Computer Project 2: Simulations for Central Limit Theorem & Bootstrap

Math 324: Probability and Statistics with Computing

Due Thursday 11/8 by 10 pm (submit online; no late submissions allowed; contributes to the 20% computer project grade)

Purpose

1. Satisfy the computing aspect of this course.
2. Solidify understanding of the CLT (central limit theorem) by simulating data and seeing it happen.
3. Learn the statistical bootstrap method for estimating variance by doing it.

Please submit the pdf generated by your RMD file. Show all R code and output.

Central Limit Theorem Simulations

Exercise 1. Generate 500 sample means for the following distributions:

I) U(0,5) (uniform from $A = 0$, to $B = 5$)
II) Binomial with # trials $= 15$, $p = .2$
III) Exp(5) (exponential with $\lambda = 5$)
IV) Poisson with $\mu = 2$

by coding up an algorithm with the following steps for sample sizes $n = 5$ and $n = 50$:

Algorithm 1 Generate 500 Sample Means from Sample Size n and for a Particular Distribution

```
1: Initialize a vector (1-dimensional array) to store 500 sample means
2: for all 500 repetitions do
3:     Generate random sample of size $n$ for the particular distribution (see I-IV)
4:     Take the average over the $n$ sampled values.
5:     Store the average into your vector from Step 1.
6: end for
```

In total, you will repeat Algorithm 1 for (4 distributions)(2 sample sizes) = 8 settings.

Here are some pseudo R code snippits and functions that may help you implement Algorithm 1.

```
## Generating samples from distributions
runif(<sample_size>, min = <A>, max = <B>) # uniform

rbinom(<sample_size>,
       size = <number_trials>, prob=<p>) # binomial

rexp(<sample_size>, rate = <lambda>) # exponential
```

```r
rpois(<sample_size>, <mu>) # Poisson


## Using a for loop: my example will just store the loop number from 1 to 100 in
## a vector (1-dimensional array) using a for loop. You will be doing something
## else with your for loop.

nReps = 100

# Initalizes a vector of 100 "NA" values which will be replaced with actual
# numbers when we run the loop
storedData = rep(NA, nReps)

# The loop itself
for (i in 1:nReps){
  # Saves the loop index into storedData vector
  storedData[i] = i
}

# Now I have a vector of numbers from 1-100. Your vector should contain 500
# sample means instead. Then use your vector to answer a-d below.
```

For each of the distributions in 1 - 4, please answer the following questions:

a. What is the average of the 500 sample means when the sample size is $n = 5$? What is the average of the 500 sample means when the sample size is $n = 50$? What are the theoretical expected values of sample means, respectively?
b. For $n = 5$ and $n = 50$, what are the variances of the 500 sample means, respectively? What are the theoretical variances of sample means, respectively?
c. Construct histogram for sample means for $n = 5$ and $n = 50$, respectively.
d. Construct normal probability plot of sample means for $n = 5$ and $n = 50$, respectively.

You will need these functions for the probability plot:

```r
# Functions to generate probability plot and straight line for it, ideas we
# discussed on W10D2.
qqnorm(<your_vector>) # the plot
qqline(<your_vector>) # the line
```

e. After answering parts a-d for all distributions I-IV, you should see some common trends across distributions. Summarize these findings for each of a-d, and use the central limit theorem to explain your findings.

(Please see page 3 for second exercise.)

Bootstrap: Estimating the Variance of a Statistic

Short introduction:

- The bootstrap method is implemented by sampling **with replacement** many, many times from a single sample.
- It is quite useful in real life for figuring out how a complicated statistic varies when you don't actually have a theoretical probability model (like we do for uniform, binomial, etc).
- The basic bootstrap gives a decent estimator of the variance long as your data is sampled independently and identically (the data come from the same distribution), and as long as you take enough bootstrapped samples. The bootstrap is considered quite computationally demanding.

Extended:

Why is it important to estimate variance and why is the bootstrap useful? If you produce a number from a sample in a complicated way (perhaps you build a statistical model to predict how long users will stay on your website and use a sample of users to get a prediction) and then wonder how reliable that number is (e.g. is this really how long most users stay?) but don't have a theoretical formula, you could use the bootstrap method to get a general idea. Intuitively, you want to figure out the variability of your statistic because if you had a different sample, and your statistic was dramatically different, then whatever number you produce from a specific sample would be useless since you want a generalization about the population (e.g. you want to know about what most users do, not what this specific sample does). If your statistic would be about the same regardless of sample, then the specific number is very helpful to you (e.g. from the sample, you can tell that most users are spending this much time on the site). The bootstrap method will estimate the variance for your statistic, and hopefully, that variance will be low so that you can generalize from sample to population.

Exercise 2.

Bootstrap a variance estimate for the sample mean of the binomial distribution with $p = .2$ and sample size $n = 50$ by coding up an algorithm with the following steps:

Algorithm 2 Use 40,000 Bootstrapped Means to Estimate Variance of the Mean

```
1: Initialize a vector to store 40,000 bootstrapped means
2: Generate random sample of size $n = 50$ for the binomial distribution with $p = .2$.
3: for all 40,000 repetitions do
4:     Resample $n = 50$ with replacement from the vector in Step 2.
5:     Take the average over the 50 sampled values.
6:     Store the average into your vector from Step 1.
7: end for
8: Calculate the variance of 40,000 bootstrapped means.
```

a. Report the variance from Step 8 and consider the theoretical variance you got in Exercise 1-II-b for $n = 50$. Is your bootstrapped estimator for the variance on the right scale? (That is, if the theoretical variance is somewhere in 1-2, is the boostrapped variance also somewhere in 1-2?)
b. If you actually could compute the theoretical variance, as you can for a sample mean of a binomial distribution, would you prefer to bootstrap an estimated variance for that mean as we did here or calculate the theoretical variance as we did in Exercise 1? Why? (Think about computational cost and accuracy.)

One more R function you'll need (you've already used it before but make sure "replace" is set to true)

```
sample(<your_vector_of_50_binomial_random_numbers>, sample = 50, replace = TRUE)
```