Probability and Statistical Data Analysis Project 2
Section 10, Semester 2 2018/2019
School of Computing, Faculty Of Engineering, Universiti Teknologi Malaysia.

Survey On Weekend Activities Among Students

Ishaq sani ishaq
A18CS0353
Universiti Technology Malaysia (UTM),
School of Computing(SC),
81310 Skudai,
Johor Bahru,
Johor, Malaysia.

Steven Yong Wei
B18CS0025
Universiti Technology Malaysia (UTM),
School of Computing(SC),
81310 Skudai,
Johor Bahru,
Johor, Malaysia.

Aleef Hyqal bin Hilmeriza
A18CS0023
Universiti Technology Malaysia (UTM),
School of Computing(SC),
81310 Skudai,
Johor Bahru,
Johor, Malaysia.

Muhammad Syukran Syahiran
A18CS0162
Universiti Technology Malaysia (UTM),
School of Computing(SC),
81310 Skudai,
Johor Bahru,
Johor, Malaysia.

Muhammad Alif Solahuddin Bin Nordim
A18CS0124
Universiti Technology Malaysia (UTM),
School of Computing(SC),
81310 Skudai,
Johor Bahru,
Johor, Malaysia.

Abstract: Probability and statistical data analysis are the process of solving the problems in the real world (Data Analysis, Statistics, and Probability n.d.). Probability used to predict the certain event that likely to happen by using the degree of accuracy which is the sample mean, sample standard deviation, or sample proportion represent the corresponding population values (R. Lyman Ott, Longnecker, M 2001). As for the statistic, it is a set of scientific principles and techniques that are essential to conclude about the population and processes when that is limited information and data (Davis R.B., Mukamal K.J 2006). In this paper, we will conclude our findings that we done during the previous project with existed method such as hypothesis testing, correlation, linear regression, Godness of fit test, Chi square of independent and analysis of variance (ANOVA). The calculation for the mentioned method will conduct through 2 software that is Microsoft Excel and SPSS Statistic 24.

Keyword: statistical, degree of accuracy, linear regression, Chi square of independent, analysis of variance (ANOVA)

1. Introduction

Activity is a thing a person does or has done. Students all across University Teknologi Malaysia take part on weekend activities. Activities include studying, playing video games, sports, university club activities and shopping. This is a short paper on our second project about the survey on weekend activities among students in University Teknologi Malaysia. We were interested in doing a survey about the weekend activities to know how students spend their time during the weekend and how useful they spend their free time. We gave out questionnaires or google forms to students across school of computing.

A total of 60 students responded to the forms. This came with different opinions from each of the students. A lot of analysis was done due to the amount of variables in the questionnaire. This gave us a little problem because we have to include all the level of measurements among the variables which are nominal, ordinal, ratio and range are all included. Bar charts, pie charts, scatter plots, steam and leaf, box plot, histogram and frequency polygon were all used to analyse the data. More over different methods of hypothesis testing were used such as correlation, Linear regression, Chi Square Test Of Independent, Goodness of fit test, Analysis of variance (ANOVA).

2. Methodology

There are two software that involved in this paper. Microsoft Excel to determine and calculate the accuracy of the data and SPSS Statistic 24 used to statistical analysis and generate tables, graphs, and E.T.C for this paper. By using the data from the previous project, there are 60 respondent's data and it will imply for the rest of the paper. 6 existed statistical method chosen and that is hypothesis testing, correlation, linear regression, Godness of fit test, Chi square of independent and analysis of variance (ANOVA) to calculate and determine the relationship of the existed data. In addition, the paper will include the tables and plot that generated by the SPSS Statistic 24 at the below.

3. Hypothesis Testing

3.1 Hypothesis testing- 2 Sample

Hypothesis testing is the method or process to determine the strength of the evidence from the sample and provide a solid support for the decision making related to the population (Davis R.B., Mukamal K.J. 2006).

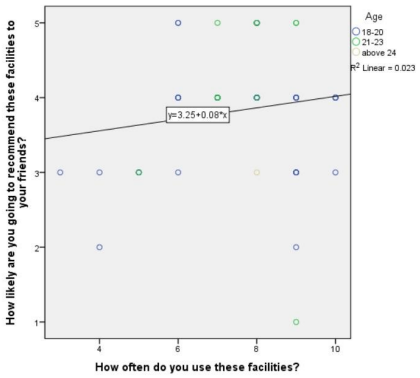
3.2 Correlation

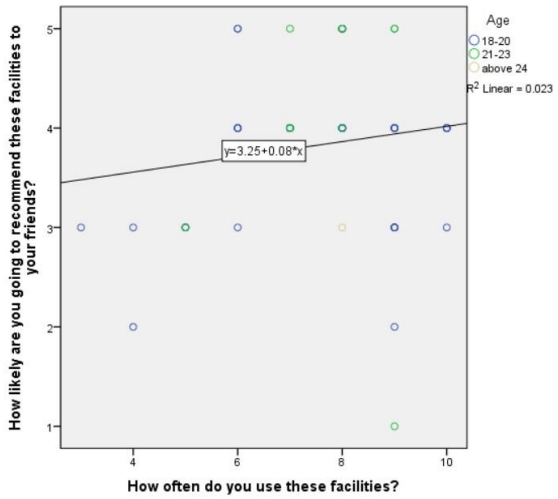
We make a research for linear relationship between how often student use facilities in their faculties and rate of facilities in their faculties with using a significance level of 0.05.

Correlation coefficient, r = 0.0034

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

r = correlation coefficient
N = number of respondent
∑xy = sum of the products of paired scores
∑x = sum of number often for use facilities in faculties scores
∑y = sum of rate of facilities in faculties scores
∑x^2= sum of squared number often for use facilities in faculties scores
∑y^2= sum of squared rate of facilities in faculties scores





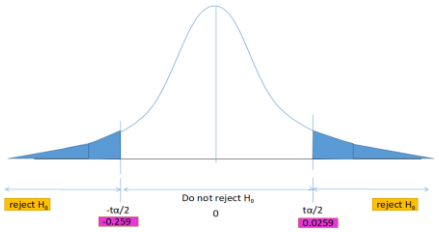
Hypothesis

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

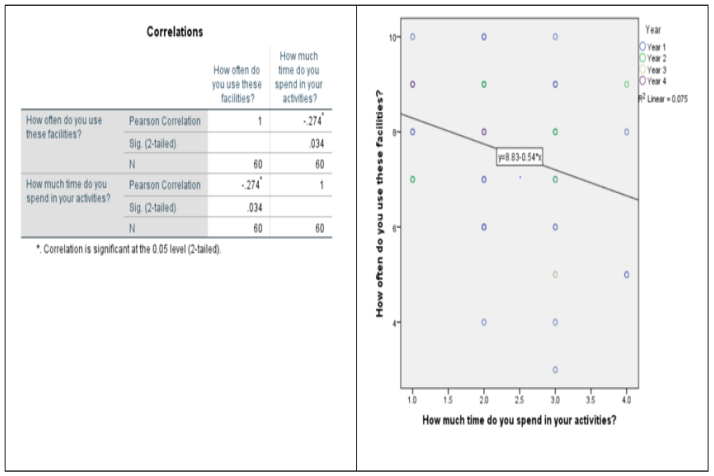
, df = 59 - 2 = 57

Test statistic:

$$t = \frac{r}{\sqrt{\frac{1-r2}{n-2}}} = \frac{0.0034}{\sqrt{\frac{1-(0.0034)2}{(60)-2}}} = 0.0259$$



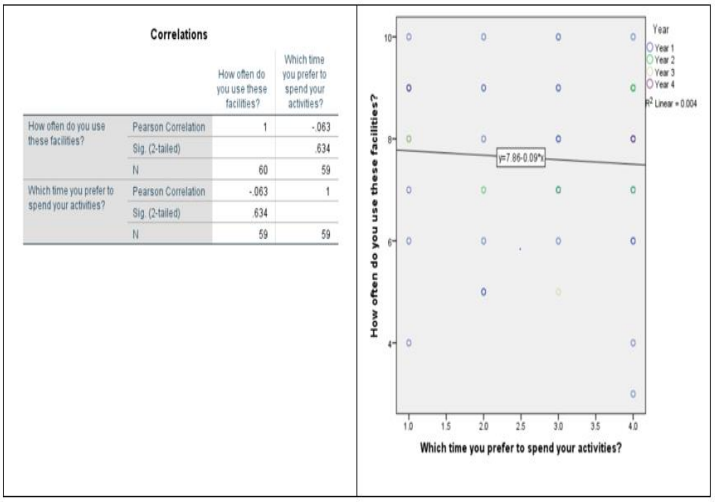
Conclusion: We reject H0. There is a sufficient evidence of a linear relationship between how often student use facilities in their faculties and rate of facilities in their faculties at 0.05 level of significance.



Correlations

		How often do you use these facilities?	How much time do you spend in your activities?
How often do you use these facilities?	Pearson Correlation	1	-.274*
	Sig. (2-tailed)		.034
	N	60	60
How much time do you spend in your activities?	Pearson Correlation	-.274*	1
	Sig. (2-tailed)	.034	
	N	60	60

*. Correlation is significant at the 0.05 level (2-tailed).



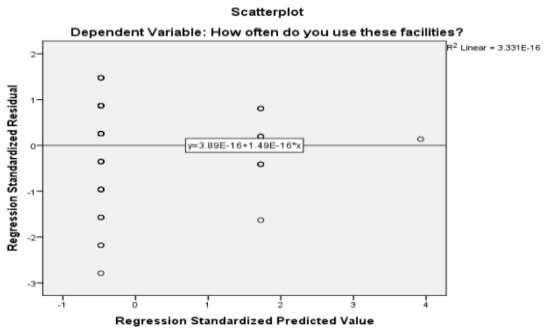
Correlations

		How often do you use these facilities?	Which time you prefer to spend your activities?
How often do you use these facilities?	Pearson Correlation	1	-.063
	Sig. (2-tailed)		.634
	N	60	59
Which time you prefer to spend your activities?	Pearson Correlation	-.063	1
	Sig. (2-tailed)	.634	
	N	59	59

3.3 Linear Regression

We find a research for linear relationship between how often student use facilities in their faculties and

rate of facilities in their faculties with using a significance level of 0.05.
Coefficient of determination, R2 = r2 = 0.0007



- Hypothesis

  H0: β1 = 0 (no linear relationship)

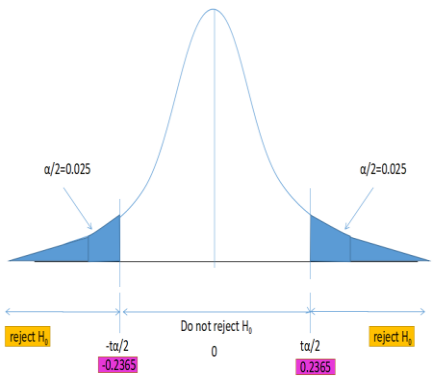
  H1: β1 ≠ 0 (linear relationship does exist)

- $\bar{y} = 1.628 + 0.821x$
- Test statistic

$$b1 = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}$$

$$t = \frac{b1 - \beta 1}{S_{\beta 1}} = \frac{0.3832 - 0}{1.62} = 0.2365$$

- df = n – 2 = 60 – 2 = 58, α = 0.05



Conclusion: We reject H0. There is a sufficient evidence that how often student use facilities in their faculties affect rate of facilities in their faculties. We find a linear relationship between how often student use facilities in their faculties and rate of facilities in their faculties at significance level of 0.05.

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	95.0% Confidence Interval for B Lower Bound	95.0% Confidence Interval for B Upper Bound
1	(Constant)	7.480	.610		12.259	.000	6.259	8.702
	Age	.098	.470	.027	.209	.835	-.843	1.040

a. Dependent Variable: How often do you use these facilities?

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	7.58	7.78	7.60	.045	60
Residual	-4.579	2.421	.000	1.628	60
Std. Predicted Value	-.477	3.924	.000	1.000	60
Std. Residual	-2.789	1.475	.000	.991	60

a. Dependent Variable: How often do you use these facilities?

3.4 Goodness of fit test

Goodness of fit test is a method to measure the compatibility of a set of random sample into the observation.(B.Aslan, G.Zech 2002)(Maydeu-Olivares, Garcia-Forero 2010)

3.5 Chi square of independent

How often do you use these facilities? * How much total time do you spend on these activities in the weekend? Crosstabulation

			How much total time do you spend on these activities in the weekend? Below 1 hour	1-2 Hour	2-3 Hour	Above Hour	Total
How often do you use these facilities?	3	Count	0	0	1	0	1
		Expected Count	.1	.5	.3	.1	1.0
		Residual	-.1	-.5	.7	-.1	
	4	Count	0	2	0	0	2
		Expected Count	.3	.9	.6	.2	2.0
		Residual	-.3	1.1	-.6	-.2	
	5	Count	0	1	1	2	4
		Expected Count	.5	1.9	1.3	.3	4.0
		Residual	-.5	-.9	-.3	1.7	
	6	Count	0	4	3	0	7
		Expected Count	.9	3.3	2.2	.6	7.0
		Residual	-.9	.7	.8	-.6	
	7	Count	2	5	3	0	10
		Expected Count	1.3	4.7	3.2	.8	10.0
		Residual	.7	.3	-.2	-.8	
	8	Count	2	8	5	1	16
		Expected Count	2.1	7.5	5.1	1.3	16.0
		Residual	-.1	.5	-.1	-.3	
	9	Count	3	5	6	1	15
		Expected Count	2.0	7.0	4.8	1.3	15.0
		Residual	1.0	-2.0	1.3	-.3	
	10	Count	1	3	0	1	5
		Expected Count	.7	2.3	1.6	.4	5.0
		Residual	.3	.7	-1.6	.6	
Total		Count	8	28	19	5	60
		Expected Count	8.0	28.0	19.0	5.0	60.0

Case Processing Summary

	Cases Valid N	Valid Percent	Missing N	Missing Percent	Total N	Total Percent
How often do you use these facilities? * How much total time do you spend on these activities in the weekend?	60	100.0%	0	0.0%	60	100.0%

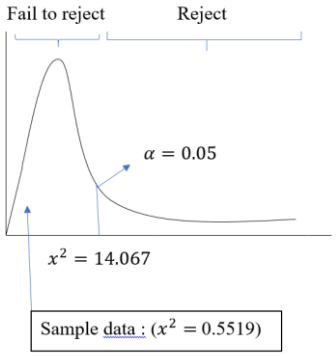
Frequencies of students using the facilities								
Number of times student use the facilities	3	4	5	6	7	8	9	10
Observed frequency	2	2	4	7	9	16	15	5

Claim : $p_3 = 1.7 \quad p_4 = 3.4 \quad p_5 = 6.8 \quad p_6 = 11.9$
$p_7 = 15.3 \quad p_8 = 27.1$
$p_9 = 25.4 \quad p_{10} = 8.5$
Statement of test hypothesis :
$H_0 : p_3 = 1.7 \quad p_4 = 3.4 \quad p_5 = 6.8 \quad p_6 = 11.9 \quad p_7 = 15.3 \quad p_8 = 27.1$
$p_9 = 25.4 \quad p_{10} = 8.5$
H_1 : At least one of the proportions is different from the claimed value
Expected frequency :
n = 60
3 times E = np = (60)(0.017) = 1.2
4 times E = np = (60)(0.034) = 2.04
5 times E = np = (60)(0.068) = 4.08
6 times E = np = (60)(0.119) = 7.14
7 times E = np = (60)(0.153) = 9.18
8 times E = np = (60)(0.271) = 16.26
9 times E = np = (60)(0.254) = 15.24
10 times E = np = (60)(0.085) = 5.1

Number of times student use the facilities	3	4	5	6	7	8	9	10
Observed frequency	2	2	4	7	9	16	15	5
Expeceted frequency	1.2	2.04	4.08	7.14	9.18	16.26	15.24	5.1
$\frac{(O-E)^2}{E}$	0.5333	0.0008	0.0016	0.0027	0.0035	0.0042	0.0038	0.0020

$x^2 = \sum \frac{(O-E)^2}{E} = 0.5519$
Critical Value $x^2 = 14.067$
(with k-1 = 8 – 1= 7 and α = 0.05)

Test statistic value ($x^2 = 0.5519$) < critical value ($x^2_{k=5,\alpha=0.05} = 14.067$), that is it does not fall within critical region. Thus, we do not reject H_0.
There is not sufficient evidence to warrant rejection of the claim that the number of times student use the facilities with the given percentages.



3.6 Analysis of variance (ANOVA)

Descriptives

How often do you use these facilities?

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
Morning	13	7.92	1.553	.431	6.98	8.86	4	10
Afternoon	9	7.11	1.965	.655	5.60	8.62	5	10
Evening	15	7.87	1.407	.363	7.09	8.65	5	10
Night	22	7.45	1.738	.371	6.68	8.23	3	10
Total	59	7.61	1.640	.214	7.18	8.04	3	10

ANOVA

How often do you use these facilities?

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	5.034	3	1.678	.611	.611
Within Groups	151.000	55	2.745		
Total	156.034	58			

Calculation :

H_0: μ1 = μ2 = μ3 = μ4

H_1: At least one mean is different

Morning	Afternoon	Evening	Night
$\bar{x} = 7.92$	$\bar{x} = 7.11$	$\bar{x} = 7.87$	$\bar{x} = 7.45$
S = 1.553	S = 1.965	S = 1.407	S = 1.738

Test statistic :

$\bar{x} = \frac{7.92+7.11+7.87+7.45}{4} = 7.5875$

$S_{\bar{x}} = \sqrt{\frac{(7.92-7.5875)^2+(7.11-7.5875)^2+(7.87-7.5875)^2+(7.45-7.5875)^2}{4-1}} = 0.382$

$ns_x^2 = 59(0.382)^2 = 8.62$

$s_p^2 = \frac{(1.553)^2 + (1.965)^2 + (1.407)^2 + (1.738)^2}{4} = 2.81$

$F = \frac{variance\ between\ samples}{variance\ within\ samples} = \frac{ns_{\bar{x}}^2}{s_p^2} = \frac{8.92}{2.81} = 3.17$

Numerator = k – 1 = 4 – 1 = 3

Denominator = k(n-1) = 4(59 – 1) = 232

F critical value = 2.605

Since $F_{test\ statistic} > F_{critical\ value}$ $(3.17 > 2.605)$, we reject the null hypothesis. There are sufficient evidence to claim that different time have different mean for student's prefered time to do their activities.

4. Discussion and Conclusion

The observation of the survey on weekend activities among students in University Technology Malaysia had been conducted. In the survey, 60 students had responded. The main reason for this survey is to know how the students spend their time during the weekend. Furthermore, SPSS is used to generate the graph for this observation.

We used hypothesis testing (one sample tests) which is test on mean and unknown variance. The z value that we get is lower than critical z value which is -5.85 < 0.499. We can conclude that we fail to reject H0.

For ANOVA, the number of student use the facilities has the highest value of mean while the mean for total time spent and time spent of activity is almost the same. We also can conclude that the total time spent and time spent of activity has the same value for standard variation. We also conducted chi square test on number of times student use the facilities. We calculated the expected frequency for each frequency. Once we calculate the expected frequency, we can get the test statistic which is 0.5519 and the critical value is 14.067. Therefore, we fail to reject H0 because test statistic is lower than critical value.

For correlation test, we make a observation between how often student use facilities in their faculties and the rate of faculties. The value for correlation coefficient is r = 0.0034. So, we reject H0 and the linear correlation is exist. Lastly, for regression test, the value for coefficient of determination is r^2 = 0.007. Thus, we reject H0 so that linear relationship does exist.

5. Reference

[1] Testpreptoolkit.com. (2019). "Data Analysis, Statistics, and Probability". Available at: https://www.testpreptoolkit.com/data-analysis-statistics-probability [Accessed 9 May 2019].

[2] R. Lyman Ott, Longnecker, M., "An Introduction to Statistical Methods and Data Analysis". Australia, Duxbury, 2001.

[3] Davis R.B., Mukamal K.J.. Hypothesis testing: means. Circulation, 114 ,2006, 1078-1082

[4] Statisticssolutions.com. (2019). Correlation (Pearson, Kendall, Spearman). [online] Available at: https://www.statisticssolutions.com/wp-content/uploads/wp-post-to-pdf-enhanced-cache/1/correlation-pearson-kendall-spearman.pdf [Accessed 9 May 2019].

[5] Kumari K, Yadav S. Linear Regression Analysis Study. J Pract Cardiovasc Sci,2018, 4, 33-6

[6] B.Aslan, G.Zech . Comparison Of Different Goodness-of-Fit Tests. In Advanced statistical techniques in particle physics. Proceedings, Conference, March 18–22, 2002, 166–175

[7] Maydeu-Olivares, Alberto, Forero, Carlos. Goodness-of-Fit Testing.International Encyclopedia of Education, 7,2010, 190-196.

[8] Statistics Solutions. (2019). Chi-Square Test of Independence - Statistics Solutions. [online] Available at: https://www.statisticssolutions.com/non-parametric-analysis-chi-square/ [Accessed 10 May 2019].

[9] F. Sawyer Steven. Analysis of variance: The Fundamental Concepts. The Journal of Manual and Manipulative Therapy. 17 (2) 27E–38E,2009.