Fitting Bell Curves to Data Distributions using Visualization

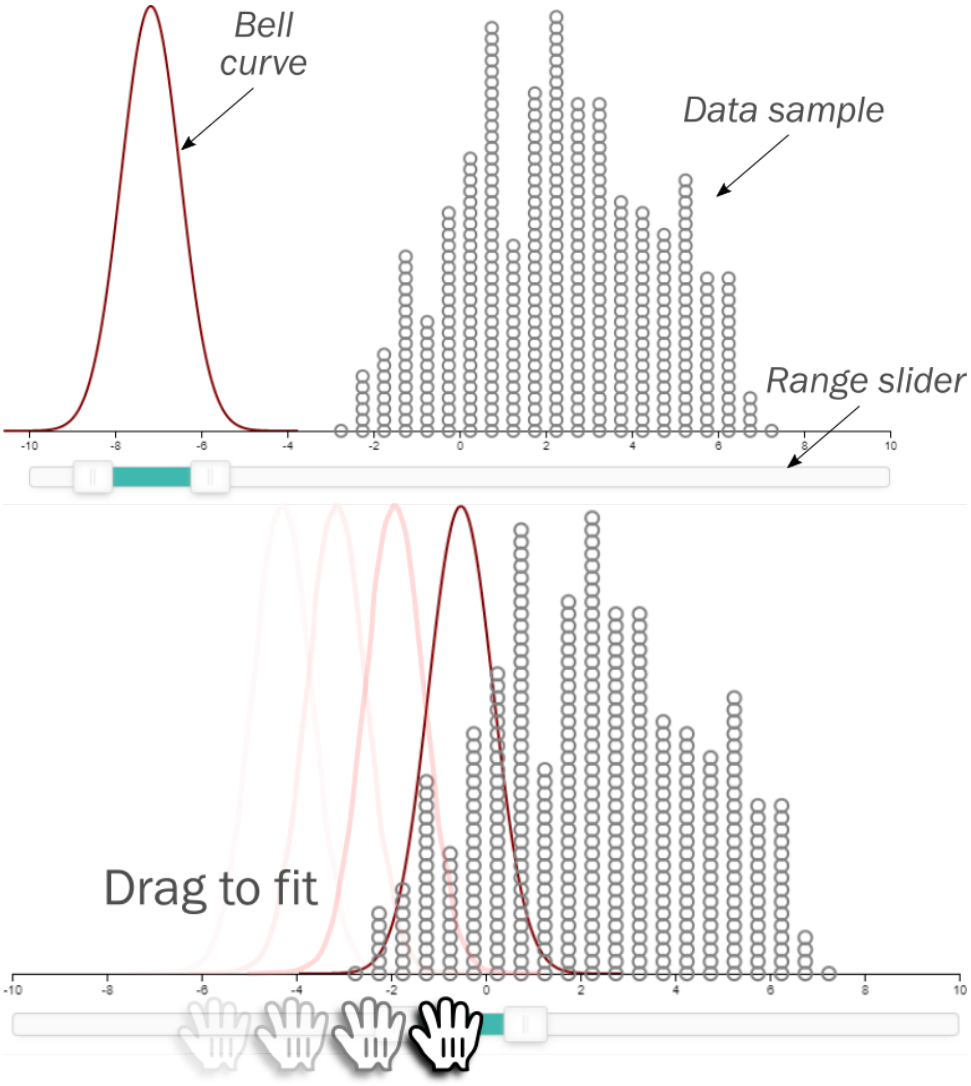
Supplementary Materials

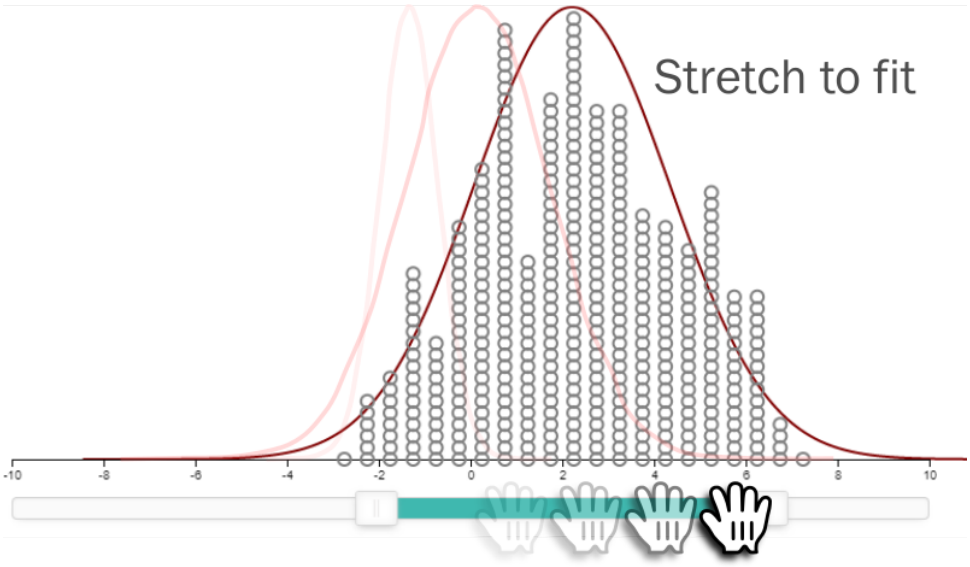
APPENDIX A
ADDITIONAL MATERIAL

All of the user study materials have been made available on our OSF website: https://osf.io/r3jpn/

APPENDIX B
STUDY INSTRUCTIONS

We are asking you to fit a smooth *bell curve* to a noisy data *sample*. The sample has been randomly drawn from a source that is itself in the shape of a bell curve; imagine that it represents the heights of a large group of people, where some individuals are tall or short, but most are in the middle. However, since the sample is small and randomly drawn, its shape can be a little rough. Your task is to fit the bell curve on top of the data sample by adjusting its center and degree of spread, as a best guess of the approximate shape of the source population.





The gray dots here represent each sample point. As you can see from the image, the goal is to "fit" the red curve on top of the gray dots. This means two things: (1) The curve should be **centered** on top of the gray dots. The curve should be **stretched** so that it approximates the shape of the dots.

When fitting, the goal is **not** that all dots are inside the curve. Instead, you want to find the best approximate fit between the dots and the curve. Once you are satisfied, click "**submit**."

APPENDIX C
ADDITIONAL VISUALIZATIONS

Figures 1 and 2 show individual participant performance for mean error and standard deviation error, respectively.

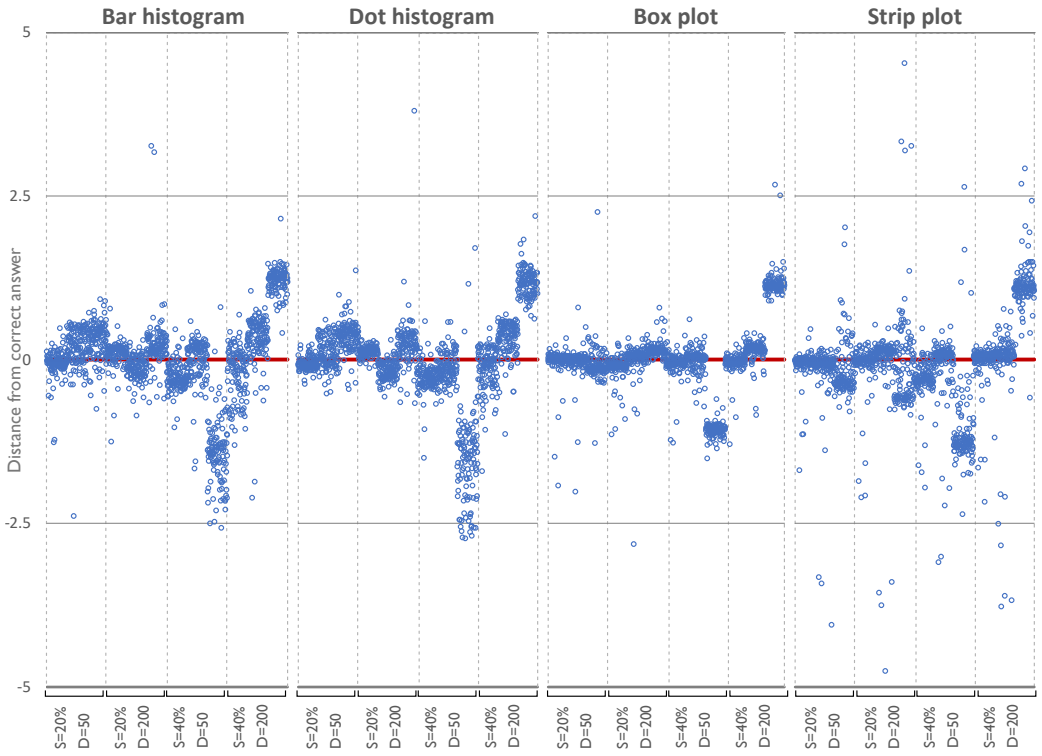


Fig. 1: **Mean error performance.** Distance between true values and individual participant estimates of dataset means. The red line denotes zero distance (correct answer). Each region for a specific spread *S* and dataset size *D* represent sets of trials, with the same datasets used for each set).

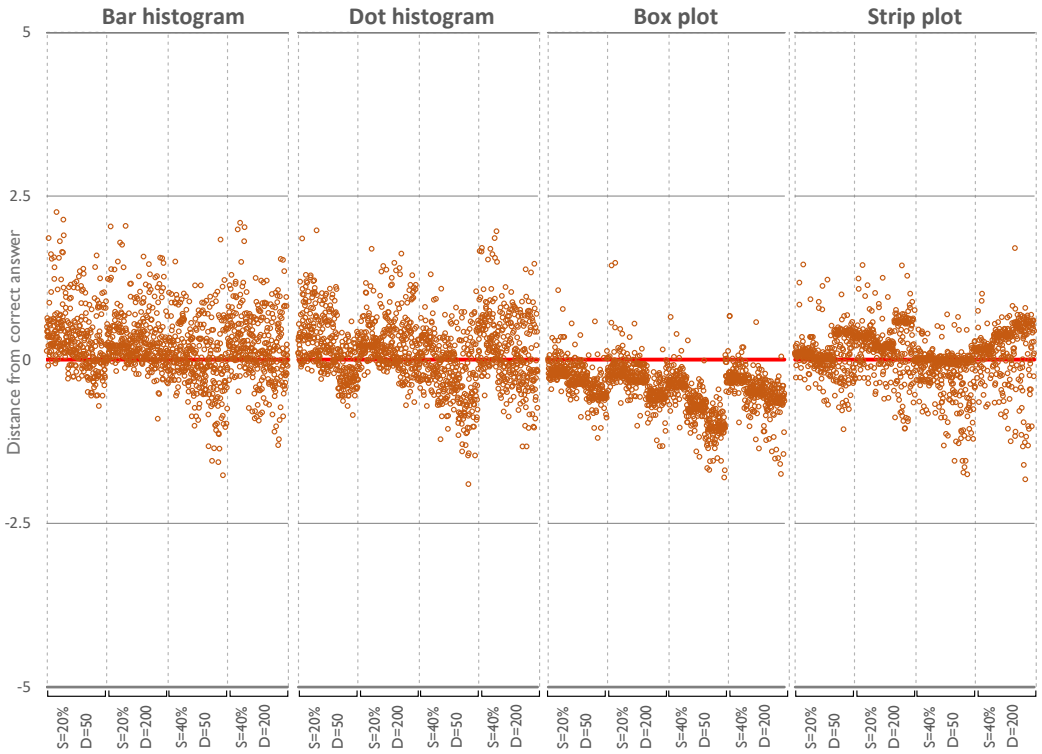


Fig. 2: **Standard deviation error performance.** Distance between true values and individual participant estimates of dataset standard deviations. The red line denotes zero distance (correct answer). Each region for a specific spread *S* and dataset size *D* represent sets of trials, with the same datasets used for each set).