**STAT 35000**

**Introduction to Statistics**

# Project 2

# Due: April 10 (Wednesday), 2024 Name: Raja Allmdar Tariq Ali

# 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)

# 

# > # \*\* Task 1 \*\*

# > # ----------------------------------------------

# > # a) Generate random samples and calculate the sample means

# > # Define the parameters

# > M <- 1000 # Number of samples

# > n <- 16 # Number of observations per sample

# > lambda <- 0.1 # Rate parameter for the exponential distribution

# > # Generate random samples

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

# > dim(XSamples)

# [1] 16 1000

# 

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

# Xbars<- colSums(XSamples)/n

**> # b) Calculate the sample mean of each 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")

**> # c) Obtain sampling distribution sample mean**

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

A graph of a number of bars

Description automatically generated

# 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)

**> # d) Adding approximate density curve to the histogram above**

**> dens <- density(Xbars)**

**> lines(dens$x, dens$y, col=2)**

A graph of a distribution of data

Description automatically generated

# 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)

**> # e) Getting some summary statistics about distribution**

**> # Calculate and display summary statistics**

**> mean\_Xbars <- mean(Xbars)**

**> var\_Xbars <- var(Xbars)**

**> sd\_Xbars <- sd(Xbars)**

**> # Display the results**

**> cat("Mean of Xbars:", mean\_Xbars, "\n")**

**Mean of Xbars: 9.979396**

**> cat("Variance of Xbars:", var\_Xbars, "\n")**

**Variance of Xbars: 6.169144**

**> cat("Standard Deviation of Xbars:", sd\_Xbars, "\n")**

**Standard Deviation of Xbars: 2.483776**

**> summary(Xbars)**

**Min. 1st Qu. Median Mean 3rd Qu. Max.**

**4.098 8.208 9.793 9.979 11.401 21.337**

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

# 9.979396 and 2.483776

# Repeat a)-f) above for , and , and summarize the result from part f) in the table below. You do not need to copy and paste the output from part a)-e). (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 | 20.03449 | 3.516455 | 3.535533 |
| 64 | 20.13281 | 2.593771 | 2.500000 |
| 128 | 20.02375 | 1.869171 | 1.767767 |
| 0.1 | 10 | 10 | 32 | 10.05886 | 1.780232 | 1.767767 |
| 64 | 9.960483 | 1.27267 | 1.25 |
| 128 | 10.01351 | 0.8927119 | 0.88388 |
| 0.5 | 2 | 2 | 32 | 1.987486 | 0.365035 | 0.353553 |
| 64 | 1.985633 | 0.2532009 | 0.25 |
| 128 | 2.006308 | 0.17339 | 0.17677 |
| 1 | 1 | 1 | 32 | 1.002061 | 0.1836427 | 0.17677 |
| 64 | 0.9996066 | 0.1193824 | 0.125 |
| 128 | 1.005024 | 0.09057692 | 0.088388 |

# 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. Again, just summarize the result from part f) 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 | 10.00425 | 0.2552946 | 0.25 |
| 25 | 10.00401 | 0.2077019 | 0.20 |
| 36 | 10.00353 | 0.1686799 | 0.16667 |
| 10 | 3 | 16 | 10.00084 | 0.7492337 | 0.75 |
| 25 | 10.03932 | 0.6026922 | 0.60 |
| 36 | 10.00073 | 0.5015264 | 0.50 |
| 20 | 1 | 16 | 19.99229 | 0.2475279 | 0.25 |
| 25 | 19.998 | 0.2090922 | 0.20 |
| 36 | 20.00698 | 0.1630882 | 0.1667 |
| 20 | 3 | 16 | 20.0022 | 0.7465781 | 0.75 |
| 25 | 20.00035 | 0.5940372 | 0.60 |
| 36 | 20.00656 | 0.5086334 | 0.50 |

# 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)

# > # A. Simulate random sample of 50 observations from this distribution

# > n <- 50

# > p\_success <- 0.2

# > XSample <- rbinom(n, 1, p\_success)

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)

**> # B. Calculate the sample proportion of successes**

**> p\_hat\_obs <- mean(XSample)**

**> cat("Observed sample proportion, p̂\_obs:", p\_hat\_obs, "\n")**

**Observed sample proportion, p̂\_obs: 0.28**

1. Is your observed supporting or against the value of  being 0.2? Explain why. Here, please compare to to see how close is to by checking whether  is within 2 standard deviation, , of the mean, of the sampling distribution of , i.e., whether . (Note: by CLT, is approximately normal with mean and variance with .)

**> # C. Evaluating observed proportion and checking if it is within 2 s.d**

**> # Calculate the standard deviation of the sampling distribution of p̂**

**> std\_error <- sqrt((p\_success \* (1 - p\_success)) / n)**

**> # Check if p̂\_obs is within 2 standard deviations of p\_0**

**> lower\_bound <- p\_success - 2 \* std\_error**

**> upper\_bound <- p\_success + 2 \* std\_error**

**> cat("Lower bound:", lower\_bound, "\n")**

**Lower bound: 0.08686292**

**> cat("Upper bound:", upper\_bound, "\n")**

**Upper bound: 0.3131371**

**> # Determine if p̂\_obs supports p\_0 = 0.2**

**> if (p\_hat\_obs >= lower\_bound && p\_hat\_obs <= upper\_bound) {**

**+ cat("The observed p̂\_obs supports the value of p being 0.2.\n")**

**+ } else {**

**+ cat("The observed p̂\_obs does not support the value of p being 0.2.\n")**

**+ }**

**The observed p̂\_obs supports the value of p being 0.2.**

# 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 , i.e., whether. This approach is related to the test for using rejection region approach.

**> # ----------------------------------------------**

**> # \*\* Task 3 \*\***

**> # ----------------------------------------------**

**> # A. Compare Pobs with P\_0 = 0.4**

**> # New hypothesized probability of success**

**> p\_new\_success <- 0.4**

**> # Calculate the new standard deviation of the sampling distribution of p̂**

**> std\_error\_new <- sqrt((p\_new\_success \* (1 - p\_new\_success)) / n)**

**> # Check if p̂\_obs is within 2 standard deviations of the new p\_0**

**> lower\_bound\_new <- p\_new\_success - 2 \* std\_error\_new**

**> upper\_bound\_new <- p\_new\_success + 2 \* std\_error\_new**

**> cat("New lower bound:", lower\_bound\_new, "\n")**

**New lower bound: 0.2614359**

**> cat("New upper bound:", upper\_bound\_new, "\n")**

**New upper bound: 0.5385641**

**> # Determine if p̂\_obs is reasonably close to the new hypothesized value p\_0 = 0.4**

**> if (p\_hat\_obs >= lower\_bound\_new && p\_hat\_obs <= upper\_bound\_new) {**

**+ cat("The observed p̂\_obs might be considered reasonably close to p\_0 = 0.4.\n")**

**+ } else {**

**+ cat("The observed p̂\_obs is not reasonably close to p\_0 = 0.4.\n")**

**+ }**

**The observed p̂\_obs might be considered reasonably close to p\_0 = 0.4.**

1. 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))

# > # B. How typical or extreme p\_obs is

# > # Simulate 10,000 random samples of size 50 each from Bernoulli distribution with p = 0.4

# > n <- 50

# > XSamples\_new <- replicate(10000, rbinom(n, 1, p\_new\_success))

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

****

# hatpk<- colSums(XSamples)/n

# > # b) Calculate the sample proportion for each sample

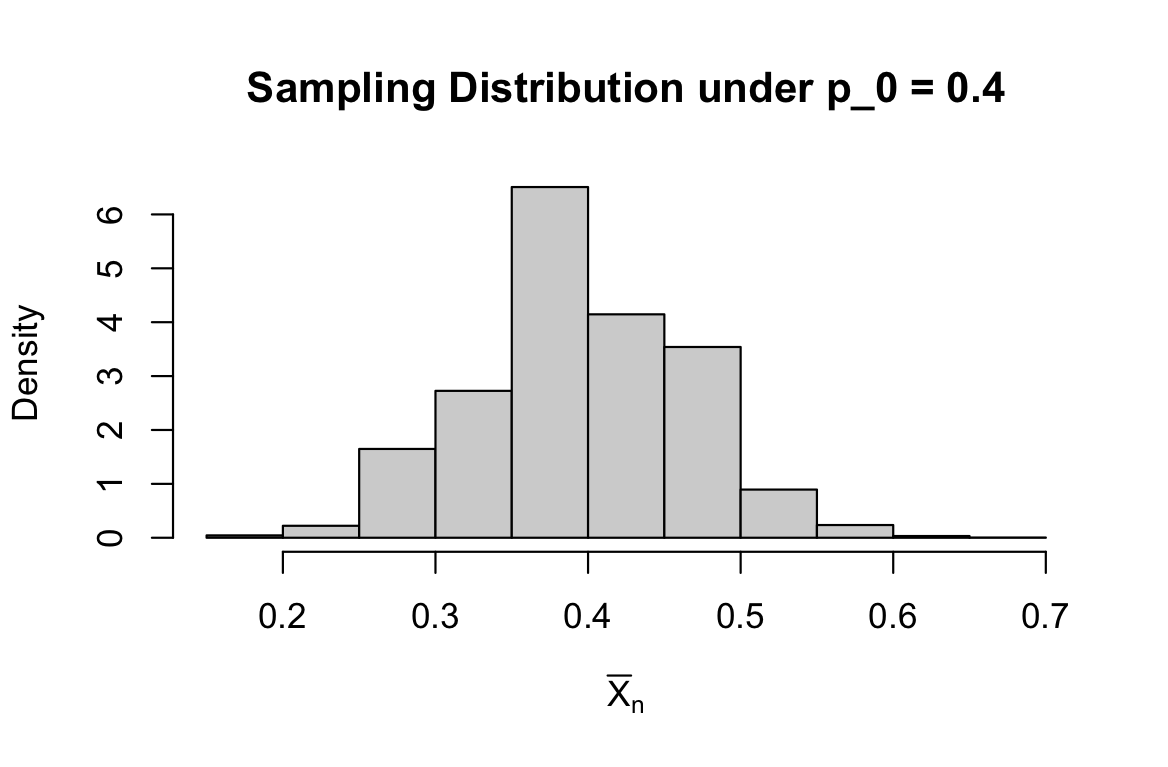
# > hatpk <- colSums(XSamples\_new) / n

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)

# > # c) Plot the histogram of these sample proportions

# > hist(hatpk, xlab=expression(bar(X)[n]), main="Sampling Distribution under p\_0 = 0.4", 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**. The smaller p-value is, the more ‘extreme’ is and more significant the evidence is to against .

# mean(hatpk <=

# (Note: Please plug in the value obtained in part B of Task2 when you write the R code.)

# > # d) Calculate the p-value: the proportion of simulated proportions ≤ observed proportion

# > p\_value <- mean(hatpk <= p\_hat\_obs) # p̂\_obs: 0.28

# > cat("P-value:", p\_value, "\n")

# P-value: 0.0529

**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?

# n = 50 (sample size)

# p\_0 = 0.4 (hypothesized probability of success)

# p\_obs = 0.28 (observed sample proportion)

# X = 50 \* 0.28 = 14

# 

# .

# 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

# Z = X - / sigma =

# 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 also some empirical continuity correction about this CLT approximation. Please conduct the continuity correction to make the approximation better. Calculate the corrected probability to earn 10 bonus points.

# Xcorrected = X + 0.5 = 14.5

# Z(with Xcorrected) = -1.5877

# continuity correction

# 