Error bars

Tae Hyon Lee

December 12th, 2018

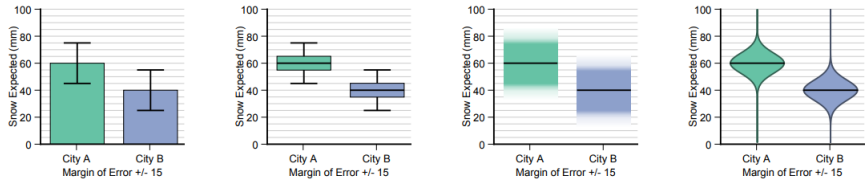
https://github.com/leeth7830/Error-Bars-User-Study

1 Paper

Error Bars Considered Harmful: Exploring Alternate Encodings for Mean and Error by Michael Correll and Michael Gleicher

2 Goal

The goal of the paper is to investigate the drawbacks with this standard encoding (Bar chart) and consider a set of alternatives (Modified box plot, gradient plot, violin plot) designed to more effectively communicate the implications of mean and error data to a general audience.



(a) **Bar chart** with error bars: the height of the bars encodes the sample mean, and the whiskers encode a 95% t-confidence interval.

(b) **Modified box plot**: The whiskers are the 95% t-confidence interval, the box is a 50% t-confidence interval.

(c) **Gradient plot**: the transparency of the colored region corresponds to the cumulative density function of a t-distribution.

(d) **Violin plot**: the width of the colored region corresponds to the probability density function of a t-distribution.

Figure 1: The four chart types presented in the paper

3 Scope

The paper presented three types of experiments. First experiment is the one-sample judgement experiment that presents participants with a single sample mean, postulates a potential outcome and asks participants to reason about the relationship this potential outcome to the sample. The second experiment is the textual one-sample judgement experiment that evaluates another potential

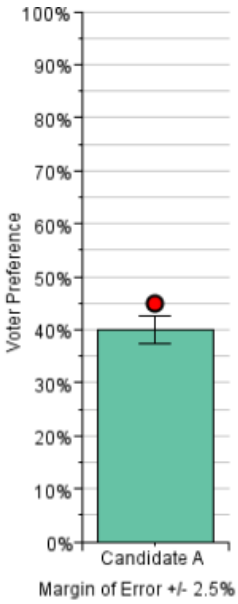
approach to mitigating within-the-bar bias, which is to abstract some of the information from the bar chart itself into text. The third experiment is the two-sample judgements which evaluates the alternative encoding in a setting that resembles how these visualizations are frequently used in practice: to compare samples and make predictive inferences about the differences in mean, given the error. For this survey, I chose to replicate the first experiment.

4 Setup

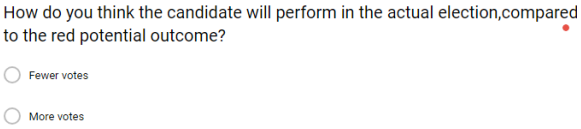
First, I contacted the author, Michael Correll, to let him know that I will be replicating the study and ask if there are any suggestions. He let me know that if he were to run the studies again today, he would included more strict exclusion criteria because of the quality of Turk participants pool. He also provided me with a website (http://graphics.cs.wisc.edu/Vis/ErrorBars/) that has more details on how they conducted the study, what images they chose and what was the result of the analysis.

Second, I chose which of the experiments to replicate. Due to limited time and budget, I can only have maximum of 100 participants so I chose to replicate the first experiment (one-sample judgement) with the same wording (Election).

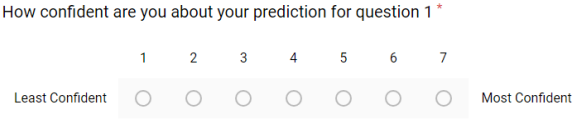
Third, I replicated the survey by identifying the questions asked and visualizations used in the paper. The type of questions that was asked in the paper was available in the paper and the website that was provided by the author. For each type of graph, I created 36 sets of questions, each of which had 3 questions and one image in a set. Participants was shown an image of a graph and was asked three questions about the graph for every set. The questions asked were the following: 1. How do you think the candidate will perform in the actual election, compared to the red potential outcome? (Fewer votes, more votes) 2. How confident are you about your prediction for question 1? (1-7) 3. How likely (or how surprising) do you think the red potential outcome is given the poll? (1-7) Each sets had varying margin of errors (2.5, 5.0, 7.5, 10.0, 12.5, 15.0) and varying differences between the red dot and the mean. For this survey, I created 4 different sets of survey for each chart types (bar chart, box plot, gradient plot, and violin plot), and asked 25 participants to answer 36 sets of questions for each type of graph. In total, we had 100 participants answering a total of 3600 sets of questions.



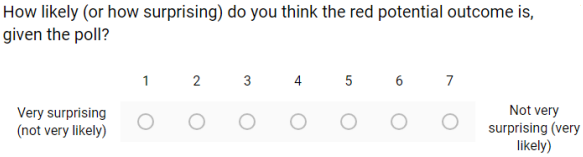
(a) image



(b) question 1



(c) question 2



(d) question 3

Figure 2: Example of questions shown to the participant

Fourth, with the help of Professor Alark from University of San Francisco, I deployed the survey on Amazon Mechanical Turk and received results within 5 hours of deployment. Due to budget constraint, we were not able to add criteria to only allow qualified participants. As a result, the quality of the results were

not great as discussed more later.

Fifth, I wrote a python script using pandas library to transform the raw data into a format that can be analyzed easily. It also created a separate file that shows the information about the users such as accuracy and time spent on the survey to filter out bad quality data. The script and output file is available in the Github link provided above. I also manually checked some of the results to identify patterns in answers such as straight line answers.

Sixth, I imported the data to Jupyter notebook to calculate overall accuracy, average accuracy for each question, standard deviation, cdf, 1-cdf, maxcdf, and pdf. Then I used statsmodels and scipy libraries to perform ANCOVA on the data. More details available in the next section.

5 Result

There are three hypothesis presented in the paper. Refer to below. For hypothesis 1, the author expects the participants' responses to follow expected behavior. To be more specific, for question 1, participants should get the similar amount of correct answers on average. For question 2, the author expected participants' confidence in answer to follow cdf of the type of question. Lastly, for question 3, the likelihood of the proposed outcome would follow pdf of the type of question. In the paper, this hypothesis was generally supported with R-squared followed of 0.805 and 0.842 for question 2 and 3. But from our experiment, the R-squared value for both was close to 0, meaning there was no correlation between the two. Refer to figure 4.

H1 Participant responses would generally follow expected behavior. That is, participant responses to question 1 would "follow the sample mean" — if the red dot is above the sample, assume the real election will be lower than the red dot, and vice versa. The answers to question 2 should correlate with the cdf of the t distribution given the data, and the answers to question 3 should correlate with the pdf. Both cdf and pdf are modulated by both the difference in value between the predicted outcome and the sample mean, and the margin of error of the sample.

H2 The non-symmetric encoding (bar charts) would exhibit within-the-bar bias — proposed outcomes within the bar would be seen as likelier than outcomes outside of the bar. Symmetric encodings (box, violin, and gradient plots) would not have this bias.

H3 The proposed encodings, which encoded the t-distribution in a non-binary way (gradient and violin plots), would provide more accurate and more confident judgments about the t-distribution than the binary encodings (bar charts and box plots).

Figure 3: The three hypothesis presented in the paper

For hypothesis 2, the author expected the non-symmetric encoding (bar charts) to exhibit within-the-bar bias. The proposed outcomes are within the bar (under the mean of bar chart) would be seen as likelier than outcomes outside of the bar. This was the case in the paper as shown in figure 5 where if the proposed outcome was within the bar, then likelihood of the outcome was

	Overall Accuracy	R^2 (confidence, cdf)	R^2 (likelihood, pdf)
Original Study	87.1%	0.805	0.842
Replication	45.1%	0.002	0.002

Figure 4: Comparison of the result of accuracy, cdf, and pdf

higher. However, as shown in figure 6, that was not the case for our experiment.

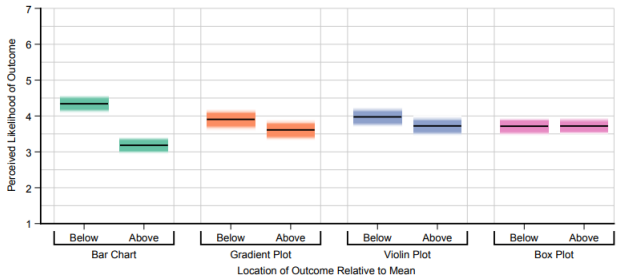
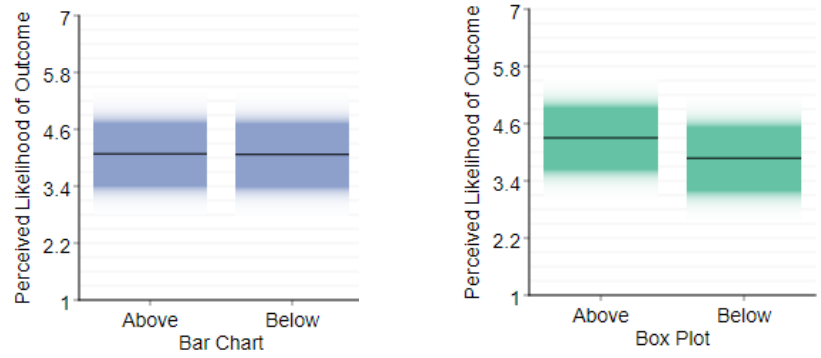
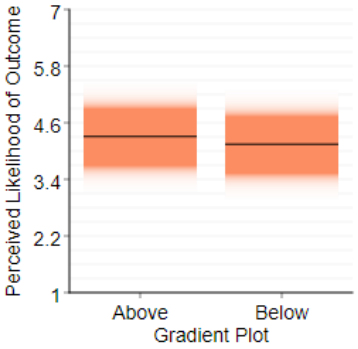


Figure 5: Comparison of the four charts when the red dot was above or below the mean in the original experiment

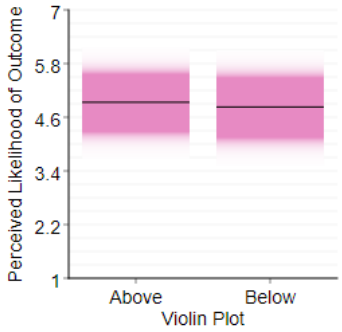


(a) bar chart

(b) box plot



(c) gradient plot



(d) violin plot

Figure 6: Perceived likelihood of outcome for bar chart, gradient plot, violin plot, and box plot when the proposed outcome was below or above the mean from our experiment

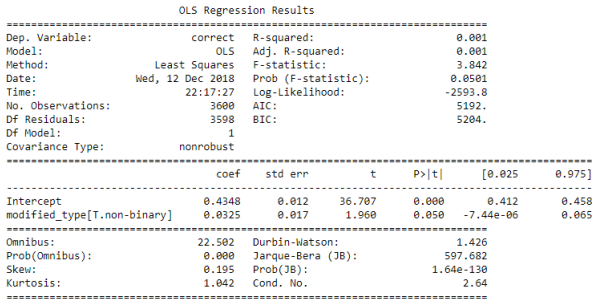
The hypothesis 3 compared the gradient and violin plot versus the bar chart and box plot. The author expected the gradient and violin plot to perform better in terms of accuracy and confidence than the bar chart and box plot. This was generally supported in the original experiment but not in our experiment as shown in figure 7 and 8.

6 Discussion

The results are significantly different from the experiment from the paper, most likely due to the fact that the data was collected from a low quality pool of participants. Nevertheless, we saw that bar chart performed the worst and violin plot performed the best out of the four, which is somewhat supported the author's third hypothesis.

	Bar Chart	Box Plot	Gradient Plot	Violin Plot
Accuracy - Original	83.2%	87.4%	88.5%	89.2%
Accuracy - Replication	37.8%	48.1%	39.7%	53.5%
Confidence - Original	4.86	4.86	5.12	5.06
Confidence - Replication	4.3	4.14	4.55	4.88

Figure 7: Comparison of the four charts for confidence and accuracy



OLS Regression Results

Dep. Variable:	correct	R-squared:	0.001
Model:	OLS	Adj. R-squared:	0.001
Method:	Least Squares	F-statistic:	3.842
Date:	Wed, 12 Dec 2018	Prob (F-statistic):	0.0501
Time:	22:17:27	Log-Likelihood:	-2593.8
No. Observations:	3600	AIC:	5192.
Df Residuals:	3598	BIC:	5204.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.4348	0.012	36.707	0.000	0.412	0.458
modified_type[T.non-binary]	0.0325	0.017	1.960	0.050	-7.44e-06	0.065

Omnibus:	22.502	Durbin-Watson:	1.426
Prob(Omnibus):	0.000	Jarque-Bera (JB):	597.682
Skew:	0.195	Prob(JB):	1.64e-130
Kurtosis:	1.042	Cond. No.	2.64

Figure 8: Summary statistics showing correlation between accuracy and binary/non-binary type. There was just enough significant difference to show that the two types were different when determining accuracy

7 Future work

If I were to run this experiment again in the future, I would have spent more time or money collecting quality data. As noted from the author earlier, the quality of online survey platform has gone down significantly over the past years. It is very important to choose the right platform and the right participant to get more accurate results.