**STAT 35000**

**Introduction to Statistics**

# Project 2

# Due: November 16 (Tuesday), 2021 Name:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# Task 1: --

# The purpose of this task is to introduce you to the term *Sampling Distribution* of a statistic. We consider a random sample of independent observations drawn from a large population with (population) mean and (population) variance . Usually, these two (population) parameters and are unknown to us and we use the observations in the sample to *estimate* these parameters. A natural *estimator* for the population mean is the sample mean

# 

# and a natural *estimator* of the population variance is the sample variance

# .

# However, since the values of these two statistics depend upon the observed in values in the sample, these sample statistics are actually, also random variables. The distributions of their values (of such sample statistics) are called *Sampling Distributions*.

# You will now use R to generate (simulate), the sampling distribution of the statistic (the sample average) when sampling from a distribution called the *Exponential distribution*.

# Generate different random samples, each with observations, from the Exponential distribution, , which has mean and a standard deviation ;

# M<-1000

# n<-16

# XSamples<- replicate(M,rexp(n,rate=0.1))

# dim(XSamples)

# Calculate the sample mean of each sample (Note: Each column represents individual sample.)

# Xbars<- colSums(XSamples)/n

# Obtain the sampling distribution of , as the histogram of all these sample averages you simulated in the previous step.

* + hist(Xbars, nclass=30, freq=F, main="Sampling Distribution of Xbar when n=16")

# From this histogram you can see that this sampling distribution is relatively symmetric, almost ‘bell-shaped’. Add ‘an approximated’ density curve to the histogram above.

# 

* + dens<-density(Xbars)
  + lines(dens$x, dens$y, col=2)

# The question however is what are the mean and the standard deviation of this (sampling) distribution? Get some summary statistics about this distribution;

* + mean(Xbars)
  + var(Xbars)
  + sd(Xbars)
  + summary(Xbars)

# Let us use denote by and the men and the standard deviation of the sampling distribution of , by part e) above, for a sample of size ;

# and

# Repeat a)-f) above for , and , and summarize the result in the table below. (Please calculate the last column by hand. Then compare column 2 to 5 and column 6 to 7 to see whether the values indeed are similar.)

# Sampling Distribution of when sampling from the Exponential Distribution

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |
| 0.05 | 20 | 20 | 32 |  |  |  |
| 64 |  |  |  |
| 128 |  |  |  |
| 0.1 | 10 | 10 | 32 |  |  |  |
| 64 |  |  |  |
| 128 |  |  |  |
| 0.5 | 2 | 2 | 32 |  |  |  |
| 64 |  |  |  |
| 128 |  |  |  |
| 1 | 1 | 1 | 32 |  |  |  |
| 64 |  |  |  |
| 128 |  |  |  |

# Repeat all the steps in A. above, but now simulate the sampling distribution of the statistic when sampling from the *Normal Distribution* . You need only to replace rexp(n, rate=??) with rnorm(n, mean=??, sd=?? ) in step A) above. Summarize the result in the table below (Please calculate the last column by hand. Then compare column 1 to 4 and column 5 to 6 to see whether the values indeed are similar.)

# Sampling Distribution of when sampling from the Normal Distribution

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
| 10 | 1 | 16 |  |  |  |
| 25 |  |  |  |
| 36 |  |  |  |
| 10 | 3 | 16 |  |  |  |
| 25 |  |  |  |
| 36 |  |  |  |
| 20 | 1 | 16 |  |  |  |
| 25 |  |  |  |
| 36 |  |  |  |
| 20 | 3 | 16 |  |  |  |
| 25 |  |  |  |
| 36 |  |  |  |

# Task 2: For the Bernoulli distribution ), figure out the following problems:

1. Obtain (simulate) a random sample of observations from this distribution but with  of ‘success’.

# n<-50

# XSample<- rbinom(n,1,0.2)

1. Calculate , the observed sample proportion  (as the proportion of the “1” in the sample you obtained above.) **You will use this value for the rest of the problems in Task 2 and 3.** 
   * mean(XSample)

1. Do you think it's reasonably good enough estimate of the value of ? Explain why. Here, Please compare to to see how close is to by checking whether is within 2 standard deviation, , of the mean, for the sampling distribution of . (Note: by CLT, is approximately normal with mean and variance with .)

# Task 3: Pretend now that you forgot the true value of the probability of success you used to generate the above sample of size 50. However, you are guessing that and you would like to see whether your guess (or hypothesized value) is supported by the data you collected in Task 2, or not. Here are two ways to proceed for your choice.

1. Compare the sample proportion you got in Task 2 with your guess 0.4. If they are reasonably close, you probably will adopt your guess 0.4. Think about how could you judge the closeness. It is similar as in part C of Task 2. Please check whether is within of the mean , of the sampling distribution of , which is approximately normal with mean and variance under . This approach is related to the test for using rejection region approach.
2. Here is the other procedure. The basic logic behind this is to examine how ‘extreme’ or ‘typical’ the actual sample proportion you obtained in Part B of Task2, if the true value **** were to be ****.

Here is the implementation.

1. Use , to simulate  random samples of each from the  distribution.

# n<-50

# XSamples<- replicate(10000, rbinom(n,1,0.4))

1. Calculate the sample proportion for each of these ****samples, denoted by

****

# hatpk<- colSums(XSamples)/n

1. Plot the histogram of all these  sample proportions  you obtained above. This is the sampling distribution of if the true value of ****.

# hist(hatpk,xlab = expression(bar(X)[n]), main = "", prob = TRUE)

1. To examine how ‘typical’ or ‘extreme’ is the value of ** you obtained in part B of Task 2**, relative to this sampling distribution, calculate the probability using the relative frequency of . This calculated probability is referred to as the **p-value**.

# mean(hatpk <

# (Note: Please plug in the value obtained in part B of Task2.)

**Please plug in the value obtained in part B of Task 2 for the following parts and then complete the calculations by hand.**

# For the above part B d), we are actually using simulation to approximate . Could you figure the probability out exactly without any approximation? What is the exact probability?

# .

# Note:

# For the above probability, we are actually also be able to approximate it without simulation. Remember by the central limit theorem for n = 50 > 30,

# can be approximated by Normal distribution with mean and variance Use this approximation fact, please calculate

# Compare the p values you obtained by the above three ways (simulation approximation, exact, CLT approximation), you should expect to see that the CLT approximation is as good as the simulation approximation. There is some empirical continuity correction about this CLT approximation. Please check out the online material <https://people.richland.edu/james/lecture/m170/ch07-bin.html> to figure out how to conduct the correction to make the approximation better. Calculate the corrected probability to earn 10 bonus points.

# continuity correction

# 