**Activity 7.3.5 – Testing Differences in Two Population Proportions**

**Do Female Students Prefer Water More than Male Students?**

Do male and female high school students differ in their beverage preferences? Suppose a researcher surveyed 50 high school students on their drinking preference. The two-way table below summarizes the results from the random sample. Does this sample data provide evidence that males and females in this population have different drinking preferences?

|  |  |
| --- | --- |
|  | **Gender** |
| **Drink Preference** | Male | Female |
| Water | 13 | 17 |
| Soft Drink | 9 | 4 |
| Other | 3 | 4 |
| **Total** | **25** | **25** |

The data from this survey enables us to test whether a difference exists between two population proportions. Similar to our tests on a single population proportion, we can use a randomization distribution to assess whether the difference in sample results is significant – that is not likely due to chance, and instead due to the fact that they originate from populations with different parameters.

The following problems detail this process.

1. Determine the sample proportion of males who prefer water and the sample proportion of females who prefer water. Label the sample proportions $\hat{p}\_{m}$ and $\hat{p}\_{f}$.
2. Find the difference in sample proportions, $\hat{p}\_{m}-\hat{p}\_{f}$. This difference is a statistic.

There are two possible scenarios that could explain this difference in sample proportions.

**Scenario 1:** The population proportions are the same (i.e. the proportion of males who prefer water equals the population proportion of females who prefer water) and the difference in sample proportions is due to chance alone or due to sampling variation.

**Scenario 2:** The population proportions are different (i.e. the proportion of males who prefer water is less than the population proportion of females who prefer water) and the difference in sample proportions reflects the difference that exists between the population parameters.

To make a decision on which scenario is most probable, we perform a randomization hypothesis test using a randomization distribution of *differences in sample proportions*.

**Randomization Test for a Difference in Population Proportions**

* Assume the population proportions are equal.
* Construct a randomization distribution of differences in sample proportions from random samples from the *same* population.
	+ Combine the data from both samples to form a population
	+ Randomly sample from the population and calculate differences in sample proportions
* Find the probability of observing a difference in sample proportions as extreme as the one we observed.
* State a conclusion about the population proportions.

**Constructing a Randomization Distribution using Random Numbers**

The image below shows how the samples are combined to form a new “population” of responses. The data are based on the Gender and Drink Preferences two-way table. “Successes” represent students who prefer water and “failures” represent students who don’t prefer water.



The population formed by combining the samples contains 50 observations: 30 “successes” and 20 “failures”, and has a population proportion of *p* = 0.60. We assume that this value is the *common* population proportion, so this value is the assumed parameter.

We will now randomly re-sample from this population, finding a sample proportion from a random sample of size 25 (for males), finding a sample proportion from a random sample of size 25 (for females), and then computing differences between the sample proportions.

We can model a population with proportion *p* = 0.60 using random numbers as follows:

* The population consists of all two-digit numbers: 00 to 99
* Numbers between 00 to 59 are “successes”
* Numbers between 60 – 99 are “failures”
1. *Random sample of males –* Randomly select 25 two-digit numbers, with replacement, for the sample of males. Count the number of two-digit numbers that represent “successes”. Find the sample proportion of “successes”. Label it $\hat{p}\_{m}$.
2. *Random sample of females* – Randomly select 25 two-digit numbers, with replacement, for the sample of females. Count the number of two-digit numbers that represent “successes”. Find the sample proportion of “successes”. Label it $\hat{p}\_{f}$.
3. Find the difference in the sample proportions from the resampled samples, $\hat{p}\_{m}-\hat{p}\_{f}$.
4. Repeat the steps in Questions 3 through 5 to compute a second difference in sample proportions, $\hat{p}\_{m}-\hat{p}\_{f}$.

1. Create a dotplot of the differences in sample proportions using all the data from your class.



1. Using technology, determine the mean and standard error of the distribution of differences in sample proportions.
2. The difference in the *observed* sample proportions was -0.16. The randomization distribution assumes that the proportion of males and females who prefer water is the same. Assuming this is true, what is the probability of obtaining two random samples of size 25 in which the difference in sample proportions that is -0.16 or less? This probability is called a *P*-value (probability value).
3. Is the difference in the observed sample proportions statistically significant?
4. What can we conclude about the proportion of male high school students and proportion of female high school students who prefer drinking water?

**Soft Drink (Soda) Drinking Preferences**

Does the data provide evidence that a greater proportion of males than females prefer drinking soft drinks. We can perform a randomization test on these data to address this question.

1. Determine the sample proportion of males who prefer soft drinks and the sample proportion of females who prefer soft drinks. Label the sample proportions $\hat{p}\_{m}$ and $\hat{p}\_{f}$.
2. Find the difference in the observed sample proportions, $\hat{p}\_{m}-\hat{p}\_{f}$.

The distribution below shows a randomization distribution of differences in sample proportions. The mean and standard deviation of the differences are shown in the upper right corner.



1. The randomization distribution assumes that the proportion of males and females who prefer soft drinks is the same. Based on this assumption, what is the probability of obtaining random samples of size 25 that have a difference in sample proportions of 0.20 or more?
2. Is the difference in the observed sample proportions statistically significant?
3. What can we conclude about the proportion of male high school students and proportion of female high school students who prefer drinking soft drinks?