Intro to Data Science - Lab 4

IST687 Section M003

Professor Anderson

Enter your name here: Hrishikesh Telang

#Select one of the below and add needed information # 1. I did this homework by myself, with help from the book and the professor. # 2. I did this homework with help from the book and the professor and these Internet sources: # 3. I did this homework with help from but did not cut and paste any code.

#The problem is this: You receive a sample containing the ages of 30 students. You are wondering whether this sample is a group of undergraduates (mean age = 20 years) or graduates (mean age = 25 years). To answer this question, you must compare the mean of the sample you receive to a distribution of means from the population. The following fragment of R code begins the solution:

```
set.seed(2)
sampleSize <- 30
studentPop <- rnorm(20000,mean=20,sd=3)
undergrads <- sample(studentPop,size=sampleSize,replace=TRUE)
grads <- rnorm(sampleSize,mean=25,sd=3)
if (runif(1)>0.5) { testSample <- grads } else { testSample <- undergrads }
mean(testSample)
```

```
## [1] 24.89729
```

#After you run this code, the variable "testSample" will contain either a sample of undergrads or a sample of grads. The line before last "flips a coin" by generating one value from a uniform distribution (by default the distribution covers 0 to 1) and comparing it to 0.5. The question you must answer with additional code is: Which is it, grad or undergrad?

#Here are the steps that will help you finish the job: #1. Annotate the code above with line-by-line commentary. In other words, you must explainwhat each of the six lines of code above actually do! You will have to lookup the meaning of some commands.

#2. Generate 10 samples from the "undergrads" dataset.

```
sample(undergrads, size=10, replace=TRUE)
```

```
##  [1] 14.13022 25.57921 25.57921 25.92212 13.99717 19.87687 21.48056 24.82276
##  [9] 25.57921 18.73168
```

#3. Generate 10 new samples and take the mean of that sample

```
sample_data <- sample(undergrads, size=10, replace=TRUE)
mean(sample_data)
```

```
## [1] 19.24712
```

#4. Repeat this process 3 times (i.e., generate a sample and take the mean 3 times, using the replicate function).

```
replicate(3, mean(sample(undergrads, size=10, replace=TRUE)), simplify=TRUE)
```

```
## [1] 19.94657 18.68297 18.99601
```

#End of the first breakout:

#5. Generate a list of sample means from the population called "undergrads" #How many sample means should you generate? Really you can create any number that you want – hundreds, thousands, whatever – but I suggest for ease of inspection that you generate just 100 means. That is a pretty small number, but it makes it easy to think about percentiles and ranks.

```
richie <- replicate(100, mean(sample(undergrads, size=10, replace=TRUE)), simplify=TRUE)
richie
```

```
##   [1] 20.36587 21.10507 19.02457 18.28870 17.27890 19.89785 19.66068 19.94193
##   [9] 20.68523 21.08060 20.07199 17.73030 21.30522 21.54053 20.53283 19.78731
##  [17] 19.05615 18.97547 18.13302 18.50421 17.97375 19.38756 20.02094 19.37616
##  [25] 19.56847 19.25992 19.10462 19.30592 17.59697 19.85861 21.78712 20.02088
##  [33] 20.81500 18.65394 19.57369 19.00040 18.96819 21.15763 19.05183 20.58691
##  [41] 19.29047 18.87327 20.93982 17.40055 18.96881 20.19355 20.17649 21.81700
##  [49] 19.37695 21.12508 19.15978 20.39494 19.98219 20.04858 21.05120 20.68634
##  [57] 20.33131 20.29818 19.41035 17.80218 18.30865 19.80892 19.72218 20.90138
##  [65] 20.37560 20.31163 17.68896 18.47158 20.26910 20.43995 18.92449 21.11918
##  [73] 20.05361 19.70654 18.69019 18.13072 19.95182 18.25051 18.70488 19.93355
##  [81] 18.72291 19.39856 18.11499 19.52592 19.51070 18.10659 17.82630 20.41680
##  [89] 19.56717 20.08732 18.23101 19.10227 18.61019 19.48643 18.99633 19.33597
##  [97] 19.06984 20.52165 19.94312 18.96525
```

#6. Once you have your list of sample means generated from undergrads, the trick is to compare mean(testSample) to that list of sample means and see where it falls. Is it in the middle of the pack? Far out toward one end? Here is one hint that will help you: In chapter 7, the quantile() command is used to generate percentiles based on thresholds of 2.5% and 97.5%. Those are the thresholds we want, and the quantile() command will help you create them.

```
quant <- quantile(richie, probs=c(0.025,0.975))
quant
```

```
##     2.5%    97.5%
## 17.64066 21.42876
```

#7. Your code should have a print() statement that should say either, "Sample mean is extreme," or, "Sample mean is not extreme."

```
if (mean(undergrads) < quant[1] || mean(undergrads) > quant[2]) {
  print("Sample mean is extreme")
} else {
  print("Sample mean is not extreme")
}
```

```
## [1] "Sample mean is not extreme"
```

#8. Add a comment stating if you think the testSample are undergrad students. Explain why or why not.

```
mean(undergrads)
```

```
## [1] 19.52719
```

```
#I personally think that the testSample are undergrad students because the mean sample is between 17 an
```

#9. Repeat the same analysis to see if the testSample are grad students. Learning Goals for this activity: #A. Generate random numbers in a normal distribution and assign a variable name. #B. Understand and use a conditional statement. #C. Reason about percentiles. #D. Reason about distributions of sample means. #E. Use R code to report the results of a test.

```
sample(grads, size=10, replace=TRUE)
```

```
##  [1] 29.34834 22.42978 24.14025 27.08501 25.64208 25.35534 20.88695 27.08501
##  [9] 25.04493 27.96066
```

```
sample_data <- sample(grads, size=10, replace=TRUE)
mean(sample_data)
```

```
## [1] 24.9171
```

```
replicate(3, mean(sample(grads, size=10, replace=TRUE)), simplify=TRUE)
```

```
## [1] 25.48172 25.00178 24.50377
```

```
richie2 <- replicate(100, mean(sample(grads, size=10, replace=TRUE)), simplify=TRUE)
richie2
```

```
##   [1] 24.66289 25.33433 24.47164 23.77301 22.90149 25.02388 25.70716 25.83604
##   [9] 25.32001 25.90285 23.85644 26.06596 25.47103 24.39127 25.55406 24.40172
##  [17] 25.54636 24.68357 25.77465 24.77859 25.19190 24.71782 23.68625 25.69709
##  [25] 25.85468 24.61507 24.66775 25.00217 24.54938 25.94769 24.66288 24.59672
##  [33] 24.53149 25.56060 24.67279 27.27502 25.22744 24.07979 27.49211 25.50489
##  [41] 25.36124 26.77254 24.05500 24.61908 25.30644 24.63031 25.49847 24.79611
##  [49] 25.17777 25.35507 26.19351 24.91199 24.44436 24.34440 25.05795 24.27256
##  [57] 24.24062 26.25192 25.80106 25.54287 26.94533 24.13525 25.86716 23.91292
##  [65] 25.26907 24.93553 25.28057 25.12139 26.47143 24.85560 26.61996 24.80568
##  [73] 23.18602 25.30736 24.31846 24.98478 23.94598 25.36377 24.50047 25.93096
##  [81] 24.41323 24.17038 24.10524 24.35074 25.00484 25.23348 26.14437 25.78555
##  [89] 23.78958 24.13385 24.71367 26.61526 23.75629 24.79398 23.41953 25.28689
##  [97] 26.29807 24.74015 26.87821 23.30800
```

```
quant <- quantile(richie2, probs=c(0.025,0.975))
quant
```

```
##     2.5%    97.5%
## 23.36098 26.91345
```

```r
if (mean(grads) < quant[1] || mean(grads) > quant[2]) {
  print("Sample mean is extreme")
} else {
  print("Sample mean is not extreme")
}
```

```
## [1] "Sample mean is not extreme"
```

```r
mean(undergrads)
```

```
## [1] 19.52719
```

```r
#I personally think that the testSample are undergrad students because the given mean sample is between
```