STAT 41700
Statistical Theory

Project 03

Task 1: For the Bernoulli distribution $X \sim Binom(1, \theta)$), figure out the following problems:

A. Obtain (simulate) a random sample of $n = 50$ observations from this distribution but with $\theta_0 = 0.2$ of 'success'. Keep it, as this will serve as your sampled data.

B. Calculate the observed sample proportion $\hat{\theta}_{obs} = \sum_1^n x_i / n$ (as the proportion of the "1" in the sample you obtained above.

C. Do you think it is reasonably good enough estimate of the value of θ_0? Explain why.

Task 2: Now Pretend that you forgot the true value of the probability of success θ_0 you used to generate the above sample of size 50. However, you are hypothesizing that $\theta_0 = 0.4$ and you would like to see whether your guess (or hypothesized value) is supported by the data you collected in Part I, or not.

Here are two ways to proceed for your choice, and which one do you think is more reasonable?

A. Compare the sample proportion you got in above Part I with your guess $\theta_0 = 0.4$. If they are reasonably close, you probably will accept your guess 0.4. Think about how could you judge the closeness.

B. Here is the other procedure. The basic logic behind this is to examine how 'extreme' or 'typical' the actual sample you obtain in Part I if the true value θ_0 were to be 0.4.
Here is the implementation.

   a) Use $\theta_0 = 0.4$, to simulate $N = 10000$ different random samples of $n = 50$ each from the $Binom(1, \theta_0 = 0.4)$ distribution.

   b) Calculate the "observed" sample proportion for each of these $N = 10000$ samples, denoted by
   $$\{\hat{\theta}_k, \quad k = 1, 2, ...10000\}$$

   c) Plot the histogram of all these $N = 10000$ sample proportions $\hat{\theta}_k$ you obtained above. This is the sampling distribution of $\hat{\theta}$ if the true value of $\theta_0 = 0.4$.

   d) To examine how 'typical' or 'extreme' is the value of $\hat{\theta}_{obs}$ you obtain in Part I, relative to this sampling distribution, calculate the probability $\hat{\theta} \le \hat{\theta}_{obs}$ using the relative frequency of $\{\hat{\theta}_k \le \hat{\theta}_{obs}, \quad k = 1, 2, ...10000\}$. This calculated probability is referred to as the **p-value. We will discuss this more in class**.