Factors that Impact the Strength of Evidence

Hubble Space Telescope Proposals and Gender

The Hubble Space Telescope is a telescope in space. Because it is above the Earth's atmosphere, it can take high quality images of planets and stars that are very far away. Scientist are able to use the Hubble Telescope if they are granted permission by NASA. Use of the telescope is a limited resource, and so NASA has an interest in proposals that have the most promise to result in learning more about our universe.

Recently, beginning in the year 2020, NASA switched to using a blind proposal strategy. In other words, they take the names off of proposals when they are deliberating about accepting the proposal. Previously, however, the names of the scientists were on the proposal. When social scientists Stephanie Johnson and Jessica Kirk were invited to listen in on NASA's conversations to determine which proposals to accept to improve the process, they found that the conversations often centered around *who* wrote the proposal, and whether they were known in the scientific community as capable, rather than about the merits of the proposal itself (Greenfieldboyce, 2022).

Before the policy change, Iain Neill Reid (2014) looked at past proposals for use of the Hubble Space telescope. He looked at Cycles 11-20 of proposals. Researchers wondered whether proposals from female principal investigators (PIs) would be less likely to be accepted.

1. What is the research question?

2. What are the observational units? What are the variables?

In Cycle 11, there were 1078 proposals. 205 of them, or 19%, were submitted by female principal investigators (PIs). Of the 1078 proposals that were submitted, 198 of them were accepted. 30 of the proposals from female PIs were accepted. In other words, 30/198, or 15%, of the accepted proposals were submitted by women.

3. Does the data suggest that proposals from female PIs are less likely to be accepted? How so?

We can think of Cycle 11 proposals as the population and the accepted proposals as a sample. While 15% is less than 19%, they are fairly close. We might wonder if we would see a difference of this size just due to chance.

4. Label the figures using the correct symbols.

   π =_____

   $\hat{p}$ = _____

5. Write the null and alternative hypotheses in words and symbols.

6. What type of sample are the accepted proposals?

While the accepted proposals are certainly not a random sample, we could investigate whether we would get a difference like the one in Cycle 11 just due to chance.

7. In the past, we have used coins to simulate the chance model. Could you use a coin toss in this case? Why or why not?

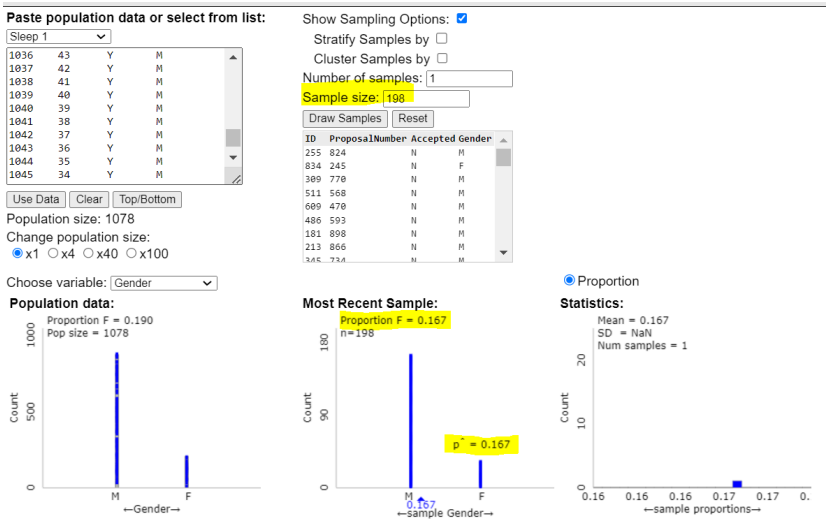
Here is a link to the data for Cycle 11. Put the data in the Sampling from Finite Population applet.

8. Under "Choose Variable" select "Gender." What proportion of the principal investigators submitting proposals were female?

9. If you were to take a sample of 198 proposals, what proportion of them would you expect to be submitted by female PIs? Would it be exactly this amount, or do you think it could vary a little?

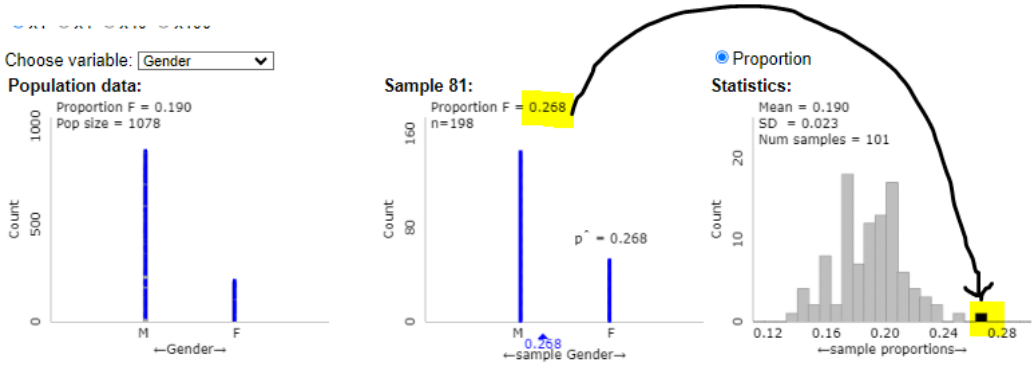
Recall that there were 198 proposals accepted in Cycle 11. Even though 19% of the proposals were written by women, only 15% of the accepted proposals were written by women. We'd like to know if it's reasonable that we could get a sample of 198 proposals with only 15% of them being written by women just due to chance.

10. Select "Show Sampling Options." Under "Sample Size," write 198. Then click the "Draw Samples" button. Look at the middle graph. What proportion of your sample is female?



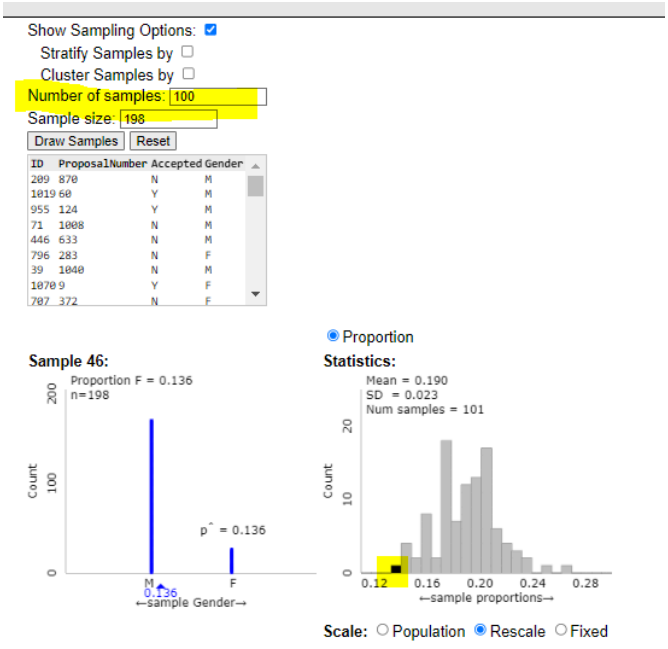
11. Compare with your classmates. How far away are you all from the population proportion, $\pi = 0.19$?

Click “Draw Samples” again. Notice in the right hand graph that the applet is recording the proportion of each sample that you take. If you click on one of the squares, it will show you a previous sample.



We could keep clicking the “Draw Samples” button over and over again to take additional simulated samples to see how far away sample proportions would typically be from the population parameter. As a short cut, we can tell the applet to take many samples at once.

12. Where it says “Number of samples” enter in 100. Notice how it creates a graph of all of the proportion of women in each of those samples on the right. What was the lowest sample proportion of women you got in your 100 samples? What was the lowest sample proportion of women your classmates got?



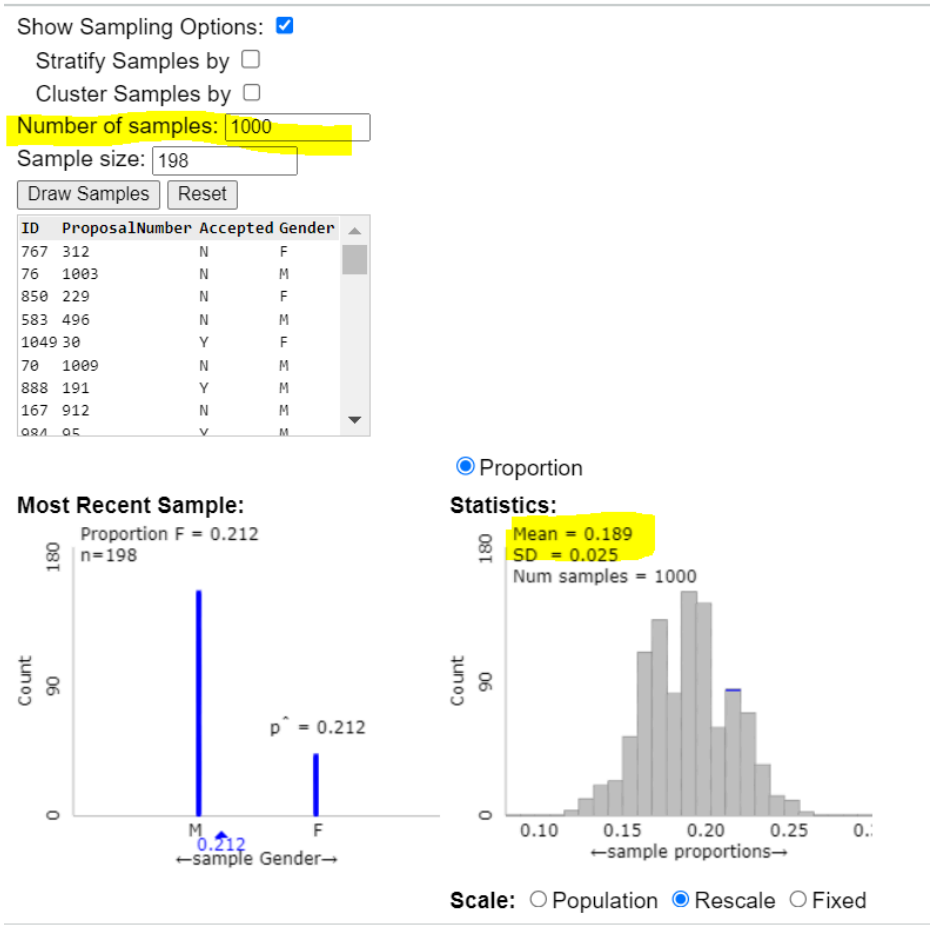
13. The graph on the right is referred to as the *sampling* distribution, or the distribution of sample proportions.

    Where is the distribution of sample proportions centered? Why does that make sense?

    What is the standard deviation of the distribution of sample proportions? (This is referred to as the **standard error.** It tells us how close the sample proportions are to the center.)

    What is the lowest and highest simulated sample proportion? Did you ever get a sample proportion as far away as 15%?

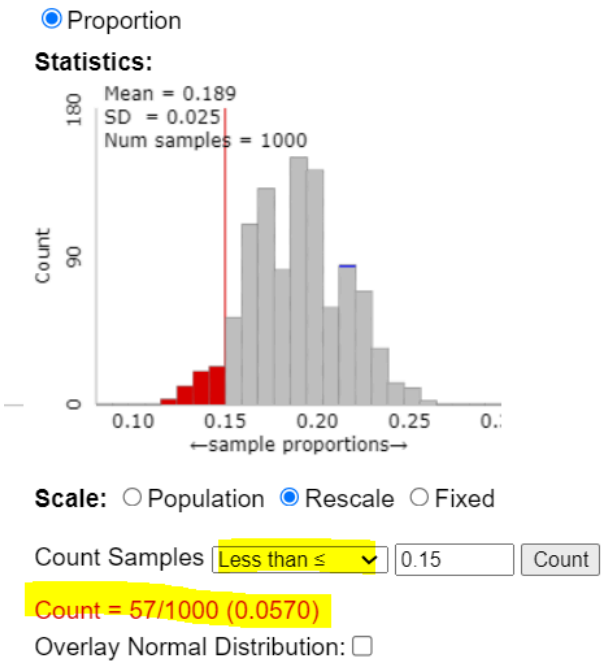
14. When you compare your results with your classmates, you may notice some slight differences. However, if we all take 1000 samples, our results will be similar. Change the **Number of Samples** box to 1000 total samples. (Keep the Sample Size at 198 proposals.) Compare your distribution of sample proportions, the far-right graph, with your classmates. Is it similar or different? Be sure to compare both the center and the spread, the **standard deviation** (standard error).



The distribution of sample proportions is centered at _____ and has a standard deviation of _____.

15. Recall that the proportion of accepted proposals that were submitted by female PIs was 15%. How many of your simulated samples have a proportion of fewer than 15% women, the proportion of accepted proposals submitted by female PIs? What proportion of your simulated samples is that? (This is your one-tailed p-value.)

    If it is difficult to count them, you can enter the value 0.15 into the "Count Samples Less Than" box, and the computer will count the number of samples for you. Compare your answers with your classmates.

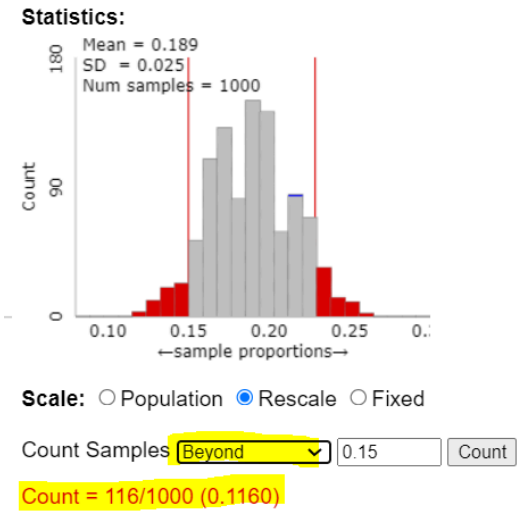


The p-value is _____. Out of _____ total simulated random samples, _____ of them, or 0._____ of them, were less than 15% women.

16. It is very common to conduct a 2-tailed test, because then we make no assumptions about the direction.

    What would the null and alternative hypotheses be if you were conducting a 2-tailed test? Write them in words and symbols.

17. Change the "Count Samples" drop down box from "Less Than" to "Beyond." What happens to the graph? What happens to the proportion of samples?



18. Given this data, could you reject the null hypothesis? Justify your answer, and explain what that means in terms of evidence about proposals to use the Hubble Space Telescope that are written by female PIs. Include the p-value in your response.

In this case, there is a difference between the proportion of proposals that are submitted by female PIs , 19%, and the proportion of proposals that are accepted that are submitted by female PIs, 15%. However, this 4% difference is the size that we might see even if we were just selecting 198 proposals randomly. It could be that accepted proposals are less likely to be submitted by female PIs, or it could be that there is some chance variation. With this data, it's unclear.

19. However, Reid looked at more than 1 cycle of data. He collected data on Cycles 11-20. The data table for all of these cycles put together is below, and you can find a link to the raw data here.
    - State the **hypotheses** and complete a 2-tailed test for this data.
    - Be sure to report the **p-value** and explain the **meaning** in terms of Hubble Space Telescope proposals and female PIs' applications.
    - Be sure to report the center and **standard deviation** of the distribution of the sample proportions, the graph on the right of the applet. Compare these values you found in question 14.

	Total number of proposals	Total number of accepted proposals	Number of Proposals from Female PIs	Number of Accepted Proposals from Female PIs
Cycles 11-20 combined	9410	2103	1953	353

Null Hypotheses: H_0: $\pi =$ ____ (notice, π is a different value than previously)

Alternative Hypotheses: H_A: $\pi \neq$ ____

$\hat{p} =$ _____

Sample size $n =$ _____

Standard error, standard deviation of sample proportions: _____

p-value:

20. How did the p-value from your first investigation using only Cycle 11 data compare to the p-value when you used Cycles 11-20 data? Why might this be the case?

When we looked at data only from Cycle 11, it seemed like the fact that the proportion of accepted proposals that had been submitted by female PIs was lower than the proportion of all of the proposals that had been submitted by female PIs could just be due to chance. After all, 15% is sort of close to 19%, and if we are taking a sample of only 198 proposals from a population where 19% of them were submitted by women, we would expect the samples to have anywhere from 11% to 26% of the proposals submitted by women. The standard deviation of the sample proportions, or standard error, was around 0.025, or 2.5%. 15%, the sample proportion, is less than 2 standard errors away from the population proportion. (19% - 15% = 4% < 2 x 2.5%).

However, when we looked at Cycles 11-20, the sample size was much bigger, 2103 proposals. You should have seen that the standard error, or the standard deviation of the sample proportions, was much smaller, 0.008, or 0.8%. Most of the sample proportions were very close to the center, the parameter value of $\frac{1953}{9410} = 20.8\%$. Therefore, seeing a sample proportion as far away as $\frac{353}{2103} = 16.8\%$, is much less likely. The population parameter and the sample proportion are still 4% away from each other (20.8% - 16.8%), but with larger samples we are less likely to get something this far away.

When we look at the larger data set, we have stronger evidence against the null hypothesis. Such a difference would not occur just due to chance. Proposals submitted by female PIs were less likely to be accepted, and this was a general trend. Notice that it was always true to that the null hypothesis was false, but when we only looked at the smaller data set, we did not have enough evidence to say the null hypothesis was false. This is why we never say that we "proved" the null hypothesis is true. We can only say whether we have evidence **against** the null hypothesis.

Reid's data proves that the difference in acceptance rates for women could not just be due to chance. It does not tell us why there is a difference. In this case, there was no explicit bias against women – none of the reviewers explicitly indicated that they did not believe projects headed by women should not be given access to the Hubble Space Telescope. Instead, there was a systematic bias. Researchers Kirk and Johnson found that in the discussion of the proposals, reviewers commonly commented on who the principal investigator was. If there was a vagueness in a proposal and the researcher was known by the reviewers, this vagueness was explicitly overlooked. If the proposal was from a lesser-known scientist, the conversation cast doubt on whether the person would be capable of completing the research. Many male scientists have had a head start on building their name recognition. Many colleges were not open to women, particularly prestigious colleges – Columbia did not accept women until as late as 1983. Given both this quantitative data from Reid and the qualitative data from Kirk and Johnson, NASA decided to begin a blind review policy, where the names of the investigators were removed from the proposal. In this way, the proposals were evaluated based on their merit, rather than the esteem of

the investigators. As reported by NPR, there was an immediate impact on this change in policy, with proposals by female principal investigators and proposals by newer researchers, those who have not yet had an opportunity to use the Hubble Space Telescope, have been accepted at a higher rate under the new policy. There are similar investigations about names on resumes for job positions and salary offerings that suggested blinding application materials may transfer toward addressing other issues of inequity as well.

Greenfieldboyce, N. (2022, January 11). Who gets to use NASA's James Webb Space Telescope? Astronomers work to fight bias. *NPR*. https://www.npr.org/2022/01/11/1071752559/who-gets-to-use-nasas-james-webb-space-telescope-astronomers-work-to-fight-bias

Reid, I. N. (2014). Gender-Correlated Systematics in HST Proposal Selection. *Publications of the Astronomical Society of the Pacific*, *126*(944), 923. https://doi.org/10.1086/678964