Summary of *P*-value Survey Research

Sharon J. Lane-Getaz, University of Minnesota
lane0139@umn.edu

Reviewing the literature across the areas of education, psychology, statistics, and statistical and mathematics education illuminates the difficulties people have with the logic of inference. Within this literature the common indicator of statistical significance, the P-value, is fraught with controversy and confusion. Thirteen P-value misconceptions have been documented in empirical studies. In fall 2004 students in first and second courses in statistics responded to survey items developed based on this review of the research literature about P-values. The survey development and validation is discussed, along with summaries of students' correct reasoning or misconceptions about P-values based on this preliminary pilot.

1. Literature Review

Misinterpretations of *P*-values appear to be common among statistics students and some experienced researchers as well. A compilation of *P*-value misconceptions from these six empirical studies is listed in Table 1. This spotlight session highlights results of a pilot of a seventeen item survey designed to diagnose students' conceptions and misconceptions about *P*-values based on the research literature.

Table 1: Compilation of P-value misconceptions documented in empirical research

P-value misunderstandings and misinterpretations
1. *Significance level confusion:* The predetermined significance level, α and *P*-value are confused (Vallecillos, et al., 1992; Williams, 1997, 1998, 1999).
2. *P-value always low:* The belief that all *P*-values are small in value (Williams, 1998, 1999).
3. *P-value versus test statistic confusion:* The student confuses or doesn't understand the relationship between *P*-values and test statistics (Williams, 1998, 1999).
4. *Confusion of the converse (Causality):* Interprets the *P*-value—which is $P(D\|H_o)$—as $P(H_o\|D)$; Interprets small *P*-values as probability *the null hypothesis is true* (Brewer, 1985; Falk, 1988; Haller & Kraus, 2002; Pollatsek, et al., 1987; Vallecillos & Holmes, 1994).
5. *Significance testing confusion*: The logic or language of significance testing presents an obstacle to understanding and interpreting *P*-values (Batanero, 2000; Vallecillos, et al., 1992).
6. *Sample size dependence or effects:* Failure to recognize *P*-values are dependent on sample size; related to power of the test or the treatment effect (Mittag & Thompson, 2000; Wilkerson & Olsen, 1997).
7. *Confusion between sample effects and population effects:* Reflects a belief that the *P*-value is the probability of sample effects, rather than population effect under the null (Mittag & Thompson, 2000)
8. *Odds-against-chance fantasy:* Reflects a belief that significant *P*-values can be used to decide to accept or reject the idea that *chance caused* the experimental results obtained (Carver, 1978).
9. *Illusion of probabilistic proof (Determinism):* Reflects belief that small *P*-values ($p < \alpha$) justify a definitive statement; i.e., *outcome approach* (Oakes, 1986; Falk, 1988; Konold, 1989; Cohen, 1990).
10. *Valid hypothesis fantasy:* Reflects a belief that the *P*-value is the probability that the research hypothesis; i.e., *alternative hypothesis is true* (Carver, 1978; Brewer, 1985; Oakes, 1986; Haller & Kraus, 2002; Vallecillos & Holmes, 1994).
11. *Reliability / Replicability fantasy:* Reflects a belief that the *P*-value is related to reliability or that the repeatability of the research results is 1 – *P*-value (Carver, 1978; Oakes, 1986; Haller & Krauss, 2002; Mittag & Thompson, 2000).
12. *Probability H_a is "wrong:"* Reflects a belief that the *P*-value is the probability that the research hypothesis is "wrong" (Oakes, 1986; Vallecillos & Holmes, 1994; Williams, 1998, 1999; Brewer, 1985).
13. *P-value and Type I error:* Failure to differentiate between the *P*-value and Type I error rates (Garfield & Ahlgren, 1988; Haller & Kraus, 2002; Mittag & Thompson, 2000).

2. Survey and Analysis

- *How do beginning and more experienced students of statistics differ in their conceptions and misconceptions about P-values?*

Tables 2 through 5 detail the number and percentage of correct responses to each of the 17 items by course. The tables mirror the four sections of the *P*-value survey: Defining *P*-values, Using *P*-values, Interpreting *P*-values and Drawing Conclusions from *P*-values. The *P*-value conception or misconception being assessed is described in the left column of these tables. In the right columns are the number of correct answers for the item by course. Each of the four survey sections has a problem context. The scenario for each section precedes the results.

Table 2: Defining P-values—results by course

Scenario 1: A research article gives a P-value of .001 in the analysis section. Do you think the following definition is true or false?

	Undergraduates		*Graduate Students*	
	Lower	*Upper*	*1st Masters*	*2nd Doctoral*
1. Null hypothesis is true (False)	41 39%	28 43%	68 63%	29 51%
2. Formal definition (True)	56 54%	44 68%	85 79%	43 75%
3. Simulation definition (True)	52 16%	33 51%	62 57%	35 61%
4. Lay (informal) definition (True)	56 54%	45 69%	57 53%	37 65%
5. Population Proportion (False)	41 40%	31 48%	86 80%	46 81%

See note below.

Table 3: Using P-values—results by course

Scenario 2: District administrators of an experimental program similar to Head Start are interested in knowing if the program had an impact on reading readiness of first graders. Assume that the historical, pre-implementation mean Reading Readiness score for all first graders is 100 and the population standard deviation is 15. A random sample of current first graders who have been through program scored a mean Reading Readiness of 102.

	Undergraduates		*Graduate Students*	
	Lower	*Upper*	*1st Masters*	*2nd Doctoral*
6. Sample size impact (Valid)	63 61%	40 62%	60 56%	35 61%
7. Results due to chance (Invalid)	37 36%	12 19%	24 22%	19 33%
8. "Odds" against chance (Invalid)	25 24%	13 20%	14 13%	6 11%
9. Stochastics definition (Valid)	69 67%	47 72%	82 76%	46 81%

See note below.

Table 4: Interpreting P-values—results by course

Scenario 3: An ethical researcher is hoping to show that his new hair growth treatment had statistically significant results. How should this researcher interpret results from the research study?

	Undergraduates		*Graduate Students*	
	Lower	*Upper*	*1st Masters*	*2nd Doctoral*
10. Rareness measure (Valid)	53 52%	36 55%	76 70%	41 72%
11. Test statistics confusion (Invalid)	40 39%	24 37%	27 25%	11 19%
12. Converse is true (Invalid)	34 33%	39 60%	57 53%	33 58%
13. Large P-value significant (Invalid)	44 43%	34 52%	54 50%	20 35%

Note: Lower = lower division undergraduate statistics course; Upper = upper division undergraduate statistics course; 1st Masters = masters level first course in statistics; 2nd Doctoral = doctoral level second course in statistics

Table 5: Drawing Conclusions from P-values—results by course

Scenario 4: A researcher conducts an appropriate hypothesis test where she compares the scores of a random sample of students' SAT scores to a national average (500). She hopes to show the students' mean score will be higher than average. The researcher finds a P-value for her sample of .03.

	Undergraduates		*Graduate Students*	
	Lower	*Upper*	*1st Masters*	*2nd Doctoral*
14. Reliability (Invalid)	30 29%	24 37%	60 56%	27 47%
15. Valid Hypothesis (Invalid)	48 49%	40 62%	83 77%	41 72%
16. Wrong (Invalid)	40 39%	36 55%	67 62%	33 58%
17. Type-I (Valid)	57 55%	46 71%	61 57%	37 65%

Note: Lower = lower division undergraduate statistics course; Upper = upper division undergraduate statistics course; 1st Masters = masters level first course in statistics; 2nd Doctoral = doctoral level second course in statistics

- *Do beginning and more experienced students of statistics share the same conceptions and misconceptions about P-values?*

The *P*-value survey data collected in the fall 2004 sheds some light on this question. Figures 1 and 2 depict the means and box plots of total correct scores cross tabulated by course. There do appear to be some differences in these samples.

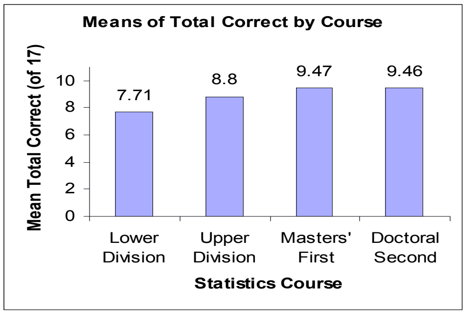

Figure 1: Mean of total items correct by course

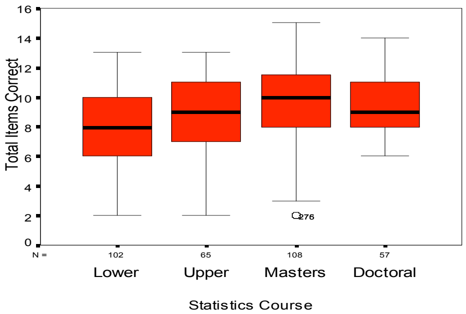

Figure 2: A comparison of medians and boxplots

Graduate students tended to get more items correct than undergraduates. There was less variation between scores among doctoral students taking their second course in statistics. In addition, there was evidence of a statistically significant difference in total scores between courses as well ($F_{.05}(3, 328) = 10.7$, $p < .001$).

3. Conclusions and implications for teaching and research

The results of this investigation indicate that this survey measures two levels of thinking: lower level statistical literacy and a higher level statistical reasoning and thinking. *P*-value misconceptions seem to require a deeper level of processing about *P*-values. Graduate respondents tended to answer correctly to more of the higher order thinking items.

There were four items that functioned counter-intuitively and require additional item development efforts. More qualitative detail is needed to shed light on why these items are discriminating so poorly. Either these items need to be improved or eliminated from the survey altogether. Future item development and modification should include cognitive interviews in which respondents "talk aloud" as they conduct the survey. These sessions can be videotaped to capture all of their reactions without intervention. This technique was used during the initial development process as well. In addition, some respondents should be interviewed after they have taken the survey to better understand how the survey is received. Some qualitative data and analysis may add the needed depth of information to further develop the instrument.

In addition to serving as a research tool, this survey can fulfill a practical role as a diagnostic tool for classroom use. For example, in spring 2005 this survey was used as a formative assessment of students' understanding of *P*-values after completion of an inference unit. The results helped the instructor determine how to target a final review on *P*-values, prior to giving students the final summative assessment. The survey identifies subtle aspects of the *P*-value that may remain elusive after instruction. By taking the survey students' misconceptions and misunderstandings are made explicit. Once they are aware of these misconceptions, the instructor has a teaching opportunity to confront and potentially overturn students' misunderstandings.

References

Batanero, C. (2000). Controversies around the role of statistical tests in experimental research. *Mathematical Thinking and Learning*, *2*(1&2), 75-97.

Brewer, J. K. (1985). Behavioral statistics textbooks: Source of myths and misconceptions? *Journal of Educational Statistics, (10)*3, 252-268.

Carver, R. P. (1993). The case against statistical significance testing, revisited. *Journal of Experimental Education, 61*(4), 287-292.

Carver, R. P. (1978). The case against statistical significance testing. *Harvard Educational Review, 48*(3), 378-399.

Cohen, J. (1994). The earth is round (p < .05). *American Psychologist, 49*(12), 997-1003.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45*(12), 1304-1312.

Dillman, D. A. (2000). *Mail and internet surveys: The tailored design method* (2nd ed.). New York: John Wiley & Sons, Inc.

Falk, R. (1988). Conditional probabilities: Insights and difficulties. In R. Davidson, & J. Swift (Eds.), *Proceedings of the Second International Conference on Teaching Statistics* (pp. 292). Victoria, BC: University of Victoria.

Falk, R., & Greenbaum, C. (1995). Significance tests die hard. *Theory & Psychology, 5*(1), 75-98.

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.

Haller, H., & Krauss, S. (2002). Misinterpretations of significance: A problem students share with their teachers? *Methods of Psychological Research, 7*(1), 1-20.

Lane-Getaz, S. J. (2005). Reasoning about *P*-values to develop the logic of inference: A review of related literature. Unpublished manuscript.

Lane-Getaz, S. J. (2005). Reasoning about *P*-values: Results of a research-based survey. Unpublished manuscript.

Mittag, K. C., & Thompson, B. (2000). A national survey of AERA members' perceptions of statistical significance tests and other statistical issues. *Educational Researcher, 29*(4),14-20.

Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioral sciences,* Chichester: Wiley.

Pollatsek, A., Well, A. D., Konold, C., Hardiman, P., & Cobb, G. (1987). Understanding conditional probabilities. *Organizational Behavior and Human Decision Processes, 40*, 255-269.

Vallecillos, J. A., & Holmes, P. (1994). Students' understanding of the logic of hypothesis testing. In *Proceedings of the Fourth International Conference on Teaching Statistics.* Marrakech, Morocco: International Statistical Institute.

Wilkerson, M., & Olson, J. R. (1997). Misconceptions about sample size, statistical significance, and treatment effect. *The Journal of Psychology, 131*(6), 627-631.

Williams, A. M. (1999). Novice students' conceptual knowledge of statistical hypothesis testing. In J. M. Truran, & K. M. Truran (Eds.), *Making the Difference: Proceedings of the Twenty-second Annual Conference of the Mathematics Education Research Group of Australasia* (pp. 554-560). Adelaide, South Australia: MERGA.

Williams, A. M. (1998). Students' understanding of the significance level concept. In L. Pereira-Mendoza (Ed.), *Statistical Education - Expanding the Network: Proceedings of the Fifth International Conference on Teaching Statistics* (pp. 744-749). Voorburg, The Netherlands: International Statistical Institute.