HW4_Vikas Sanil

Vikas Sanil

Due Date 3/30

Packages

```
library("tigerstats")
```

Part ONE: Review the approach to location and scale problems for one and two populations (6 points, 2 points each question)

- For inference on population mean, which of the following could we potentially use? *Choose all apply*

A. The normal distribution (with the Z statistic)

B. The t-distribution (with the T statistic)

C. The chi-square distribution (with the χ^2 statistic)

D. The F-distribution (with the F statistic)

> Answer: The normal distribution (with the Z statistic) when Standard Deviation of population is known and The t-distribution (with the T statistic) when Standard Deviation of population is not known.

- For inference on population variance, which of the following distributions will be useful?

  A. normal

  B. T

  C. χ^2

  D. F

  > Answer: χ^2

- For comparing variances between two populations, which of the following distributions will be useful?

  A. normal

  B. T

  C. χ^2

  D. F

  > Answer: F

Part Two: Confidence Interval and Hypothesis Testing (30 points)

Problem 1. An electrical firm manufactures light bulbs that have a length of life that is approximately normally distributed, with mean equal to 800 hours and a standard deviation of 40 hours. A random sample of 16 bulbs will have an average life of 775 hours. (15 points)

We wish to test

$$H_0 : \mu = 800,$$
$$H_1 : \mu \neq 800.$$

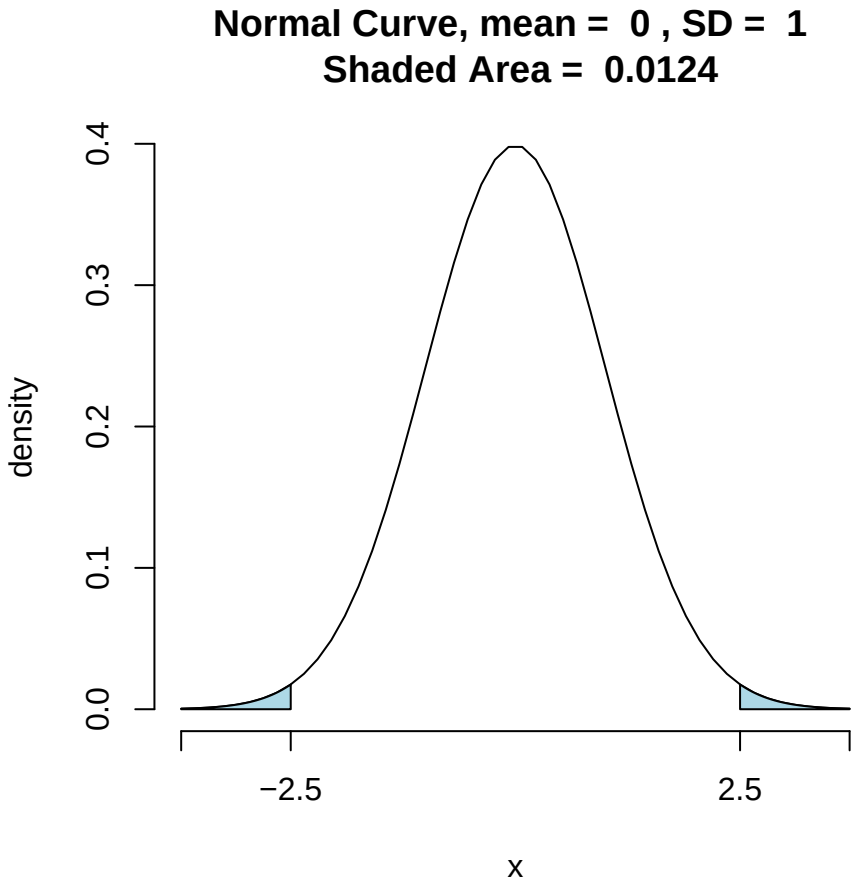
Will you reject H_0 suppose $\bar{X} = 775$? Justify your answer.

Pro-tip: You may use either R or Python or hand-calculations to answer the computational part of this question, however, you do need to–in any case–explain and justify your answer.

```
mu<-800
sigma<-40
n<-16
xbar<-775

zstats<-(xbar-mu)/(sigma/sqrt(n))
zstats
```

```
[1] -2.5
```

```
pnormGC(c(zstats,-zstats), region = "outside", graph=TRUE)
```




```
[1] 0.01241933
```

Answer: Reject H_0 since the z-score falls beyond 95% confidence interval and same can be seen in the graph.

Problem 2.Checking out some small data sets that come with R (15 points)

In this problem, you will load and work with the `mtcars` data set in R.

Two data samples are independent if they come from unrelated populations and the samples does not affect each other. Here, we assume that the data populations follow the *normal distribution*. In the data frame column `mpg` (which stands for "miles per galon") of the data set `mtcars`, there are gas mileage data of various 1974 U.S. automobiles. Let's take a look:

```
mtcars$mpg
```

```
 [1] 21.0 21.0 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 17.8 16.4 17.3 15.2 10.4
[16] 10.4 14.7 32.4 30.4 33.9 21.5 15.5 15.2 13.3 19.2 27.3 26.0 30.4 15.8 19.7
[31] 15.0 21.4
```

Meanwhile, another data column in `mtcars`, named `am`, indicates the transmission type of the automobile model (0 = automatic, 1 = manual):

```
mtcars$am
```

```
 [1] 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 0 0 0 0 0 1 1 1 1 1 1 1
```

In particular, the gas mileage for manual and automatic transmissions are two independent data populations.

Assume that the data in `mtcars` follows the normal distribution, let us look for whether the **difference between the mean gas mileage** of manual and automatic transmissions seems to be statistically significant.

a. Please construct a hypothesis test for ratio of the variances of these two populations. Then choose the appropriate test to make a conclusion.

b. Based on the result in a), construct a hypothesis test for the means of these two populations. Show your conclusion.

Hints and shortcuts The gas mileage for automatic transmission can be listed as follows:

```
L = mtcars$am == 0
 mpgAuto = mtcars[L,]$mpg
 mpgAuto                    # automatic transmission mileage
```

```
 [1] 21.4 18.7 18.1 14.3 24.4 22.8 19.2 17.8 16.4 17.3 15.2 10.4 10.4 14.7 21.5
[16] 15.5 15.2 13.3 19.2
```

By applying the negation of L, we can find the gas mileage for manual transmission:

```
mpgManual = mtcars[!L,]$mpg
 mpgManual                  # manual transmission mileage
```

```
 [1] 21.0 21.0 22.8 32.4 30.4 33.9 27.3 26.0 30.4 15.8 19.7 15.0 21.4
```

Now you should be able to finish solving the problem. Enjoy! :)

```
var.test(mpgAuto,mpgManual)
```

```
    F test to compare two variances

data:  mpgAuto and mpgManual
F = 0.38656, num df = 18, denom df = 12, p-value = 0.06691
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.1243721 1.0703429
sample estimates:
ratio of variances
         0.3865615
```

Answer a: The ratio of variances is 0.3866 and p-value of 0.067. We can assume the variance are equal as it falls within 95% confidence interval. I am performing two sample t-test with condition var.equal = TRUE since we are assuming that the data populations follow the *normal distribution*. The hypothesis being test:
$H_0 : \mu_{auto} = \mu_{manual}$, $H_1 : \mu_{auto} \neq \mu_{manual}$

```
t.test(mpgAuto,mpgManual, var.equal = TRUE, conf.level = .95)
```

```
    Two Sample t-test

data:  mpgAuto and mpgManual
t = -4.1061, df = 30, p-value = 0.000285
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -10.84837  -3.64151
sample estimates:
mean of x mean of y
 17.14737  24.39231
```

Answer b: Rejecting H_0 since the probablity value is 0.000285 which is not within 95% confidence interval.

Part Three: Working With Data (60 points)

Let's revisit cybersecurity breach report data downloaded 2015-02-26 from the US Health and Human Services. From the *Office for Civil Rights* of the *U.S. Department of Health and Human Services*, I obtained the following information:

"As required by section 13402(e)(4) of the HITECH Act, the Secretary must post a list of breaches of unsecured protected health information affecting 500 or more individuals.

"Since October 2009 organizations in the U.S. that store data on human health are required to report any incident that compromises the confidentiality of 500 or more patients / human subjects (45 C.F.R. 164.408). These reports are publicly available. Our data set was downloaded from the Office for Civil Rights of the U.S. Department of Health and Human Services, 2015-02-26."

Load this data set and store it as `cyberData`, using the following code:

```
cyberData<-read.csv(url("https://vincentarelbundock.github.io/Rdatasets/csv/Ecdat/HHSCyberSecurityBreach
```

As you know, this data set contains *all* reports regarding health information data breaches from 2009 to 2015. Let's pretend this is just a *sample* from the population of *all data breaches*, related or not to health information.

Question 1. (30 points)

- Compare the number of individuals affected by data breaches (column `Individuals.Affected`) in two states, Arkansas (`State=="AR"`) and California (`State=="CA"`). This can be done by performing a test of difference in means, for example. Assume the individuals affected follows an **approximately normal distribution**. (15 points)

Please note, in order to answer this question completely, you will need to run several lines of code, extract subsets of the data appropriately, run a statistical hypothesis test, and interpret the results. Draw a conclusion. Partial answers to the question are insufficient.

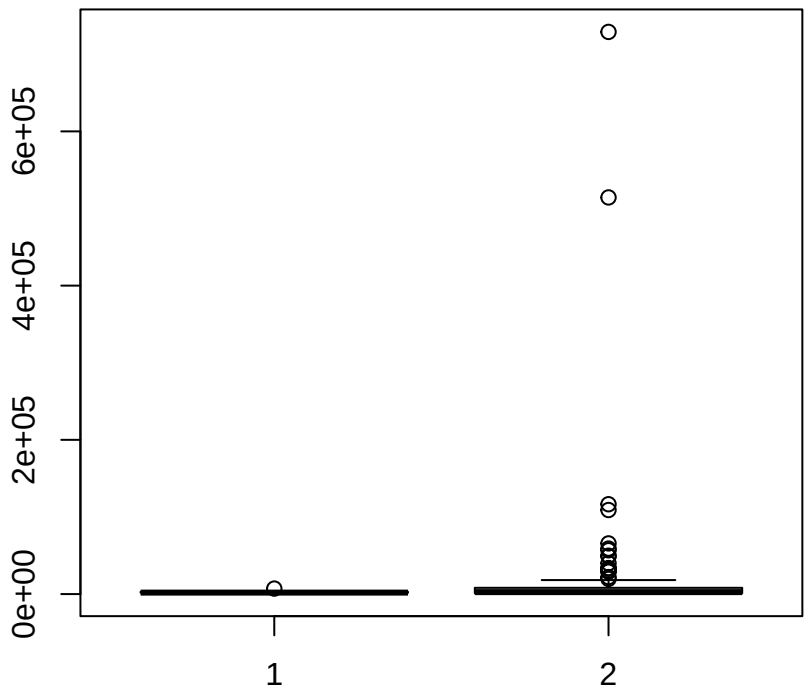
```
individualsAffected_AR = subset(cyberData, State=="AR")
individualsAffected_CA = subset(cyberData, State=="CA")
var.test(individualsAffected_AR$Individuals.Affected,individualsAffected_CA$Individuals.Affected)
```

```
    F test to compare two variances

data:  individualsAffected_AR$Individuals.Affected and individualsAffected_CA$Individuals.Affected
F = 0.00066857, num df = 6, denom df = 127, p-value = 2.814e-09
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.0002664288 0.0032769357
sample estimates:
ratio of variances
      0.0006685688
```

```
t.test(individualsAffected_AR$Individuals.Affected ,individualsAffected_CA$Individuals.Affected ,var.eq
```

```
    Welch Two Sample t-test

data:  individualsAffected_AR$Individuals.Affected and individualsAffected_CA$Individuals.Affected
t = -2.2841, df = 129.71, p-value = 0.02399
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -30145.686  -2161.579
sample estimates:
mean of x mean of y
  2769.00  18922.63
```

```
boxplot(individualsAffected_AR$Individuals.Affected ,individualsAffected_CA$Individuals.Affected)
```



Answer:
- Since variance is unknown, I have ran var.test and found the variance is not equal for given data set and even mean is also not equal.
- Based on the test results and boxplot we can conclude the average individuals affected by data breach in State CA is greater than average individuals affected in State AR.

- Repeat the same test for another pair of states, California ("CA") and Illinois ("IL"). (15 points)

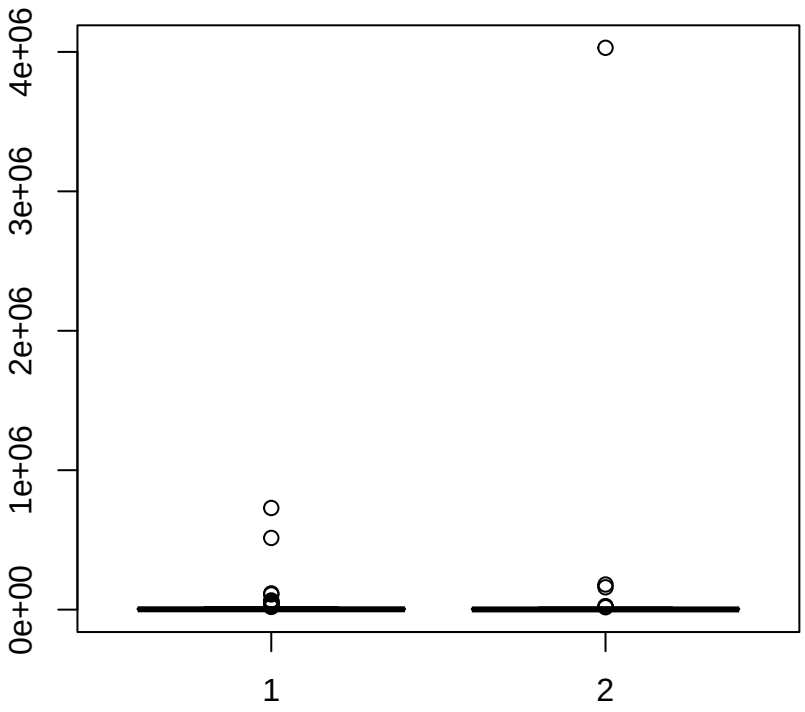
```
individualsAffected_IL = subset(cyberData, State=="IL")
var.test(individualsAffected_CA$Individuals.Affected ,individualsAffected_IL$Individuals.Affected)
```

```
    F test to compare two variances

data:  individualsAffected_CA$Individuals.Affected and individualsAffected_IL$Individuals.Affected
F = 0.02224, num df = 127, denom df = 56, p-value < 2.2e-16
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.01392852 0.03413975
sample estimates:
ratio of variances
        0.02223981
```

```
t.test(individualsAffected_CA$Individuals.Affected ,individualsAffected_IL$Individuals.Affected ,var.equ
```

```
    Welch Two Sample t-test

data:  individualsAffected_CA$Individuals.Affected and individualsAffected_IL$Individuals.Affected
t = -0.87104, df = 57.112, p-value = 0.3874
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -203969.48   80308.12
sample estimates:
mean of x mean of y
 18922.63  80753.32
```

```
boxplot(individualsAffected_CA$Individuals.Affected ,individualsAffected_IL$Individuals.Affected)
```



Answer:
- Since variance is unknown, I have ran var.test and found the variance is not equal for given data set and even mean is also not equal.
- Based on the test results and boxplot we can conclude the average individuals affected by data breach in State IL is greater than average individuals affected in State CA.

Question 2. (30 points)

Explore the variable `Type.Of.Breach` collected in this data set:

- What proportion of data entries in `cyberData` have `Type.of.Breach == "Hacking/IT Incident"` ? (6 points)

```
numOfHITI<-length(which(cyberData$Type.of.Breach  == "Hacking/IT Incident"))
prop1 <- numOfHITI / nrow(cyberData)
prop1
```

```
[1] 0.06689835
```

Answer: 0.0668983 proportion of data entries in 'cyberData' have 'Type.of.Breach == "Hacking/IT Incident" '.

- What are all the different values of `Type.Of.Breach` reported in the data set? How many are hacking/IT incidents? (6 points)

```
typeOfBreach<-cyberData$Type.of.Breach
numTypeofBreach<- length(unique(typeOfBreach))
table(typeOfBreach)
```

```
typeOfBreach
                                        Hacking/IT Incident
                                                         77
                                   Hacking/IT Incident, Other
                                                          2
Hacking/IT Incident, Other, Unauthorized Access/Disclosure
                                                          1
                                   Hacking/IT Incident, Theft
                                                          1
Hacking/IT Incident, Theft, Unauthorized Access/Disclosure
                                                          3
       Hacking/IT Incident, Unauthorized Access/Disclosure
                                                         10
                                          Improper Disposal
                                                         42
                                     Improper Disposal, Loss
                                                          3
                               Improper Disposal, Loss, Theft
                                                          3
  Improper Disposal, Theft, Unauthorized Access/Disclosure
                                                          1
         Improper Disposal, Unauthorized Access/Disclosure
                                                          2
                                                       Loss
                                                         79
                                                 Loss, Other
                                                          2
                                           Loss, Other, Theft
                                                          1
```

```
                                        Loss, Theft
                                                 15
                  Loss, Unauthorized Access/Disclosure
                                                  5
 Loss, Unauthorized Access/Disclosure, Unknown
                                                  1
                                       Loss, Unknown
                                                  2
                                              Other
                                                 89
                                        Other, Theft
                                                  5
  Other, Theft, Unauthorized Access/Disclosure
                                                  2
         Other, Unauthorized Access/Disclosure
                                                  7
                                      Other, Unknown
                                                  2
                                              Theft
                                                577
         Theft, Unauthorized Access/Disclosure
                                                 24
Theft, Unauthorized Access/Disclosure, Unknown
                                                  1
                       Unauthorized Access/Disclosure
                                                183
                     Unauthorized Access/Disclosure
                                                  1
                                            Unknown
                                                 10
```

Answer: 29 types of Breach reported in the data set. 77 are hacking/IT incidents.

- What type of breach is reported in the 748th row of `cyberData`? How about 349th row? Was row 349 counted in the proportion of Hacking/IT incident breaches you computed above? Why or why not? (8 points)

```
breach748<-cyberData$Type.of.Breach[748]
breach349<-cyberData$Type.of.Breach[349]
tb. <- strsplit(typeOfBreach, ', ')
table(unlist(tb.))
```

```
            Hacking/IT Incident                Improper Disposal
                             94                               51
                           Loss                            Other
                            111                              111
                          Theft  Unauthorized Access/Disclosure
                            633                              240
Unauthorized Access/Disclosure                          Unknown
                              1                               16
```

Answer: - Row 748th has Loss, Theft breach in cyberData. - Row 349th has Hacking/IT Incident, Unauthorized Access/Disclosure breach in cyberData. - No row 349 was

not counted in the proportion of Hacking/IT incident breaches computed above. - I was seraching for unique entry of Hacking/IT Incident breaches. Didn't account the combination of multiple breaches.

- Perform a hypothesis test on whether there is a difference in proportion of Hacking/IT incidents between the state of Illinois and the state of California. Write your conclusion interpreting the results of the statistical test. (10 points)

```
tb_IL<- strsplit(individualsAffected_IL$Type.of.Breach, ', ')
tb_CA<- strsplit(individualsAffected_CA$Type.of.Breach, ', ')
table(unlist(tb_IL))
```

```
           Hacking/IT Incident              Improper Disposal
                             8                              2
                          Loss                          Other
                             5                             10
                         Theft Unauthorized Access/Disclosure
                            25                             10
```

```
table(unlist(tb_CA))
```

```
           Hacking/IT Incident              Improper Disposal
                             7                              3
                          Loss                          Other
                            13                             13
                         Theft Unauthorized Access/Disclosure
                            81                             22
```

```
numOfHITI_IL <- 8
numOfBreach_IL <- nrow(individualsAffected_IL)
numOfHITI_CA <- 7
numOfBreach_CA <- nrow(individualsAffected_IL)
prop.test(x = c(numOfHITI_IL,numOfHITI_CA), n = c(numOfBreach_IL,numOfBreach_CA), alternative = "greate
```

```
    2-sample test for equality of proportions with continuity correction

data:  c out of cnumOfHITI_IL out of numOfBreach_ILnumOfHITI_CA out of numOfBreach_CA
X-squared = 0, df = 1, p-value = 0.5
alternative hypothesis: greater
95 percent confidence interval:
 -0.1041159  1.0000000
sample estimates:
   prop 1    prop 2
0.1403509 0.1228070
```

Answer: Based on p-value 0.5 we can say proporition of Hacking/IT incidents in State IL is equal to State CA.