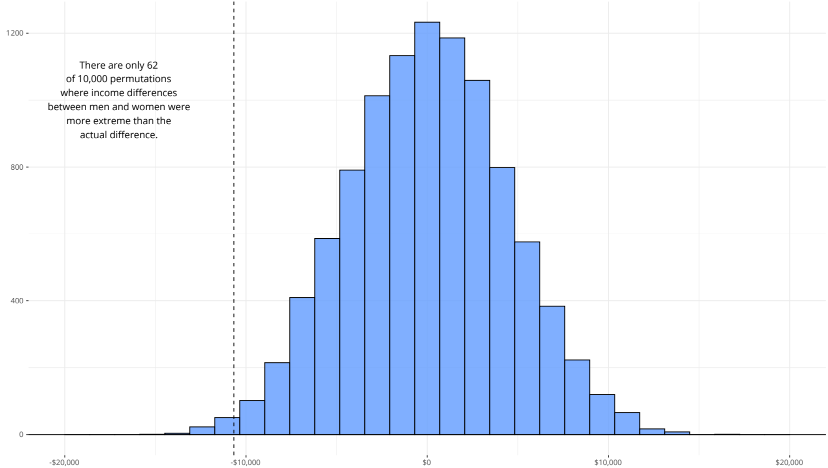
## The Distribution of Possible Average Income Differences Between Men and Women

This histogram has 1,000 different permutation means and a vertical line representing the actual difference (-$10,670).



## The Distribution of Possible Average Income Differences Between Men and Women

This histogram has 10,000 different permutation means and a vertical line representing the actual difference (-$10,670).



There are only 62
of 10,000 permutations
where income differences
between men and women were
more extreme than the
actual difference.

*A Distribution of Possible Average Income Differences between Men and Women (in 2019 USD)*

# Inference by Computation/Permutations

Recall the skeptic's argument: there are no meaningful differences by gender; the observed difference is due to chance. We retort:

- We simulated this argument through 10,000 permutations.
- The observed difference is very rare. Only 62 of 10,000 permutations yielded more extreme differences.

We reject the skeptic's argument. The advocate is right to note an important difference.

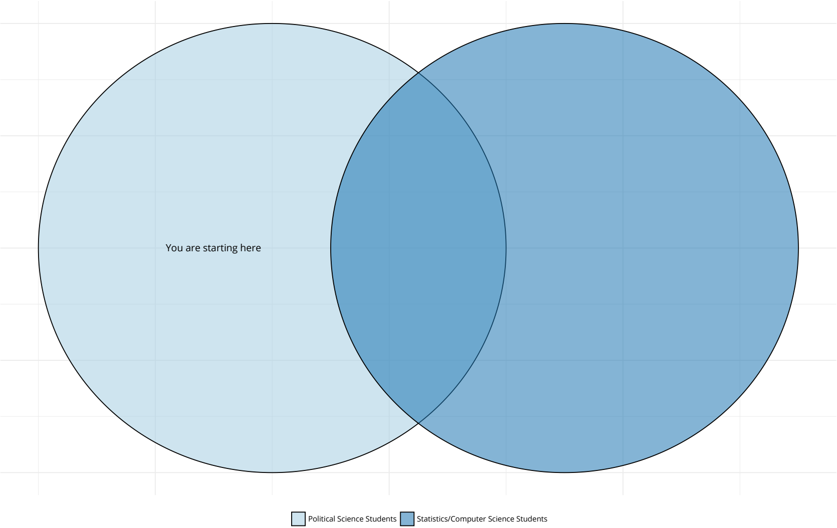
# Takeaways

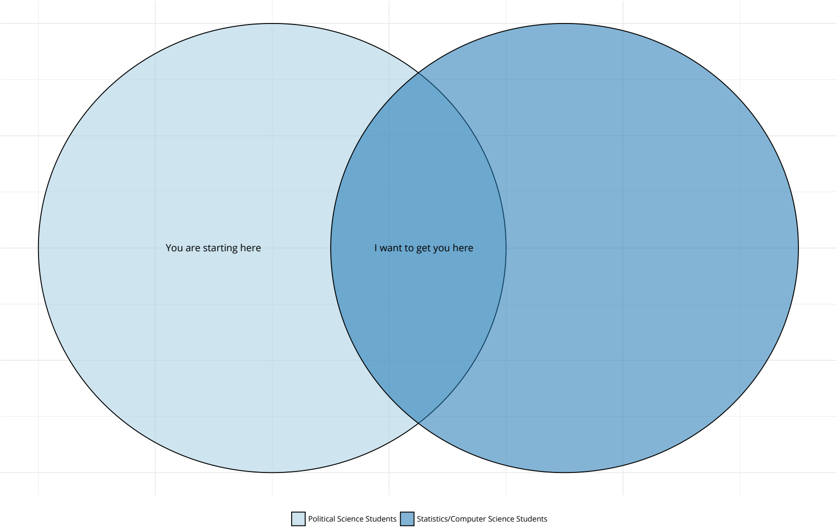
To do statistics in an intuitive way, you'll need the ability to:

1. follow a simple logical argument
2. randomize/shuffle data
3. iterate

You were born with the ability to do the first.

- Any decent programming language will help you with the last two.


You are starting here
Political Science Students
Statistics/Computer Science Students


You are starting here
I want to get you here
Political Science Students
Statistics/Computer Science Students

# Recommended Reading

Check my blog! (`svmiller.com/blog`)

- "Permutations and Inference with an Application to the Gender Pay Gap in the General Social Survey"
- "What Do We Know About British Attitudes Toward Immigration? A Pedagogical Exercise of Sample Inference and Regression"
- "The Normal Distribution, Central Limit Theorem, and Inference from a Sample"

Check out the presentation as well (`svmiller.com/presentations`).

# Table of Contents

Introduction

The Gender Pay Gap

- The STAT 101 (analytical) Method
- The Computational Method

Conclusion/Takeaways