Submitted by Mayank Sharma & Maulishree Pandey

SI-544: Introduction to Statistics and Data Analysis

11 December 2016

Comparison of Numerical and Visual Representations of Confidence Intervals

ABSTRACT

In this project, we compared two different representations of confidence intervals - numerical and visual. Our research was inspired by Robert Kosara's and Drew Skau's study which compared pie and donut charts[1]. There has been some research in the area of diagrammatic representation of confidence intervals[2], however, less research has been done in the comparison of visual and text representations of CI's. We use two different representations of CI's (Confidence intervals) to analyze the relationship between representation and interpretation. Interpretation is quantified as "scores" on a survey provided to controlled and treatment groups. The said relationship is studied by applying linear regression and rank sum test. Due to small sample size, the results are insignificant. However, data collected points to visual representation being more easy to interpret.

---

[1] Kosara, R., & Skau, D. (2016). Judgment Error in Pie Chart Variations. Eurographics Conference on Visualization (EuroVis) 2016, 1-6

[2] Frank S. T. Hsiao. "The Diagrammatical Representation of Confidence-Interval Estimation and Hypothesis Testing." The American Statistician, vol. 26, no. 5, 1972, pp. 28–29. www.jstor.org/stable/2683779.

1.0 INTRODUCTION

We decided to conduct a controlled experiment, and collect data by means of an survey. This was in line with the past research in the field. Plus, topic chosen for our study sat at the intersection of human computer interaction (HCI) and statistics. Thus, an experiment conducted directly with the users made perfect sense.

Our study utilized the concepts of confidence intervals, linear regression and experiment design from the SI544 course. Experiment design as a skill is much used in both statistics and HCI. By putting together these concepts, our study made for an interesting learning exercise as well as had the potential of providing utility to both disciplines. Studying interpretation and usability of confidence intervals would be useful to researchers and academicians from both domains. Additionally, if one representation fares better than the other, then it can be used to educate students about confidence intervals in introductory courses such as ours.

2.0 METHODOLOGY

Experiment Design

For the experiment, it was important that individuals in both groups had similar knowledge of statistics and confidence intervals. To appropriate for this requirement, we decided to conduct the study only within the students of University of Michigan's "Introduction to Statistics" or SI544 course. The course was conducted in "lecture-discussion" style format. Thus, it had a common lecture component, attended by all 40 students together. It had two discussion sessions for lecture and assignments review, with almost equal number of students in each discussion group. Both sessions were instructed

by the same GSI (graduate student instructor), and had the same format for instruction. Hence, there was no difference in type of instruction each student received. Given that it was an introductory course, we could safely assume that only students with little to no prior knowledge of statistics took the course. Since all students received the same instruction, regardless of the discussion session, and covered the same material, we could safely assume that they had similar exposure to statistics at the time of our study. This ensured that "knowledge", as a random variable, was kept constant across both groups. The only thing that varied was the representation itself.

To ensure randomness further we selected each discussion session as controlled and treatment group. Students decided at the very beginning of the course the session they wished to attend. Thus, composition of discussion sessions was independent of course performance, gender, knowledge or any such variables. By regulating the sample selection for controlled and treatment groups in this fashion, we controlled all qualitative and quantitative independent random variables. The only thing varying between both groups was the "representation of confidence intervals".

Survey Design

We designed a survey of questions for our study. There were eight questions in total, presented in multiple formats - true/false, multiple choice and single choice questions. Each question was worth one point on the survey. We made a conscious choice to divide the questions into two distinct categories. Four general questions related to

students' knowledge of confidence intervals; four questions related specifically to the confidence interval representation. Scores obtained on the former set of questions was represented by *bgScore,* while scores on the latter was represented by repS*core.* The reason behind this decision was to validate our assumption about randomness of sample in each group.

Lastly, the controlled group's survey comprised numerical representations of confidence intervals (fig 1). The survey provided to treatment group consisted of same questions, but with a visual representation (fig 2) of confidence intervals. The representations used in the survey were adopted from a previous research study[3] on confidence intervals. The questions were based on a population of normally distributed scores.The population standard deviation (σ) was unknown and the population mean (μ) was 50. Six samples (labeled A to F) are randomly drawn from the population. Every sample included 20 scores. The sample scores were provided along side the 95% CI. Based on this information, the following questions were asked. The questions used in the survey were as follows -

1. *In which of these CIs, population mean is contained? Check all that apply.*

   (This questions consisted of an image for the experiment group which showed different CI's in form of bars and simple intervals in text format for the control group)

[3] Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E. (n.d.). Robust Misinterpretation of Confidence Intervals.

2. *What do you think is true about the sample labeled E? Check all that apply*
3. *The CIs in which population mean is contained may have following characteristics (check all that apply*).
4. *What do the above 6 CIs suggest? Check all that apply.*
5. *Will 95/2 i.e. 47.5% CI be half or twice in length of 95% CI?*
6. *Would the 90% CI for the true mean be bigger than 95% CI ?*

Limitations of survey

The survey was conducted using Google forms. This constrained us from capturing more granular details like time taken to interpret the CI, time taken for each question, etc. We also faced a severe limitation with regard to capturing participant details. Due to participants' discomfort with disclosing their identity, we had to anonymize the survey. Thus, we couldn't study relationships between demographics variables and the final survey scores.

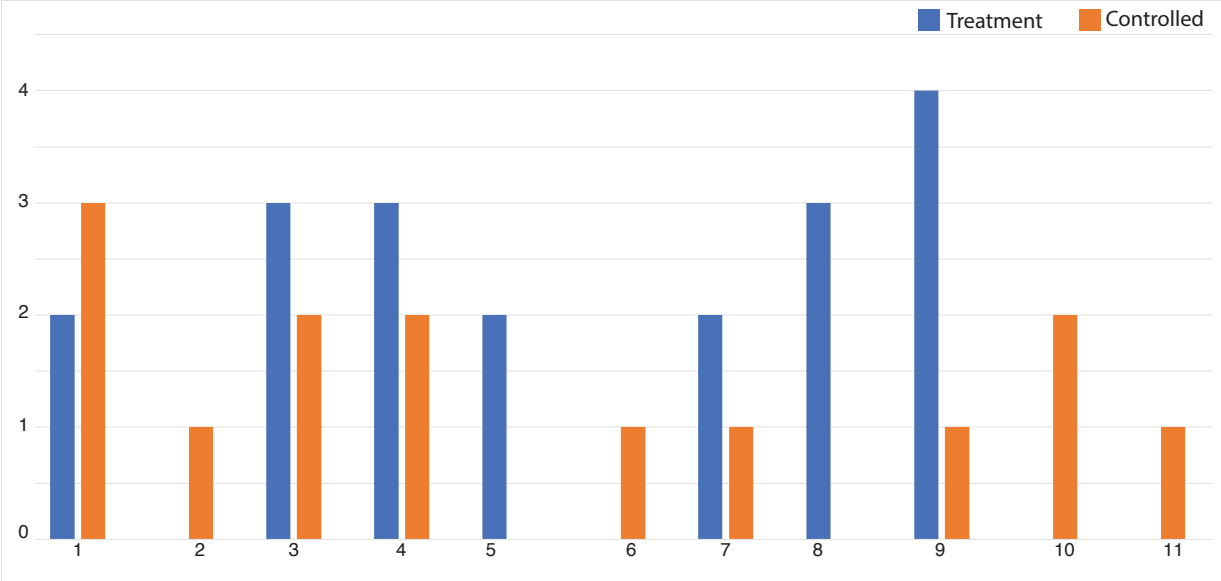
3.0 DISCUSSION OF RESULTS

Descriptive Analysis

We received a total of 20 responses from control (9) and experimental groups (11). The mean *repScore* i.e. score on questions specific to representations was 1.27 and 2.11 for controlled and experimental group respectively. The mean *bgScore* i.e. score on questions testing background knowledge was 2 and 2.44 respectively.

Rank-sum Test

Despite our efforts to keep the two groups similar in terms of background knowledge, there was a difference between bgScores. Before proceeding to analyzing



Graph 1 : Results of Wllcoxon Rank Sum Test on bgScores

repScores, we wanted to confirm the difference wasn't significant. A significant difference would imply that experiment group was better at their understanding of confidence intervals and thus interpreted the representations better.

Given the small size, we couldn't assume if the scores were normally distributed. Thus, instead of using T-test, we conducted Wilcoxon rank sum test on bgScores, which is a nonparametric test. Our null hypothesis was that there wasn't significant difference between background knowledge. Rank sum test on *bgScore* gave p-value *0.26*. Therefore,

```
Wilcoxon rank sum test with continuity correction

data:  x1 and y1
W = 35, p-value = 0.2558
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
 -1.000013e+00  7.151047e-05
sample estimates:
difference in location
             -0.999964
```

Table 1 : Results of Wllcoxon Rank Sum Test on bgScores

```
Call:
lm(formula = yscore ~ Rep + bscore + Rep * bscore, data = ci)

Residuals:
     Min       1Q   Median       3Q      Max
-2.28378 -0.40602  0.03256  0.72727  1.71622

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.2727     1.2044   0.226    0.824
Rep1          1.0786     1.6015   0.674    0.510
bscore        0.5000     0.5766   0.867    0.399
Rep1:bscore  -0.1892     0.7030  -0.269    0.791

Residual standard error: 1.153 on 16 degrees of freedom
Multiple R-squared:  0.1986,  Adjusted R-squared:  0.04837
F-statistic: 1.322 on 3 and 16 DF,  p-value: 0.302
```

Table 2 : Results of linear regression with repScores as dependent variable

we could safely conclude the background knowledge was not significantly different among individuals for both groups.

Linear Regression

We conducted linear regression with repS*core* as a dependant variable on b*gScore*, dummy variable *rep* (1 for image representation and 0 for numeric representation) and interaction term of both variables. The regression yielded the following equation:

$$Yscore \sim 0.5*bgScore + 1.0786*Rep - 0.1892*bgScore*Rep + 0.2727 \quad (1)$$

For image representation, the equation was:

$$Yscore \sim 0.311*bgScore + 1.3513 \quad (2)$$

For text representation, the equation was:

$$Yscore \sim 0.5*bgScore + 0.2727 \quad (3)$$

The results suggest that if background knowledge is same, then an individual may interpret visual representation more easily. This is explained by higher intercept for visual representation as compared to numerical representation. However, none of the coefficients are significant. This could be attributed to low sample size. We also conducted a rank sum test on *repScores* to check if mean scores were significantly different (1.27 for controlled, 2.11 for experiment) . The p-value was 0.1178, which couldn't be rejected at any reasonable level of significance. Thus, although there were hints, we couldn't conclude if visual representation did make a significant difference in interpretation and understanding of confidence intervals.

4.0 FUTURE WORK

The small sample size posed a limitation on our study. For concrete results, we would require a larger sample size in both groups. Apart from background knowledge, we also want to study time taken to interpret the confidence intervals in either of the representations. Other variables worth analysis include background in information and data visualization, demography, gender, etc. We intend to expand upon our work to include the aforementioned and conduct a more comprehensive survey with students.

Lastly, we would also use "think aloud" techniques with members of our cohort who were part of the experiment group. This would help us understand their "mental model" behind interpretation of the visual representation. We believe the results would be useful in designing the representation itself for better usability.

5.0 CONCLUSION

We believe this research has potential of being useful in multiple research domains. By incorporating more variables in the survey, we can get conclusive results. Having said that, we acquired much experience in way of experiment design and statistics with this study. We intend to take this forward in the ensuing semesters to expand our knowledge, and to contribute to the field of HCI and statistics.

Appendix

Numerical representation based survey

Confidence Intervals Survey

In the question below, a population of normally distributed scores has been created.The population standard deviation (σ) is unknown and the population mean (μ) is 50.This population is shown in the figure below. Six samples (labeled A to F) are randomly drawn from the population. Every sample includes 20 scores. The sample scores are provided along side the CI. Their 95% Confidence interval is presented along with the circle. Based on this information, please answer the questions below.

Note : Confidence interval is abbreviated as CI in questions below.

A : [45 , 65] Sample score : 55

B : [44 , 64] Sample score : 54

C : [37 , 57] Sample score : 47

D : [38 , 64] Sample score : 51

E : [35 , 47] Sample score : 41

F : [43 , 63] Sample score : 53

1. In which of these CIs, population mean is contained? Check all that apply.

- [ ] A
- [ ] B
- [ ] C
- [ ] D
- [ ] E
- [ ] F
- [ ] None of these

2. What do you think is true about the sample labeled E? Check all that apply

- [ ] It has a relatively low std error than other CIs
- [ ] Sample mean is close to 40
- [ ] Population mean is in the range 35 to 46

3. The CIs in which population mean is contained may have following characteristics (check all that apply)

- [ ] Lower degrees of freedom
- [ ] More representative sample
- [ ] Sample mean close to 50

4. What do the above 6 CIs suggest? Check all that apply.

- [ ] 95% of these CIs may contain the population mean
- [ ] 5% of these CIs may not contain the population mean
- [ ] The population mean doesn't change even if the CI changes

5. Would the 90% CI for the true mean be bigger than 95% CI ?

- ○ It will be bigger
- ○ It will be smaller
- ○ None of these

6. Will 95/2 i.e. 47.5% CI be half or twice in length of 95% CI?

- ○ It will be twice in length
- ○ It will be half in length
- ○ None of these

Mean: B, Confidence interval : [a, c]

7. In the CI above, what does 'c' represent? Check all that apply.

- [ ] Upper limit of confidence interval in which population mean can lie
- [ ] Sample mean + 1.96*std error
- [ ] None of these

Mean: B, Confidence Interval : [a, c]

8. In the CI above, what does 'B' represent? Check all that apply.

- [ ] Sample mean
- [ ] Population mean
- [ ] None of these
- [ ] All of these

SUBMIT

Visual representation based survey

Confidence Intervals Survey

In the question below, a population of normally distributed scores has been created.The population standard deviation (σ) is unknown and the population mean (μ) is 50.This population is shown in the figure below. Six samples (labeled A to F) are randomly drawn from the population. Every sample includes 20 scores. The sample scores are shown with small green circles. Their 95% Confidence interval is presented along with the circle. Based on this information, please answer the questions below.

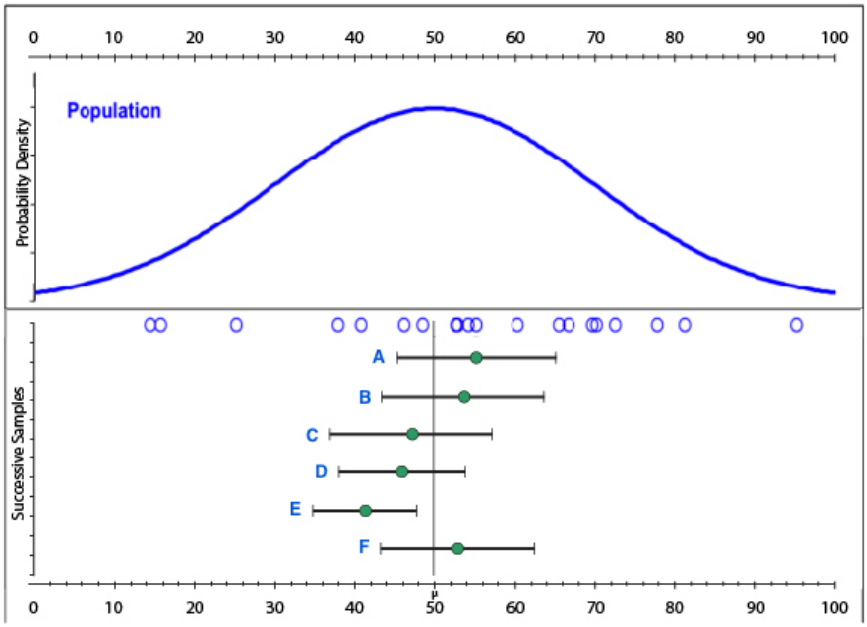
Note : Confidence interval is abbreviated as CI in questions below.

Email address *

Valid email address

This form is collecting email addresses. Change settings

Image title



1. In which of these CIs, population mean is contained? Check all that apply.

- [ ] A
- [ ] B
- [ ] C
- [ ] D
- [ ] E
- [ ] F
- [ ] None of these

2. What do you think is true about the sample labeled E? Check all that apply

- [ ] It has a relatively low std error than other CIs
- [ ] Sample mean is close to 40
- [ ] Population mean is in the range 35 to 46

3. The CIs in which population mean is contained may have following characteristics (check all that apply)

- [ ] Lower degrees of freedom
- [ ] More representative sample
- [ ] Sample mean close to 50

4. What do the above 6 CIs suggest? Check all that apply.

- [ ] 95% of these CIs may contain the population mean

- [ ] 5% of these CIs may not contain the population mean
- [ ] The population mean doesn't change even if the CI changes

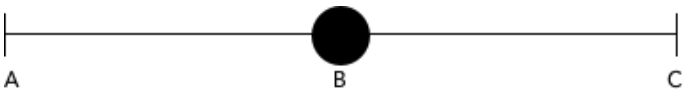
5. Would the 90% CI for the true mean be bigger than 95% CI ?

- ( ) It will be bigger
- ( ) It will be smaller
- ( ) None of these

6. Will 95/2 i.e. 47.5% CI be half or twice in length of 95% CI?

- ( ) It will be twice in length
- ( ) It will be half in length
- ( ) None of these

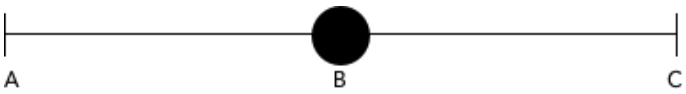
95% Confidence Interval



7. In the CI given above, what does C represent?

- [ ] Upper limit of confidence interval in which population mean can lie
- [ ] Sample mean + 1.96*std error
- [ ] None of these

8. In the CI, what does B represent



- Sample mean
- Population mean
- None of these
- All of these

Data from the survey

Table :

Rep	yscore	bscore	Total_Score
1	2	1	3
1	0	2	2
1	3	2	5
1	3	3	6
1	2	3	5
1	0	3	3
1	2	1	3
1	3	4	7
1	4	3	7
0	3	3	6
0	1	3	4
0	2	1	3
0	2	2	4
0	0	2	2
0	1	2	3
0	1	2	3
0	0	1	1
0	1	2	3
0	2	2	4
0	1	2	3

Rep – Dummy variable representing visual (1) and numeric (0) representation
yscore – Score for Representation based questions
bscore – Score for background related questions

R Code

```
#Reading the file
ci <- read.csv("confintervals.csv")
ci$Rep <- factor(ci$Rep)
summary(ci)

#mean of text and images
mbtext <- mean(ci$bscore[ci$Rep == 0])
var_mbtext <- var(ci$bscore[ci$Rep == 0])
mbimg <- mean(ci$bscore[ci$Rep == 1])
var_mbimg <- var(ci$bscore[ci$Rep == 1])

mytext <- mean(ci$yscore[ci$Rep == 0])
myimg <- mean(ci$yscore[ci$Rep == 1])

#Ranksum test
x1 <- ci$bscore[ci$Rep == 0]
y1 <- ci$bscore[ci$Rep == 1]
wilcox.test(x1,y1,mu =0, alt = "two.sided", conf.int = TRUE,
paired = F, exact = F, correct = T)



#Linear Regression
score.lm <- lm(yscore ~ Rep + bscore + Rep*bscore, data = ci)
summary(score.lm)
```