ITMD/ITMS/STAT 514 Homework 3

GURUTEJA KANDERI

Due Date 3/9

Packages

```
# Import needed packages
library(tigerstats)
```

Part I. Confidence Interval and Hypothesis Testing (60 points)

Question 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 less than 775 hours. (15 points)

```

# Given values
mu <- 800
sigma <- 40
n <- 16
sample_mean <- 775

# Calculate z-score
z <- (sample_mean - mu) / (sigma / sqrt(n))


prob_800 <- pnorm(z)


mu <- 760
z <- (sample_mean - mu) / (sigma / sqrt(n))
prob_760 <- pnorm(z)

# Print the results
prob_800
```

a. Give a probabilistic result that indicates how rare an event $\bar{X} \leq 775$ is when $\mu = 800$. (Hint: Calculate the probability $P(\bar{X} \leq 775)$ when $\mu = 800$). On the other hand, how rare would it be if μ truly were, say, 760 hours?

```
## [1] 0.006209665
```

```
prob_760
```

```
## [1] 0.9331928
```

Answer: prob_800= 0.006209665 prob_760= 0.9331928

```r
# Given values
sample_mean <- 775
sigma <- 40
n <- 16
conf_level <- 0.95

# Calculate the critical value (z-score) for the confidence level
z_crit <- qnorm(1 - (1 - conf_level) / 2)

# Calculate the margin of error
margin_of_error <- z_crit * (sigma / sqrt(n))

# Calculate the confidence interval
conf_int <- c(sample_mean - margin_of_error, sample_mean + margin_of_error)

# Print the confidence interval
cat("95% Confidence Interval:", conf_int, "\n")
```

b. Please construct a 95% confidence interval on μ with $\bar{X} = 775$. Is 800 inside the interval?

```
## 95% Confidence Interval: 755.4004 794.5996
```

```r
# Check if mu = 800 falls within the interval
if (800 >= conf_int[1] && 800 <= conf_int[2]) {
  cat("mu = 800 falls within the confidence interval.")
} else {
  cat("mu = 800 does not fall within the confidence interval.")
}
```

```
## mu = 800 does not fall within the confidence interval.
```

> Answer: 95% Confidence Interval: 755.4004 794.5996 mu = 800 does not fall within the confidence interval.

Question 2. A maker of a certain brand of low-fat cereal bars claims that the average saturated fat content is 0.5 gram. In a random sample of 8 cereal bars of this brand, the saturated fat content was 0.6, 0.7, 0.7, 0.3, 0.4, 0.5, 0.4, and 0.2. ***Assume a normal distribution.*** **(15 points)**

```r
saturated_fat <- c(0.6, 0.7, 0.7, 0.3, 0.4, 0.5, 0.4, 0.2)  # Saturated fat content in grams
sample_mean <- mean(saturated_fat)
sample_sd <- sd(saturated_fat)
n <- length(saturated_fat)

# Calculate critical value (t) from t-distribution for 95% confidence level
t <- qt(0.975, df = n - 1)  # 0.975 corresponds to (1 - (1 - 0.95)/2)

# Calculate margin of error
margin_of_error <- t * (sample_sd / sqrt(n))

# Calculate lower and upper bounds of the confidence interval
lower_bound <- sample_mean - margin_of_error
upper_bound <- sample_mean + margin_of_error

# Construct the confidence interval
```

```
confidence_interval <- c(lower_bound, upper_bound)

# Print the confidence interval
confidence_interval
```

a. Please construct a 95% confidence interval on the average saturated fat content.

```
## [1] 0.32182 0.62818
```

```
cat("95% confidence interval for average saturated fat content: (", lower_bound, ",", upper_bound, ") g
```

```
## 95% confidence interval for average saturated fat content: ( 0.32182 , 0.62818 ) grams
```

> Answer: 0.32182 0.62818 95% confidence interval for average saturated fat content: (0.32182 , 0.62818) grams

b. Would you agree with the claim? Justify your answer.

> Answer: Yes, we can agree with the claim that the average saturated fat content is 0.5 gram since the claimed value of 0.5 falls within the calculated 95% confidence interval [0.3165347, 0.6334653].

Question 3. In this problem, you will load and work with the mtcars data set in R. (30 points)

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 gallon") 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 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.

```

# Split the data based on transmission type
auto_mpg <- mtcars$mpg[mtcars$am == 0]
manual_mpg <- mtcars$mpg[mtcars$am == 1]

# Perform the variance ratio test
var_ratio_test <- var.test(auto_mpg, manual_mpg, ratio = 1, conf.level = 0.95)
var_ratio_test
```

a. Please construct a 95% confidence interval on the ratio of the variances of gas milage between auto and manual.

```
##
##  F test to compare two variances
```

```
##
## data:  auto_mpg and manual_mpg
## 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: 95% Confidence Interval for the Ratio of Variances:0.1243721 1.0703429

F test to compare two variances

data: auto_mpg and manual_mpg 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

b. Please construct a hypothesis test for ratio of the variances of these two populations with the significance level $\alpha = 0.05$. Then choose the appropriate test to make a conclusion.

Hypothesis Test Here

```
# Load the mtcars dataset
data(mtcars)

# Separate gas mileage data for automatic and manual transmissions
mpg_auto <- mtcars$mpg[mtcars$am == 0]
mpg_manual <- mtcars$mpg[mtcars$am == 1]

# Calculate sample variances
s_auto <- var(mpg_auto)
s_manual <- var(mpg_manual)

# Sample sizes
n_auto <- length(mpg_auto)
n_manual <- length(mpg_manual)

# Degrees of freedom
df1 <- n_auto - 1
df2 <- n_manual - 1

# Test statistic
F_statistic <- s_auto / s_manual

# Significance level
alpha <- 0.05

# Calculate the critical values
F_crit_lower <- qf(alpha / 2, df1, df2)
F_crit_upper <- qf(1 - alpha / 2, df1, df2)

# Print the test results
cat("Test Statistic (F):", F_statistic, "\n")
```

```
## Test Statistic (F): 0.3865615
```

```r
cat("Critical Values (F_crit_lower, F_crit_upper):", F_crit_lower, F_crit_upper, "\n")
```

```
## Critical Values (F_crit_lower, F_crit_upper): 0.3611567 3.108106
```

```r
# Make a conclusion
if (F_statistic < F_crit_lower || F_statistic > F_crit_upper) {
  cat("Reject the null hypothesis (H0) at the significance level of 0.05.\n")
  cat("The variances of gas mileage for automatic and manual transmissions are not equal.\n")
} else {
  cat("Fail to reject the null hypothesis (H0) at the significance level of 0.05.\n")
  cat("There is no evidence to suggest that the variances of gas mileage for automatic and manual transm
}
```

```
## Fail to reject the null hypothesis (H0) at the significance level of 0.05.
## There is no evidence to suggest that the variances of gas mileage for automatic and manual transmissi
```

Answer: Test Statistic (F): 0.3865615 Critical Values (F_crit_lower, F_crit_upper): 0.3611567 3.108106 Fail to reject the null hypothesis (H0) at the significance level of 0.05. There is no evidence to suggest that the variances of gas mileage for automatic and manual transmissions are different.

c. Based on the result in b), construct a hypothesis test for the means of these two populations with the significance level $\alpha = 0.05$. Show your conclusion.

Hypothesis Test Here

```r
t_test_result <- t.test(mpg ~ am, data = mtcars, var.equal = TRUE)

t_statistic <- t_test_result$statistic
p_value <- t_test_result$p.value

# Determining if the null hypothesis should be rejected
alpha <- 0.05
reject_null <- p_value < alpha

cat("Two-Sample T-Test:\n")
```

```
## Two-Sample T-Test:
```

```r
cat("Test Statistic:", t_statistic, "\n")
```

```
## Test Statistic: -4.106127
```

```r
cat("P-Value:", p_value, "\n")
```

```
## P-Value: 0.0002850207
```

```r
cat("Reject Null Hypothesis (at alpha = 0.05):", reject_null, "\n")
```

```
## Reject Null Hypothesis (at alpha = 0.05): TRUE
```

Answer: Two-Sample T-Test: Test Statistic: -4.106127 P-Value: 0.0002850207 Reject Null Hypothesis (at alpha = 0.05): TRUE Conclusion: Based on the two-sample t-test conducted with a significance level of alpha = 0.05, we found a test statistic of -4.106127 and a p-value of 0.0002850207. Since the p-value is smaller than the significance level, we reject the null hypothesis. Therefore, we conclude that there is a significant difference in the means of the two populations represented by the “mpg” variable, classified by the “am” variable in the mtcars dataset.

Two-Sample T-Test: Test Statistic: -4.106127 P-Value: 0.0002850207 Reject Null Hypothesis (at alpha = 0.05): TRUE

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
```

Part II. Working With Data (45 points)

In this part, we work on the cyber security breach report data downloaded 2015-02-26 from the US Health and Human Services.

To understand what the data represents, here is some information from the *Office for Civil Rights* of the *U.S. Department of Health and Human Services*:

- "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 save it as `cyberData`, using the following code:

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

In this homework, we focus more on **how to construct hypothesis testing**.

Question 1. (10 points)

Check the type of the `Breach.Submission.Date` column: is it a numeric? What type is it?

```
# Load the dataset
cyberData <- read.csv(url("https://vincentarelbundock.github.io/Rdatasets/csv/Ecdat/HHSCyberSecurityBrea

# Check the type of the Breach.Submission.Date column
typeof(cyberData$Breach.Submission.Date)
```

```
## [1] "character"
```

Answer: "character"

Let us change it to a numeric and extract *the year only*. The code that does this is `as.numeric(format(as.Date(.....),"%Y")` Let us use this code to break up the data to before and after 2013, like this:

```
# Subset the data before and after 2013
before2013 <- subset(cyberData, as.numeric(format(as.Date(Breach.Submission.Date), "%Y")) <= 2013)
after2013 <- subset(cyberData, as.numeric(format(as.Date(Breach.Submission.Date), "%Y")) > 2013)
```

```r
# Calculate the number of observations in each subset
num_before2013 <- nrow(before2013)
num_after2013 <- nrow(after2013)

# Print the number of observations in each subset
num_before2013
```

```
## [1] 848
```

```r
num_after2013
```

```
## [1] 303
```

How many observations are in each subset of the population?

> *Type your answer here* # Print the number of observations in each subset num_before2013 [1] 848 num_after2013 [1] 303

Question 2. (15 points)

a. What is the number of observation in `before2013` have `Type.of.Breach == "Hacking/IT Incident"` ? Hint: Recall the functions we used in HW 2.

```r
# Subset before2013 data frame where Type.of.Breach is "Hacking/IT Incident"
hacking_incidents_before2013 <- subset(before2013, Type.of.Breach == "Hacking/IT Incident")

# Count the number of observations
num_hacking_incidents_before2013 <- nrow(hacking_incidents_before2013)

# Print the result
num_hacking_incidents_before2013
```

```
## [1] 49
```

> Your Answer: 49

```r
# Subset after2013 data frame where Type.of.Breach is "Hacking/IT Incident"
hacking_incidents_after2013 <- subset(after2013, Type.of.Breach == "Hacking/IT Incident")

# Count the number of observations
num_hacking_incidents_after2013 <- nrow(hacking_incidents_after2013)

# Print the result
num_hacking_incidents_after2013
```

b. What is the number of observation in `after2013` have `Type.of.Breach == "Hacking/IT Incident"` ?

```
## [1] 28
```

> Answer: 28

```r
# Count the number of observations with Type.of.Breach == "Hacking/IT Incident" before and after 2013
count_before2013 <- sum(before2013$Type.of.Breach == "Hacking/IT Incident")
count_after2013 <- sum(after2013$Type.of.Breach == "Hacking/IT Incident")
```

```
# Total number of observations before and after 2013
total_before2013 <- nrow(before2013)
total_after2013 <- nrow(after2013)

# Use prop.test to compare proportions
prop_test_result <- prop.test(c(count_before2013, count_after2013), c(total_before2013, total_after2013)

# Print the result of prop.test
print(prop_test_result)
```

c. Please use `prop.test` to check if the proporations of data entries having `Type.of.Breach == "Hacking/IT Incident"` are the same before and after 2013. Show you conclusion.

```
##
##  2-sample test for equality of proportions without continuity correction
##
## data:  c out of ccount_before2013 out of total_before2013count_after2013 out of total_after2013
## X-squared = 4.2877, df = 1, p-value = 0.03839
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.070819329  0.001566885
## sample estimates:
##     prop 1     prop 2
## 0.05778302 0.09240924
```

```
# Conclusion
if (prop_test_result$p.value < 0.05) {
  cat("Conclusion: Based on the prop.test, we reject the null hypothesis.\n")
  cat("There is evidence to suggest that the proportions of data entries having 'Type.of.Breach == \"Ha
} else {
  cat("Conclusion: Based on the prop.test, we fail to reject the null hypothesis.\n")
  cat("There is insufficient evidence to conclude that the proportions of data entries having 'Type.of.I
}
```

```
## Conclusion: Based on the prop.test, we reject the null hypothesis.
## There is evidence to suggest that the proportions of data entries having 'Type.of.Breach == "Hacking
```

> Conclusion: Based on the prop.test, we reject the null hypothesis. The p-value (0.03839) is smaller than the significance level (0.05)

2-sample test for equality of proportions without continuity correction

data: c(count_before2013, count_after2013) out of c(total_before2013, total_after2013) X-squared = 4.2877, df = 1, p-value = 0.03839 alternative hypothesis: two.sided 95 percent confidence interval: -0.070819329 0.001566885 sample estimates: prop 1 prop 2 0.05778302 0.09240924

Question 3. Conducting Hypothesis Testing (20 points)

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.

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**.

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.

```r
# Load data
cyberData <- read.csv("https://vincentarelbundock.github.io/Rdatasets/csv/Ecdat/HHSCyberSecurityBreaches

# Subset the data for Arkansas and California
data_ar <- subset(cyberData, State == "AR")
data_ca <- subset(cyberData, State == "CA")

# Perform F-test to compare variances
f_test_result <- var.test(data_ar$Individuals.Affected, data_ca$Individuals.Affected)

# Display F-test results
print(f_test_result)
```

```
##
##  F test to compare two variances
##
## data:  data_ar$Individuals.Affected and data_ca$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
```

```r
# Check if variances are significantly different
if (f_test_result$p.value < 0.05) {
  cat("The variances of individuals affected in AR and CA are significantly different.\n")
} else {
  cat("There is no significant difference in variances of individuals affected between AR and CA.\n")
}
```

```
## The variances of individuals affected in AR and CA are significantly different.
```

```r
# Perform Welch's t-test assuming variances are not equal
t_test_result <- t.test(data_ar$Individuals.Affected, data_ca$Individuals.Affected, var.equal = FALSE)

# Display t-test results
print(t_test_result)
```

```
##
##  Welch Two Sample t-test
##
## data:  data_ar$Individuals.Affected and data_ca$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
```

```r
# Interpret t-test results and conclude
if (t_test_result$p.value < 0.05) {
  cat("Based on the two-sample t-test, we reject the null hypothesis.\n")
  cat("There is a statistically significant difference in the mean number of individuals affected by dat
  cat("Mean of Individuals Affected in Arkansas:", mean(data_ar$Individuals.Affected), "\n")
  cat("Mean of Individuals Affected in California:", mean(data_ca$Individuals.Affected), "\n")
} else {
  cat("Based on the two-sample t-test, we fail to reject the null hypothesis.\n")
  cat("There is no statistically significant difference in the mean number of individuals affected by da
}
```

```
## Based on the two-sample t-test, we reject the null hypothesis.
## There is a statistically significant difference in the mean number of individuals affected by data bi
## Mean of Individuals Affected in Arkansas: 2769
## Mean of Individuals Affected in California: 18922.63
```

> Your answer: Based on the two-sample t-test, we reject the null hypothesis.The F-test revealed significantly different variances of individuals affected by data breaches in Arkansas and California (F = 0.00066857, p < 0.05), indicating variance inequality between the states.Proceeding with Welch's two-sample t-test due to unequal variances, a significant difference in mean individuals affected was found (t = -2.2841, p = 0.02399).Arkansas (mean = 2769) had significantly fewer individuals affected compared to California (mean = 18922.63), with a 95% confidence interval (-30145.686, -2161.579) indicating a substantial difference.In summary, there is a statistically significant discrepancy in the mean number of individuals affected by data breaches between Arkansas and California, with California experiencing a significantly higher impact.

F test to compare two variances

data: data_ar$Individuals.Affectedanddata_ca$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

The variances of individuals affected in AR and CA are significantly different.

`Welch Two Sample t-test`

data: data_ar$Individuals.Affectedanddata_ca$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

Based on the two-sample t-test, we reject the null hypothesis. There is a statistically significant difference in the mean number of individuals affected by data breaches between Arkansas and California. Mean of Individuals Affected in Arkansas: 2769 Mean of Individuals Affected in California: 18922.63