Statistics Instructors' Beliefs and Misconceptions About *p*-values

Robyn Reaburn
University of Tasmania
<Robyn.Reaburn@utas.edu.au>

It is well known that students of inferential statistics find the hypothetical, probabilistic reasoning used in hypothesis tests difficult to understand. Consequently, they will also have difficulties in understanding *p*-values. It is not unusual for these students to hold misconceptions about *p*-values that are difficult to remove. In this study, 19 Australian tertiary statistics educators were surveyed about their beliefs about *p*-values, and it was found that some of these instructors held misconceptions about the nature of *p*-values. These findings suggest that professional learning of statistics instructors is urgently required so that instructors may have their beliefs challenged and corrected.

It is well known that the processes of inferential statistics are difficult for students to understand and, as a consequence, these students resort to blindly following procedures (Garfield & Ahlgren, 1988). One of these procedures is that of the Null Hypothesis Significance Test (NHST). Briefly, the procedure involves making a hypothesis about a parameter of a population, such as the null hypothesis. A sample is then taken from the population and the appropriate sample statistic is calculated. For example, if the hypothesis is about the population mean, then the sample mean would be determined. Then, assuming that the null hypothesis is true, the probability of obtaining the sample mean or one even more extreme would be calculated. For the situation regarding the population mean, this probability is:

$$P((|\bar{x} - \mu| \geq 0)|H_o = \mu)$$

If this resultant probability (known as the *p*-value) is greater than a preset value, then the sample data are considered to be consistent with the hypothesised population parameter and the null hypothesis is not rejected (according to some texts the null hypothesis is accepted). If this p-value is less than the preset value, the sample data are considered not to be consistent with the hypothesised population parameter and the hypothesis is rejected.

This form of hypothetical, probabilistic reasoning is not one with which students are generally familiar, and misunderstandings are common. In fact, in contrast to deterministic reasoning, it is not unusual for students to find any form of probabilistic reasoning difficult (Kahneman, & Tversky, 1982; Tversky & Kahneman, 1982a, 1982b). To complicate matters, full understanding of the NHST process also requires understanding of randomness, probability, sampling, sampling distributions, and variability, about which students may also have misconceptions (Castro-Sotos, Vanhoof, Van den Noorgate, & Onghena, 2007). As a consequence of these difficulties, students develop misunderstandings about the nature of the *p*-value. Previous research has shown that students may believe that the *p*-value is the probability that the null hypothesis is true (Reaburn, 2014; Gliner, Leech, & Morgan, 2002). Students may also believe that the same *p*-value will be obtained if the experiment were to be repeated (Cumming, 2006; Mittag & Thompson, 2000; Nickerson, 2000). Further problems occur because students may not fully understand what not rejecting or rejecting a null hypothesis actually means.

Oakes (1986, as cited in Haller & Krauss, 2002) found that there was also widespread misunderstanding about *p*-values among academic psychologists. Building on that research, Haller and Krauss (2002) surveyed 113 staff and students (including 30 statistics instructors) from psychology departments from six German universities and found that

(2017). In A. Downton, S. Livy, & J. Hall (Eds.), *40 years on: We are still learning! Proceedings of the 40th Annual Conference of the Mathematics Education Research Group of Australasia* (pp. 428-433). Melbourne: MERGA.

some of the instructors and students had misconception, such as believing that NHSTs prove or disprove hypotheses. Some of these participants also indicated that they believed that the *p*-value indicated the probability of the null hypothesis being true. They also indicated that they believed that the value of 1- *p* gives the relative frequency of the null hypothesis being rejected if the experiment were to be repeated. The aim of this study was to determine if there is a similar problem of misinterpretation of *p*-values by statistics instructors in Australian universities.

Method

In the study, I adopted an exploratory paradigm (Mackenzie & Knipe, 2006) to investigate the understanding of *p*-values of statistics instructors in Australian universities. It was a quantitative study in which I used a survey to collect data. The initial participants constituted a sample of convenience, and a snowball recruiting strategy was used (Babbie, 2014). The initial potential participants were recruited from nine universities from New South Wales, Queensland, South Australia, Tasmania, and the Northern Territory. These potential participants were selected because their email addresses and details about their teaching responsibilities were freely available on their respective university websites. Academics who were teaching statistics courses were sent emails with a link to an online survey platform with a request to complete a survey. Because the links were able to be sent to other potential participants, no identifying information was collected and this ensured that the data remained anonymous.

The survey administered included the scenario and six statements utilised by Haller and Krauss in their 2002 study, with small alterations (Figure 1).

Suppose you have a treatment that you suspect may alter the performance on a certain task. You compare the means of your control and experimental groups. You then use a 2-sample independent means *t*-test and your *p*-value is 0.01. What can you conclude? Please mark the following as "true" or "false".

1. You have disproved the null hypothesis that there is no difference between population means.
2. You have found the probability of the null hypothesis being true.
3. You have proved your experimental hypothesis. That is, there is a difference between the population means.
4. You can deduce the probability of your null hypothesis being true.
5. If you decide to reject the null hypothesis, you know the probability that you are making the wrong decision.
6. You have a reliable experimental finding in the sense, that, if, hypothetically the experiment were repeated a number of times, you would obtain a significant result on 99% of occasions.

Figure 1. Survey scenario and questions.

The Haller and Krauss scenario included the details about the *t*-statistic and degrees of freedom, and their wording of Statement 5 was slightly different to that shown in Figure 1. The statements represented common misconceptions of the meaning of a significant test result. Statements 1 and 3 addressed the belief that NHSTs can give definitive proof – the illusion of certainty (Gigerenzer, 2004). Statements 2 and 4 addressed the beliefs that probabilities can be assigned to hypotheses. Statement 5 was similar to the definition of a Type I error, but did not include the proviso that the null hypothesis must be true.

Statement 6 addressed the replication fallacy (Gigerenzer, 2004). Also included in the survey were some questions to collect data about the units into which the academics taught, the courses in which the units were embedded, and the number of years of experience the academics had in teaching statistics.

Results

Fifty emails were sent to potential participants, as described in the Method section. Of these, 19 completed surveys were received. Eleven participants indicated that they were lecturers in a statistics unit, two were tutors in a statistics unit, one was a lecturer and tutor, and one was a statistician embedded in a science faculty. The status of the others is not known. Two participants indicated that they were teaching in units that were part of a mathematics or statistics major, and 14 indicated that they were teaching in service units where the students were not expected to become professional mathematicians or statisticians. The other participants did not answer this question. The academics' years of teaching statistics ranged from two years to 40 years, with a mean of 18 years and a median of 13 years.

At least three participants agreed with each of the statements about the hypothetical two-sample *t*-tests. The number of participants who agreed with each statement is indicated in Table 1. The greatest level of agreement was with Statement 6, with the next highest level of agreement with Statement 5, followed by Statements 4 and 2. Five participants did not agree with any of the statements.

Table 1
Number of Participants Who Agreed with Each Statement

Statement about *t*-test *p*-value	Number who indicated "true" (n = 19)
You have disproved the null hypothesis that there is no difference between population means.	3
You have found the probability of the null hypothesis being true.	5
You have proved your experimental hypothesis. That is, there is a difference between the population means.	3
You can deduce the probability of your null hypothesis being true.	6
If you decide to reject the null hypothesis, you know the probability that you are making the wrong decision.	7
You have a reliable experimental finding in the sense, that, if, hypothetically the experiment were repeated a number of times, you would obtain a significant result on 99% of occasions.	8

Discussion

It has to be admitted that the sample size in this study is small. However, it is concerning to see that all of these misunderstandings about the nature of *p*-values were held by at least three of the participants, all of whom, apart from one, identified themselves as instructors in statistics. In this study, 42% of the participants agreed with Statement 6.

This is a description of the replication fallacy, the idea that 1-*p* represents the proportion of significant results if the experiment were to be repeated many times. However, as Cumming (2013) so engagingly demonstrates in his *Dance of the p-values*, replications of experiments can result in widely varying *p*-values. In the study by Haller and Krauss (2002), 37% of statistics instructors also held this misunderstanding.

Statement 5 appears to describe the definition of a Type I error; that is, it is the probability of rejecting the null hypothesis if this null hypothesis is true. A close reading, however, shows that the important proviso that the null hypothesis is actually true was not included. In this study, the proportion of participants who agreed with this statement (41%) is much less than that obtained in the study by Haller and Krauss (2002), where 73% of the instructors agreed with this statement.

Statement 4 is incorrect as it is not possible to assign a probability to a hypothesis in the NHST process (Haller & Krauss, 2002). Statements 1 and 3 are also untrue as we cannot be sure about any conclusion about a population when only a sample is taken. The NHST process is one way that the uncertainty of conclusions based on samples can be addressed; it is disturbing to see that this fundamental issue of uncertainty may be missed by some instructors of statistics. Statement 2 has been shown to be a common misunderstanding among students, and it is concerning to see that 29% of the participants also hold this misunderstanding. This compares with 17% of the instructors in the study by Haller and Krauss (2002).

These results beg the question: How do such misunderstandings arise? Reaburn (2014) noted that the misunderstanding indicated by Statement 2 was not in any of the materials supplied to the students in her study. She proposed that this misunderstanding may arise from the students' attempts to rationalise the difficult material with which they were dealing. This study, along with that of Haller and Krause (2002), suggests that students may also be gaining their misunderstandings from their instructors. It is reasonable to posit that if these students in turn become instructors, they will pass on these misunderstandings to their students. In addition, Gigerenzer (2004), Haller and Krauss (2002), and Pollard and Richardson (1987) have described examples where textbooks have given inaccurate descriptions of the interpretation of NHSTs.

It is also possible that misunderstandings arise because of the teaching methods that are chosen by the instructors. It has been suggested that despite the availability of computers, instructional methods have remained substantially unchanged (Garfield, delMas, & Zieffler, 2012). Modern computers allow much more exploration of the data than was formerly available. In particular, computers allow repeated randomization and simulation. Computers allow students "to create models, repeatedly simulate data from the model, and then use the resulting distribution of [the] computed statistic to draw statistical inferences" (Garfield et al., 2012, p. 883). As they do this, the students can see what "sort of results are typical, and what should be considered unusual" (Cobb, 2007, p. 3).

Whether or not *p*-values should be used at all is debated in the literature. This is partly due to their tendency to vary, as noted above. This is problematic in the scientific endeavour where replicable experiments are desired. Varying *p*-values may also be obtained with the same data depending on the method of analysis chosen by the researcher and on whether a one- or two-tailed test is chosen (Hubbard & Lyndsay, 2008). In addition, *p*-values do not indicate the effect size; a study with a large effect with a small sample may result in the same *p*-value as a study with a small effect size and a large sample (Hubbard & Lindsay, 2008; Wagenmakers, 2007). In contrast, confidence intervals, which give an estimate of the range within which the value of the parameter of interest may be found,

give an idea of the precision of the estimate, make it easier to determine the effect size, and also have the same metric as the point estimate (Cumming 2010; Wagenmakers 2007). Their use also avoids the complicated logic used in the NHST. Whatever one's view on this debate, the NHST is still widely used in the scientific literature, and even if it is replaced by the use of confidence intervals or other methods in the future, there are many years of literature using *p*-values that need to be understood.

Whether not the use of *p*-values continues into the future, and whatever the source of misunderstandings, the findings of this study suggest that professional development for statistics instructors in our tertiary institutions is urgently required so that their misunderstandings are corrected. The findings of this study also suggest that research is needed to see if similar misunderstandings are held by school teachers who teach in pre-tertiary subjects where *p*-values are used. Examples of such subjects include Specialist Mathematics in Victoria (Victorian Curriculum and Assessment Authority, 2015) and Mathematics Applied in Tasmania (Tasmanian Qualifications Authority, 2013). If future researchers are to produce the best possible research, and interpret their findings accurately, then they must be able to understand the statistical procedures they use in their work.

Acknowledgements

I would like to thank Dr. Noleine Fitzallen for her helpful advice in the writing of this paper.

References

Babbie, E. (2014). *The basics of social research* (6th ed.). Belmont, CA: Wadsworth Cengage Learning.

Castro-Sotos, A. E. C., Vanhoof, S., Van den Noortgate, V., & Onghena, P. (2007). Students' misconceptions of statistical inference: A review of the empirical evidence from research on statistics education. *Educational Research Review, 2*, 98-113.

Cobb, G. (2007). The introductory statistics course: A Ptolemaic curriculum? *Technology Innovation in Statistics Education, 1*, Article 1.

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 societ:, Proceedings of the 7th International Conference on Teaching Statistics*.

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.

Cumming, G. (2013). *Intro Statistics 9 Dance of the p Values* [Video file]. Retrieved from https://www.youtube.com/watch?v=5OL1RqHrZQ8

Garfield, J., & Ahlgren, A. (1988). Difficulties in learning basic concepts in probability and statistics: Implications for research. *Journal for Research in Mathematics Education, 19*(1), 44-63.

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.

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

Gliner, J., Leech, N., & Morgan, G. (2002). Problems with null hypothesis significance testing (NHST): What do the textbooks say? *The Journal of Experimental Education, 71*(1), 83-92.

Haller, H., & Krauss, S. (2002). Misinterpretations of significance: A problem students share with their teachers? *Methods of Psychological Research Online, 7*(1), Article 1.

Hubbard, R., & Lindsay, R. (2008). Why *P* values are not a useful measure of evidence in statistical significance testing. *Theory and Psychology, 18*(1), 69-88.

Kahneman, D., & Tversky, A. (1982). Subjective probability: A judgement of representativeness. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgement under uncertainty: Heuristics and biases* (pp. 32-47). New York, NY: Cambridge University Press.

Mackenzie, N., & Knipe, S. (2006). Research dilemmas: Paradigms, methods and methodology. *Issues in Educational Research, 16*(2), 193-205.

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.

Nickerson, R. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods, 5*(2), 241-301.

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

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

Tasmanian Qualifications Authority. (2013). *Mathematics applied*. Hobart, TAS: Author.

Tversky, A., & Kahneman, D. (1982a). Availability: A heuristic for judging frequency and probability. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgement under uncertainty: Heuristics and biases* (pp. 163-178). New York, NY: Cambridge University Press.

Tversky, A., & Kahneman, D. (1982b). Belief in the law of small numbers. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgement under uncertainty: Heuristics and biases* (pp. 23-31). New York, NY: Cambridge University Press.

Victorian Curriculum and Assessment Authority (2015). *Victorian Certificate of Education, Mathematics: Study design 2016-2018*. Melbourne: State Government of Victoria.

Wagenmakers, E. (2007). A practical solution to the pervasive problems of *P* values. *Psychonomic Bulletin and Review, 14*(5), 779-804.