Project 1

STAT 109: Introductory Biostatistics

Project 1

STAT 109 Project 1: Testing a Hypothesis from Observations of a Binomial Random Process

Instructions: In this project, you will write a one-page summary of how you used the Binomial Distribution to make a conclusion about an infinite process from a dataset that you collected. For these calculations to be valid, it is important that the data you collect satisfies the 4 conditions of a Binomial Random Process. The following steps will guide you through making a plan, collecting data and analyzing the data to support the conclusion in your report.

Please follow these steps before Thursday, Feb. 19th:

1. **THINK** of a Binomial Random Process you could observe by Thursday, Feb. 19^{th}. Carefully consider each of the four conditions of a BRP until you are sure that each condition is valid for your chosen process. Please do not use a process from a class lecture example, such as the coin tosses or ESP. The four conditions are a) binary outcome in each trial (success or failure), b) independent outcomes between trials, c) fixed probability of success in a single trial ($= p$) and d) fixed number of trials ($= n$).
2. **BEFORE** you observe the process, make your best guess about whether p, the probability of one success in a single trial, is larger or smaller or just not equal to some number. Write your guess in the form of null and alternative **hypotheses**, that is,
   H_0: $p =$ **some number** and
   H_a: p ($>$ or $<$ or $\neq$) **some number**.
   Remember that your guess for p must be *some number* between 0 and 1.

Be ready to share your two hypotheses in-class Thursday, Feb. 19th so you can make a plot of what to expect (step 4) in lab.

3. **OBSERVE** n=25 trials from your chosen Binomial Random Process. How many "successes" occurred in these n trials (that is, what is the value of x)? Use these quantities to compute the observed proportion of successes out of n trials in your dataset, $\widehat{p} = x/n$.
4. Use the `dbinom` function and `geom_col` to create a bar plot of the probability of each possible value of X, that is, P(X=x) for x=0,1,2,...n for your value of n and the *some number* guess for p in H_0. Note that we can easily calculate exactly the right answer for these probabilities using `dbinom` so we use `geom_col` to make the plot, rather than `geom_bar` and a set of simulated tosses.
5. Add a vertical line to your plot in step 4 using `geom_vline(xintercept = ...)` at the value of x you observed in step 3, and then compute the p-value, i.e., the probability of observing a new value of X in n trials as extreme as the observed value of x, assuming the null hypothesis p=*some number* is true.
6. Based on the size of your p-value from step 4, is the null hypothesis consistent with the data from your observations of the process? Remember that a small p-value means that visually your data (the vertical line) is far from where you'd be most likely to see new data if your guess for the value of p is correct (the part of the x-axis with the highest probability on the y-axis).

I'll give you time in class on Thursday, Feb 19th to create your plot and your p-value.

7. Write up your study in a one-page summary with the following components:

Project 1: One Page Summary of Binomial Test — Grading Rubric

Total Points: 35

Format (2 pts)

Your summary should be organized into 4 sections (1 section with 2 subsections) with the following headings:

1. Introduction

2. Data Collection

3. Data Analysis
   - 3a. Descriptive Statistics
   - 3b. Inferential Statistics

4. Conclusion

Style (3 pts)

- Your summary should fit within one side of an 8.5 by 11 inch paper. (1 point)
- Your summary should be concise and written in complete sentences with proper grammar. (2 points)

Introduction Section (5 pts)

In this section, you should:

1. Briefly describe what question you are trying to answer. (1 point)

2. Briefly describe the motivation you have for trying to answer the question. (1 point)

3. Describe the parameter of interest p in one sentence. (1 point)

4. State the null and alternative hypotheses for your research question using appropriate notation and value for p. (2 points)

Data Collection Section (7 pts)

- In this section, you should describe how, when and where you collected data in enough detail that someone else could collect data in the same way. (3 points)
- You should also explicitly explain why you believe each of the 4 conditions of a Binomial Random Process was satisfied by your method of data collection. (4 points) This explanation should be in the context of your data collection, NOT a generic restatement of the 4 conditions.

Descriptive Statistics Subsection (6 pts)

- Please include the total number of trials n, the number of successes x and the percentage of successes you observed. (3 points)
- Please include a well-labelled pie chart or bar chart (column chart) of this data that you made in Google Sheets or Excel. (3 points)

Inferential Statistics Subsection (8 pts)

In this subsection, please include:

- A graph of the Binomial Distribution for all possible values of X and the value of p from the null hypothesis. You should create this graph as a column chart in Google Sheets after using the BINOMDIST function to compute the probability of each value of x. (4 points)
- Compute the p-value, which is the probability of seeing a new X value as extreme as your x if the null hypothesis is true. (2 points)
- Explain how you calculated this p-value. (1 point)
- Indicate where your data x is on your graph of the binomial distribution. (1 point)

Conclusion Section (4 pts)

- In this section, write a one-sentence summary of what your p-value means in the context of your research question and data. (2 points) (See my sample project for an example.)
- Based on the size of this p-value, state what you conclude the answer to your research question is. (2 points)