Readiness of Final Semester Diploma in Statistics Students to Progress to the Next Level: A Trial Exit Test

Nor RashidahPaujah @ Ismail, FadzilahAbdolRazak, Elizabeth Arul

Abstract—Diploma in Statistics graduates choose to either to pursue further education or begin their careers as assistant statisticians. In either case, they are expected to have a strong of understanding in statistics and should be able to apply the knowledge and skill that they have learned when dealing with real data sets. This paper aims to assess the readiness of final semester Diploma in Statistics students to meet this expectation by analyzing their responses on a trial exit test. The test was given during the final (fifth) semester before graduation. The results showed that these students have the capability to compute the statistics values correctly. However, they still lack an understanding of the appropriate statistics that should be used as they failed to respond correctly to question on the application of statistics. They were not competent enough to understand the formula, and could not see how certain data could not appropriately be represented by certain statistics. For instance, it is not appropriate to use mean for a data set which contains outliers. This deficiency should be catered for and resolved by educators, so that the institution may provide skilled and knowledgeable statistics graduates who are able to meet the requirement set by educators and other stakeholders.

Keywords: Correlation and regression, descriptive statistics, measures of central tendency, measures of variation.

I. INTRODUCTION

According to [1], the goal in introductory statistics courses is to ingest a large amount of data that is generated every day, to do a critical analysis of the data and make good decision based on the analysis of the data. The second goal is to develop research scientist skills in students by promoting the use of the scientific method to collect data, determine and apply appropriate statistical tools to interpret the data, and communicate and share results.

To produce graduates who are going to fulfill the job requirements as a statistician, UniversitiTeknologi MARA continues to offer Diploma in Statistics as one of the popular programs that has a high demand. One of the program outcomes is to produce assistant statisticians who are able to apply fundamental statistical knowledge to solve statistical problem. To determine whether this outcome is achieved, this research focuses on investigating students' knowledge and understanding in elementary statistical concepts upon completion of their study in Diploma in Statistics. A trial exit test consisting of a variety of questions which covered different statistics topics that have been taught throughout five semesters of study was given to final semester students before they left the university. Specifically, the objectives were set as follows:

i. To investigate students' knowledge and understanding in numerical descriptive statistics (measures of central tendency and variation).

ii. To investigate students' knowledge and understanding in basic concepts of correlation and regression analysis.

iii. To investigate the association between students' academic performance and the accuracy score in the trial exit test.

II. LITERATURE REVIEW

Descriptive statistics is a branch of statistical methods that helps us to describe data and its properties, while inferential statistics helps us to draw conclusions about the characteristics of the population from a given data sample. Descriptive statistics can be divided into numerical and graphical methods while numerical descriptive statistics can be further divided into measures of central tendency and variability.

A. Measures of central tendency

A measure of central tendency is a single value that describes a set of data by identifying the central location within that set of data. Measures of central tendency can be divided into mean, median and mode and have many applications in everyday life. The computational aspects of mean, median and mode is simple to apply but the conceptual understanding can be quite difficult. In [2],the authorconcluded that most people can calculate the value of mean using the correct formula but have problems relating to it as representative of the data set. However, in [3] found that the concept of mean is complex and a conceptually vital process. The students appear to face difficulties in comprehending mean and show a tendency to calculate the algorithm in order to resolve statistical problems. In agreement with the above, in [4]in her finding summarized that the inability of students to solve problems related to the mean is due to a lack of deep understanding of the underlying concepts.

In [5],astudy of the development of children's understanding of properties of the arithmetic mean, found that some of the properties of the mean such as "the sum of the deviations from the average is zero", "when one

Revised Manuscript Received on September 14, 2019.

Nor Rashidah Paujah @ Ismail, Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Perak Branch, Tapah Campus, Tapah Road, 35400 Perak, Malaysia.

Fadzilah Abdol Razak, Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Perak Branch, Tapah Campus, Tapah Road, 35400 Perak, Malaysia.

Elizabeth Arul, Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Perak Branch, Tapah Campus, Tapah Road, 35400 Perak, Malaysia

calculates the average, a value of zero, if it appears, must be taken into account" and "the average is representative of the values that are averaged" are difficult for students to understand.

Another interesting finding was by [6]who examined and analysed teacher views about the difficulties in learning and teaching middle school statistics subjects. The difficulties mentioned by teachers in relation to central tendency are as follows:

- students are confused about the concepts of median and mode.
- students attempt to find the median without sorting the data.
- students make the mistake of dividing by a fixed number when calculating the mean.
- students tend to mainly use the arithmetic mean to represent a data set.
- students are unable to interpret the mode, median and mean.

In [7]also study students' academic difficulties in learning a statistics and probability course from the point of view of the instructors. He prepared a 39-item questionnaire to determine the academic difficulties that students experienced in learning statistics and probability. The level of difficulty for each item was identified from the point of view of the researcher. The following is some of the items that were classified as high in difficulty level:

- Calculate the median of a frequency distribution table.
- Calculate the mode of a frequency distribution table.
- Realize the relationship between the mean, median, and mode.
- Recognize the indication of symbols when calculating the quartile or percentile.
- Realize that Q1=P25, median = Q2= P50, Q3=P75.

B. Measures of variation

Standard deviation is one of the common measures of variation that is widely used to measure how far data depart from the central tendency. A study by [6] found that there was confusion among students in the calculation of standard deviation as they were unsure whether to use the formula for population or sample. This finding was in line with [7] who found that to calculate a variance (application of the rule and indication of symbols) was highly difficult for students. In [8] however believes that ability to calculate the standard deviation does not translate to ability to understand the meaning of the standard deviation, what it measures, or how it is used. A study by [9] on how educators apply concepts related to the idea of variation come to an agreement with this as the result from her study revealed that the participants knew how to compute the values of variance or standard deviation. However, the participants did not know how to analyse these variation measures and failed to construct the concept of variation around the mean.

In [6]also found that students were being unable to calculate the interquartile range.

C.*Correlation and regression analysis*

Correlation and regression analysis are statistical techniques for investigating and modeling the relationship between variables [10]. Regression models identify the type of mathematical relationships that exist between a dependent variable and an independent variable, thus the changes in the dependent variable (Y) can be predicted based on the changes in the independent variable (X). The correlation coefficient measures the strength of the linear relationship between two quantitative variables, and it does not depend upon which variable is labeled X and which variable is labeled Y.

In [11], the authorinvestigated students' abilities in interpreting the correlation coefficient and regression coefficient. The findings discovered that only 19.43% of students managed to correctly interpret the value of Pearson's correlation coefficient while 33.18% of the students were able to interpret the value of the regression slope completely and correctly. These findings support a previous study by [7]who found that to explain the value of the correlation coefficient was highly difficult for students. Additionally, in [7]the authorfound that students have high difficulty in recognizing the symbol used for Pearson's Coefficient of correlation, to calculate the slope of the linear regression equation and to estimate the value of the dependent variable when the independent variable was defined.

In a more detailed discussion of regression, in [12]explored students' ability to write a correct hypothesis based on the statement involving to regression coefficients. The results indicated that students were able to write the proper hypothesis statement for a regression coefficient that directly refers to the slope of the variable. However, they failed to provide the correct hypothesis statement when they had to translate the definition of the slope. In a similar study, in [13]explored students' ability in computing the confidence interval of the regression slope and whether they were able to explain clearly the meaning of their computation. Findings showed that only 48% of the students managed to compute the confidence interval of the slope correctly. 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%.

III. METHODOLOGY

This study was conducted at the end of the Sept – Dec 2018 semester, involving final semester students taking Diploma in Statistics (CS111) at UiTM Perak Branch, Tapah Campus. The number of respondents involved in the study was 50 out of 73 final semester students. Students in the Diploma in Statistics program have to enroll and pass nine statistics courses, three mathematics courses, three computer science courses, one operations research course and the remaining courses are compulsory university courses. Students need to complete 90 credit hours within a minimum of five semesters in order to graduate.

An instrument which was later defined as a trial exit test was developed to test students' understanding in some of the statistics concepts that they had been taught throughout their five semesters of study. The instrument was divided into two parts. The first part comprised of 25 questions which focused on basic knowledge in the measures of central tendency and measures of variation as well as the basic of correlation and regression. The second part focused on students' knowledge and understanding in a variety of statistical methods, consisting of 18 questions. However, only results from the first part will be discussed in this paper.

During the data collection process, students were gathered at the same place and time and were required to answer all questions independently using the online Quizizz tool which was used to gain information about how the class as a whole is doing in understanding content material. Quizizz was chosen as the tool in this study as it can produce the report for the test in the form of an Excel file which can be easily downloaded and analyzed. Data was then, transferred to SPSS to obtain more statistics on the results produced.

The instrument developed was answered by two experts beforehand in order to test it. The average time to complete the instrument by these experts was 10 minutes, therefore the students were allotted 30 minutes to complete the online test. As a general rule of thumb, students need triple the amount of time required by experts. All the questions were designed as multiple choice questions; true/false questions or student were required to fill in the blanks. Students were allowed use their calculators and statistical tables to answer questions.

Quizizz produced accuracy scores based on percentage of correct answers. Accuracy score is one of the interest variables in this study. The second variable of interest is cumulative grade point average (CGPA). According to [14], CGPA is usually used to measure overall students' academic performance where it considers the average of all examination grades for all semesters during the student's tenure in university.

IV. RESULTS AND DISCUSSION

Table 1 summarizes the demographic information of the respondents. Male respondents comprised 20% of the sample and the remaining 80% were female.

Table- I: Gender demography

Variable	Frequency (%)
Gender	
Male	10 (20%)
Female	40 (80%)

Table- II: Items regarding numerical descriptive statistics

No.	Item	Number of Correct Responses (%)
1	Identify the mode from the data below: 7, 5, 0, 7, 8, 5, 5, 4, 1, 5 Answer: *5*	50 (100%)
2	The number of pages that Amirah wrote in her journal each day from Monday to Friday is: 9, 8, 12, 6, 10. What is the mean number of pages she wrote per day? Answer: *9*	
3	Which measurements are used to analyze the centrality of data? Answer: *Mean, Median and Mode*	49 (98%)
4	In order to find the median, the data must be sorted from least to greatest first. Answer: *True*	48 (96%)
5	What is the symbol for median? Answer: $\tilde{x}$	47 (94%)
6	Roslan bowled 7 games last weekend. His scores are: 155, 165, 138, 172, 127, 193, 142. What is Roslan's median score? Answer: *155*	46 (92%)
7	Which of the following is a measure of variability? Answer: *standard deviation*	45 (90%)
8	What is the symbol for mode? Answer: $\hat{x}$	43 (86%)
9	A set of data on how many hours per week each student sleeps was collected. The standard deviation calculated was 0. Which of the statement below is true? Answer: *All students slept the same amount of time.*	39 (78%)
10	Standard deviation can be negative. Answer: *False*	39 (78%)
11	Which of the options given is true for the statement below? "Variance of a data set is computed using formula of variance with the denominator of (n – 1)". Answer: *Data set is a sample.*	39 (78%)
12	Refer to the histogram. Which statement is true? Answer: *The histogram is bimodal.*	38 (76%)
13	Any data set might have more than one modal value. Answer: *True*	33 (66%)
14	Standard deviation is not affected by outliers. Answer: *False*	31 (62%)
15	Which of these options is true, given that a data is skewed to the right with a median of 50? Answer: *Both A and C*	31 (62%)
16	If a number is added to a set that is far away from the mean, this would affect the standard deviation as it will_____. Answer : *increase*	28 (56%)
17	_____ would change if any data in the set changes. Answer: *Mean*	27 (54%)

As expected, most of the students had no problems in calculating the values of mean, median and mode as can be seen in Table II (Items 1, 2, 4 and 6). They knew that these measures are used to analyze the central tendency of data (Item 3). However, of the three symbols representing mean,

median and mode, 14% of the students failed to recognize the correct symbol for a sample mode which is $\hat{x}$ (Item 8). This might have been because most students were used to the words median and mode, instead of the symbols in answering the questions involving calculation of these central tendencies.

The concept and usage of measures of central tendency and measures of dispersion were poorly mastered by students (Items 9-17). For example, for item 16, 44% of the students failed to identify the effect of extreme values on a standard deviation (the standard deviation will increase) while 38% of them were unable to respond correctly that standard deviation is affected by extreme values (Item 14). 46% of the students failed to understand the concept behind the process to obtain the values of mean, median and mode as they could not recognise that only the mean always changes if a single value in the data changes (Item 17). This finding suggests that the students were good in using the correct formula to find the statistic, however, they failed to see how the formula related to the nature of the data. Consequently, the students may not be able to suggest an appropriate statistic when dealing with a real data set in future.

Table III: Items regarding correlation and regression analysis

No.	Item	Number of Correct Responses (%)
1	Choose the graph that represent a strong positive linear relationship? Answer: *option 2 (image)*	50 (100%)
2	Given a fitted equation of y = 120 - 5x where x represents height and y represents weight. The slope of the above equation suggests that the weight is expected to _____ for 1 inch increase of height. Answer: *decrease by 5 pound.*	44 (88%)
3	A vertical distance of a point from a regression line in a scatter plot is known as _____. Answer : *residual*	38 (76%)
4	Which of the following statement is true? Answer: C*oefficient of determination is the square of coefficient of correlation.*	31 (62%)
5	What would happen when more variables were added to a linear regression model? Answer: *The coefficient of determination might increase or remain unchanged while the adjusted coefficient of determination might increase or decrease.*	29 (58%)
6	Given that there is a very high correlation betweenmathematics test scores and amount of physical exercise done by a student per week, which of the inferences below is most suitable for the above statement? A. High correlation implies that doing physical exercise results high test scores. B. Correlation does not imply causation. C. Correlation measures the strength of the linear relationship between amount of physical exercise and test scores. Answer: *B & C.*	28 (56%)
7	The correlation coefficient between variable P and variable Q is 0.65. If we multiply all the values in variable P with 100, the correlation coefficient _____. Answer: *will stay the same.*	26 (52%)
8	If the correlation coefficient between the amount of saturated fat consumed per day and the cholesterol level of men is 0.78, what percentage of variability in cholesterol level is explained by the amount of fat consumption? Answer: *61%*	20 (40%)

Table III suggests that students had no problem identifying the type and direction of relationship between two variables, from a scatter plot (Item 1). 88% of students who answered item 2 correctly, showed that they were also aware of the meaning of slope in the regression equation. However, they showed a lack of understanding in questions involving the coefficient of determination (R^2). At least 40% of students gave wrong answers on items 4 and 8. This indicates that the concept and usage of coefficient of determination were poorly mastered by students as they failed to relate coefficient of determination and coefficient of correlation. Student also failed to understand that the variability in the dependent variable explained by the variability of independent variables can be represented as the coefficient of determination.

For item 6, 22 out of 50 students (44%) failed to understand that the word correlation does not imply causation or cause of effect relationship. Many of them believed that "High correlation implies that doing physical exercise results high test scores" is a true statement.

The researchers also found that students were not aware that the value of *r* does not depend on the unit of measurement (Item 7). Only 52% of the students gave the correct answer while the majority of them chose "increase" as their answer.

Table- IV: Descriptive statistics for CGPA and accuracy score

Variable	Mean	Std. Deviation	*n*
CGPA	3.3682	0.34412	50
Accuracy score	75.92	13.389	50

Table IV shows that the mean CGPA of students was 3.368 (SD = 0.34412) while the mean for accuracy score was 75.9 (SD = 13.389). CGPA scores were obtained after the students finished their Diploma studies in December 2018 and based on 4.0 CGPA scale, while the accuracy scores were obtained from this study and was calculated based on percentage of correct answers.

Next, the relationship between students' CGPA and the accuracy scores was investigated using Pearson product-moment correlation coefficient. Preliminary analyses were performed to ensure no violation of the assumptions of normality, linearity and homoscedasticity. The findings in Table V suggests that there is no significant correlation between the two variables [r = -0.040, n = 50, p = 0.784 > 0.05].

Table V: The association between CGPA and accuracy score (n=50)

		CGPA	Accuracy Score
CGPA	Pearson Correlation	1	-0.040
	Sig. (2-tailed)		0.784
	N	50	50
Accuracy score	Pearson Correlation	-0.040	1
	Sig. (2-tailed)	0.784	
	N	50	50

V. CONCLUSION

In this study, students were required to answer a variety of questions which tested their knowledge and understanding of basic statistics, correlation, regression, and statistical methods. These topics had been taught since they were in the first semester. As Diploma in Statistics graduates, students are expected to have strong understanding of the statistics courses as they would need to apply all this knowledge in their careers or further studies.

To be a statistician, it is very important to know how to choose a correct statistic and/or method to represent the data. Therefore, to understand the theory and successfully apply correct methods are crucial. However, the results obtained from this study suggest that the students did not have strong understanding of the theoretical part of the statistics courses. Even though they could successfully compute the statistics value using the correct formula, they failed to understand that the representation of statistics, for example the mean, would be affected and may not be appropriate anymore to represent the data when there is an outlier in the data. Students knew how to use the formula, but they did not know when to use the correct statistic when they dealing with a real data set.

Students also show limited understanding of regression analysis. They were unable to relate coefficient of correlation with the meaning of coefficient of determination. Some were still confused with the effect of the unit of the data on the correlation coefficient value. As expected, there is no relationship between students' CGPA and accuracy score. This could be because the CGPA measures their performance for each semester, when the students were fully prepared for the exam. However, for this test there was no preparation and the questions were based on randomly selected topics that they have previously studied.

This was a preliminary study to investigate how far students were able to understand basic concepts of statistics and basic regression. The instrument may be further refined with more related questions in order to get a clearer view on how the students understand statistics concepts.

VI. ACKNOWLEDGMENT

The researchers would like to thank the Faculty of Computer and Mathematical Sciences, UiTM Perak Branch, Tapah Campus for providing the funding for the Scopus publication. Special thanks are also extended to all those who contributed directly or indirectly to this study.

REFERENCES

1 D. J. Rumsey, "Statistical literacy as a goal for introductory statistics courses," Journal of Statistics Education, 10(3), 2002, pp. 1-12.
2 M. R. Leon, "Use of arithmetic mean: An investigation of four properties issues and preliminary results," III International Conference on Teaching Statistics, 1990, pp. 302-306.
3 E. Chatzivasileiou, I. Michalis, and C. Tsaliki, "Elementary school students' understanding of concept of arithmetic mean," 8th International Conference on Teaching Statistics, 2010, pp. 1-4.
4 C. Batanero, J. D. Godino, A. Vallecillos, D. E. Green, and P. Holmes, "Errors and difficulties in understanding elementary statistical concepts," International Journal of Mathematical Education in Science and Technology, 25(4), 1994, pp. 527-547.
5 S. Strauss and E. Bichler, "The development of children's concepts of the arithmetic average," Journal for Research in Mathematics Education, 19(1), 1988, pp. 64-80.
6 T. Koparan, "Difficulties in learning and teaching statistics: Teacher views," International Journal of Mathematical Education in Science and Technology, 46(1), 2015, pp. 94-104.
7 R. A. A. Kandeel, "Students' academic difficulties in learning a statistics and probability course: The instructors' view," Journal of Education and Practice, 10(9), 2019, pp. 43-52.
8 I. Gal, Adult Numeracy Development: Theory, Research, Practice. New Jersey: Hampton Press, 2000.
9 C.B. da Silva, "The variation concept: A study with secondary school mathematics teachers," 7th International Conference on Teaching Statistics, 2006, pp. 1-4.
10 D. C. Montgomery, E. A. Peck, and G. G. Vining, Introduction to Linear Regression Analysis. New Jersey: John Wiley and Sons, 2012.
11 A. R. Fadzilah, N. Baharun, N. A. Deraman, and N. R. P. Ismail, "Assessing students' abilities in interpreting the correlation and regression analysis," Journal of Fundamental and Applied Sciences, 9(5S), 2017, pp. 644-661.

12 A. R. Fadzilah, N. R. P. Ismail, N. Baharun, and N. A. Deraman, "Hypothesis testing on regression: Investigating students' skill," International Journal of Engineering and Technology, 7(4.33), 2018, pp. 45-48.
13 N. R. P Ismail, A. R. Fadzilah, N. Baharun, and E. S. G. Arul, "Investigating students' difficulties in understanding confidence intervals in linear regression models," International Journal of Engineering and Technology, 7(4.33), 2018, pp. 60-64.
14 A. Norhidayah, J. Kamaruzaman, A. Syukriah, M. Najah, and S. A. S Azni, "The factors influencing students' performance at Universiti Teknologi MARA Kedah, Malaysia," Management Science and Engineering, 3(4), 2009, pp. 81-90.

AUTHORS PROFILE

Nor RashidahPaujah @ Ismail,is a senior lecturer in the Faculty of Computer and Mathematical Sciences, UniversitiTeknologi MARA Perak Branch, Tapah Campus. She has a bachelor's degree and master's degree in Applied Statistics. Her research interests include Statistics Education, Regression Analysis and Operations Research.

FadzilahAbdolRazak is a senior lecturer in the Faculty of Computer and Mathematical Sciences, UniversitiTeknologi MARA Perak Branch, Tapah Campus. She has a bachelor's degree and master's degree in Applied Statistics. Her research interests include Statistics Education, Regression Analysis and Applied Statistics.

Elizabeth G. Arulis a senior lecturer in the Faculty of Computer and Mathematical Sciences, UniversitiTeknologi MARA Perak Branch, Tapah Campus. She has a Bsc in Mathematics and MS in Applied Mathematics. Her research interests include Mathematics Anxiety and Mathematics Education.