STUDENTS' INFERENTIAL REASONING ABOUT SAMPLE SIZE

Ethan Brown
Educational Psychology
University of Minnesota—Twin Cities
brow3821@umn.edu

Intuitions that guide judgments about sample size and sampling variability have been studied for over 40 years, but the implications for statistics instructors are not clear. The administration of the second version of Goals and Outcomes Associated with Learning Statistics (GOALS-2), with 1,165 US undergraduate statistics students participating in a pilot test in December 2013, provides opportunities for an exploration of students' inferential reasoning about sample size across multiple contexts. The assessment contains items regarding the impact of sample size on sampling variability, confidence interval widths, and p-values. For each item, about half of participants answered the item correctly, and performance was positively correlated across items. However, correlations were small and patterns of responses revealed little consistency across formally analogous distractors. More fundamental research on the linkages between these concepts appears to be needed.

INTRODUCTION

Randomization-based statistical curricula use empirical sampling distributions as building blocks for inference. For instance, Cobb (2007, p. 12) influentially proposed the "three Rs of inference": Randomize, Repeat, and Reject, where students produce an empirical distribution of sample statistics based on a null hypothesis (Randomize and Repeat), enabling students to evaluate their model against the data (Reject). Others have built on this framework (e.g., Garfield, delMas, & Zieffler, 2012; Tintle, VanderStoep, Holmes, Quisenberry, & Swanson, 2011), but the outlines are similar.

How strong is the link between understanding the empirical sampling distribution and its inferential consequences? This study investigates this link in terms of students' reasoning about sample size. Based on the e-ATLAS project funded by the National Science Foundation (DUE 1044812 & 1043141), a 23-item instrument, GOALS-2, was developed to assess statistical reasoning, with 1,165 U.S. students participating in a pilot test at the conclusion of an introductory statistics course. The instrument contains three items that assess students' reasoning about sample size by comparing the sampling variability, p-values, and confidence interval widths produced under two different sample sizes.

PREDICTIONS FROM PREVIOUS RESEARCH

Research on reasoning about sample size has identified processes by which students a) correctly respond that larger samples will have smaller expected sampling variability, or b) incorrectly respond that sample size will not change the expected sampling variability.

Normative reasoning about these tasks should lead to recognition that the sampling variability is expected to be smaller for larger sample sizes, and thus the *p*-value is expected to be smaller and confidence interval narrower. Students can use formal strategies from their statistics courses that lead to correct solutions for these problems, even if their intuitions are otherwise incorrect (e.g., Watson, 2000). Another plausible pathway to a correct solution is the well-documented intuition that larger samples should be regarded as more accurately representing the population (e.g., Nisbett, Krantz, Jepson, & Kunda, 1983), known as the *size-confidence intuition* (Sedlmeier, 1999).

There is substantial evidence from psychology that people tend to ignore sample size information in certain contexts (for a recent review, see Lem, Van Dooren, Gillard, & Verschaffel, 2011). This suggests that many students expect differently-sized samples to have the same sampling variability, *p*-values, and confidence interval widths.

The original explanation for sample size neglect is the *representativeness heuristic* (Kahneman & Tversky, 1972), which posits that people judge the probability of the sample by the degree to which it resembles the population and the sampling process. Since sample size does not

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 [© 2014 ISI/IASE]

affect this resemblance, people will tend to ignore sample size. This construct has been heavily criticized for its vagueness about when representativeness will apply (e.g., Gigerenzer, 1996), but it is still widely cited and studied (e.g., Watson & Callingham, 2013).

An alternative explanation for sample size neglect is that people may simply interpret empirical sampling distributions as frequency distributions (Well, Pollatsek, & Boyce, 1990) and may not activate the size-confidence intuition in the context of empirical sampling distributions (e.g., Sedlmeier, 1998).

A few studies have found that some subjects reason that larger sample size leads to *more* sampling variation, but this error has not been frequently noted in the literature. The reasoning that leads to this conclusion is not clear (but see Chance, delMas, & Garfield, 2004).

MATERIALS AND METHODS

GOALS-2 is a 23-item instrument developed in Fall 2013 to assess students' statistical reasoning after a first course in statistics. It is a refinement of GOALS-1 (see Sabbag, 2013), which itself was a modification of the *Comprehensive Assessment of Outcomes in Statistics* (CAOS; delMas, Garfield, Ooms, & Chance, 2007). Pilot data were gathered from students enrolled in introductory statistics courses at more than 30 institutions (N = 1,165).

The analysis presented in this paper focuses on three items from GOALS-2 that assess students' reasoning about sample size.

The "sampling variability" item presents three candidate pairs of dotplots to represent the empirical sampling distributions of a Bernoulli process (p = .50) with N = 20 and N = 100: one option where the two dotplots have the same variability, one where the smaller sample has more variability, and one where the larger sample has more variability.

Similarly, the "confidence interval width" item presents three candidate pairs of 95% confidence intervals for the same population mean with N = 50 and N = 250: again, one option where the two confidence intervals have the same width, one where the smaller sample has a wider confidence interval, and one where the larger sample has a wider confidence interval.

The "*p*-value" item asks students to assess whether the *p*-value would be the same, larger, or smaller if they received the same difference in proportions utilizing a larger sample (n = 100 in each group) compared to a smaller sample (n = 20 in each group).

The full text and graphics of the 3 tasks can be viewed at: http://www.tc.umn.edu/~brow3821/icots-2014/

RESULTS

About 50% of students correctly answered the "sampling variability" item, choosing the candidate dotplot pair where the smaller sample had more sampling variability in sample means. As expected from a similar item administered on CAOS and the original GOALS-1, as well as from the sample size neglect literature, the most frequent misconception (34%) was to choose the option where the two candidate dotplots have the same variability. Overall percentages for the "*p*-value" item were similar, with about half of students answering the item correctly and 34% of students responding that a larger sample size would lead to the same *p*-value.

Unexpectedly, 34% of students chose a response indicating that a larger sample size has a *wider* expected confidence interval on the "confidence interval width" item. Similar to the other two items, about half of the students correctly answered the item.

The picture becomes more complex when examining students' responses across the three items. Students who tended to correctly assess the impact of sample size in the "sampling variability" item were somewhat more likely to correctly note that larger sample size is expected to lead to narrower confidence intervals (φ = .28, 95% CI [.22, .33]) and smaller *p*-values (φ = .13, [.08, .19]). Also, students indicating that a larger sample size is expected to lead to narrower confidence intervals were more likely to also indicate that this would lead to smaller *p*-values (φ = .24, [.18, .29]).

Associations between formally equivalent misconceptions were positive but fairly small in magnitude. The correlations between choosing the response options associated with sample size neglect, coded as a dichotomous variable, were all around φ = .10. Results for response options associated with larger samples having more sampling variability were comparable, although there was a negligible correlation between thinking larger samples have more variability and smaller *p*-values (φ = .04, 95% CI [–.01, 0.10]).

DISCUSSION

Statistical reasoning includes coordinating the related concepts of sampling variability, confidence intervals, and *p*-values. Students' responses to three GOALS-2 items related to the conceptual understanding of the impact of sample size suggest that students may not be reasoning consistently across these concepts at the conclusion of an introductory statistics course.

Hypotheses based on the results of previous research predicted that students' response patterns for the three items would be similar (i.e., the percentage of students answering correctly would be similar across the three items). It was also hypothesized that most students would either respond correctly or neglect sample size. Although the three items did have a similar percentage of students who responded correctly, sample size neglect was not as frequently observed on the "confidence interval width" item as it was on the other two items. Furthermore, although there were associations found in performance across these items, prediction was modest at best, and the choice of distractor was only mildly predictive (highest was φ = .12) of choosing a similar distractor on the other items.

This could be evidence that many students, even those who correctly respond to the "sampling variability" item, may not be drawing appropriate connections between sampling variability and its inferential implications for *p*-values and confidence intervals.

One limitation to the study is that the instructional content of the participating courses was not analyzed, nor were participant characteristics. It may well be that students were directly taught issues of sample size in some contexts, or were able to recall formulas that were relevant to some of the three items. Another limitation of this analysis is that students' inconsistency across concepts may reflect contextual differences, low item-level reliability, or other non-task-relevant error. Preliminary unpublished results from another e-ATLAS assessment, *Models of Statistical Thinking*, suggest substantial inconsistencies in students' reasoning about *p*-values and sample size even across nearly identical items.

This negative result may also be indicative of difficulties with fundamental concepts of variability and distribution. In a series of studies of college students' understanding of the sampling distribution, Chance et al. (2004) encountered several confusions founded in fundamental conceptual gaps in these concepts. These gaps may intrude in different ways on the concepts assessed in these GOALS-2 items.

Following that idea further, it may be useful to more directly study the links between the size-confidence intuition, variability, and distribution in the context of sample size reasoning. These GOALS-2 results beg the question of what processes are actually leading to the student responses. A planned follow-up experiment is being designed to provide more insight into these connections.

REFERENCES

Chance, B., delMas, R., & Garfield, J. (2004). Reasoning about sampling distributions. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 295–323). Dordrecht, the Netherlands: Springer.

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

delMas, R., Garfield, J., Ooms, A., & Chance, B. (2007). Assessing students' conceptual understanding after a first course in statistics. *Statistics Education Research Journal, 6*(2), 28–58.

Garfield, J., delMas, R., & Zieffler, A. (2012). Developing statistical modelers and thinkers in an introductory, tertiary-level statistics course. *ZDM*, *44*(7), 883–898.

Gigerenzer, G. (1996). On narrow norms and vague heuristics: A reply to Kahneman and Tversky. *Psychological Review*, *103*(3), 592–596.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, *3*(3), 430–454.

Lem, S., Van Dooren, W., Gillard, E., & Verschaffel, L. (2011). Sample size neglect problems: A critical analysis. *Studia Psychologica*, *53*(2), 123–135.

Nisbett, R. E., Krantz, D. H., Jepson, C., & Kunda, Z. (1983). The use of statistical heuristics in everyday inductive reasoning. *Psychological Review*, *90*(4), 339–363.

Sabbag, A. G. (2013). *A Psychometric Analysis of the Goals and Outcomes Associated with Learning Statistics (GOALS) instrument.* (Unpublished Master's thesis). University of Minnesota, Twin Cities.

Sedlmeier, P. (1998). The distribution matters: two types of sample-size tasks. *Journal of Behavioral Decision Making*, *11*(4), 281–301.

Sedlmeier, P. (1999). *Improving statistical reasoning: Theoretical models and practical implications*. Mahwah, New Jersey: Lawrence Erlbaum.

Tintle, N., VanderStoep, J., Holmes, V.-L., Quisenberry, B., & Swanson, T. (2011). Development and assessment of a preliminary randomization-based introductory statistics curriculum. *Journal of Statistics Education*, *19*(1), n1.

Watson, J. (2000). Preservice mathematics teachers' understanding of sampling: intuition or mathematics. *Mathematics Teacher Education and Development*, *2*, 121–135.

Watson, J., & Callingham, R. (2013). Likelihood and sample size: The understandings of students and their teachers. *The Journal of Mathematical Behavior*, *32*(3), 660–672.

Well, A. D., Pollatsek, A., & Boyce, S. J. (1990). Understanding the effects of sample size on the variability of the mean. *Organizational Behavior and Human Decision Processes*, *47*(2), 289 – 312.