Practice Session 6

Introduction to Hypothesis Testing:

Generally speaking, hypothesis testing is a way for researchers to answer questions in a meaningful way. The methodology leverages statistical reasoning and probability to either "reject" or "fail to reject" a statement about the population as a whole. Hypothesis testing in statistics often focuses on questions about population parameters (e.g., μ, π). We do not write hypotheses in terms of the sample statistics (e.g., $\bar{x}$, $\hat{p}$). This aligns with the general theme of the course, which focuses on using samples to better understand the population.

Hypothesis testing can be used to answer questions about a single sample, or to compare samples from multiple groups. Although the calculations and assumptions may differ, the overarching framework is roughly the same in each case.

Hypothesis Testing Terms and Definitions:

Terms:

- Null Hypothesis
- Alternative Hypothesis
- P-value

Definitions:

- The probability of observing results at least as extreme as what we observed, assuming the null hypothesis is true.
- A statement of interest that represents the *status quo*, or that there is no effect or difference.
- A statement of interest that proposes there is an effect or difference.

Answers

- Null Hypothesis: A statement of interest that represents the *status quo*, or that there is no effect or difference.
- Alternative Hypothesis: A statement of interest that proposes there is an effect or difference.
- P-value: The probability of observing results at least as extreme as what we observed, assuming the null hypothesis is true.

Hypothesis Testing for a Single Proportion

Example 1: AP Multiple Choice

Does the answer choice "C" occur **less** frequently than expected? Answer this question with a hypothesis test. Use the `APMultipleChoice` data set from the `Lock5Data` library. Make sure to follow all 5 steps of hypothesis testing. *Hint:* when defining the null hypothesis, consider how likely each answer choice would be if answers were selected at random.

1.) State the null and alternative hypothesis in words:

- H0: $\pi = 0.2$
- H1: $\pi < 0.2$

2.) Calculate the observed statistic of interest

```
library(SDS1000)
library(Lock5Data)
p_hat_ap = get_proportion(APMultipleChoice$Answer, "C")
p_hat_ap
```

```
     C
0.1975
```

3.) Create a null distribution in R using `do_it()` and `rflip_count()`. Plot the null distribution, and add a vertical red line at the observed number of questions that had 'C' as the answer. Note that there were 400 multiple-choice questions in the data set.

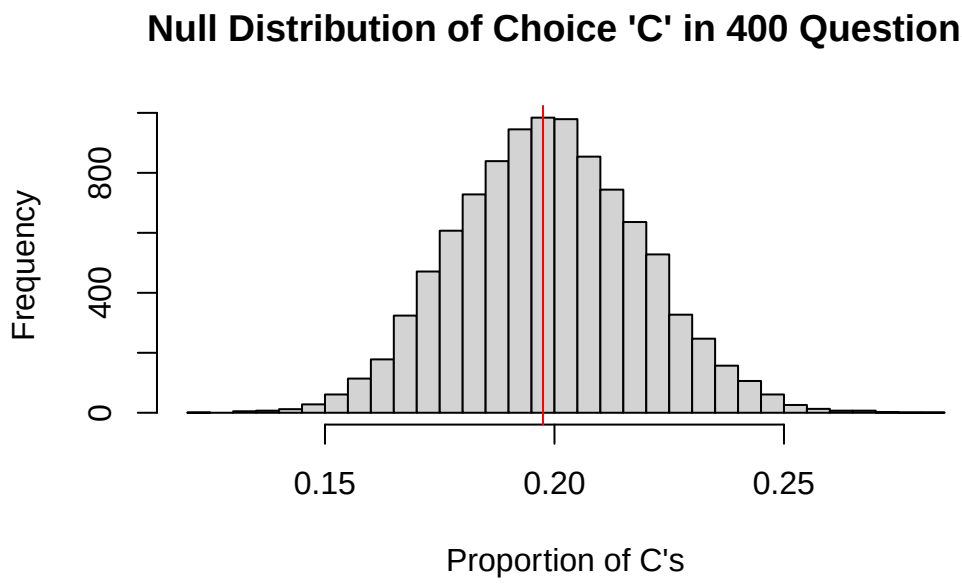
```
# Number of questions to simulate
n_questions <- 400

# Simulate the null distribution
null_dist_ap <- do_it(10000) * {
  rflip_count(n_questions, 0.2) / 400
```

```r
}

# Visualize the null distribution
hist(null_dist_ap, breaks = 50,
     main = "Null Distribution of Choice 'C' in 400 Questions",
     xlab = "Proportion of C's")

abline(v = p_hat_ap, col = "red")
```



4.) Calculate the p-value (probability of obtaining a result as or more extreme than what we observed) using the `pnull()` function.

```r
pnull(p_hat_ap, null_dist_ap, lower.tail = T)
```

```
[1] 0.4807
```

5.) Report the p-value. Make a decision on whether the results are statistically significant, and state your conclusion.

Since our p-value is greater than the standard 0.05, we will fail to reject the null hypothesis. We therefore have evidence that the answer choice 'C' is occurring as expected.

Example 2: Preferred Water Brand

A national sales bureau believes that 30% of consumers prefer Fiji water as their first choice for water. Specifically, they choose it over Aquafina, Sam's Choice, and tap water. The Fiji company claims that this proportion is greater than 30%. Using the `WaterTaste` data set from the Lock5Data library, run a hypothesis test to check this claim. Look at the `help` page for the data set to see which variable is appropriate to use.

1.) State the null and alternative hypothesis in words:

- H0: $\pi = 0.3$
- H1: $\pi > 0.3$

2.) Calculate the observed statistic of interest

```
p_hat_fiji = get_proportion(WaterTaste$First, "Fiji")
p_hat_fiji
```

```
Fiji
0.41
```

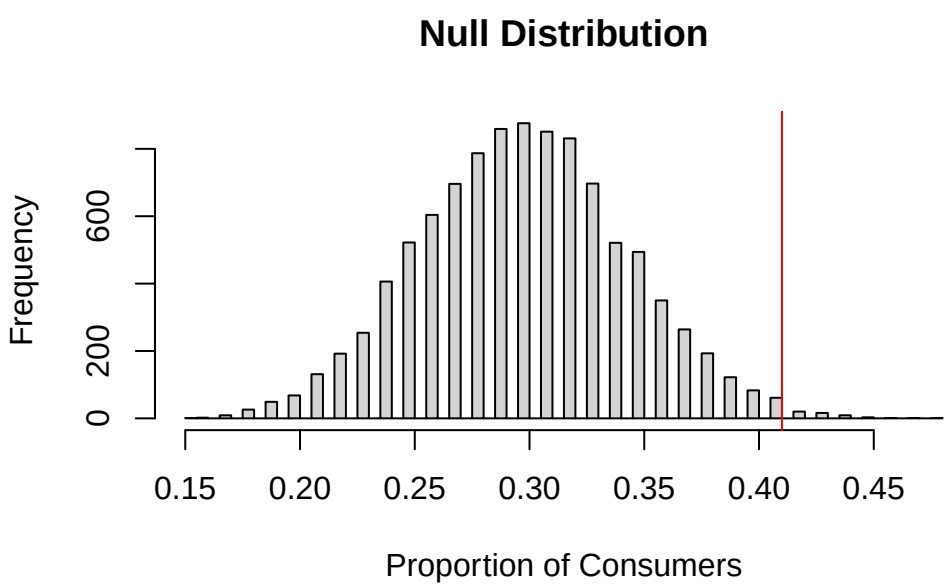
3.) Create a null distribution in R using `do_it()` and `rflip_count()`. Plot the null distribution, and add a vertical red line at your point estimate.

```
# Number of questions to simulate
n_water <- 100

# Simulate the null distribution
null_dist_fiji <- do_it(10000) * {
  rflip_count(n_water, 0.3) / 100
}

# Visualize the null distribution
hist(null_dist_fiji, breaks = 50,
     main = "Null Distribution",
     xlab = "Proportion of Consumers")

abline(v = p_hat_fiji, col = "red")
```



4.) Calculate the p-value (probability of obtaining a result as or more extreme than what we observed) using the `pnull()` function.

```
pnull(p_hat_fiji, null_dist_fiji, lower.tail = F)
```

```
[1] 0.0112
```

5.) Report the p-value. Make a decision on whether the results are statistically significant, and state your conclusion.

Since our p-value is less than the standard 0.05, we will reject the null hypothesis. We therefore have evidence that over 30% of consumers would select Fuji water as their first choice of water.

Example 3: UFO Shapes

UFOs, or Unidentified Flying Objects have become commonplace in media and popular culture. Some witnesses claim that the objects have a cylindrical shape, while others believe they are more spherical. Load the `ufo.csv` data set, and run a hypothesis test to check if over 75% of UFOs are believed to be spherical.

1.) State the null and alternative hypothesis in words:

- H0: $\pi = 0.75$
- H1: $\pi > 0.75$

2.) Calculate the observed statistic of interest (obs_stat)

```
ufo_data = read.csv("ufo_data.csv")
p_hat_sphere = get_proportion(ufo_data$shape, "sphere")
```

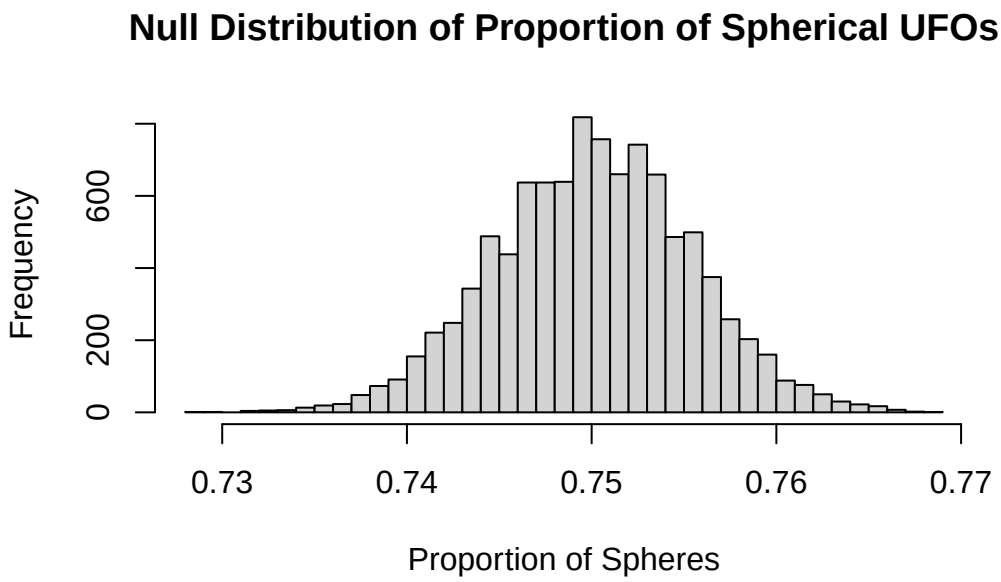
3.) Create a null distribution in R using `do_it()` and `rflip_count()`. Plot the null distribution, and add a vertical red line at your observed total number of spheres.

```
# Number of questions to simulate
n_ufos <- 6670

# Simulate the null distribution
null_dist_sphere <- do_it(10000) * {
  rflip_count(n_ufos, 0.75) / 6670
}

# Visualize the null distribution
hist(null_dist_sphere, breaks = 50,
     main = "Null Distribution of Proportion of Spherical UFOs",
     xlab = "Proportion of Spheres")

abline(v = p_hat_sphere, col = "red")
```



4.) Calculate the p-value (probability of obtaining a result as or more extreme than what we observed) using the `pnull()` function.

```
pnull(p_hat_sphere, null_dist_sphere, lower.tail = F)
```

```
[1] 0
```

5.) Report the p-value. Make a decision on whether the results are statistically significant, and state your conclusion.

Since our p-value is less than the standard 0.05, we will reject the null hypothesis. We therefore have evidence that more than half of the UFO sightings had a cylinder shape.

Exploring the `pnull()` Function

In this exercise, we will practice using some basic R coding to calculate a p-value without using the `pnull()` function.

1.) Consider null distribution and point estimate from the `AP Multiple Choice` exercise. Using the `<=` operator, write a single line of code that will show which values in the null distribution are **less than** the point estimate. Run this code and observe what the values look

like. Save this result to a vector. *Hint:* don't over think this part. Just write the inequality as you would in math class.

```
true_false_total = null_dist_ap <= p_hat_ap
```

2.) Using your vector from above, count how many values from the null distribution are **less than** the point estimate. Use the `sum()` function for this. *Hint:* the `==` operator can be used within the `sum()` function to compare values.

```
sum(true_false_total == TRUE)
```

```
[1] 4807
```

3.) Repeat the exercise above simply using the `sum()` function on your vector from part 1.). What do you notice? What happens when you run the code `sum(TRUE)`?

```
sum(true_false_total)
```

```
[1] 4807
```

4.) Using your result from part 2.), find the percentage of observations in the null distributions that are greater than the point estimate. Compare to the number from `pnull()` earlier.

```
sum(true_false_total) / 10000
```

```
[1] 0.4807
```