International Journal of Engineering & Technology, 7 (4.33) (2018) 60-64

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Investigating Students' Difficulties in Understanding Confidence Intervals in Linear Regression Models

Nor Rashidah Paujah @ Ismail, Fadzilah Abdol Razak*, Norhayati Baharun, Elizabeth Selvamani G. Arul

Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Perak Branch, Tapah Campus, Tapah Road, 35400 Perak, Malaysia
**Corresponding author E-mail: fadzilah.ar@perak.uitm.edu.my*

Abstract

How do tertiary students perform when finding confidence intervals of linear regression models? Do they have strong understanding on how to compute the interval and provide good explanation on the interval obtained? To answer these questions, 197 answer scripts were examined to investigate students' ability to calculate the confidence interval of the regression slope and their ability to make comprehensive interpretation afterwards. It was found that only 48% of the students managed to compute the confidence interval correctly. The errors made by most of the students were caused by the failure to identify the correct degrees of freedom and the failure to evaluate the correct value of the standard error of the slope. Of those who were able to compute the correct values, the percentage that were able to give complete and correct interpretation dropped to only 7.1%. 68.5% of them provided incorrect interpretations which showed their inability to understand the concept of regression slope. It is hoped that this study will give some ideas to educators in providing better understanding on computing and interpreting the confidence interval among students.

Keywords: *confidence interval; regression slope; regression model; students' difficulties.*

1. Introduction

Confidence interval is one of the important topics in statistics courses. When data is not available for the entire population, the sample statistic (point estimate) calculated from the sample data is used to estimate the values of population parameters. A confidence interval is constructed around this value of point estimate and this can be easily calculated using respective formulae. The point estimate gives a best estimate of the true population value based on the sample, but it provides no indication of how precise the estimate is [1]. Hence, confidence intervals are used to estimate the values of population parameters and offer more information than the point estimates since they provide a range of values at each side of the observed sample statistics. In [2] states that under equal circumstances, a larger sample will give a better estimation of the parameter; but also, a wider interval is determined by a higher confidence level and therefore it has more chance of success, while a smaller interval with the same confidence level offers a more accurate estimation, but with less chance of success.

1.1. Confidence Interval in Regression Analysis

Regression analysis is a statistical methodology that utilizes the relation between two or more quantitative variables so that a response (dependent) variable can be predicted from one or more independent variables. The relationship between a dependent variable Y and an independent variable X is expressed as a linear model:

$$Y = \beta_0 + \beta_1 X + \varepsilon \quad (1)$$

The parameters β_0 and β_1 in the regression model are called regression coefficients. β_1 is the slope of the regression line which can be interpreted as the change in the mean of Y produced by a unit change in X. Usually, the parameters β_0 and β_1 are unknown and must be estimated using sample data. The method of least squares is used to obtain the estimated value of parameters β_0 and β_1. The point estimate gives the best estimate of the true population value based on the sample. In addition to point estimates of β_0 and β_1, we may also obtain confidence interval estimates of these parameters using the following formula:

$$b_i \pm t_{\frac{\alpha}{2}, n-2} s(b_i) \quad (2)$$

where b_i are the regression coefficients, $s(b_i)$ are the standard deviation of the estimated regression coefficients and t is a value obtained from a standard t-table at a certain value of α and degrees of freedom [3].

It has been observed in students' work that students tend to focus more on the calculation part and ignore the writing part. Most students fail to write correct and meaningful statistical statements where they need to make conclusions and interpretations based on the value calculated. Certain topics require students to have a deep understanding of a preceding topic in order to interpret the calculated value. For example, in order to interpret the meaning of confidence interval of the regression slope, students need to understand the meaning of the regression slope first. Again, if students focus only to the calculation of the regression slope and fail to explain the meaning of the value obtained, they will carry forward their ignorance when it comes to the topic on confidence interval of the regression slope.

Copyright © 2018 Authors. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Therefore, this study aims to investigate students' understanding on the concept of confidence interval of the regression slope. Specifically, the objectives were set as follows:

i. To investigate students' ability to calculate the value of confidence interval of the regression slope.
ii. To investigate students' ability to write the correct interpretation of confidence interval of the regression slope.
iii. To investigate the association between students' ability to interpret the regression slope and the ability to interpret the confidence interval of the regression slope.
iv. To compare the ability to interpret the confidence interval of the regression slope between students enrolled in Diploma in Actuarial Science and Diploma in Statistics programs.

2. Literature Review

2.1. Introduction

Confidence interval is one of the topic in statistical inferences and is very useful to estimate population parameters from a sample of values of the population [4]. According to [5], confidence interval provides an interval describing a plausible range of values of population parameters from which a random sample might produce the observed sample statistics at a given level of confidence (usually 95%). 95 % confidence level means that, if sampling is repeated a sufficient number of times, 95% of intervals generated, will capture the population parameter. Confidence intervals provide additional information on the variability of an observed sample statistic (i.e. its precision) and on its probable relationship to the value of this statistic in the population from which the sample was drawn (i.e. its accuracy) [6]. Other than estimation, confidence interval is widely used in deciding whether or not to reject the null hypothesis as an alternative to the conventional hypothesis testing. According to [7], teaching inference via confidence intervals instead of null hypothesis significance testing may lessen misconceptions associated with statistical significance and lead to more accurate conclusions and more justified decisions.

2.2. Students' Difficulties in Understanding Confidence Interval

It has been observed that the right understanding of confidence intervals seems to be difficult for university students taking statistics courses. According to [2], understanding confidence interval requires knowledge of mathematical objects related with it, like population, sample, statistic, parameter, standard error, sampling distribution, critical value and theoretical models of sampling distribution. Several studies have reported that students tend to interpret the confidence level associated with a confidence interval as the probability that the parameter value will be between the lower and upper interval limits [1, 4-5, 8-10]. A study conducted by [4] revealed that both students and professors possess a misconception regarding confidence intervals proving that confidence intervals is a difficult topic not only for students but also for the experts. Findings by [7] revealed that students fail to realize the inferential nature of confidence intervals and tend to interpret the confidence intervals as a descriptive statistic only and hold a variety of misconception about how the aspects of a confidence interval relate to each other (relationship between confidence interval width and sample size). In [11] observed that the difficulties in understanding the confidence interval can be divided into two: the difficulty related to the randomness of intervals in repeated samples and the difficulty related to the non-randomness of a confidence interval after the sample has been taken.

2.3. How Students Learn and Are Assessed

In [11] suggested the use of "Wheel of Fortune" as a device to understand the concept of confidence intervals. In this game, a student has to decide how many sectors (of equal size) the wheel has to be partitioned into. If he wants to work with a procedure that produces "good" intervals (containing π with probability 95%), he may ask for a division of the wheel into 100 sectors. This game can help students gain a better understanding of what is going on with confidence intervals and hence, be able to differentiate between the word "confidence" and "probability" in interpreting a confidence interval. In [8] provided an example of assessment items used to assess students' understanding, reasoning and thinking after students studied a unit on confidence intervals. While this type of assessment does not produce a score that is entered into a student's record, it does provide "just in time" feedback that can help a student determine whether he has attained understanding or needs additional help and information. In [12] suggested an easy way to teach the notion of a "confidence interval". It involves a construction of a series of simulations with samples of size 3 from the population of digits with uniform distribution, and form confidence intervals for the median from the minimum and maximum of each sample. The simulations can be done quickly and is also easy to understand and display and hence, can provide a clear and effective introduction to the concept of a confidence interval.

3. Methodology

This study was conducted on semester 4 students enrolled Diploma in Statistics (CS111) and Diploma in Actuarial Science (CS112) at UiTM Perak Branch, Tapah Campus during the December 2016 – April 2017 session. The number of respondents involved in the study was 197. In order to fulfil the requirement for the diploma program, students from both programs were required to take Fundamentals of Regression Analysis course. This course was taught in face-to-face sessions using 3 methods which are lectures, computer labs and tutorials. The class met five hours a week over a 14-week (1 semester) period. The assessment methods such as tests (*Test 1* and *Test 2*), quizzes and lab assignments were used to assess the attainment of the learning outcomes throughout the semester.

The students were given 2 common tests to measure their learning attainment. The first test was given after 7 weeks of lectures and the second test was given after 13 weeks of lectures. The question in Fig. 1 is one of the questions tested on *Test 2*, which was designed to assess students' ability in interpreting confidence interval of the regression slope. In this question, students were required to calculate the value of the confidence interval and interpret the meaning of the interval in the context of the problem.

Student answer scripts for *Test 2* were graded based on clarity of working as given in the marking scheme. For the purpose of this study, each answer script was observed to identify students' ability to compute and interpret the confidence interval of the regression slope. Data was recorded and analyzed using IBM SPSS Statistics 24.

A psychologist is interested in investigating the effect of demographic factors and the amount of time spent sleeping per day on the depression level (measured using the Beck Depression Inventory – BDI) of a person. The independent variables used were age (in years), gender (1 = female, 0 = male) and time spent sleeping (hours/day). Data was collected from a sample of 25 respondents and the following regression output was obtained.

ANOVA

Model		Sum of Squares	df	Mean Square	F
1	Regression	4936.593	3	1645.531	9.965[b]
	Residual	3467.765	21	165.132	
	Total	8404.358	24		

a. Dependent Variable: Beck Depression Inventory
b. Predictors: (Constant), age, gender, hour sleep

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t
		B	Std. Error	Beta	
1	(Constant)	67.310	9.150	.127	7.357
	Gender	4.664	1.146	-.744	4.070
	Hour sleep	-8.462	1.744	-.026	-4.852
	Age	-.044	.260		-.168

a. Dependent Variable: Beck Depression Inventory

Obtain a 90% confidence interval for the slope of the variable "time spent sleeping". Interpret the value obtained in the context of the problem.

Fig. 1: Item used to assess students' ability to calculate and interpret confidence interval of the slope

4. Results and Discussion

Table 1: Demographic information of respondent (N = 197)

Variable	Frequency (%)
Gender	
Male	52 (26.4%)
Female	145 (73.6%)
Program	
Diploma in Statistics	98 (49.8%)
Diploma in Actuarial Science	99 (50.2%)

Table 1 summarizes the demographic information of the respondents. Male respondents comprised 26.4% of the population and the remaining 73.6% were female. 49.8% of the respondents were from the Diploma in Statistics program, while 50.2% were enrolled in the Diploma in Actuarial Science program.

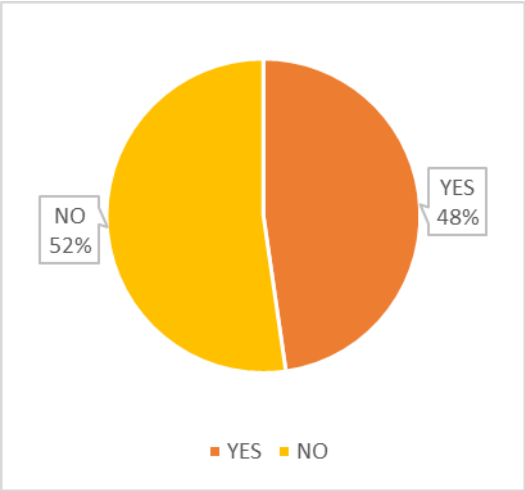

Fig. 2: Students' ability in calculating the value of confidence interval of the regression slope (N = 197)

First, the ability of students on procedural literacy was observed where they had to calculate a 90% confidence interval for the slope of the variable "time spent sleeping". To answer the question, students needed to recall the correct formula, plug in the correct values into the formula and carry out computation. The correct value of confidence interval was (-11.4634, -5.4606). Fig. 2 shows that 48% of the respondents (94 students) successfully computed the value of the confidence interval and 52% failed to compute the value correctly. Most students failed to compute the value because they failed to determine the correct critical value as they were unable to identify the correct degrees of freedom or failed to calculate the value of standard deviation of the estimated slope, $s(b_1)$ correctly.

Next, the study continued to investigate the aspect of statistics reasoning, i.e. how students interpret the value of the confidence interval of the regression slope. Based on students' answers, the ability to interpret the value of confidence interval was divided into three main categories, namely complete interpretation, incomplete interpretation and wrong interpretation (an interpretation that is completely incorrect or some calculation with no interpretation provided).

Table 2: Students' ability in interpreting the value of confidence interval of the regression slope (N = 197)

	Complete Interpretation	Incomplete Interpretation	Wrong or No Interpretation
Ability to interpret the value of confidence interval of the regression slope	14 (7.1%)	48 (24.4%)	135 (68.5%)

This study revealed that only 7.1% of the students interpreted the value of confidence interval of the slope correctly, 24.4% gave incomplete interpretation and 68.5% gave wrong interpretation (or left the answer blank) (Table 2). Some of students' interpretation that was considered incomplete or wrong is listed in Table 3. From the question, the most appropriate interpretation is "We are 90% confident that the depression level would decrease between 5.46 and 11.46 units for every one hour increase of time spent sleeping of a person, keeping all other variables constant". Another interpretation that can be used is "Since the 95% confidence interval does not include zero, we can reject the null hypothesis and conclude that the depression level would decrease between 5.46 and 11.46 unit for every one hour increase of time spent sleeping of a person, keeping all other variables constant".

This study has revealed that students do have difficulties in interpreting the confidence interval of the regression slope. Is this difficulty caused by their inability to understand the result of the confidence interval of the slope, or their failure to interpret the slope effectively, making it more challenging to combine this interpretation with the confidence interval concept? A previous study conducted by [13] on the same set of students found that most of the students (66.82%) were unable to provide the correct interpretation of regression slope. Fig. 3 illustrates the comparison between the ability of students to interpret the value of regression slope and confidence interval of the regression slope. It can be seen that the number of students who gave wrong or no interpretation of "slope" and "confidence interval of slope" is relatively high compared to other categories. Only 67 out of 197 students gave complete interpretation of the slope. However, this number dropped to 14 (out of 197) who were able to interpret the confidence interval of the slope completely. The pattern continues where we can see that more students failed to interpret the value of confidence interval of the slope (135 students) compared to students who failed to interpret the value of the slope (98 students).

Table 3: Examples of incomplete and wrong interpretation of confidence interval of the regression slope

No.	Students' Answer	Comment
1	We have enough evidence to conclude for everyone hour increases in time spent sleeping of a person, the depression level will decrease by -0.852, keeping the other variables constant.	Double negative is used.
2	We are 90% confidence that the minimum time spent sleeping is -11.4365 and the maximum of time spent sleeping is -5.4675.	Inability to understand the concept of slope.
3	The value of time spent sleeping is between -11.463 hour/day and 5.4606 hour/day.	Inability to understand the concept of slope.
4	We are 90% confidence that the interval for the slope of the variable "time spent sleeping" lies between -11.4634 and -5.4606.	Inability to understand the concept of slope.
5	90% confident interval for the variable time spent sleeping is lies on around -11.4634 and -5.4606.	Inability to understand the concept of slope.
6	We are 90% confident that the "time spent sleeping" are between -0.115 and -16.812	Inability to understand the concept of slope.
7	We can conclude that 90% confidence interval for the slope of variable spent sleeping are between 62.77 and 71.49	Inability to understand the concept of slope.
8	We are 90% confident that the slope of the variable "time spent sleeping" is in between -11.46 and -5.46	Inability to understand the concept of slope.
9	We are 90% confidence that the variable "time spent sleeping" values of a person decrease between 11.46 34 and 5.4606 hour per day.	Inability to understand the concept of slope.
10	We are 90% confidence that for every 1 hour increase in time spent sleeping, the depression level will lies between -3.3476	Inability to understand the concept of

	and 2.1552.	slope.
11	We are 90% confident the slope is between -11.4634 and -5.4606.	Interprets the interval
12	Since there is -11.451 and -5.473 confidence interval, there is difference between Beck Depression Inventory and time spent sleeping.	Inability to interpret the confidence interval of the slope.

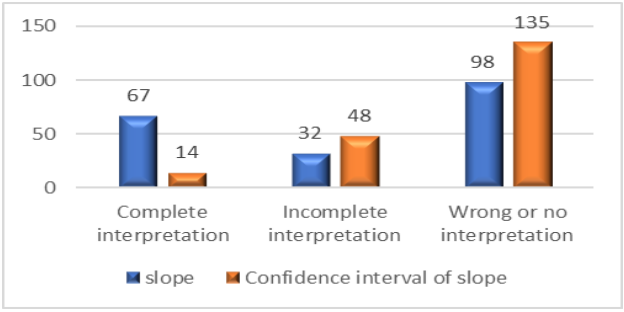

Fig. 3: Graphical comparison between ability of students to interpret the value of the regression slope and confidence interval of the regression slope (N = 197)

Table 4: The association between students' ability to interpret the value of regression slope and confidence interval of the regression slope (N = 197)

		Ability to Interpret the Value of Confidence Interval of Slope			χ^2 (df)	p Value
		Complete Interpretation	Incomplete Interpretation	Wrong Or No Interpretation		
Ability to interpret the value of slope	Complete Interpretation	6 (9%)	17 (25%)	44 (66%)	7.50 (4)	0.112
	Incomplete interpretation	5 (16%)	9 (28%)	18 (56%)		
	Wrong or no interpretation	3 (3%)	22 (22%)	73 (75%)		

Next, a test was performed to investigate the association between "ability to interpret the value of slope" and "ability to interpret the value of confidence interval of the slope" among the students. A Chi-square test for independence (see Table 4) indicated non-significant association between "ability to interpret the value of slope" and "ability to interpret the value of confidence interval of the slope" χ^2 (4, N = 197) = 7.50, p = 0.112.

The multiple bar chart (Fig. 4) shows the comparison of abilities to interpret the confidence interval between the two diploma programs. It can be seen that most of the students fall in the "wrong or no interpretation" group for both programs (CS111 and CS112). Only 14 out of 197 students could interpret the value of the confidence interval correctly of which 5 were from Diploma in Statistics and 9 from Diploma in Actuarial Science. It is clear that most of the students in both programs have difficulties in interpreting the confidence interval of slope.

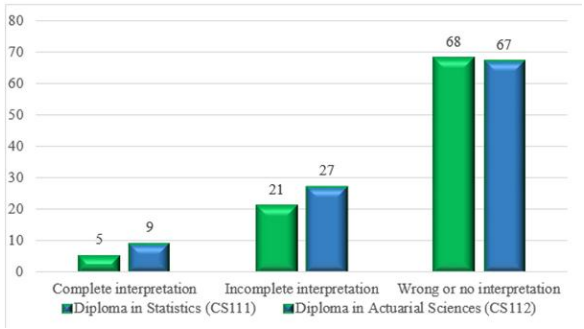

Fig. 4: Graphical comparison of abilities to interpret the value of confidence interval of the regression slope between the two diploma programs.

Table 5: Association between students' ability in interpreting the value of confidence interval of the regression slope and program (N = 197)

		Ability to Interpret the Value of Confidence Interval of the Regression Slope			χ^2 (df)	p Value
		Complete interpretation	Incomplete interpretation	Wrong or no interpretation		
Program	CS111	5 (5.3%)	21 (22.4%)	68 (72.3%)	1.49 (3)	0.474
	CS112	9 (8.8%)	27 (26.2%)	67 (65%)		

Further analysis was done to investigate the association between "ability to interpret the value of confidence interval of the slope" and "programs". A Chi-square test for independence (see Table 5) indicated non-significant association between "ability to interpret the value of regression slope" and "programs" χ^2(3, N = 197) = 1.492, p = 0.474. This test confirmed that students from both programs have difficulties in interpreting the confidence interval of slope.

5. Conclusion

This study has shown that students have difficulty not only in interpreting the confidence interval of the regression slope, but also in computing the confidence interval of the regression slope. Only 48% of students successfully calculated the value of the confidence interval while 52% failed to complete the calculation correctly. Most students failed to compute the value because they failed to determine the correct critical value for multiple regression or failed to calculate the standard deviation of the estimated slope, $s(b_1)$. The most critical finding in this study is: only 7.1% of the students could interpret the value of the confidence interval completely and correctly while 24.4% could interpret but not completely, and 68.5% failed to interpret the value of the confidence interval of the slope. Educators should find a strategy on how to help students understand the principles behind confidence interval of the slope. Since the medium of instruction is English and English is only a second language for these students, it is believed that proficiency in the language is also a contributing factor in students' difficulty in understanding and interpreting the confidence interval of the regression slope. Further research can be conducted to investigate the strategies that can be used to teach the confidence interval of the regression slope and how to best educate statistics students.

Acknowledgement

The study is financially funded by the Academic and Research Assimilation Grant, Universiti Teknologi MARA, 600-IRMI/DANA 5/3/ARAS (0180/2016). Special thanks are also extended to those participated in this study.

References

[1] Finch S, & Cumming G (2008), Putting research in context: Understanding confidence intervals from one or more studies. *Journal of Pediatric Psychology* 34(9), 903-916.

[2] Andrade L & Fernánde F (2016), Interpretation of confidence interval facing the conflict. *Universal Journal of Educational Research* 4(12), 2687-2700.

[3] Kutner MH, Nachtsheim C & Neter J (2004), Applied linear regression models. *McGraw-Hill/Irwin*.

[4] Canal GY & Gutiérrez RB (2010), The confidence intervals: A difficult matter, even for experts. *Proceedings of the 8th International Conference on Teaching Statistics*, pp. 1-4.

[5] Reaburn R (2014), Students' understanding of confidence intervals. *Proceedings of the 9th International Conference on Teaching Statistics*, pp. 1-4.
[6] Sim J & Reid N (1999), Statistical inference by confidence intervals: Issues of interpretation and utilization. Physical Therapy 79(2), 186-195.
[7] Fidler F, & Cumming G (2005), Teaching confidence intervals: Problems and potential solutions. *Proceedings of the 55th International Statistics Institute Session*, pp. 1-5.
[8] DelMas RC (2002), Statistical literacy, reasoning, and learning: A commentary. *Journal of Statistics Education* 10(3), 1-11.
[9] Andrade L, Fernández F & Álvarez I (2014), Fostering changes in confidence intervals interpretation. *Proceedings of the 9th International Conference on Teaching Statistics*, pp. 1-4.
[10] Henriques A (2016), Students' difficulties in understanding of confidence intervals. In D. Ben-Zvi, & K. Makar (Eds.), Teaching and Learning of Statistics: International Perspectives. Basel: Springer International Publishing, pp. 129–138.
[11] Callaert H (2007), Understanding confidence intervals. Working Group 5. Stochastic Thinking, pp. 685-692.
[12] Holte JM (2003), Teaching confidence intervals. Proceedings of the NCS-MAA Spring Meeting, pp. 1-6.
[13] Fadzilah AR, Baharun N, Deraman NA & Ismail NRP (2017), Assessing students' abilities in interpreting the correlation and regression analysis. *Journal of Fundamental and Applied Sciences* 9(9S), 644-661.