Measuring Statistical Literacy Surrounding Confidence Intervals

Susan Lloyd, Matthew Beckman
The Pennsylvania State University

Science Achievement Graduate Fellows

Introduction

Motivation:
- With the proliferation in data, an introductory statistics course is a requirement for many degree programs
- Hypothesis testing is often presented as a rote procedure
- Many research communities falsely believe that only statistically significant results are meaningful
- With the growing desire to shift away from a dichotomous formal decision framework, interval estimation becomes a preferred way to communicate the findings of a study

Methods:
- Develop instrument using a fractional factorial design
- Deploy instrument to collect data from Penn State students enrolled in introductory statistics
- Analyze the data using multivariate techniques

Implications:
- Meaningful changes to curriculum and instruction
- Investigate misunderstandings and identify ways to effectively communicate about data using confidence intervals
- Prompt the formation of similarly designed instruments

Research Questions of Interest

RQ1: What degree of understanding do students have about confidence intervals after completing an introductory statistics course using simulation-based inference methods?

RQ2: How do individual students' backgrounds shape their interpretation and understanding of statistical concepts, including confidence intervals?

Methods

Subject:
- Statistical literacy (multidimensional construct): the ability to effectively read, construct schema, and use basic statistical information to solve problems and communicate findings
- Content area described by 5 attributes: confidence level, error, sample size, p-value, conditions of the data
- Testlet structure with mixed formats
  - 3 selected response items to assess reading, constructing schema, and problem-solving; 1 constructed response item to assess communicating

Procedure:
- A 2^{5-1} fractional factorial design assigned attributes to testlets
- The items within each testlet were constructed such that, collectively, they assessed the attributes assigned, and did not assess the attributes not assigned, to the testlet
- Items were revised to logically fit in the test blueprint

Intended Remaining Procedure:
- Items will be revised based on feedback provided from other expert statistics education researchers
- Think-aloud interviews with undergraduate and graduate statistics students will guide further revisions to the items
- The reliability & validity of the final instrument will be verified
- Plan to collect data at-large from Penn State students enrolled in an introductory statistics course

Figures & Tables

Table 1: Alias Table for Resolution V Fractional (1/2) Factorial Design

Alias Table (Resolution V)		
Factors	**Alias Structure**	
A	I + ABCDE	AB + CDE
B	A + BCDE	AC + BDE
C	B + ACDE	AD + BCE
D	C + ABDE	AE + BCD
E	D + ABCE	BC + ADE
	E + ABCD	BD + ACE
		BE + ACD
		CD + ABE
		CE + ABD
		DE + ABC

Figure 1: Test Blueprint to Develop Items for the Instrument Assessing Statistical Literacy Surrounding Confidence Intervals

Objectives / Content	Use Power Analysis	Calculate (Estimate) Error	Build / Identify a CI	Compare / Discuss widths of the CI	Interpret in Context	Identify Plausible Values	Understand definitions and meanings of terms
Confidence Level				**S1Q3:** How increasing / decreasing confidence level affects CI width **S4Q15:** How decreasing confidence level affects width **S8Q32:** Use formula for CI and relationship between confidence level and width of CI to explain relationship between confidence level and multiplier **S11Q42:** Determine appropriate multiplier based on knowledge of confidence level and width of CI	**S1Q4:** Interpret 95% CI for difference in proportions **S4Q16:** In the long run, how many intervals are expected to capture the true parameter **S7Q26:** Recognize parameter is fixed, process of constructing CI is random, confidence level determines success rate **S7Q27:** Interpret 99% CI for difference in means **S7Q28:** Discuss relationship between confidence level & CI **S14Q56:** Discuss trade-offs when determining the confidence level to choose	**S1Q2:** Identify inferential procedure and recognize CI represents range of reasonable values **S1Q4:** Determine if there is evidence of a difference between two groups **S6Q23:** Determine if a value is plausible based on a CI with a specified level of confidence **S9Q33:** Identify the confidence level	**S4Q13:** Understand the meaning of 90% confidence level in context **S7Q25:** Understand the meaning of level of confidence in general
Error	**S8Q31:** Determine sample size needed to obtain the desired margin of error	**S12Q47:** Calculate margin of error from CI **S16Q62:** Calculate standard error for difference in means **S16Q63:** Estimate standard error from bootstrap distribution	**S4Q14:** Identify a CI as sample statistic +/ margin of error **S6Q22:** Construct CI using bootstrapping to estimate standard error **S16Q64:** Build a CI using the standard error estimated using a bootstrap distribution	**S2Q7:** How standard error affects CI width **S4Q15:** How decreasing margin of error affects CI width **S8Q30:** How standard error is affected by sample size, and in turn affects CI width **S9Q35:** How pairing affects standard error, and in turn, CI width **S10Q39:** How decreasing sample size affects error and CI width			**S2Q5:** Understand meaning of error in context **S10Q37:** Understand meaning of margin of error **S16Q61:** Definition of standard error in general
Sample Size	**S8Q31:** Determine sample size needed to obtain the desired margin of error		**S2Q8:** Determine if the sample size is sufficiently large to justify use of CLT to form a CI **S6Q24:** Determine if the sample size is sufficiently large to use bootstrapping for constructing CI	**S1Q3:** How increasing / decreasing sample size affects CI width **S8Q30:** How sample size affects standard error, and in turn affects CI width **S10Q39:** How decreasing sample size affects CI width **S11Q44:** How sample size affects CI width and precision **S13Q50:** Identify inferential procedure and recognize impact of small sample size on CI width	**S13Q51:** Recognize implications of small sample size **S13Q52:** Discuss whether the conditions are met based on the sample sizes **S6Q24:** Discuss how methods produce similar results (2SD method and percentile method) when sample size is large	**S1Q1:** Identify a possible sample based on context **S3Q9:** Identify the sample size in a given scenario **S13Q49:** Identify the sample size for each group	**S8Q29:** Definition of sample size in context
P-values			**S5Q18:** Choose reasonable form for CI and reasonable interval based on p-value		**S3Q10:** Interpret p-value in context **S5Q19:** How p-value changes going from CI (two-sided) to one-sided test **S10Q40:** Discuss relationship between p-value and CI	**S3Q11:** Determine if a parameter is plausible based on the two-sided p-value **S4Q15:** Discuss how the two-sided p-value compares to significance level based on plausible values **S5Q18:** Identify a plausible CI based on the p-value **S5Q20:** Determine if predictor is effective based on CI and discuss how p-value would compare to significance level **S6Q23:** How changing the null value affects p-value **S10Q38:** Identify percentage of statistics in each tail **S11Q43:** Choose a reasonable CI based on the given p-value **S12Q48:** Determine if there is a difference in two groups based on the p-value **S14Q54:** Determine if values are (un)likely to occur and how this reflects itself in the p-value **S14Q55:** Identify percentage of statistics in each tail and whether values are reasonable based on CI	**S5Q17:** Understand meaning of effective predictor in relation to p-value **S11Q41:** Definition of p-value in general **S14Q53:** Interpret random chance in relation to interpretation of p-value
Conditions of the Data			**S2Q6:** Build a CI using the 95% rule **S2Q8:** Use CLT to justify building a CI **S3Q12:** Determine if conditions are met for using normal distribution to build CI **S9Q34:** Use 95% rule to build CI using margin of error **S12Q46:** Determine if bootstrapping is appropriate to build a CI	**S1Q3:** How outliers affect CI width	**S6Q24:** Discuss how CI's constructed using percentile and standard error method compare **S15Q59:** Recognize if 95% rule is appropriate in a given situation	**S9Q36:** How outliers affect the range of plausible values and resulting CI **S15Q58:** Identify multiplier using symmetry of the normal distribution **S15Q60:** Identify range of plausible values as those that fall within 2 SD of statistic	**S6Q21:** Definition of skewness in general **S12Q45:** Definition of outliers in general **S14Q53:** Meaning of whether or not a sample statistic is an outlier **S15Q57:** Understand meaning of a distribution being normal to use 95% rule for constructing CI

Figure 2: Example of a Testlet from the Instrument

Instructions: Questions 1 – 4 pertain to the following scenario. Please read, and only consider, the following scenario when responding to these questions. Please provide a response to all questions. For multiple choice items, select the correct response from the options provided. For open-ended items, respond to the prompt using complete sentences. The answers to questions 1 – 4 are not necessarily related.

Scenario 1 (confidence level, sample size, conditions of the data; NOT error, p-value):

Two of the most popular flavors of ice cream at the Penn State Berkey Creamery are chocolate and vanilla. During one of the monthly sales meetings, a worker mentions that female students tend to prefer chocolate ice cream, whereas male students tend to prefer vanilla ice cream. As a result, the Creamery is interested in formally answering the following question: How different are the proportions of female and male college students who prefer chocolate ice cream at the Creamery?

Q1 (Reading). To construct a confidence interval for this study, the sample should consist of:
- (A) All college students
- (B) A subset of college students
- (C) All Berkey Creamery workers
- (D) A subset of Berkey Creamery workers

Q2 (Constructing Schema). The most appropriate inferential procedure to use in this scenario is a [answer 1] which would show the range of [answer 2] values for the parameter at the given level of confidence.
- (A) Confidence interval for a difference in means; unreasonable
- (B) Confidence interval for a difference in proportions; unreasonable
- (C) Confidence interval for a difference in means; reasonable
- (D) Confidence interval for a difference in proportions; reasonable

Q3 (Problem-Solving). Suppose that a researcher collects data on the sex and favorite ice cream flavor of a random sample of 100 Penn State students and finds the resulting 95% confidence interval of interest to be (0.177, 0.543). The researcher is concerned that the interval is too wide. The researcher could make the confidence interval narrower by:
- (A) Decreasing the confidence level from 95% to 90%
- (B) Decreasing the sample size from 100 students to 50 students
- (C) Increasing the confidence level from 95% to 99%
- (D) Increasing the number of outliers in the dataset

Q4 (Communicating). In the context of the research question of interest, interpret the 95% confidence interval found by the researcher: (0.177, 0.543). Is it plausible that there is a difference in female and male college students who prefer chocolate ice cream at the Creamery?

We are 95% confident that the proportion of female college students who prefer chocolate ice cream at the Creamery is between 0.177 and 0.543 higher than the proportion of male college students who prefer chocolate ice cream at the Creamery. Since this interval does not contain zero, there is significant evidence of a difference in female and male college students who prefer chocolate ice cream at the Creamery.

Future Work

- Use a natural language processing algorithm to score the open-ended responses at scale
- Study optimal feedback to provide to students based on their performance on the instrument
- Disseminate the instrument for use by statistics instructors in any environment for more generalizable results about the instrument's ability to effectively gauge students' understanding of confidence intervals

Acknowledgements

- Please see primary author for comprehensive list of references
- I would like to acknowledge the support of the Science Achievement Graduate Fellowship at Penn State
- Scan the following QR code to provide feedback or to indicate your interest in potential involvement with this project