UTM
UNIVERSITI TEKNOLOGI MALAYSIA

FACULTY OF ENGINEERING

SCHOOL OF COMPUTING

SEMESTER 1/20212022

PROJECT 2

(SECI2143:PROBABILITY AND STATISTICAL DATA ANALYSIS)

NAMA PENSYARAH:

Dr. Nor Azizah Ali

GROUP NO: 1

Name	Matric ID
ABDUL MUHAIMIN BIN ABDUL RAZAK	A21EC0002
HAFIZULSHAH BIN SHAROM	A21EC0027
IZZAT HAQEEMI BIN HAIRUDIN	A21EC0033
MUHAMMAD HAZIM BIN SALMAN	A21EC0078

Table of content

1) INTRODUCTION AND BACKGROUND

For this project, our main objective is to test out our practical skills in conducting inferential statistics, using all the method and analytical skills learned from SECI2143 PROBABILITY AND STATISTICAL DATA ANALYTICS. For this project, we have acquired a few datasets that will be analyzed by us. Our chosen dataset is

1. Global video gaming market value
2. Number of degree and diploma graduates
3. Titanic dataset

This dataset has taken our interest because gaming is becoming a top contender in the current business world, therefore, knowing its yearly revenue and annual income may help us in deciding our future career path and considering alternative paths if we are unable to follow what data engineering has in store for us. We expect to see some sort of high market value since the contender for this industries has become worldwide and the earnings from the gaming industry itself has reach billions of dollars in the recent year

For the second dataset, we are intrigued to know how many portion of university students that manage to graduate per year to see how we can evaluate our performance to not just be able to graduate, but also to be able to be a life-ready graduate with proper attributes to face the world. We hope to discover any relation with certain attribute or factor that allows a student to graduate, so that we may follow these examples.

Lastly, the titanic dataset is taken as a tribute to the disastrous shipwreck case which took many lives with it. We are curious to see how the survivor behavior within the ship allows them to live and we are considering all the possible factors as to why these survivors manage to stay alive , such as taking different class seats or other factors, to provide a countermeasure, if this case ever repeats itself.

2) DATASET

For hypothesis 2 sample, our chosen dataset is global video game market value from 2020 to 2025, by region in million U.S. dollars. Based on this dataset description, it stated that the North American gaming market will increase up to 80.9 billion U.S. dollars annually in 2025,starting from 56.8 billion U.S dollars in 2021. It is also expected that the region is going to stay strong in this area even though there is significant growth in the Asian region. This has piqued our interest in finding out whether there will be some significant difference in other regions compared to North America. For the test, we are planning to compare Far East and China to North America since the dataset shows us that the Far East and China region as another top competitors in the gaming market value against North America. We will be using these regions for global video game market mean value from 2020 to 2025, and we are expecting it to not have any or much significant difference.

For both correlation and regression test analysis, our chosen dataset is a test on the number of graduates for Degree students in Malaysia in between 2016 to 2020. In this dataset we want to see whether there is a relation between the number of graduates in line with the increase in each year. There are different numbers of graduates between male and female through the data. Furthermore, from the data we would see the number of graduates rising from 2174600 to 2828000 . Thus, the level of measurement for variables that we use in this dataset is interval.

Lastly for the optional test, the test that we conduct is Chi Square Test Of Independence In Two Way Contingency Table. The Titanic dataset is chosen to carry out this test. In this dataset, it has many variables such as passenger id, class of passenger's seat, surviving passengers etc. For the test, we used the variable class of passenger's seat and the survival of passengers to show whether the survivability of passengers depends on what class of seat they choose. We will be using all the passengers as our sample which are 418 passengers

No.	Variables	Answer	Level Of measurement
1.	Passenger seat's class & survival condition	Survivability rate (0=dead, 1=alive)	Nominal
2.	Education Level	Degree , Diploma	Nominal
3.	Market value for 6 year	Mean	Ratio

3) DATA ANALYSIS

a) 3 COMPULSORY TEST

Hypothesis 1 or 2 sample test

For hypothesis 2 sample testing, we conducted a test on the global video game market value from the year 2020 to 2025 in Million US dollar. In our test, we wish to determine whether there is any difference between the mean of the global game market value in 5 years for north america and far east and china at the 0.05 level of significance.

μ_1: mean of global game market value for Far east and China
μ_2 : mean of global game market value for far North America

$H_0 : \mu_1 == \mu_2$
$H_1 : \mu_1 =/= \mu_2$

Dataset :

Far east and China	North America
47,558.42	49,071.61
54,099.67	56,802.18
59,314.21	62,342.99
64,771.16	67,535.31
70,788.57	73,736.24
78,491.99	80,922.57
Mean ,$\bar{x}$ = 62,504.00 S_1 = 10,275.93	Mean,$\bar{x}$ = 65,058.98 S_2 = 11,540.73

Based on our calculation :

	value
T_0	$\frac{62504.00-65058.98-0}{\sqrt{\left(\frac{10275.93^{\ 2}}{6}+\frac{11540.73^{\ 2}}{6}\right)}}$ =-0.405
v	$\frac{\left(\frac{10275.93^{\ 2}}{6}+\frac{11540.73^{\ 2}}{6}\right)^{\ 2}}{\frac{\left(\frac{10273.93^{\ 2}}{6}\right)}{6-1}+\frac{\left(\frac{11540.73^{\ 2}}{6}\right)}{6-1}}$ $= 9.9 \approx 10$
$T_{0.025,10}$	-2.228

Criteria for rejection:
H_0 will be rejected if
$T_0 > T_{0.025,10} = 2.228$ OR $-T_0 < -T_{0.025,10} = -2.228$

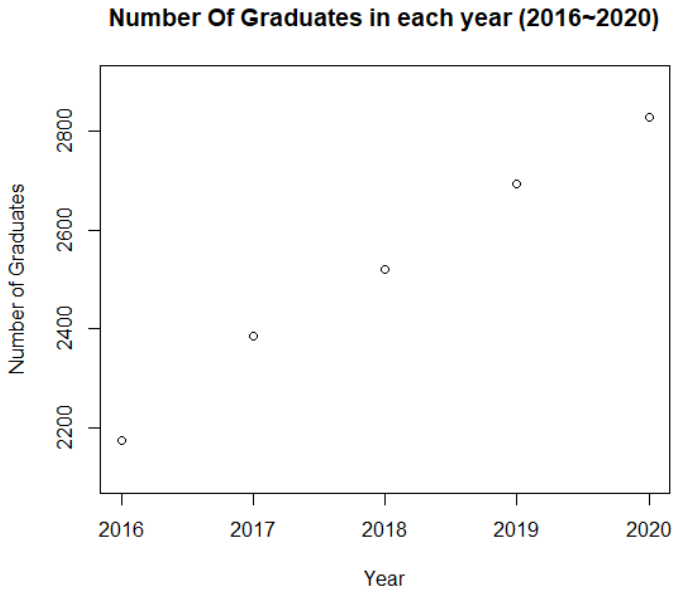
Conclusion

- Since $-0.405 > -T_{0.025,10} = -2.228$, fail to reject H_0
- That is, at 0.05 level of significance, we do not have strong evidence to conclude that the mean of global game market value for Far East and China differ from mean of global game market value for far North America.

Correlation test

For the Correlation Test , we conducted a test on the number of graduates for Degree students in Malaysia in between 2016 to 2020. In this analysis, we are using variables year and number of graduates for Degree students, where we will test whether there is a linear relationship between both variables using Pearson's Product-Moment Correlation Coefficient, at 95% confidence level. Correlation analysis is used to measure the strength of association(linear relationship) between two variables.

Year	Number of Graduates for Degree ('000)
2016	2174.6
2017	2384.6
2018	2520.5
2019	2693.4
2020	2828.0



* Noted : Number of Graduates ('000)

From the scatterplot above , it indicates that there is a positive correlation between the number of graduates for Degree's students and year , that is the higher the year, the higher the number of graduates for Degree's students. There are no outliers for this data.

1. Calculate the sample correlation coefficient using Pearson's method by

Formula to calculate the sample correlation coefficient, r:

$$r = \frac{\sum xy - \left(\sum x \sum y\right)/n}{\sqrt{\left[\left(\sum x^2\right) - \left(\sum x\right)^2/n\right]\left[\left(\sum y^2\right) - \left(\sum y\right)^2/n\right]}}$$

where:
- r = Sample correlation coefficient
- n = Sample size
- x = Value of the independent variable
- y = Value of the dependent variable

```
> cor(data$year,data$Degree)
[1] 0.9970448
```

Calculate r using R-studio

By using RStudio, we get a sample correlation coefficient, r = 0.9970448 , which indicates that there is a relatively strong positive linear correlation between the number of graduates for Degree's student and year.

2. Significance Test for Correlation

➢ Hypothesis Statement:

H0: ρ = 0 (no linear correlation)
H1: ρ ≠ 0 (linear correlation exists)

➢ Calculate test statistic by:

```
n <- 5
r <- cor(x,y)
t <- r/(sqrt((1-(r^2))/(n-2)))
```

```
[1] 22.47975
```

By using R-Studio , test statistic t = 22.47975

➢ Find critical value, using $\alpha = 0.05$, df = n-2 = 3 ;

From t-table, since this is a two-tailed test, there are two critical values:

Lower tail critical value $-t_{\alpha/2=0.025,\ df=3} = -3.182$

Upper tail critical value $t_{\alpha/2=0.025,\ df=3} = 3.182$

Hence, if test statistics > 3.182 / test statistics < -3.182, reject H0. Otherwise fail to reject H0.

➢ State the decision :

Since test statistics t = 22.4798 > upper tail critical value $t_{\alpha/2=0.025,\ df=3} = 3.182$, we reject the null hypothesis. There is sufficient evidence to conclude that there is a linear relationship between number of graduates for a Degree's student and year at $\alpha = 0.05$.

```
        Pearson's product-moment correlation

data:  data$year and data$Degree
t = 22.48, df = 3, p-value = 0.0001928
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9537781 0.9998149
sample estimates:
      cor
0.9970448
```

Regression test

For the Regression Test , we conducted a test on the number of graduates for Degree students in Malaysia in between 2016 to 2020. In this analysis, we are using variables year and number of graduates for Degree students, where we will test whether the linear relationship between both variables is positive or negative. Regression analysis is used to measure the average increment of number of graduates for Degree students per year in 5 years (2016 - 2020)

Year	Number of Graduates for Degree ('000)
2016	2174.6
2017	2384.6
2018	2520.5
2019	2693.4
2020	2828.0

Manual Calculation :

$\Sigma x = 10090$

$\Sigma y = 12601.1$

$\bar{x} = 2018$

$\bar{y} = 2520.22$

$\Sigma xy = 25430635.4$

$\Sigma x^2 = 20361630$

$n = 5$

$$b_1 = \frac{\Sigma xy - \frac{\Sigma x \Sigma y}{n}}{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}$$

$$\Sigma xy - \frac{\Sigma x \Sigma y}{n} = 25430635.4 - \frac{(10090)(12601.1)}{5}$$

$$= 1615.6$$

$$\Sigma x^2 - \frac{(\Sigma x)^2}{n} = 20361630 - \frac{(10090)^2}{5}$$

$$= 10$$

$$b_1 = 1615.6 / 10 = 161.56$$

$b_0 = \bar{y} - b_1 \bar{x}$

$= 2520.22 - (161.56)(2018)$

$= -323507.86$

Estimated Regression Model : $\hat{y} = -323507.86 + 161.56x$

Calculation Using R Programming :

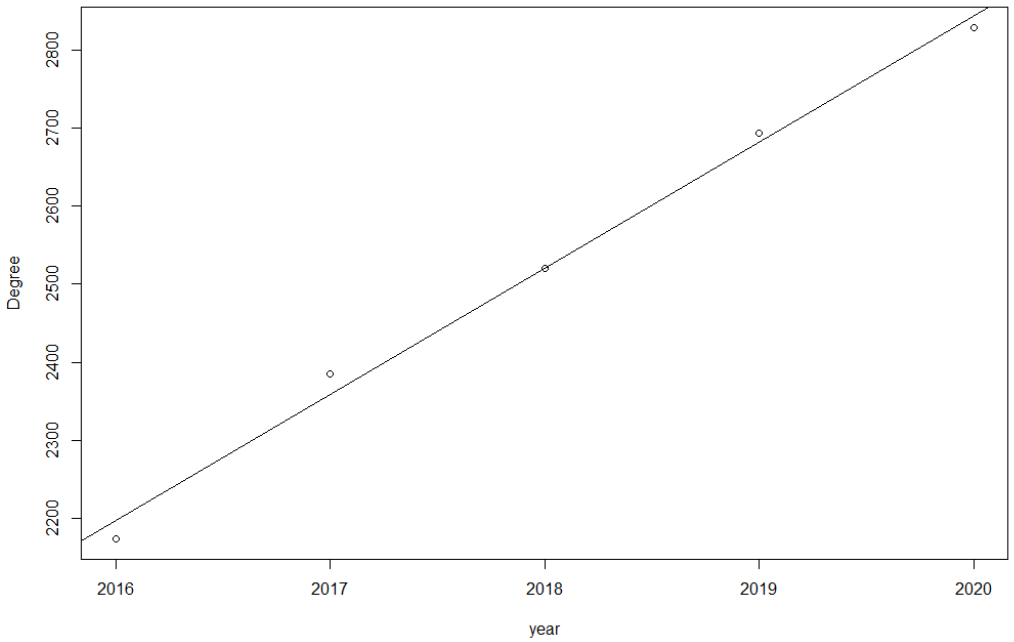
```
Console   Terminal   Jobs
R 4.1.3 · ~/
> # grouping data: x-year, y-degree
> x <- c(2016, 2017, 2018, 2019, 2020)
> y <- c(2174.6, 2384.6, 2520.5, 2693.4, 2828)
> # Create a linear regression model
> model <- lm(y~x)
> # Print regression model
> print(model)

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
  -323507.9        161.6

> # graph plotting
> with(data,plot(year,Degree))
> abline(lm(y~x))
```

Estimated Regression Model using R Programming : $\hat{y} = -323507.9 + 161.6x$

b_0 = is the estimated average value of number of graduate for Degree students when the value of year is zero (if year = 0 is in the range of observed year value)

b_1 interpretation :

Here, b_1 = 161.6 tells us that the average value of number of graduates for Degree students increased by 161.6 on average, for each additional year



This is a Positive Linear Relationship

Optional Tests

Chi Square Test Of Independence In Two Way Contingency Table

We conducted a Chi Square Test to test the relationship between passenger seat class affects the possibility of survival of the passenger at 5% level of significance ($\alpha = 0.05$) with random sample of 418 passengers.

Seat Class	If survived 1 , if dead 0	
	0	1
1	57	50
2	63	30
3	146	72

Hypothesis statement:

H_0: The passengers' seat class does not have any relationship with the possibility of survival of the passenger.
H_1: The passengers' seat class and possibility of survival of the passenger have a relationship between them.

Critical value:

$\alpha = 0.05$
$df = (3 - 1)(2-1) = 2$
Critical value = 5.991

Expected values:

Seat Class	If survived 1 , if dead 0				Total
	0		1		
	Observed	Expected	Observed	Expected	
1	57	69.09	50	38.91	107
2	63	59.18	30	33.82	93
3	146	138.72	72	79.27	218
Total	266	266	152	152	418

Expected Values Formula:

$$e_{ij} = \frac{(i^{th}\ \text{Row total})(j^{th}\ \text{Column total})}{\text{Total sample size}}$$

Test statistic values:

Cell, ij	Observed Value,O	Expected Value,E	$[\ Oij\ -\ Eij\ ]^2 / Eij$
1,1	57	69.09	2.12
1,2	50	38.91	3.16
2,1	63	59.18	0.25
2,2	30	33.82	0.43
3,1	146	138.72	0.38
3,2	72	79.27	0.67
		X^2	7.01

The decision:

Components	Values/Explanations
Test statistics	$X^2 = 7.01$
Critical Value	$X^2, k = 2, \alpha = 0.05 = 5.991$
Decision	Since, test statistic value (7.01) > critical value (5.991), thus reject hypothesis null , H_0 at $\alpha = 0.05$.
Conclusion	As a conclusion, there is evidence that the passengers' seat class does have relationship with the possibility of survival of the passenger

Initially a chi-square test is a statistical test used to compare observed results with expected results. The purpose of this test is to determine if a difference between observed data and expected data is due to chance or if it is due to a relationship between the variables we are taking into consideration. In this case, we are using the variables on whether the relationship between passenger seat class affects the possibility of survival of the passenger. By analyzing the data using a two-way contingency table, we can conclude that the possibility of the survival of the passengers depends on the passenger seat's class.

4) CONCLUSION

As a conclusion, firstly we would say that choosing the correct dataset was quite tough as there are many dataset available though the internet but not all of them we can make the hypothesis test. Next, after choosing the correct dataset we would consult with Dr.Nor Azizah Ali to double check our data during class sessions to make sure that the analysis that we are going to do is correct. As for the analysis process, we have divided each task between our group members and during the analysis we somehow faced some difficulties as we needed to choose the variables correctly but fortunately Dr.Nor Azizah Ali was there to help with this project. Our projected strong evidence is not supported by the data we retrieved and examined. Upon finishing our research, we realized that we had gained many new skills for studying data and doing analyses, including how to use R programming to make calculating the mean, standard deviation, and other statistics for a large set of data easier. We discovered that using practical methods to conduct analytical study was more simple. We developed a greater sense of curiosity and observance of our surroundings after examining the data that we are employing. Thanks to our analytical study, we can examine anything we choose and comprehend the true causes of problems. Thus, it broadens our perspective on nature

5) APPENDIX

Video gaming market size worldwide 2020-2025, by region								
Global video game market value from 2020 to 2025, by region (in million U.S. dollars)								
	North America	Latin America	West Europe	al & East Europe	ar East & China	an Subcontinent	st of Asia Pacific	a & Middle East
2020	49,017.61	7,819.72	32,415.35	8,886.54	47,558.42	1,948	5,833.14	2,411.34
2021*	56,802.18	8,464.12	38,618.05	9,570.73	54,099.67	2,031.68	6,208.13	2,574.27
2022*	62,342.99	8,953	45,347.34	9,934.14	59,314.21	2,075.26	6,448.05	2,694.75
2023*	67,532.31	9,228.89	53,784.44	10,209.52	64,771.16	2,151.42	6,591.51	2,790.33
2024*	73,736.24	9,195.46	64,446.85	10,416.60	70,788.57	2,275.16	6,612.49	2,840.58
2025*	80,922.57	9,260.03	77,441.75	10,618.45	78,491.99	2,491.05	6,676.98	2,902.38

Sex	Education le	Age Group	Ye	No. of graduates ('00
Male	Degree	less than or equal 24	2016	60.1
Male	Degree	25 - 34	2016	399.1
Male	Degree	35 - 44	2016	268
Male	Degree	greater than or equal 45	2016	297.3
Female	Degree	less than or equal 24	2016	90.2
Female	Degree	25 - 34	2016	576.7
Female	Degree	35 - 44	2016	300.4
Female	Degree	greater than or equal 45	2016	182.8
Male	Degree	less than or equal 24	2017	58.5
Male	Degree	25 - 34	2017	447.5
Male	Degree	35 - 44	2017	291.5
Male	Degree	greater than or equal 45	2017	325.2
Female	Degree	less than or equal 24	2017	98.3
Female	Degree	25 - 34	2017	640.4
Female	Degree	35 - 44	2017	319.1
Female	Degree	greater than or equal 45	2017	204.1
Male	Degree	less than or equal 24	2018	58.6
Male	Degree	25 - 34	2018	438.7
Male	Degree	35 - 44	2018	316.3
Male	Degree	greater than or equal 45	2018	346.5
Female	Degree	less than or equal 24	2018	114.1
Female	Degree	25 - 34	2018	669.9
Female	Degree	35 - 44	2018	357.6
Female	Degree	greater than or equal 45	2018	218.8
Male	Degree	less than or equal 24	2019	66.1
Male	Degree	25 - 34	2019	472.5
Male	Degree	35 - 44	2019	334.8
Male	Degree	greater than or equal 45	2019	361.5
Female	Degree	less than or equal 24	2019	123.5
Female	Degree	25 - 34	2019	677.5
Female	Degree	35 - 44	2019	402.1
Female	Degree	greater than or equal 45	2019	255.4
Male	Degree	less than or equal 24	2020	66.4
Male	Degree	25 - 34	2020	467
Male	Degree	35 - 44	2020	367.5
Male	Degree	greater than or equal 45	2020	394.6
Female	Degree	less than or equal 24	2020	100.6
Female	Degree	25 - 34	2020	697.9
Female	Degree	35 - 44	2020	429.8
Female	Degree	greater than or equal 45	2020	304.2