

STAT 217: Permutation Test Activity - A non-parametric approach

Lab tests with animals have shown that exposure to tobacco smoke is harmful in many ways. To make connections to humans has been more of a challenge. One dataset which might help us connect tobacco use of pregnant women to birth weights of their babies comes from a large set of data on births in North Carolina. We will examine a random sample of size 200 from the much larger dataset. The two variables provided are habit (either smoker or nonsmoker) and weight (babys weight at birth measured in pounds).

1. Could there be some physiological reason why birthweights for the children of the 28 smokers might be lower than the birth weights of the babies born to nonsmokers? Write down what you and your group know about smoke and nicotine to hypothesize a connection to birthweight.
2. What is the response variable in this study? Is it quantitative or categorical?
3. Is there an explanatory variable in this study? If so, name it and tell which type of variable it is.
4. Go to <http://lock5stat.com/statkey/> and click on **Test For Difference in Means**. Press the **Edit Data** button. This is where you will paste the birth weight data. The data is on D2L. Open it in a text editor and copy and paste it in.
5. What is the average birth weight for smokers? Non-smokers?
6. What is the difference in average birth weights between non-smokers and smokers?

You will conduct a permutation test to find out how likely it would be to see this large a difference in sample means if the two populations really have the same overall mean birth weight. By doing the reassignment many times, we can see what results are expected when the populations really have the same distribution of responses.

7. Write the null and alternative hypotheses for your permutation test.
8. Now generate 1 reshuffled sample by clicking **Generate 1 Sample**. What is the mean birth weight to smokers from this simulated trial? Answers will vary. What is the mean birth weight to nonsmokers for this single simulated trial? What is the difference in means between these two groups?

9. Now click **Generate** 1000 samples. In the 1001 reshuffled samples, what was the smallest difference in means? What was the largest?
10. How were each of these samples created?
11. Where is the histogram centered and why?
12. Go back and consider the difference of means you observed in the original sample. Based on the null distribution you just simulated, is it **possible** that the null hypothesis is true?
13. Is it **likely** that the null hypothesis is true?
14. Now let's find the approximate p-value. Check the box for either right-tail or left-tail depending on your alternative hypothesis. After doing so, a box with a number in it will pop up on your x-axis. Click on this box and enter the observed difference in means from the original sample. The p-value is then displayed. What is your p-value? Answers may vary.
15. Write a conclusion in context of the problem.