UTM UNIVERSITI TEKNOLOGI MALAYSIA

SCHOOL OF COMPUTING
Faculty of Engineering

Semester II 2021/2022

SECI 2143 - 01

(Probability & Statistical Data Analysis)

TEAM MEMBERS:

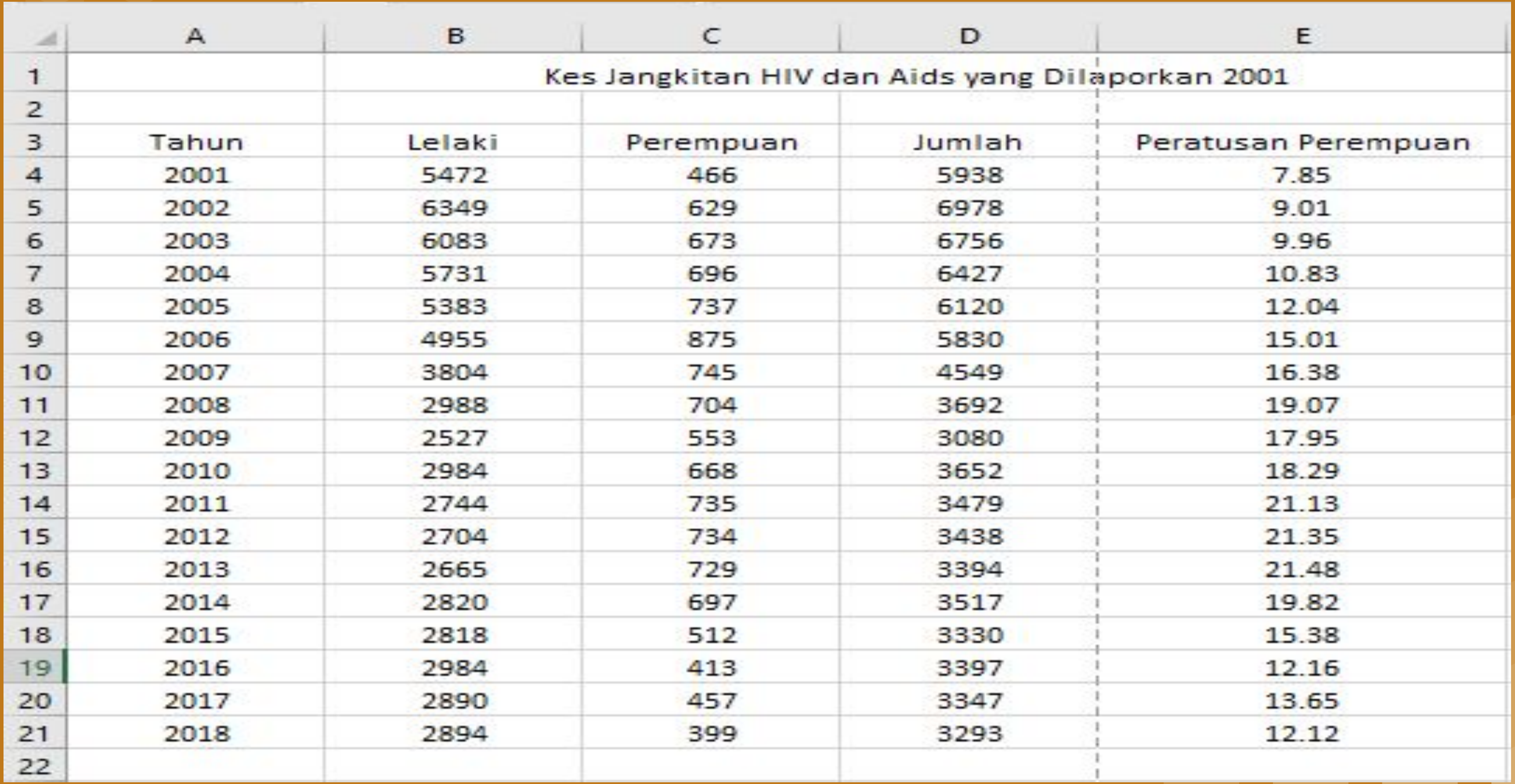
FARAH NABILAH BINTI NAJMUDIN (A21EC0023)
MAATHUREE A/P VEERABALAN (A21EC0051)
MUHAMMAD SAIFUDDIN BIN ISMAIL (A21EC0093)
NUR SYAFIKA BINTI MOHD SALMIZI (A21EC0115)

INTRODUCTION

The dataset chosen in this study is “Kes Jangkitan HIV dan Aids yang dilaporkan dari 2001 hingga 2018” was collected from a long term study of cases through 8 years. The main purpose in this project is to identify the increment and how widespread the reported cases of the virus are. Based on the dataset chosen , it included several records from each gender, male, female, total cases per year and percentage of infected females.

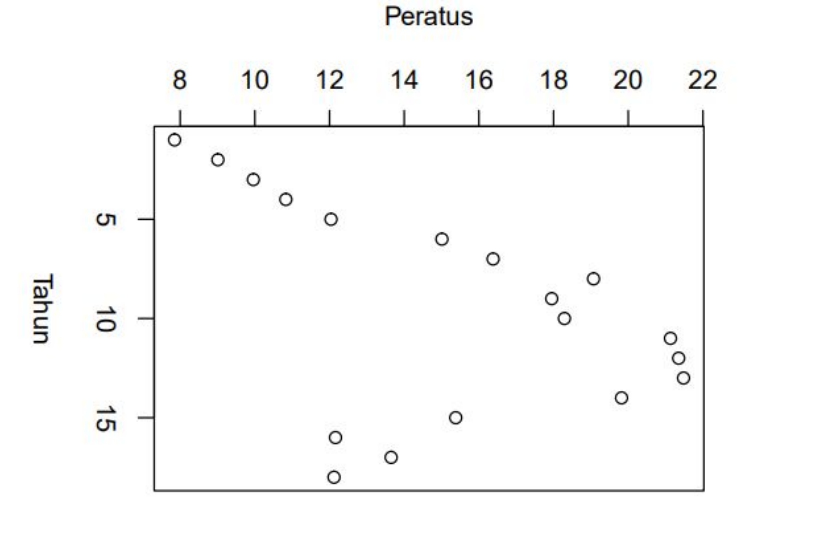
As we all know, HIV and Aids infection is a virus that seems to have become obsolete in our society. Moreover, it is a very dangerous virus as it is often infected by women . The first HIV/AIDS case in Malaysia was first detected in 1986 and since then, the virus has become one of the most serious health and development challenges in the country. As of the end of 2015, there are an estimated 93,000 people living with HIV in the country.

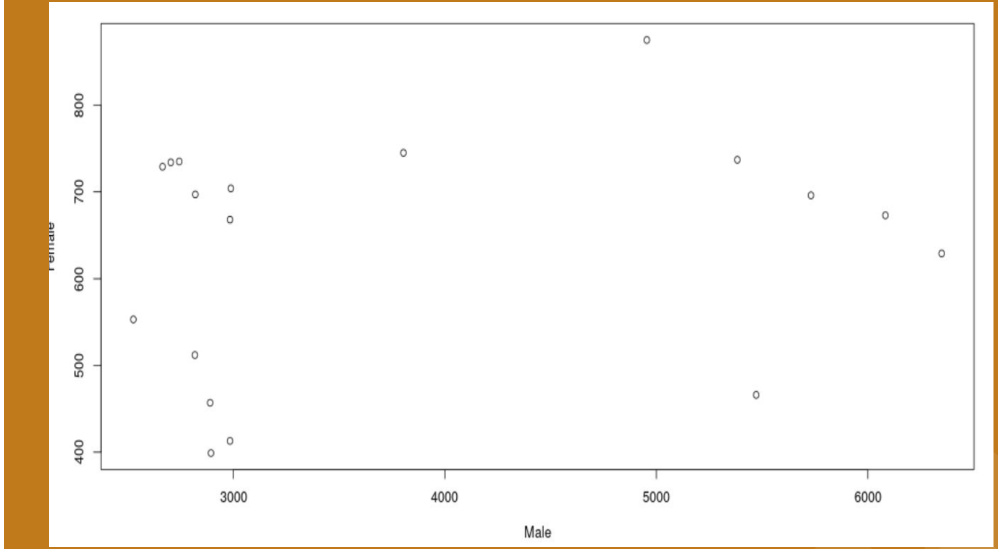
Data chosen



	A	B	C	D	E
1		Kes Jangkitan HIV dan Aids yang Dilaporkan 2001			
2					
3	Tahun	Lelaki	Perempuan	Jumlah	Peratusan Perempuan
4	2001	5472	466	5938	7.85
5	2002	6349	629	6978	9.01
6	2003	6083	673	6756	9.96
7	2004	5731	696	6427	10.83
8	2005	5383	737	6120	12.04
9	2006	4955	875	5830	15.01
10	2007	3804	745	4549	16.38
11	2008	2988	704	3692	19.07
12	2009	2527	553	3080	17.95
13	2010	2984	668	3652	18.29
14	2011	2744	735	3479	21.13
15	2012	2704	734	3438	21.35
16	2013	2665	729	3394	21.48
17	2014	2820	697	3517	19.82
18	2015	2818	512	3330	15.38
19	2016	2984	413	3397	12.16
20	2017	2890	457	3347	13.65
21	2018	2894	399	3293	12.12
22					

Hypothesis Testing - 1 Sample Test



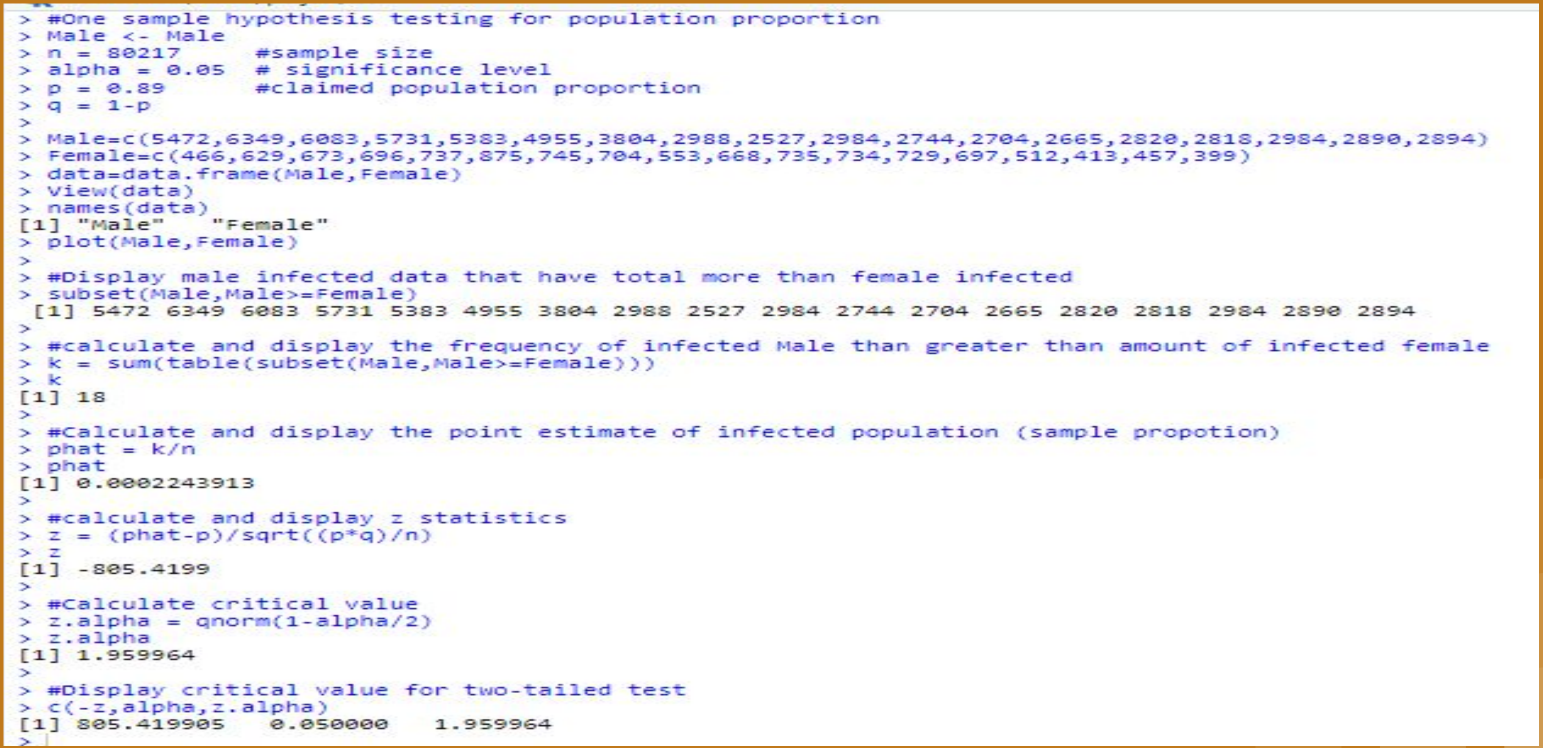


. A survey of a representative sample of 80217 cases among females and males reported that males surveyed had higher cases of infection in each year from 2001 to 2018. However, the result from opposite gender, females, has lower cases than expected. Figure 1 shows the percentage of infected females reported throughout the years.

H_0 : p = 0.89

H_1 : p != 0.89

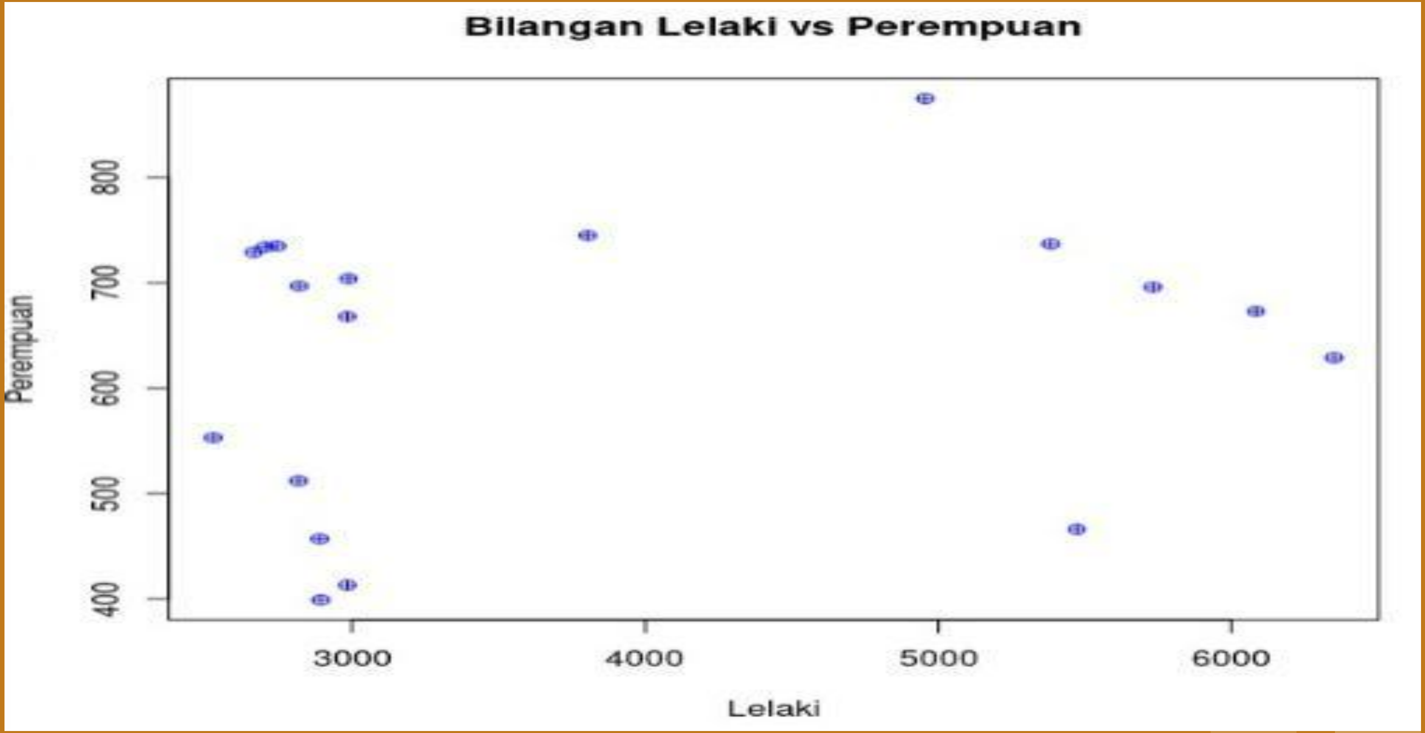
Hypothesis Testing - 1 Sample Test



```
> #One sample hypothesis testing for population proportion
> Male <- Male
> n = 80217      #sample size
> alpha = 0.05   # significance level
> p = 0.89       #claimed population proportion
> q = 1-p
>
> Male=c(5472,6349,6083,5731,5383,4955,3804,2988,2527,2984,2744,2704,2665,2820,2818,2984,2890,2894)
> Female=c(466,629,673,696,737,875,745,704,553,668,735,734,729,697,512,413,457,399)
> data=data.frame(Male,Female)
> View(data)
> names(data)
[1] "Male"   "Female"
> plot(Male,Female)
>
> #Display male infected data that have total more than female infected
> subset(Male,Male>=Female)
 [1] 5472 6349 6083 5731 5383 4955 3804 2988 2527 2984 2744 2704 2665 2820 2818 2984 2890 2894
>
> #calculate and display the frequency of infected Male than greater than amount of infected female
> k = sum(table(subset(Male,Male>=Female)))
> k
[1] 18
>
> #Calculate and display the point estimate of infected population (sample propotion)
> phat = k/n
> phat
[1] 0.0002243913
>
> #calculate and display z statistics
> z = (phat-p)/sqrt((p*q)/n)
> z
[1] -805.4199
>
> #Calculate critical value
> z.alpha = qnorm(1-alpha/2)
> z.alpha
[1] 1.959964
>
> #Display critical value for two-tailed test
> c(-z,alpha,z.alpha)
[1] 805.419905   0.050000   1.959964
>
```

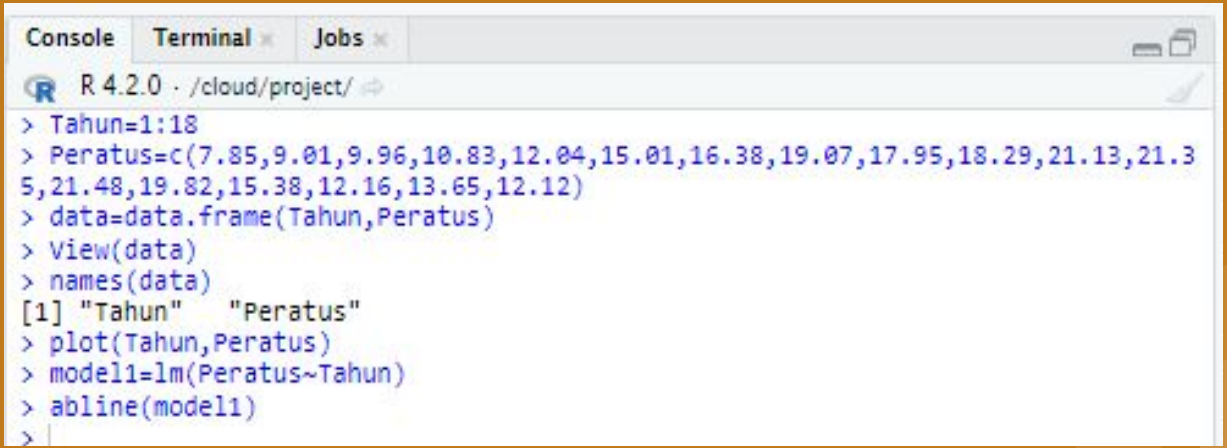
Correlation

The plot for correlation below shows that scatterplot has weak positive correlation . As we can see, the relationship between infected females does not depend on the infected male cases reported.



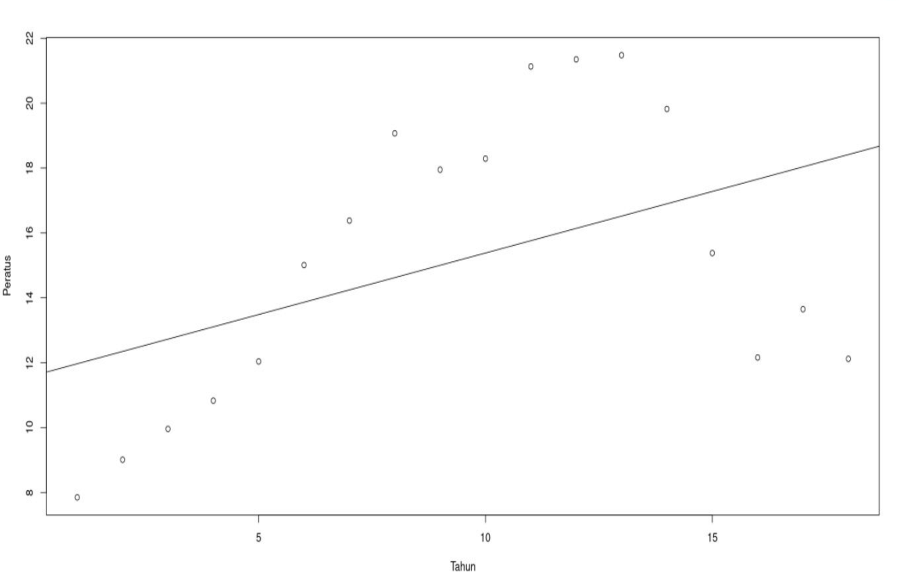
Regression

Regression analysis is a set of statistical methods used for the estimation of relationships between a dependent variable and one or more independent variables. It can be utilized to assess the strength of the relationship between variables and for modeling the future relationship between them. It is always risky to extrapolate the graph because the relationship between x and y may change or some kind of cut off point may exist.

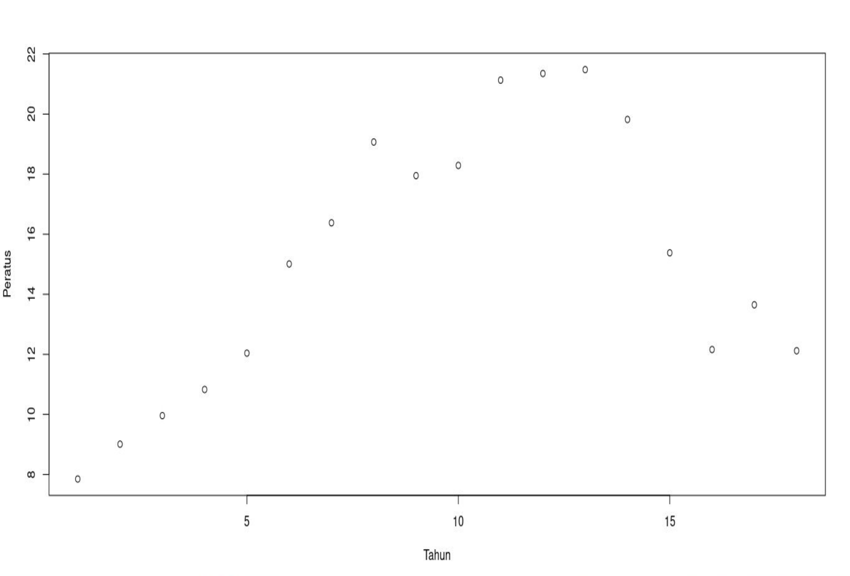


```
> Tahun=1:18
> Peratus=c(7.85,9.01,9.96,10.83,12.04,15.01,16.38,19.07,17.95,18.29,21.13,21.3
5,21.48,19.82,15.38,12.16,13.65,12.12)
> data=data.frame(Tahun,Peratus)
> View(data)
> names(data)
[1] "Tahun"    "Peratus"
> plot(Tahun,Peratus)
> model1=lm(Peratus~Tahun)
> abline(model1)
> 
```

Regression

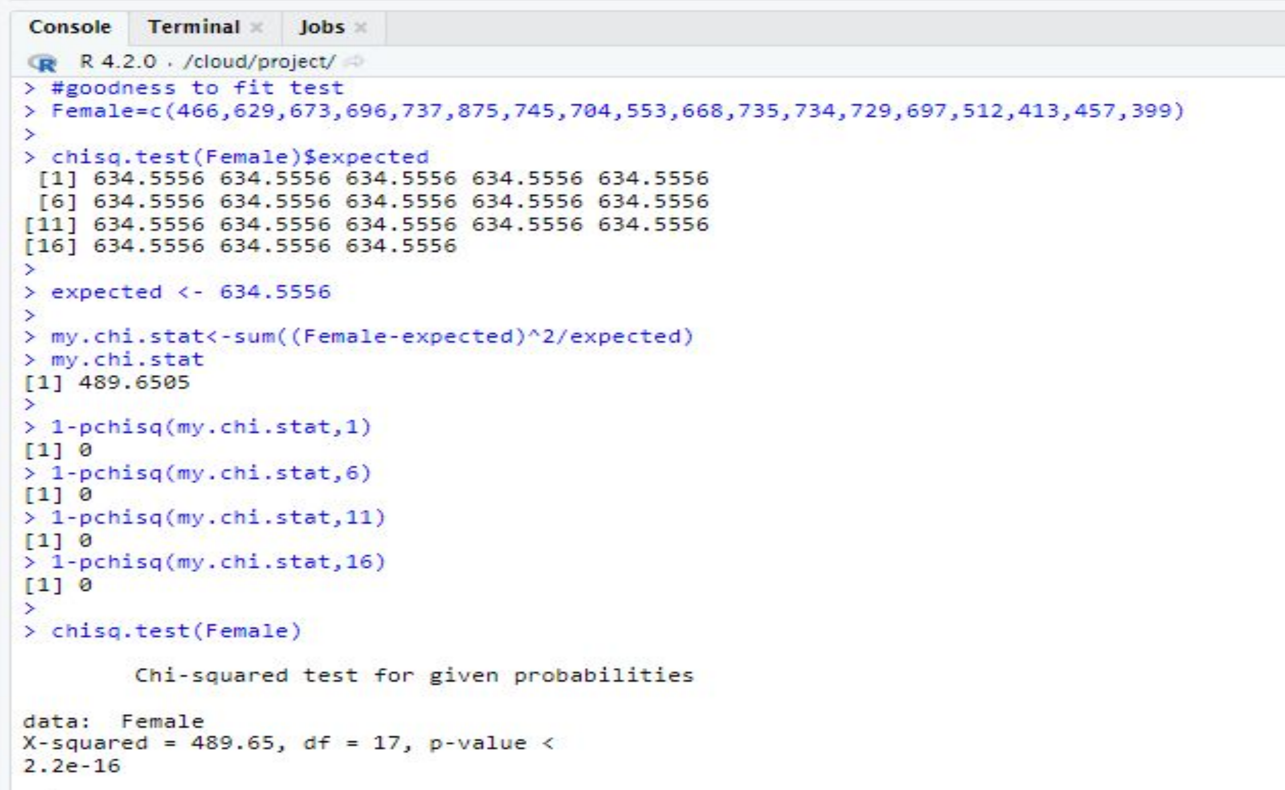


Peratus
8
10
12
14
16
18
20
22
5
10
15
Tahun



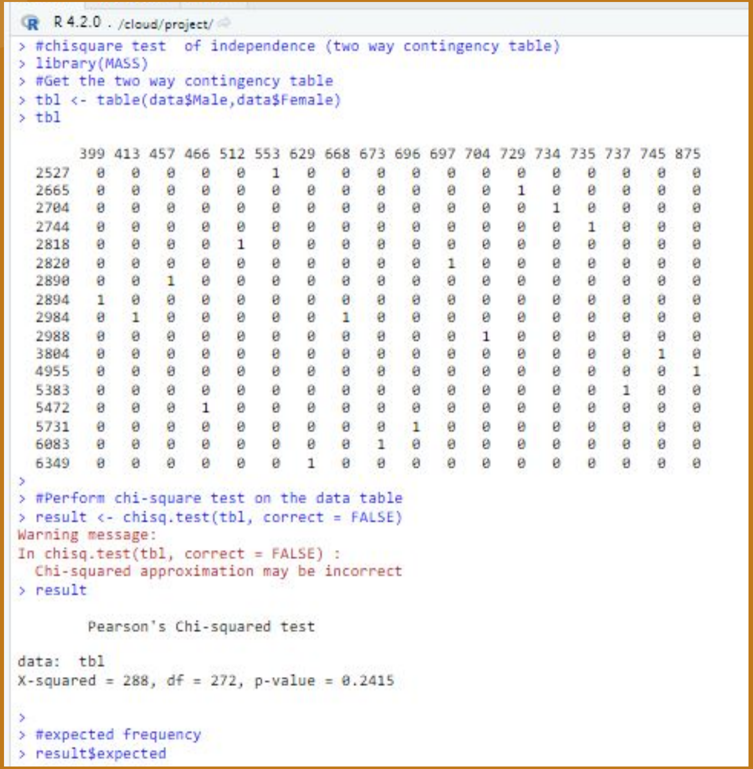
Peratus
8
10
12
14
16
18
20
22
5
10
15
Tahun

Goodness Fit Test



```
> #goodness to fit test
> Female=c(466,629,673,696,737,875,745,704,553,668,735,734,729,697,512,413,457,399)
>
> chisq.test(Female)$expected
 [1] 634.5556 634.5556 634.5556 634.5556 634.5556
 [6] 634.5556 634.5556 634.5556 634.5556 634.5556
[11] 634.5556 634.5556 634.5556 634.5556 634.5556
[16] 634.5556 634.5556 634.5556
>
> expected <- 634.5556
>
> my.chi.stat<-sum((Female-expected)^2/expected)
> my.chi.stat
[1] 489.6505
>
> 1-pchisq(my.chi.stat,1)
[1] 0
> 1-pchisq(my.chi.stat,6)
[1] 0
> 1-pchisq(my.chi.stat,11)
[1] 0
> 1-pchisq(my.chi.stat,16)
[1] 0
>
> chisq.test(Female)

	Chi-squared test for given probabilities

data:  Female
X-squared = 489.65, df = 17, p-value <
2.2e-16
```

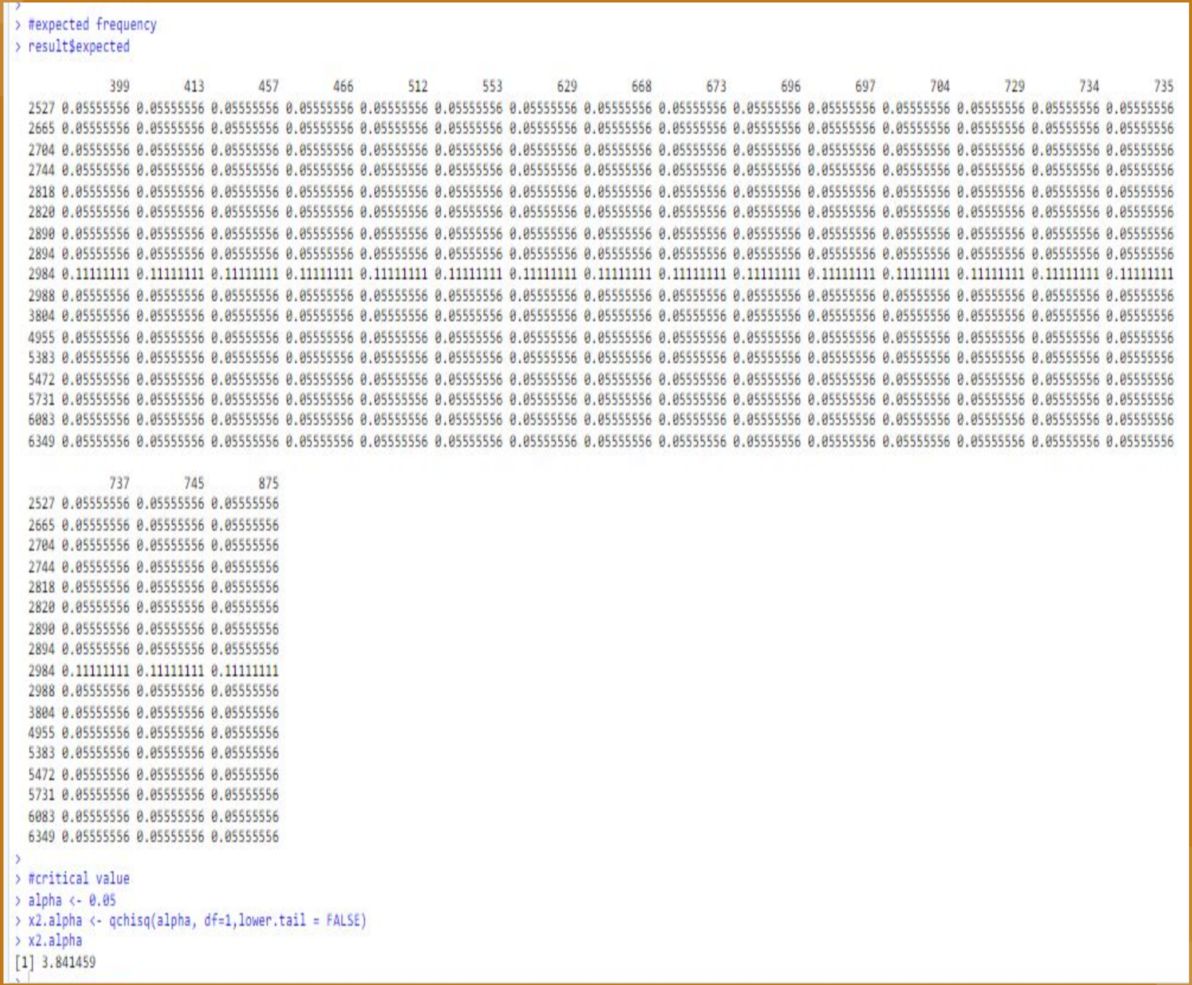
Chi Square test of Independence



```
R 4.2.0 · /cloud/project/
> #chisquare test  of independence (two way contingency table)
> library(MASS)
> #Get the two way contingency table
> tbl <- table(data$Male,data$Female)
> tbl

       399 413 457 466 512 553 629 668 673 696 697 704 729 734 735 737 745 875
  2527   0   0   0   0   0   1   0   0   0   0   0   0   0   0   0   0   0   0
  2665   0   0   0   0   0   0   0   0   0   0   0   0   1   0   0   0   0   0
  2704   0   0   0   0   0   0   0   0   0   0   0   0   0   1   0   0   0   0
  2744   0   0   0   0   0   0   0   0   0   0   0   0   0   0   1   0   0   0
  2818   0   0   0   0   1   0   0   0   0   0   0   0   0   0   0   0   0   0
  2820   0   0   0   0   0   0   0   0   0   0   1   0   0   0   0   0   0   0
  2890   0   0   1   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
  2894   1   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
  2984   0   1   0   0   0   0   0   1   0   0   0   0   0   0   0   0   0   0
  2988   0   0   0   0   0   0   0   0   0   0   0   1   0   0   0   0   0   0
  3804   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   1   0
  4955   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   1
  5383   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   1   0   0
  5472   0   0   0   1   0   0   0   0   0   0   0   0   0   0   0   0   0   0
  5731   0   0   0   0   0   0   0   0   0   1   0   0   0   0   0   0   0   0
  6083   0   0   0   0   0   0   0   0   1   0   0   0   0   0   0   0   0   0
  6349   0   0   0   0   0   0   1   0   0   0   0   0   0   0   0   0   0   0
>
> #Perform chi-square test on the data table
> result <- chisq.test(tbl, correct = FALSE)
Warning message:
In chisq.test(tbl, correct = FALSE) :
  Chi-squared approximation may be incorrect
> result

	Pearson's Chi-squared test

data:  tbl
X-squared = 288, df = 272, p-value = 0.2415

>
> #expected frequency
> result$expected
```

Chi-square tests are often used in hypothesis testing. The chi-square statistic compares the size of any discrepancies between the expected results and the actual results, given the size of the sample and the number of variables in the relationship.

For these tests, degrees of freedom are utilized to determine if a certain null hypothesis can be rejected based on the total number of variables and samples within the experiment. As with any statistic, the larger the sample size, the more reliable the results.



```
>
> #expected frequency
> result$expected

            399        413        457        466        512        553        629        668        673        696        697        704        729        734        735
2527 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556
2665 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556
2704 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556
2744 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556
2818 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556
2820 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556
2890 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556
2894 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556
2984 0.11111111 0.11111111 0.11111111 0.11111111 0.11111111 0.11111111 0.11111111 0.11111111 0.11111111 0.11111111 0.11111111 0.11111111 0.11111111 0.11111111 0.11111111
2988 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556
3804 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556
4955 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556
5383 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556
5472 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556
5731 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556
6083 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556
6349 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556 0.05555556

            737        745        875
2527 0.05555556 0.05555556 0.05555556
2665 0.05555556 0.05555556 0.05555556
2704 0.05555556 0.05555556 0.05555556
2744 0.05555556 0.05555556 0.05555556
2818 0.05555556 0.05555556 0.05555556
2820 0.05555556 0.05555556 0.05555556
2890 0.05555556 0.05555556 0.05555556
2894 0.05555556 0.05555556 0.05555556
2984 0.11111111 0.11111111 0.11111111
2988 0.05555556 0.05555556 0.05555556
3804 0.05555556 0.05555556 0.05555556
4955 0.05555556 0.05555556 0.05555556
5383 0.05555556 0.05555556 0.05555556
5472 0.05555556 0.05555556 0.05555556
5731 0.05555556 0.05555556 0.05555556
6083 0.05555556 0.05555556 0.05555556
6349 0.05555556 0.05555556 0.05555556
>
> #critical value
> alpha <- 0.05
> x2.alpha <- qchisq(alpha, df=1,lower.tail = FALSE)
> x2.alpha
[1] 3.841459
>
```

THAT’S ALL,THANK YOU EVERYONE!