PROJECT 3

Abstract

The scope of this project is dealing with Chapter 11 of Bain, L.J. and Engelhardt M. Introduction to Probability and Mathematical Statistics 1992.

Question 1:

1. Problem Statement:

Discuss the use of the pivotal quantity and bootstrap methods for obtaining confidence intervals for a population parameter θ, considering the following outcomes:

1.1 What does a confidence interval with a confidence level of γ represent:

In statistics, a confidence interval (CI) is a type of estimate computed from the observed data. This gives a range of values for an unknown parameter (for example, a population mean). The interval has an associated confidence level chosen by the investigator. For a given estimation in a given sample, using a higher confidence level generates a wider (i.e., less precise) confidence interval. In general terms, a confidence interval for an unknown parameter is based on sampling the distribution of a corresponding estimator. This means that the confidence level represents the theoretical long-run frequency (i.e., the proportion) of confidence intervals that contain the true value of the unknown population parameter. Wikipedia (2021)

1.2 Which requirements should be met to determine a confidence interval for θ with the pivotal quantity method:

Firstly: $Q = q(X_1, ..., X_n; \theta)$ should be a function of $\{\underline{x}, \underline{\theta}\}$ only.

Secondly: PDF of Q should not depend on θ or any other unknown parameters.

Then we can go and apply the Pivotal Quantity Method according to *Definition 11.3.1* from Notes (2021).

1.3 Which steps should be followed to determine a confidence interval for θ with the pivotal quantity method:

STEP 1: Determine the distribution of our pivotal quantity and insure that it does not depend on θ or any other unknown parameters.

STEP 2: Take note of our tables and for which distribution we can find percentiles for readily. This will generally entail applying a transformation as per Chap 8 of Notes (2021) to get our pivotal quantity in either a χ^2, F or t distribution such that we can use our tables.

STEP 3: Setup the confidence interval as is required, taking note of the type: equal tailed or single tailed which can be an one sided upper or lower confidence limit.

STEP 4: With the pivotal quantity in the middle of the inequality reduce it such the the parameter for which the confidence interval is required is left sandwiched between two values which will be our resultant confidence interval.

1.4 Why do we use bootstrapping in statistics:

Since many things can affect how well a sample reflects the population; and therefore, how valid and reliable the conclusions will be, we need a way to mitigate this - Joseph (2020).

Bootstrapping is a statistical procedure that resamples a single data set to create many simulated samples. This process allows for the calculation of standard errors, confidence intervals, and hypothesis testing - Frost (2020).

A bootstrapping approach is an extremely useful alternative to the traditional method of hypothesis testing as it is fairly simple and it mitigates some of the pitfalls encountered within the traditional approach. As with the traditional approach, a sample, say S, of size n is drawn from the population within the bootstrapping approach. Now, rather than using theory to determine all possible estimates, the sampling distribution is created by resampling observations with replacement from S, m times, with each resampled set having n observations. Now, if sampled appropriately, S should be representative of the population. Therefore, by resampling S m times with replacement, it would be as if m samples were drawn from the original population, and the estimates derived would be representative of the theoretical distribution under the traditional approach - Joseph (2020).

Another advantage of bootstrapping is that we can calculate confidence intervals for statistics other than the mean - WiseStatistics (2018).

1.5 How do we obtain bootstrap samples:

A bootstrap sample is a smaller sample that is "bootstrapped" from a larger sample. Bootstrapping is a type of resampling where large numbers of smaller samples of the same size are repeatedly drawn, with replacement, from a single original sample. Bootstrapping is loosely based on the law of large numbers, which states that if you sample over and over again, your data should approximate the true population data. The process generally follows three steps - Glen (2016):

1) Resample a data set x times,

2) Find a summary statistic (called a bootstrap statistic) for each of the x samples,

3) Estimate the standard error for the bootstrap statistic using the standard deviation of the bootstrap distribution.

1.6 Which steps should be followed to obtain a bootstrap confidence interval using the percentile method:

With the percentile method, a certain percentage (e.g. 5% or 10%) is trimmed from the lower and upper end of the sample statistic (e.g. the mean or standard deviation). Which number you trim depends on the confidence interval you're looking for. For example, a 90% confidence interval would generate a 100% – 90% = 10% trim (i.e. 5% from both ends) - Glen (2016).

A bootstrap percentile confidence interval of $\widehat{\theta}$,an estimator of θ, can be obtained as follows:

1) Generate B number of random bootstrap samples,
2) Calculate a parameter estimate from each bootstrap sample,
3) Order all B bootstrap parameter estimates from the lowest to highest,
4) Construct the confidence interval: $[\widehat{\theta}_{lower\,limit}, \widehat{\theta}_{upper\,limit}]$=$[\widehat{\theta}^*_j, \widehat{\theta}^*_k]$ such that $\widehat{\theta}^*_j$ denotes the j'th quantile (the lower limit) and $\widehat{\theta}^*_k$ denotes the k'th quantile (the upper limit). Also $j = [\frac{\alpha}{2} \times B]$ and $k = [(1 - \frac{\alpha}{2}) \times B]$. Thus, a 95% percentile bootstrap CI with 1,000 bootstrap samples is the interval between the 25th quantile value and the 975th quantile value of the 1,000 bootstrap parameter estimates - Jung et al. (2019).

1.7 What are the advantages and disadvantages of the bootstrap percentile method, by referring to the sample size and symmetry of the statistic of interest:

Advantages:

1) As the sample size increases, bootstrapping converges on the correct sampling distribution under most conditions - Frost (2020).
2) Bootstrapping does not make assumptions about the distribution of your data - Frost (2020).
3) You can use bootstrapping for a wider variety of distributions, unknown distributions, and smaller sample sizes. Sample sizes as small as 10 can be usable - Frost (2020).

Disadvantages:

1) Resampling involves reusing your one data set many times under the central assumption that the original sample accurately represents the actual population, which may not always be the case - Frost (2020).
2) The choice of the percentage trimmed from the upper and lower end of the sample statistic could have an effect on the simulation.
3) There are several conditions when bootstrap percentile method is not appropriate, such as when the population variance is infinite, or when the population values are discontinuous at the median - Frost (2020).

Question 2:

1. Problem Statement:

Consider a random sample, $\{X_1, ..., X_n\}$, with $X_i \sim N(\mu, \sigma^2)$ $\forall i = 1, ..., n$ with both population parameters $\{\mu, \sigma^2\}$ are unknown. Design a simulation study to empirically show that the confidence intervals for both $\{\mu, \sigma^2\}$, as given by Theorems 11.3.1b and 11.3.1c, as in Notes (2021), are indeed $100(1-\alpha)\%$ confidence regions. Consider a small and large sample size. You can choose the population parameters.

2. Methodology and Design of Simulation Study:

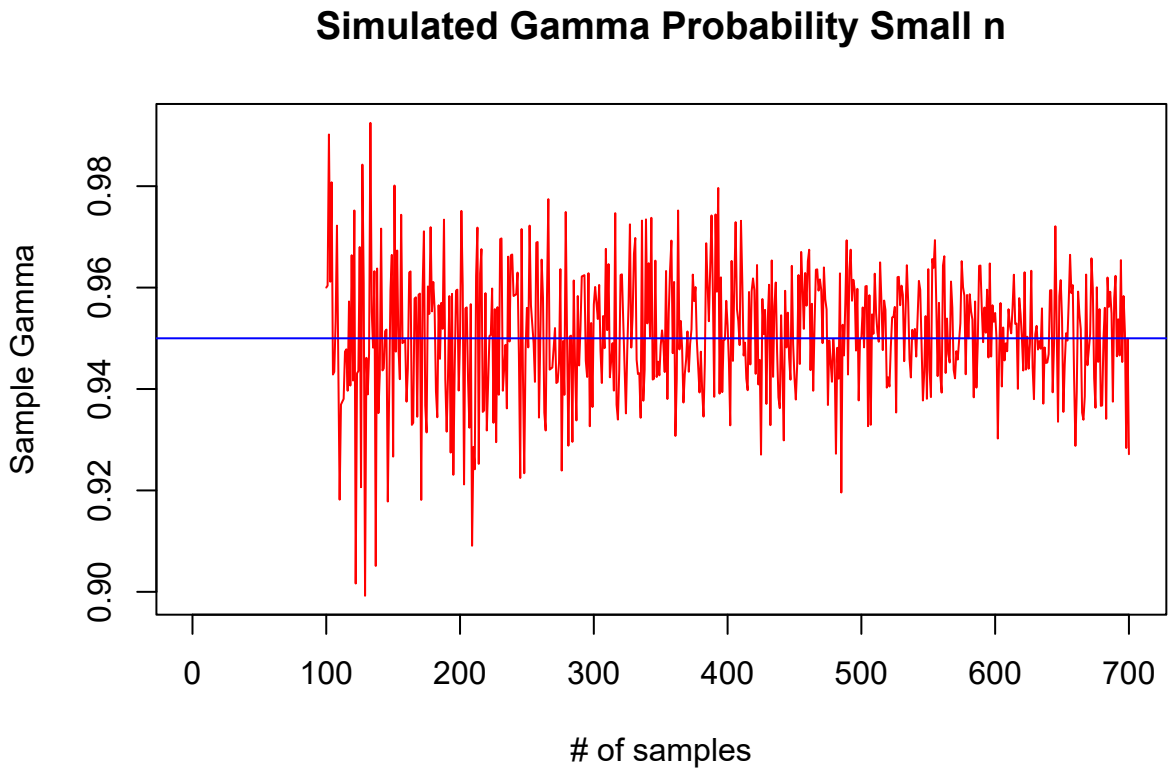
The methodology behind the simulation study is as follows:

1) I will generate a population of normally distributed random deviates.
2) Simulate sampling of various sizes from a size of 100 to a size of 700 for both a large and small value of n in our population in 1.
3) Calculate the required confidence interval based on the theory of Theorem 11.3.1b and c
4) Calculate the proportion of the population mean or population variance that fall within these bounds an plot the results
5) Show that this proportion results in the desired 100(1-α)% confidence level

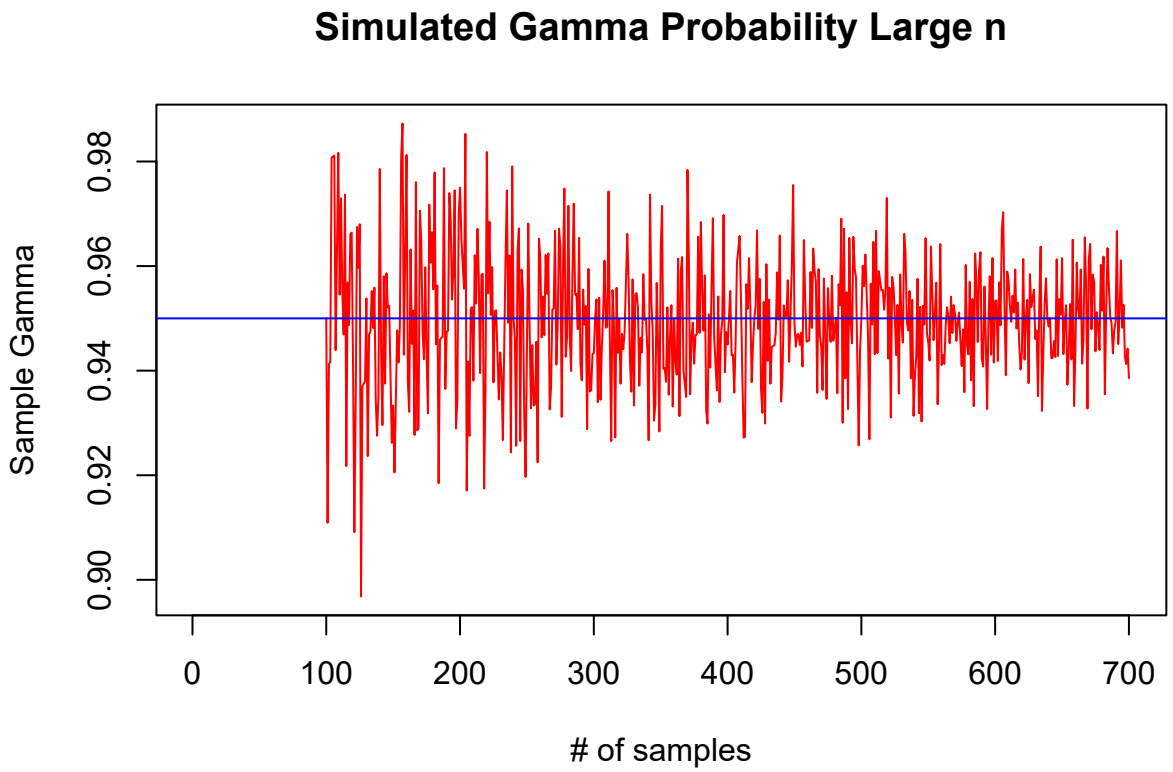
Simulating a 95% confidence interval for μ:

```
alpha=0.05
popMean = 3
popVar = 4

#SIMULATING CI FOR THE MEAN
############################
lower = c()
upper = c()
test = c()
prop=0

propFun = function (n,k){
  for (i in 1:n) {
    samples <- rnorm(k, popMean, popVar)
    lower[i] <- mean(samples) - qt(1-alpha/2, df= k-1) * sqrt(var(samples))/sqrt(k)
    upper[i] <- mean(samples) + qt(1-alpha/2, df= k-1) * sqrt(var(samples))/sqrt(k)

    test[i] = between(popMean, lower[i], upper[i])
    prop = sum(test)/n
  }
```

```
  return(prop)
}
simProp1 = c()
simProp2 = c()
for (i in 100:700) {
  simProp1[i] = propFun(i,50)
  simProp2[i] = propFun(i,5000)
}
plot(simProp1,type = "l", col = "red", xlab = "# of samples", ylab = "Sample Gamma",
     main = "Simulated Gamma Probability Small n")
abline(h=0.95, col="blue")
```



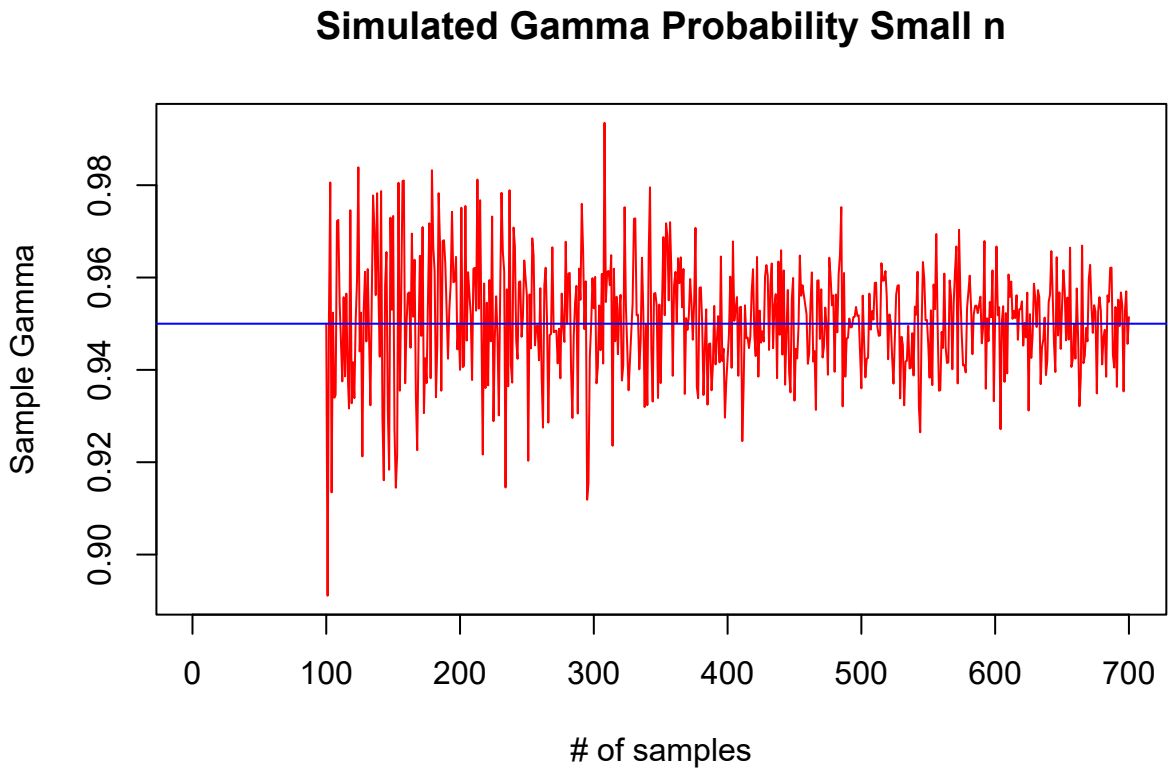

```
plot(simProp2,type = "l", col = "red", xlab = "# of samples", ylab = "Sample Gamma",
     main = "Simulated Gamma Probability Large n")
abline(h=0.95, col="blue")
```



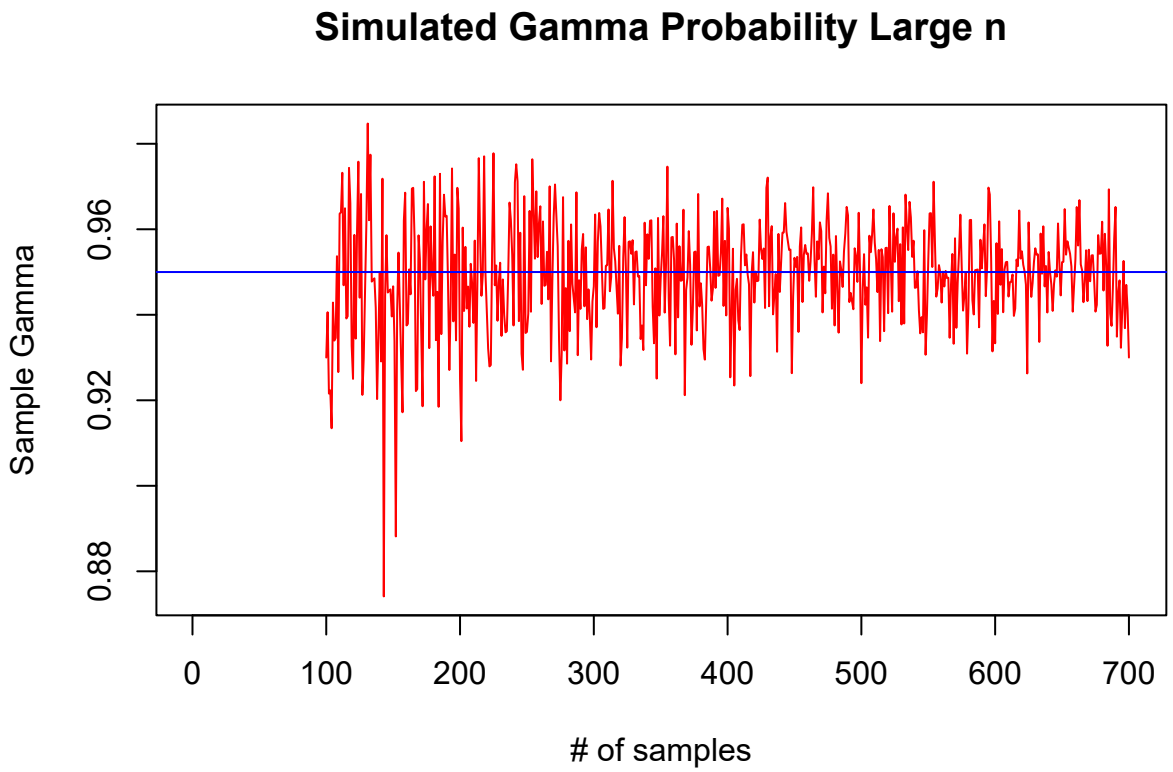
Simulating a 95% confidence interval for σ^2:

```
alpha=0.05

popMean = 2
popVar = 4

#SIMULATING CI FOR THE VARIANCE
################################
lower = c()
upper = c()
test = c()
sampleVar = c()
prop=0

propFun = function (n,k){
  for (i in 1:n) {
    samples = rnorm(k, popMean, sqrt(popVar))
    lower[i] = (k-1)*var(samples)/qchisq(0.975, df= k-1)
    upper[i] = (k-1)*var(samples)/qchisq(0.025, df= k-1)

    test[i] = between(popVar, lower[i], upper[i])
    prop = sum(test)/n
  }
```

```
  return(prop)
}

simProp1 = c()
simProp2 = c()
for (i in 100:700) {
  simProp1[i] = propFun(i,50)
  simProp2[i] = propFun(i,5000)
}
plot(simProp1,type = "l", col = "red", xlab = "# of samples", ylab = "Sample Gamma",
     main = "Simulated Gamma Probability Small n")
abline(h=0.95, col="blue")
```




```
plot(simProp2,type = "l", col = "red", xlab = "# of samples", ylab = "Sample Gamma",
      main = "Simulated Gamma Probability Large n")
abline(h=0.95, col="blue")
```



3. Discussion fo Results Obtained:

In both cases for μ and σ^2 we saw that through the simulation with the confidence interval limits obtained in Theorem 11.3.1b and c that we do indeed get the desired $100(1-\alpha)\%$ confidence region. We also see that as we go from a small to a large sample size, that the variability is less for a large sample size as can be seen in the graphs above. Also it is evident that the envelope within which the simulated value lie is much tighter in the case of a larger value fo n.

References

Frost, Jim. 2020. "*Introduction to Bootstrapping in Statistics with an Example.*". http://statisticsbyjim.com/hypothesis-testing/bootstrapping [Accessed: 2021/10/27].

Glen, Stephanie. 2016. "*Bootstrap Sample: Definition, Example.*". http://www.statisticshowto.com/bootstrap-sample/ [Accessed: 2021/10/27].

Joseph, Trist'n. 2020. "*Bootstrapping Statistics. What it is and why it's used.*". http://towardsdatascience.com/bootstrapping-statistics-what-it-is-and-why-its-used-e2fa29577307 [Accessed: 2021/10/27].

Jung, Lee, Gupta and Cho. 2019. "*Comparison of Bootstrap Confidence Interval Methods for GSCA Using a Monte Carlo Simulation.*". http://www.frontiersin.org/articles/10.3389/fpsyg.2019.02215/full#h3 [Accessed: 2021/10/27].

Wikipedia. 2021. "*Confidence interval*.". http://en.wikipedia.org/wiki/Confidence_interval [Accessed: 2021/10/25].

WiseStatistics. 2018. "*WISE Statistics: Introduction to Bootstrapping*.". http://www.youtube.com/watch?v=G8sUVMoAj7s [Accessed: 2021/10/27].

.