Tutorial 5

TBA2102 2020/2021 Semester 2

STRUCTURE OF TUTORIALS

Duration:

45 mins

Content:

- Tutorial 4 assignment
- Recap on Key Concepts for Hypothesis Testing
- Tutorial 5 Questions

Tutorial 4 Assignment

PURCHASE ORDER

The `Purchase Orders.xlsx` data set contains data on all items that an aircraft component manufacturing company has purchased over the past 4 months. Each of the column is defined as follows:

- `Supplier`	Supplier of items purchased
- `Order No.`	Order Number of the items purchased
- `Item No.`	A categorical variable used to identify the item
- `Item Description`	Description of the item purchased
- `Item Cost`	Item unit cost
- `Quantity`	Number of items bought in the purchase order
- `Cost per order`	Total cost of the order
- `A/P Terms (Months)`	Suppliers' Accounts Payable (A/P) terms
- `Order Date`	Items order date
- `Arrival Date`	Items arrival date

LET'S TAKE A LOOK AT THE DATA

- glimpse is from dplyr package
- An alternative to str()

```
glimpse(PO)
```

```
Rows: 94
Columns: 11
$ Supplier            <chr> "Hulkey Fasteners", "Alum Sheeting", "Fast-Tie Aerospace"...
$ `Order No.`         <chr> "Aug11001", "Aug11002", "Aug11003", "Aug11004", "Aug11005...
$ `Item No.`          <dbl> 1122, 1243, 5462, 5462, 5319, 5462, 4312, 7258, 6321, 546...
$ `Item Description`  <chr> "Airframe fasteners", "Airframe fasteners", "Shielded Cab...
$ `Item Cost`         <dbl> 4.25, 4.25, 1.05, 1.05, 1.10, 1.05, 3.75, 90.00, 2.45, 1....
$ Quantity            <dbl> 19500, 10000, 23000, 21500, 17500, 22500, 4250, 100, 1300...
$ `Cost per order`    <dbl> 82875.00, 42500.00, 24150.00, 22575.00, 19250.00, 23625.0...
$ `A/P Terms (Months)` <dbl> 30, 30, 30, 30, 30, 30, 30, 45, 30, 30, 30, 30, 30, 3...
$ `Order Date`        <dttm> 2011-08-05, 2011-08-08, 2011-08-10, 2011-08-15, 2011-08-...
$ `Arrival Date`      <dttm> 2011-08-13, 2011-08-14, 2011-08-15, 2011-08-22, 2011-08-...
$ `Arrival Time`      <dbl> 8, 6, 5, 7, 11, 6, 7, 3, 10, 8, 11, 4, 6, 8, 9, 9, 5, 9, ...
```

Other functions that can help you explore the data

- View(PO)
- str(PO)
- head(PO)
- lapply(PO,class) --- check the data type of all the variables

QUESTION 2A

Last tutorial, you were told the manager would like to understand more about the items purchased in the last 4 months. More specifically, he is interested in the following purchase order information: Supplier, Item Description, Cost per order, Arrival Time and Arrival Time is the difference between Arrival Date and Order Date.

i. Now that we have learnt how to identify outliers, let's explore to see if there are any outliers for Arrival Time. Plot the boxplot with range at 1.5 and 3 to help you with the identification. For each boxplot, how many outliers do you detect? Extract the outlier record(s) (i.e. row in the dataframe) for range=1.5 and range=3 to dataframes outlier1.5 and outlier3 respectively. [3 marks]

ii. Generate the descriptive statistics for Arrival Time and Cost per order in a table. Note that despite the outliers detected in (i), the manager would like to keep all the data. In addition, he expressed that he is only interested in the following descriptive statistics so only have these displayed in the table: n (or number of observations), mean, sd, median, min, max, skew, kurtosis. [2 marks]

iii. What can you conclude about the shape of the distribution for the two variables from the coefficient of skewness and coefficient of kurtosis? [1 mark]

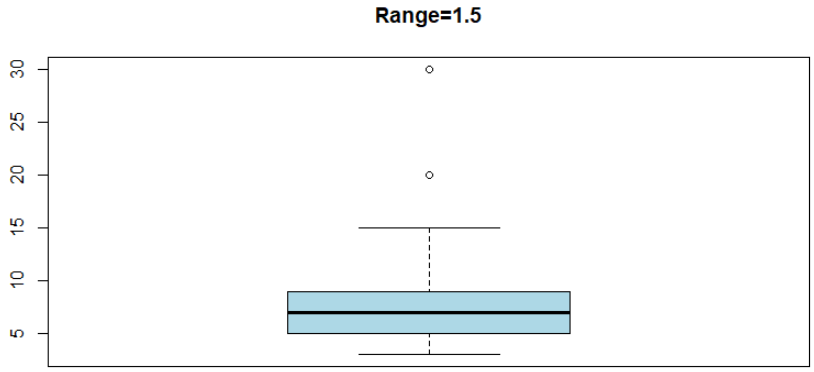
iv. The manager would like to assess if Cost per order and Arrival Time are linearly related. He also would like to assess the same for Quantityand Arrival Time. Compute and display the statistical measure for the manager. From the results, what can you conclude? [2 marks]

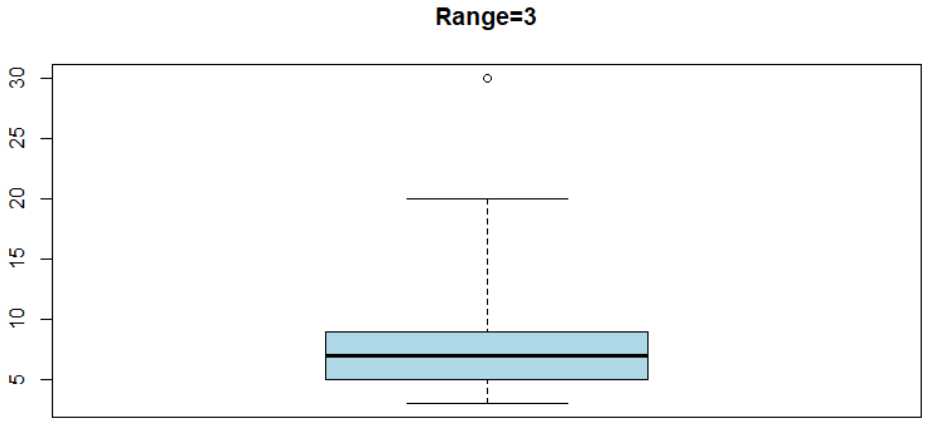
QUESTION 2A(I)

Now that we have learnt how to identify outliers, let's explore to see if there are any outliers for Arrival Time.

- Plot the boxplot with range at 1.5 and 3 to help you with the identification.

```
PO$`Arrival Time`<-as.numeric(PO$`Arrival Date`-PO$`Order Date`)
boxplot(PO$`Arrival Time`,
        range = 1.5,
        col = "light blue",
        main = "Range=1.5")

boxplot(PO$`Arrival Time`,
        range = 3,
        col = "light blue",
        main = "Range=3")
```





QUESTION 2A(I)

Now that we have learnt how to identify outliers, let's explore to see if there are any outliers for Arrival Time.

- Extract the outlier record(s) (i.e. row in the dataframe) for range=1.5 and range=3 to dataframes outlier 1.5 and outlier3 respectively. [3 marks]

```
outlier3<- PO %>% filter(`Arrival Time`>=21)   # Q3+3IQR
outlier1.5<-PO %>% filter(`Arrival Time`>=15)  # Q3+1.5IQR
```

```
boxplota <- boxplot(PO$`Arrival Time`,
        range = 1.5,
        col = "light blue",
        main = "Range=1.5")

boxplota$out
```

20 20 30 20 30

```
boxplotb <-boxplot(PO$`Arrival Time`,
        range = 3,
        col = "light blue",
        main = "Range=3")

boxplotb$out
```

30 30

QUESTION 2A(II)

Generate the descriptive statistics for Arrival Time and Cost per order in a table. Note that despite the outliers detected in (i), the manager would like to keep all the data. In addition, he expressed that he is only interested in the following descriptive statistics so only have these displayed in the table: n (or number of observations), mean, sd, median, min, max, skew, kurtosis. [2 marks]

```
dfat <- describe(PO$`Arrival Time`)
dfcost <- describe(PO$`Cost per order`)
dfatcost<-rbind(dfat,dfcost)
# remove columns not needed
dfatcost$range <- dfatcost$trimmed <- dfatcost$mad <- dfatcost$se <- NULL
dfatcost$vars[1]<-"Arrival Time"
dfatcost$vars[2]<-"Cost per order"
kable(dfatcost, row.names = FALSE, digits = 2,
      caption = "Descriptive Statistics for `Arrival Time` and `Cost per order`") %>%
  kable_classic(full_width = F, html_font = "Arial")
```

- If you use Rstudio, you need the kableExtra package and kable_classic function
- kable () works just fine in Rmarkdown

Descriptive Statistics for `Arrival Time` and `Cost per order`

vars	n	mean	sd	median	min	max	skew	kurtosis
Arrival Time	94	8.41	4.58	7.00	3.00	30	2.68	9.01
Cost per order	94	26295.32	29842.83	15656.25	68.75	127500	1.61	1.80

QUESTION 2A(III)

What can you conclude about the shape of the distribution for the two variables from the coefficient of skewness and coefficient of kurtosis? [1 mark]

Descriptive Statistics for `Arrival Time` and `Cost per order`

vars	n	mean	sd	median	min	max	skew	kurtosis
Arrival Time	94	8.41	4.58	7.00	3.00	30	2.68	9.01
Cost per order	94	26295.32	29842.83	15656.25	68.75	127500	1.61	1.80

- Both distributions are right/positively skewed. This can be inferred from the positive coefficients of skew.
- Both distributions are peaked (as opposed to flat), especially Arrival Time. This can be inferred from the positive coefficients of kurtosis.

QUESTION 2A(III)

The manager would like to assess if Cost per order and Arrival Time are linearly related. He also would like to assess the same for Quantity and Arrival Time. Compute and display the statistical measure for the manager. From the results, what can you conclude? [2 marks]

```
round(cor(PO$`Arrival Time`, PO$`Cost per order`),4)
```

-0.0752

```
round(cor(PO$`Arrival Time`, PO$Quantity),4)
```

-0.0298

Q2.(B) SUPPLIER ANALYSES DASHBOARD

Recall the manager would like to have a deeper analyses of Supplier. He would like to compare the mean and standard deviation of Arrival Time for each of the Supplier.

i. Generate the table and chart that will allow the manager to be able to compare the means and standard deviations of Arrival Time for each Supplier. (Note: There is limited screen space so the manager would like just one table and one chart for this.) [4 marks]

ii. From the chart, describe your conclusion on the suppliers with the highest and lowest mean and variance in Arrival Time. [1 mark]

QUESTION 2B(I)

i. Generate the table and chart that will allow the manager to be able to compare the means and standard deviations of Arrival Time for each Supplier. (Note: There is limited screen space so the manager would like just one table and one chart for this.) [4 marks]

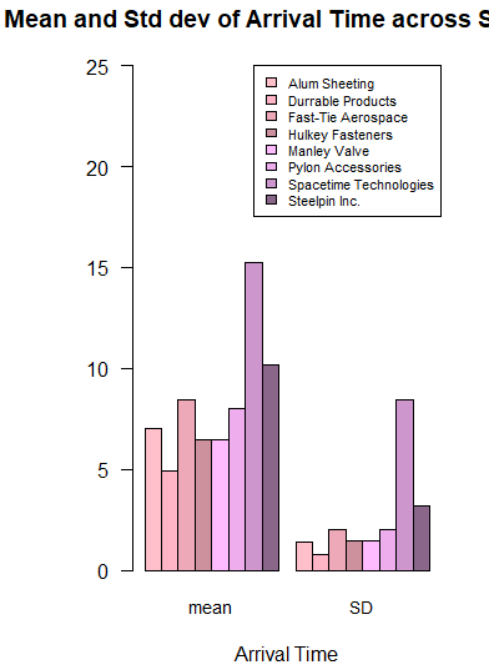
```
tab.2b <- PO %>%
  group_by(`Supplier`) %>%
  summarise(mean=mean(`Arrival Time`), SD=sd(`Arrival Time`))
kable(tab.2b,digits = 2)
```

Supplier	mean	SD
Alum Sheeting	7.00	1.41
Durrable Products	4.92	0.76
Fast-Tie Aerospace	8.47	2.03
Hulkey Fasteners	6.47	1.46
Manley Valve	6.45	1.44
Pylon Accessories	8.00	2.00
Spacetime Technologies	15.25	8.44
Steelpin Inc.	10.20	3.17

QUESTION 2B(I)

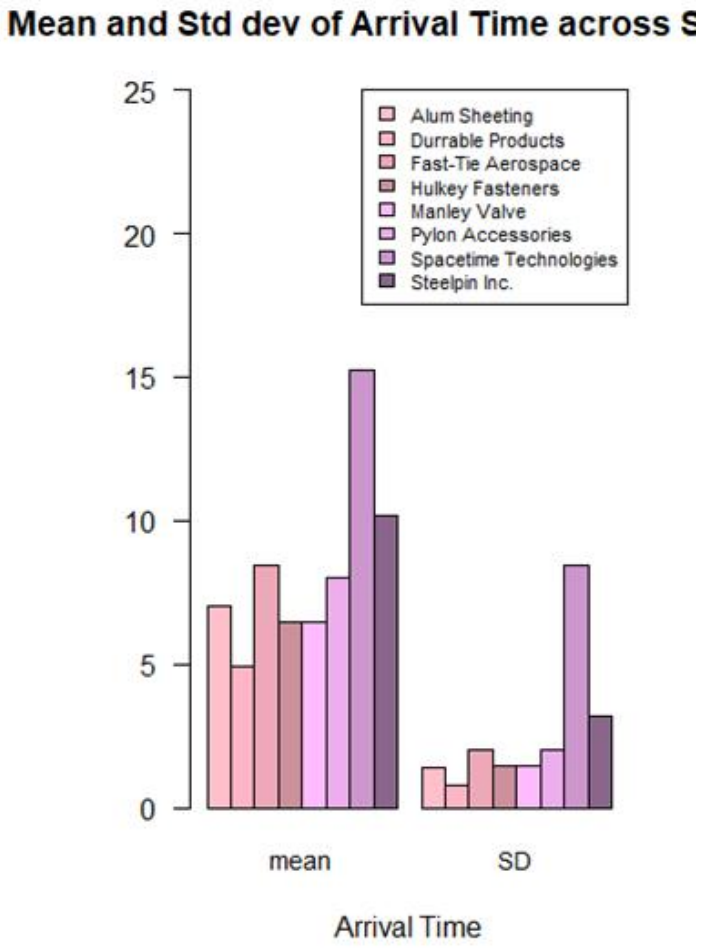
Generate the table and chart that will allow the manager to be able to compare the means and standard deviations of Arrival Time for each Supplier. (Note: There is limited screen space so the manager would like just one table and one chart for this.) [4 marks]

```
par(mar=c(5,10,4,2))
bar.2b<-as.matrix(tab.2b[,c(2:3)])
col.2b<-c("pink","pink1","pink2","pink3","plum1","plum2","plum3","plum4")
barplot(bar.2b,
        beside= TRUE,
        col =col.2b,
        main=" Mean and Std dev of Arrival Time across Supplier",
        cex.names=0.9,
        ylim=c(0,25),
        las=1,
        xlab="Arrival Time")
legend("topright", cex=0.7, fill=col.2b, tab.2b$Supplier)
```



QUESTION 2B(II)

From the chart, describe your conclusion on the suppliers with the highest and lowest mean and variance in Arrival Time. [1 mark]



- Spacetime technologies has the highest mean arrival time and highest standard deviation. This means that on average, it takes the longest to deliver. However, there is relatively high variation in arrival times meaning that we are least certain about their expected arrival times.
- In contrast, Durrable products has the shortest mean arrival time and smallest standard deviation. This means that we can be certain that it will deliver in a time span relative to the other suppliers.

QUESTION 2C

The manager wants to gain more insights into the trends of purchasing costs across the 4 months of data.

i. There are a total of 13 different item types (defined by Item Description) that the company has purchased. The manager would like to have a dashboard to visualize the trend of monthly purchasing cost spent on each of the 13 items over the 4-month period. Using the Order Date to compute the month of purchase, generate an appropriate table and chart for the manager to view the trends. [4 marks]

ii. From the chart, describe two trends you observe. [1 mark]

QUESTION 2C(I)

There are a total of 13 different item types (defined by Item Description) that the company has purchased. The manager would like to have a dashboard to visualize the trend of monthly purchasing cost spent on each of the 13 items over the 4-month period. Using the Order Date to compute the month of purchase, generate an appropriate table and chart for the manager to view the trends. [4 marks]

```
# create order month variable
PO$Month<-format(as.Date(PO$`Order Date`), "%m")
```

QUESTION 2C(I)

There are a total of 13 different item types (defined by Item Description) that the company has purchased. The manager would like to have a dashboard to visualize the trend of monthly purchasing cost spent on each of the 13 items over the 4-month period. Using the Order Date to compute the month of purchase, generate an appropriate table and chart for the manager to view the trends. [4 marks]

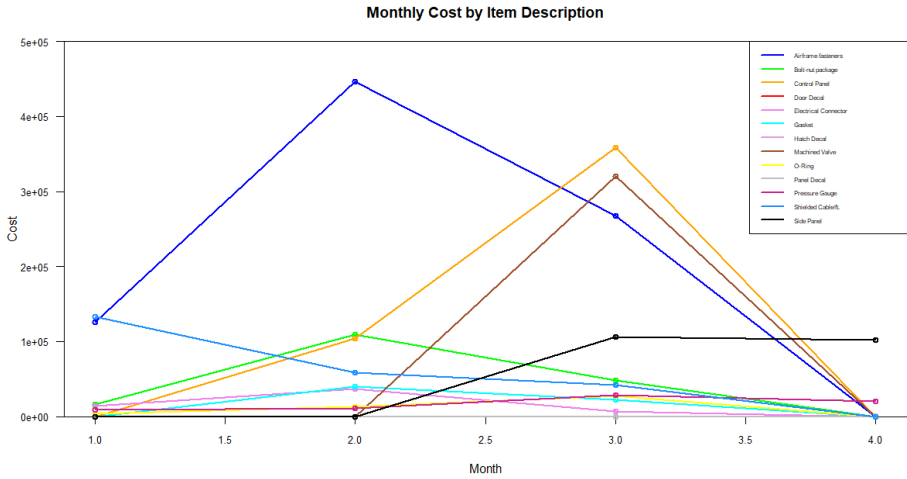
```
PO.2c1 <- PO %>%
  group_by(`Month`) %>%
  summarise(total.cost=sum(`Cost per order`))

PO.2c2 <- PO %>%
  group_by(`Month`,`Item Description`) %>%
  summarise(total.item.cost=sum(`Cost per order`))

PO.2c2.spread<- PO.2c2 %>%
  spread(key="Item.Description",value=total.item.cost)

PO.2c2.spread[is.na(PO.2c2.spread)]<-0 #convert NA to 0 value
PO.2c2.spread$Total.Cost <- as.numeric(apply(PO.2c2.spread[,2:14], 1, sum))
kable(PO.2c2.spread, caption = "Contingency Table for Supplier & Order Month")%>%
  kable_classic(full_width = F, html_font = "Arial")
```

Contingency Table for Supplier & Order Month

Month	Airframe fasteners	Bolt-nut package	Control Panel	Door Decal	Electrical Connector	Gasket	Hatch Decal	Machined Valve	O-Ring	Panel Decal	Pressure Gauge	Shielded Cable/ft.	Side Panel	Total.Cost
08	125375	15937.5	0	0.00	14425.00	0.00	375.0	0.0	3185.0	0	9000.0	133135	0	301432.5
09	446425	109536.8	103530	0.00	36558.75	40027.50	0.0	0.0	12770.0	0	10800.0	58525	0	818173.0
10	267750	47755.0	358500	0.00	7062.50	22465.75	0.0	320587.5	27722.5	0	28642.5	42250	106125	1228860.8
11	0	0.0	0	151.25	0.00	0.00	467.5	0.0	0.0	525	20425.0	0	101725	123293.8

QUESTION 2C(I)

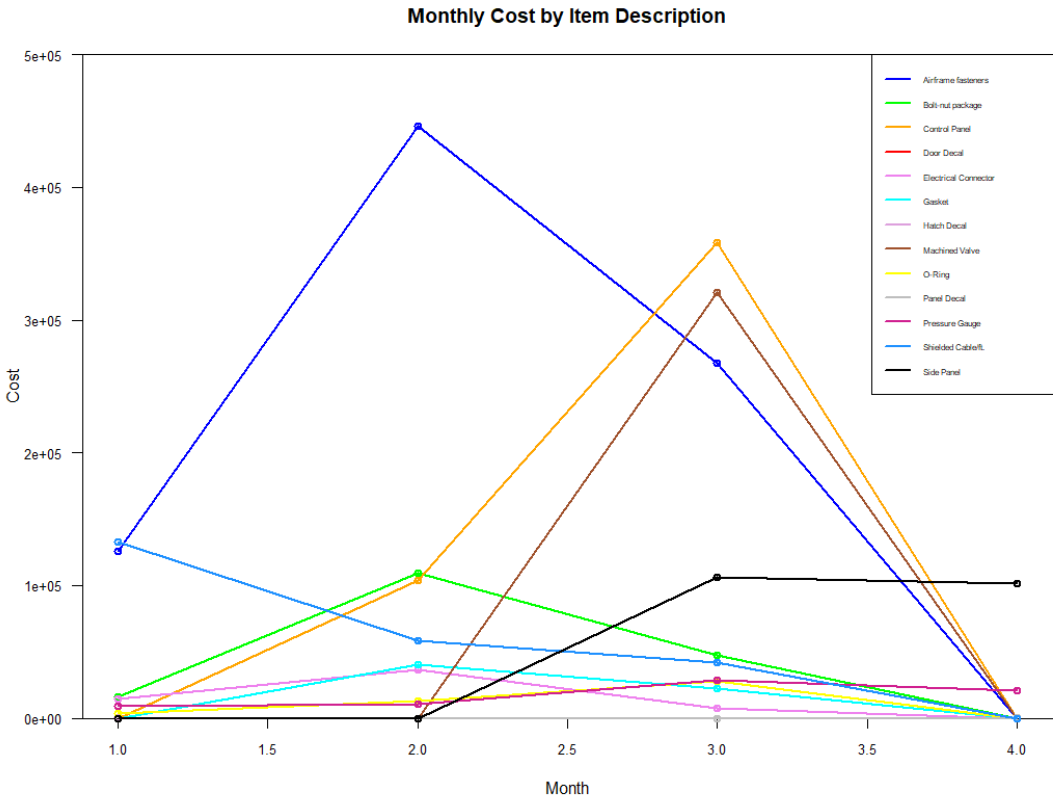
There are a total of 13 different item types (defined by Item Description) that the company has purchased. The manager would like to have a dashboard to visualize the trend of monthly purchasing cost spent on each of the 13 items over the 4-month period. Using the Order Date to compute the month of purchase, generate an appropriate table and chart for the manager to view the trends. [4 marks]




```
# line chart
par(mar=c(5,10,4,0))
plot(PO.2c2.spread$`Airframe fasteners`, type="o", col="blue", xlab="Month",
     ylab="Cost",
     ylim= c(0, 500000),
     las=1,
     cex.axis= 0.8,
     lwd=2,
     main="Monthly Cost by Item Description")
lines(PO.2c2.spread$`Bolt-nut package`, type="o", col="green", lwd=2)
lines(PO.2c2.spread$`Control Panel`, type="o", col="orange", lwd=2)
lines(PO.2c2.spread$`Door Decal`, type="o", col="red", lwd=2)
lines(PO.2c2.spread$`Electrical Connector`, type="o", col="violet", lwd=2)
lines(PO.2c2.spread$Gasket, type="o", col="cyan", lwd=2)
lines(PO.2c2.spread$`Hatch Decal`, type="o", col="plum", lwd=2)
lines(PO.2c2.spread$`Machined Valve`, type="o", col="sienna", lwd=2)
lines(PO.2c2.spread$`O-Ring`, type="o", col="yellow", lwd=2)
lines(PO.2c2.spread$`Panel Decal`, type="o", col="gray", lwd=2)
lines(PO.2c2.spread$`Pressure Gauge`, type="o", col="violetred", lwd=2)
lines(PO.2c2.spread$`Shielded Cable/ft.`, type="o", col="dodgerblue", lwd=2)
lines(PO.2c2.spread$`Side Panel`, type="o", col="black", lwd=2)
# add a legend
item<-sort(unique(PO.2c2$`Item Description`))
legend("topright", legend=item,
       col=c("blue","green","orange","red", "violet", "cyan", "plum", "sienna", "yellow", "gray", "violetred", "dodgerblue", "black"),
       lty=1, cex=0.5, lwd=2 )
```

QUESTION 2C(I)

From the chart, describe two trends you observe. [1 mark]



Airframe fasteners costs the most in the month of September but the cost incurred from this supplier decreased in the following months (to 0 in the month of December).

Machined Valve and O-ring were generally low in cost. However, in the month of October, there was a short-term spike in cost incurred from these 2 suppliers.

Another possibility: most of the items (other than the 3 described above) incurred low cost throughout the months

QUESTION 2D: COMPUTING PROBABILITIES

The manager would like to use the existing data to compute the probability of the following events:

i. Arrival Time for an order with Alum Sheeting being less than 7 days [1 mark]
ii. Cost of an order of O-Ring being more than $12000. [1 mark]

Please compute the probabilities and type your answer below.

```
df.alum <- PO %>%
  filter(Supplier=="Alum Sheeting")
df.alumat <- df.alum %>%
  filter(`Arrival Time`<7)
nrow(df.alumat)/nrow(df.alum)
```

0.375

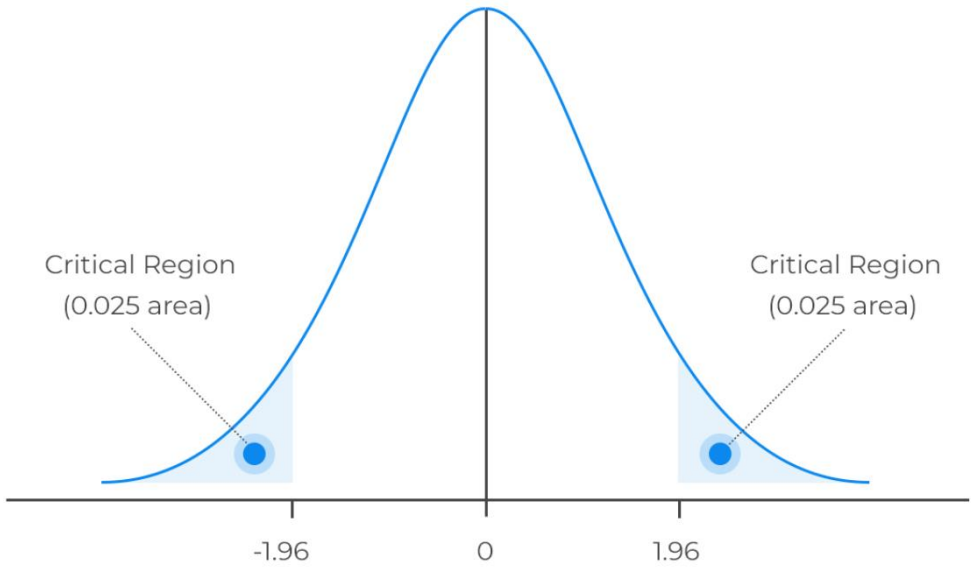
```
df.or <- PO %>%
  filter(`Item Description`=="O-Ring")
df.orc <- df.or%>%
  filter(`Cost per order`>12000)
nrow(df.orc)/nrow(df.or)
```

0

Tutorial 5

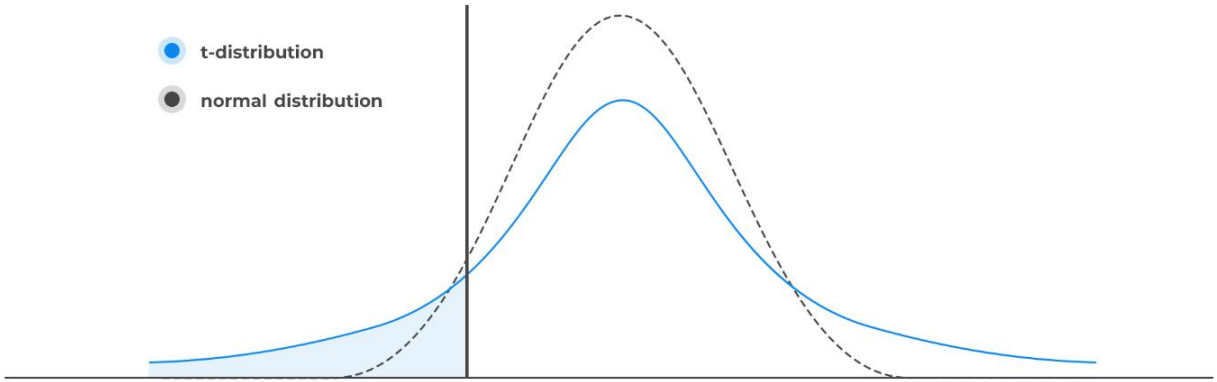
A Recap of Important Concepts

WHAT IS A Z-STATISTIC



In R:
qnorm(.025)
qnorm(.975)

HOW IS Z-STATISTIC DIFFERENT FROM T-STATISTIC?



In R:
qt(.025, df = n-1)
qt(.975, df = n-1)

USE Z-STATISTIC OR T-STATISTIC?

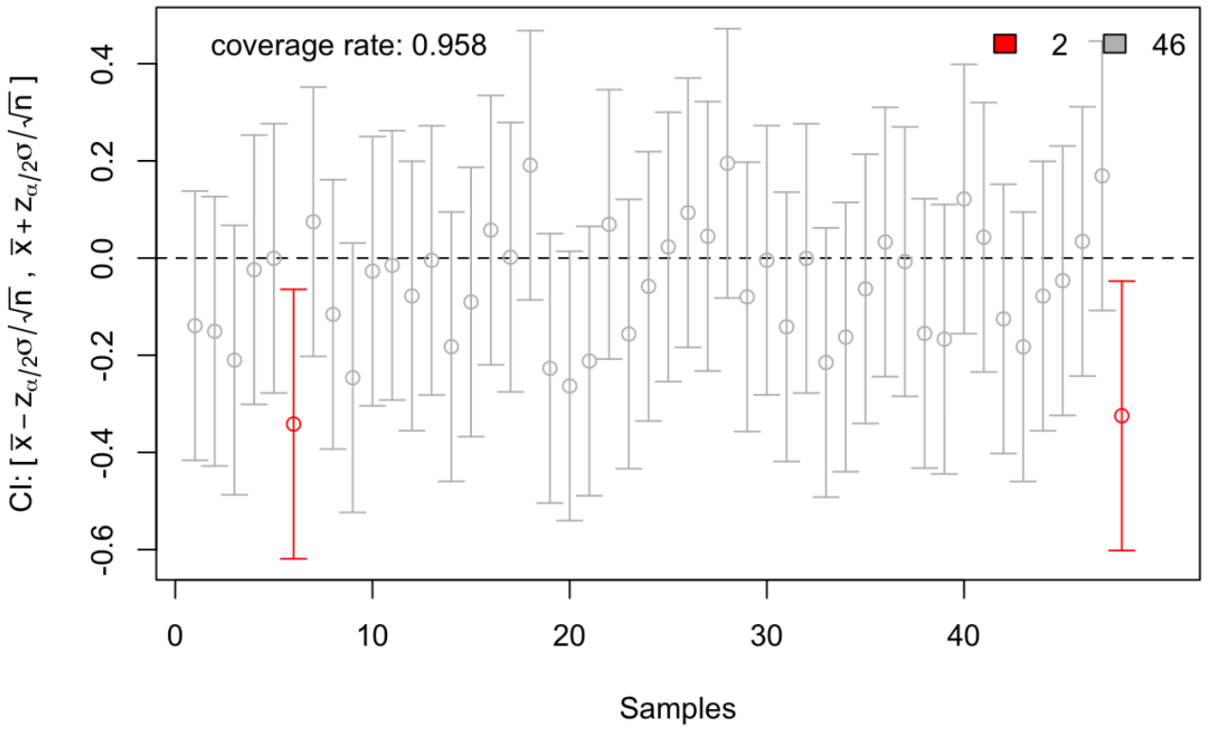
- Can we confidently apply the Central Limit Theorem?
- Remember: CLT only applies to sampling distribution of the mean, not to the distribution of the random variable itself
- When population variance is known, we use the z-statistic
- When population variance is unknown, we use the t-statistic

CONFIDENCE INTERVALS

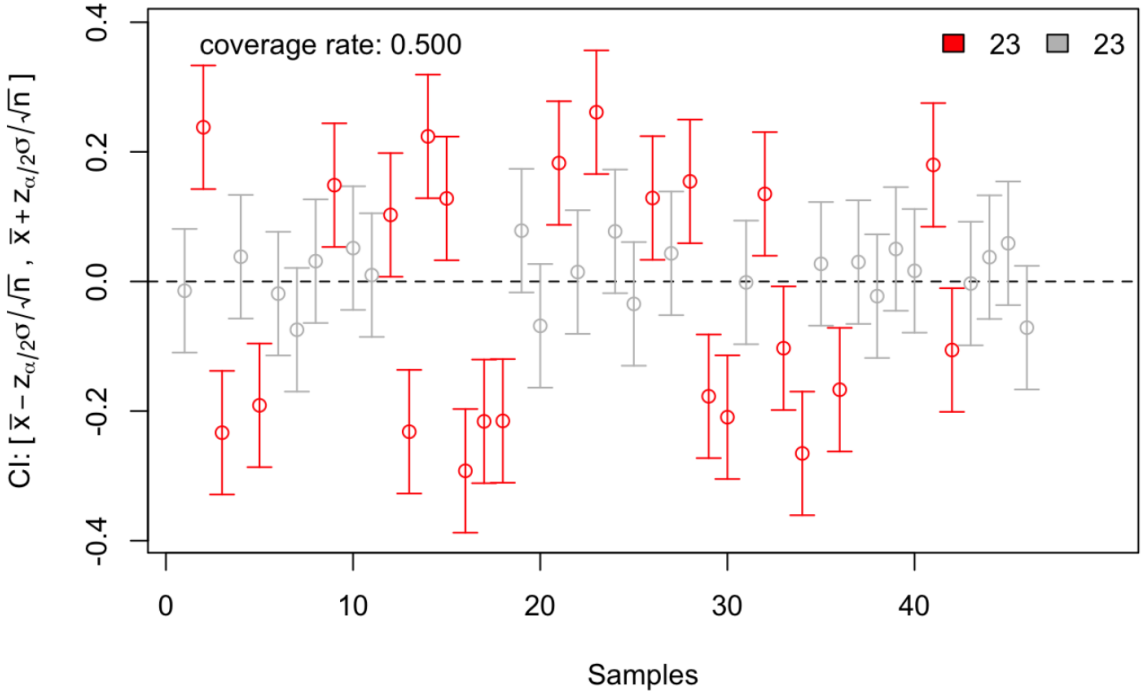
Formulae: $\bar{X} \pm Z_{\alpha/2} (\sigma/\sqrt{n})$; (σ is known)

Things you should know:
- What is a confidence interval and how is it typically calculated?
- Conceptually, what does a 95% confidence interval mean?

95% CONFIDENCE INTERVALS



50% CONFIDENCE INTERVALS



HYPOTHESIS TESTING

Things you should know:

- What is hypothesis testing?
- What can we infer from these tests?
- What are the common statistical tests that can be used for null-hypothesis testing?
- What assumptions do these tests make and how do you verify them?

HYPOTHESIS TESTING

H_0: Null Hypothesis (e.g. $\mu = 0$; $x >= 0$)
H_1: Alternative Hypothesis (e.g. $\mu \neq 0$, $x < 0$)

- We start by assuming the null hypothesis is true and check what is the probability of observing the data given that the null hypothesis is true.

- If the probability is low, we reject H_0, else we retain H_0.

COMMON STATISTICAL TESTS

T-test : test if there is a difference in **mean** between 2 groups

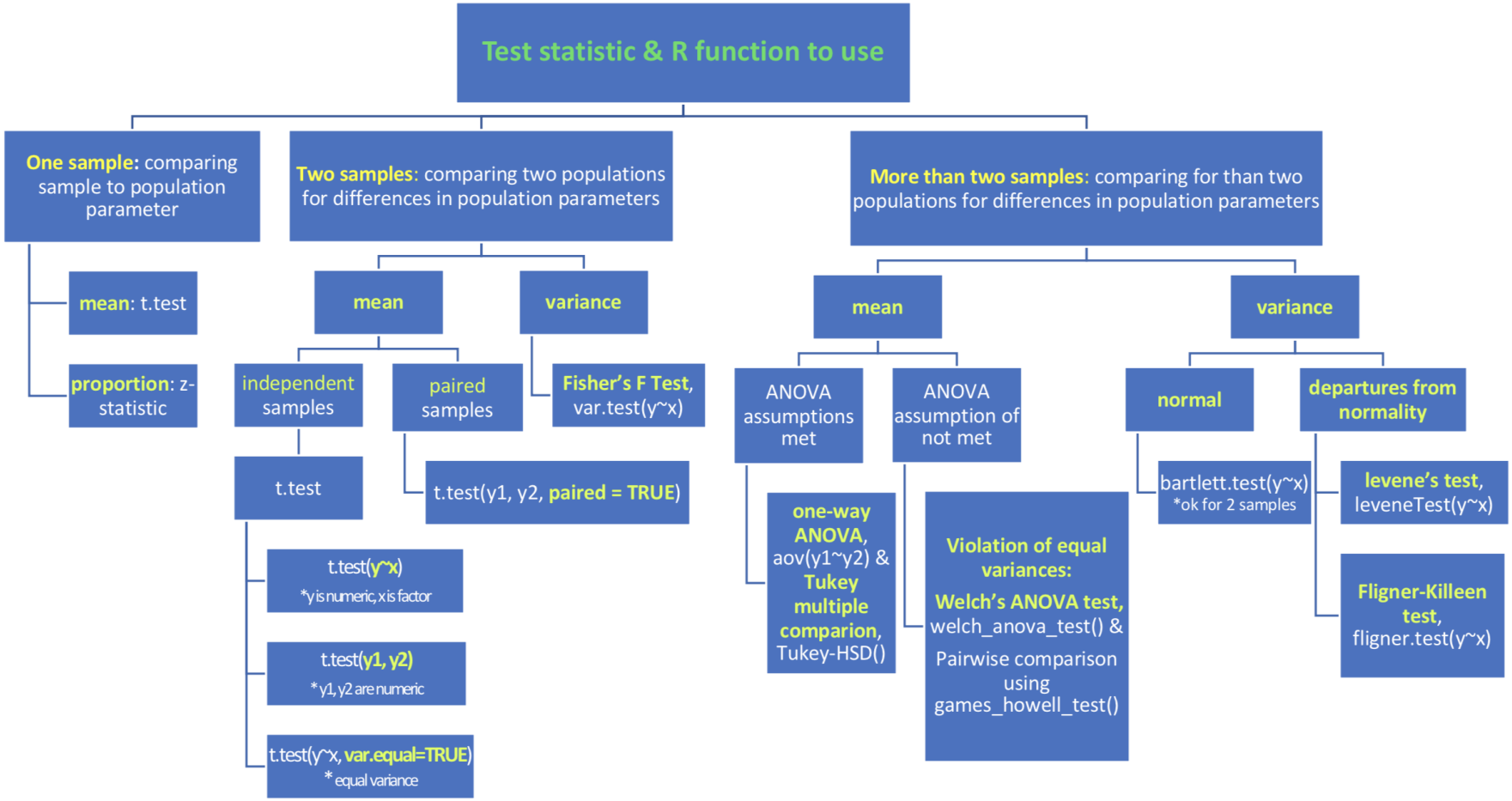
- t.test(outcome ~ group); R defaults to Welch T-test
- alt: t.test(group1.scores, group2.scores)

Analysis of Variance: test if there is a difference in **mean** between more than 2 groups

- aov(outcome ~ group)

What assumptions do these tests make?

OTHER TESTS


Test statistic & R function to use
One sample: comparing sample to population parameter
mean: t.test
proportion: z-statistic
Two samples: comparing two populations for differences in population parameters
mean
variance
independent samples
paired samples
Fisher's F Test, var.test(y~x)
t.test
t.test(y1, y2, paired = TRUE)
t.test(y~x)
*y is numeric, x is factor
t.test(y1, y2)
*y1, y2 are numeric
t.test(y~x, var.equal=TRUE)
*equal variance
More than two samples: comparing for than two populations for differences in population parameters
mean
variance
ANOVA assumptions met
ANOVA assumption of not met
normal
departures from normality
one-way ANOVA, aov(y1~y2) & Tukey multiple comparion, Tukey-HSD()
Violation of equal variances:
Welch's ANOVA test, welch_anova_test() &
Pairwise comparison using games_howell_test()
bartlett.test(y~x)
*ok for 2 samples
levene's test, leveneTest(y~x)
Fligner-Killeen test, fligner.test(y~x)

OTHER TESTS (2 GROUPS)

Pairwise t-test (when groups are not independent)

- t.test(outcome ~ group, paired = TRUE)

Fisher's F-test (comparing variances)

- var.test(outcome ~ group)

OTHER TESTS (MORE THAN 2 GROUPS)

Welch's ANOVA test(when ANOVA assumptions not met)

- welch_anova_test(outcome ~ group)

Levene's test (comparing variances)

- LeveneTest(outcome ~ group)

Tutorial questions

Tutorial 5: Question 1

DATASET REQUIRED

- Dataset required: `Sales Transactions.xlsx`

`Sales Transactions.xlsx` contains the records of all sale transactions for a day, July 14. Each of the column is defined as follows:

- `CustID` : Unique identifier for a customer
- `Region` : Region of customer's home address
- `Payment` : Mode of payment used for the sales transaction
- `Transction Code` : Numerical code for the sales transaction
- `Source` : Source of the sales (whether it is through the Web or email)
- `Amount` : Sales amount
- `Product` : Product bought by customer
- `Time Of Day` : time in which the sale transaction took place.

In the last tutorial, you have been tasked to conduct some descriptive analytics on the dataset, to identify and understand any interesting patterns from the sales transaction data, and to develop dashboards to make visualization of these patterns better. This week, we will conduct sampling estimation and hypotheses testing with the data.

Note for this tutorial to round off your answers to the following: If the answer is greater than 1, round off to 2 decimal places. If the answer is less than 1, round off to 3 significant numbers. When rounding, also take note of the natural rounding points, for example, costs in dollars would round off to 2 decimal places.

QUESTION 1A: COMPUTING INTERVAL ESTIMATES

i. compute the 95% confidence interval for the mean of Amount.
ii. compute the 99% confidence interval for proportion of book sales transactions with sales amount being greater than $50.
iii. compute the 90% predictive interval for Amount for orders of DVD.

What do each of the interval estimates above tell us? Type your answer below. Indicate clearly what parameter or value are you estimating and which sample or population the estimates are for.

1AI – CI FOR MEAN

i. compute the 95% confidence interval for the mean of Amount.

Formulae: $\bar{X} \pm t_{\alpha/2,\, n-1} (s/\sqrt{n})$

```
# Enter your codes here
#compute manually 95% CI for mean `Amount`
uCIamt95<- mean(ST$Amount) - qt(0.025,df=nrow(ST)-1)*sd(ST$Amount)/sqrt(nrow(ST))
lCIamt95 <- mean(ST$Amount) + qt(0.025,df=nrow(ST)-1)*sd(ST$Amount)/sqrt(nrow(ST))
print(cbind(lCIamt95, uCIamt95), digits=4)
```

```
##      lCIamt95 uCIamt95
## [1,]    34.76    45.13
```

1AII – CI FOR PROPORTION

ii. compute the 99% confidence interval for proportion of book sales transactions with sales amount being greater than $50.

Formulae: $p \pm Z\ \alpha/2\ \sqrt{p(1-p)/n}$

```
#compute 95% CI for proportion (Age>50)
book<-ST %>% filter(Product=="Book")
bk50<- book %>% filter(Amount>50)
pbk50<-nrow(bk50)/nrow(book)
lCIpbk50 <- pbk50 + (qnorm(0.005)*sqrt(pbk50*(1-pbk50)/nrow(book)))
uCIpbk50 <- pbk50 - (qnorm(0.005)*sqrt(pbk50*(1-pbk50)/nrow(book)))
print(cbind(lCIpbk50, uCIpbk50),digits=3)
```

```
##      lCIpbk50 uCIpbk50
## [1,]    0.139    0.267
```

1AIII - COMPUTE THE 90% PREDICTIVE INTERVAL FOR AMOUNT FOR ORDERS OF DVD.

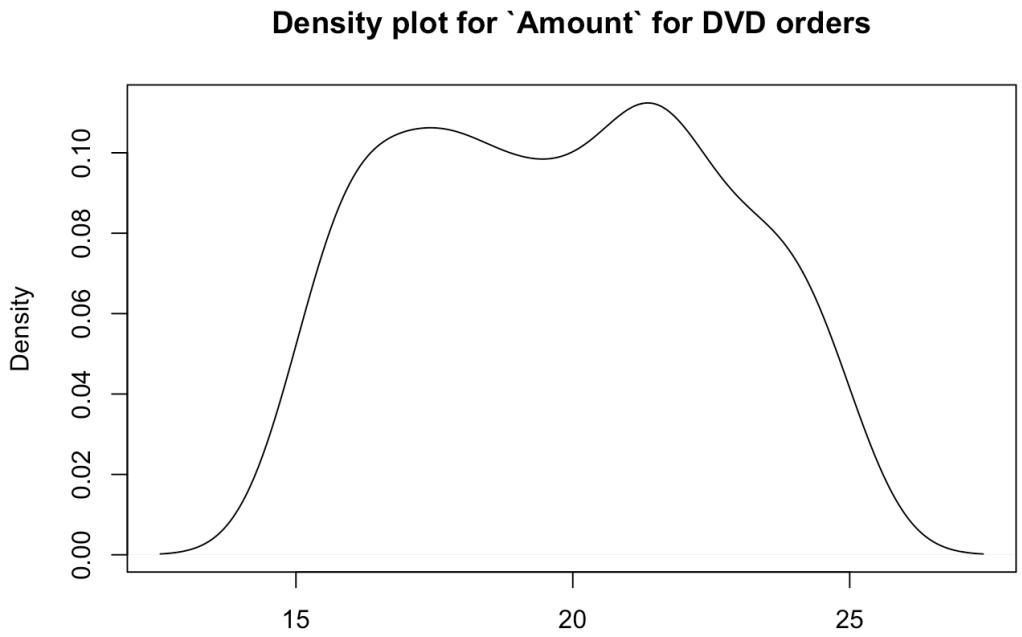
How are Predictive Intervals different from Confidence Intervals?

1AIII - COMPUTE THE 90% PREDICTIVE INTERVAL FOR AMOUNT FOR ORDERS OF DVD.

iii. compute the 90% predictive interval for Amount for orders of DVD.

Why should we check for normality here?

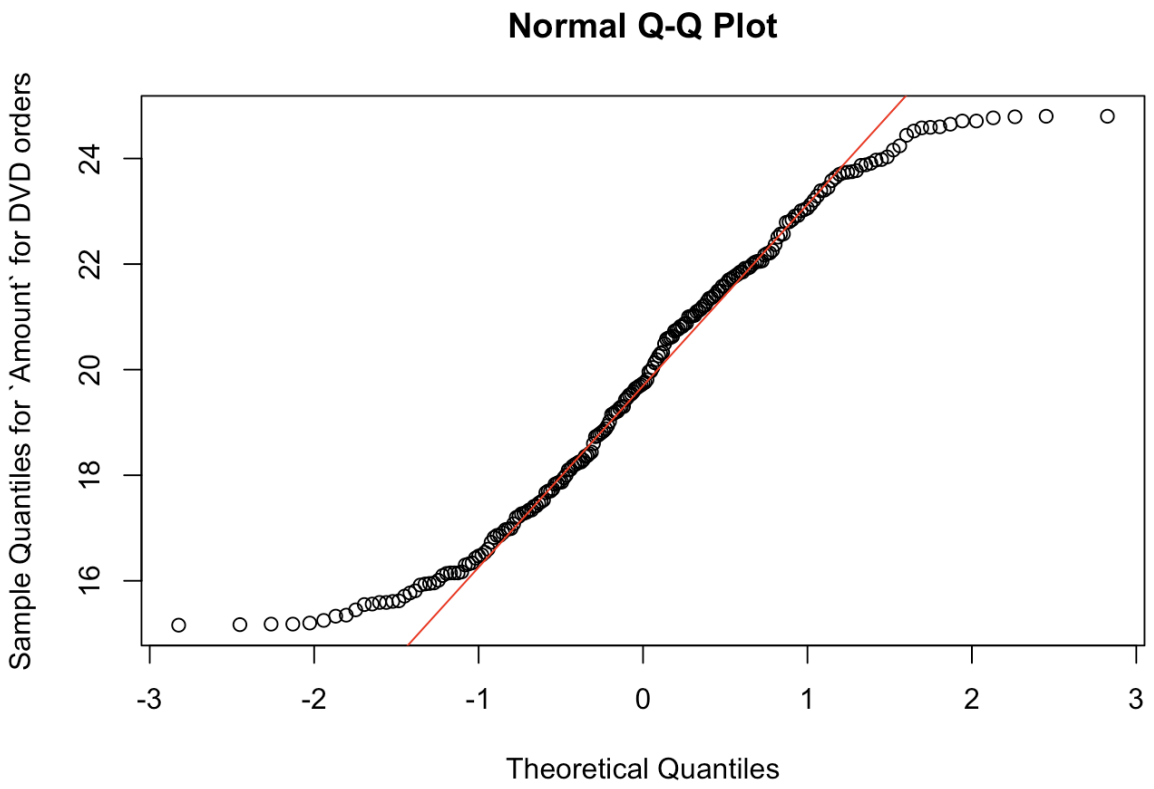
```
#compute 90% predictive interval for `Amount` for DVD
dvd<-ST %>% filter(Product=="DVD")
# check normality
plot(density(dvd$Amount),main="Density plot for `Amount` for DVD orders")
```



1AIII - QQNORM

iii. compute the 90% predictive interval for Amount for orders of DVD.

```
qqnorm(dvd$Amount,
       ylab="Sample Quantiles for `Amount` for DVD orders")
qqline(dvd$Amount,
       col="red")
```



1AIII

iii. compute the 90% predictive interval for Amount for orders of DVD.

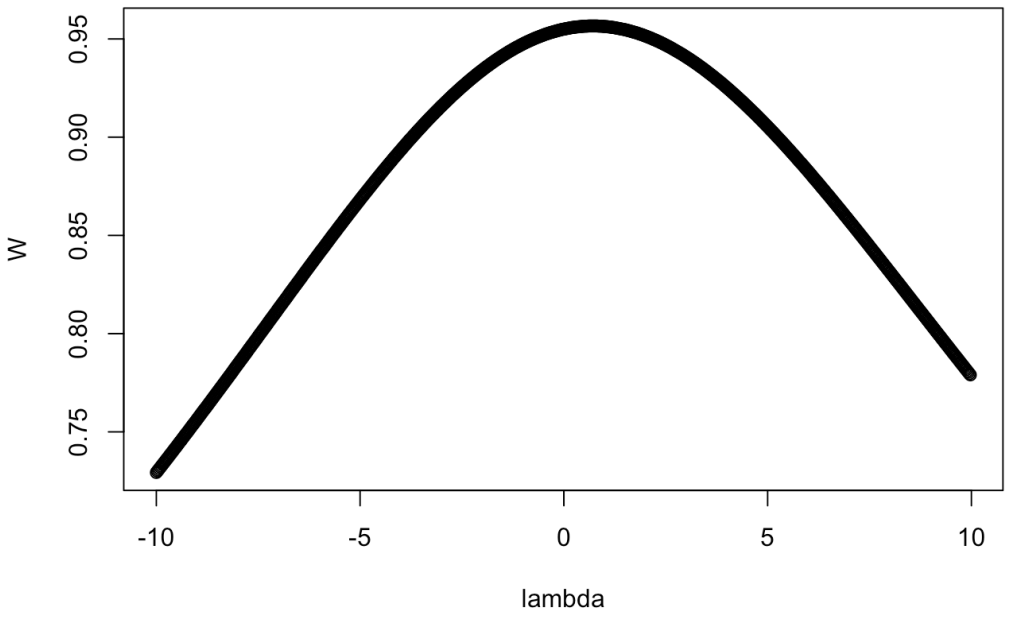
```
shapiro.test(dvd$Amount)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dvd$Amount
## W = 0.95635, p-value = 4.703e-06
```

1AIII – TRANSFORM TO NORMAL

iii. compute the 90% predictive interval for Amount for orders of DVD.

```
#transform data to normal distribution using transformTukey
dvd$Amt.t = transformTukey(dvd$Amount, plotit=TRUE)
```




```
    lambda      W Shapiro.p.value
425    0.6 0.9566       4.959e-06

if (lambda >  0){TRANS = x ^ lambda}
if (lambda == 0){TRANS = log(x)}
if (lambda <  0){TRANS = -1 * x ^ lambda}
```

1AIII – REVERSE TRANSFORM

iii. compute the 90% predictive interval for Amount for orders of DVD.

Formulae:

$$\bar{X} \pm t_{\alpha/2,n-1}\,(s\,\sqrt{(1+(1/n))}$$

```
#output lambda=0.6

#using x ^ lambda where lambda = 0.6
mnamt.t <- mean(dvd$Amt.t)
sdamt.t <- sd(dvd$Amt.t)
lPI.amtt <- mnamt.t + (qt(0.05, df = (nrow(dvd)-1))*sdamt.t*sqrt(1+1/nrow(dvd)))
uPI.amtt <- mnamt.t - (qt(0.05, df = (nrow(dvd)-1))*sdamt.t*sqrt(1+1/nrow(dvd)))
cbind(lPI.amtt, uPI.amtt)
```

```
##      lPI.amtt uPI.amtt
## [1,]  5.13671 6.837365
```

```
#reverse transform; comments below is to derive the formula
# y= x^lamda
# y = x^0.6
# x = y^(1/0.6)
lPI.amt <- lPI.amtt^(1/0.6)
uPI.amt<- uPI.amtt^(1/0.6)

cbind(lPI.amt,uPI.amt) # reverse transform
```

```
##       lPI.amt  uPI.amt
## [1,] 15.29238 24.63095
```

QUESTION 1B: HYPOTHESIS TESTING

The manager would like to draw some conclusions from the sales transaction data. He does not believe there are outliers in the data so you may leave all the data as is. He would like your help to set up and test the following hypotheses.

i. Is the proportion of book sales transactions with Amount greater than $50 at least 10 percent of book sales transactions?
ii. Is the mean sales amount for books the same as for dvds?
iii. Is the mean sales amount the same across all 4 regions?

1BI

Is the proportion of book sales transactions with Amount greater than $50 at least 10 percent of book sales transactions?

What is the null and alternative hypothesis here?
Is it a two-sided or one-sided hypothesis test?

Null: P(Amount > 50) >= .1
Alt: P(Amount > 50) < .1
--> Left one-sided hypothesis test

1BI

Is the proportion of book sales transactions with Amount greater than $50 at least 10 percent of book sales transactions?

```
# compute z-statistic for proportion. The next 3 lines have been executed earlier.
#book<-ST %>% filter(Product=="Book")
#bk50<- book %>% filter(Amount>50)
#pbk50<-nrow(bk50)/nrow(book)

z <- (pbk50 - 0.10) / sqrt(0.1*(1-0.1)/nrow(book))
z
```

```
## [1] 5.550227
```

```
#compute critical value
cv95<-qnorm(0.05)
cv95
```

```
## [1] -1.644854
```

```
z<cv95
```

```
## [1] FALSE
```

1BII – COMPARING TWO GROUPS

ii. Is the mean sales amount for books the same as for dvds?

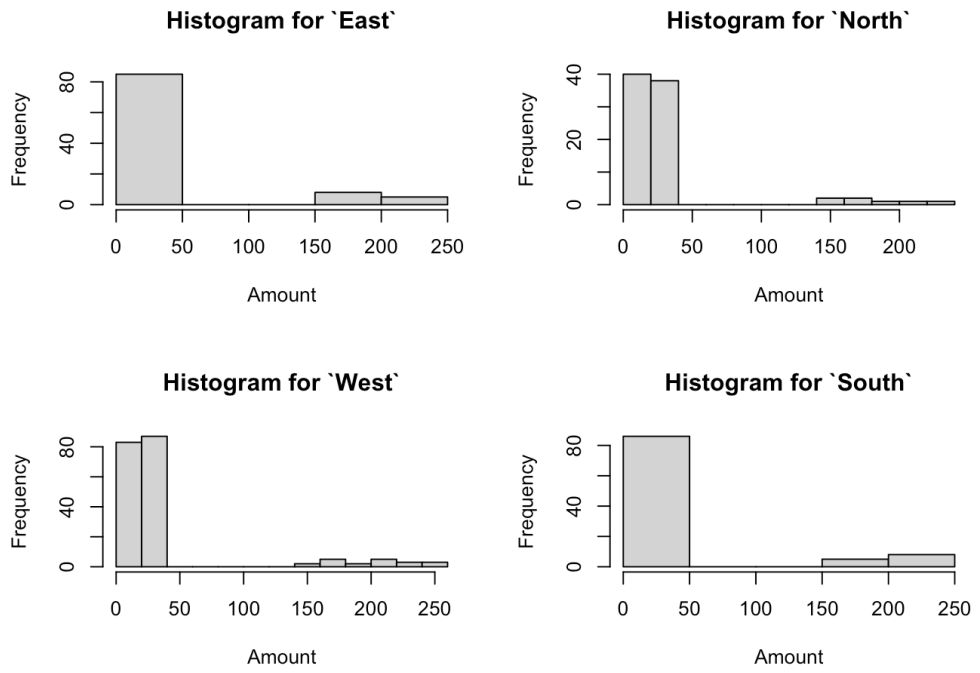
```
# ii)
t.test(ST$Amount~ST$Product)
```

```
##
##  Welch Two Sample t-test
##
## data:  ST$Amount by ST$Product
## t = 8.0304, df = 260.96, p-value = 3.344e-14
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  27.47079 45.31916
## sample estimates:
## mean in group Book  mean in group DVD
##           56.21559            19.82062
```

1BIII – NORMALITY ASSUMPTION

iii. Is the mean sales amount the same across all 4 regions?

```
# iii) Check if ANOVA assumptions are met
# check normality
# plot histogram
par(mfcol=c(2,2))
E<-ST %>% filter(Region=="East")
W<-ST %>% filter(Region=="West")
N<-ST %>% filter(Region=="North")
S<-ST %>% filter(Region=="South")


hist(E$Amount, main="Histogram for `East`", xlab="Amount")
hist(W$Amount, main="Histogram for `West`", xlab="Amount")
hist(N$Amount, main="Histogram for `North`", xlab="Amount")
hist(S$Amount, main="Histogram for `South`", xlab="Amount")
```



1BIII – VARIANCE ASSUMPTION

iii. Is the mean sales amount the same across all 4 regions?

```
# check sample sizes across regions
table(ST$Region)
```

```
##
##  East North South  West
##    98    85    99   190
```

```
# check equal variance assumption
fligner.test(Amount~ Region, ST)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Amount by Region
## Fligner-Killeen:med chi-squared = 5.9472, df = 3, p-value = 0.1142
```

1BIII - ANOVA

iii. Is the mean sales amount the same across all 4 regions?

```
# Conduct Anova
aov.amt<-aov(ST$Amount ~ as.factor(ST$Region)) #note the group variable should be a factor
summary(aov.amt)
```

```
##                       Df  Sum Sq Mean Sq F value Pr(>F)
## as.factor(ST$Region)   3    6549    2183   0.663  0.575
## Residuals            468 1540965    3293
```

1BIIII - POSTHOC

iii. Is the mean sales amount the same across all 4 regions?

```
TukeyHSD(aov.amt) # since anova result is not significant, by right there is no need to do post-hoc
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = ST$Amount ~ as.factor(ST$Region))
##
## $`as.factor(ST$Region)`
##                  diff       lwr      upr     p adj
## North-East  -9.035040 -30.96347 12.89339 0.7126439
## South-East   2.112268 -18.96949 23.19403 0.9939520
## West-East   -3.541668 -21.94137 14.85804 0.9598923
## South-North 11.147307 -10.72962 33.02423 0.5545926
## West-North   5.493372 -13.81228 24.79902 0.8835323
## West-South  -5.653936 -23.99223 12.68436 0.8567450
```

THANK YOU. SEE YOU NEXT WEEK.