Error Bars Considered Harmful

Michael A. Correll*
University of Wisconsin-Madison

Michael Gleicher†
University of Wisconsin-Madison

ABSTRACT

Confidence intervals, standard error, and generalized error rates are typically visualized with "error bars" – thin strokes that are superimposed over the mean. This work presents the results of crowd-sourced experiments which illustrate that viewers misinterpret these encodings even at the most basic level (where one would hope larger margins of error reduce the confidence in judgments about means). We then present evaluations of three alternate (or supplemental) visual encodings for the same task and show that choice of visual encoding can result in viewers who make decisions which are better informed by the margins of error.

Index Terms: I.3.6 [Computing Methodologies]: Computer Graphics—Methodology and Techniques

1 INTRODUCTION

The power of statistical inferences is highly connected to the notion of error. Error is often visualized as "error bars" centered around a sample mean. Commonly, the mean is encoded as the height of a bar chart – e.g., of the 15 papers of InfoVis 2012 which presented statistical results graphically, 13 employed some form of error bars, and 7 used bar charts to encode the mean of a sample. The ability of viewers to interpret statistical information is highly sensitive to the choice of encoding [4]. In this poster I will present the results of a crowd-sourced experiment illustrating that viewers habitually misinterpret error bars and bar charts when making decisions, but that different visual encodings of the same data can inform decision-making which is more in-line with statistical expectations.

Both components of the bar chart with error bars are prone to misinterpretation by audiences. Prior work has shown that using bars to represent the mean of a sample creates a false perception that values "within" the visual area of the bar are likelier than values outside of the bar [7]. Even without an associated bar chart, error bars themselves have several issues: they are ambiguous (the same error bar can represent a confidence interval, standard error, min/max, interquartile range, standard deviation, or some other error property; these are often not labeled in the diagram itself), they rely on sophisticated knowledge of the statistics behind them (even self-professed experts do not correctly interpret the extent to which error bars are connected with inferential certainty [1]), and they do not carry any semiotic connection with uncertainty (they use solid lines and width to encode value, instead of something like blurriness or transparency [5]).

Violin plots [3] are an alternate visual encoding which preserve details about the underlying distribution that are lost in encodings such as box-and-whiskers diagrams. Unfortunately making a judgment about statistical confidence would require extracting the cumulative probability, which visually would mean performing partial integration on the "violin." Figure 1b presents an example.

Another alternate encoding we propose is the "Gradient plot." The Gradient plot relies on the existing semiotic associations of α-blended colors and uncertainty. Emanating from the mean is a color which is gradually blended with the background. The extent of the α-blending is related to how wide a particular confidence interval (in this case a t-confidence interval) would have to be in order to include a value that distance from the mean. To preserve the notion of a recommended p-value (the statistical, rather than graphical, notion of an α value), all values within the 95% t-confidence interval are fully opaque, with a fall off to the (fictional) "100%" confidence interval which would be fully transparent. Figure 1c presents an example. While α channel is not sufficiently perceptually uniform or discriminate to afford precise readings of continuous values, the *semiotic* encoding (where the closer we get to the sample, the more likely a population mean is to be in that area) is generally preserved.

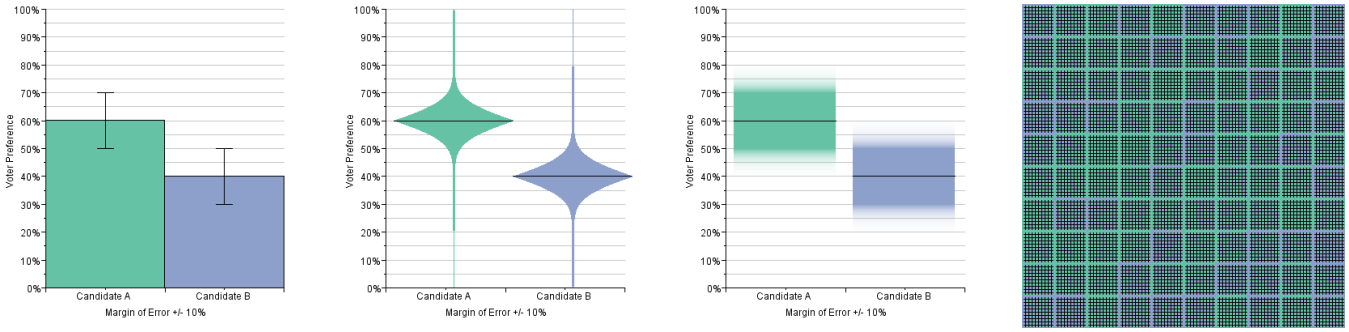
The last encoding functions as a secondary, supplemental encoding which communicates the possibility space rather than the sample means. We call these encodings "Pangloss plots" as they show a quasi-random set of "best possible worlds." The example presented in this poster is that of an election, but the encoding is extensible to a wider set of samples drawn from multiple categories. One hundred potential population means are drawn at random from a probability distribution (in this application a t-distribution). The p-value of a one-tailed t-test is used as a metric to see if the sampling procedure over-sampled one category to "win" more often than the other, and samples are discarded and generated until the number of times a particular category "wins" is in line with expectations. Each "mock election" is then encoded as an internally permuted block of one hundred "voters." A colored stroke is drawn around the entire mock election reflecting the overall winner. The result is a "quilt" of different possible outcomes. Figure 1d presents an example. Prior research has shown that viewers can accurately estimate the mean values of particular sections of this quilt [2].

While no choice of encoding is a substitute for detailed statistical literacy, by avoiding some of the known deficiencies in standard encodings of error it is possible that the lay audience will be better informed about the statistical inferences being made in visual presentations of data. One basic metric of this improvement is the strength of the correlation between the p-value of a particular statistical inference and the self-reported confidence of a viewer in their own inferences about the data. That is, given two data sets which are equal in all other respects, we would hope that viewers would be *less confident* in decisions made using the data set with a *larger margin of error*. This confidence can be measured both explicitly with self-reported indicators of decision confidence, but also implicitly in how often a viewer will *refrain from deciding* about the relationships present in the data.

2 EXPERIMENT

In order to gauge how well different visual encodings of margins of error inform decision making, we conducted an experiment on Amazon's Mechanical Turk platform. Participants were given a brief tutorial about margins of error and sampling, and then introduced to a sample decision problem: given a graphic of sample polling data from a fictional election, which of two candidates is likely to win the election? Participants were asked to make a prediction about the general election from the sample (or say that the election was too close to call), and then state their confidence in this prediction on a Likert scale from 1 (most confident) to 7 (least confident). Participants were exposed to one type of encoding as

*e-mail: mcorrell@cs.wisc.edu
†e-mail: gleicher@cs.wisc.edu



(a) Bar chart with error bars. The baseline encoding for this task.

(b) "Violin Plot" encoding the likely distribution of population means with width.

(c) "Gradient Plot" encoding the likely distribution of population means with color.

(d) "Pangloss Plot" showing potential election outcomes given the sample data.

Figure 1: The experimental conditions in this work. The experimental task was to decide which of two political candidates (A or B) was most likely to win the election given sample polling data, and then to state one's confidence in this decision. Participants saw only one type of encoding (1a,1c,1b) as a between-subjects factor. Another between-subjects factor was whether or not the participant was also presented with a Pangloss plot showing simulated election results (1d) in addition to the primary encoding, for a total of 6 levels of the encoding factor.

a between-subjects factor (see Fig. 1 for a list of conditions). We recruited 6 for each of the 6 levels of encoding, for a total of 36 participants. There were two within-subjects factors: the difference between means (voter preferences of +/- 0,5,10, and 25% between candidates "A" and "B"), and the margin of error (m.o.e. of 5,10, and 15%). Participants saw one example of each combination of mean difference and margin of error, for a total of 21 different stimuli per participants. After the task, we had participants self-report general demographic data, as well as risk aversion using an established scale (General Risk Assessment, with a previously measured coefficient α of 0.72 [6]).

Our hypotheses were generated optimistically based on the intended meaning of margins of error for the mean prediction task (margins are larger, predicting the population mean becomes more difficult). We also had a hypothesis about the inclusion or exclusion of a Pangloss plot: by explicitly showing the possibility space (including examples where a sample winner was *not* the population winner), participants would be less confident in the predictions they did make. Our primary hypotheses were:

H1 As the margins of error *increase*, participants will be *less likely* to make a prediction.

H2 As the margins of error *increase*, participants will be *less confident* when they do make a prediction.

H3 If the secondary Pangloss encoding was *included*, participants would be *less confident* in their predictions.

3 Results & Discussion

Our results *support* **H1** – a two-way analysis of variance (ANOVA) on the likelihood of participants to make a decision had a main effect of size of margin ($F(2,996)=3.05$, $p = 0.047$). This effect was consistent across encoding type (encoding was not a main effect, $F(2,996)=0.362$, $p = 0.70$).

Our results *support* **H2** – a two-way ANOVA on decision confidence had a main effect of size of margin ($F(2,693)=10.4$, $p \leq 0.0001$). *However*, choice of encoding was also a main effect ($F(2,693)=10.1$, $p \leq 0.0001$) – a post-hoc Tukey's test of Honest Significant Difference shows that participants who saw a gradient plot were significantly less confident than those who saw the other encodings. There was also a marginal interaction between encoding and margin ($F(4,693)=2.17$, $p = 0.07$). For standard bar charts there was no statistically significant difference in confidence between 10% and 15% margins of error, and for Violin plots there was no statistically significant difference in confidence between *any* of the levels of margin of error. Post-hoc comparison of means using one-tailed t-test showed that **only for Gradient plots was there a monotonic, statistically significant difference in confidence as margins increased** (although with a Bonferroni correction, difference between margins of 5% and 10% becomes marginal ($p = 0.03$)).

Our results *partially support* **H3** – the presence of a Pangloss plot was not a main effect ($F(1,693)=0.82$, $p = 0.36$), but there was a marginal interaction between encoding type and the presence of a Pangloss plot ($F(2,693)=2.98$, $p = 0.05$). A post-hoc t-test shows that adding a Pangloss plot to supplement a Violin plot results in significantly less confidence in predictions overall ($p = 0.02$).

In summary, the standard encoding for showing sample means with associated error – bar charts with error bars – does not correctly inform viewers about the relative likelihood of particular outcomes. Violin plots, although they do present more information about likely distributions of means, do not perform any better. Gradient plots correct some of the known deficiencies of error bars, and utilize existing semiotic connections to guide lay audiences to statistically-informed decisions.

Acknowledgements

This work was supported in part by NSF awards CMMI-0941013 and IIS-1162037 and NIH award R01 AU974787.

References

[1] S. Belia, F. Fidler, J. Williams, and G. Cumming. Researchers misunderstand confidence intervals and standard error bars. *Psychological methods*, 10(4):389–96, Dec. 2005.

[2] M. Correll, D. Albers, S. Franconeri, and M. Gleicher. Comparing averages in time series data. In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems*, pages 1095–1104. ACM, may 2012.

[3] J. Hintze and R. Nelson. Violin plots: a box plot-density trace synergism. *The American Statistician*, 1998.

[4] H. Ibrekk and M. G. Morgan. Graphical communication of uncertain quantities to nontechnical people. *Risk analysis*, 7(4):519–529, 1987.

[5] A. M. MacEachren, R. E. Roth, J. O'Brien, B. Li, D. Swingley, and M. Gahegan. Visual semiotics & uncertainty visualization: An empirical study. *Visualization and Computer Graphics, IEEE Transactions on*, 18(12):2496–2505, 2012.

[6] C. A. Mandrik and Y. Bao. Exploring the concept and measurement of general risk aversion. *Advances in Consumer Research*, 32:531, 2005.

[7] G. E. Newman and B. J. Scholl. Bar graphs depicting averages are perceptually misinterpreted: The within-the-bar bias. *Psychonomic bulletin & review*, 19(4):601–607, 2012.