SCSI2143: Probability & Statistical Data Analysis

2018/2019 – Semester 2

Report project 2

Due date – 12th May 2019

Group Members

1. DARISYAM BIN DARISMAN	A18CS0052
2. TAN ZHI QUAN	A18CS0260
3. JULIA BINTI JASMIN	B18CS0005
4. NUR SYAWANI BINTI HAMDAN	B18CS0022
5. MUHAMMAD FAYYADH BIN MOHD SALEHODDIN	A18CS0137

LECTURER

Assoc. Prof Dr Azlan bin Mohd Zain

Table of Contents

Introduction

In this project, after conducting the survey on Project 1, we want to observe the accuracy and consistency of data collected. These were done by using various methods. First is hypothesis testing with two samples. For this method, we tested the hypothesis of the average time for male gender spent on sports activities is higher than female. The next method is correlation analysis. This method is used to measure of the statistical relationship between two comparable variables or quantities(bivariate data). We also used regression analysis. From both of these method, we gathered and compared the relationship between cost spent on sport activities and time spent on sport in a session. Goodness of fit test is to test the hypothesis that the observed proportion claim that the number of respondent from different college are same. Chi Square Test were used to show if a relationship exists between the number of student spending their time in sport for a week claim that the average time spent on spot for the student is independent on the college they live. Last but least, we used ANOVA to test the significant difference between means. In this case, we compared 3 colleges which are KTDI, KTF and KTHO and relate them to the average time spent on sport activities.

Hypothesis 2 Samples

The average amount of time males spend playing sports each day is believed to be greater than female. An experiment is done; data is collected and resulting in the table below.

	Sample Size	**Standard Deviation**	**Average Time Spent Per Week (Mean)**
Male	34	4.969	8.2647
Female	26	3.213	4.6154

Conduct a hypothesis test at the 5% level of significant to test the claim.

Solution:

H_0: $\mu_1 - \mu_2 = 0$;8.8.26

$H_1: \mu_1 > \mu_2$. We want to reject H_0 if the average amount of time males spend playing sports each day is greater than female.

Given, $\alpha = 0.05$. The test statistic is

$$Z_0 = \frac{\bar{x}_1 - \bar{x}_2 - 0}{\sqrt{\frac{\delta^2_1}{n1} + \frac{\delta^2_2}{n2}}}$$

where $\delta^2_1 = (4.969)^2 = 24.70$ and $\delta^2_2 = (3.213)^2 = 10.32$, n1 = 34 and n2 = 26 ,

$Z_{0.05} = 1.645$

Reject H_0 if $Z_0 > Z_{0.05} = 1.645$

Computations: Since $\bar{x}_1 = 8.2647$ and $\bar{x}_2 = 4.6154$, we have

$$Z_0 = \frac{8.2647 - 4.6154 - 0}{\sqrt{\frac{24.70}{34} + \frac{10.32}{26}}}$$

$= 3.44$

Conclusion: Since $Z_0 = 3.44 > 1.645$, we reject H_0 at the level 0.05 level and conclude that there is sufficient evidence to support that average amount of time males spend playing sports each day is greater than female.

```
> xbar1 = 8.2647
> xbar2 = 4.6154
> n1 = 34
> n2 = 26
> sigma1 = 4.969
> sigma2 = 3.213
> variance1 = sigma1^2
> variance2 = sigma2^2
> z = ((xbar1-xbar2-0)/(sqrt((variance1/n1)+(variance2/n2))))
> alpha = 0.05
> z.apla = qnorm(alpha,lower.tail = FALSE)
```

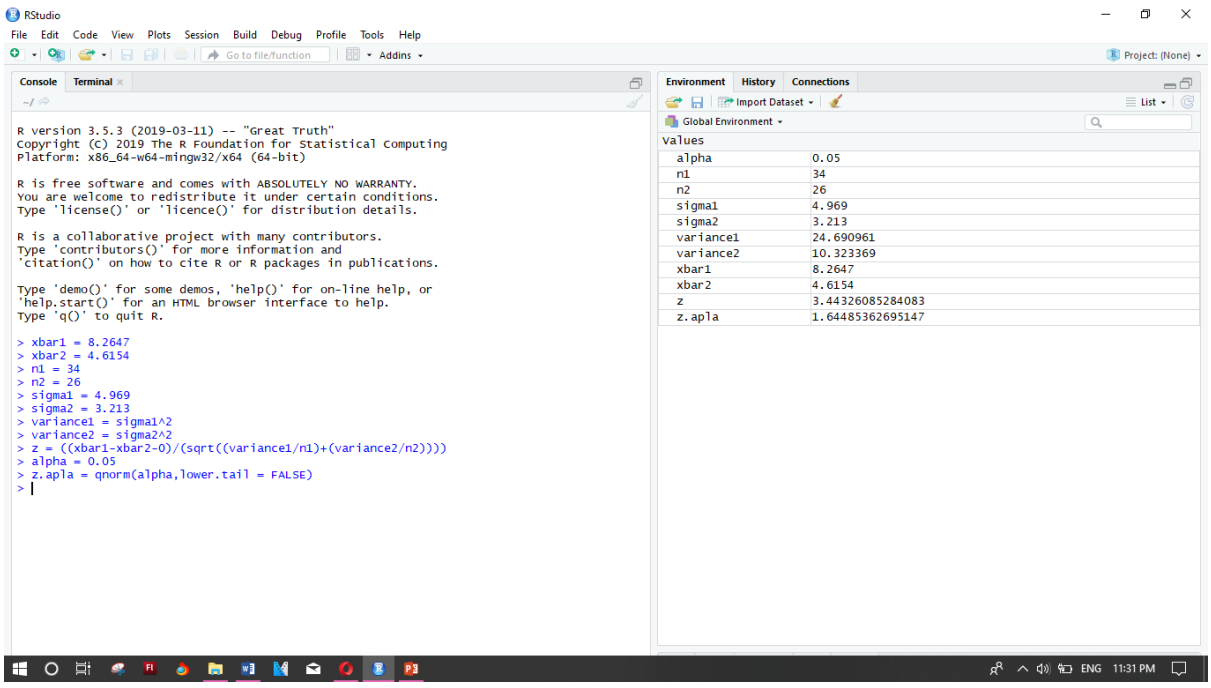
Environment | History | Connections

Import Dataset | List

Global Environment

Values

alpha	0.05
n1	34
n2	26
sigma1	4.969
sigma2	3.213
variance1	24.690961
variance2	10.323369
xbar1	8.2647
xbar2	4.6154
z	3.44326085284083
z.apla	1.64485362695147



Correlation

The data used to measure the relationship between variables:

x = Time spent on sport in a session (hour(s))

y = Cost spent on sport activities (RM)

x	y	x^2	y^2	x.y
2	20	4	400	40
1	50	1	2500	50
2	10	4	100	20
1	10	1	100	10
1	10	1	100	10
2	5	4	25	10
2	10	4	100	20
3	100	9	10000	300
3	10	9	100	30
2	10	4	100	20
3	20	9	400	40
1	10	1	100	10
1	0	1	0	0
2	5	4	25	10
2	5	4	25	10
2	10	4	100	20
2	10	4	100	20
3	10	9	100	30
1	20	1	400	20
2	20	4	400	40
2	10	4	100	20
2	0	4	0	0
2	10	4	100	20
2	20	4	400	40
2	50	4	2500	100
3	5	9	25	15

1	20	1	400	20
2	20	4	400	40
2	30	4	900	60
1	30	1	900	30
3	0	9	0	0
2	5	4	25	10
1	8	1	64	8
4	20	16	400	80
2	10	2	100	20
1	3	1	9	3
3	50	9	2500	150
2	0	4	0	0
5	4	25	16	20
4	50	16	2500	200
3	5	9	25	15
5	3	25	9	15
4	200	16	40000	800
3	0	9	0	0
1	15	1	225	15
2	2	2	4	4
3	10	9	100	30
3	2	9	4	6
3	5	9	25	15
1	50	1	2500	50
2	9	4	81	18
2	7	4	49	14
2	500	4	250000	1000
4	20	16	400	80
3	30	9	900	90
2	50	4	2500	100
2	100	4	10000	200
4	20	16	400	80

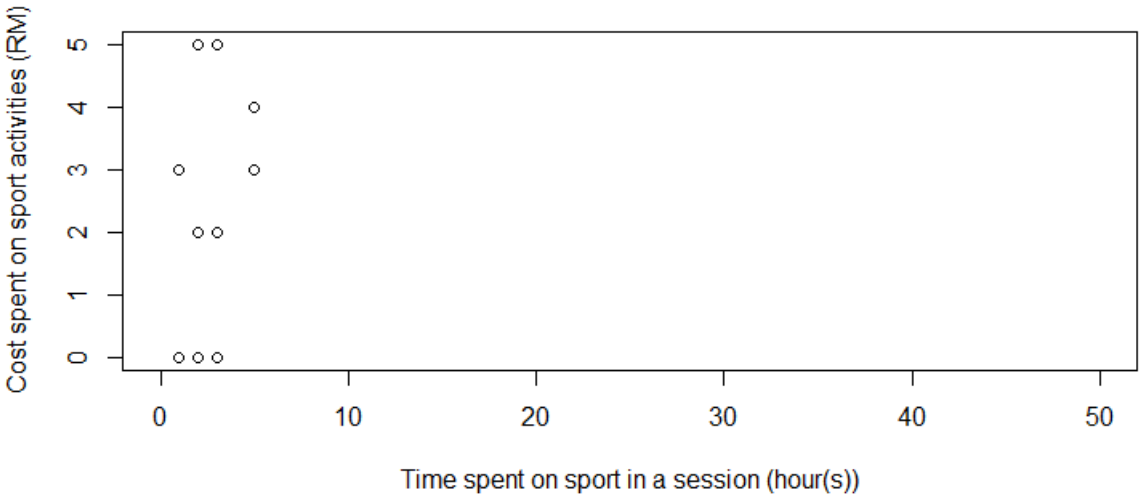
2	20	4	400	40
2	5	4	25	10
Σ = 137	Σ = 1743	Σ = 367	Σ = 334161	Σ = 4128

CODING CORRELATION:

```
> x<-c(2,1,2,1,1,2,2,3,3,2,2,1,1,2,2,2,2,3,1,2,2,2,2,2,2,3,1,2,2,1,3,2,1,4
,2,1,3,2,5,4,3,5,4,3,1,2,3,3,3,1,2,2,2,4,3,2,2,4,2,2)
> y<-c(20,50,10,10,10,5,10,100,10,10,20,10,0,5,5,10,10,10,20,20,10,0,10,20
,50,5,20,20,30,30,0,5,8,20,10,3,50,0,4,50,5,3,200,0,15,2,10,2,5,50,9,7,500
,20,30,50,100,20,20,5)
> cor(x,y)
[1] 0.0437979
> plot (x, y, xlim = c(0,50),ylim = c(0,5), xlab = "Time spent on sport in
 a session (hour(s))", ylab = "Cost spent on sport activities (RM)")
```

SCREENSHOT CODING:

```
> x<-c(2,1,2,1,1,2,2,3,3,2,2,1,1,2,2,2,2,3,1,2,2,2,2,2,2,3,1,2,2,1,3,2,1
,4,2,1,3,2,5,4,3,5,4,3,1,2,3,3,3,1,2,2,2,4,3,2,2,4,2,2)
> y<-c(20,50,10,10,10,5,10,100,10,10,20,10,0,5,5,10,10,10,20,20,10,0,10,
20,50,5,20,20,30,30,0,5,8,20,10,3,50,0,4,50,5,3,200,0,15,2,10,2,5,50,9,7
,500,20,30,50,100,20,20,5)
> cor(x,y)
[1] 0.0437979
> plot (x, y, xlim = c(0,50),ylim = c(0,5), xlab = "Time spent on sport
in a session (hour(s))", ylab = "Cost spent on sport activities (RM)")
```



From the table, we can substitute the data into the formula below:

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

Where,

n = 60

$\Sigma x = 137$

$\Sigma y = 1743$

$\Sigma x^2 = 367$

$\Sigma y^2 = 334161$

$\Sigma xy = 4128$

and we get:

$$r = \frac{4124 - (137)(1743)/60}{\sqrt{[(367) - (137)^2/60]\,[(334161) - (1743)^2/60]}}$$

$$= \frac{144.15}{15361484.73}$$

$$= 0.00000938$$

The conclusion we get from the calculation is that r = 0.00000938, it is relatively weak positive linear association between the cost spent on sport activities and the time spent on sport in a session.

Regression

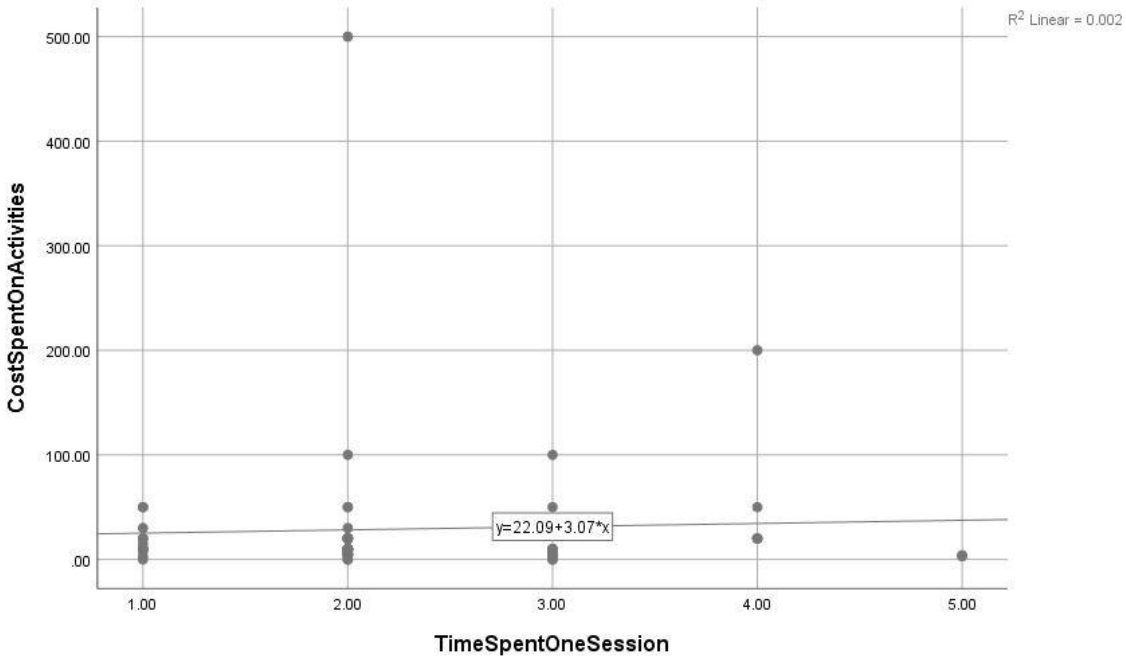


Figure above shows the cost spent on sport activities against the number of hours spent on one sport session.

In figure above, the simple regression analysis introduced. This regression model is a linear regression model. The number of hours spent on one sport session used as independent variable, x and the cost spent on sport activities used as dependent variable, y. Firstly, the sample regression line equation need to be obtained. The sample regression line equation, $\hat{y} = b_0 + b_1 x$, where

$\hat{y}$ = Estimated y value

b_0 = Estimate of the regression intercept

b_1 = Estimate of the regression slope x

x = Independent variable

The value of b_0 and b_1 can be calculated using the equation, $b_0 = \bar{y} - b_1\bar{x}$, and $b^1 = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}$. In our statistic analysis, we use SPSS to calculate the sample regression line equation and obtain the result as $y = 22.09 + 3.07x$. From the equation, we can conclude

that the variation of number of hours spent on one sport session, x will affect the cost spent on sport activities, y.

The value of 22.09 corresponding to b_0 represent the estimated average value of cost spent on sport activities is 22.09 when number of hours of spent on one sport session is 0. The value of 3.07 which corresponding to b_1 represent that the estimated change in the average value of cost spent on sport activities will be increased by 3.07 when the number of hours spent on one sport session is increased by 1.

From figure above, we can also calculate the value of coefficient of determination, R^2 which used to indicate the linear relationship between the cost spent on sport activities and the number of hours spent on one sport activity. The value R^2 calculated using SPSS showed a value of 0.002. Therefore, we able to prove that 0.2% of the variation in the cost spent on sport activities is explained by variation in the number of hours spent on one sport session by the students.

Goodness Of Fit Test

```
> obs<-c(16,5,8,4,9,5,4,3,3,3)
> chisq.test(obs,correct=F)

        Chi-squared test for given probabilities

data:  obs
X-squared = 25, df = 9, p-value = 0.002971

> output<-chisq.test(obs,correct=F)
> output$expected
 [1] 6 6 6 6 6 6 6 6 6 6
```

Goodness of fit test is used to test the hypothesis that the observed proportion claim that the number of responden from different colleges are same at a significant level of 0.05. The expected value will then be calculated by using data of observed proportion. The null hypothesis and the alternative hypothesis is as follow:

H_0 : The proportions for the number of responden from different colleges is the same

H_1 : At least one of the proportions for number of responden from different colleges are different from others.

Since expected value is the same, the value of expected proportion can be calculated by using the formula of mean, $Ei = \frac{n}{k}$ where k is sample size and n is number of category. The Ei is then calculated with n=60 and k=10, thus $Ei = 6$. The chi-square result is then calculated by using Rstudio with $Ei = 6.65$ by the test statistic, X^2 formula:

$$X^2 = \frac{(Oi-Ei)^2}{Ei}$$

From the result of Rstudio, the X^2 =25. The critical value then can be found in the Chi-Square Distribution table with degree of freedom=9, and significant level of 0.05. The critical value is found to be approximately equal to 16.919. Since the test statistics falls within the critical region, then we reject the null hypothesis. Hence, we have enough evidence to reject that The proportions for the number of responden from different colleges is the same.

Chi-Square

Chi-Square Test for k proportions

Chi-Square Test is used to test the hypothesis where the observed proportion for the number of student spending their time in sport for a week claim that the average time spent on spot for the student is independent on the College they live at a significance level of 0.05. The observed frequency is then compared with the expected values where we can calculate it by using observed data. The null hypothesis, H_0 and alternative hypothesis, H_a are as follow:

H_0 : The time spent in spot for the student is independent to the college they live.

H_a : The time spent in spot for the student is dependent to the college they live.

```
> obs<-matrix(c(15,4,7,3,8,3,2,3,3,3,1,1,1,1,1,2,2,0,0,0),byrow=F,ncol=2)
> colnames(obs)<-c("0<t<11","12<t<23")
> rownames(obs)<-c("KTDI","K9&K10","KTF","KTC","KTHO","KRP","KTR","KP","KD
J","KSDE")
> obs
       0<t<11 12<t<23
KTDI       15       1
K9&K10      4       1
KTF         7       1
KTC         3       1
KTHO        8       1
KRP         3       2
KTR         2       2
KP          3       0
KDOJ        3       0
KSDE        3       0
> chisq.test(obs,correct=F)

        Pearson's Chi-squared test

data:  obs
X-squared = 9.4009, df = 9, p-value = 0.4011
```

```
> output$expected
       0<t<11 12<t<23
KTDI    13.60    2.40
K9&K10   4.25    0.75
KTF      6.80    1.20
KTC      3.40    0.60
KTHO     7.65    1.35
KRP      4.25    0.75
KTR      3.40    0.60
KP       2.55    0.45
KDOJ     2.55    0.45
KSDE     2.55    0.45
```

Figure show the calculation of the Chi-Square test obtained by the RStudio.

From the calculation made by the RStudio, the value of Chi-Square is obtained which is equal to 9.4009. The Chi-Square value can also be double checked by adding all the multiple of each column and row number per total sample size. The critical value of this data can be obtained by using Chi-Square Distribution table, with degree of freedom (10-1)(2-1) and the significance level of 0.05(two-tailed). The critical value is then found to be approximately equal to 2.262. Since the Chi-Square value is greater than the critical value or fall within the critical region. Then, we can conclude that there is sufficient evidence to reject that the time spent in spot for the student is independent to the college they live.

ANOVA

```
> KTDI<-c(6,5,6,4,6)
> KTF<-c(2,1,10,6,6)
> KTHO<-c(4,8,3,3,6)
> Combined_Groups<-data.frame(cbind(KTDI,KTF,KTHO))
> Combined_Groups #shows spreadsheet like results
  KTDI KTF KTHO
1    6   2    4
2    5   1    8
3    6  10    3
4    4   6    3
5    6   6    6
> summary(Combined_Groups) # min,median,mean,max
      KTDI           KTF           KTHO
 Min.   :4.0   Min.   : 1   Min.   :3.0
 1st Qu.:5.0   1st Qu.: 2   1st Qu.:3.0
 Median :6.0   Median : 6   Median :4.0
 Mean   :5.4   Mean   : 5   Mean   :4.8
 3rd Qu.:6.0   3rd Qu.: 6   3rd Qu.:6.0
 Max.   :6.0   Max.   :10   Max.   :8.0
> Stacked_Groups<-stack(Combined_Groups)
> Stacked Groups # shows the table Stacked_Groups
Error: unexpected symbol in "Stacked Groups"
> Stacked_Groups # shows the table Stacked_Groups
   values  ind
1       6 KTDI
2       5 KTDI
3       6 KTDI
4       4 KTDI
5       6 KTDI
6       2  KTF
7       1  KTF
8      10  KTF
9       6  KTF
10      6  KTF
11      4 KTHO
12      8 KTHO
13      3 KTHO
14      3 KTHO
15      6 KTHO
> Anova_Results<-aov(values~ind,data=Stacked_Groups)
> summary(Anova_Results)#shows Anova_Results
            Df Sum Sq Mean Sq F value Pr(>F)
ind          2   0.93   0.467   0.076  0.928
Residuals   12  74.00   6.167
```

ANOVA method of testing is used to determine the equality of 3 or more population means for testing the significant differences between means by analyzing sample variances. By using One-way ANOVA with equal sample, we assume that the population have normal distribution, variance and the samples are random and independent of each other. Figure show the average time of students spent in the sport from different colleges. The dataset is distributed into 3 different colleges where is KTDI, KTF and KTHO. We use the 0.05 significant level to test the null hypothesis that the average spent on sport by different college student has the same mean. The null hypothesis and alternative hypothesis is as below:

$H_0 : \mu_1 = \mu_2 = \mu_3$

H_1 : at least one mean is different

Test statistic, F is calculated using Rstudio and the formula is as below:

$$F = \frac{variance\ between\ sample\ (nS\frac{2}{x})}{variance\ within\ sample\ (S_p^2)}$$

By using this formula, the variance between sample is calculated which is 0.467, while variance within sample is 6.167. Then test statistic, F=0.0757. The numerator, k-1=3-1=2, while the denominator, k(n-1)=3(5-1)=12, then F(2,12)=0.0757. The critical value of F with α =0.05 from F-distribution table is F (2,12) = 3.885. Since $F_{\text{test statistic}} < F_{\text{critical value}}$ (0.0757 < 3.885), the test statistic does not fall within the critical region, therefore we fail to reject the null hypothesis. There is sufficient evidence to claim that the average time spent on sport by different colleges have the same mean.

Conclusion

In conclusion, from this project, we learnt more about analysis of data. There are a lot of method we can use to obtain similar result. Learning all the methods is important for us not only during our studies in university but also in the future. Diving into software like SPSS, RStudio and many more, we now knew the functions they have for our future references. Mastering such software can benefit and prepare us in our future workplace.