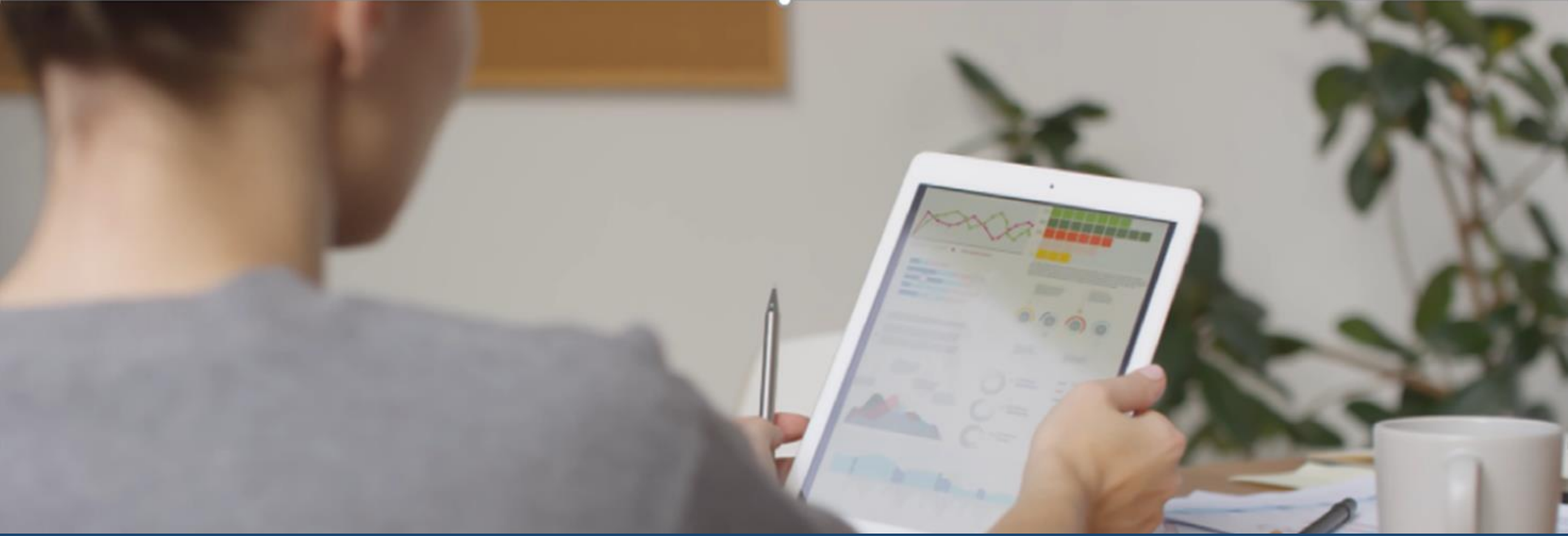
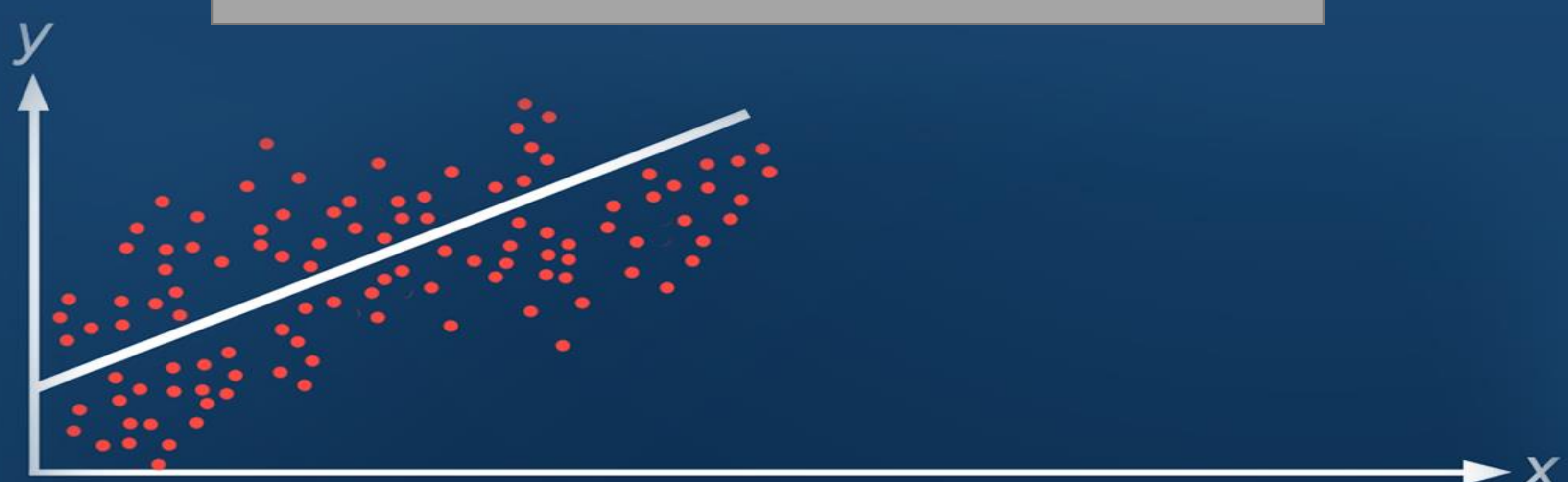


TBA2102 2020/2021 Semester 2
Tutorial 6



STRUCTURE OF TUTORIALS

Duration:

45 mins

Content:

- Tutorial 5 assignment
- Tutorial 6 (Qn 1)

Tutorial 5 Assignment

DATASET REQUIRED

The Purchase Orders 2.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
Arrival Time	Number of days for order to arrive (Arrival Date -Order Date)

Q2.(A) COMPUTING INTERVAL ESTIMATES

Using this 4-month sample data, assist the manager of this company to:

i. Compute the 90% confidence interval for the mean of Cost per order for orders with A/P Terms of 30 Months. (2 marks)

ii. Compute the 95% confidence interval for proportion of orders with Arrival Time less than 6 days. (2 marks)

iii. Compute the 99% predictive interval for Arrival Time of a new order with A/P Terms of 30 months. (4 marks)

Explain to the manager what each of these interval estimates mean. In your explanation, you must indicate clearly what parameter or value are being estimated and whether the estimates are for the population or sample. (3 marks)

QUESTION 2A

Compute the 90% confidence interval for the mean of Cost per order for orders with A/P Terms of 30 Months. (2 marks)

Why do we use t distribution?

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

What is the value of ∝ here?

```
ap30<-PO2%>%filter(`A/P Terms (Months)`==30)
lCIc90<- mean(ap30$`Cost per order`) + qt(0.05,df=nrow(ap30)-1)*sd(ap30$`Cost per order`)/sqrt(nrow(ap30))
uCIc90 <- mean(ap30$`Cost per order`) - qt(0.05,df=nrow(ap30)-1)*sd(ap30$`Cost per order`)/sqrt(nrow(ap30))
print(cbind(lCIc90, uCIc90), digits=7)
```

```
       lCIc90   uCIc90
[1,] 22366.69 34840.73
```

QUESTION 2A

Compute the 95% confidence interval for proportion of orders with Arrival Time less than 6 days. (2 marks)

z or t distribution?

$$\hat{p} \pm \text{[redacted]} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

```
ar6<- PO2 %>% filter(`Arrival Time`<6)
par6<-nrow(ar6)/nrow(PO2)
lCIpar6 <- par6 + (qnorm(0.025)*sqrt(par6*(1-par6)/nrow(PO2)))
uCIpar6 <- par6 - (qnorm(0.025)*sqrt(par6*(1-par6)/nrow(PO2)))
print(cbind(lCIpar6, uCIpar6),digits=3)
```

```
     lCIpar6 uCIpar6
[1,]   0.177   0.355
```

QUESTION 2A III

Compute the 99% predictive interval for Arrival Time of a new order with A/P Terms of 30 months. (4 marks)

Need to check Arrival Time distribution?

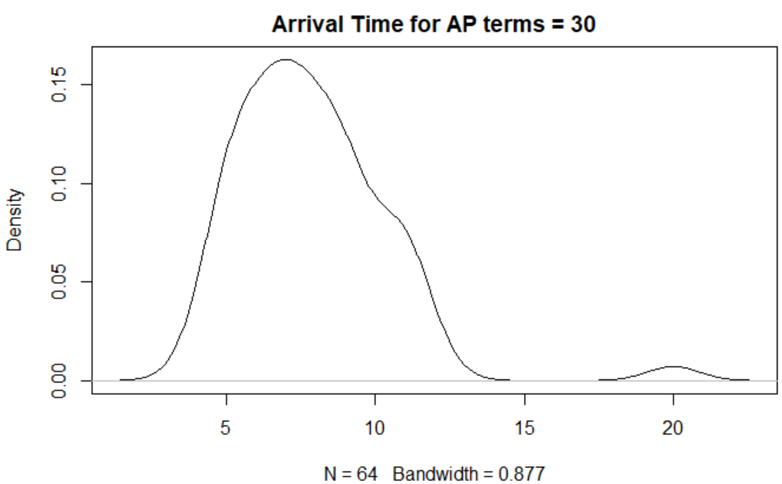
```
plot(density(ap30$`Arrival Time`),main="Arrival Time for AP terms = 30")

qqnorm(ap30$`Arrival Time`,
       ylab="Sample Quantiles for Arrival Time")
qqline(ap30$`Arrival Time`,
       col="red")

shapiro.test(ap30$`Arrival Time`)
```

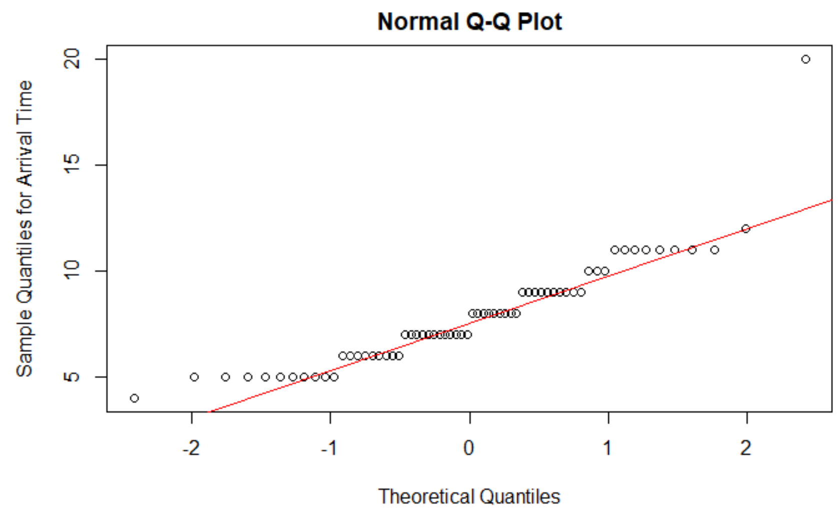
Outputs next slide

QUESTION 2A III

Compute the 99% predictive interval for Arrival Time of a new order with A/P Terms of 30 months. (4 marks)

Check Arrival Time distribution






```
Shapiro-Wilk normality test

data:  ap30$`Arrival Time`
W = 0.85854, p-value = 3.015e-06
```

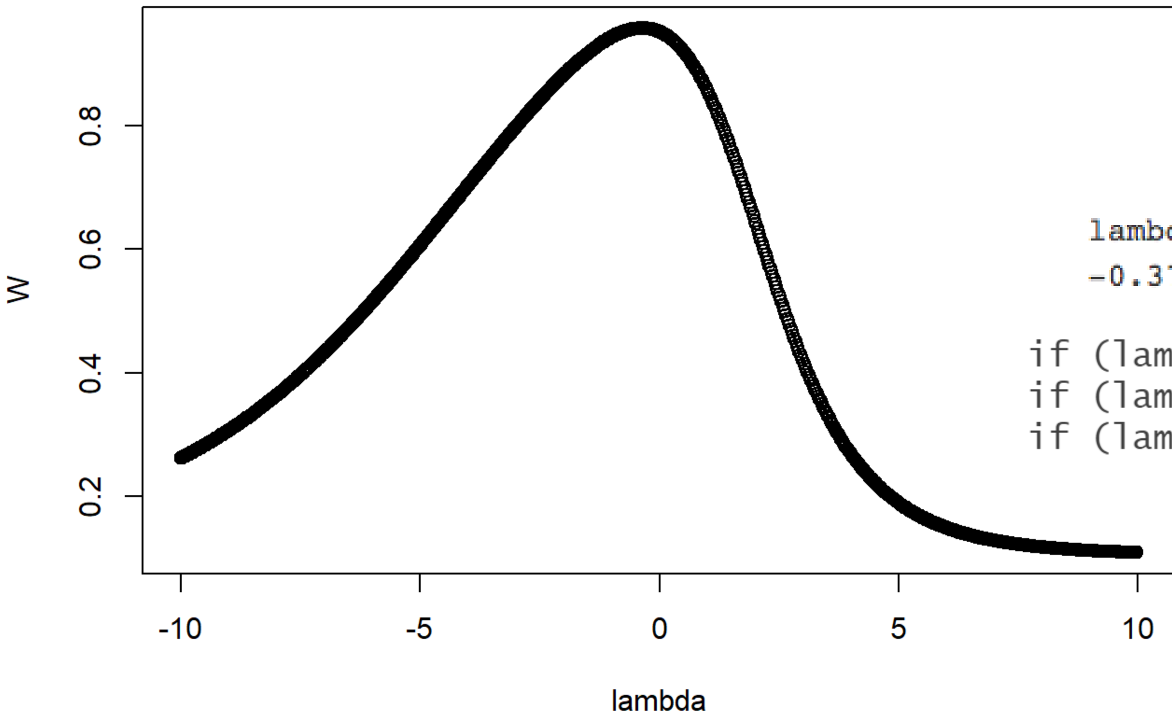
Transform data?

QUESTION 2A III

Compute the 99% predictive interval for Arrival Time of a new order with A/P Terms of 30 months. (4 marks)

Transform Arrival Time

```r
#transform data to normal distribution using transformTukey
ap30$atime.t = transformTukey(ap30$`Arrival Time`, plotit=TRUE)
```



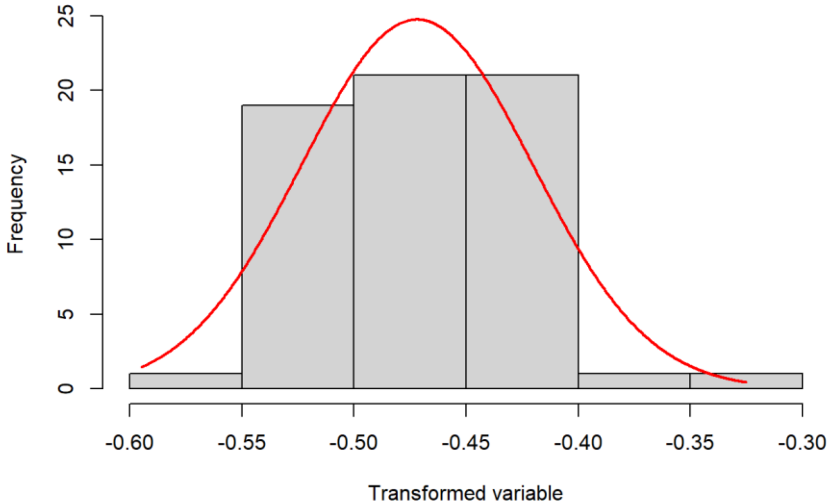
HOW?

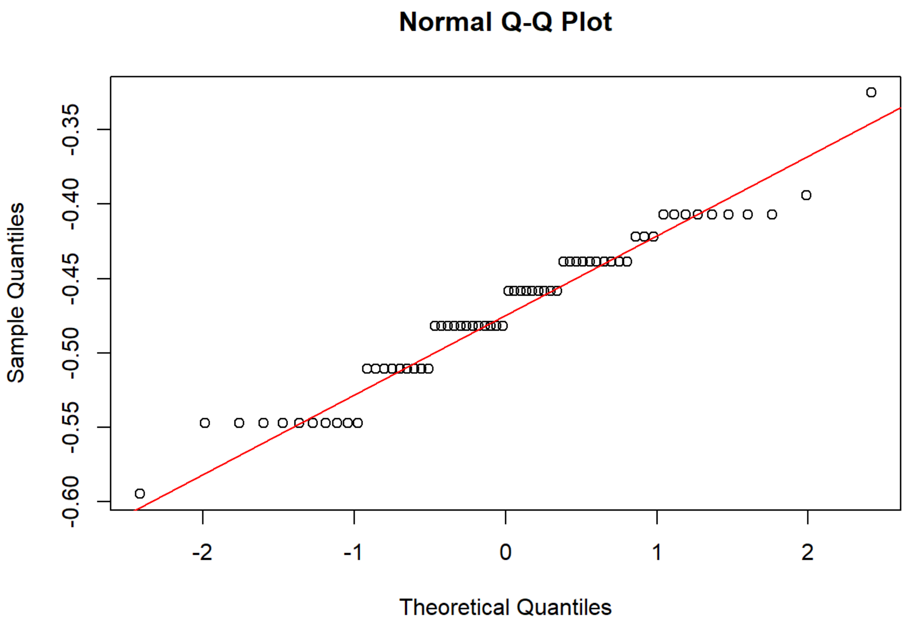
```
lambda      W Shapiro.p.value
-0.375 0.9587         0.03123
```

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

QUESTION 2A

Transformed Arrival Time





QUESTION 2A III

Compute the 99% predictive interval for Arrival Time of a new order with A/P Terms of 30 months. (4 marks)

```
mnap30.t <- mean(ap30$atime.t)
sdap30.t <- sd(ap30$atime.t)

nap30<-nrow(ap30)

lPI.ap30t <- mnap30.t + (qt(0.005, df = (nap30-1))*sdap30.t*sqrt(1+1/nap30))
uPI.ap30t <- mnap30.t - (qt(0.005, df = (nap30-1))*sdap30.t*sqrt(1+1/nap30))
print(cbind(lPI.ap30t, uPI.ap30t), digits=3)
```

$$\bar{x} \pm t_{\alpha/2,n-1}\left(s\sqrt{1+\frac{1}{n}}\right)$$

```
     lPI.ap30t uPI.ap30t
[1,]     -0.61    -0.334
```

```
lPI.ap30 <- (-1/lPI.ap30t)^(1/0.375)
uPI.ap30<- (-1/uPI.ap30t)^(1/0.375)

print(cbind(lPI.ap30,uPI.ap30),digits=3) # reverse transform
```

```
     lPI.ap30 uPI.ap30
[1,]     3.74     18.7
```

Q2.(B) HYPOTHESES TESTING

The manager of the company would like to draw some conclusions from the sample data he has. He too does not wish to remove any outlier data so you may use the datset as is to test the hypotheses below.

i. The proportion of orders with Arrival Time less than 6 days is greater than 15 percent of all orders. (4 marks)

ii. Is the mean Arrival Time for order with AP Terms of 30 months different from those with other AP terms? [Hint: Here you can treat orders with A/P Terms of 30 months to be one group where those with A/P Terms of 15, 25 and 45 months to be another group. You can create a new variable, e.g. ap.gp and assign it a value of “1” when A/P Terms (Months) is 30; and assign it a value of 2 if A/P Terms (Months) is 15, 25 or 45.] (4 marks)

iii. Is the mean Arrival Time the same across these 3 groups of suppliers. The suppliers are grouped according to the manager’s definition: “1” includes Fast-Tie Aerospace and Hulkey Fasteners; “2” includes Alum Sheeting, Durrable Products, and Manley Valve; 3 includes Steelpin Inc.,Spacetime Technologies,Pylon Accessories? [Hint: To help you with the hypothesis testing, ou would want to create a new variable s.gp (you can use a different name) to group the records by Suppliers first. You would assign the values of “1”, “2”, “3” tos.gp depending on the value of Suppliers for the records.] (6 marks)

QUESTION 2B

The proportion of orders with Arrival Time less than 6 days is greater than 15 percent of all orders. (4 marks)

Hypotheses

H_0 : p <= 0.15
H_1 : p > 0.15

Let p be probability of Arrival Time < 6
rejection region is right tailed

Test statistic

```
z <- (par6 - 0.15) / sqrt(0.15*(1-0.15)/nrow(PO2))
z
```

3.148527

Critical Value

```
cv95<-qnorm(0.95)
cv95
```

1.644854

Apply decision rule

```
z>cv95
```

TRUE

QUESTION 2B

Is the mean Arrival Time for order with AP Terms of 30 months different from those with other AP terms? [Hint: Here you can treat orders with A/P Terms of 30 months to be one group where those with A/P Terms of 15, 25 and 45 months to be another group. You can create a new variable, e.g. ap.gp and assign it a value of "1" when A/P Terms (Months) is 30; and assign it a value of 2 if A/P Terms (Months) is 15, 25 or 45.] (4 marks)

Hypotheses

H_0 : Mean Arrival time for Group 1 = Mean Arrival time for Group 2
H_1 : Mean Arrival time for Group 1 ≠ Mean Arrival time for Group 2

```
PO2$ap.gp[PO2$`A/P Terms (Months)`==30]<-1
PO2$ap.gp[PO2$`A/P Terms (Months)`!=30]<-2
t.test(PO2$`Arrival Time` ~ PO2$ap.gp)
        Welch Two Sample t-test

data:  PO2$`Arrival Time` by PO2$ap.gp
t = -1.2572, df = 32.474, p-value = 0.2176
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -4.430937  1.047604
sample estimates:
mean in group 1 mean in group 2
       7.875000        9.566667
```

- What is the NULL and ALTERNATIVE hypotheses?
- What is the decision rule?

QUESTION 2B

Is the mean Arrival Time the same **across these 3 groups of suppliers**. The suppliers are grouped according to the manager's definition: "1" includes Fast-Tie Aerospace and Hulkey Fasteners; "2" includes Alum Sheeting, Durrable Products, and Manley Valve; 3 includes Steelpin Inc.,Spacetime Technologies,Pylon Accessories?
[Hint: To help you with the hypothesis testing, you would want to create a new variable s.gp (you can use a different name) to group the records by Suppliers first. You would assign the values of "1", "2", "3" tos.gp depending on the value of Suppliers for the records.] (6 marks)

Hypotheses

H_0 : Mean arrival time is the same across the 3 groups
H_1 : At least one of the mean arrival time is different

- What is the NULL and ALTERNATIVE hypotheses?
- What test to conduct?

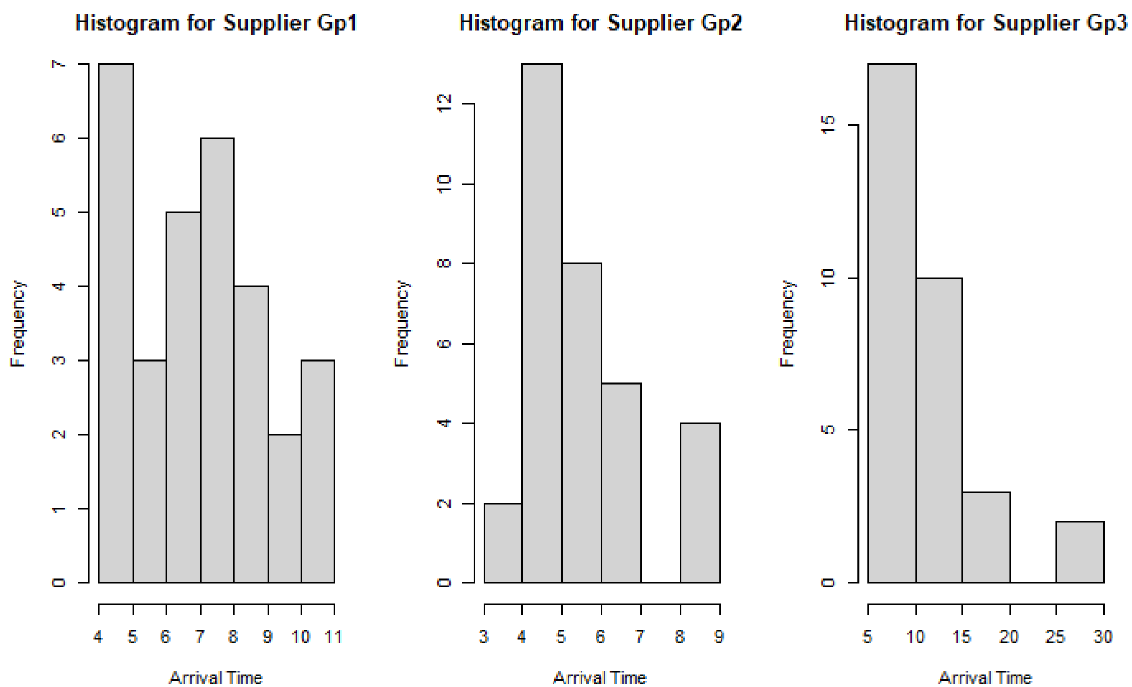
```
# Create grouping variable
PO2$s.gp[PO2$Supplier=="Fast-Tie Aerospace" | PO2$Supplier=="Hulkey Fasteners"]<-1
PO2$s.gp[PO2$Supplier=="Alum Sheeting" | PO2$Supplier=="Durrable Products" |
PO2$Supplier=="Manley Valve"]<-2
PO2$s.gp[PO2$Supplier=="Steelpin Inc." | PO2$Supplier=="Spacetime Technologies" |
PO2$Supplier=="Pylon Accessories"]<-3
```

```
sgp1<-PO2%>%filter(s.gp==1)
sgp2<-PO2%>%filter(s.gp==2)
sgp3<-PO2%>%filter(s.gp==3)
```

NORMALITY ASSUMPTION

Is the mean Arrival Time the same **across these 3 groups of suppliers**. The suppliers are grouped according to the manager's definition: "1" includes Fast-Tie Aerospace and Hulkey Fasteners; "2" includes Alum Sheeting, Durrable Products, and Manley Valve; 3 includes Steelpin Inc.,Spacetime Technologies,Pylon Accessories?

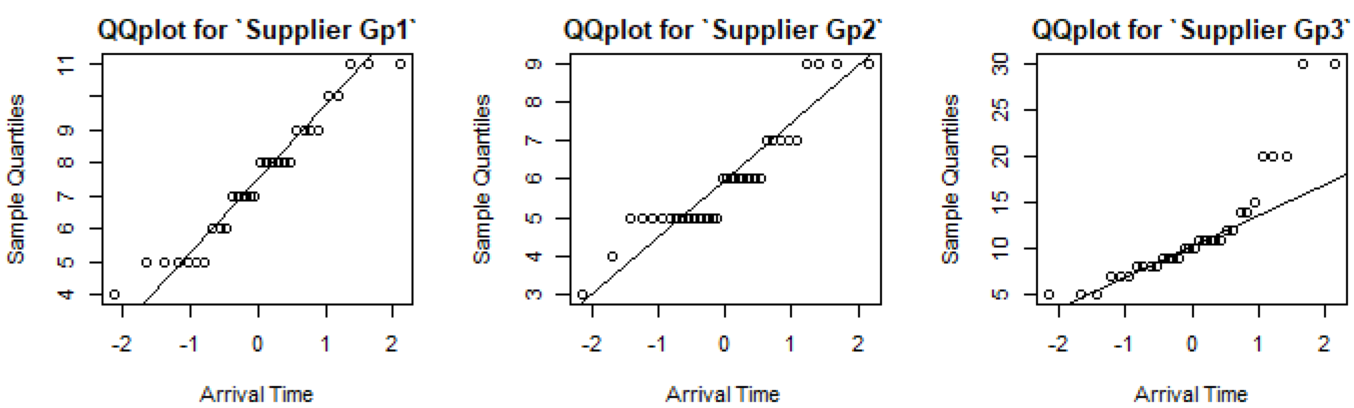
```
par(mfcol=c(1,3))
hist(sgp1$`Arrival Time`, main="Histogram for Supplier Gp1", xlab="Arrival Time")
hist(sgp2$`Arrival Time`, main="Histogram for Supplier Gp2", xlab="Arrival Time")
hist(sgp3$`Arrival Time`, main="Histogram for Supplier Gp3", xlab="Arrival Time")
```



NORMALITY ASSUMPTION

Is the mean Arrival Time the same **across these 3 groups of suppliers**. The suppliers are grouped according to the manager's definition: "1" includes Fast-Tie Aerospace and Hulkey Fasteners; "2" includes Alum Sheeting, Durrable Products, and Manley Valve; 3 includes Steelpin Inc.,Spacetime Technologies,Pylon Accessories?

```
par(mfrow=c(2,3))
qqnorm(sgp1$`Arrival Time`, main="QQplot for `Supplier Gp1`", xlab="Arrival Time")
qqline(sgp1$`Arrival Time`)
qqnorm(sgp2$`Arrival Time`, main="QQplot for `Supplier Gp2`", xlab="Arrival Time")
qqline(sgp2$`Arrival Time`)
qqnorm(sgp3$`Arrival Time`, main="QQplot for `Supplier Gp3`", xlab="Arrival Time")
qqline(sgp3$`Arrival Time`)
```



NORMALITY ASSUMPTION

Is the mean Arrival Time the same **across these 3 groups of suppliers**. The suppliers are grouped according to the manager's definition: "1" includes Fast-Tie Aerospace and Hulkey Fasteners; "2" includes Alum Sheeting, Durrable Products, and Manley Valve; 3 includes Steelpin Inc.,Spacetime Technologies,Pylon Accessories?

```
# Shapiro Wilkins Test
lapply(list(sgp1,sgp2,sgp3),
       function(sa)
         {
         shapiro.test(sa$`Arrival Time`)
         })
```

```
[[1]]

        Shapiro-Wilk normality test

data:  sa$`Arrival Time`
W = 0.94325, p-value = 0.1113
```

```
[[2]]

        Shapiro-Wilk normality test

data:  sa$`Arrival Time`
W = 0.86562, p-value = 0.0009216
```

```
[[3]]

        Shapiro-Wilk normality test

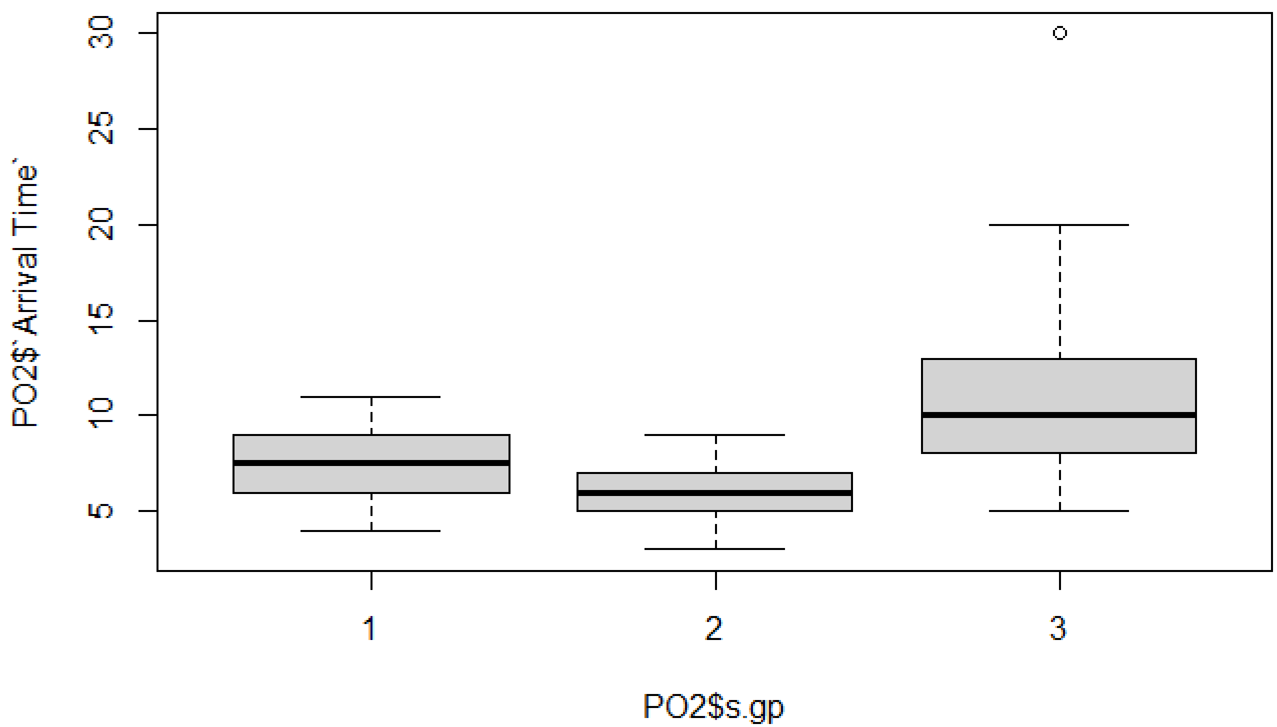
data:  sa$`Arrival Time`
W = 0.79914, p-value = 3.927e-05
```

OUTLIERS

Is the mean Arrival Time the same **across these 3 groups of suppliers**. The suppliers are grouped according to the manager's definition: "1" includes Fast-Tie Aerospace and Hulkey Fasteners; "2" includes Alum Sheeting, Durrable Products, and Manley Valve; 3 includes Steelpin Inc.,Spacetime Technologies,Pylon Accessories?

Not necessary but just to show how boxplots can be used too

```
# plot boxplots
boxplot(PO2$`Arrival Time` ~ PO2$s.gp)
```



HOMOGENEITY OF VARIANCE ASSUMPTION

Is the mean Arrival Time the same **across these 3 groups of suppliers**. The suppliers are grouped according to the manager's definition: "1" includes Fast-Tie Aerospace and Hulkey Fasteners; "2" includes Alum Sheeting, Durrable Products, and Manley Valve; 3 includes Steelpin Inc.,Spacetime Technologies,Pylon Accessories?

```
# check if samples are equal
table(PO2$s.gp)
```

1	2	3
30	32	32

What does this tell us?

Since samples are almost equal in size, it is ok not to worry about the equal variance assumption and proceed to conduct ANOVA

ANOVA

Is the mean Arrival Time the same **across these 3 groups of suppliers**. The suppliers are grouped according to the manager's definition: "1" includes Fast-Tie Aerospace and Hulkey Fasteners; "2" includes Alum Sheeting, Durrable Products, and Manley Valve; 3 includes Steelpin Inc.,Spacetime Technologies,Pylon Accessories?

```
aov.time<-aov(PO2$`Arrival Time` ~ as.factor(PO2$s.gp))
summary(aov.time)
```

```
                    Df Sum Sq Mean Sq F value   Pr(>F)
as.factor(PO2$s.gp)  2  574.4  287.19   18.96 1.31e-07
Residuals           91 1378.4   15.15

as.factor(PO2$s.gp) ***
Residuals
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- What is the decision rule?

ANOVA: POSTHOC ANALYSIS

```
TukeyHSD(aov.time)
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = PO2$`Arrival Time` ~ as.factor(PO2$s.gp))

$`as.factor(PO2$s.gp)`
         diff       lwr       upr     p adj
2-1 -1.497917 -3.854568 0.8587342 0.2889516
3-1  4.283333  1.926682 6.6399842 0.0001124
3-2  5.781250  3.462921 8.0995788 0.0000002
```

how do we interpret this?

Tutorial 6

DATASET REQUIRED

Tutorial6_WorldBankData.csv
(Note: This dataset comes from a publically available dataset from The World Bank. https://databank.worldbank.org/source/world-development-indicators.)

There are 8 variables in this (real) dataset, from 258 countries in 2016/2017:

- **Human.Capital.Index :** Unitless number that goes from 0 to 1.
- **GDP.per.capita.PPP:** In $. This is GDP per capita, but taking into account the purchasing power of the local currency, by comparing how much it costs to buy a basket of goods (e.g. food) compared to the reference currency (USD). (PPP stands for Purchasing Power Parity)
- **Health.Expenditure.per.capita**. In $.
- **Tertiary.Education.Expenditure.per.student**. In $.
- **Population**. In people.
- **Life.Expectancy.at.birth**. In years.
- **Diabetes.Prevalence**. In units of % of population ages 20 to 79.
- **Years.of.Compulsory.Education**. In years.

This being a real dataset, there is lots of missing data. Be wary of this!

EXPLORING THE DATASET REQUIRED: WORLDBANKDATA.CSV

Understand the key variables e.g., Human Capital Index

How is it defined?

Let's look at descriptive stats

Are there any missing values? Is there something interesting about the missingness?

Measures which countries are best in mobilizing the economic and professional potential of its citizens.

- Takes on values 0 to 1
- The lowest and highest HCI values are 0.293 and 0.884.

describe(dta_wb$Human.Capital.Index)

- Recall that the describe function is from the psych package

Can use the filter function

Study NAs and see if there might be something interesting there.

QUESTION 1A

First, let's investigate Human.Capital.Index. As noted by Prime Minister Lee in his 2019 National Day Rally, Singapore topped the world on this Human Capital Index in 2018. Let's try to see what are some of the possible variables that correlate with this.

Start off by plotting Human.Capital.Index (on the y-axis) versus GDP.per.capita.PPP on the x-axis. What do you notice? What type of relationship exists between the two variables? Is it linear?

Make a new variable that is the base10 logarithm of GDP.per.capita.PPP. Plot Human.Capital.Index against this new variable. What do you notice now?

For fun: try to add color to the dot that represents Singapore (i.e., make Singapore a red dot).

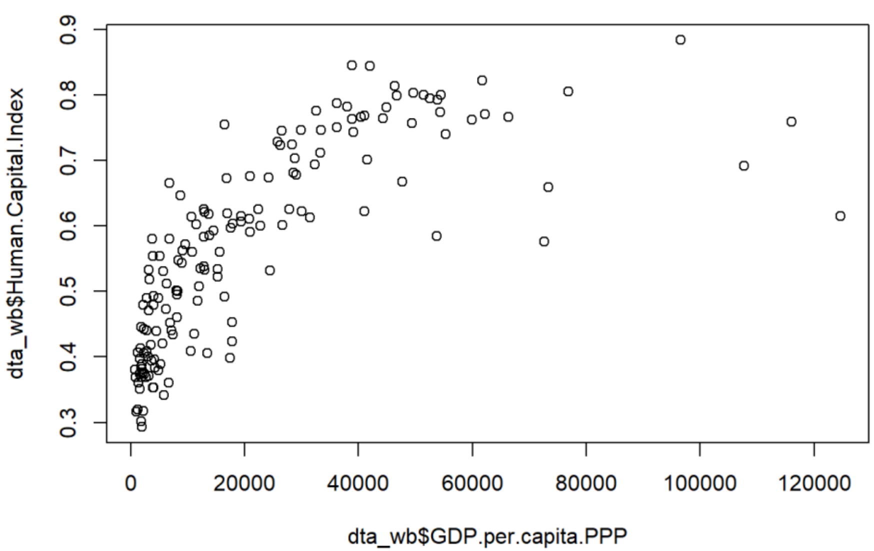
QUESTION 1A

First, let's investigate Human.Capital.Index. As noted by Prime Minister Lee in his 2019 National Day Rally, Singapore topped the world on this Human Capital Index in 2018. Let's try to see what are some of the possible variables that correlate with this.

(1a) Start off by plotting Human.Capital.Index (on the y-axis) vs. GDP.per.capita.PPP on the x-axis. What do you notice? What type of relationship exists between the two variables? Is it linear?

````
```{r q1a-plot, echo=TRUE}
plot(dta_wb$GDP.per.capita.PPP, dta_wb$Human.Capital.Index
```
````

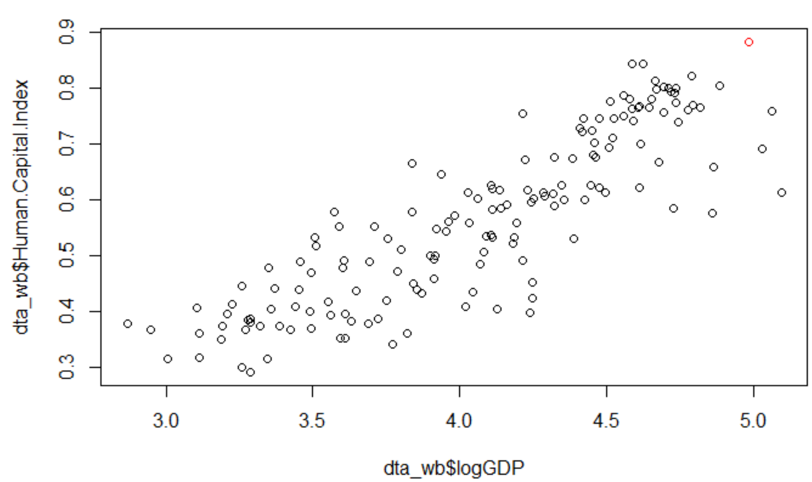
- GDP per capita seems related to Human Capital Index:
- Countries with higher GDP per capita also tend to have higher Human Capital Index.
- Non-linear relationship



QUESTION 1A

Make a new variable that is the base10 logarithm of GDP.per.capita.PPP. Plot Human.Capital.Index against this new variable. What do you notice now?

For fun: try to add color to the dot that represents Singapore (i.e., make Singapore a red dot).

```
```{r q1b-plot, echo=TRUE}
# NOTE: log() is the natural log; log10() is logarithm of base 10.
dta_wb$logGDP = log10(dta_wb$GDP.per.capita.PPP)
plot(dta_wb$logGDP, dta_wb$Human.Capital.Index, col =
[redacted]
```
```

log10 is more interpretable than ln
Eg: log10 GDP is 5.0
→ GDP is 10^5 or $100,000

Linear relationship between Human Capital Index and log-GDP-per-capita



QUESTION 1B

Now that you have a plot of a linear relationship, run a linear regression using lm(), predicting Human Capital Index. Run summary(...) on the lm object to produce an output table. Interpret the output of the lm(). What do the b coefficients mean? (Interpret them and try to make sense of the numbers, even if they might seem weird at first.)

How many countries made it into this regression? (What happened to the rest?)

Comment on the goodness-of-fit statistics.

QUESTION 1B

```
```{r q1c-lm, echo=TRUE}
summary(lm(Human.Capital.Index ~ logGDP, dta_wb))
```
```

```
Call:
lm(formula = Human.Capital.Index ~ logGDP, data = dta_wb)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.21270 -0.04959  0.01103  0.06164  0.15487 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.43264    0.04726  -9.155 3.03e-16 ***
logGDP       0.24602    0.01153  21.335  < 2e-16 ***
---
Signif. codes:  
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.07666 on 155 degrees of freedom
  (101 observations deleted due to missingness)
Multiple R-squared:  0.746,	Adjusted R-squared:  0.7443 
F-statistic: 455.2 on 1 and 155 DF,  p-value: < 2.2e-16
```

The intercept (b0) is -0.43. This shows when a country has logGDP=0 the average Human Capital Index is -0.43.

QUESTION 1B

```
```{r q1c-lm, echo=TRUE}
summary(lm(Human.Capital.Index ~ logGDP, dta_wb))
```
```

```
Call:
lm(formula = Human.Capital.Index ~ logGDP, data = dta_wb)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.21270 -0.04959  0.01103  0.06164  0.15487 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.43264    0.04726  -9.155 3.03e-16 ***
logGDP       0.24602    0.01153  21.335  < 2e-16 ***
---
Signif. codes:  
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.07666 on 155 degrees of freedom
  (101 observations deleted due to missingness)
Multiple R-squared:  0.746,	Adjusted R-squared:  0.7443 
F-statistic: 455.2 on 1 and 155 DF,  p-value: < 2.2e-16
```

The slope (b1) is 0.246. When the log10-GDP of a country increases by 1 unit, we expect to see an average increase of Human Capital Index by 0.246.

t-value is very large, and the p-value is less than 0.05. What's the implication?

We can reject the null hypothesis that this slope $\beta 1 = 0$, i.e., $\beta 1$ is statistically significantly different from zero. (why $\beta 1$ and not b1?)

QUESTION 1B

How many countries made it into this regression? (What happened to the rest?)

```
Call:
lm(formula = Human.Capital.Index ~ logGDP, data = dta_wb)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.21270 -0.04959  0.01103  0.06164  0.15487 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.43264    0.04726  -9.155 3.03e-16 ***
logGDP       0.24602    0.01153  21.335  < 2e-16 ***
---
Signif. codes:  
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.07666 on 155 degrees of freedom
  (101 observations deleted due to missingness)
Multiple R-squared:  0.746,	Adjusted R-squared:  0.7443 
F-statistic: 455.2 on 1 and 155 DF,  p-value: < 2.2e-16
```

There are a total of 258 (`nrow(dta_wb)`) countries.

`summary()` shows **101 observations were deleted due to missingness** leaving only 157 countries used in the regression analysis.

QUESTION 1B

Comment on the goodness-of-fit statistics.

```
Call:
lm(formula = Human.Capital.Index ~ logGDP, data = dta_wb)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.21270 -0.04959  0.01103  0.06164  0.15487 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.43264    0.04726  -9.155 3.03e-16 ***
logGDP       0.24602    0.01153  21.335  < 2e-16 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.07666 on 155 degrees of freedom
  (101 observations deleted due to missingness)
Multiple R-squared:  0.746,     Adjusted R-squared:  0.7443 
F-statistic: 455.2 on 1 and 155 DF,  p-value: < 2.2e-16
```

R-square is 0.746, which means that this **model explains almost 75% of the total variation of `Human.Capital.Index`.**

F-test's p-value is significant implies "**at least one of βs is not zero**".
Given only one predictor → $\beta 1$ is statistically not zero.

THANK YOU. SEE YOU NEXT WEEK.