UNIVERSITI TEKNOLOGI MALAYSIA
SCHOOL OF COMPUTING
SESSION 2019/2020 SEMESTER 2

COURSE CODE
SECI2143 – Probability & Statistical Data Analysis

LECTURE’S NAME
DR CHAN WENG HOWE

INDIVIDUAL ASSIGNMENT
TITLE

PROJECT 2

STUDENT’S NAME
AZRIANA BINTI ZAINAL ABIDIN

MATRIC NO
A19EC0027

SECTION
SECTION 02

INTRODUCTION

This dataset is collected by Jakki that is from Hyderabad, Telangana, India . The purpose of this dataset is to show the difference of multiple students based on gender and their preparation for an examination. Other than that this dataset also shown their group race ethnicity, gender and what is their type of lunch. The type of lunch are free/reduced and standard lunch. For the examination it is divided to three paper which are math score, writing score and reading score. Lastly, the final variable is parent level education. The reason I choose this dataset is to examine what actually affect the student score. This dataset have many probability which are lunch type affect score, test preparation affect score or even parent level od education affect score.

HYPOTHESIS TESTING

For this dataset I choose :

- One Sample Hypothesis Testing Mean
- Correlation
- Regression
- T-Test
- Chisquare
- One Way Contingency Test
- Two Way Conitngency Test

ONE SAMPLE HYPOTHESIS TESTING MEAN

The main of one sample hypothesis testing is test a hypothesis. For example it is to test whether a population mean is significantly different from some hypothesized values.

Ho: population mean for math score = 80

Hypothesis null :population mean for math score equal to 80

H1: population mean for math score ≠ 80

Alternative hypothesis mean for math score not equal to 80

```
> #ONE SAMPLE HYPOTHESIS TESTING MEAN (math score)
```

Method 1: Critical Region

```
> #population variance unknown
> mu=80    #null Ho
> #H1=mu >80
> alpha = 0.1
> z=(xbar-mu)/(stdDev/sqrt(n)) #test statistic
> z.alpha = qnorm(1-alpha)    #critical value
> z            #tesrtresult
[1] -9.663916
> z.alpha         #cv
[1] 1.281552
```

Since the test statistic,z = -9.663916 < critical value,z.alpha = 1.281552. We fail to reject Ho null hypothesis. There is sufficient evidence to conclude that the population mean of math score is equal to 80.

Method 1: P-value

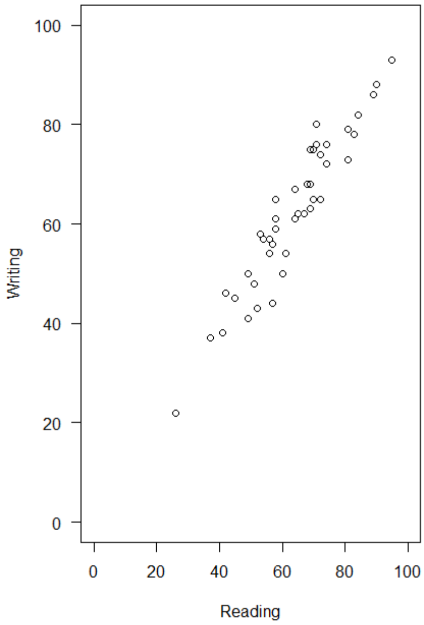
```
> pval<-pnorm(z, lower.tail=FALSE)
> pval
[1] 1
> alpha
[1] 0.1
```

Since the p-value, pval = 1 > significance level, alpha = 0.10, we fail to reject the null hypothesis, H_0. There is sufficient evidence to conclude that the population mean of math score is equal to 80.

CORRELATION

Correlation analysis is to measure the strength of relationship between reading and writing.

```
> #CORRELATION    : cor(x,y)   x - reading, y - writing
> cor(dataL$reading.score,dataL$writing.score,method = "pearson")
[1] 0.9481469
>
> plot(reading.score,writing.score, xlim=c(0,100), ylim = c(0,100), xlab="
Reading", ylab="Writing", las=1, pch=1)
```



From the correlation test that I have run in Rstudio the correlation coefficient value between reading score and writing score is 0.9481469. As we know correlation value that near to 1.0 means that the strength of correlation coefficient relationship between the 2 variables is strong. So now I can conclude that the strength of relationship between writing score variable and reading score variable is strong since my correlation value is 0.9481469.

Besides that, to represent correlation I used scatterplot. A scatterplot is used to represent a correlation between two variables. Scatterplot can be interpreted by looking at the direction of the line of best fit and how far the data point lie away from the line of best fit. So based on my scatterplot above I can conclude that the scatterplot is positive linear association and has strong relationship.

```
> cor.test(dataL$reading.score,dataL$writing.score, method = "pearson", co
nf.level = 0.90)

	Pearson's product-moment correlation

data:  dataL$reading.score and dataL$writing.score
t = 19.333, df = 42, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
90 percent confidence interval:
 0.9148071 0.9686530
sample estimates:
      cor
0.9481469
```

r = 0.9481469	t=19.333
df = 42	α = 0.10
Ho : p = 0	p-value = $2.2\text{e} \times 10^{-16}$
H1 : p ≠ 0	

Method 1: P-value

The null hypothesis, Ho states that no linear correlation while the alternative hypothesis, H1 states the linear correlation exists. Since p-value = 2.2×10^{-16} < significance level, α = 0.10, we reject the null hypothesis, H_0. There is sufficient evidence to support that there is linear correlation between reading score and writing score.

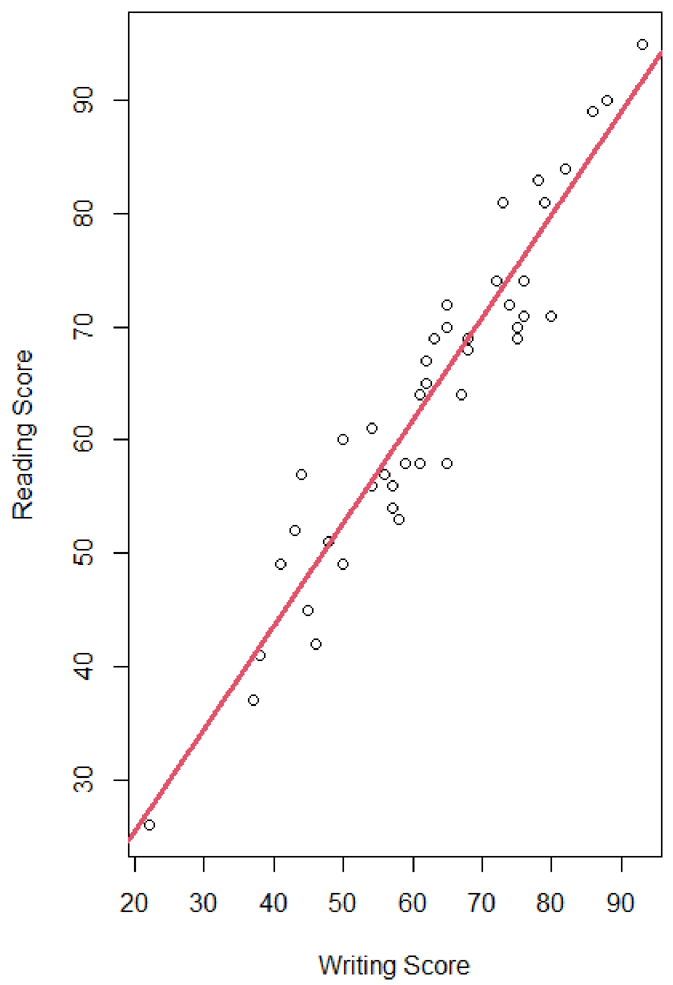
Method 2: T-test

The null hypothesis, Ho states that no linear correlation while the alternative hypothesis, H1 states the linear correlation exists. Since test statistics, t = 19.333 > critical t value = 1.282, we reject the null hypothesis, H_0. There is sufficient evidence to support that there is linear correlation between reading score and writing score.

REGRESSION

Regression analysis is a powerful statistical method that allows you to examine the relationship between two or more variables of interest.

```
> #REGRESSION (y: dependent(witing), x: independent(reading), lim(y,x)
> regr<-lm(reading.score~writing.score)
> regr

Call:
lm(formula = reading.score ~ writing.score)

Coefficients:
  (Intercept)  writing.score
       7.1266         0.9088
> plot(dataL$writing.score,dataL$reading.score)
> plot(dataL$writing.score,dataL$reading.score,xlab="Writing Score", ylab=
"Reading Score")
> abline(regr) #build regression line
> abline(mod,col=2,lwd=3) # to change line colors
```



The sign of a regression coefficient tells you whether there is a positive or negative correlation between each independent variable the dependent variable.

The regression test that has been done,the coeffiecient of determination between reading score and writing score is 0.948 which tells us that its shows positive correlation between writing score and reading score.

A positive coefficient indicates that as the value of the independent variable increases, the mean of the dependent variable also tends to increase. As a conclusion ,the regression line show that the reading score increases as the writing score increases.

```
> summary(regr)

Call:
lm(formula = reading.score ~ writing.score)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.8266 -3.6106 -0.0603  3.5784  9.8884 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)     7.1266     3.0056   2.371   0.0224 *  
writing.score   0.9087     0.0470  19.333   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.676 on 42 degrees of freedom
Multiple R-squared:  0.899,	Adjusted R-squared:  0.8966 
F-statistic: 373.8 on 1 and 42 DF,  p-value: < 2.2e-16
```

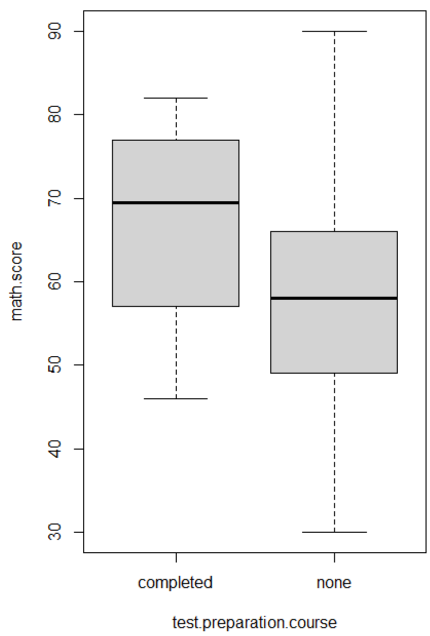
Ho: $\beta 1 = 0$	p-value(intercept): 0.02244
H1: $\beta 1 \neq 0$	p-value(slope) : $2e \times 10^{-16}$

Null hypothesis stated that there is no linear relationship between reading score and writing score while alternative hypothesis ,H1 stated that the linear relationship exist.
Since those two p-value < significance level = 0.1 , we reject null hypothesis ,Ho. There is sufficient evidence to support that there is linear relationship between reading score and wr iting score. Furthermore, there is sufficient evidence to support that reading score affect wri ting score.

T-TEST

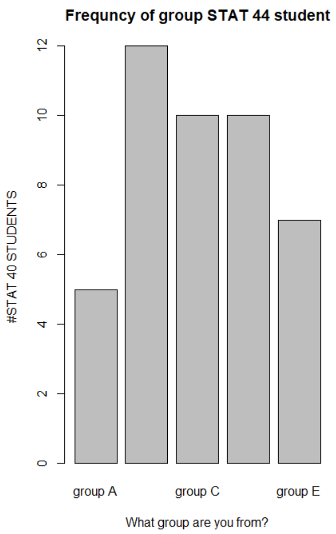
A **t-test** is a type of inferential **statistic** used to determine if there is a significant difference between the means of two groups,

```
> #T TEST
> t.test(math.score~test.preparation.course,mu=0,alt="two.sided",conf=0.05
,var.eq=F,paired=F)

        Welch Two Sample t-test

data:  math.score by test.preparation.course
t = 2.1332, df = 15.984, p-value = 0.04876
alternative hypothesis: true difference in means is not equal to 0
5 percent confidence interval:
 9.256266 9.826087
sample estimates:
mean in group completed      mean in group none
               67.60000                58.05882

> boxplot(math.score ~ test.preparation.course)
```

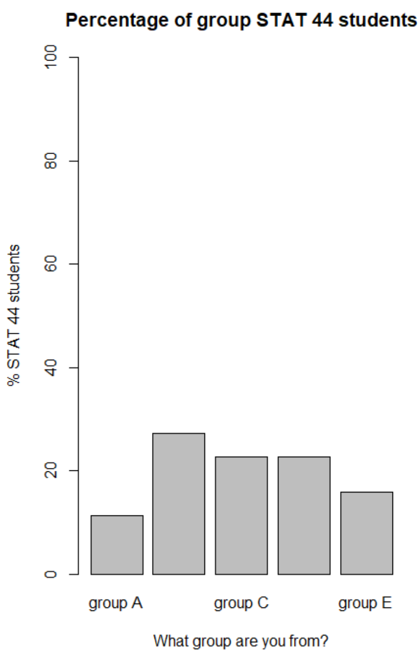


CHI-SQUARE TEST

```
> #CHI-SQUARE TEST
> names(dataL)
[1] "gender"                    "race.ethnicity"              "parental.
level.of.education"
[4] "lunch"                     "test.preparation.course"     "math.scor
e"
[7] "reading.score"             "writing.score"
> race.tab<-table(race.ethnicity)
> race.tab
race.ethnicity
group A group B group C group D group E
      5      12      10      10      7
> barplot(race.tab,xlab = "What group are you from?",ylab = "#STAT 40 STUD
ENTS",main="Frequncy of group STAT 44 student",beside = TRUE)
```




```
                          >
> race.proportion.tab<-(race.tab/sum(race.tab))
> race.proportion.tab        #will show percentage
race.ethnicity
  group A   group B   group C   group D   group E
0.1136364 0.2727273 0.2272727 0.2272727 0.1590909
> barplot((race.proportion.tab)*100,xlab="What group
are you from?",ylab="% STAT 44 students",main="Perce
ntage of group STAT 44 students",beside=TRUE,ylim=c(
0,100))
```



ONE WAY CONTINGENCY TEST

The Chi Square distribution can be used to test whether observed data differ significantly from theoretical expectations.

Ho: Parent Level Education is indepedent

H1: Parent Level Education is dependent(related)

```
> #ONEWAY CONTINGENCY
> #number of education
> table1<-table(parental.level.of.education)
> table1
parental.level.of.education
       high school associate's degree  bachelor's degree        high schoo
l    master's degree 
                 1                 17                  5                 1
4                  7 
> k=5
> 
> #insert value from table
> numofedu<- c(1,17,5,14,7)
> 
> #expected probability
> expectprob <-sum(numofedu)/5
> 
> #write expectprob k times
> expectEdu <-c(expectprob,expectprob,expectprob,expectprob,expectprob)
> 
> #test statistic
> exp1 <-((numofedu-expectEdu)^2)/expectEdu
> x1 <- sum(exp1)   #result test
> x1
[1] 19.63636

> 
> #critical value
> alpha2 <- 0.1
> x1.alpha2 <- qchisq(alpha2,df=4)
> x1.alpha2 <- qchisq(alpha2,df=4,lower.tail = FALSE)
> x1.alpha2
[1] 7.77944

> output <- chisq.test(numofedu,correct = FALSE)
> output

	Chi-squared test for given probabilities

data:  numofedu
X-squared = 19.636, df = 4, p-value = 0.0005891
```

∴ Since the test statistic = 19.63636 > critical value = 7.77944. We will reject Ho, null hypothesis . There is sufficient evidence to conclude that variable parent level education is dependent.

TWO WAY CONTINGENCY TEST

The chi-square test provides a method for testing the association between the row and colu mn variables in a two-way table

Ho: There is no association between gender and lunch variables

Hypothesis null :gender variable does not vary according to the lunch variables

H1: There is association exist between gender and lunch variables

Alternative hypothesis : gender variable vary according to the lunch variables

```
> #TWO WAY CONTINGENCY
> #bwtween gender and lunch
> table2 <- table (gender,lunch)
> table2
        lunch
gender   free/reduced standard
  female            7       15
  male             12       10
> freereduced <-c(7,12)
>
> standard <-c(15,10)
> eat<-data.frame(freereduced,standard)
> chisq.test(eat,correct = FALSE)

        Pearson's Chi-squared test

data:  eat
X-squared = 2.3158, df = 1, p-value = 0.1281

>
>
> #critical value
> alpha3 <-0.1
> x1.alpha3 <- qchisq(alpha3,df=1,lower.tail = FALSE)
> x1.alpha3
[1] 2.705543
```

∴ Since the test statistic = 2.3158 < critical value = 2.705543. We fail to reject Ho, null hypot hesis . There is sufficient evidence to conclude that there is no association between gender and lunch variables

INTERPRETATION

In one sample hypothesis testing:

Hypothesis Statement

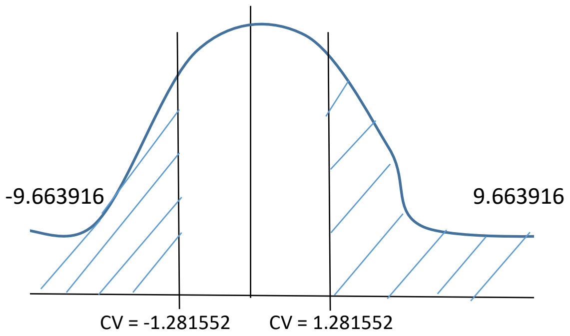
Ho: population mean for math score = 80

Hypothesis null :population mean for math score equal to 80

H1: population mean for math score ≠ 80

Alternative hypothesis mean for math score not equal to 80-9.663916 < critical value = 1.281552.

Execution Test



RSTUDIO

```
> #ONE SAMPLE HYPOTHESIS TESTING MEAN (math score)
> #population variance unknown
> mu=80    #null Ho
> #H1=mu >80
> alpha = 0.1
> z=(xbar-mu)/(stdDev/sqrt(n)) #test statistic
> z.alpha = qnorm(1-alpha)   #critical value
> z           #tesrtresult
[1] -9.663916
> z.alpha        #cv
[1] 1.281552
```

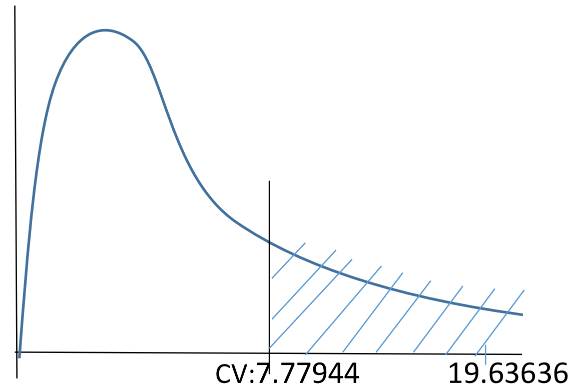
In one way contingency test :

Hypothesis Statement

Ho: Parent Level Education is indepedent

H1: Parent Level Education is dependent(related)

Execution Test



RSTUDIO

```
> #ONE SAMPLE HYPOTHESIS TESTING MEAN (math score)
> #population variance unknown
> mu=80    #null Ho
> #H1=mu >80
> alpha = 0.1
> z=(xbar-mu)/(stdDev/sqrt(n)) #test statistic
> z.alpha = qnorm(1-alpha)    #critical value
> z            #tesrtresult
[1] -9.663916
> z.alpha       #cv
[1] 1.281552
```

In two way contingency test

Hypothesis Statement

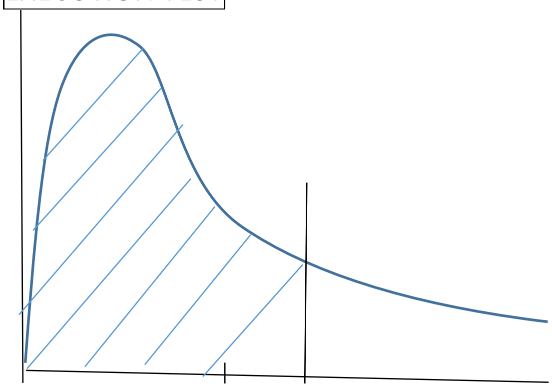
Ho: There is no association between gender and lunch variables

Hypothesis null :gender variable does not vary according to the lunch variables

H1: There is association exist between gender and lunch variables

Alternative hypothesis : gender variable vary according to the lunch variables

EXECUTION TEST




```
> #TWO WAY CONTINGENCY
> #bwtween gender and lunch
> table2 <- table (gender,lunch)
> table2
          lunch
gender    free/reduced standard
  female             7       15
  male              12       10
> freereduced <-c(7,12)
>
> standard <-c(15,10)
> eat<-data.frame(freereduced,standard)
> chisq.test(eat,correct = FALSE)

        Pearson's Chi-squared test

data:  eat
X-squared = 2.3158, df = 1, p-value = 0.1281

>
>
> #critical value
> alpha3 <-0.1
> x1.alpha3 <- qchisq(alpha3,df=1,lower.tail = FALSE)
> x1.alpha3
[1] 2.705543
```

DISCUSSION

For the first thing, we will talk about the one-sample hypothesis testing mean. For this test, I already set the null hypothesis testing and alternative hypothesis which will be used to dete rmine if we want to reject or fail to reject. The null hypothesis is the population mean for m ath score equal to 80 while the alternative hypothesis is the population mean is not equal to 80. As we can see I use 2 methods to show if we want to reject or reject the null hypothesis. Both of this method I got fail to reject null hypothesis. As a conclusion we can said that popu lation mean for math score is equal to 80.

Second, we will talk about correlation For this correlation test the null hypothesis is no linea r correlation exist between reading score and writing score while for the alternative hypothe sis is there is a linear correlation between reading score and writing score. As we can see I u se 2 method to determine which is the accurate hypothesis. Both of this method which are method 1:p-value and method2:t-test reject the null hypothesis. As a conclusion there is suf ficient evidence to support that there is linear correlation between reading score and writin g score. Based on the correlation scatter plot is positive linear association and has strong rel ationship. As we know correlation value that near to 1.0 means that the strength of correlati on coefficient relationship between the 2 variables is strong. So now I can conclude that the strength of relationship between writing score variable and reading score variable is strong s ince my correlation value is 0.9481469.

Next is about regression, the null hypothesis stated that there is no linear relationship betw een reading score and writing score while alternative hypothesis ,H1 stated that the linear r elationship exist. Since those two p-value < significance level = 0.1 , we reject null hypothesi s ,Ho. There is sufficient evidence to support that there is linear relationship between readin g score and writing score. Furthermore, there is sufficient evidence to support that reading s core affect writing score. The regression test that has been done, the coefficient of determin ation between reading score and writing score is 0.948 which tells us that its shows positive correlation between writing score and reading score. Based on my regression scatterplot I b uild regression line to indicate what is the regression coefficient. A positive coefficient indica tes that as the value of the independent variable increases, the mean of the dependent vari able also tends to increase. As a conclusion ,the regression line show that the reading score i ncreases as the writing score increases. As a conclusion, the reading score will increase if the writing score increase as well.

Next is about one way contingency test, the null hypothesis parent level education is indepe ndent while the alternative hypothesis is parent level education is dependent. Since the test statistic = 19.63636 > critical value = 7.77944. We will reject Ho, null hypothesis . There is suf ficient evidence to conclude that parent level education is dependent.

Lastly I will be discuss about two way contingency test, the null hypothesis there is no associ ation between gender and lunch variables while the alternative hypothesis is there is associ ation exist between gender and lunch variables. Since the test statistic = 2.3158 < critical val ue = 2.705543. We fail to reject Ho, null hypothesis . There is sufficient evidence to conclude that there is no association between gender and lunch variables. It means that that gender d oes not depend on lunch variable.

CONCLUSION

In conclusion, I have identify the main problem and reason of the students bad score in their examination. Sometimes lunch do affect the students examination score. Furthermore to gain good examination score student need to prepare for the examination not just do half of preparation but student must make full preparation for their upcoming examination. Overall, we can see that the man of population of math score is equal to 80. There is strong relationship between reading score and writing score . Other than that, we can also see that the two variables are dependent or related. Lastly, I learnt a lot about students perfromance because sometimes it was affected by lunch or test preparation or parent level education. I think this case study really useful to help me in the future.