

Your class just performed a modified version of the famous “John/Jenn” study* conducted by researchers at Yale University. You will now perform calculations with the resulting data and investigate if your findings are **statistically significant**.

*Moss-Racusin, C., Dovidio, J., et al. “Science faculty’s subtle gender biases favor male students.” PNAS October 9, 2012 109 (41) 16474-16479; <https://doi.org/10.1073/pnas.1211286109>

Surprise! The resumé you evaluated was fake. You just took part in an experiment. Each resumé the class evaluated was completely identical except for one component: the first name of the applicant. Half of the class received a resumé with the name “John Miller,” and the other half received one for “Jennifer Miller.” Let’s see if the name, alone, was enough to influence your class’s ‘hireability’ estimates.

Ratings for John

Ratings for Jennifer

John mean rating:

Jennifer mean rating:

Difference (John – Jenn):

Is the difference between these mean ratings **statistically significant**? To decide this, we will do a simulation with the data.

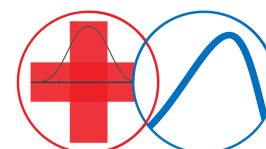
Simulation:

Step 1: We will use index cards to represent the different ratings in this experiment. On each card, write one of the ratings listed in the above tables (for both John and Jennifer). For example, if there were eight ratings of ‘3’ above, then eight index cards should have a ‘3’ on them.

Step 2: Shuffle the cards and deal two even piles. This represents randomly assigning the ratings to the two name groups (assuming gender doesn’t affect the rating). The first pile of cards represents the “John” group, and the second pile cards represents the “Jennifer” group.

Step 3: Fill in the table below with your simulated data.

John																		
Jennifer																		



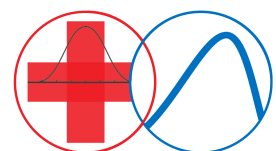
1. Find the mean for each group in your simulation and subtract the means (John - Jenn).
2. Add your difference between means to the dotplot on the board. Sketch the dotplot below.

What does each dot represent?

3. What percentage of the dots are greater than or equal to the difference in means we found in our experiment?

Interpret this percentage:

4. Do you think the difference between mean 'hireability' ratings we found from our experiment is due to the name or has it occurred purely by chance? Explain.
5. In the original Yale study, applications were sent to STEM faculty at various Universities for a lab manager position. Among the 127 faculty that were randomly assigned either Jenn/John materials, the mean 'hireability' rating for John was 3.78 and for Jenn was 2.93. Is this evidence **more convincing** or **less convincing** of gender bias than the data from our class experiment? Explain.



Inference for Experiments

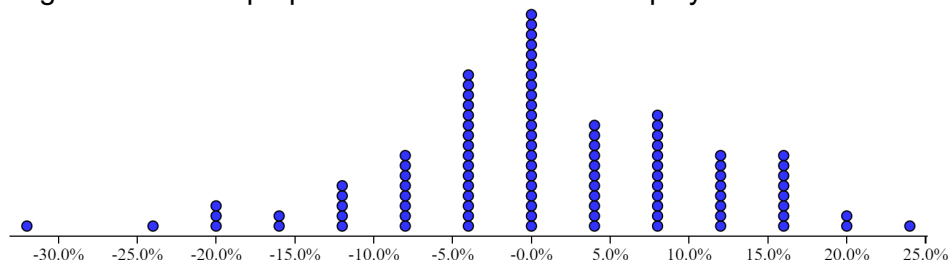
Important Ideas:

Check Your Understanding:

The following is an example of A/B testing, which is a common data science practice among tech and marketing companies. The marketing team at Yelp came up with two new formats for displaying ads on its site - format A and format B. Before deciding on which format to use across its entire site, Yelp wants to test which format would get a higher rate of ad clicks from users. To test this, the company randomly assigns 50 user accounts to see format A and 50 user accounts to see format B. After an hour, format A had 21 users click on an ad ($\hat{p} = \frac{21}{50} = 42\%$). Format B had 19 users click on an ad ($\hat{p} = \frac{19}{50} = 38\%$).

- (1) Calculate the difference (A – B) in proportion of users who clicked on an ad.
- (2) In your own words, explain what ‘statistically significant’ means. Before doing further calculations, do you believe the difference you calculated in part (a) is statistically significant? Explain.

The following simulation was performed assuming format A and format B are equally effective: the individual outcomes from this experiment (‘click’ or ‘not clicked’) were randomly sorted into two groups (representing formats A and B). The proportion who ‘clicked’ were calculated in each group and then the difference was taken between the two groups’ proportions. This was repeated 100 times. The resulting differences in proportions who clicked are displayed here:



- (3) Using the results from the simulation, do you believe the difference in proportions of users who clicked between formats A and B is statistically significant? Explain.

