**R Exercise**

Requirements

1. Complete the entire assignment using this Word document.
2. If a question requires R scripts, you copy your R scripts and the corresponding outputs from your RStudio and paste them here in response to the question.
3. Submit your work as an attachment via the assignment link which will be available on the blackboard.
4. You may find the following resources helpful:

* Elementary Statistics with R (<http://www.r-tutor.com/elementary-statistics>).
  + Most R related problems follow this R tutorial.
  + In addition to the R Tutorial video for beginners (See blackboard), the section “R Introduction” will give you a starting point for learning R.
* <https://www.youtube.com/watch?v=VGKz3Jkx-9I&t=326s> (This video helps with joint frequency distribution, a.k.a. cross-tabulation or contingency table).

**Nivethida Kumarasamy**

**SB3893**

**Problem 1**

Nearly 1.8 million bachelor’s degrees and over 750,000 master’s degrees are awarded annually by U.S. postsecondary institutions. The department of Education tracks the field of study for these graduates in the following categories: Business (B), Computer Sciences and Engineering (CSE), Education (E), Humanities (H), Natural Sciences and Mathematics (NSM), Social and Behavioral Sciences (SBS), and Other (O). Use the data in the file “Majors.csv” to answer the following questions in R.

1. Find the frequency distribution of the Bachelor’s degree.

table(majorsTable$Bachelor)

output:

B CSE E H NSM O SBS

21 9 6 16 8 24 16

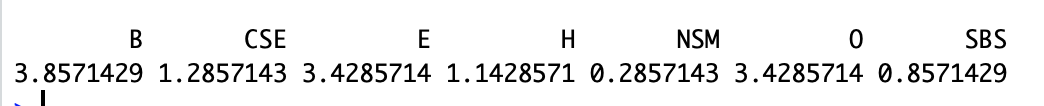
1. Find the relative frequency distribution of the Master’s degree.

MastersDegree <- table(majorsTable$Master)

MastersRelativeFreq <- MastersDegree/nrow(MastersDegree)

print(MastersRelativeFreq)

output:



1. Find the joint frequency distribution of the Bachelor’s degree and the Master’s degree.

table(majorsTable$Bachelor,majorsTable$Master)

output:

A close up of a logo

Description automatically generated

1. Find the joint relative frequency distribution of the Bachelor’s degree and the Master’s degree.

JointTable <- table(majorsTable$Bachelor,majorsTable$Master)

JointTableRelativeFreq <- JointTable/nrow(JointTable)

print(JointTableRelativeFreq)

output:

A screenshot of a cell phone

Description automatically generated

1. Create a bar chart of the Bachelor’s degree.

BarchartData <- table(majorsTable$Bachelor)

barplot(BarchartData, xlab="Bachelor Degree",ylab="Count")

output:

A close up of a logo

Description automatically generated

1. Create a pie chart of the Master’s degree.

PieChartData <- table(majorsTable$Master)

pie(PieChartData)

output:

A picture containing drawing

Description automatically generated

**Problem 2**

The 32 teams in the National Football League (NFL) are worth, on average, $1.17 billion, 5 percent more than last year. The data file "NFLTeamValue.csv" shows the annual revenue ($ millions) and the estimated team value ($ millions) for the 32 NFL teams (Forbes website, February 28, 2014).

Please read the data from "NFLTeamValue.csv" and complete the following tasks. Note: In total, there are 32 NFL teams. Therefore, the data contains the entire population. Make sure you use appropriate formulas (for population, not for sample) when computing the values needed.

1. Find the frequency distribution of the team revenue from 200 to 550 with the length of each interval being 50.

TeamRevenue <- NFLTable$Revenue

RevenueClassWidth = seq(200, 550, by=50)

TeamRevenue.cut <- cut(TeamRevenue, RevenueClassWidth, labels = c("200-250","250-300","300-350","350-400","400-450","450-500","500-550"))

TreamRevenue.freq = table(TeamRevenue.cut)

print(TreamRevenue.freq)

output:

A close up of a logo

Description automatically generated

1. Create a histogram of the team revenue based on the frequency distribution above.

output <- table(TeamRevenue.cut)

hist(output, plot = FALSE)

plot(output, type="s")

output:

A screenshot of a cell phone

Description automatically generated

1. Find the relative frequency distribution of the team value from 750 to 2500 with the length of each interval being 250.

TeamValue <- NFLTable$Value

TeamValueWidth = seq(750, 2500, by=250)

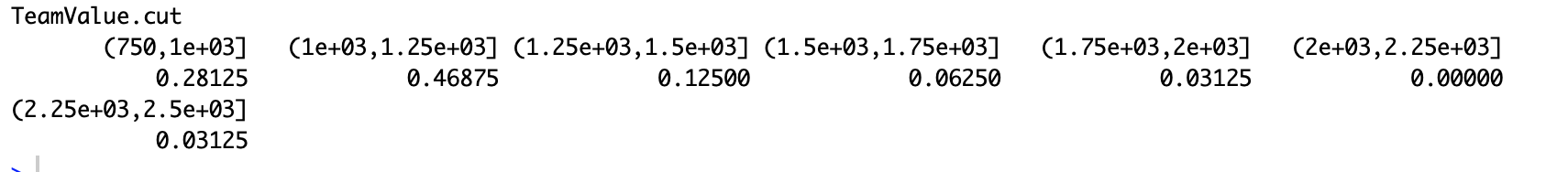
TeamValue.cut <- cut(TeamValue, TeamValueWidth)

teamValule.freq <- table(TeamValue.cut)

teamValue.relativefreq <- teamValule.freq/nrow(NFLTable)

print(teamValue.relativefreq)

output:

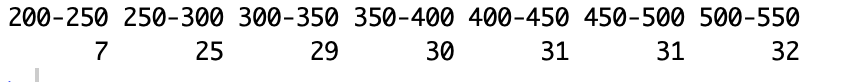
****

1. Find the cumulative frequency distribution of the team revenue.

TeamRevenue.cumulativefreq = cumsum(TreamRevenue.freq)

print(TeamRevenue.cumulativefreq)

**output:**

****

1. Create a cumulative frequency graph of the team revenue.

cumrelfreqGraphTeamRevenue = c(0, TeamRevenue.cumulativefreq)

plot(cumrelfreqGraphTeamRevenue,

main="cumulative frequency graph of the team revenue")

lines(cumrelfreqGraphTeamRevenue)

A screenshot of a cell phone

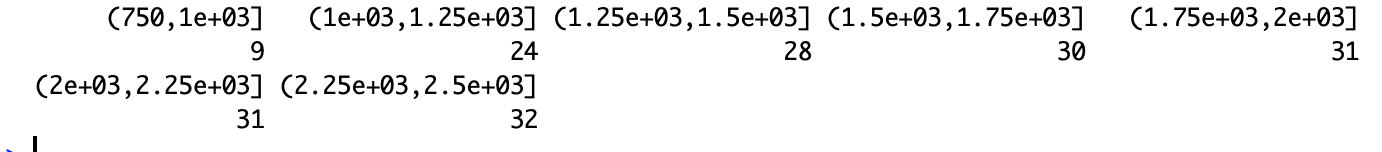
Description automatically generated

1. Find the cumulative relative frequency distribution of the team value.

teamValue.cumulativefreq = cumsum(teamValule.freq)

print(teamValue.cumulativefreq)

output:



1. Create a cumulative relative frequency graph of the team value.

cumrelfreqGraphTeamValue = c(0, teamValue.cumulativefreq)

plot(cumrelfreqGraphTeamValue,

main="cumulative frequency graph of the team value")

lines(cumrelfreqGraphTeamValue)

output:

A screenshot of a cell phone

Description automatically generated

1. Create a stem-and-leaf plot of the team revenue.

TeamRevenue.duration = (TeamRevenue)

stem(TeamRevenue.duration)

output:

A screenshot of a cell phone

Description automatically generated

1. Create a scatter plot of the team revenue against the team value.

plot(x = TeamRevenue,y = TeamValue,

xlab = "Revenue",

ylab = "Value",

main = "Scatter plot of the team revenue against the team value.")

A screenshot of a cell phone

Description automatically generated

1. Find the mean, median, all quartiles, 30th, 90th and 95th percentiles, range, and IQR of the team revenue.

TeamRevenue.mean = mean(TeamRevenue)

print(TeamRevenue.mean)

286.4688

TeamRevenue.median = median(TeamRevenue)

print(TeamRevenue.media)

269

TeamRevenue.quartiles = quantile(TeamRevenue)

print(TeamRevenue.quartiles)

A picture containing drawing

Description automatically generated

TeamRevenue.quartilesspecific = quantile(TeamRevenue,c(.30,.90,.95))

print(TeamRevenue.quartilesspecific)

A picture containing table

Description automatically generated

TeamRevenue.range = max(TeamRevenue)-min(TeamRevenue)

print(TeamRevenue.range)

310

TeamRevenue.iqr = IQR(TeamRevenue)

print(TeamRevenue.iqr)

40.75

1. Note that the dataset includes the entire population. Compute the variance of the team revenue. Define a function named "sd.p" that computes the population standard deviation. Use "sd.p" to compute the standard deviation of the team revenue.

varianceFunc = function(TeamRevenue) {

RevenueVarience = var(TeamRevenue)\*(NROW(TeamRevenue) - 1)/

NROW(TeamRevenue)

return(RevenueVarience)

}

sd.p = function(x) {

standardDeviation = sqrt(varianceFunc(x))

return(standardDeviation)

}

sd.p(TeamRevenue)

print(sd.p(TeamRevenue)

output: 59.88

1. Plot a box plot of the team value.

boxplot(TeamValue, horizontal = TRUE)

**A screenshot of a cell phone

Description automatically generated**

1. Define a function named "cov.p" that compute the population covariance. Use "cov.p" to compute the covariance of the team revenue and the team value. Compute the correlation coefficient of the team revenue and the team value.

cov.p = function(TeamRevenue, TeamValue) {

populationCov = cov(TeamRevenue, TeamValue)\*(NROW(TeamRevenue) - 1)/NROW(TeamRevenue)

return(populationCov)

}

corelationCo = cov.p(TeamRevenue, TeamValue)/(sd.p(TeamRevenue)\*sd.p(TeamValue))

print(corelationCo)

output: 0.9633148

**Problem 3**

Market-share-analysis company Net Applications monitors and reports on Internet browser usage. According to Net Applications, Google’s Chrome browser exceeded a 20% market share for the first time, with a 20.37% share of the browser market. For a randomly selected group of 20 Internet browser users, answer the following questions.

* 1. Compute the probability that exactly 8 of the 20 Internet browser users use Chrome as their Internet browser.

print(dbinom(8, size=20, prob=0.4))

output: 0.1797058

* 1. Compute the probability that at least 3 of the 20 Internet browser users use Chrome as their Internet browser.

print(1-dbinom(3, size=20, prob=0.15))

Output: 0.7571711

* 1. For the sample of 20 Internet browser users, compute the expected number, the variance and standard deviation of the number of Chrome users.

ExpectedNumber = 20\*.1

print(ExpectedNumber)

output: 2

varience = ExpectedNumber\*.9

print(varience)

output: 1.8

standardDeviation = sqrt(varience)

print(standardDeviation)

output: 1.341641

**Problem 4**

Over 500 million tweets are sent per day. Assume that the number of tweets per hour follows a Poisson distribution and that Bob receives on average 7 tweets during his lunch hour.

1. What is the probability that Bob receives no tweets during his lunch hour?

print(ppois(0,lambda = 7))

output:0.000911882

1. What is the probability that Bob receives at least 4 tweets during his lunch hour?

lessthan4 = ppois(0,lambda = 7)+ppois(1,lambda = 7)+ppois(2,lambda = 7)+ppois(3,lambda = 7)

print(1-lessthan4)

output: 0.8803915

1. What is the expected number of tweets Bob receives during the first 30 minutes of his lunch hour? What is the probability that Bob receives no tweets during the first 30 minutes of his lunch hour?

print(ppois(0,lambda = 3.5))

0.03019738

**Problem 5**

Television viewing reached a new high when the Nielsen Company reported a mean daily viewing time of 8.35 hours per household. Use a nor­ mal probability distribution with a standard deviation of 2.5 hours to answer the following questions about daily television viewing per household.

1. What is the probability that a household views television more than 3 hours a day?

mean = 8.35

sd = 2.5

z = (3-mean)/sd

print(1 - pnorm(z))

output: 0.9838226

1. What is the probability that a household spends 5 – 10 hours watching television more a day?

z1 = (5-mean)/sd

z2 = (10-mean)/sd

print(pnorm(z2)-pnorm(z1))

output: 0.6552504

1. How many hours of television viewing must a household have in order to be in the top 3% of all television viewing households?

z3 = qnorm(0.97)

print(z3\*sd+mean)

output: 13.05198

**Problem 6**

Comcast Corporation is the largest cable television company, the second largest Internet service provider, and the fourth largest telephone service provider in the United States. Generally known for quality and reliable service, the company periodically experiences unexpected service interruptions. On January 14, 2014, such an interruption occurred for the Comcast customers living in southwest Florida. When customers called the Comcast office, a recorded message told them that the company was aware of the service outage and that it was anticipated that service would be restored in two hours. Assume that two hours is the mean time to do the repair and that the repair time has an exponential probability distribution.

1. What is the probability that the cable service will be repaired in one hour or less?

print(pexp(1, rate=1/2))

output: 0.3934693

1. What is the probability that the repair will take between one hour and two hours?

exp1 = pexp(1, rate=1/2)

exp2 = pexp(1, rate=2/2)

print(exp2-exp1)

output: 0.2386512

1. For a customer who calls the Comcast office at 1:00 p.m., what is the probability that the cable service will not be repaired by 5:00 p.m.?

print(1-pexp(1, rate=4/2))

output: 0.1353353

**Problem 7**

The mean preparation fee H&R Block charged retail customers last year was $183. Use this price as the population mean and assume the population standard deviation of preparation fees is $50.

1. Now we randomly select 30 H&R Block retail customers. What are the values of the mean and the standard deviation of the sampling distribution of the sample mean?

populationMean = 183

populationSd = 50

sampleSd = populationSd/sqrt(30)

print(sampleMean)

print(sampleSd)

output: sample mean: 183

sample sd: 9.128709

1. What is the probability that the mean price for a sample of 30 H&R Block retail customers is within $8 (this value is generally called margin of error) of population mean? What is the probability that the mean price for a sample of 50 H&R Block retail customers is within $8 of population mean? What is the probability that the mean price for a sample of 100 H&R Block retail customers is within $8 of population mean?

# For sample size n = 30

print(pnorm(183+8, 183, populationSd/sqrt(30)) - pnorm(183-8, 183, populationSd/sqrt(30)))

output: 0.6191635

# For sample size n = 50

print(pnorm(183+8, 183, populationSd/sqrt(50)) - pnorm(183-8, 183, populationSd/sqrt(50)))

output: 0.742101

# For sample size n = 100

print(pnorm(183+8, 183, populationSd/sqrt(100)) - pnorm(183-8, 183, populationSd/sqrt(100)))

output: 0.8904014

1. What is the impact of sample size based on b) above?

When the sample size is larger, the sample mean is closer to the population mean. They are directly proportional.

1. What sample size would you recommend to have at least a .95 probability that the sample mean is within $8 of population mean?

margin = 8

sampleSize = (qnorm(1-.05/2)\*populationSd)^2/margin^2

print(sampleSize)

output: 150.057

1. Now, let’s assume we don’t know the population mean; but we still know the population standard deviation to be σ=$50. We randomly sampled 40 H&R Block retail customers and the mean price is $183. What is the probability that the population mean is within $5 of the sample mean?

sampleMeanProb =pnorm(5/(populationSd/40)) - pnorm(-5/(populationSd/40))

print(sampleMeanProb)

output: 0.9999367

1. We randomly sampled 100 H&R Block retail customers and the mean price is $183, assuming the population standard deviation is still $50. Construct a 90%, 95%, and 99% confidence interval of the population mean, respectively.

# For 90% confidence interval

marginOfError = qnorm(0.95)\*5

print(183+c(-marginOfError,marginOfError))

output: 174.7757 191.2243

# For 95% confidence interval

marginOfError = qnorm(0.975)\*5

print(183+c(-marginOfError,marginOfError))

output: 173.2002 192.7998

# For 99% confidence interval

marginOfError = qnorm(0.995)\*5

print(183+c(-marginOfError,marginOfError))

output: 170.1209 195.8791

1. Provide a practical interpretation of the above 90% confidence interval. What conclusions can you draw based on the 90%, 95%, and 99% confidence intervals you constructed above?

174.7757 191.2243 is the interval estimate when confidence coefficient is .90, It denotes that 90% of the parameters will lie within this interval. From confidence intervals 90,90,99 we observe that When the sample size is larger, the margin of error will be smaller hence confidence interval will be smaller.

1. When the population standard deviation is unknown, the best we can do is to replace it with the sample standard deviation, s. Just like the sample mean is a random variable, so is the sample standard deviation s. The replacement of σ with s adds more variability. Some adjustment to the Central Limit Theorem is thus necessary. It turns out that when the population standard deviation is unknown and the sample size n is sufficiently large, the sample statistic t=(X ̅-μ)/(s/√n) approximately follows a t distribution with a degree of freedom n-1. As a result, when we look for a confidence interval with σ unknown, we will replace normal distributions with t distribution. Use this result to find a 92% confidence interval for the population mean price that the retail customers pay for, given the sample mean is $183, the sample standard deviation is $50, and the sample size is 36. Comparing this result to that in question g), you should notice that the margin of error is slightly larger when the population standard deviation is unknown.

marginOfError = qt(0.96, 35)\*50/sqrt(36)

print(183+c(-marginOfError, marginOfError))

Output: 167.9748 198.0252

**Problem 8**

Suppose you have been hired by the Better Business Bureau (BBB) to investigate the settlement ratio of the complaints they have received. You plan to select a sample of n complaints to estimate the proportion of complaints the BBB is able to settle. We use p to denote the percentage or proportion of complaints settled among all the complaints that the BBB has received.

1. Let Y be the random variable, which indicates whether a complaint is settled. Without loss of generality, let Y be 1 if a complaint is settle, the probability of which is p; 0 if not settled. What probability distribution does Y follow? Compute its mean and standard deviation.

For a complaint to be settled or not it takes probability is 1 or 0, so it follows Bernoulli distribution.

Mean = µ

Standard deviation = sqrt(p(1-p))

1. Now suppose you select a random sample of n complaints and find that p ̅ of them have been settled (not surprisingly, p ̅ is called the sample proportion). Assume the sample size n is sufficiently large. What do we know about the probability distribution of p ̅ (sampling distribution of the sample proportion)?

The distribution can be approximated by normal distribution.

1. Let’s apply the results above and derive some confidence intervals. Note that the population proportion p is unknown. In order to compute the standard error , we substitute for p. As long as the sample size n is sufficiently large, a normal distribution would approximate the sample distribution of the sample proportion well enough. Suppose the sample proportion you’ve found is 0.6. Find a 95% confidence interval of the population proportion, if the sample size is 36, 100, and 400, respectively. What effect does the sample size n have on the resulting confidence interval?

# For sample size 36

print(0.6 - (qnorm(1 - (0.05/2))) \* sqrt (0.6 \* (1-0.6)/36))

output: 0.4399696

# For sample size 100

print(0.6 - (qnorm(1 - (0.05/2))) \* sqrt (0.6 \* (1-0.6)/100))

output: 0.5039818

# For sample size 400

print(0.6 - (qnorm(1 - (0.05/2))) \* sqrt (0.6 \* (1-0.6)/400))

output: 0.5519909

When sample size increases, the width of the confidence interval decreases for

Population proportion.

1. It is often the case that we have a target for margin of error in mind and we want to know the sample size needed to guarantee such a margin of error when the confidence level is given. Use the formula to compute the sample sizes needed when the respective value of m is 1%, 3%, and 5% and the respective confidence level is 90%, 95%, and 99%. You may fill out the table below and round your answers up to an integer.

print(((qnorm(1-.10/2))^2)/4\*0.01\*0.01)

print(((qnorm(1-.10/2))^2)/4\*0.03\*0.03)

print(((qnorm(1-.10/2))^2)/4\*0.05\*0.05)

print(((qnorm(1-.05/2))^2)/4\*0.01\*0.01)

print(((qnorm(1-.05/2))^2)/4\*0.03\*0.03)

print(((qnorm(1-.05/2))^2)/4\*0.05\*0.05)

print(((qnorm(1-.01/2))^2)/4\*0.01\*0.01)

print(((qnorm(1-.01/2))^2)/4\*0.03\*0.03)

print(((qnorm(1-.01/2))^2)/4\*0.05\*0.05)

|  |  |  |  |
| --- | --- | --- | --- |
| Confidence | Margin of error |  |  |
| levels | m = 1% | m = 3% | m = 5% |
| 90% | 6.763859e-05 | 0.0006087473 | 0.001690965 |
| 95% | 9.603647e-05 | 0.0008643282 | 0.002400912 |
| 99% | 0.0001658724 | 0.001492852 | 0.00414681 |

**Problem 9**

A consumer research group is interested in testing an automobile manufacturer’s claim that a new economy model will travel at least 25 miles per gallon of gasoline.

1. Provide a null and alternative hypothesis for the test.

Null Hypothesis Ho : µ >= 25

Alternate Hypothesis Ha : µ < 25

1. Suppose a test on 25 cars of this model indicates an average of 24 mpg, with a **sample** standard deviation of 3 mpg. Compute the value for the test statistic and the p-value.

sampleSize = 25

sampleMean = 24

populationMean = 25

sampleSd = 3

testStatistics = (sampleMean - populationMean)/(sampleSd/sqrt(sampleSize))

print(testStatistics)

output : -1.666667

pValue = pt(testStatistics, 24)

print(pValue)

output: 0.05429006

1. Suppose the significance level is 5%. Compute the critical value for the test statistic. What conclusion should we draw for the test? Provide a practical interpretation for this conclusion.

qt(0.05 , 24)

Output: -1.710882

Reject Ho when z ≥ zα

From above z-value is -1.666667 -1.666667 ≥ -1.710882 hence the null hypothesis is rejected at 0.05 level of significance.

1. Compute the critical value for the sample mean and determine the when we should reject H0 and when we should accept H0.

From above c, we know that z score is -1.666667

criticalValue=-1.666\*SD/sqrt(sampleSize) +25

print(criticalValue)

output: 24.0004

Reject Ho if Z <= - Za

1. Provide a practical interpretation of Type II error in this case.

Type 11 error occurs when we accept the null hypothesis when it is actually false.

When we accept that the mileage is atleast 25, when its actually less that 25 then type 11 error occurs.

1. Compute the probability of committing a Type II error (denoted as ) if the actual mileage is 23 mpg as well as the power of the test.

sampleSize = 25

standardDeviation = 3

SE = standardDeviation/sqrt(sampleSize)

alpha = .05

givenMean = 25

qValue = givenMean + qt(alpha, df=n-1)\*SE

actualMean = 23

beta = pt((q - actualMean)/SE, df=n-1, lower.tail=FALSE)

powerOfTest = 1-beta

print(powerOfTest)

output: beta = 0.05888562

powerOfTest = 0.9411144

probability of making type 11 error is 5%

**Problem 10**

Par, Inc., is a major manufacturer of golf equipment. Management believes that Par’s market share could be increased with the introduction of a cut-resistant, longer-lasting golf ball. Therefore, the research group at Par has been investigating a new golf ball coating designed to resist cuts and provide a more durable ball. The tests with the coating have been promising. One of the researchers voiced concern about the negative effect of the new coating on driving distances. Par would like the new cut-resistant ball to offer driving distances no worse than those of the current-model golf ball. To compare the driving distances for the two balls, 40 balls of both the new and current models were subjected to distance tests. The testing was performed with a mechanical hitting machine so that any difference between the mean distances for the two models could be attributed to a difference in the two models. The results of the tests, with distances measured to the nearest yard, are in the file “Golf.csv”. Let the current-model golf balls be population 1 and the new cut-resistant balls be population 2. Complete the following.

1. Formulate and present the rationale for a hypothesis test that Par could use to compare the driving distances of the current and new golf balls.

There are 40 balls from both new and current models are under the test.

Mean distance of current balls: µ1

Mean distance of new balls: µ2

Ho: µ1 = µ2

H1: µ1 != µ2

1. Provide descriptive statistical summaries of the data for each model; in particular, the sample mean, the sample standard deviation, and the sample size for each model.

golfTable <- read.csv("Golf.csv", header = TRUE)

currentBalls<-golfTable$Current

newBalls<-golfTable$New

mean(currentBalls)

mean(newBalls)

sd(currentBalls)

sd(newBalls)

NROW(currentBalls)

NROW(newBalls)

A screenshot of a cell phone

Description automatically generated

1. Compute the standard error for your test.

print(sqrt((sd(currentBalls)^2/NROW(currentBalls))+(sd(newBalls)^2/NROW(newBalls))))

output: 2.08904

1. Compute the degree of freedom for your test.

Current ball sample size n1 = 40

New ball sample size n2 = 40

Degree of freedom = n1+n2-2 = 78

1. Compute the test statistic for your test.

SE = sqrt((sd(currentBalls)^2/NROW(currentBalls))+(sd(newBalls)^2/NROW(newBalls)))

print(SE)

testStatistic = (mean(currentBalls)-mean(newBalls))/SE

print(testStatistic)

output: 1.328362

1. Compute the p value for your test.

z = (mean(currentBalls)-mean(newBalls))/SE

pvalue = 2\*pnorm(z, lower.tail=FALSE)

print(pvalue)

Output: .1840587

1. Suppose the significance level is set at 5%. What is your conclusion? Provide a practical interpretation of your conclusion in this case.

The rejection rule is if p-value < = significance level reject the Ho. P-value is 1.18 which is greater than significance level which is 0.05 . Hence we do not reject the p-value at 0.05 level of significance

1. Use the function t.test() in R to run the test directly to confirm your results above are correct.

t.test(currentBalls, newBalls, var.equal = TRUE)

A picture containing bird, flower

Description automatically generated

Comparing with the t.test, the descriptive statistics of mine matches with it.

1. What is the 95% confidence interval for the population mean driving distance of the current model?

t.test(currentBalls, var.equal = TRUE)

output:

95 percent confidence interval:

267.4757 273.0743

1. What is the 95% confidence interval for the population mean driving distance of the new model?

t.test(newBalls, var.equal = TRUE)

output:

95 percent confidence interval:

264.3348 270.6652

1. What is the 95% confidence interval for the difference between the means of the two populations?

t.test(currentBalls, newBalls)

A picture containing bird

Description automatically generated

the 95% confidence interval for the difference between the means of the two populations is -1.384937 6.934937

**Problem 11**

The variance in a production process is an important measure of the quality of the process. A large variance often signals an opportunity for improvement in the process by finding ways to reduce the process variance. The following sample data show the weight of bags (in pounds) produced on two machines: machine 1 and 2.

m1 = (2.95, 3.45, 3.50, 3.75, 3.48, 3.26, 3.33, 3.20, 3.16, 3.20, 3.22, 3.38, 3.90, 3.36, 3.25, 3.28, 3.20, 3.22, 2.98, 3.45, 3.70, 3.34, 3.18, 3.35, 3.12)

m2 = (3.22, 3.30, 3.34, 3.28, 3.29, 3.25, 3.30, 3.27, 3.38, 3.34, 3.35, 3.19, 3.35, 3.05, 3.36, 3.28, 3.30, 3.28, 3.30, 3.20, 3.16, 3.33)

1. Provide descriptive statistical summaries of the data for each model; in particular, the sample variance and the sample size for each machine.

> mean(m1)

[1] 3.3284

> mean(m1)

[1] 3.3284

> mean(m2)

[1] 3.278182

> var(m1)

[1] 0.048889

> var(m2)

[1] 0.005901299

> length(m1)

25

> length(m2)

22

> sd(m1)

[1] 0.2211086

> sd(m2)

[1] 0.07681991

1. Conduct a statistical test to determine whether there is a significant difference between the variances in the bag weights for two machines. First, clearly formulating your hypotheses below.

H0 : σ1^2=σ2^2

Ha : σ1^2 !=σ2^2

Null Hypothesis is there is no difference between the variances in the bag weights . Alternative Hypothesis is there is difference between the variances in the bag weights.

1. Compute the test statistic.

testStatistic=var(m1)/var(m2)

print(testStatistic)

output: 8.284448

1. Compute the p value.

pValue = 2\*pf(testStatistic,df1 =length(m1)-1, df2 = length(m2)-1, lower.tail = FALSE)

print(pValue)

output: 7.22158e-06

1. Use a .05 level of significance to compute both critical values for your test statistic.

significanceLevel=0.05

criticalValueLower=qf(significanceLevel/2, df1 = length(m1)-1, df2 = length(m2)-1)

criticalValueUpper=qf(1-significanceLevel/2, df1 = length(m1)-1, df2 = length(m2)-1)

print(criticalValueLower)

output: 0.4327282

print(criticalValueUpper)

output: 2.367526

1. Use a .05 level of significance. What is your conclusion?

Reject Ho if F>=Falpha/2. The test statistic “8.284448” is greater than the critical value “2.367526”. Hence reject Null Hypothesis. Which shows there is difference between the variances in the bag weights.

1. Use the function var.test() in R to run the test directly to confirm your results above are correct.

Var.test(m1,m2)

A picture containing bird, flower

Description automatically generated

The results for above and my calculations are same.

1. Construct a 95% confidence interval for the variance of the weight of bags produced on machine 1.

confidanceInterval = 0.05

criticalValueLower <-qchisq(confidanceInterval/2,df=length(m1)-1)

criticalValueUpper<-qchisq(1-confidanceInterval/2,df=length(m1)-1)

confidenceIntervalLower<-(length(m1)-1)\*var(m1)/ criticalValueUpper

confidenceIntervalUpper <-(length(m1)-1)\*var(m1)/ criticalValueLower

print(confidenceIntervalLower)

print(confidenceIntervalUpper)

output: 0.02980728 ,0.09461509

1. Construct a 95% confidence interval for the standard deviation of the weight of bags produced on machine 2.

confidanceInterval = 0.05

criticalValueLower <-qchisq(confidanceInterval/2,df=length(m2)-1)

criticalValueUpper<-qchisq(1-confidanceInterval/2,df=length(m2)-1)

confidenceIntervalLower<-(length(m2)-1)\*var(m2)/criticalValueUpper

confidenceIntervalUpper <-(length(m2)-1)\*var(m2)/criticalValueLower

print(sqrt(confidenceIntervalLower))

print(sqrt(confidenceIntervalUpper))

output: 0.0591015 , 0.1097806

1. Which machine, if either, provides the greater opportunity for quality improvements?

The machine having larger variance will provide provides the greater opportunity for quality improvements hence it will be machine 1.

**Problem 12**

A Bloomberg Businessweek subscriber study asked, “in the past 12 months, when traveling for business, what type of airline ticket did you purchase most often?” A second question asked if the type of airline ticket purchased most often was for domestic or international travel. Sample data obtained are shown in the following table.

| **Type of Ticket** | **Domestic Flight** | **International Flight** |
| --- | --- | --- |
| First class | 29 | 22 |
| Business class | 95 | 121 |
| Economy class | 518 | 135 |

1. The study wants to test whether the type of ticket is independent of the type of flight. Clearly state the null and alternative hypotheses.

Ho: type of ticket and type of flight are independent.

Ha: type of ticket and type of flight are not independent.

1. Compute the expected frequencies by completing the table below.

Expected frequency is the product of column and row total divided by total sample size n.

|  |  |  |  |
| --- | --- | --- | --- |
| **Type of Ticket** | **Domestic Flight** | **International Flight** | **Total** |
| **First Class** | 35.5 | 15.4 | 51 |
| **Business Class** | 150.7 | 65.2 | 216 |
| **Economy Class** | 455.6 | 197.3 | 653 |
| **Total** | 642 | 278 | 920 |

1. Compute the test statistic.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| fij | eij | | fij-eij | (fij-eij)^2 | | | ((fij-eij)^2)/eij | |
| 29 | 35.5 | | -6.5 | 42.25 | | | 1.19 | |
| 95 | 150.7 | | -55.7 | 3102.49 | | | 20.58 | |
| 518 | 455.6 | | 62.4 | 3893.7 | | | 8.54 | |
| 22 | 15.4 | | 6.6 | 43.56 | | | 2.82 | |
| 121 | 65.2 | | 55.8 | 3113.6 | | | 47.75 | |
| 135 | 197.3 | | -62.3 | 3881.29 | | | 19.67 | |
|  | |  | | |  |  | |  | |  |  |

Test statistic = 100.55

1. At 5% significance level, compute the critical value for the test statistic and the p value for the test. Draw your conclusion.

testStatistic = 100.55

pValue<-pchisq(testStatistic, df=2, lower.tail=FALSE)

print(pValue)

output: 1.465025e-22

significanceLvel=0.05

criticalValue<-qchisq(significanceLvel, df=2, lower.tail=FALSE)

print(criticalValue)

output: 5.991465

Null hypothesis is rejected as test statistic (100.55) is greater than critical value(5.99).

1. Use the function chisq.test() in R to run the test directly to confirm your results above are correct.

BloombergData<- as.table(rbind(c(29,22), c(95,121),c(518,135)))

print(BloombergData)

chisq.test(BloombergData)

A picture containing knife, table

Description automatically generated

**Problem 13**

As part of a long-term study of individuals 65 years of age or older, sociologists and physicians at the Wentworth Medical Center in upstate New York investigated the relationship between geographic location and depression. A sample of 60 individuals, all in reasonably good health, was selected; 20 individuals were residents of Florida, 20 were residents of New York, and 20 were residents of North Carolina. Each of the individuals sampled was given a standardized test to measure depression. The data collected follow; higher test scores indicate higher levels of depression. These data are contained in the file “Medical1.csv”. A second part of the study considered the relationship between geographic location and depression for individuals 65 years of age or older who had a chronic health condition such as arthritis, hypertension, and/or heart ailment. A sample of 60 individuals with such conditions was identified. Again, 20 were residents of Florida, 20 were residents of New York, and 20 were residents of North Carolina. The levels of depression recorded for this study follow. These data are contained in the file named “Medical2.csv”.

For the rest of this problem, we will solely use “Medical1.csv”. If you wish to look further, particularly into two-way ANOVA, feel free to utilize the second data file on your own.

1. Use descriptive statistics to summarize the data from the first study; in particular, the sample mean for each location, the grand mean, the sample size for each location, the sample size for the entire dataset, the sample standard deviation for each location, and the sample standard deviation for the entire dataset.

install.packages("fBasics")

library(fBasics)

medicalTable <- read.csv("Medical1.csv", header = TRUE)

basicStats(medicalTable)

A screenshot of a cell phone

Description automatically generated

1. Clearly state the hypotheses being tested.

Null hypothesis, Ho : The depression test scores of healthy people in the given three given locations are equal.

Ho : µ1= µ 2= µ 3

Alternate hypothesis, Ha : The depression test scores of healthy people in the given three given locations are not equal.

Ha : µ1≠ µ 2≠ µ 3

1. Use the descriptive summary data from part a) to compute SSE, its degree of freedom, and MSE.

Sum of Squared estimate of Errors, SSE = (20-1)(2.139233)2 + (20-1)(2.200478)2 + (20-1)(2.837252)2 = 331.08

Degree of Freedom, DOF = nT-k

=60-3

=57

Mean Squared Error, MSE = SSE/DOF

= 331.9/57

= 5.80

1. Use the descriptive summary data from part a) to compute SSTR, its degree of freedom, and MSTR.

Mean of all three locations = (5.55+8.00+7.05)/3 = 6.86

SSTR = 20(5.55-6.866667)2 + 20(8-6.866667)2 + 20(7.05-6.866667)2

= 61.03333

DOF = k-1 = 3-1

= 2

MSTR = SSTR/DOF

=61.03333/2=30.51667

1. Use the descriptive summary data from part a) to compute SST and confirm that it is equal to the sum of SSE and SSTR.

SST = SSTR+SSE = 61.03333+331.9

= 392.9333

SD for three locations = 2.5866

SST = (60-1)\*(2.58) \*\*2 = 392.7276

Thus they are equal.

1. Compute the test statistic for your test.

F = MSTR/MSE = 30.51/5.80

= 5.2603

1. Compute the p value for your test.

F = 5.2603

p\_value <- 1-pf(F,df1=2, df2=57)

output:

0.008007539

1. Suppose the significance level is set at 5%. What is your conclusion? Provide a practical interpretation of your conclusion in this case.

Alpha = 0.05

p-value = 0.00800

The null hypothesis is rejected when p-value is <= alpha. Hence we reject the null

Hypothesis, ie. not all three locations have same test scores.

1. Use the function aov() in R to run the test directly to confirm your results above are correct. If you need help with the function aov(), see <http://www.r-tutor.com/elementary-statistics/analysis-variance/completely-randomized-design>

mat = as.matrix(medicalTable)

print(mat)

A screenshot of a cell phone

Description automatically generated

r = c(t(mat))

print(r)

[1] 3 8 10 7 11 7 7 9 3 3 7 5 8 8 11 8 7 8 8 8 4 5 4 3 5 13 7 2 10 8 6 6 8 2 8 7 6 12 3 6 8 9 9 6 8 7 8

[48] 12 5 5 6 4 7 3 7 7 8 3 8 11

f = c("Florida", "New.York", "North.Carolina")

# number of treatments

k = 3

# observations in one treatment

n = 20

tm = gl(k, 1, n\*k, factor(f))

print(tm)

A screenshot of a cell phone

Description automatically generated

av = aov(r ~ tm)

summary(av)

A screenshot of a cell phone

Description automatically generated

**Problem 14**

As part of a study on transportation safety, the U.S. Department of Transportation collected data on the number of fatal accidents per 1000 licenses and the percentage of licensed drivers under the age of 21 in a sample of 42 cities. Data collected over a one-year period follow. These data are contained in the file named “Safety.csv”.

1. Find the sample mean and standard deviation for each variable. Round your answers to the nearest thousandth.

safetyTable = read.csv("Safety.csv", header = TRUE)

basicStats(safetyTable)

A screenshot of a cell phone

Description automatically generated

1. Use the function lm() in R to run a simple linear regression model on the data provided. Use the function summary() in R to generate the regression output. Use the function anova() in R to generate the corresponding ANOVA table. You ought to be able to determine which is the dependent variable and which is the independent variable in this SLR model. If you need help with lm() function, see <http://www.r-tutor.com/elementary-statistics/simple-linear-regression/estimated-simple-regression-equation>

safetyTable

lm.safetyTable = lm(Fatal.Accidents.per.1000 ~ Percent.Under.21, data =safetyTable)

summary(lm.safetyTable)

anova(lm.safetyTable)

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

1. Write down the estimated regression function below and provide a practical interpretation of the coefficient of the independent variable.

0.2871\*X -1.5974 (percentage of drivers under 21)

For percentage of drivers under age 21, the number of fatal accidents per 1000 licenses increases by 0.2871.

1. Please find a 95% confidence interval for the coefficient of the independent variable and provide a practical interpretation of this interval.

0.228 to 0.347

For every additional percentage of drivers under 21, there is a 95% chance that increase in the average of fatal accidents per 1000 licenses is between 0.228 and 0.347.

1. At the 5% level of significance, is there a significant relationship between the two variables? Why or why not?

The p-value 3.794e-12 is less that the level of significance 0.05 predicting that there is significant relationship between the two variables. Thus the null hypothesis is rejected.

1. What is the value of the coefficient of determination for this simple linear regression model? Provide a brief interpretation of this value.

Coefficient of determination, r2  = 33.15/47.03 = 0.7048

70.48% of variation in the fatal accidents can be explained by the linear relationship between drivers below 21 and the fatal accidents.

1. Use the information from the ANOVA table to compute the standard error of estimate, a.k,a, residual standard error. This value must match the residual standard error in the regression summary.

Residual standard error: 0.5894, calculated from question b.

Sqrt(0.347334) = 0.5894

1. What is the point estimate of the **expected** number of fatal accidents per 1000 licenses if there are 10% drivers under age in a city? (**Show your work; Round to the nearest thousandth**)

Intercept = - 1.5974

Std = 0.2871

y = 0.2871x - 1.5974

= 0.2871(10) – 1.5974

= 1.274

1. Suppose we want to develop a 95% confidence interval for the average number of fatal accidents per 1000 licenses for all the cities with 10% of drivers under age 21. What is the estimate of the standard deviation for this confidence interval? (**Show your work; Round to the nearest thousandth**)

A picture containing object, clock

Description automatically generated

= 0.59 \*

= 0.59\*0.19422299

= 0.114

1. Suppose we want to develop a 95% confidence interval for the average number of fatal accidents per 1000 licenses for all the cities with 10% of drivers under age 21. Compute the t value and the margin of error needed for this confidence interval. (**Show your work; Round to the nearest thousandth**)

qt((1+0.95)/2, 40)

output:

t-value = 2.021

moe = 2.021\*0.114 = 0.231

1. Provide a 95% confidence interval for the average number of fatal accidents per 1000 licenses for all the cities with 10% of drivers under age 21 and a practical interpretation to this confidence interval.

A close up of a logo

Description automatically generated

Margin of error = 0.230394

Confidence interval = 1.274 plus or minus -0.231

= 1.043 to 1.504

It is 95% confident that the average number of fatal accidents per 1000 licenses for all the cities with 10% of drivers under age 21 is between 1.043 and 1.504.

1. Suppose we want to develop a 95% prediction interval for the number of fatal accidents per 1000 licenses for a city with 10% of drivers under age 21. What is the estimate of the standard deviation for this prediction interval? (**Show your work; Round to the nearest thousandth**)

A picture containing object, clock

Description automatically generated

= 0.59\*

= 0.59\*1.0186

= 0.600

1. Suppose we want to develop a 95% prediction interval for the number of fatal accidents per 1000 licenses for a city with 10% of drivers under age 21. Compute the margin of error needed for this prediction interval. (**Show your work; Round to the nearest thousandth**)

moe = -2.021\*0.600

=-1.216

1. Provide a 95% prediction interval for the number of fatal accidents per 1000 licenses for a city with 10% of drivers under age 21 and a practical interpretation to this prediction interval.

A close up of a logo

Description automatically generated

Prediction Interval = 1.274 plus or minus -1.216

= 0.059 to 2.489