A GRAPHICAL APPROACH TO EXAMINE INFERENTIAL REASONING DEVELOPMENT

Sharon Lane-Getaz
St. Olaf College, Northfield, Minnesota USA
lanegeta@stolaf.edu

This observational study examines inferential reasoning development in students taking a randomization-based introductory college course (n = 38). The proportions of students answering each item correctly are compared from Pretest to Posttest using a simple descriptive scatterplot. This "canoe plot" includes a superimposed 95% confidence band which differentiates items with no statistical difference between the proportion answering correctly on the Pretest and Posttest from those that do differ. A brief discussion follows highlighting the content for those items where inferential reasoning differs from Pretest to Posttest. The Reasoning about P-values and Statistical Significance (RPASS) scale is used to measure students' inferential reasoning outcomes and gains. Directions for future research are discussed.

INTRODUCTION

Despite the importance of understanding inference to be "statistically educated," inferential reasoning has historically remained elusive for many introductory statistics students at the end of their first course. The many difficulties people have using or interpreting statistical tests and confidence intervals and related inferential reasoning have been summarized in the literature (e.g., Lane-Getaz, 2007; Sotos, Vanhoof, Noortgate, & Onghena, 2009; Utts, 2003). Cobb (2007) posits that students' difficulties with formal inferential reasoning may be tied to an over-emphasis on asymptotic tests (e.g., the normal and *t*-distributions). He suggests bringing the "logic of inference" to the center of the introductory course using randomization-based curricula. Cobb suggests that the introductory course focus on the "Three Rs of inference:" Randomize data, Replicate to create a null model, and Reject any model that puts the observed data in the null model's tail. (For an illustration: http://sharonlanegetaz.efoliomn.com/research). To this end, statistics educators are designing curricula with more randomization-based content (e.g., Lock, Lock, Lock, Lock, & Lock, 2012; Chance & Rossman, 2006; Tintle, VanderStoep, Holmes, Quisenberry, & Swanson, 2011). Statistics education researchers must examine the impact of these new randomization-based curricula on students' inferential reasoning.

In related research, Tintle and colleagues (2011) compared outcomes from a randomization-based introductory course to those from an algebra-based course. They used the *Comprehensive Assessment of Statistics Outcomes in a first Statistics course* (CAOS) to compare student learning outcomes (delMas, Garfield, Ooms, & Chance, 2007). The results suggest that randomization-based content produced better learning outcomes than the algebra-based course (Tintle, et al., 2011). The current study focuses on students' inferential reasoning before and after taking an introductory course. The *Reasoning about P-values and Statistical Significance* scale, which has established evidence of reliability and validity, is used to measure students' inferential reasoning outcomes (see Lane-Getaz, 2013).

METHODS

This observational study was conducted at a liberal arts college in the Midwest United States during spring 2010. Students were enrolled in two sections of an introductory-level statistics course with a calculus prerequisite. This course fulfills a quantitative reasoning requirement and is designed for students majoring in the natural sciences or mathematics. Of 49 students enrolled, 38 (20 females, 18 males) completed the Pretest and Posttest and consented to participate (78%). The sample includes (5) freshmen, (17) sophomores, (7) juniors and (9) seniors.

The textbook, *Investigating Statistical Concepts, Applications and Methods* (ISCAM, Chance & Rossman, 2006), is a discovery-oriented, technology-rich curriculum designed to employ randomizations and simulations to introduce asymptotic tests. For example, categorical data analysis proceeds from two-way table simulations to Fisher's exact test to normal approximations. Quantitative response analysis proceeds from randomization tests to one- and two-sample *t*-tests.

In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS9, July, 2014), Flagstaff, Arizona, USA.* Voorburg, The Netherlands: International Statistical Institute. iase-web.org

Bivariate analysis proceeds from sampling lines to simple linear regression. Course sections met three times per week in a computer-equipped classroom—Mondays, Wednesdays, and Fridays—for 55 minutes. Online applets and *R* software were used for simulations and data analysis.

Introductory students' correct conceptions, misconceptions and difficulties with inferential concepts were measured using the online RPASS-8 scale, a reliable and valid measure of inferential reasoning (see Lane-Getaz, 2013). The 35 RPASS-8 items include the 34 items with previously reported psychometric properties and a 35th item (Item 4b-6) added to assess whether students understand how increasing sample size versus increasing replications impact the variability of a randomization distribution.

In addition to reporting the RPASS-8 Pretest and Posttest means and standard deviations, the Pretest and Posttest proportion of students answering each item correctly are plotted on "canoe plots," named for the canoe-shaped 95% confidence band around the $\pi_{posttest}=\pi_{pretest}$ line. The 95% confidence band differentiates items with significant differences between the Pretest and Posttest proportions (outside the band) from those with insignificant differences (within the band). The margin of error for the difference in proportions includes a Wilson adjustment for each proportion $\tilde{p}_i = (X_i + 1)/(n_i + 2)$ to maintain a 95% nominal rate (see Agresti & Caffo, 2000). No family-wise correction is made since this graph is intended to be used for descriptive purposes.

RESULTS

On average, students answered 79% of the 35 RPASS-8 items correctly on the Posttest (n = 38). An average of seven more items were answered correctly on the Posttest (M = 27.5, SD = 2.7) compared to the Pretest (M = 20.9, SD = 4.4). To better understand how responses differed by item, the "canoe plot" in Figure 1 juxtaposes the Pretest and Posttest proportion of students answering each of the 35 items correctly. The item icons are color coded to reflect if the item assessed a correct conception, a misconception or if the multiple choice (MC) options assessed correct conceptions with common misconceptions as distractors. Two items are just below the confidence band, 22 are above the band, and 11 are within. Respondents seemed to use better reasoning on the Pretest than on the Posttest for the two items below the band, Item 2-5 and Item 4b-5. For Item 2-5 Pretest respondents indicated that confidence intervals can be used to assess statistical significance, much like p-values are used when hypothesis testing (proportion difference Posttest – Pretest = -.24). For Item 4b-5 (proportion difference = -.23) Pretest respondents indicated that larger sample sizes can yield statistically significant results which are not necessarily of practical importance. The 22 items above the band (Table 1) appear in descending order based on the difference in proportions along with the item number and a description of the reasoning assessed. Table 2 lists the 11 items within the band for which the course had little, if any, impact.

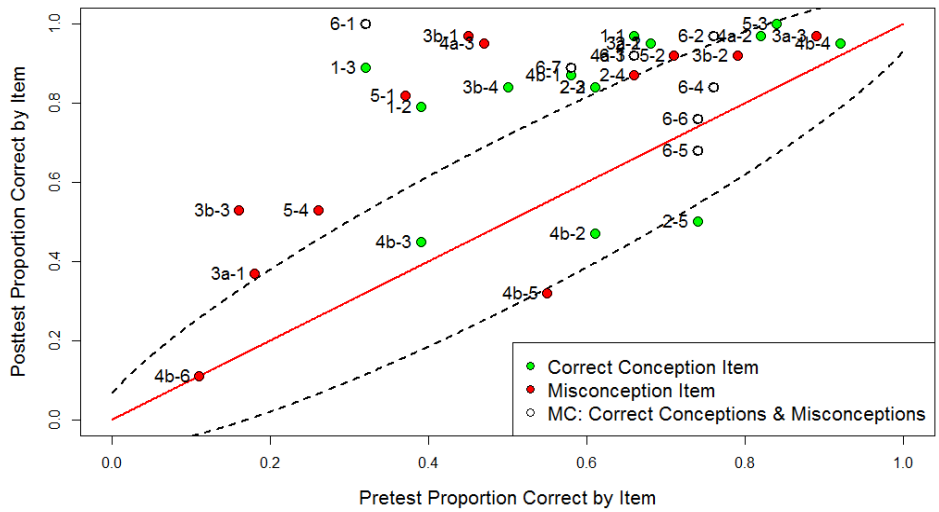

Figure 1: "Canoe plot" showing the proportion of students answering each item correctly on the Pretest versus the Posttest, with a 95% confidence band for equal proportions ($\pi_{posttest} - \pi_{pretest} = 0$).

Table 1: RPASS-8 items, where Posttest Proportions exceed the Pretest Proportions (22 items)

RPASS-8 item	Brief description of inferential reasoning concept or difficulty being assessed	Proportion Difference
6-1	Selects a textbook definition of a *p*-value given multiple choices.	.68
1-3[b]	Understands magnitude of *p*-value depends if conducting one- or two-sided test.	.57
3b-1[a]	Uses a density curve and an observed value to estimate if the observed value (or more extreme) is statistically significant.	.52
4a-3[b]	Believes causal conclusion can be drawn from small *p*-values regardless of study design.	.48
5-1	Misinterprets a *p*-value as the probability that the null hypothesis is false.	.45
1-2	Recognizes an informal description of the *p*-value embedded in context.	.40
3b-3	Believes the *p*-value is always a low number (or is always desired to be a low number).	.37
3b-4[ab]	Recognizes incorrect direction is specified for shading *p*-value.	.34
6-7	Differentiates between concepts of Type I and Type II error.	.31
1-1	Recognizes a formal textbook definition of the *p*-value without a context.	.31
4b-1[ab]	Understands *P*-value is computed in the direction hypothesized by the researcher.	.29
5-4	Confuses whether statistically significant results refer to a sample or a population.	.27
2-1[b]	Recognizes the *p*-value in terms of sampling variation in a sampling distribution.	.27
3a-2	Understands the *p*-value as a rareness measure.	.27
4a-1	Confuses the *p*-value with the significance level α.	.26
6-3	Understands the stronger the evidence of a difference or effect, the smaller the *p*-value.	.26
2-2	Understands the stronger the evidence of a difference or effect, the smaller the *p*-value.	.23
2-3[b]	Understands the *p*-value as being conditioned on the null hypothesis being true.	.23
5-2	Believes that the *p*-value is the probability that the alternative hypothesis is true.	.21
2-4	Believes the *p*-value is the probability observed results are due to chance or caused by chance, if the null is true.	.21
6-2	Understands that a small *p*-value suggests the results are statistically significant.	.21
3a-1[ab]	Believes statistics provide definitive proof; misuses the deterministic Boolean logic of contrapositive proof.	.19

Note. [a]Items associated with a sampling or randomization distribution. [b]Requests an explanation of reasoning.

Table 2: RPASS-8 items, where Posttest and Pretest Proportions are Statistically Equal (11 items)

RPASS-8 item	Brief description of inferential reasoning concept or difficulty being assessed	Proportion Difference
5-3	Reasons that the smaller *p*-value, the stronger the evidence of a difference or effect.	.16
4a-2	Reasons that the greater evidence of a difference or effect, the smaller the *p*-value.	.15
3b-2[ab]	Believes a *p*-value is always a low number (or is always desired to be a low number).	.13
3a-3	Believes a *p*-value is always a low number (or is always desired to be a low number).	.08
6-4	Reasons about the impact of a small sample size on statistical significance.	.08
4b-3[a]	Assesses statistical significance based on a depicted randomization distribution.	.06
4b-4[a]	Uses significance level α to assess whether a *p*-value is rare or unusual enough to indicate statistical significance.	.03
6-6	Understands that in order to conduct a significance test the necessary conditions must be met.	.02
4b-6[ab]	Understands the impact of increasing the number of replications in a simulation versus the impact of increasing the sample size.	.00
6-5	Understands a small *p*-value does not necessarily mean a practical difference or effect.	-.06
4b-2[ab]	Understands large *p*-values provide insufficient evidence to reject the null hypothesis.	-.14

Note. [a]Items associated with a sampling or randomization distribution. [b]Requests an explanation of reasoning.

DISCUSSION

Twenty-two RPASS-8 items had a significantly higher proportion of students answering correctly on the Posttest compared to the Pretest. Students' statistical literacy improved: recognizing formal and informal definitions of the *p*-value (6-1, 1-2, 1-1, 2-1, 3a-2, 2-3). Students also showed evidence of improved inferential reasoning: correctly assessing significance graphically, assessing the strength of statistical evidence, assessing the impact of alternative hypotheses on *p*-values and interpreting small *p*-values and Type I and II errors correctly (3b-1, 3b-4, 6-3, 2-2, 6-2, 1-3, 4b-1, 6-7.) Students also overturned some known difficulties and

misconceptions about inference (5-1, 3b-3, 5-4, 4a-1, 5-2, 2-4, 3a-1, 4a-3). Of particular importance is that students overturned a common misconception that given a small *p*-value a causal conclusion can be drawn, regardless of study design (4a-3).

Little improvement could be expected on four items for which students exhibited good inferential reasoning on the pretest ($p_{Pretest} > .80$)—this may be an effect of statistics being part of the common core standards. These four items include reasoning about: relationships between *p*-values and effects (5-3, 4a-2), the impact of sample size on *p*-values (4b-4) or overturning the belief that *p*-values should always be small (3a-3). However, there was room for improvement that was not realized on three of six items requiring graphical interpretation of a randomization distribution (4b-2: differentiating the impact of increasing replications from increasing sample size, 4b-3: assessing statistical significance, 4b-6: making a rejection decision). Students seem to have difficulties with multiple choice items that mixed correct conceptions with misconceptions as distractors: the impact of sample size on *p*-values (6-4), differentiating statistical and practical importance (6-5) and attending to conditions for inference (6-6).

CONCLUSION

This study illustrates a simple graphical method to examine how inferential reasoning changed in an introductory statistics course from Pretest to Posttest. The item level comparison shows that students improved their statistical literacy, inferential reasoning and graphical interpretation skills and overturned some common misconceptions. Students did not develop the level of statistical thinking needed to synthesize how inferential concepts relate to one another.

The item-level, graphical comparison method demonstrated in this study will be used in future research to compare how inferential reasoning differs in a course taught with and without randomization content, and to examine how targeted activities impact inferential reasoning outcomes. To broaden generalizability of findings, some future studies will include collaboration with instructors at various types of institutions.

REFERENCES

Agresti, A, & Caffo, B. (2000). Simple and Effective Confidence Intervals for Proportions and Differences of Proportions result from Adding Two Successes and Two Failures. *The American Statistician*, *54*(4), 280–288.

Chance, B. L., & Rossman, A. J. (2006). *Investigating Statistical Concepts, Applications, and Methods*, Belmont, CA: Brooks/Cole – Thomson Learning.

Cobb, G. (2007), The Introductory Statistics Course: A Ptolemaic Curriculum?. *Technology Innovations in Statistics Education*, *1*,(1). http://repositories.cdlib.org/uclastat/cts/tise/

delMas, R. C., Garfield, J. B., Ooms, A., & Chance, B. (2007). Assessing Students' Conceptual Understanding after a First Course in Statistics. *Statistics Education Research Journal* [online], *(6)*2, 28-58. http://www.stat.auckland.ac.nz/serj

Lane-Getaz, S. J. (2013). Development of a Reliable Measure of Students' Inferential Reasoning Ability. *Statistics Education Research Journal (SERJ)*, *12*(1), 20-47. http://iase-web.org/documents/SERJ/SERJ12(1)_LaneGetaz.pdf

Lane-Getaz, S. J. (2007). Toward the Development and Validation of the Reasoning about *P*-values and Statistical Significance Scale. In B. Phillips & L. Weldon (Eds.), *Proceedings of the ISI / IASE Satellite Conference on Assessing Student Learning in Statistics*, Voorburg, The Netherlands: ISI. http://www.stat.auckland.ac.nz/~iase/publications/sat07/Lane-Getaz.pdf

Lock, R., Lock, P., Lock, K., Lock, E., & Lock, D. (2012). *Statistics: Unlocking the Power of Data*. Wiley. http://www.lock5stat.com/index.html

Sotos, A., Vanhoof, S., den Noortgate, W., & Onghena, P. (2009). How Confident are Students in their Misconceptions about Hypothesis Tests?. *Journal of Statistics Education*, *17*(2). http://www.amstat.org/publications/jse/v17n2/castrosotos.html

Tintle, N, VanderStoep, J, Holmes, V., Quisenberry, B. & Swanson, T. (2011). Development and Assessment of a Preliminary Randomization-based Introductory Statistics Curriculum. *Journal of Statistics Education*, *(11)*1.

Utts, J. (2003). What Educated Citizens Should Know about Statistics and Probability. *The American Statistician*, *57*(2), 74-79.