# Fitting Bell Curves to Data Distributions using Visualization

Eric Newburger, Michael Correll, and Niklas Elmqvist, *Senior Member, IEEE*

**Abstract**—Idealized probability distributions, such as normal or other curves, lie at the root of confirmatory statistical tests. But how well do people understand these idealized curves? In practical terms, does the human visual system allow us to match sample data distributions with hypothesized population distributions from which those samples might have been drawn? And how do different visualization techniques impact this capability? This paper shares the results of a crowdsourced experiment that tested the ability of respondents to fit normal curves to four different data distribution visualizations: bar histograms, dotplot histograms, strip plots, and boxplots. We find that the crowd can estimate the center (mean) of a distribution with some success and little bias. We also find that people generally overestimate the standard deviation—which we dub the "umbrella effect" because people tend to want to cover the whole distribution using the curve, as if sheltering it from the heavens above—and that strip plots yield the best accuracy.

**Index Terms**—Graphical inference, visual statistics, statistics by eye, fitting distributions, crowdsourcing.

## 1 Introduction

Many crucial visualization tasks rely on not just an assessment of individual values, but a holistic assessment of the overall *distribution*: identifying points as outliers, judging variability, assessing the appropriateness of various parameterized statistical tests, or even just building up a picture of likely and unlikely values: all require fitting, implicitly or explicitly, distributions to sample data. While prior work has examined aggregate or ensemble graphical perception tasks [1], the specific fitting of curves of idealized distributions we believe is both understudied and has the potential to inform the design of statistical graphics, especially for novice users. We choose the fitting of normal distributions to sample data as a graphical perception "fruit fly" [2]: a relatively simple and well-understood problem that is nonetheless important for a variety of tasks in statistics.

Our work stems from an open question in visualization: the extent to which human visual judgments from statistical graphics can operate as a sort of "visual statistics" [3] or "graphical inference" [4]: that is, the extent that we can rely on our ability to read or aggregate information in charts to assess effect sizes, provide evidence against null hypotheses, or assess trends in our data. If these are abilities are robust, they point to the possibility to augment or even replace formal statistical tests with visual appraisals. Yet, there is also evidence that in some ways the visual estimation of statistical properties is insufficient, non-anologous, or otherwise disconnected from the results of formal statistical tests. Bias [5], satisficing strategies [6], and perceptual proxies [7], [8] can produce mismatches between statistical assessments of data and human judgments or estimates. It therefore is necessary to perform empirical assessment of human performance and measurement of human biases in reading statistical graphics before promoting their use as inferential or confirmatory tools.

In this paper, we present results from a preregistered and crowdsourced user study investigating how well members of the general population are able to fit normal curves to data distributions when represented in a variety of graphic forms. For each trial, participants were shown a visualization of the data and were asked to move the centerpoint (mean) and width (spread or standard deviation) of a Gaussian curve overlaying the sample to create the best possible fit. The visualizations studied included bar histograms, Wilkinson dotplot histograms, strip plots, and boxplots (Figure 1).

## 2 Related Work

Beyond its use for communication, visualization is often touted as a tool primarily for *exploratory data analysis* [9] due to its investigative and data-driven nature. However, visualization can also be used for some forms of *confirmatory data analysis* [10], [11], at least as a complement. We review these topics in detail.

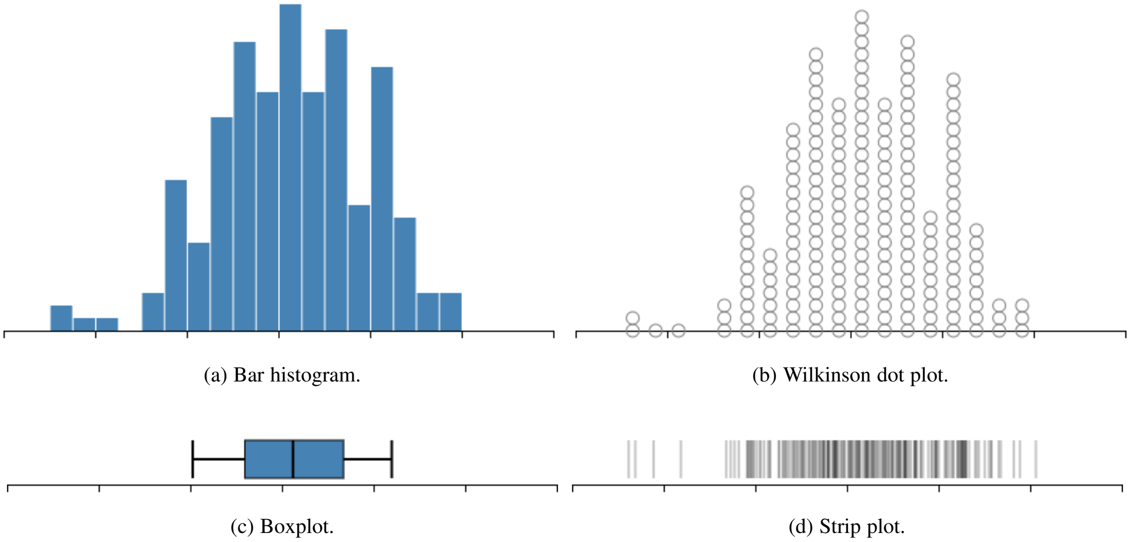
### 2.1 Graphical Inference

Creating graphical representations of data is a common and natural part of statistical workflows [12], and even central to some, such as exploratory data analysis [9]. Accordingly, making inferences from these graphical representations—i.e., *graphical inference*—is a commonplace complement to formal statistical tests. Early examples of such practice date back to Scott et al. [13] validating astronomical models by generating artificial star charts using model parameters and then asking people to compare them to real charts.

However, it is only recently that the community has begun to ask how graphical representations of data can support higher-order tasks beyond merely reading values, trends, and outliers. Buja et al. [4] proposed frameworks for *visual statistics*, where the visual representations provide the test statistic and human cognition is the statistical test. They demonstrate this approach using a "Rohrschach" test of random data, as well as a lineup of small multiples, only one of which uses real data. In followup work, Wickham et al. [14] adapted the idea to the visualization

- *Eric Newburger and Niklas Elmqvist are with University of Maryland, College Park, MD, United States. E-mail: enewburg@terpmail.umd.edu, elm@umd.edu*
- *Michael Correll is with Tableau Research, Seattle, WA, United States. E-mail: mcorrell@tableau.com*

*Manuscript received XXX XX, 2021; revised XXX XX, 2021.*



(a) Bar histogram.

(b) Wilkinson dot plot.

(c) Boxplot.

(d) Strip plot.

Fig. 1: **Experimental setup.** One-dimensional dataset visualized using the four different visual representations that we studied in our experiment. In the experiment, participants fitted a normal curve on these representations.

community, describing how these protocols can be used with common visualizations to uncover new findings while avoiding false positives. Beecham et al. [15] applied this "lineup" protocol for graphical inference to geographic clustering visualizations. Correll et al. [16] used it to investigate whether common distribution graphics were effective in displaying outliers or gaps.

Research has also been conducted on how well people can retrieve aggregate statistics from visualizations. Correll et al. [17] studied how line graphs and other visualizations can be designed to enable accurate comparisons of averages in time-series data. Albers et al. [18] generalized this idea to six aggregate tasks for eight different time-series visualizations. Furthermore, Aigner et al. [19] enriched line graphs with color to better support visual statistics, and Fuchs et al. [20] derived line glyphs to support higher-level aggregate tasks. Correll and Heer [21] studied people's ability to fit trend lines to bivariate visualizations in a crowdsourced experiment.

Deriving aggregate statistics from other visual representations is also relevant to our work. Fouriezos et al. [22] asked participants to compare the average height of two groups of bar charts, yielding high accuracy improved by the number of bars, but impaired by variance. Gleicher et al. [23] study mean value judgments in multi-class scatterplots, finding that performance is reliably high independent of the number of points and conflicting encodings. Based on results from crowdsourced experiments, Correll and Gleicher [5] propose redesigns of error bars in bar charts, showing how violin or gradient plots produce insights more aligned with statistical inference. Finally, Nguyen et al. [24] used a crowdsourced experiment to explore how different visual aggregations might impact users' perception of summary statements about sample populations. They found no impact of visual aggregation strategy on participant accuracy, but noted that participants that were shown the full data were less confident and less prone to engaging in dichotomous thinking. This suggests that dis-aggregated visualizations give a richer and more nuanced view of the underlying data.

### 2.2 Ensemble Processing

A central aspect of our research is how people visually estimate set characteristics in a visualization. A common trend in vision research is to use abstract dot clusters because these allow for precisely controlling the visual stimulus [25]. Morgan and Glennerster [26] studied perception of centroids in such clusters and found very high accuracy—with some individual differences. Ariely confirmed these results, finding that people generally are quite accurate in estimating means and ranges in a set of points even if they quickly forget specific details [27]. This suggests that the visual system extracts statistical rather than individual details of sets.

Naturally, human perception may behave differently for geometric shapes than for random dot clusters. Melcher and Kowler [28] study eye movements (saccades) during centroid estimation on shapes formed using outlines created by such dots. They found that people in general are quite proficient at establishing the centerpoint of a target shape's area, even when presented with skewed distributions of dots as well as distracting clusters. This suggests that the shape itself—as defined by its outline—guides perception rather than individual point primitives.

But how do we from many individual dots to entire shapes when our perceptual system is limited to perceiving only a handful of entities? The answer may lie in *ensemble processing* [29], which deals with how the visual system computes averages of many types of visual features to handle complex shapes and configurations. This is easier for actual objects in the real world, which tend to have regular shapes, than for artificial dot clusters. This kind of computing has also been known as *perceptual averaging.* Chong and Treisman [30] presented results from an experiment showing evidence that participants conducted size averaging even for clusters of 12 shapes. Albrecht and Scholl [31] extended these results to dynamically changing displays with similar outcomes.

Some work exists on studying these phenomena in visualization. Most relevant to our work, Szafir et al. [1] show how ensemble processing can support a wider variety of relevant tasks in statistical graphics, including *summarization* of values and *estimation* of structures. As the visual complexity rises, however, precision often decreases. Yuan et al. [8] found that estimating graphical attributes of shapes in a visualization during multi-value comparison often reduces to primitive perceptual cues, yielding lower accuracy than when perceiving single values. These perceptual cues, or *proxies* [7],

[8], provide shortcuts for when the visual system is asked to compare multiple shapes in parallel. In contrast, our work in this paper focuses on a more holistic task of fitting a mean and spread of an idealized bell curve to a single visualization of data distributions.

### 2.3 Visualizing Distributions

One of the most common visualizations for univariate distributions is the histogram, which aggregates data occurrences into discrete ranges ("bins") and visualizes the resulting counts using bars. However, histograms have flaws, most of them related to bin size and bin number [16]. Alternate representations include strip plots, density plots, violin plots, and gradient plots [32], but these lack the easy familiarity of histograms. Even disregarding binning aspects, the aggregating nature of histograms can both be a strength and a weakness: a strength, because the representation is robust in visualizing large datasets, but a weakness because the bars convey information about the relative, rather than the absolute, number of cases in each bin. Interested users must look to the axis for absolute counts—a reading task rather than a mere seeing task. For this reason, Wilkinson dot-histograms [33], where each discrete item in a bin is represented as a circle in a stack of circles, may improve on the traditional bar-based design [16], [34].

Uncertainty visualization is a closely related topic, since we often calculate uncertainty as an idealized distribution of potential variance around a point estimate, and significant progress has been made recently on designing visualization techniques where the uncertainty is intrinsic to the representation. In one example, this approach yielded a significant confidence improvement when estimating outcomes using a so-called quantile dotplot [32]. Hypothetical outcome plots (HOPs) [35] use animated draws to illustrate uncertainty, and have shown superior performance compared to violin plots and error bars.

Finally, our methodology in this paper is inspired by recent work in uncertainty visualization that uses *graphic elicitation* [36] by asking participants to draw visual representations of data as part of the evaluation. For one thing, doing this improves recall and comprehension; Kim et al. [37] found that graphically eliciting visual forms of a participant's prior knowledge and observed data helped them remember and reason about this. In follow-up work, Hullman et al. [34] asked participants to sketch their predictions of uncertainty distributions using both continuous and discrete representations prior to seeing the actual distributions. Their findings show that participant predictions are also improved by such graphic elicitation. Finally, Kim et al. [38] elicit prior and posterior beliefs for participants to derive a Bayesian cognition model for how people interpret data visualizations. Our work uses similar elicitation methods to ask participants to fit a graphical representation of a continuous curve to visualizations, but since our distribution is Gaussian, we ask only for mean and spread.

## 3 Study: Fitting Bell Curves

The goal of our study was to understand how well, or even if, untrained people would be able to fit normal curves to a data sample drawn from a normal distribution. To begin exploring this question, we conducted a crowdsourced user study on Amazon Mechanical Turk where participants were asked to control the position (mean, or "center point") and width (standard deviation, or "spread") of a Gaussian curve to fit a data sample. Samples were represented on screen using four different visualizations: a bar histogram, a dot histogram, a strip plot, and a boxplot. The variety of representations made it less likely that the characteristics of any one distribution graph type would bias the experimental results, and also allowed for the possibility of deriving design recommendations for creating effective data distribution graphics. Our design varied the size of the random sample ($n = 50$ or 200), the noise in the data (coefficients of variation = 0.2 or 0.4), and the visual representation of the data samples. To minimize the possibility of a "left-to-right bias," [39] where respondents get into a habit of always moving the curve from left to right on the screen, the means of some data samples were adjusted to move their center points to the left on the screen (values below zero). These moves did not affect the shape of the data, and coefficient of variation (CV) calculations were based upon the original positions of datasets.

An original version of this study only included Wilkinson dot histograms, and was preregistrered on OSF.[1] Since that initial study, we have expanded our experiment to include three additional visual representations: traditional bar histograms, strip plots, and boxplots. The new preregistration can be found on OSF.[2] The discussion below only concerns the new experiment; the original data for 200 participants are not included in this paper and are thus not reported. Below we review our methods, followed by our results next.

### 3.1 Participants

Because this study focused on low-level perceptual tasks that require no specific training or prior data visualization expertise, we conducted our study using Amazon Mechanical Turk (MTurk). While the use of MTurk means that we have little control over participant demographics and expertise as well as their computer hardware, prior work has shown that simple visual tasks are particularly amenable to this kind of crowdsourced study [40].

In our experiments, all experimental factors were within-participants. We planned to recruit a total of 100 participants. We limited participation to the United States due to tax and compensation restrictions imposed by our IRB. We screened participants to ensure at least a working knowledge of English; this was required to follow the instructions and task descriptions in our testing platform. Participants were prevented from participating in the experiment more than once. All participants were ethically compensated at a rate consistent with an hourly wage of at least $10/hour (the U.S. federal minimum wage in 2020 was $7.25). More specifically, the payout was $2.50 per session, and with a typical completion time of 14 minutes and 54 seconds, this yielded an hourly wage of approximately $10.00/hour.

### 3.2 Apparatus

Because of our crowdsourced setting, we were unable to control the specific computer equipment that the participants used. The study was distributed through the user's web browser. We required that all devices were personal computers (laptop or desktop) or touch tablets; smartphones were disallowed due to the limited screen space available on such devices. We also required participants to use browser windows of at least 1280 × 800 pixels.

### 3.3 Task

The tasks consisted of fitting a normal (Gaussian) curve onto a data visualization using a range slider [41] that controlled the spread (i.e., standard deviation) of the curve using the width of the interval

[1] https://osf.io/behwz
[2] https://osf.io/a9b48

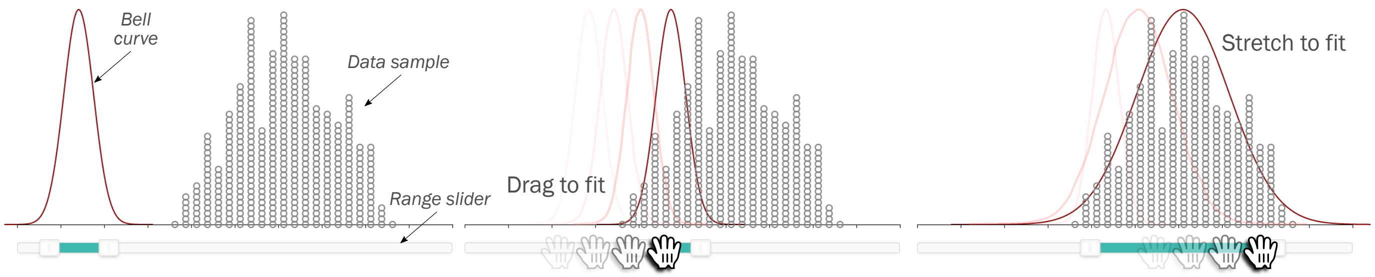
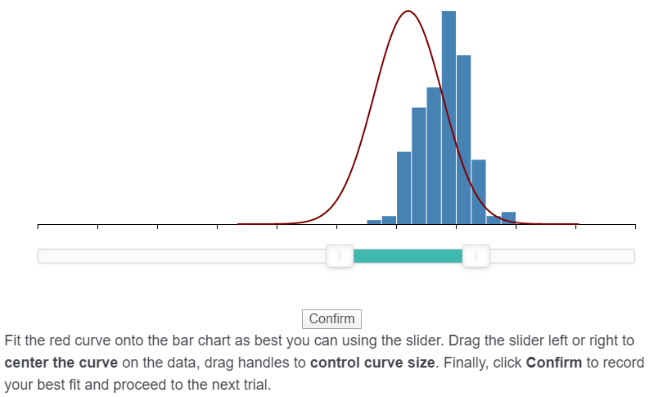


Fig. 2: **Curve fitting task.** Typical sequence in our crowdsourced curve fitting experiment. Participants controlled a normal curve using a range slider. They fit this continuous curve on top of a data sample (here represented by a Wilkinson dotplot histogram). Our evaluation varied the visualizations as well as the number of data points and the coefficient of variation for samples.



Fit the red curve onto the bar chart as best you can using the slider. Drag the slider left or right to **center the curve** on the data, drag handles to **control curve size**. Finally, click **Confirm** to record your best fit and proceed to the next trial.

Fig. 3: **Curve fitting task (bar histogram).** The red Gaussian curve is controlled by the user. The bar histogram represents the sample to fit. The range slider below the axis controls the spread using the width of the range, and the centerpoint using its position on the slider. Once the participant is satisfied with the fit, they click the “Confirm” button to finish and then proceed to the next trial.

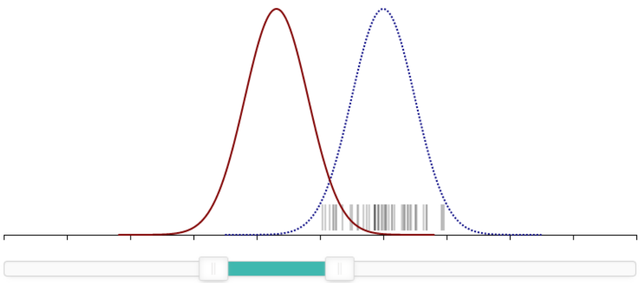


Fig. 4: **Training trial.** Participants were asked to match the curve (red line) with a correctly fitted curve (blue line). This example shows a strip plot; we included training trials for all visualizations. These trials also served as attention trials for our experiment.

and its center point (i.e., mean) by moving the position of the interval on the slider. Figure 3 shows a screenshot of a typical task. Participants were instructed to find the “best fit” between the curve and the visualized data. In a training trial for each visualization block (which we also used as attention trials; see below for details), participants were shown a perfect fit using a curve with a contrasting color, and were asked to match their own curve with the correct answer (Figure 4).

The testing platform was implemented in JavaScript using D3 [42] and embedded into a Qualtrics survey accessed using the participant’s web browser. We used `noUiSlider`[3] for the range slider implementation. Participants moved the center of the range slider both by moving one of the range endpoints, or by dragging the center area of the range to move both endpoints simultaneously. In other words, participants were able to directly control both the spread and the mean of the curve. However, participants were unable to drag the visual representation of the curve itself.

### 3.4 Dataset Generation

Data for all trials was controlled so that all participants saw the same datasets during a session. All datasets were randomly drawn from a normal distribution with varying degrees of spread, and then iteratively jittered until the standard deviation fell within 1% of the intended values for the spread $S$ (see below). Datasets were generated so that they fell within $[-3,7]$, with the horizontal axis fixed at $[-10,10]$. However, no value labels were shown for the axis to minimize number bias. The histogram used 50 bins across the horizontal axis, but actual trials typically used between 4 and 8 bins in total (based on spread).

We chose not to introduce any datasets drawn from any other distribution than the normal one, as our purpose with this experiment is to fit normal curves rather than having participants determine which distribution to use. Regardless, by varying the experimental factors (below), we are able to get distributions that are sufficiently noisy in appearance. However, it means that our datasets were all more or less symmetric.

### 3.5 Experimental Factors

We chose to model three factors in our experiment:

- **Data Size** ($D$): The number of samples in the dataset being fitted. As the number of samples increases, a histogram will begin to approach the idealized distribution. We chose two levels: 50 and 200 samples. While people may want to fit smaller sample sizes in practice, we chose 50 items as the smallest level because the Student t-distribution [43]—which we are interested in supporting using graphical inference—only begins to conform to the normal distribution for larger samples (30 items or more).
- **Spread** ($S$): The standard deviation of the Gaussian distribution being fitted. We chose two levels for this factor expressed

[3] https://refreshless.com/nouislider/

as the *coefficient of variation* (CV) (or relative standard deviation), i.e., as a ratio between the standard deviation $\sigma$ and the mean $\mu$ ($\sigma/\mu$): 20% and 40%. We settled on these values through pilot testing to ensure that this yielded both relatively noisy (for high values of $S$) as well as relatively tight (for low values of $S$) samples. Because our data generation is stochastic, datasets are only guaranteed to have specific spreads within a tolerance of 1%.

- **Visualization** ($V$): The visualization type used to represent the dataset. Based on our review of the literature, we chose four distinctive visualization techniques (Figure 1):
  - **Bar histogram:** A "classic" histogram where the aggregated number of data items for each bin is represented using a bar of uniform width (Figure 1a).
  - **Wilkinson dot histogram:** A variant of histograms initially proposed by Wilkinson [44], dot histograms are unit visualizations [45] that organize individual dots (circles) for each item into bars for each bin (Figure 1b).
  - **Boxplot:** A box-and-whisker plot as pioneered by John W. Tukey [46], where a central rectangle contains the middle half of the data (from the $25^{th}$ to the $75^{th}$ percentile), the median ($50^{th}$ percentile) is marked with a line, and the "whiskers" mark borders of wider percentiles, in this case the upper 10% and lower 10% of the data (the $10^{th}$ and $90^{th}$ percentiles) (Figure 1c). Boxplots in this study were **not** augmented with icons like dots or stars to indicate outliers.
  - **Strip plot:** A unit visualization [45] where each item is drawn as a short vertical line with opacity on the horizontal axes (i.e., with no vertical data encoding), yielding a representation similar to a barcode (Figure 1d).

While other visual representations exist, we felt this to be a representative selection on the spectrum of data aggregation: strip plots draw all values, bar and dot histograms bin them, and boxplots discard them all in favor of descriptive statistics on the sample.

Furthermore, the number of bins is an important parameter for histograms [16]. We opted not to model this factor directly, instead keeping the extents of the horizontal axis constant (at $[-10, 10]$) and the number of bins constant (50), thus indirectly controlling the number of bins used depending on spread $S$.

### 3.6 Experimental Design

We used a within-participant design, where each participant saw all data sizes, spreads, and visualizations. The relatively small total number of conditions enabled us to keep sessions shorter than 20 minutes in duration to minimize fatigue and maximize attention for crowdworkers. The order of trials was randomized for each individual participant. This yielded the following design:

| | | |
|---|---|---|
| | 2 | **Data Size** $D$ (50, 200 samples) |
| × | 2 | **Spread** $S$ (0.2, 0.4) |
| × | 4 | **Visualization** $V$ (bar, dot, box, strip) |
| × | 3 | repetitions |
| | 48 | trials per participant |

For 100 participants, we planned to collect a total of 4,800 trials.[4] For each trial, we captured the completion time as well as the accuracy. The completion time was measured from when the trial was displayed to the participant until the participant submitted an answer. The accuracy measure was based on two metrics:

- **Mean error:** If $\mu_c$ is the mean of the normal curve fitted by the participant and $\mu_s$ is the actual mean of the sample, we calculate the mean error as $\mu_c - \mu_s$. In most of our analysis, we take the absolute value of this metric.
- **Standard deviation error (%):** If $\sigma_c$ is the standard deviation of the normal curve fitted by the participant and $\sigma_s$ is the actual standard deviation of the sample, we calculate the standard deviation error as $(\sigma_c - \sigma_s)/\sigma_s$, expressed as a percentage (since this measure is independent of screen resolution, a ratio provides more information). In most of our analysis, we take the absolute value of this metric.

### 3.7 Procedure

All recruitment was conducted via Amazon Mechanical Turk. Participants that fit the eligibility criteria opened the survey in a separate browser window. At the end of their participation, they copied a unique completion code back into the Mechanical Turk interface, and were later paid as their work was checked.

Each session started with a consent form with waived signed consent. Failing to give consent terminated the experiment. Participants were instructed that they could abandon their session at any point in time. Unfortunately, we were unable to pay participants who only completed a partial session. We informed participants of this fact in the consent form when starting the session.

After consenting, participants were asked demographic questions about their age, education level, and knowledge of statistical concepts. We also asked participants to reaffirm that they were using a tablet or computer to participate. Then participants were shown practice trials for each visualization type where they were given instructions on how to read the visualization and complete the task. For each such practice trial, a correctly fitted curve was shown in a contrasting blue color. These practice trials also served as "attention trials." The purpose of these attention trials was to eliminate responses from crowdworkers who did not pay attention to the task. Any session where the participant responded with an error of more than 3 standard deviations from the actual mean for these attention trials were discarded from analysis. We included information about this fact in the consent form.

Each individual trial started with the display of the dataset and the curve (Figure 3) and ended when participants clicked the "confirm" button. Participants were unable to confirm a trial before interacting at least once with the range slider. Completion time was measured from the display of the trial, to this button-click. Participants were instructed to use the intermission between visualization blocks if they needed to rest between trials. A progress bar at the top of the screen showed the study progress.

Typical sessions lasted between 14 and 15 minutes. A few participants used much more time to complete their sessions, but our logs indicate that these participants took long breaks between trials (presumably due to interruptions). We believe that the effective time spent on the experiment was less than 15 minutes.

### 3.8 Hypotheses

We preregistered the following hypotheses about our experiment:

- *Estimation of means will be more accurate than estimation of spread.* Our intuition from our own experience as well as our literature review (e.g., that people are able to visually

[4] Note that due to a miscommunication within the research team, our preregistration called for 100 participants but we ended up recruiting 150. We discuss this later in our section on deviations from the preregistration.

determine averages with high accuracy) [23] is that people will be capable of determining the average with high accuracy, but fitting the curve over the sample will be less accurate.

- *Participants will be more accurate at estimating both mean and spread as the number of points increases.* For larger datasets, the impact of sampling error will be lower and the overall shape of the distribution more well-defined. This should make it easier to perceptually estimate the mean.
- *There will be non-uniformity in performance across visualization types.* In particular, we anticipate that:
  - *Participants will be more accurate at estimating means with boxplots.* Boxplots directly encode the median of the distribution in its visual representation, which is close to or identical to the mean in most of our stimuli.
  - *Participants will be less accurate at estimating means with strip plots.* The use of opacity and the impact of overplotting to encode density makes precise estimation difficult.
- *Participants will consistently overestimate the spread of distributions, except for in boxplots.* Based on prior work suggesting the use of "perceptual proxies" for ensemble processing and visual statistical tasks [7], we anticipate that participants will attempt to cover all of the visible marks inside the main density mass of the idealized curve. This would result in overestimation of spread for all visualization types except for boxplots, where covering the central rectangle—which only contains 50% of the data—result in underestimation.

## 4 Results

We analyzed the results from our study using bootstrapping [47] ($N = 1{,}000$ repetitions) to compute 95% confidence intervals (CIs) [48]. Our summary calculations do not adjust for interactions between factors; see Section 4.2 for our analysis of interaction effects. We report effect sizes based on these intervals.

We collected data from 146 participants who completed 48 trials for a total of 7,008 individual trials. After discarding the 19 participants who failed the four attention trials, we were left with 127 participants. Upon inspecting the data, we found that an additional 10 participants appeared to have misunderstood the curve fitting task for an entire block or more of the experiment. More specifically, these participants had moved the position of the curve to fit the mean, but had not changed the width of the curve to fit the spread. We speculate that this problem arose because our training trials failed to require respondents to adjust the spread. In retrospect, we should have forced participants to change the spread of a curve before they were able to submit a trial (we only forced them to interact with the slider in some way). Since we are unable to assess the impact of this misleading training, we opted to remove the data from all of those 10 participants from our analysis. This means we ended up with a total of 117 participants.

Based on our preregistration, we eliminated outlier trials (not participants) with a completion time higher than $3\sigma$; this removed a total of 79 trials (i.e., 1.3% of all trials). We assume these trials represent situations when the participant was interrupted mid-trial; most of these lasted for hundreds of seconds (the maximum was 698 seconds). We argue that eliminating such trials based on completion time is valid both because (a) data collection using online crowdsourcing is much less controlled than in laboratory settings, thus requiring accommodations due to participant inattention, latency, and external interruptions [40], and (b) none of our hypotheses are based on completion times.

The final dataset, after removing outliers, had 5,527 trials. The overall average completion time was 10.5 (s.d. 7.99) seconds. The overall absolute average mean error was 11.2% (s.d. 25.2%). The overall absolute average standard deviation error was 28.9% (s.d. 30.4%). Below we analyze participant performance and then go into detail on the characteristics of the different factors.

### 4.1 Averages and Individual Analysis

Figure 5 shows the data distribution for all combinations of spread, data size, and visualization type for both mean estimation error and standard deviation error. There are a few clear trends visible. There is a tendency for several visualizations to cause participants to overestimate the standard deviation. The exception is boxplots, which appear to instead yield underestimation. This overestimation is especially clear in Figure 6, which shows centered fitted curves for two specific trials chosen as representative examples of "tight" (peaked) and "loose" (flatter) distributions.

Figure 7 compares performance for individuals. Each dot represents a respondent's average error across all trials for that graph type. Outliers larger than $3\sigma$ have been removed from this data. Respondent performance in estimating standard deviations shows considerable variation. Absolute Pearson $r$ correlation values in performance between pairs of visual representations are also shown in the figure. For absolute mean error, boxplots and strip plots appear most closely correlated ($|r| = 0.31$), closely followed by dotplots and bar histograms ($|r| = 0.29$). On the other hand, performance on histograms does not appear to correlate well with performance on strip plots or boxplots ($|r| < 0.13$).

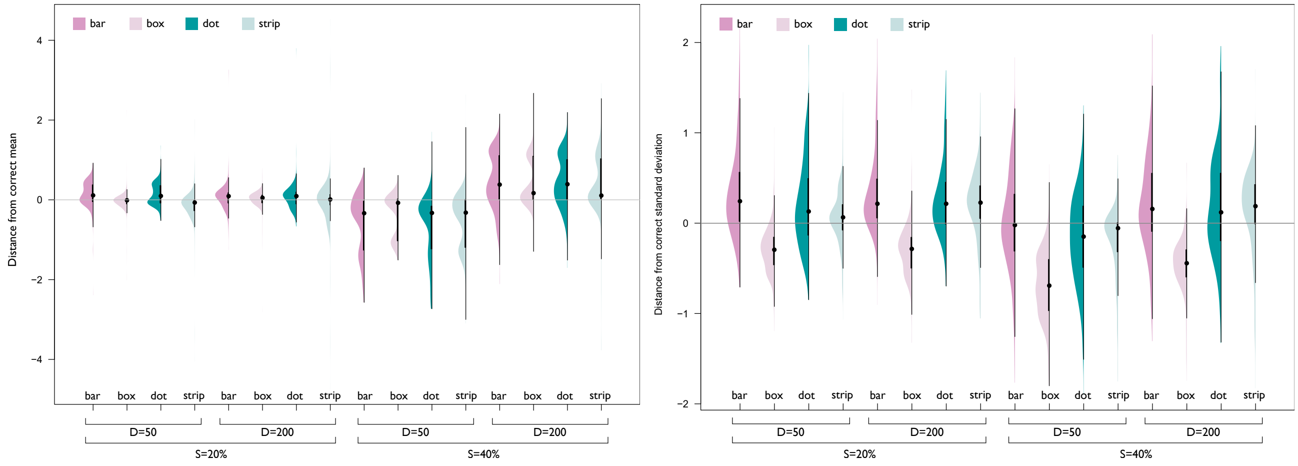
For absolute standard deviation error, absolute correlations are much higher—for example, bar histograms and dotplots are highly correlated ($|r| = 0.82$); dotplots vs. strip plots and bar histograms vs. strip plots also exhibit correlation (both $|r| = 0.53$).

TABLE 1: **Effect sizes.** Absolute mean error and absolute standard deviation error (%) for Visualization types $V$. The 95% confidence intervals were calculated using bootstrapping; see Section 4.

| **Visualization $V$** | **mean** | **95% CIs** | **s.d.** | **Cohen's $d$** |
|---|---|---|---|---|
| *Absolute mean error:* | | | | |
| – Bar histogram | 0.312 | [0.29, 0.34] | 0.480 | −0.001 |
| – Boxplot | 0.242 | [0.20, 0.28] | 0.713 | −0.140 |
| – Dotplot | 0.328 | [0.30, 0.36] | 0.581 | 0.031 |
| – Strip plot | 0.367 | [0.33, 0.41] | 0.833 | 0.110 |
| *Absolute s.d. error (%):* | | | | |
| – Bar histogram | 32.2 | [30.2, 34.4] | 39.3 | 0.148 |
| – Boxplot | 30.7 | [29.9, 31.7] | 17.1 | 0.082 |
| – Dotplot | 30.6 | [28.9, 32.5] | 34.6 | 0.078 |
| – Strip plot | 21.9 | [20.6, 23.1] | 24.6 | −0.308 |

### 4.2 Analysis by Characteristics of Factors

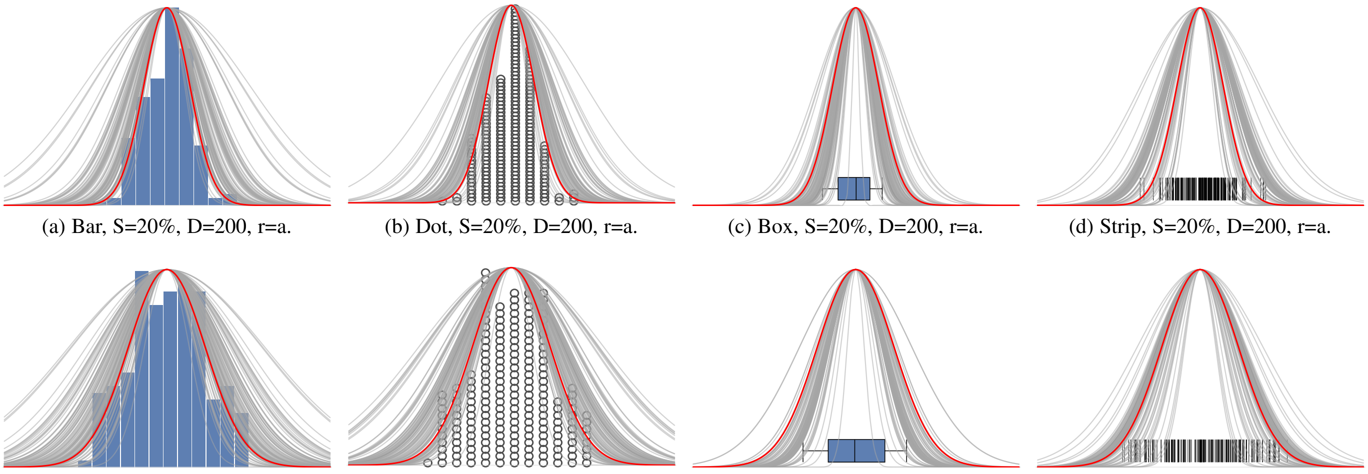
The first three rows of Figure 8 summarizes performance for all three measures based on Visualization $V$, Data Size $D$, and Spread $S$ using 95% confidence intervals (calculated using bootstrapping as discussed above). Furthermore, Table 1 summarizes the effect sizes for absolute mean error and standard deviation error (%). We used the classic Cohen's $d$ formulation, i.e., one independent of the experimental design and thus amenable to comparison with other experiments. For Visualization (the first row), error in estimating the mean was lower for boxplots (absolute mean error=0.24, Cohen's $d = -0.14$) compared to the other chart



(a) Mean estimation.

(b) Standard deviation estimation.

Fig. 5: **Detailed performance per visualization.** Distribution of mean (left) and standard deviation (right) error in all trials organized by spread $S$, data size $D$, and visualization $V$.



(a) Bar, S=20%, D=200, r=a.

(b) Dot, S=20%, D=200, r=a.

(c) Box, S=20%, D=200, r=a.

(d) Strip, S=20%, D=200, r=a.

(e) Bar, S=40%, D=200, r=a.

(f) Dot, S=40%, D=200, r=a.

(g) Box, S=40%, D=200, r=a.

(h) Strip, S=40%, D=200, r=a.

Fig. 6: **Curve fitting examples.** All fitted curves for two specific trials: $S$=20%, $D$=200, rep=a (top row) and $S$=40%, $D$=200, rep=a (bottom row). The curves have been centered around the mean value. The red curve represents the actual Gaussian of the given data sample. The respective visualizations have been added to the background for context.

types. Strip plots had slightly worse performance (absolute mean error=0.37, $d = -0.11$), but were comparable in performance to bar and dotplots. Participants were considerably more accurate in estimating standard deviation with strip plots compared to others (absolute percentage s.d. error=0.21, Cohen's $d = -0.31$). Bar charts were the least accurate for estimating standard deviation (absolute percentage s.d. error=0.32, Cohen's $d = -0.15$).

The second row in Fig. 8 summarizes measures for the data size. Although larger data size was associated with better performance, these effects were small both for mean error (0.32 for $D = 50$, 0.30 for $D = 200$, Cohen's $d = 0.02$) and standard deviation error (29.4% for $D = 50$, 28.4% for $D = 200$, Cohen's $d = 0.03$).

On the final row of Figure 8, we see the same data for spread. Here, while larger spread yields minimally higher mean error (0.307 for 20%, 0.317 for 40%, Cohen's $d = 0.0142$), it was associated with a larger relative performance gain for standard deviation (34.1% for 20%, 23.7% for 40%, Cohen's $d = 0.341$).

Finally, we present the interactions between spread and data size with visualization in Figure 9. Some interesting observations:

- *For interactions between data size and visualization*, absolute mean error appears to decrease with larger data size. The exception is boxplots, which have consistently low absolute mean error for both 50 and 200 items. As for absolute standard deviation error, the same trend recurs—the error appears to decrease with higher data sizes. Strip plots exhibit consistently lower error than the other techniques for both data sizes.
- *For interactions between spread and visualization*, the absolute mean error appears to increase with higher spread; however, this effect does not persist for boxplots, which exhibit similar absolute mean error for both levels of spread. Furthermore, for this same interaction, all visualizations yield lower standard deviation error for higher (40%) compared to lower (20%) spread, again except for boxplots; boxplots have largely unchanged standard deviation error for both conditions.

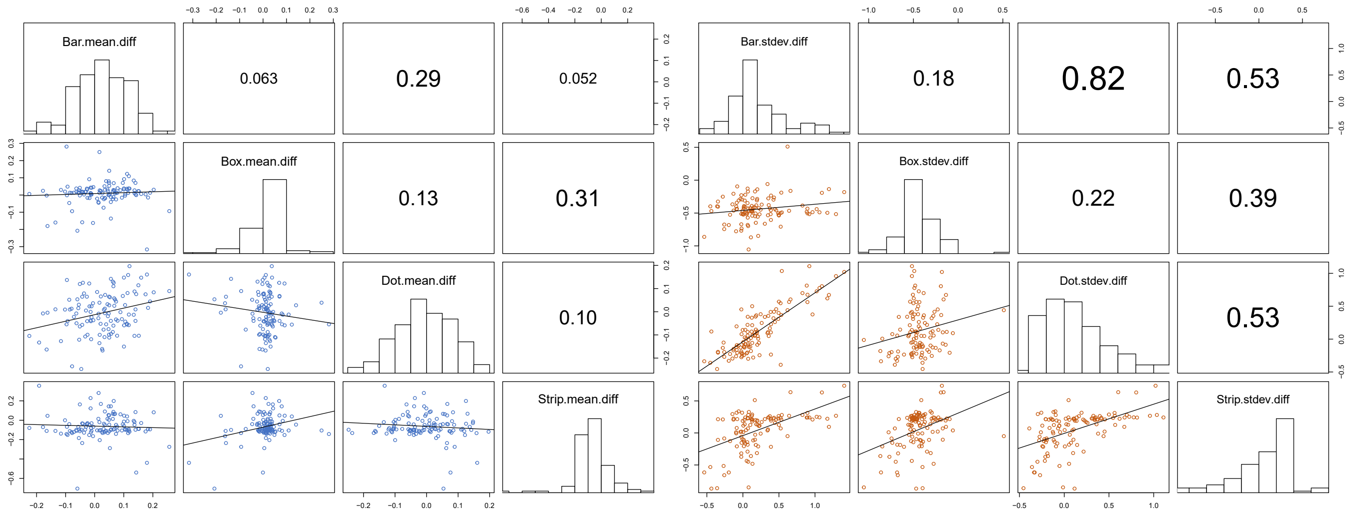


Fig. 7: **Performance comparison.** Comparing average performance across graph types of individual respondents. Each dot represents a respondent's average absolute error in finding the mean or standard deviation of data sets in all trials with the given graph type. The cells of the matrices facilitate comparisons of respondent performance on each possible graph type pairing. Respondent performance in finding means are in blue (left), standard deviations in range (right). Absolute magnitude of Pearson *r* pairwise correlation for each pair of visualization types is given above the diagonal.

### 4.3 Demographics and Participant Feedback

Figure 10 summarizes results from our demographics survey. These are more or less consistent with prior attempts at understanding the Amazon Mechanical Turk population [49]. Importantly, very few of our respondents (8) had professional statistics experience, and only about 1 in 7 had moderate statistics experience.

Free-form feedback was provided by 37 out of the 117 participants. The feedback was generally positive, with several participants noting that the task was fun and engaging. Wrote participant P21812, *"It was almost gamified, like something you'd find on a tablet or phone."* Similarly, *"This was fun! I kept wanting to fit everything in under the bell curve, though the instructions didn't state that."* (P51429) Many other participants also noted that fitting curves was challenging: *"Much of the data was not in a normal distribution!"* (P15279), *"Some of those were tough, just trying to eyeball them."* (P28558), and *"Not sure exactly how these were supposed to fit, some were too weird."* (P49775)

Finally, participant P21883 had a more general comment: *"I took high school statistics but I don't remember estimating the curves this way. I just remember doing tons of calculations."* P43443 went even deeper: *"I think I would have had better luck at proving string theory than doing this exercise with any degree of accuracy."* Fortunately, our results disagree.

### 4.4 Deviations from the Preregistration

In our preregistration, we stated that we would collect data from 100 participants. For the actual study, we recruited a total of 150 participants (and ended up with 146 participants, and eventually 117 after filtering), which was an unintentional deviation from the preregistration. The cause for this deviation was miscommunication within the research team. Though not presented here, we did a separate analysis on the first 100 respondents which indicated the same effects as the full results we have reported.

Because of the misleading training, where participants were not required to change the spread to fit the blue curve, we also removed data for 10 participants that we deemed to have misunderstood the curve fitting task. We classified such misunderstanding when at least one full block of 48 trials for a participant had a spread of 1.0 (the starting spread). We provide our unfiltered data on OSF.

Additional deviations include the following:

- Our spread values were said to be "determined by pilot testing." This is incorrect—we had determined the values based on experiences from the prior version of this experiment.
- Our preregistration included a fifth hypothesis: "There will be non-uniformity in performance across individual participants." This hypothesis is underspecified and ambiguous, and we opted to discard it in our analysis.
- Several of our plots and analyses—including our analysis of Pearson correlations and Cohen's effect sizes as well as the scatterplot matrix in Figure 7—were not included in our preregistration. We include them here in the interest of providing a richer analysis and reporting of our results.

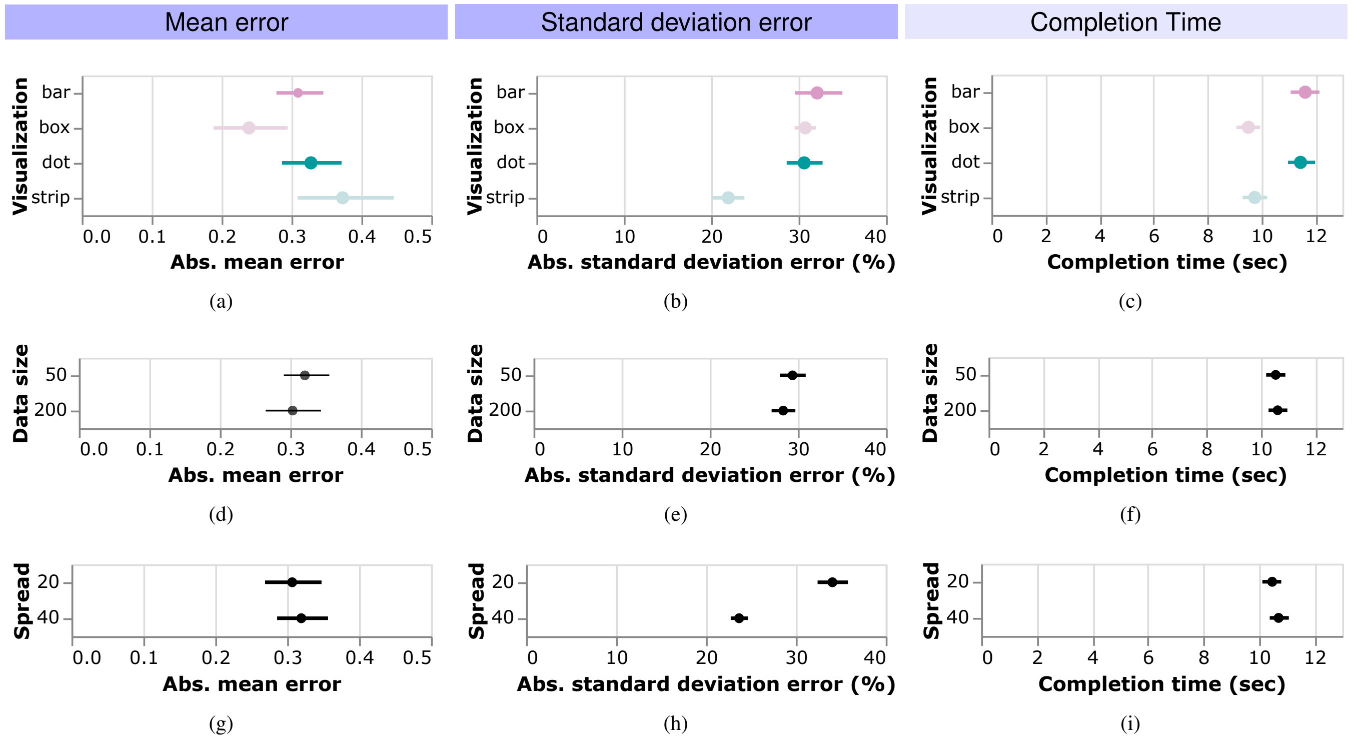
## 5 Discussion

Our study gave rise to several interesting findings while confirming our basic intuitions about this task. First we review our hypotheses in light of these results. Then we explain our results in detail and discuss how they generalize. We close by discussing the overarching research vision motivating this work.

### 5.1 Reviewing Hypotheses

We find strong evidence in our results **that the estimation of means is more accurate than estimation of spread**. Overall, the left and center columns of plots in Figure 8 indicate that people have a much easier time estimating the average values in a dataset rather than its spread, regardless of visualization technique, data size, or spread. This supports our first hypothesis.

We observed that errors in estimating both mean and standard deviation decrease as data size increases for both bar and dotplot histograms. However, for other visualizations, there appears to be



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Fig. 8: **Overall performance.** Bootstrapped per-participant 95% C.I.s showing the effect of Visualization $V$, Data Size $D$, and Spread $S$ on mean error, standard deviation error, and completion time. (Note that completion time is only included for the sake of completeness.)

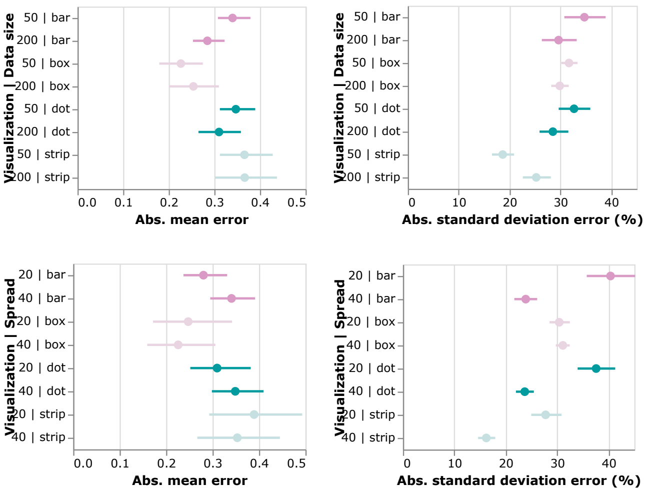


Fig. 9: **Interactions between factors.** Bootstrapped per-participant 95% C.I.s showing the effect of Visualization $V$ with Data Size $D$ and Spread $S$ on absolute mean and standard deviation error.

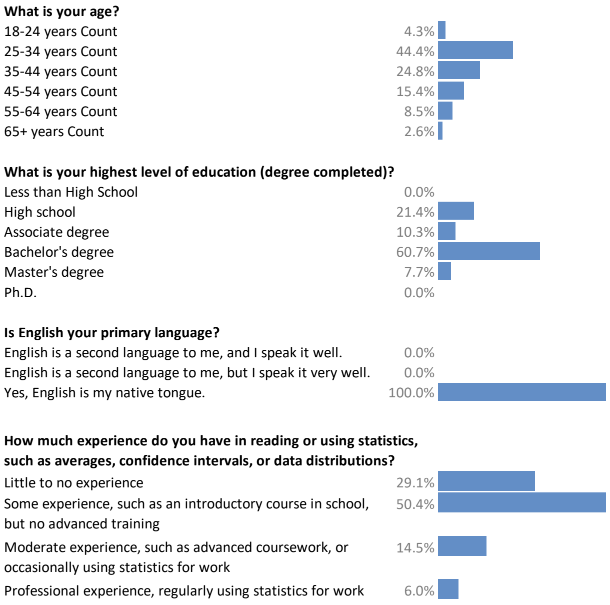


Fig. 10: **Demographics.** Summary of participant demographics.

no such effect, and for strip plots, the error actually increases as data size increases. While this could be said to partially support our second hypothesis, we think it rather suggests that this hypothesis—more data equals better estimation—is overly simplistic. For example, boxplots do not even encode data size, and strip plots seem to not scale well as the number of strips increases.

In keeping with our third hypothesis, the top row of Figure 8 illustrates some non-uniformity in performance across visualization types. As per our expectations, **boxplots yielded the lowest error and strip plots the highest for estimating the mean.**

Our fourth hypothesis dealt with the assumption that participants would seek to fit as many of the visible data points in the sample as possible under the curve. We call this an "umbrella effect," as if people are protecting the data from rain. If so, this would result in systematic overestimation of standard deviation for visualizations other than boxplots. Figures 5, 6, and 7 show that people routinely overestimated the spread for bar and dot histogram plots, but this phenomenon does not appear to extend to boxplots and strip plots. This suggests the presence of the umbrella effect for these plots, supporting the fourth hypothesis. It may also apply for the boxplot, as the underestimation may indicate participants are fitting the whiskers, which would exclude some data, or even boxplot rectangle, which only contains 50% of the data.

### 5.2 Explaining the Results

We note that the error in estimating means does not increase with higher spread $S$; basically, higher spread derived from more noise in the sample should make it harder to determine an accurate mean for the histogram. Similarly, we our results indicate the error is unchanged as the data size $D$ increases. We would have expected that an increasing number of data points would build a fuller picture of a distribution, easing the perceptual task of finding the midpoint. We see some evidence of this, as discussed at the end of Section 4.2, but it is not nearly as strong as expected.

Our results on standard deviation error are also noteworthy. First of all, the impact of data size $D$ on this error is small. This would usually be interesting as one would guess that increasing the number of items would yield better accuracy because the sample becomes more regular. However, we controlled the relative noise of the samples when we selected CV levels. On the other hand, we do see that standard deviation error decreases as the spread $S$ increases. This is counterintuitive, but may possibly be explained by the aforementioned “umbrella effect.” Figure 6 tells this story. This is further supported by our results that indicate that most trials were overestimates, i.e., with standard deviations larger than necessary. Results for boxplots were the reverse, showing a strong negative bias in errors of standard deviation estimation, a reverse umbrella. We speculate that this may reflect participants attempting to fit the curve directly onto the figure. The excellent results for estimating means with boxplots may simply reflect how on an individual basis we cannot compete with the precision offered by automated calculations, despite our innate ability to identify the mean of a distribution. In samples from normal distributions, the precalculated medians denoted by the center lines of boxplots are very close to the distribution mean. The other points making up a boxplot represent similar precomputed values. Boxplots thus present distributions with an appearance of having very little random noise.

A normal curve goes even further, being a completely noiseless and idealized representation of a data sample. Upon reflection, it seems only reasonable that participants would have greater ease applying a normal curve to the most noise-free visualization among the four we tested. As data visualization researchers, we tend to assume there is useful information in deviations of data from some idealized model; we look for meaning in the details. Yet that same information—outliers, unexpected correlations, gaps in the data—can apparently act as a distraction from ”eyeballing” traditional summary measures. It may be that the idealized forms traditional statistics focus upon sometimes make a poor match with our intuitions for messier, real world data. A normal curve is a structure we impose on smaller samples whenever we calculate a mean and standard deviation, rather than an obvious fit. This may highlight the importance of getting a sense of the data by looking at it before it is abstracted or summarized. Yet it also argues for the importance of summary calculations that can precisely identify (often critical) measures of central tendency, since noise in the data may distract our visual capacity.

### 5.3 Generalizing the Results

Our participants included a high concentration of younger adults, with a higher ratio of university degrees than in the U.S. general population. However, the level of statistics knowledge participants report was relatively basic. This low level of statistics training gives us confidence that the findings may generalize more broadly.

We believe that the sample sizes of our trial datasets (50 and 200 data items) represent typical sample sizes for people in many fields of study. The smaller size does raise an issue for the applicability of these results to future research into visual inference, since below 100 cases, the Student’s t-distribution is not totally equivalent to the normal distribution [43]. However, at $n = 50$, the two curves are similar, and we expect that the challenge-to-fit presented by the noisiness of the small sample would be the more important effect.

The visualizations we tested varied in their degree of aggregation, spanning the full range from zero aggregation (strip plots), to partial spatial aggregation (dot histograms), moderate aggregation (bar histograms), and finally a high degree with boxplots. We believe that other techniques for visualizing distributions, for example, violin plots or density plots, might be similarly scored along this dimension. By so doing, future researchers might use the results of this study to inform performance predictions for these or other graphics. However, other techniques may exhibit entirely different behavior. For instance, hypothetical outcome plots (HOPs) [35], for example, use animation to convey uncertainty, and so, while they are not aggregated, rely on spatial memory and other visual mechanisms that we did not investigate in this study.

As shown in our study of interactions (Section 4.2), aggregation is a double-edged sword. While other visualization types showed variation in performance in estimating standard deviation as the data size and spread fluctuated, boxplots had consistent performance across these properties. Yet, this *consistent* error was relatively *high*: boxplots were relatively robust in appearance regardless of the distribution, but this simplicity could hide or mislead viewers about more fine-grained properties of the distribution.

### 5.4 Limitations

Even though the results presented here represents the second incarnation of our study, our experiment has some weaknesses. For one thing, we made several deviations from our preregistration (Section 4.4), and a few of our analyses were also not explicitly detailed in the preregistration. In out future work, we will endeavor to be more precise and prescriptive even in its planning stages.

Our training trials primed participants to answer our questions by showing a prefitted idealized curve on top of a data distribution for each visualization type. However, it is conceivable that such training teaches people less about fitting curves to data than fitting curves to a specific visualization. This is an entirely fair point, and is essentially also the purpose of our experiment: to understand how well different visualizations convey the mean and standard deviation of a normal data sample. Still, we tried to avoid training people to mechanically fit curves using a prescribed pattern by only giving participants a single training trial per visualization type.

Furthermore, our study only involved normally distributed data. This is a clear limitation to our experiment, and more work is needed to chart these waters in the future. In addition, we opted not to include the number and configuration of bin sizes, which have been shown to be important factors in histogram design [16], in our experiment to keep the size of the experimental design manageable, and instead held the number of bins (50) and their size ($20.0/50 = 0.40$ per bin) constant. This is another limitation of our work, as two of our techniques—bar and dot histograms—clearly are affected by this choice, whereas the other two—strip plots and boxplots—are not. While we think that the bin sizes were more or less appropriate given the dataset properties, this is nevertheless an important factor that we hope will be studied in the future.

Finally, our work here is focused on fitting curves for use in classic parametric statistical tests. You might argue that basing this work on such classic tests is counterproductive in light of more modern tests potentially better suited to graphical inference. While we do not necessarily disagree with this sentiment, we note that (1) fitting normal distributions on discrete samples is a common task for many basic statistical operations—not the least the ability to simply characterize a distribution—and that (2) t-tests and other parametric statistical tests are still fundamental statistical tools.

### 5.5 Future Work

The overarching vision motivating our work in this paper is to lower the threshold of using real statistical methods so that anyone with little mathematical or statistical background can use them. This vision is inspired by findings from perception and visualization (e.g., [5], [18], [23]) that shows how appropriate visual representations can enable even sophisticated statistics.

Our work in this paper represents one step towards such a vision. Our findings suggest that people with minimal training can gain statistically useful information from visual displays. More importantly, our findings show that people are *consistent*; they tend to overestimate or underestimate the standard deviation depending upon the visualization used. This gives us hope that we can design visual interventions to overcome these biases as the following step, and suggest fruitful avenues of future empirical work such as exploring more complex scenarios like comparisons of multiple distributions or constructing intervals.

Looking to the future, we view the simple estimates of statistical moments in single distributions as an important and necessary stepping stone for the project of improving “visual statistics” [3] as well as charting the limits and potential pitfalls in “graphical inferences” [4]. For example, the heterogeneity of performance across visualization types in our experiment suggests that a space for potential interventions to improve an audience’s grasp of the properties of a distribution, such as “hybrid” visualizations that combine multiple univariate visualizations [16].

## 6 Conclusion

We have presented results from a preregistered crowdsourced user study on fitting normal curves onto more or less noisy data samples visualized using different representations. We find that our participants are good (approximately 12%-13% error) at accurately determining the mean of a data sample and that they can determine the standard deviation of the sample with 28%-29% error. We think that these results are encouraging for visualization designers that rely on their audiences having a good grasp of mean and spread—especially for cases where these designers rely on implicit estimates or satisficing strategies [6].

## References

[1] D. A. Szafir, S. Haroz, M. Gleicher, and S. Franconeri, “Four types of ensemble coding in data visualizations,” *Journal of Vision*, vol. 16, no. 5, pp. 11–11, 03 2016. [Online]. Available: https://doi.org/10.1167/16.5.11

[2] R. A. Rensink, *On the Prospects for a Science of Visualization*. New York, NY: Springer New York, 2014, pp. 147–175. [Online]. Available: https://doi.org/10.1007/978-1-4614-7485-2_6

[3] M. Correll, “Improving Visual Statistics,” Ph.D. dissertation, University of Wisconsin – Madison, 2015.

[4] A. Buja, D. Cook, H. Hofmann, M. Lawrence, E.-K. Lee, D. F. Swayne, and H. Wickham, “Statistical inference for exploratory data analysis and model diagnostics,” *Philosophical Transactions of the Royal Society*, vol. 367, no. 1906, pp. 4361–4383, 2009. [Online]. Available: https://doi.org/10.1098/rsta.2009.0120

[5] M. Correll and M. Gleicher, “Error bars considered harmful: Exploring alternate encodings for mean and error,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 12, pp. 2142–2151, 2014. [Online]. Available: https://doi.org/10.1109/TVCG.2014.2346298

[6] A. Kale, M. Kay, and J. Hullman, “Visual reasoning strategies for effect size judgments and decisions,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 27, no. 2, pp. 272–282, 2021. [Online]. Available: https://doi.org/10.1109/TVCG.2020.3030335

[7] B. D. Ondov, F. Yang, M. Kay, N. Elmqvist, and S. Franconeri, “Revealing perceptual proxies with adversarial examples,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 27, no. 2, pp. 1073–1083, 2021. [Online]. Available: https://doi.org/10.1109/TVCG.2020.3030429

[8] L. Yuan, S. Haroz, and S. Franconeri, “Perceptual proxies for extracting averages in data visualizations,” *Psychonomic Bulletin & Review*, vol. 26, no. 2, pp. 669–676, 2019. [Online]. Available: https://doi.org/10.3758/s13423-018-1525-7

[9] J. W. Tukey, *Exploratory Data Analysis*. Reading, MA, USA: Addison-Wesley, 1977.

[10] E. L. Lehmann and J. P. Romano, *Testing Statistical Hypotheses*, 3rd ed. New York, NY, USA: Springer, 2005.

[11] M. J. Schervish, *Theory of Statistics*. New York, NY, USA: Springer Verlag, 1995.

[12] W. S. Cleveland, *Visualizing Data*. Summit, NJ, USA: Hobart Press, 1993.

[13] E. L. Scott, C. P. D. Shane, and M. D. Swanson, “Comparison of the synthetic and actual distribution of galaxies on a photographic plate,” *Astrophysics*, vol. 119, pp. 91–112, 1954. [Online]. Available: https://doi.org/10.1086/145799

[14] H. Wickham, D. Cook, H. Hofmann, and A. Buja, “Graphical inference for infovis,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 16, no. 6, pp. 973–979, 2010. [Online]. Available: https://doi.org/10.1109/TVCG.2010.161

[15] R. Beecham, J. Dykes, W. Meulemans, A. Slingsby, C. Turkay, and J. Wood, “Map lineups: Effects of spatial structure on graphical inference,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, no. 1, pp. 391–400, 2017. [Online]. Available: https://doi.org/10.1109/TVCG.2016.2598862

[16] M. Correll, M. Li, G. L. Kindlmann, and C. Scheidegger, “Looks good to me: Visualizations as sanity checks,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 25, no. 1, pp. 830–839, 2019. [Online]. Available: https://doi.org/10.1109/TVCG.2018.2864907

[17] M. Correll, D. Albers, S. Franconeri, and M. Gleicher, “Comparing averages in time series data,” in *Proceedings of the ACM Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, 2012, pp. 1095–1104. [Online]. Available: https://doi.org/10.1145/2207676.2208556

[18] D. Albers, M. Correll, and M. Gleicher, “Task-driven evaluation of aggregation in time series visualization,” in *Proceedings of the ACM Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, 2014, pp. 551–560. [Online]. Available: https://doi.org/10.1145/2556288.2557200

[19] W. Aigner, A. Rind, and S. Hoffmann, “Comparative evaluation of an interactive time-series visualization that combines quantitative data with qualitative abstractions,” *Computer Graphics Forum*, vol. 31, no. 3, pp. 995–1004, 2012. [Online]. Available: https://doi.org/10.1111/j.1467-8659.2012.03092.x

[20] J. Fuchs, F. Fischer, F. Mansmann, E. Bertini, and P. Isenberg, “Evaluation of alternative glyph designs for time series data in a small multiple setting,” in *Proceedings of the ACM Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, 2013, pp. 3237–3246. [Online]. Available: https://doi.org/10.1145/2470654.2466443

[21] M. Correll and J. Heer, “Regression by eye: Estimating trends in bivariate visualizations,” in *Proceedings of the ACM Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, 2017, pp. 1387–1396. [Online]. Available: https://doi.org/10.1145/3025453.3025922

[22] G. Fouriezos, S. Rubenfeld, and G. Capstick, “Visual statistical decisions,” *Perception & Psychophysics*, vol. 70, no. 3, pp. 456–464, 2008. [Online]. Available: https://doi.org/10.3758/PP.70.3.456

[23] M. Gleicher, M. Correll, C. Nothelfer, and S. Franconeri, “Perception of average value in multiclass scatterplots,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2316–2325, 2013. [Online]. Available: https://doi.org/10.1109/TVCG.2013.183

[24] F. Nguyen, X. Qiao, J. Heer, and J. Hullman, "Exploring the effects of aggregation choices on untrained visualization users' generalizations from data," *Computer Graphics Forum*, vol. 39, no. 6, pp. 33–48, 2020. [Online]. Available: https://doi.org/10.1111/cgf.13902

[25] D. Whitaker and H. Walker, "Centroid evaluation in the vernier alignment of random dot clusters," *Vision Research*, vol. 28, no. 7, pp. 777–784, 1988. [Online]. Available: https://doi.org/10.1016/0042-6989(88)90024-7

[26] M. J. Morgan and A. Glennerster, "Efficiency of locating centres of dot-clusters by human observers," *Vision Research*, vol. 31, no. 12, pp. 2075–2083, 1991. [Online]. Available: https://doi.org/10.1016/0042-6989(91)90165-2

[27] D. Ariely, "Seeing sets: representation by statistical properties," *Psychological Science*, vol. 12, no. 2, pp. 157–162, 2001. [Online]. Available: https://doi.org/10.1111/1467-9280.00327

[28] D. Melcher and E. Kowler, "Shapes, surfaces and saccades," *Vision Research*, vol. 39, no. 17, pp. 2929–2946, 1999. [Online]. Available: https://doi.org/10.1016/S0042-6989(99)00029-2

[29] G. A. Alvarez, "Representing multiple objects as an ensemble enhances visual cognition," *Trends in Cognitive Sciences*, vol. 15, no. 3, pp. 122–131, 2011. [Online]. Available: https://doi.org/10.1016/j.tics.2011.01.003

[30] S. C. Chong and A. Treisman, "Representation of statistical properties," *Vision Research*, vol. 43, no. 4, pp. 393–404, 2003. [Online]. Available: https://doi.org/10.1016/s0042-6989(02)00596-5

[31] A. R. Albrecht and B. J. Scholl, "Perceptually averaging in a continuous visual world: extracting statistical summary representations over time," *Psychological Science*, vol. 21, no. 4, pp. 560–567, 201. [Online]. Available: https://doi.org/10.1177/0956797610363543

[32] M. Kay, T. Kola, J. R. Hullman, and S. A. Munson, "When (ish) is my bus?: User-centered visualizations of uncertainty in everyday, mobile predictive systems," in *Proceedings of the ACM Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, 2016, pp. 5092–5103. [Online]. Available: https://doi.org/10.1145/2858036.2858558

[33] K.-W. Moon, "Wilkinson dot plot," in *Learn ggplot2 Using Shiny App*. New York, NY, USA: Springer, Cham, 2016, pp. 103–109. [Online]. Available: https://doi.org/10.1007/978-3-319-53019-2_12

[34] J. Hullman, M. Kay, Y. Kim, and S. Shrestha, "Imagining replications: Graphical prediction & discrete visualizations improve recall & estimation of effect uncertainty," *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 1, pp. 446–456, 2018. [Online]. Available: https://doi.org/10.1109/TVCG.2017.2743898

[35] J. Hullman, P. Resnick, and E. Adar, "Hypothetical outcome plots outperform error bars and violin plots for inferences about reliability of variable ordering," *PLOS ONE*, vol. 10, no. 11, pp. 1–25, 11 2015. [Online]. Available: https://doi.org/10.1371/journal.pone.0142444

[36] N. Crilly, A. F. Blackwell, and P. J. Clarkson, "Graphic elicitation: using research diagrams as interview stimuli," *Qualitative Research*, vol. 6, no. 3, pp. 341–366, 2006. [Online]. Available: https://doi.org/10.1177/1468794106065007

[37] Y. Kim, K. Reinecke, and J. Hullman, "Explaining the gap: Visualizing one's predictions improves recall and comprehension of data," in *Proceedings of the ACM Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, 2017, pp. 1375–1386. [Online]. Available: https://doi.org/10.1145/3025453.3025592

[38] Y. Kim, L. A. Walls, P. M. Krafft, and J. Hullman, "A bayesian cognition approach to improve data visualization," in *Proceedings of the ACM Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, 2019, pp. 682:1–682:14. [Online]. Available: https://doi.org/10.1145/3290605.3300912

[39] T. M. Spalek and S. Hammad, "The left-to-right bias in inhibition of return is due to the direction of reading," *Psychological Science*, vol. 16, no. 1, pp. 15–18, 2005. [Online]. Available: https://doi.org/10.1111/j.0956-7976.2005.00774.x

[40] J. Heer and M. Bostock, "Crowdsourcing graphical perception: using Mechanical Turk to assess visualization design," in *Proceedings of the ACM Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, 2010, pp. 203–212. [Online]. Available: https://doi.org/10.1145/1753326.1753357

[41] C. Ahlberg, C. Williamson, and B. Shneiderman, "Dynamic queries for information exploration: An implementation and evaluation," in *Proceedings of the ACM Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, 1992, pp. 619–626. [Online]. Available: https://doi.org/10.1145/142750.143054

[42] M. Bostock, V. Ogievetsky, and J. Heer, "D$^3$: Data-driven documents," *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 12, pp. 2301–2309, Dec. 2011. [Online]. Available: https://doi.org/10.1109/TVCG.2011.185

[43] Student, "The probable error of a mean," *Biometrika*, vol. 6, no. 1, pp. 1—25, 1908. [Online]. Available: https://doi.org/10.1093/biomet/6.1.1

[44] L. Wilkinson, "Dot plots," *The American Statistician*, vol. 53, no. 3, pp. 276–281, 1999. [Online]. Available: https://doi.org/10.1080/00031305.1999.10474474

[45] D. Park, S. M. Drucker, R. Fernandez, and N. Elmqvist, "Atom: A grammar for unit visualizations," *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 12, pp. 3032–3043, 2018. [Online]. Available: https://doi.org/10.1109/TVCG.2017.2785807

[46] H. Wickham and L. Stryjewski, "40 years of boxplots," 2011, unpublished. [Online]. Available: http://vita.had.co.nz/papers/boxplots.pdf

[47] B. Efron, "Bootstrap methods: Another look at the jackknife," in *Breakthroughs in Statistics*. New York, NY, USA: Springer Verlag, 1992, pp. 569–593. [Online]. Available: https://doi.org/10.1007/978-1-4612-4380-9_41

[48] P. Dragicevic, "Fair statistical communication in HCI," in *Modern Statistical Methods for HCI*, J. Robertson and M. Kaptein, Eds. New York, NY, USA: Springer Verlag, 2016, pp. 291–330. [Online]. Available: https://doi.org/10.1007/978-3-319-26633-6_13

[49] J. Ross, L. Irani, M. S. Silberman, A. Zaldivar, and B. Tomlinson, "Who are the crowdworkers?: shifting demographics in Mechanical Turk," in *Extended Abstracts of the ACM Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, 2010, pp. 2863–2872. [Online]. Available: https://doi.org/10.1145/1753846.1753873

**Eric Newburger** received the masters degree in Applied Economics in 1995 from the University of Wisconsin – Madison, in Madison, WI, USA. He is a Ph.D. candidate and adjunct lecturer in the College of Information Studies, University of Maryland, College Park in College Park, Maryland, USA. He is also a student member of the Human-Computer Interaction Laboratory (HCIL) at UMD.

**Michael Correll** received the Ph.D. degree in 2015 from the University of Wisconsin – Madison in Madison, WI, USA. He is a Senior Research Scientist at Tableau Research as part of Tableau Software in Seattle, WA, USA. His research interests include data ethics, communicating statistics to mass audiences, and investigating biased or misleading data visualizations.

**Niklas Elmqvist** received the Ph.D. degree in 2006 from Chalmers University of Technology in Göteborg, Sweden. He is a professor in the College of Information Studies, University of Maryland, College Park in College Park, MD, USA. He is also a member of the Institute for Advanced Computer Studies (UMIACS) and formerly the director of the Human-Computer Interaction Laboratory (HCIL) at UMD. He is a senior member of the IEEE and the IEEE Computer Society.