**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).

**Name: Shruti Ravindra Joshi**

**NET Id: jd4923**

**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.

test1 <-read.csv("/Users/shrutijoshi/Documents/Majors.csv",header=T)

test1

table(test1$Bachelor)

bachelor.freq <- table(test1$Bachelor)

bachelor.freq

bachelor.freq <- bachelor.freq[order(bachelor.freq, decreasing =TRUE)]

bachelor.freq

**Output:**

> bachelor.freq <- table(test1$Bachelor)

> bachelor.freq

B CSE E H NSM O SBS

21 9 6 16 8 24 16

> bachelor.freq <- bachelor.freq[order(bachelor.freq, decreasing =TRUE)]

> bachelor.freq

O B H SBS CSE NSM E

24 21 16 16 9 8 6

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

test2 <- read.csv("/Users/shrutijoshi/Documents/Majors.csv",header=T)

test2

table(test2$Master)

master.freq <- table(test2$Master)

master.freq

masterrelfreq <- master.freq/ nrow(master.freq)

masterrelfreq

**Output:**

B CSE E H NSM O SBS

27 9 24 8 2 24 6

> masterrelfreq <- master.freq/ nrow(master.freq)

> masterrelfreq

B CSE E H NSM

3.8571429 1.2857143 3.4285714 1.1428571 0.2857143

O SBS

3.4285714 0.8571429

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

jointtable <- table(test1$Bachelor, test2$Master)

jointtable

**Output:**

B CSE E H NSM O SBS

B 3 2 6 2 0 5 3

CSE 4 0 4 0 0 1 0

E 1 1 2 1 0 1 0

H 4 2 4 1 2 3 0

NSM 2 0 3 0 0 2 1

O 9 4 1 1 0 7 2

SBS 4 0 4 3 0 5 0

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

jointtable <- table(test1$Bachelor, test2$Master)

jointrelfreq <- jointtable/ nrow(jointtable)

jointrelfreq

**Output:**

B CSE E H NSM O SBS

B 0.4285714 0.2857143 0.8571429 0.2857143 0.0000000 0.7142857 0.4285714

CSE 0.5714286 0.0000000 0.5714286 0.0000000 0.0000000 0.1428571 0.0000000

E 0.1428571 0.1428571 0.2857143 0.1428571 0.0000000 0.1428571 0.0000000

H 0.5714286 0.2857143 0.5714286 0.1428571 0.2857143 0.4285714 0.0000000

NSM 0.2857143 0.0000000 0.4285714 0.0000000 0.0000000 0.2857143 0.1428571

O 1.2857143 0.5714286 0.1428571 0.1428571 0.0000000 1.0000000 0.2857143

SBS 0.5714286 0.0000000 0.5714286 0.4285714 0.0000000 0.7142857 0.0000000

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

Barchartdata <- table(test1$Bachelor)

barplot(Barchartdata, xlab = "Bachelors 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(test2$Master)

lbls <- paste(names(Piechartdata), "\n", Piechartdata, sep="")

pie(Piechartdata, labels = lbls, main = "PieChart")

**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.

Answer the following questions in R.

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

test1 <- read.csv("/Users/shrutijoshi/Documents/Study Docs/BAN 602/R Assignment/NFLTeamValue.csv",header=T)

test1

table(test1$Revenue)

revenue.freq <- table(test1$Revenue)

revenue.freq

range(revenue.freq)

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

breaks

**Output:**

range(revenue.freq)

[1] 1 2

> breaks = seq(200, 550, by=50)

> breaks

[1] 200 250 300 350 400 450 500 550

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

revenue.freq <- test1$Revenue

hist(revenue.freq)

**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 <- test1$Value

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

breaks

teamvalue.freq = table(test1$Value)

teamvalue.freq

teamvalue.relfreq = teamvalue.freq / nrow(faithful)

teamvalue.relfreq

**Output:**

>teamvalue <- test1$Value

> breaks = seq(750, 2500, by=250)

> breaks

[1] 750 1000 1250 1500 1750 2000 2250 2500

> teamvalue.freq = table(test1$Value)

> teamvalue.freq

825 840 870 875 900 924 933 949 961 1004 1005 1007 1009 1055 1057 1067 1074 1081 1118

1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

1161 1183 1200 1224 1227 1252 1314 1380 1450 1550 1700 1800 2300

1 1 1 1 1 1 1 1 1 1 1 1 1

> teamvalue.relfreq = teamvalue.freq / nrow(faithful)

> teamvalue.relfreq

825 840 870 875 900 924 933 949

0.003676471 0.003676471 0.003676471 0.003676471 0.003676471 0.003676471 0.003676471 0.003676471

961 1004 1005 1007 1009 1055 1057 1067

0.003676471 0.003676471 0.003676471 0.003676471 0.003676471 0.003676471 0.003676471 0.003676471

1074 1081 1118 1161 1183 1200 1224 1227

0.003676471 0.003676471 0.003676471 0.003676471 0.003676471 0.003676471 0.003676471 0.003676471

1252 1314 1380 1450 1550 1700 1800 2300

0.003676471 0.003676471 0.003676471 0.003676471 0.003676471 0.003676471 0.003676471 0.003676471

>

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

revenue.freq = table(test1$Revenue)

revenue.freq

revenue.cumfreq = cumsum(revenue.freq)

revenue.cumfreq

**Output:**

>revenue.freq = table(test1$Revenue)

> revenue.freq

229 234 239 245 248 250 252 253 255 256 260 264 266 267 268 270 271 276 282 283 292 298 306 320

1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 2 1 2 1 1 1 1 1 1

321 338 381 408 539

1 1 1 1 1

> revenue.cumfreq = cumsum(revenue.freq)

> revenue.cumfreq

229 234 239 245 248 250 252 253 255 256 260 264 266 267 268 270 271 276 282 283 292 298 306 320

1 2 3 4 5 7 8 9 10 11 12 13 14 15 16 18 19 21 22 23 24 25 26 27

321 338 381 408 539

28 29 30 31 32

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

revenue.freq = table(test1$Revenue)

revenue.freq

revenue.cumfreq = cumsum(revenue.freq)

revenue.cumfreq

plot(revenue.cumfreq, xlab = "Team Value Revenue", ylab = "Cum. Frequency", main = "cumulative frequency graph of the team revenue")

lines(revenue.cumfreq)

**Output:**

A close up of a map

Description automatically generated

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

teamvalue.freq = table(test1$Value)

#teamvalue.freq

teamvalue.cumfreq = cumsum(teamvalue.freq)

#teamvalue.cumfreq

teamvalue.cumrelfreq = teamvalue.cumfreq / nrow(test1)

teamvalue.cumrelfreq

**Output:**

>teamvalue.freq = table(test1$Value)

> #teamvalue.freq

> teamvalue.cumfreq = cumsum(teamvalue.freq)

> #teamvalue.cumfreq

> teamvalue.cumrelfreq = teamvalue.cumfreq / nrow(test1)

> teamvalue.cumrelfreq

825 840 870 875 900 924 933 949 961 1004 1005 1007

0.03125 0.06250 0.09375 0.12500 0.15625 0.18750 0.21875 0.25000 0.28125 0.31250 0.34375 0.37500

1009 1055 1057 1067 1074 1081 1118 1161 1183 1200 1224 1227

0.40625 0.43750 0.46875 0.50000 0.53125 0.56250 0.59375 0.62500 0.65625 0.68750 0.71875 0.75000

1252 1314 1380 1450 1550 1700 1800 2300

0.78125 0.81250 0.84375 0.87500 0.90625 0.93750 0.96875 1.00000

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

teamvalue.freq = table(test1$Value)

#teamvalue.freq

teamvalue.cumfreq = cumsum(teamvalue.freq)

#teamvalue.cumfreq

teamvalue.cumrelfreq = teamvalue.cumfreq / nrow(test1)

#teamvalue.cumrelfreq

cumrelfreq0 = c(0, teamvalue.cumrelfreq)

plot(cumrelfreq0, main = "Cumulative Relative Frequency graph of Team Value", xlab = "Team Value",

ylab = "Cumulative Rel. Freq")

lines(cumrelfreq0)

**Output:**

**A screenshot of a cell phone

Description automatically generated**

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

revenue = test1$Revenue

stem(revenue)

**Output:**

The decimal point is 2 digit(s) to the right of the |

2 | 334

2 | 555555666677777788889

3 | 01224

3 | 8

4 | 1

4 |

5 | 4

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

revenue = test1$Revenue

teamvalue = test1$Value

plot(revenue, teamvalue, xlab = "Team Revenue", ylab = "Team Value", main = "Plot of Team Revenue against Team Value")

**Output:**

**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.

revenue = test1$Revenue

mean(revenue)

median(revenue)

quantile(revenue)

quantile(revenue, c(.3, .9, .95))

max(revenue) - min(revenue)

IQR(revenue)

**Output:**

>mean(revenue)

[1] 286.4688

> median(revenue)

[1] 269

> quantile(revenue)

0% 25% 50% 75% 100%

229.00 252.75 269.00 293.50 539.00

> quantile(revenue, c(.3, .9, .95))

30% 90% 95%

255.30 336.30 393.15

> max(revenue) - min(revenue)

[1] 310

> IQR(revenue)

[1] 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.

sd.p <- function(a){

sqrt(sum((a-mean(a))^2/length(a)-1))

}

revenue = test1$Revenue

sd.p(revenue)

**Output:**

> sd.p(revenue)

[1] 59.61962

1. Plot a box plot of the team value.

teamvalue = test1$Value

boxplot(teamvalue, horizontal = TRUE, main ="Box Plot of Team Value")

**Output:**

**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(a,b,c)

{

cov(a,b)\*((nrow(c)-1)/nrow(c))

}

revenue = test1$Revenue

teamvalue = test1$Value

cov.p(revenue, teamvalue, test1)

sd.p=function(a)

{

sd(a)\*(sqrt((length(a)-1)/length(a)))

}

cov\_rev\_val=cov.p(revenue, teamvalue, test1)

cor\_rev\_val=cov\_rev\_val/( sd.p(revenue) \* sd.p(teamvalue))

cor\_rev\_val

**Output:**

> cov.p(revenue, teamvalue, test1)

[1] 17974.78

> cor\_rev\_val

[1] 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.

pbinom(8, 20, prob = 0.4)

**Output:**

> pbinom(8, 20, prob = 0.4)

[1] 0.5955987

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

dbinom(0, size =20, prob = 0.15)+

+ dbinom(1, size = 20, prob = 0.15)+

+ dbinom(2, size = 20, prob = 0.15)+

+ dbinom(3, size = 20, prob = 0.15)

**Output:**

[1] 0.6477252

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

ENumber = mean(rbinom(20, 1, 0.2037))

ENumber

variance = sum(((rbinom(20, 1, 0.2037) - ENumber)^2)/19)

variance

standarddev = sqrt(variance)

standarddev

**Output:**

> ENumber = mean(rbinom(20, 1, 0.2037))

> ENumber

[1] 0.3

>

> variance = sum(((rbinom(20, 1, 0.2037) - ENumber)^2)/19)

> variance

[1] 0.1789474

>

> standarddev = sqrt(variance)

> standarddev

[1] 0.4230217

**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?

ppois(0, lambda = 7) #lower tail

ppois(0, lambda = 7, lower = FALSE) #upper tail

**Output:**

> ppois(0, lambda = 7) #lower tail

[1] 0.000911882

>

> ppois(0, lambda = 7, lower = FALSE) #upper tail

[1] 0.9990881

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

ppois(4, lambda = 7, lower = FALSE)

**Output:**

> ppois(4, lambda = 7, lower = FALSE)

[1] 0.8270084

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?

#Expected number during 30 mins

ExNum = (7/60)\*30

ExNum

#No tweets during 30 mins of lunch

ppois(0, lambda = 3.5)

**Output:**

> ExNum = (7/60)\*30

> ExNum

[1] 3.5

>

> #No tweets during 30 mins of lunch

> ppois(0, lambda = 3.5)

[1] 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?

pnorm(3, mean = 8.35, sd = 2.5, lower.tail = FALSE)

**Output:**

> pnorm(3, mean = 8.35, sd = 2.5, lower.tail = FALSE)

[1] 0.9838226

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

pnorm(10, mean = 8.35, sd = 2.5) - pnorm(5, mean = 8.35, sd = 2.5)

**Output:**

> pnorm(10, mean = 8.35, sd = 2.5) - pnorm(5, mean = 8.35, sd = 2.5)

[1] 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?

mean = 8.35

sd = 2.5

z = qnorm(0.97)

hours = (z\*sd+mean)

hours

**Output:**

> hours

[1] 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?

pexp(1, rate = 1/2)

**Output:**

> pexp(1, rate = 1/2)

[1] 0.3934693

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

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

**Output:**

> pexp(2, rate = 1/2) - pexp(1, rate = 1/2)

[1] 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.?

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

prob

**Output:**

> prob = 1-(pexp(4, rate =1/2))

> prob

[1] 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?

#Mean is Sample mean as sample size is 30

mean = 183

sampmean = mean

sampmean

#Standard Error

x = runif(30)

x

length(x)

sd.p <- 50

stderror = function(x)

{

sd.p/sqrt(length(x))

}

stderror(x)

**Output:**

> sampmean

[1] 183

>

> #Standard Error

> x = runif(30)

> x

[1] 0.611464965 0.437431434 0.414362035 0.336446240 0.721063387 0.917999131 0.907546820

[8] 0.339225816 0.845337087 0.427990918 0.260482426 0.599086100 0.732871598 0.664345292

[15] 0.583024533 0.542785482 0.808638457 0.200056649 0.244474856 0.596933120 0.839688310

[22] 0.253130396 0.433468720 0.952766549 0.020222540 0.003060182 0.879819887 0.506084046

[29] 0.079907203 0.325968093

> length(x)

[1] 30

> sd.p <- 50

> stderror = function(x)

{

sd.p/sqrt(length(x))

}

> stderror(x)

[1] 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

sd.p = 50

print(pnorm(183+8, 183, sd.p/sqrt(30)) - pnorm(183-8, 183, sd.p/sqrt(30)))

**Output:**

> print(pnorm(183+8, 183, sd.p/sqrt(30)) - pnorm(183-8, 183, sd.p/sqrt(30)))

[1] 0.6191635

# For sample size n = 50

sd.p = 50

print(pnorm(183+8, 183, sd.p/sqrt(50)) - pnorm(183-8, 183, sd.p/sqrt(50)))

**Output:**

> print(pnorm(183+8, 183, sd.p/sqrt(50)) - pnorm(183-8, 183, sd.p/sqrt(50)))

[1] 0.742101

# For sample size n = 100

sd.p = 50

print(pnorm(183+8, 183, sd.p/sqrt(100)) - pnorm(183-8, 183, sd.p/sqrt(100)))

**Output:**

> print(pnorm(183+8, 183, sd.p/sqrt(100)) - pnorm(183-8, 183, sd.p/sqrt(100)))

[1] 0.8904014

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

The sample size and sample mean & population mean are directly proportional. When the sample size is larger the sample mean is nearer the population mean.

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

z = qnorm(1-0.05/2)

sigma = 50

E = 8

n = z^2 \* sigma^2 / E^2

n

**Output:**

> z = qnorm(1-0.05/2)

> sigma = 50

> E = 8

> n = z^2 \* sigma^2 / E^2

> n

[1] 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?

psd = 50

MeanProb =pnorm(5/(psd/40)) - pnorm(-5/(psd/40))

print(MeanProb)

**Output:**

> print(MeanProb)

[1] 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

moe = qnorm(0.95)\*5

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

# For 95% confidence interval

moe = qnorm(0.975)\*5

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

# For 99% confidence interval

moe = qnorm(0.995)\*5

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

**Output:**

# For 90% confidence interval

> moe = qnorm(0.95)\*5

> print(183+c(-moe,moe))

[1] 174.7757 191.2243

>

> # For 95% confidence interval

> moe = qnorm(0.975)\*5

> print(183+c(-moe,moe))

[1] 173.2002 192.7998

>

>

> # For 99% confidence interval

> moe = qnorm(0.995)\*5

> print(183+c(-moe,moe))

[1] 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?

The confidence interval at 90% is 170.1209 to 195.8791. It interprets that 90% parameters will lie within this range. The confidence intervals at 90%, 95% and 99% interpret that when the sample size is larger, the margin of error will be smaller and hence the 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.

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

n = (183+c(-moe, moe))

n

**Output:**

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

> n = (183+c(-moe, moe))

> n

[1] 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.

Output:

Y follows Bernouli distribution which is discrete probability distribution and takes either 1 or 0 with probability p or 1-p.

Mean: E(Y) = µ

Standard Deviation: sd = sqrt(p\*(1-p))

Mean of Sample: P1

Standard Deviation: SD = sqrt(P1\*(1-p)/n)

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)?

Output:

The Probability distribution is approximately a Normal Distribution.

Hence, if sample size is large, it will be

n\*p >= 5

n\*(1-p) >= 5

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?

As the sample size increases, the width of the confidence interval also increases.

# For 95% confidence interval for n=36 & p=0.6

#moe = z \* sqrt(p\*(1-p)/n)

moe = qnorm(sqrt(0.6\*0.4/36))

moe

print(0.6+c(-moe,moe))

# For 95% confidence interval for n=100

moe = qnorm(sqrt(0.6\*0.4/100))

moe

print(0.6+c(-moe,moe))

# For 95% confidence interval for n=400

moe = qnorm(sqrt(0.6\*0.4/400))

moe

print(0.6+c(-moe,moe))

**Output:**

# For 95% confidence interval for n=36 & p=0.6

> #moe = z \* sqrt(p\*(1-p)/n)

> moe = qnorm(sqrt(0.6\*0.4/36))

> moe

[1] -1.394061

> print(0.6+c(-moe,moe))

[1] 1.9940606 -0.7940606

>

> # For 95% confidence interval for n=100

> moe = qnorm(sqrt(0.6\*0.4/100))

> moe

[1] -1.654728

> print(0.6+c(-moe,moe))

[1] 2.254728 -1.054728

>

> # For 95% confidence interval for n=400

> moe = qnorm(sqrt(0.6\*0.4/400))

> moe

[1] -1.96868

> print(0.6+c(-moe,moe))

[1] 2.56868 -1.36868

When the 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.

The sample sizes computed using the formula is 1691, 267 and 166.

m = 0.01

α = 0.1

n = (qnorm(1 - α / 2)) ^ 2 / (4 \* m) ^ 2

n

m = 0.03

α = 0.1

n = (qnorm(1 - α / 2)) ^ 2 / (4 \* m) ^ 2

n

m = 0.05

α = 0.1

n = (qnorm(1 - α / 2)) ^ 2 / (4 \* m) ^ 2

n

m=0.01

α = 0.05

n = (qnorm(1 - α / 2)) ^ 2 / (4 \* m) ^ 2

n

m = 0.03

α = 0.05

n = (qnorm(1 - α / 2)) ^ 2 / (4 \* m) ^ 2

n

m = 0.05

α = 0.05

n = (qnorm(1 - α / 2)) ^ 2 / (4 \* m) ^ 2

n

m=0.01

α=0.01

n = (qnorm(1 - α / 2)) ^ 2 / (4 \* m) ^ 2

n

m=0.03

α=0.01

n = (qnorm(1 - α / 2)) ^ 2 / (4 \* m) ^ 2

n

m=0.05

α=0.01

n = (qnorm(1 - α / 2)) ^ 2 / (4 \* m) ^ 2

n

|  |  |  |  |
| --- | --- | --- | --- |
| 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.

H0 ​: μ ≥ 25

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.

xbar = 24

mu0 = 25

sigma = 3

n = 25

z = (xbar - mu0) / (sigma / sqrt (n))

z

pval = pt(z, 24)

pval

**Output:**

> z

[1] -1.666667

> pval = pt(z, 24)

> pval

[1] 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.

The test statistic -1.66 is less than the critical value -1.710882. Hence at significance level 0.05, we reject the claim that a new economy model will travel at least 25 miles per gallon of gasoline.

alpha = 0.05

qt(0.05 , 24)

**Output:**

> qt(0.05 , 24)

[1] -1.710882

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

The test statistic -1.66 is greater than the critical value -1.710882. Hence at significance level 0.05, we reject the claim that a new economy model will travel at least 25 miles per gallon of gasoline.

We can accept the claim that a new model will travel at least 25 miles per gallon of gasoline if the test statistic is lesser than the critical value.

z = -1.666

n = 25

sigma = 3

criticalvalue = z\*sigma/sqrt(n)+25

criticalvalue

**Output:**

> criticalvalue = z\*sigma/sqrt(n)+25

> criticalvalue

[1] 24.0004

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

A type II erroroccurs if the hypothesis test based on a random sample fails to reject the null hypothesis even when the true population mean *μ*is in fact less than *μ*0.

In this case if the new model fails to reject that the model travels 25 miles per gallon even when the population mean (25) is less than the *μ*0.

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.

The probability of committing a Type II error is 6%.

n = 25

s = 3

SE = s / sqrt(n) #Standard Error

SE

alpha = 0.05

mu0 = 25

q = mu0 + qt(alpha, df = n-1) \* SE

q

#Probability of type II error is

mu = 23

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

powerOfTest = 1-beta

powerOfTest

**Output:**

> SE

[1] 0.6

> alpha = 0.05

> mu0 = 25

> q = mu0 + qt(alpha, df = n-1) \* SE

> q

[1] 23.97347

> mu = 23

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

> powerOfTest = 1-beta

> powerOfTest

[1] 0.9411144

The probability of making a Type II 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.

H0: μ1 = μ2

Ha: μ1 not equal to μ2.

Mean 1 and Mean 2 are the means of driving distances for two populations

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.

golf <- read.csv("/Users/shrutijoshi/Documents/Study Docs/BAN 602/R Assignment/Golf.csv",header=T)

golf

current <- golf$Current

new <- golf$New

mean(current)

mean(new)

sd(current)

sd(new)

NROW(current)

NROW(new)

**Output:**

> mean(current)

[1] 270.275

> mean(new)

[1] 267.5

> sd(current)

[1] 8.752985

> sd(new)

[1] 9.896904

> NROW(current)

[1] 40

> NROW(new)

[1] 40

1. Compute the standard error for your test.

se <- (sqrt((sd(current)^2/NROW(current))+(sd(new)^2/NROW(new))))

se

**Output:**

> se <- (sqrt((sd(current)^2/NROW(current))+(sd(new)^2/NROW(new))))

> se

[1] 2.08904

1. Compute the degree of freedom for your test.

Current sample size n1 = 40

New sample size n2 = 40

Degree of freedom = n1+n2-2 = 78

1. Compute the test statistic for your test.

data <- read.csv("Golf.csv")

data

c <- table(data$Current)

n <- table(data$New)

SE = sqrt((sd(c)^2/NROW(c))+(sd(n)^2/NROW(n)))

testStatistic = (mean(c)-mean(n))/SE

testStatistic

**Output:**

> testStatistic

[1] 1.328362

1. Compute the p value for your test.

z = (mean(c)-mean(n))/SE

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

print(pvalue)

**Output:**

> z = (mean(c)-mean(n))/SE

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

> print(pvalue)

[1] .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 then 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(c, n, var.equal = TRUE)

**Output:**

Two Sample t-test

data: c and n

t = 1.3284, df = 76.852, p-value = 0.188

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-1.384937 6.934937

sample estimates:

mean of x mean of y

270.275 267.500

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

t.test(c, 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(n, 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(c,n)

**Output:**

Welch Two Sample t-test

data: c and n

t = 1.3284, df = 76.582, p-value = 0.188

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-1.384937 6.934937

sample estimates:

mean of x mean of y

270.275 267.500

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.

m1 = c(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 = c(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)

mean(m1)

mean(m2)

var(m1)

var(m2)

length(m1)

length(m2)

sd(m1)

sd(m2)

**Output:**

>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)

[1] 25

> length(m2)

[1] 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

1. Compute the test statistic.

t = var(m1)/var(m2)

t

**Output:**

> t = var(m1)/var(m2)

> t

[1] 8.284448

1. Compute the p value.

df1 = length(m1)-1

df1

df2 = length(m2)-1

df2

pValue = 2\*pf(t, df1, df2, lower.tail = FALSE)

pValue

**Output:**

> df1 = length(m1)-1

> df1

[1] 24

> df2 = length(m2)-1

> df2

[1] 21

>

> pValue = 2\*pf(t, df1, df2, lower.tail = FALSE)

> pValue

[1] 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, df2)

criticalValueLower

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

criticalValueUpper

**Output:**

> criticalValueLower

[1] 0.4327282

>

> criticalValueUpper=qf(1-significanceLevel/2, df1, df2)

> criticalValueUpper

[1] 2.367526

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

The test statistic “8.284448” is greater than the critical value “2.367526”. Hence, we reject the 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)

**Output:**

F test to compare two variances

data: m1 and m2

F = 8.2844, num df = 24, denom df = 21, p-value = 7.222e-06

alternative hypothesis: true ratio of variances is not equal to 1

95 percent confidence interval:

3.499201 19.144689

sample estimates:

ratio of variances

8.284448

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

confidenceInterval = 0.05

df1

df2

cvlower <-qchisq(confidenceInterval/2,df1)

cvupper <-qchisq(1-confidenceInterval/2,df1)

cvlower

cvupper

cilower <-df1\*var(m1)/ cvupper

ciupper <-df1\*var(m1)/ cvlower

cilower

ciupper

**Output:**

> cvlower

[1] 12.40115

> cvupper

[1] 39.36408

> cilower <-df1\*var(m1)/ cvupper

> ciupper <-df1\*var(m1)/ cvlower

> cilower

[1] 0.02980728

> ciupper

[1] 0.09461509

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

confidenceInterval = 0.05

df2

cvlower2 <-qchisq(confidenceInterval/2,df2)

cvupper2 <-qchisq(1-confidenceInterval/2,df2)

cvlower2

cvupper2

cilower2 <-(df2\*var(m2)/cvupper2)

ciupper2 <-(df2\*var(m2)/cvupper2)

cilower2

ciupper2

sdlower <- (sqrt(cilower2))

sdlower

sdupper <- (sqrt(ciupper2))

sdupper

**Output:**

> cvlower2 <-qchisq(confidenceInterval/2,df2)

> cvupper2 <-qchisq(1-confidenceInterval/2,df2)

> cvlower2

[1] 10.2829

> cvupper2

[1] 35.47888

> cilower2 <-(df2\*var(m2)/cvupper2)

> ciupper2 <-(df2\*var(m2)/cvupper2)

> cilower2

[1] 0.003492988

> ciupper2

[1] 0.003492988

> sdlower <- (sqrt(cilower2))

> sdlower

[1] 0.0591015

> sdupper <- (sqrt(ciupper2))

> sdupper

[1] 0.0591015

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

The machine 1 having larger variance will provide the greater opportunity for quality improvements.

**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 is independent of type of flight

Ha: type of ticket is not independent of type of flight

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

|  |  |  |  |
| --- | --- | --- | --- |
| **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.

Total(fij) = 920

Total(eij) = 919.7

Total(fij-eij) = 0.3

Total((fij-eij)^2) = 14076.89

Total(((fij-eij)^2)/eij) = 100.55 = Test Statistic

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)

pValue

significancelevel=0.05

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

criticalValue

**Output:**

> testStatistic = 100.55

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

> pValue

[1] 1.465025e-22

> significancelevel=0.05

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

> criticalValue

[1] 5.991465

The Null hypothesis is rejected ,as the test statistic = 100.55 is greater than the critical value = 5.99.

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

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

Data

chisq.test(Data)

**Output:**

> Data

A B

A 29 22

B 95 121

C 518 135

> chisq.test(Data)

Pearson's Chi-squared test

data: Data

X-squared = 100.43, df = 2, p-value < 2.2e-16

**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.

medicaltable <- read.csv("Medical1.csv", header = T)

medicaltable

install.packages("fBasics")

library(fBasics)

basicStats(medicaltable)

**Output:**

>basicStats(medicaltable)

Florida New.York North.Carolina

nobs 20.000000 20.000000 20.000000

NAs 0.000000 0.000000 0.000000

Minimum 2.000000 4.000000 3.000000

Maximum 9.000000 13.000000 12.000000

1. Quartile 3.750000 7.000000 4.750000

3. Quartile 7.000000 8.250000 8.250000

Mean 5.550000 8.000000 7.050000

Median 6.000000 8.000000 7.500000

Sum 111.000000 160.000000 141.000000

SE Mean 0.478347 0.492042 0.634429

LCL Mean 4.548808 6.970144 5.722125

UCL Mean 6.551192 9.029856 8.377875

Variance 4.576316 4.842105 8.050000

Stdev 2.139233 2.200478 2.837252

Skewness -0.233891 0.534963 -0.048041

Kurtosis -1.285546 -0.116777 -1.176695

1. Clearly state the hypotheses being tested.

Ho: µ1= µ 2= µ 3

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

Ha : µ1≠ µ 2≠ µ 3

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

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

SSE = Sum ((n-1) var\*2)

= (20-1)(2.139233)2 + (20-1)(2.200478)2 + (20-1)(2.837252)2

= 331.08

Degree of Freedom, = Total(n) – k

= 6-3 = 57

MSE = SSE / Degree of Freedom

= 331.08/57 = 5.80

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

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

SSTR = 20(5.55-6.86)2 + 20(8-6.86)2 + 20(7.05-6.862

= 61.03

Degree of Freedom = k – 1 = 3-1 = 2

MSTR = SSTR / DOF = 61.03 / 2 = 30.515

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 = 61.03 + 331.9 = 392.93

SD for three locations = 2.5866

SST = (60-1)\*(2.58)\*\*2 = 392.73

Hence, SST is equal to SSE+SSTR.

1. Compute the test statistic for your test.

F = MSTR / MSE = 30.51 / 5.80 = 5.26

1. Compute the p value for your test.

F = 5.26

df1 = 2

df2 = 57

pvalue <- 1-pf(F,df1=2, df2=57)

pvalue

**Output:**

> F = 5.26

> df1 = 2

> df2 = 57

> pvalue <- 1-pf(F,df1=2, df2=57)

> pvalue

[1] 0.008009567

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

The Null hypothesis is rejected as all the three locations do not have the same test scores, i.e., alpha = 0.05 and p-value = 0.008 and hence we notice its not equal.

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>

matrix = as.matrix(medicaltable)

matrix

r = c(t(matrix))

r

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

k = 3

n = 20

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

t

a = aov(r ~ t)

summary(a)

**Output:**

matrix = as.matrix(medicaltable)

> matrix

Florida New.York North.Carolina

[1,] 3 8 10

[2,] 7 11 7

[3,] 7 9 3

[4,] 3 7 5

[5,] 8 8 11

[6,] 8 7 8

[7,] 8 8 4

[8,] 5 4 3

[9,] 5 13 7

[10,] 2 10 8

[11,] 6 6 8

[12,] 2 8 7

[13,] 6 12 3

[14,] 6 8 9

[15,] 9 6 8

[16,] 7 8 12

[17,] 5 5 6

[18,] 4 7 3

[19,] 7 7 8

[20,] 3 8 11

> r = c(t(matrix))

> 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

[32] 6 8 2 8 7 6 12 3 6 8 9 9 6 8 7 8 12 5 5 6 4 7 3 7 7 8 3 8 11

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

> k = 3

> n = 20

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

> t

[1] Florida New York North Carolina Florida New York North Carolina

[7] Florida New York North Carolina Florida New York North Carolina

[13] Florida New York North Carolina Florida New York North Carolina

[19] Florida New York North Carolina Florida New York North Carolina

[25] Florida New York North Carolina Florida New York North Carolina

[31] Florida New York North Carolina Florida New York North Carolina

[37] Florida New York North Carolina Florida New York North Carolina

[43] Florida New York North Carolina Florida New York North Carolina

[49] Florida New York North Carolina Florida New York North Carolina

[55] Florida New York North Carolina Florida New York North Carolina

Levels: Florida New York North Carolina

> a = aov(r ~ t)

> summary(a)

Df Sum Sq Mean Sq F value Pr(>F)

t 2 61.0 30.517 5.241 0.00814 \*\*

Residuals 57 331.9 5.823

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

**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 = T)

safetytable

basicStats(safetytable)

Output:

basicStats(safetytable)

Percent.Under.21 Fatal.Accidents.per.1000

nobs 42.000000 42.000000

NAs 0.000000 0.000000

Minimum 8.000000 0.039000

Maximum 18.000000 4.100000

1. Quartile 9.250000 1.017500

3. Quartile 14.750000 2.810750

Mean 12.261905 1.922405

Median 12.000000 1.881000

Sum 515.000000 80.741000

SE Mean 0.483238 0.165257

LCL Mean 11.285987 1.588661

UCL Mean 13.237823 2.256149

Variance 9.807782 1.147019

Stdev 3.131738 1.070990

Skewness 0.195570 0.179586

Kurtosis -1.232961 -1.096173

>

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>

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

summary(lm.safetytable)

anova(lm.safetytable)

**Output:**

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

> summary(lm.safetytable)

Call:

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

Residuals:

Min 1Q Median 3Q Max

-1.23412 -0.26441 0.00772 0.44362 1.49099

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -1.59741 0.37167 -4.298 0.000107 \*\*\*

Percent.Under.21 0.28705 0.02939 9.767 3.79e-12 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5894 on 40 degrees of freedom

Multiple R-squared: 0.7046, Adjusted R-squared: 0.6972

F-statistic: 95.4 on 1 and 40 DF, p-value: 3.794e-12

> anova(lm.safetytable)

Analysis of Variance Table

Response: Fatal.Accidents.per.1000

Df Sum Sq Mean Sq F value Pr(>F)

Percent.Under.21 1 33.134 33.134 95.397 3.794e-12 \*\*\*

Residuals 40 13.893 0.347

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

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

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 – Confidence Interval for coefficient of the variable

For every additional percentage of drivers under 21, there is a 95% chance that there is an 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 level of significance is 0.05 and the p-value is 3.794e-12, we notice that p-value is less than level of significance and so there is a significant relationship between the two. Hence, we can reject the Null Hypothesis.

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,

r^2 = 33.15 / 47.03 = 0.7043

We can interpret that 70% of variation in fatal accidents can be determined by the linear relationship between drivers below 21 and 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 , which is derived from the ANOVA table

Residual Standar Error = sqrt(0.3473) = 0.5894 , which is calculated

Thus, the values match.

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 = b1 = -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**)

= 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**)

sd = 0.114

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

tvalue

moe = tvalue\*sd

moe

**Output:**

> sd = 0.114

> tvalue = qt((1+0.95)/2, 40)

> tvalue

[1] 2.021075

> moe = tvalue\*sd

> moe

[1] 0.2304026

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.

Margin of error , moe = 0. 2304026

Confidence interval = 1.274 + or – (-0. 2304026)

= 1.044 to 1.504

We can interpret that average number of fatal accidents per 1000 licenses for all cities with 10% of drivers under the age of 21 is between 1.044 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**)

Standard deviation of pred = 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**)

Margin of Error, 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.

Prediction Interval = 1.274 + or – (-1.216)

= 0.059 to 2.489