

Understanding the Concept of P-Value from Perspectives of two Distinct Definitions

Charles K. Assuah

Abstract— Most lecturers still struggle to confidently explain the concept of *p*-value to their students in introductory research method courses. Often, they use the terms *p*-value and probability interchangeably in their explanations. This misconception, in many cases, is transferred to their students, and sometimes feature very prominently, even among seasoned instructors and researchers. This study provides a conceptual understanding of first-year postgraduate mathematics education students about the *p*-value from perspectives of two distinct definitions. The first defines *p*-value as the sum of three parts: The probability random chance would result in the observation, the probability of observing something else that is equally rare, and, the probability of observing something rarer or more extreme. The second defines *p*-value as the probability of obtaining results as extreme as the observed results of a statistical hypothesis, assuming that the null hypothesis is true. Forty-five (45) postgraduate mathematics education students from a mid-sized university in Ghana, were purposely selected for the study, after they had answered some diagnostic questions on the concept of the *p*-value. By working cooperatively in groups of three (3) using action research design, the students applied the two definitions in the tasks assigned to them by their instructor. The results indicated that the students' conceptual understanding of *p*-value was enhanced when their instructor, through cooperative learning method, helped the students to apply the definitions in their concept acquisition. The implication to teaching and learning is that instructors should consciously introduce concepts that enable students to apply definitions, theorems, corollaries and lemmas. The use of a cooperative learning method enhanced the students' learning outcome. The study concludes that a clear distinction between *p*-value and probability would be better understood and conceptualized if instructors teach their students concepts by applying definitions.

Index Terms— *p*-value, probability, conceptual understanding, concept, definitions.

I. INTRODUCTION

Statistical cognition is concerned with how people understand statistical concepts and interpret them (Beyth-Marom, Fidler, & Cumming, 2008). Even though the *p*-value is an important concept in statistics, its interpretation is often misconstrued among some researchers—It is indeed a measure of statistical evidence that appears in a lot of research. The *p*-value is probably the most widely-used, misunderstood, misinterpreted, and miscalculated index in research (Goodman, 2008). It is the probability of the observed result or a more extreme result of a statistical hypothesis if the null hypothesis is true (Fisher, 1925; Gill, 1999; Hubbard & Lindsay, 2008; Johnson, 1999). Although this definition is clear and precise, some instructors' and researchers' incorrect interpretations of the *p*-value still exist

(Cumming, 2012; Verdam, Oort, & Sprangers, 2014; Wagenmakers, 2007). Consequently, the *p*-value's conceptual meaning is widely and often misunderstood by students. A popular misconception is a false belief that the probability of a conclusion being in error can be calculated from the data in a single experiment without reference to external evidence of the underlying mechanism (Goodman, 2008). An incorrect interpretation of the *p*-value affects researchers' decisions and weakens the quality of psychological interventions and the accumulations of valid scientific knowledge (Badenes-Ribera, Frias-Navarro, Monterde-i-Bort, & Pascual-soler, 2015). Although introductory statistics is taken by millions of students every year, little is known about how textbooks explain the *p*-value and how that relates to what researchers do (Cumming, 2010).

II. LITERATURE REVIEW

A lot of misconceptions and misunderstandings about the *p*-value are rife in literature (Cummings, 2012). Some of these misunderstandings come from instructors who do not competently explain this concept to their students (Haller & Krause, 2002). Consequently, students develop their understanding of the *p*-value, which affects the judgements they make in statistical inference. When these students eventually become instructors, they also transfer these misunderstandings to their students, leading to streams of misconceptions. Some students believe that the *p*-value is the probability that the null hypothesis is true (Reaburn, 2014; Gliner, Leech, & Morgan, 2002). Others believe that the *p*-value can be obtained if an experiment were repeated (Cumming, 2006; Mittag & Thompson, 2000; Nickerson, 2000). These misconceptions and incorrect interpretations of the *p*-value are widespread, even among academic psychologists and seasoned researchers (Haller & Krauss, 2002). Some of the incorrect interpretations of the *p*-value are the “inverse probability” fallacy, the “replication” fallacy, “effect size” fallacy, and the “clinical or practical significance” fallacy (Carver, 1978; Cohen, 1994; Falk & Greenbaum, 1995; Fidler, 2005; Gill, 1999; Goodman, 2008; Johnson, 1999; Kirk, 1996; Kline, 2004; Levine, Weber, Hullett, Sun, & Massi, 2008, 2013; Nickerson, 2000; Thompson, 1996).

Even though the use of computers now allows students to explore and simulate data, create models, and compute statistics to draw statistical inferences (Garfield et al., 2012), numerous examples in textbooks still have inaccurate descriptions of the interpretation of Null Hypothesis Significant Testing (NHST) (Gigerenzer, 2004; Haller & Krauss, 2002; Pollard & Richardson, 1987). Despite the proliferation in the use of computers in schools, most teachers' instructional methods have not changed (Garfield,

Charles K. Assuah, Department of Mathematics, Ghana

delmas, & Zieffler, 2012). Instructors and researchers should vary their methods of analysis between, for example, a one-or two-tailed tests (Hubbard & Lyndsay, 2008). As they do this, they would identify inferential patterns that are usual and those that are unusual (Cobb, 2007).

Cooperative learning is a pedagogical practice that promotes socialization and learning among students across different subject domains. By this practice, students work together to achieve common goals or complete tasks that would be difficult for individual members to complete by themselves. Placing students into groups without any direction and control from instructors could create disharmony and discord among them, which could derail the process of achieving their objectives if they are not properly managed (Johnson & Johnson, 1990). To avoid these pitfalls, Johnson and Johnson (2009) propose that groups should embed in their structure five key components for the successful implementation of cooperative learning.

The first component involves structuring positive interdependence for group members to work together harmoniously to achieve a common goal. Through this process, they learn to understand and synchronize their efforts to meet this goal. For this group, success and failure are not attributable to the effort of a member, but to all members in the group. Positive interdependence could be established when all students in a group understand that they are all responsible for completing the task assigned to them. Teachers could ensure that this occurs by delegating different parts of the task to different students to complete (Johnson & Johnson, 2002).

The second component builds promotive interaction among students. Students can provide help to each other; share needed resources; provide effective feedback to group members; challenge others' conclusions and reasoning; and, work constructively together to attain mutual goals (Johnson & Johnson, 1990). It also encourages and facilitates each student's efforts to complete assigned tasks. The willingness to engage with other students benefits and encourages students to reorganise and restructure the information in their minds. Students also develop clearer and more elaborate cognitive understandings of the concepts (Webb & Mastergeorge, 2003). When teachers allow students to sit close to each other, they can interact with themselves and participate fully in the group's discussion. When students interact with their peers during the group's discussions, they read each other's non-verbal language, respond to social cues, and engage in collective efforts (Gillies, 2003a, b).

The third component is students' accountability or responsibility to ensure that they complete their share of the work, and ensure that others complete theirs too. The more students perceive they are linked together, the more they feel personally responsible for contributing to the collective effort of the group. Johnson and Johnson (1990) claim that teachers can establish individual accountability by structuring positive interdependence among students and holding them personally responsible for completing their tasks.

The fourth component is good interpersonal skills among students. These skills enable students to work effectively in their groups in managing disagreements among them. But these skills need to be explicitly negotiated. In studies that have investigated the effects of cooperative learning on

students' behaviours and interactions, students who were trained to cooperate and help each other were more inclusive of others; respectful and considerate of others' contributions; and, provided more detailed explanations to assist each other's learning than students who were not (Gillies, 2003a, 2003b, 2004, 2006, 2008; Gillies & Ashman, 1996, 1998). The social skills that facilitate students' interactions, listening to each other actively; sharing ideas and resources; commenting constructively on other's ideas; accepting responsibility for others behaviours; and, making decisions democratically.

The fifth component is group processing. It allows students to reflect on their progress and their working relationships. To stimulate this reflective process, the following questions are often asked: What have members of the group achieved so far? What would they still need to achieve? How might they do this?

Group processing involves ensuring that every student in the group engages in summarizing ideas and information, participates in the discussion, and checks to see that decisions made by the group are supported by all. Additionally, group processing enhances respect among students and contributes to an increase in students' collective identification (Johnson & Johnson, 2009). For cooperative groups to work effectively, the composition of the groups and their sizes is very paramount. In a meta-analysis of 66 studies examined, Lou, Abrami, Spence, Poulsen, Chambers, and d'Apollonia (1996) found that students achieved higher outcomes when they work in small groups than in traditional whole-class settings. Students worked better and achieved more when they worked in groups of 3-4 members than in groups of 5-7 members. In group composition, low-ability students learn best in either heterogeneous or mixed-ability groups, while medium-ability students benefit significantly in homogeneous groups. However, there was no difference in high-ability group composition.

Instructors' role in establishing cooperative learning experiences in their classrooms includes structuring the groups and the tasks so that students can know of their responsibilities and expectations. It also bestows a great responsibility to teachers that they have a role in promoting student interactions during their group discussions. Helping students to interact and work together, does not only enable them to learn from each other, but it also enables them to accept responsibility for the tasks they have to complete. Students elaborate on information when their teachers ask thought-provoking questions that spontaneously draw upon prior knowledge, with some relevant external guidance (King, 2002). The purpose of this study was to enable postgraduate mathematics education students to understand conceptually and interpret the p -value. It also provided some practical ways to enhance students' understanding of the concept. The study was guided by the following questions:

How did the use of the two statistical definitions enabled students understand the concept of p -value?

What role does the cooperative learning method play in students' understanding of the concept of p -value?

How would a lecturer demonstrate to his/her students an understanding of the concept of p -value using concrete examples?

What practical significance do researchers derive from the concept of p -value?

III. MATERIALS AND METHODS

Design and Participants

An Action research design using cooperative learning as an instructional method was adopted for this study. The participants were forty-five (45) (thirty (30) males and fifteen (15) females) first-year mathematics education postgraduate students from a mid-sized university in Ghana. They were purposively sampled because they had basic knowledge in statistics. The sample included twenty (20) students from different universities and twenty-five (25) from the university where the study was undertaken. All of them had completed a bachelor's degree in mathematics and had taken statistics courses as part of their undergraduate preparation in mathematics. The average age of the students was twenty-six (26) years.

Diagnostic test

Table 1 Results of Diagnostic Test

Item	Frequency	
	True	false
1 If $p = .05$, the null hypothesis has only a 5% chance of being true.	35	10
2 A nonsignificant difference (e.g., $p \geq .05$) means there is no difference between groups.	34	11
3 A statistically significant finding is very important.	36	8
4 Studies with p values on opposite sides of .05 are conflicting.	33	12
5 Studies with the same p value provide the same evidence against the null hypothesis.	35	10
6 $p = .05$ means that a researcher has observed data that would occur only 5% of the time under the null hypothesis.	30	15
7 $p = .05$ and $p \leq .05$ mean the same thing.	31	14
8 P values are properly written as inequalities (e.g., " $p \leq .02$ " when $p = .015$)	29	16
9 $P = .05$ means that if you reject the null hypothesis, the probability of a type I error is only 5%.	30	15
10 With a $p = .05$ threshold for significance, the chance of a type I error will be 5%.	34	11
11 A researcher should use a one-sided p value when he/she does not care about a result in one direction, or a difference in that is impossible.	30	15
12 A scientific conclusion or treatment policy should be based on whether or not the p value is significant.	32	13

(Goodman, 2008)

Tasks and Instructions

Fifteen (15) groups, each comprising three (3) students, were formed. All the students met as a whole class, and their lecturer informed them about their roles and responsibilities as they worked in their respective groups. He informed them that any progress a group made in the learning process, was deemed to have come from the assistance, contribution and cooperation of all members of the group, rather than from individual members. Every group was assigned two (2) tasks. For the first task, each group member tossed a coin six times and recorded the outcomes for the first, second, third, fourth, fifth, and sixth tosses. With reference to the first definition of the p-value as the sum of three parts: The probability random chance would result in the observation, the probability of observing something else that is equally rare, and the probability of observing something rarer or more extreme, the instructor meticulously explained the terminologies in the definition to the students and then demonstrated

A diagnostic test on the application and interpretation of the p-value was given to the students to assess their entry-level understanding of the concept. Table 1 shows the results of the forty-five (45) mathematics postgraduate students on the diagnostic test. The results indicated that for each of the twelve (12) questions, most of the students selected "yes" as their answer. This was contrary to the general expectation of a "no" answer. Among those who chose "no" as their answer, they struggled to justify why they chose that answer. An interview conducted among the students revealed that their inability to answer the questions correctly stemmed from the fact that they lacked conceptual understanding of the p-value, thus making it difficult for them to apply the concept to solve problems. As they struggled to understand the p-value concept, they conceptualized their understanding, leading to series of misconceptions arising from poor introduction of the concept. If these misconceptions are not corrected, their knowledge in statistics would be affected gravely, since many decisions in statistics now rely heavily on the p-value.

computations of probabilities and p-values to the students. For the second task, this definition of p-value was used: The p-value as the probability of obtaining results as extreme as the observed results of a statistical hypothesis, assuming that the null hypothesis is true, or, the probability of obtaining a sample "more extreme" than the one observed in a sample data, assuming the null hypothesis is true. The instructor took time to explain the terminology "more extreme" to the students before letting them work some examples. For the purposes of this study, the following pseudonyms were used for the participants: L= Instructor; G1 = A member from group 1; G2 = A member from group 2; G3 = A member from group 3; G4 = A member from group 4; G5 = A member from group 5; G6 = A member from group 6; G7 = A member from group 7; G8 = A member from group 8; G9 = A member from group 9; G10 = A member from group 10, G11: A member from group 11; G12 = A member from group 12; G13 = A member from group; G14 = A member from group G14; and, G15 = A member from group G15, respectively. The

vignette below shows a lecturer mediating his students' understanding of the concept of p -value using the cooperative learning method:

L: G2, what is a rare event?

G2: An event that does not happen often.

L: Good, can you explain a rare event in probability, G9?

G9: Since it has a small chance of occurring, the probability will be very small.

L: Good, a rare event has a small probability of occurrence. For such an event, the probability threshold needs to be pre-specified before one can call it a rare event. The typical threshold used in most Statistics courses is 0.005. So, an event is rare if its probability of occurrence is less than 0.005.

L: G5, what then qualifies an event to be rarer?

G5: A rarer event will have probability less than that of a rare event.

L: Is this the same as observing an event more extreme?

G7: Yes.

L: G6, what is a random event?

G6: When one tosses a coin, it could land heads or tails. The two events, heads or tails, are random events.

L: Good job, a random event therefore lacks pattern or predictability. It has no order and does not follow an intelligible pattern. G8, list the sample space when a coin is tossed.

$\{TTTT, TTTH, TTHT, THTT, HTTT, TTHH, THTH, THHT, HTTH, HTHT, HHTT, THHH, HTHH, HHTH, HHHT, HHHH\}$

Obtaining two tails and two heads is as rare as either obtaining three tails and one head or one tail and three heads. The rarer events are either obtaining four tails or four heads. Thus, the p -value is $6/16+4/16+4/16+1/16+1/16=1$, and the probability is $6/16=3/8$.

G8: $\{H, T\}$

L: What is the probability of obtaining a head?

G4: $1/2$

L: let's apply the definition to calculate the p -value. Note that, obtaining a head is equally as rare as obtaining a tail. Thus, the rarer event here has a probability of 0. Therefore, by definition, the p -value of obtaining a head is $1/2+1/2+0=1$. In your individual groups, calculate the p -value of obtaining two heads in two tosses of a coin.

G3: The sample space in two tosses of the coin is $\{HH, HT, TH, TT\}$. Getting two heads is equally as rare as getting two tails in the two tosses. Since the rarer event has a probability of 0, the p -value is $1/4+1/4+0=1/2$.

L: Good work done, all the groups should calculate the p -value of obtaining two tails and one head in three tosses.

G2: The sample space is $\{TTT, TTH, THT, HTT, THH, HTH, HHT, HHH\}$.

Obtaining two heads and one tail is as rare as obtaining one tail and two heads. The rarer events are either obtaining three tails or three heads. Therefore, the p -value is $3/8+3/8+1/8+1/8=1$, and the probability is $3/8$.

L: G1, compute the p -value for obtaining two tails and two heads in four tosses of the coin.

G1: The sample space is

L: You all seem to understand the p -value. In the case of the coin tosses where the outcomes follow some symmetry, it is easier to intuitively figure out the rare and rarer events. Note again that, rare and rarer events have small chances of occurring. Similar computations for the p -values can be done for tosses of five and six. The outcome(s), probabilities and p -values are indicated in Table 2 and 3.

Table 2 Outcomes for five tosses of a coin

Number of Tosses	OUTCOME(S)
1	T, H
2	$HH, \left. \begin{matrix} TH \\ HT \end{matrix} \right\}, TT$
3	$\left. \begin{matrix} TTH \\ TTT, THT \end{matrix} \right\}, \left. \begin{matrix} TTH \\ HTH \end{matrix} \right\}, HHH$ $\left. \begin{matrix} HTT \\ HHT \end{matrix} \right\}$
4	$\left. \begin{matrix} TTHH \\ TTTH \\ TTTT, \left. \begin{matrix} TTHT \\ THTT \end{matrix} \right\}, \left. \begin{matrix} TTHH \\ HTTH \end{matrix} \right\}, \left. \begin{matrix} TTHH \\ HTHH \end{matrix} \right\}, HHHH$ $\left. \begin{matrix} THTH \\ THHT \\ HTTT \\ HHTT \end{matrix} \right\}, \left. \begin{matrix} THTH \\ HHTT \\ HHTT \end{matrix} \right\}, \left. \begin{matrix} TTHH \\ HTHH \\ HHTT \end{matrix} \right\}, HHHH$
5	$\left. \begin{matrix} TTTHH \\ TTHTH \\ TTHHT \\ TTTTH \\ TTTHT \\ THTTT \\ HTTTT \end{matrix} \right\}, \left. \begin{matrix} THTTH \\ THTHT \\ THHTT \\ HTTTH \\ HTTHT \\ HTHTT \\ HHTTT \end{matrix} \right\}, \left. \begin{matrix} TTHHH \\ THTHH \\ THHHT \\ HHTHH \\ HHTTH \\ HHTHT \\ HHTTT \end{matrix} \right\}, \left. \begin{matrix} TTHHH \\ THTHH \\ THHHT \\ HHTHH \\ HHTTH \\ HHTHT \\ HHTTT \end{matrix} \right\}, HHHHH$

Understanding the Concept of P-Value from Perspectives of two Distinct Definitions

Item	Number of Tosses	Outcome	Probability	Rare outcome	Rarer Outcome	P- value
		and 1 Heads			Heads	$0.0125+0.0125=0.3750$
17	5	3 Tails and 2 Heads	0.3125	2 Tails and 3 Heads or 4 Tails and 1 Heads or 1 Tails and 3 Heads	5 Tails or 5 Heads	$0.3125+0.3125+0.3125+0.03125+0.15625=1$
18	5	2 Tails and 3 Heads	0.3125	3 Tails and 2 Heads or 4 Tails and 1 Heads or 1 Tails and 4 Heads	5 Tails or 5 Heads	$0.3125+0.3125+0.3125+0.03125+0.15625=1$
19	5	1 Tails and 4 Heads	0.15625	4 Tails and 1 Heads	5 Tails and 5 Heads	$0.15625+0.15625+0.0125+0.0125=0.375$
20	5	5 Heads	0.03125	5 Tails	0	$0.03215+0.03215+0=0.0625$
21	6	6 Tails	0.015625	6 Heads	0	$0.015625+0.015625+0=0.03125$
22	6	5 Tails and 1 Heads	0.09375	1 Tails and 5 Heads	6 Tails or 6 Heads	$0.09375+0.09375+0.015625+0.015625=0.21875$
23	6	4 Tails and 2 Heads	0.234375	2 Tails and 4 Heads or 4 Tails and 1 Heads or 1 Tails and 4 Heads	6 tails or 6 Heads	$0.234375+0.234375+0.09375+0.09375+0.015625+0.015625=0.6875$
24	6	3 Tails and 3 Heads	0.3125	0	4 Tails and 2 Heads or 2 Tails and 4 heads or 5 Tails and 1 Heads or 1 Tails and 5 Heads or 6 Tails or 6Heads	$0.3125+0+0.234375+0+0.234375+0.09375+0.09375+0.015625+0.015625=1$
25	6	2 Tails and 4 Heads	0.234375	4 Tails and 2 Heads or 5 Tails and 1 Heads or 1 Tails and 5 Heads	5 Tails or 5 Heads	$0.234375+0.234375+0.09375+0.09375+0.015625+0.015625=0.6875$
26	6	1 Tails and 5 Heads	0.09375	5 Tails and 1 Heads	5 Tails or 5 Heads	$0.09375+0.09375+0.015625+0.015625=0.21875$
27	6	6 Heads	0.015625	6 Tails	0	$0.015625+0.015625+0=0.03125$

L: For the second definition of *p*- value, we will consider left-tailed, right-tailed, and two-tailed tests. Can any group differentiate among them?

G2: A left-tailed test is a hypothesis test where the rejection region is located to the extreme left of a distribution. A right-tailed test is a hypothesis test where the rejection region is located to the extreme right of a distribution. A two-tailed test is a hypothesis test where the rejection regions are located to the extreme left and right of a distribution.

L: Let's consider a sample with mean, \bar{x} , standard deviation, *s*, and sample size, *n*, selected from a population of mean, μ , then $z = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$. All the groups should come up with an

interpretation of *z*, which group will start first?

G6: *z* indicates the number of standard errors from the mean.

L: Good. So, does it mean *z* can assume negative and positive values?

G8: Yes.

L: Let's now discuss what "observed results" really mean in the definition. The observed result of a statistical hypothesis is the *z* we obtain from a sample and its population parameters. In this case different samples of the same sample size could give a value equal to, less than, or greater than *z*, on the standard normal curve. Therefore, values either to the left or right of *z*, are either as extreme as *z* or more extreme than *z*, for a left-tailed and a right-tailed tests respectively. For a two-tailed test, and because of symmetry of the normal curve, *z* could be indicated at two points on the curve (i.e., to the left and right). The extreme values occur either to the left or right of *z*. The *p*-values are the areas under the normal curve to the left of *z*, for a left-tailed test, to the right of *z*, for a right-tailed test, and to the left and right of *z*, for a two-tailed test. Figures 1, 2, and 3 show the *z*, extreme values and *p*-values for left-tailed, right-tailed, and two-tailed tests.

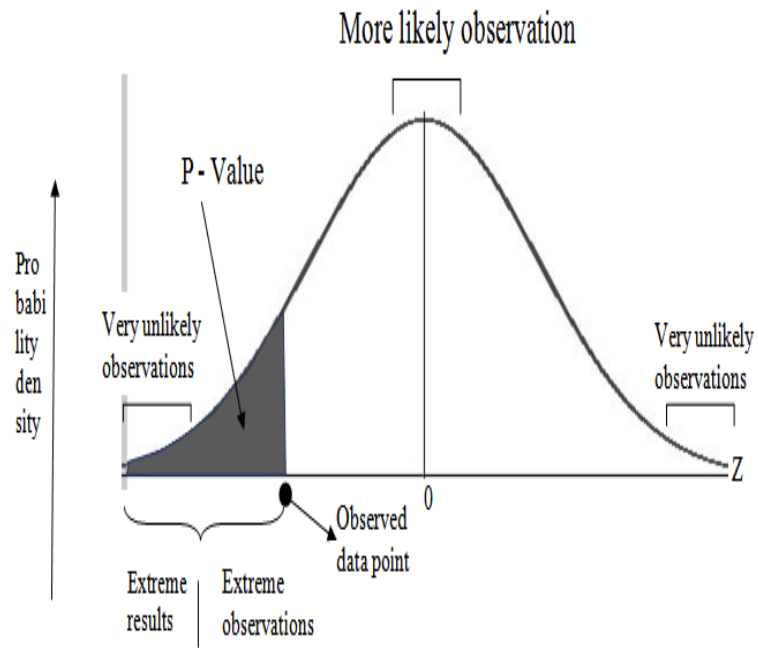


Fig. 1 Left-tailed test

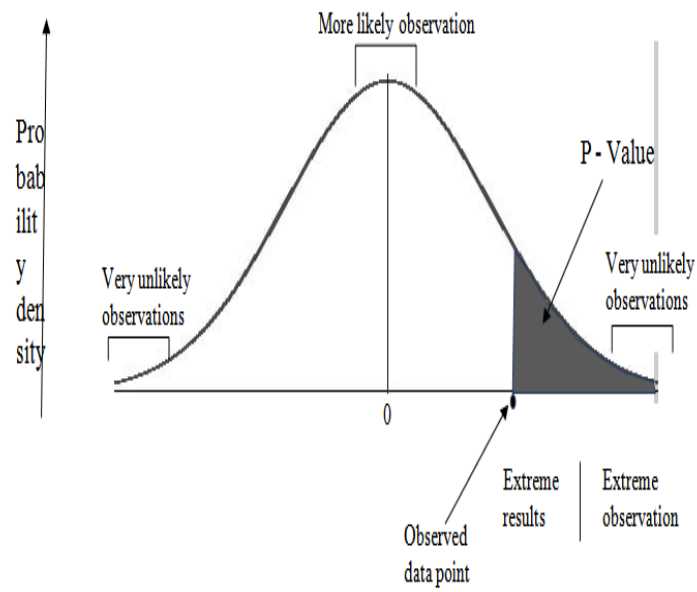


Fig. 2 Right-tailed test

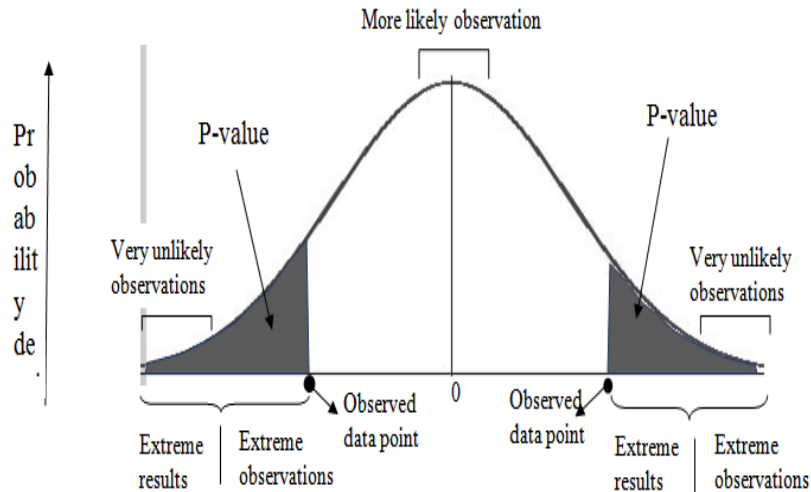


Fig. 3 Two-tailed test

L: To all the groups, can a p -value be greater than 0.5?

G3: Yes, that depends on the value of z on the normal curve, and whether the test is a left-tailed, right-tailed, or two-tailed.

L: Do the other groups share similar opinion?

G7: Yes

L: What is the largest p -value one can obtain in a hypothesis test?

G4: p -value can be as large as 1, this is the area under the normal curve.

L: Great, can you please explain in detail?

G4: z can assume many values on the normal curve. If z becomes large and positive, the area to the left of z becomes large. This area has many values less than z , by our definition. These values are extreme to the left of z . For example, on the normal curve, $z = 3.49$ corresponds to an area of 0.9998. However, several values of z are extreme to $z = 3.49$, to the left. This area, or probability is almost 1.

L: Excellent. Let's take an example to calculate p -values.

Concrete Example

A large insurance company has received numerous complaints about the length of time taken to process accident claims. In the past the company has determined that the average length of time it takes to process an accident claim has been 6 days. A new claim-processing procedure has been developed and the company hopes that this will alter the length of time needed to process a claim. After the new procedure is implemented, a random sample of 100 claims processed using the new procedure produces an average

processing time of 5.5 days, with a standard deviation of 2 days. Find the p -value if the new procedure (a) reduced (b) increased (c) did not alter the average length of time required to process an accident claim.

L: First, we need to calculate z using the formula $z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$.

Let's all identify the variables in the equation, and calculate z . Can a group begin?

G4: From the given information, $\bar{x} = 5.5 \text{ days}$, $\mu = 6 \text{ days}$, $\sigma = 2 \text{ days}$, and $n = 100$. Therefore, $z = \frac{5.5 - 6}{\frac{2}{\sqrt{100}}} =$

-2.5 (5.5 is 2.5 standard deviations below 6). From the standard normal table, the corresponding p -value = 0.00621, for a left-tailed test.

L: G7, calculate the p -values for a right-tailed and two-tailed tests.

G7: The p -value = $1 - 0.00621 = 0.99379$, for a right-tailed test. For a two-tailed test, the p -value = $2(0.00621) = 0.01242$.

L: Good. Can all groups now calculate p -values when given the necessary information?

Groups: All groups answered in the affirmative.

L: Why is the p -value very significant in hypothesis testing, G2?

G2: It enables researchers to determine the level of significance, α , to be able to reject the null hypothesis, H_0 . We rely on this conclusion for a p -value, i.e., if $p \leq \alpha \Rightarrow \text{Reject } H_0$, and, if $p > \alpha \Rightarrow \text{Fail to reject } H_0$.

L: Figures 4, 5, and 6 indicate the p -values for a left-tailed, right-tailed, and two-tailed tests.

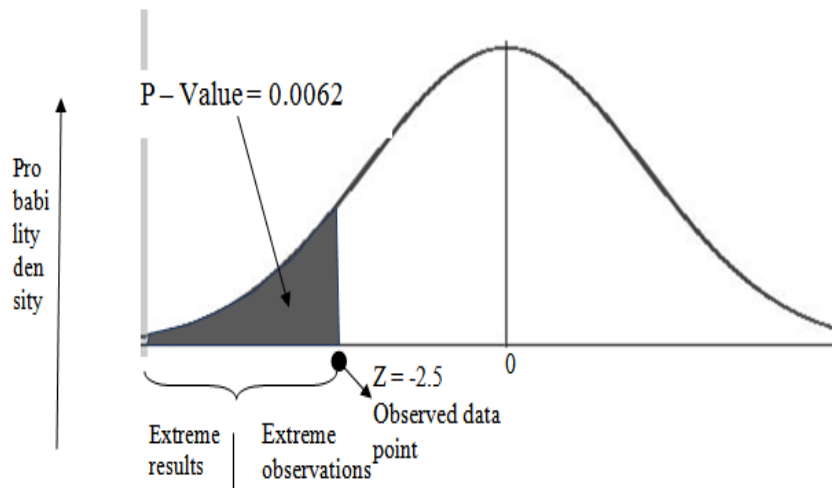


Fig. 4 Left-tailed test

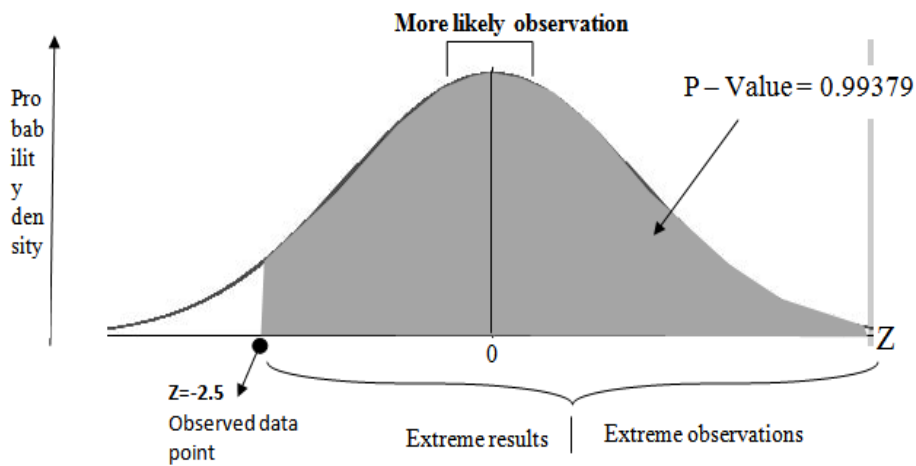


Fig. 5 Right-tailed test

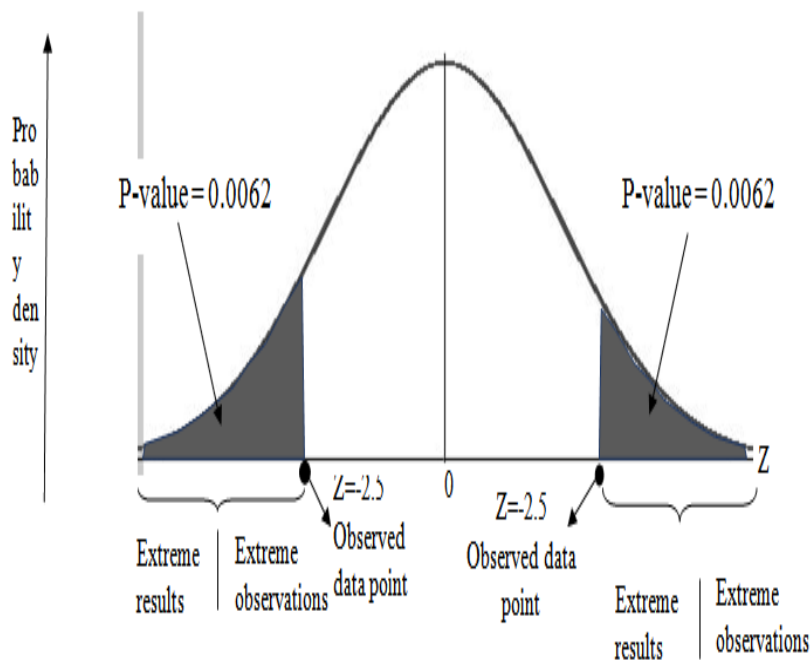


Fig. 6 Two-tailed test

IV. RESULTS AND DISCUSSIONS

The use of the two definitions enhanced the students understanding of p-value. Understanding mathematical definitions are crucial for students to understand mathematical concepts (Lipson, 2007). Instructors could play very vital roles in explaining not-easy-to-understand terms to students to avoid any possible confusion among them and set them on the path of building confidence in their concept acquisition (Wilkins, 2000). The use of definitions would allow instructors and students to discharge their responsibilities effectively. When instructors have assigned tasks to students, students must interact with their friends to complete the tasks. The ability to work well with members of their groups builds teamwork, nurtures good organizational skills, builds excellent written and verbal communication skills, and nurtures excellent presentation skills (Lim, 2002). However, they have the freedom to consult with their instructors when the need arises. In that regard, students can work cooperatively to understand every word, term, phrase, and sentence in all statistical definitions. When solving mathematical problems, students could constantly refer to these definitions to clear their doubts. These definitions give precise meaning to new terms by describing a condition that unambiguously qualifies what a mathematical term is and is not. Instructors should demonstrate an ability to make the subject they teach stimulating and enjoyable for their students.

Previous studies of cooperative learning as a successful pedagogical method are replete in literature (Johnson & Johnson, 2002; Roseth et al., 2008; Slavin, 1989). When students work together to achieve a common goal, they produce a higher achievement and greater productivity than working alone. In schools, opportunities for students to work in situations where they experience positive interdependence seems a better choice than situations based on no independence. However, it is a known fact that students on their own do not cooperate willingly during group work, strict supervision from their instructors is needed to structure the groups so that the five components for successful cooperative learning mediation are evident: Establishing positive interdependence among group members; facilitating promotive interaction; encouraging individual accountability; teaching the appropriate social skills; and, encouraging groups to reflect on both the processes involved in managing the task and interacting with their peers.

Instructors should ensure that these components are alive and present in the be groups. In this regard, students should accept personal responsibility for their contributions to the group and their behaviours towards group members. They should respect other contributions and commit themselves to resolve disagreements democratically.

They must work constructively towards managing the task and maintaining effective working relationships. Instructors have a role in promoting interaction among students because research indicates that students rarely provide quality explanations or engage in high-level discourse unless they teach them to do so (King, 2002). Instructors can also teach their students to talk, reason, and problem-solve. All these contribute to their reasoning, problem-solving and learning (Gillies, 2004, 2006, 2008; Mercer et al., 2004). The significance of the p-value in research is that it is used as an

alternative to rejection points to provide the smallest level of significance to reject the null hypothesis. A smaller p-value means that there is stronger evidence in favour of the alternative hypothesis.

Conclusion

This study has demonstrated that the use of definitions by instructors have the benefit of promoting a deeper conceptual understanding of mathematics and statistics. The use of definitions in this study promoted a deeper conceptual understanding of the p-value (Wilson, 1990). Cooperative learning, used as an intervention, allowed the students to interact, discuss and share ideas and enhanced their thinking processes. It taught the students several social and emotional skills and allowed them to learn from each other. A plethora of studies now abound that show that peer learning in which students explain concepts and ideas to each other improve comprehension considerably (Boud, 2002). Thus, cooperative learning produced critical and lifelong experiences that other learning methods may not provide.

Implication for Teaching and Learning

Instructors should teach statistical concepts using definitions, with cooperative learning as an intervention, for students to believe that statistics can be enjoyable. Thorough knowledge of the concept of p-value is crucial for beginning researchers and students because its wrong conception could have dire consequences on the decisions students make in statistics.

REFERENCES

- [1] Badenes-Ribera, L., Frías-Navarro, D., Monterde-i-Bort, H., & PascualSoler, M. (2013). Informar e interpretar el tamaño del efecto en Psicología y Educación [Reporting and interpreting the effect size in Psychology and Education]. XIV Congreso Virtual de Psiquiatría.com. Interpsiquis, 2013: 1-28 February.
- [2] Beyth-Marom, R., Fidler, F., & Cumming, G. (2008). Statistical cognition: Towards evidence-based practice in statistics and statistics education. *Statistics Education Research Journal*, 7, 20-39.
- [3] Boud, D. (2002). "What is peer learning and why is it important?" *Tomorrow's teaching and learning*. Stanford University.
- [4] Carver, R.P. (1978). The case against statistical significance testing. *Harvard Educational Review*, 48, 378-399.
- [5] Cobb, G. (2007). The introductory statistics course: A Ptolemaic curriculum? *Technology Innovation in Statistics Education*, 1, Article 1.
- [6] Cohen, J. (1994). The earth is round ($p < 0.05$). *American Psychologist*, 49, 997-1003.
- [7] Cumming, G. (2006). Understanding replication: Confidence intervals, p-values, and what's likely to happen next time. In B. Phillips (Ed.), *Developing a statistically literate society: Proceedings of the 7th International Conference on Teaching Statistics*.
- [8] Cumming, G. (2010). Understanding, teaching and using p-values. In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society: Proceedings of the 8th International Conference on Teaching Statistics*. Voorburg, The Netherlands: International Statistics Institute.
- [9] Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York, NY: Routledge
- [10] Falk, R., & Greenbaum, C.W. (1995). Significance tests die hard: The amazing persistence of a probabilistic misconception. *Theory & Psychology*, 5, 75-98.
- [11] Fidler, F. (2005). *From statistical significance to effect estimation: Statistical reform in psychology, medicine and ecology*. PhD Thesis History and Philosophy of Science. Melbourne, Australia.

Department of History and Philosophy of Science. University of Melbourne.

[14] Fisher, R.A. (1925). *Statistical methods for research workers*. Edinburgh, UK. Oliver and Boyd.

[15] Garfield, J., delMas, R., & Zieffler, A. (2012). Developing statistical modelers and thinkers in an introductory, tertiary-level statistics course. *ZDM Mathematics Education*, 44, 883-898.

[17] Garfield, J., Zieffler, A., Kaplan, D., Cobb, G., Chance, B., & Holcomb, J.P. (2011). Rethinking assessment of student learning in statistics courses. *The American Statistician*, 65, 1-10.

[18] Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, 33, 587-606.

[19] Gill, J. (1999). The insignificance of null hypothesis significance testing. *Political Research Quarterly*, 52, 647-674.

[20] Gillies, R. (2003a). The behaviors, interactions, and perceptions of junior high school students during small-group learning. *Journal of Educational Psychology*, 95, 137-147.

[21] Gillies, R. (2003b). Structuring cooperative group work in classrooms. *International Journal of Educational Research*, 39, 35-49.

[22] Gillies, R. (2004). The effects of cooperative learning on junior high school students during small group learning. *Learning and Instruction*, 14, 197-213

[23] Gillies, R. (2006). Teachers' and students' verbal behaviours during cooperative and small-group learning. *British Journal of Educational Psychology*, 76, 271-287.

[24] Gillies, R. (2008). The effects of cooperative learning on junior high school students' behaviours, discourse, and learning during a science-based learning activity. *School Psychology International*, 29, 328-347.

[25] Gillies, R., & Ashman, A. (1996). Teaching collaborative skills to primary school children in classroom-based work groups. *Learning and Instruction*, 6, 187-200.

[26] Gillies, R., & Ashman, A. (1998). Behavior and interactions of children in cooperative groups in lower and middle elementary grades. *Journal of Educational Psychology*, 90, 746-757.

[27] Gliner, J.A., Leech, N.L., & Morgan, G.A. (2002). Problems with null hypothesis significance testing (NHST): What do the textbooks say? *The Journal of Experimental Education*, 71, 83-92

[28] Goodman, S. (2008). A dirty Dozen: Twelve P-value misconceptions. *Seminars in Hematology*, 45, 135- 140.

[29] Haller, H., & Krauss, S. (2002). Misinterpretations of significance: A problem students share with their teachers? *Methods of Psychological Research Online [On-line serial]*, 7, 120. Retrieved July 30, 2014, from <http://www.metheval.uni-jena.de/lehre/0405-ws/evaluationuebung/haller.pdf>.

[30] Hubbard, R., & Lindsay, R.M. (2008). Why p values are not a useful measure of evidence in statistical significance testing. *Theory & Psychology*, 18, 69-88

[32] Johnson, D.H. (1999). The insignificance of statistical significance testing. *Journal of Wildlife Management*, 63, 763-772.

[33] Johnson, D., & Johnson, F. (2009). *Joining together: Group theory and group skills* (10th ed.). Upper Saddle River, NJ: Pearson Education.

[34] Johnson, D., & Johnson, R. (2009). An educational psychology success story: Socialinterdependence theory and cooperative learning. *Educational Researcher*, 38, 365-379.

[35] Johnson, D., & Johnson, R. (2002). Learning together and alone: Overview and meta-analysis. *Asia Pacific Journal of Education*, 22, 95-105.

[36] Johnson, D., & Johnson, R. (1990). Cooperative learning and achievement. In S. Sharan(Ed.), *Cooperative learning: Theory and research*(pp. 23-37). New York: Praeger.

[37] Kline, R.B. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research*. Washington, DC: American Psychological Association

[38] Kirk, R.E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement*, 56, 746-759.

[39] King, A. (2002). Structuring peer interaction to promote high-level cognitive processing. *Theory into Practice*, 41, 33-40. http://dx.doi.org/10.1207/s15430421tip4101_6

[40] Levine, T.R., Weber, R., Hullett, C., Sun, H. & Massi, L.L. (2008). A critical assessment of null hypothesis significance testing in quantitative communication research. *Human Communication Research*, 34, 171-187.

[41] Lim, C. S.(2002). Practice makes perfect? An insight into the culture of mathematics learning in two Chinese primary schools. *Proceedings of Mathematics Education National Seminar*, 163-171.

[42] Lipson, K. (2007). Assessing understanding in statistics. Swinburne University of Technology, Lilydale Australia. Retrieved from: <http://www.stat.auckland.ac.nz>.

[43] Lou, Y., Abrami, P., Spence, J., Poulsen, C., Chambers, B., & d'Apollonia, S. (1996). Within-class grouping: A meta-analysis. *Review of Educational Research*, 66, 423-458. <http://dx.doi.org/10.3102/00346543066004423>

[44] Mercer, N., Dawes, L., Wegerif, R., & Sams, C. (2004). Reasoning as a scientist: Ways of helping children to use language to learn science. *British Educational Research Journal*, 30, 359-377.

[45] Mittag, K., & Thompson, B. (2000). A national survey of American Educational Research Association members' perceptions of statistical significance tests and other statistical issues. *Educational Researcher*, 29(4),14-20.

[46] Nickerson, R.S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, 5, 241-301.

[47] Pollard, P., & Richardson, J. (1987). On the probability of making Type I errors. *Psychological Bulletin*, 102, 159-163.

[48] Reaburn, R. (2014). Introductory statistics course tertiary students' understanding of p-values. *Statistics Education Research Journal*, 13(1), 53-65.

[49] Roseth, C., Johnson, D., & Johnson, R. (2008). Promoting early adolescents' achievement and peer relationships: The effects of cooperative, competitive, and individualistic goal structures. *Psychological Bulletin*, 134, 223-246.

[50] Slavin, R. (1989). Cooperative learning and student achievement. In R. Slavin (Ed.), *School and classroom organization* (pp. 129-156). New Jersey: Lawrence Erlbaum.

[52] Slavin, R. E. (1994). *Student team learning: A practical guide to cooperative learning* (3rd ed.). National Education Association

[53]

[54] Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher*, 25, 26-30.

[55] Verdam, M.G.E., Oort, F.J., & Sprangers, M.A.G. (2014). Significance, truth and proof of p values: Reminders about common misconceptions regarding null hypothesis significance testing. *Quality of Life Research*, 23, 5-7.

[56] Webb, N., & Mastergeorge, A. (2003). Promoting effective helping in peer-directed groups. *International Journal of Educational Research*, 39, 73-97.

[57] Wagenmakers, E.J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, 14, 779-804.

[58] Wilkins, J.L.M. (2000). Preparing for the 21st century: The status of quantitative literacy in the United States. *School Science and Mathematics*, 100(8), 405-418.

[59] Wilson, P. (1990). Inconsistent ideas related to definitions and examples. *Focus on Learning Problems in Mathematics*, 12(3, 4), 31-47.