ABSTRACT

Title of Dissertation:	FROM EXPLORATORY TO CONFIRMATORY: TOWARDS DATA VISUALIZATION AS A COMPLETE ANALYSIS TOOL
	Eric Charles Newburger, Doctor of Philosophy, 2023
Dissertation directed by:	Niklas Elmqvist Professor, College of Information Studies (iSchool) Affiliate Professor, Department of Computer Science University of Maryland, College Park, MD, USA

Confirmatory statistics tests, performed and written with equations, are a standard in scientific publications, but may represent a barrier to entry for novice analysts who have less familiarity with purely calculative methods. Data visualization, often touted as useful for sharing completed analyses with lay audiences, is often used for early-stage exploratory analysis. Could visualization support hypothesis confirmation? Do people have the visual intuitions to make use of such a tool? What would a visual statistical test look like, and what features would it require for acceptance by the scientific community?

This research begins with a crowd-sourced experiment which asked respondents to fit a normal curve to a series of data samples, displayed as bar histograms, dot histograms, box plots, or strip plots. The results suggest people have

visual intuitions – though biased toward overestimating spread – for linking idealized probability distributions with real sample data. People performed differently depending upon graphic form, suggesting design choices for subsequent experiments.

A second experiment tested whether novice users might be able to perform a statistical test (T-Test) using a visual analogue – two overlapping distributions (shown as overlapping normal curves, box plots, strip plots, bar histograms, or dot histograms). Respondents had some capacity for this task, performing best with normal curves than with more detailed graphics like histograms.

The final investigation of this research paired the design lessons garnered during experiments 1 & 2 with an interview study of experienced statisticians to explore the design requirements for creating acceptable visual tools for inferential statistics. The interviews uncovered three design foci: that the tool must display multiple, **contrasting facets of analysis**; the tool should **connect the test back to raw data;** and include **a visual representation of real effect sizes compared to the p-value** of the test statistic.

The final chapter of this dissertation uses the design principles determined by these three investigations to propose a prototype visual tool for conducting a two-sample t-test, along with suggested variations for other inferential statistics.

FROM EXPLORATORY TO CONFIRMATORY: TOWARDS DATA VISUALIZATION AS A COMPLETE ANALYSIS TOOL

by

Eric Charles Newburger

Dissertation proposal submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2023

Advisory Committee:
Professor Niklas Elmqvist, Chair
Professor David Weintrop
Professor Tammy Clegg
Professor Ming C Lin, Dean's representative
Michael Correll, PhD.

Dedication

To

Caroline Day Walasik

&

Vivienne Aubrey Rose Newburger.

Acknowledgements

I would like to express my gratitude to the members of my degree committee: Dr. Michael Correll, Professor Tamara Clegg, Professor Ming Lin, and Professor David Weintrop, and the committee chair, Professor Niklas Elmqvist. I am indebted to each member for their advice and guidance, most especially Professor Elmqvist, whose expert mentorship led me through this process during these past five years. Professor Elmqvist also gave me the opportunity to teach, thereby showing me what I wanted to do for the rest of my life.

I would also like to thank Venkata Sai Pramod Chundury for his expert programming help through projects in this research, and Professor Dan Levy for his advice on Study 3.

I am grateful to Professors Mega Subramaniam and Joel Chan, who introduced me to the field of Information Science during my first year, both its expansive intellectual range and where my interests might fit within it. I am grateful as well to Professor Ben Shneiderman for first introducing me to the HCIL at UMD, and supporting me through my application process.

I thank Beth Newburger Schwartz and Richard Schwartz for their encouragement to pursue advanced studies, and their unflagging support when I did.

Finally, I would like to express my deep appreciation, gratitude, and thanks, to Jennifer Cheeseman Newburger, whose loving understanding, patience, and occasional professional advice, have been invaluable to me for quite a lot longer than the past five years. I would never have made it into this program to start with, much less successfully completed it, without her.

Table of Contents

Chapter 1: Introduction

1.1 Motivation for this Research

Data visualization provides a methodology for conducting quantitative analyses while eschewing mathematical notation, and, in so doing, could provide a potential inroad to data analysis for users who lack perfect fluency in equational math. During the current era of data ubiquity, a visual, "statistics for all," toolkit – aimed at the laymen who finds themselves with potentially informative data, but who lacks long training or experience as an analyst – might be especially powerful.

Yet in fields which rely heavily upon quantitative analysis, traditional, purely calculative, equation-based methods (such as null hypothesis statistical testing, or "NHST"), remain the standard for achieving publication, despite some ongoing criticism (Cumming 2014, Sullivan 2012, Belia 2005). Data visualization, if considered at all, is seen as a method only for early, exploratory analysis (Tukey 1977), or as a tool for displaying findings derived through equation-based methods. Thus, entry into these quant-heavy fields requires at least some knowledge of traditional statistics. Indeed, people may still favor traditional tools even where they have found a workable substitute. Bartram et. al. found that people actively working in analytic occupations, but using only visual interactive models of iterative calculation (spreadsheets) as their analytic tool, shared a general sense that they weren't doing 'real' analysis (Bartram 2021). This hints at a cultural bias toward equation-based methods as legitimate forms of quantitative analysis, and other methods as lessor.

At the same time, statistics represents a hurdle for most people; novices often struggle with choosing methods, understanding assumptions, interpreting results, and implementing tests, even when in analyzing and understanding data they deal with in their everyday lives (Mustafa 1996). Students who succeed at an introductory statistics course often achieve only the minimum competence required to pass, and many fail to retain even that modest understanding by the time they reach the next course in the series (author's personal experience as an instructor). This may contribute to the high rates of attrition among college students in science, technology, engineering, and mathematics (STEM) majors compared to non-STEM fields (Chen 2018).

This pattern persists after graduation; only about one quarter of adults who graduated with a bachelor's degree from a STEM major are putting their lessons to use in STEM occupations (Census data - https://www.census.gov/library/visualizations/interactive/from-college-to-jobs-stem.html). And even among professionals publishing papers founded upon quantitative analyses, there is widespread misunderstanding of the assumptions made by the very methods they used (Belia 2005, Hoekstra 2014).

Quantitative data analysis, like any field of endeavor, rests upon intellectual foundations which the practitioner should have in place prior to beginning work (Machlup 1983, Bowker 2000, Star 1989, Van House 2004, Bates 1999, Middendorf 2004, Nelson 1982, Borgman 2015). A student wishing to enter a given field must, at some point, integrate these foundations into their thinking, or face great difficulties in understanding the workings of the more technical aspects of their pursuit (English

2016, Bowker 2000, Middendorf 2004, Nelson 1982). For example, in at least one empirical study of alternative curricula, Kuo (2019) found that physics students instructed in, "mathematical sensemaking", that is, focusing explicitly on connecting standard physics equations to the conceptual models of reality they represent, performed better during end-of-semester assessments than similar students provided with more traditional exercise-focused instruction.

A visual statistics-for-all system, that is, a visual analytics system which poses lower barriers to entry to lay-analysts, would need to:

A) Provide affordances that replace the need for total training in all the assumptions of quantitative analysis;

B) Present users with legitimate analytic techniques that cover the full range of a quantitative research process, from hypothesis formation to testing.

This research will attempt to demonstrate whether it is possible to use visualization to provide such affordances in quantitative analyses, with particular regard for confirmatory visual statistics, which, heretofore, have had much less support in the literature than visual exploratory statistics.

1.2 Definitions

1.2.1 Gathering 'data' as instinctive process

I use a definition of "Data" offered by Borgman (2015), "…data are representations of observations, objects, or other entities used as evidence of phenomena for the purposes of research or scholarship." They go on to add that, "Entities become data only when someone uses them as evidence of a phenomenon,

and the same entities can be evidence of multiple phenomena." I use the word "Analyses" to denote the processes by which scholars and researchers attempt to use entities as evidence to further understanding of their study subjects. Thus, "Data analysis" in the following refers to a collection of highly varied processes which transform observations, objects, or entities into evidence. "Quantitative data analysis" further specifies that those transformations entail quantitative abstractions, that is, abstractions which go beyond identity, to include ordination and (usually) measurement.

The measurement aspect of quantitative analysis, paired with the above definition of data as a component of evidence, implies a specific epistemological foundation. A person seeking to understand the world better through quantitative data analysis first assumes that the world is subject to study: it exists with or without us and will exhibit patterns which remain consistent before and after we recognize them. This conception of the world as a place where tests can be replicated gives data, as defined here, traction (Gattei 2004); data become the raw materials for reasoning about the world.

This definition of data also reflects a particular kind of physical environment – one in which patterns of events tied to specific physical conditions reliably repeat when those conditions themselves occur again. Cosmides suggested that this kind of environment would lead a species already specializing in increasing cognition to evolve a statistical sense, that is, a sense of the world which factors in the probability of future events (Cosmides & Tooby 1996). This evolutionary explanation for our statistical sensibilities makes the explicit claim: that at the base of quantitative

analyses are instinctive intuitions about how the world works, intuitions that formal training in statistics can sharpen (Nisbett et al 1983), but that are always present in at least their nascent forms.

1.2.2 Quantitative Analysis as a tool for making us smarter

Cosmides and Tooby explored whether the instinctive intuitions that undergird quantitative analysis can be at least partially engaged through visual data displays. In their research, the choice of form in visual data presentation had a strong effect on how their subjects performed analytic tasks upon those data, as well as their success rates (Cosmides & Tooby 1996). This suggests that some visualizations may take better advantage of our statistical intuitions than others, but the larger point here is that visual data displays trigger these intuitions at all. A data display can apparently serve as a tool for thinking, triggering mental machinery evolved for dealing with physical problems in the world.

This is "External Cognition," wherein entities external to a single human brain form part of a larger cognitive system, essentially an extended mind (Scaife & Rogers 1996, Hutchins 1995, Clark 2008, Norman 2013). The external entities can be anything which:

1. Performs some cognitive function, such as memory or calculation;
2. Provides an interface for the easy exchange of resultant information among the components of the extended cognitive system, and, ultimately, with the human brain that makes decisions for the extended mind.

Modern tools of formal statistics are replete with such external entities, from hand calculators to computers. However, even statistical tools predating widely

available electronic computing engines constitute entities of external cognition, for example, tables of known probability distributions, logarithmic tables for simplifying hand calculations, or statistical notation to ease paper and pencil calculations. All these examples have information embedded within them; it may free users from having to remember (or recompute afresh) commonly needed values, or might free users from having to perfectly recall every step of the statistical methods they aid. Tools also come with embedded meta-data which guides analysts through using them, for example, the regular location of footnotes, or the tabular format common to so many of these resources.

External entities thus take on memory and calculative tasks. They pass on know-how. They allow us to perform long and difficult mental work more rapidly and more accurately.

To paraphrase Don Norman, objects can make us smarter.

1.2.3 Data Visualization as an External Analogue of Internal Cognition

Scaife and Rogers theorize about how visualizations might operate, proposing that its function as part of an extended cognition system is central (Scaife & Rogers 1996).

They begin by describing that our brains make internal models of the external world, and assert it is these models we experience rather than direct, whole, unprocessed reality (Scaife 1996, Hollan 2000, and also Chong et. al. 2003). We work with necessarily limited information – for example, we can only look in one direction at any given moment – to build a picture of our surroundings. We map objects, classify them based on our prior experience with similar objects, assign

agency to those objects likely to be thinking entities, and perform many other cognitive functions (Ware 2019, Ware 2010, Kosslyn 2006, Few 2009). We do all these simultaneously, in real time, and so automatically that usually we can remain incognizant of the degree to which the world we think we perceive is actually a construct of our internal mental processes (Scaife 1996).

For example, we think we see the world in focus – sharp edges, clear distinctions from one object to the next – rather than blurred together. Yet, the physics of our eyes are such that most of our field of view is blurry, everything but an in-focus central spot about the size of a silver dollar held at arm's length (Ware 2019, Cairo 2012). We use this spot – our eyes darting about constantly – to take samples, and from these in-focus pictures our brains stitch together composites of what the world would look like were we able to see it all clearly. Thus, the in-focus world we ultimately perceive is an internal model of external reality.

We build models using every sense, pairing them with our memories, prior knowledge, and expectations (Cosmides 1996, Few 2009, and implied by Van House 2004 with their discussion of the intuitive nature of Bayesian priors). We make errors, too: we suffer from change blindness, misperceptions in depth, see patterns where none exist, and miss other patterns actually present in the data (Shermer 2011, Cairo 2012, Ware 2010, Kosslyn 2006, Few 2009). Acknowledging the degree to which our perceived reality is a limited, mind-forged model – that is, a useful simplification based upon our interpreting always-partial sense information – is prerequisite to any kind of analysis (Cosmides 1996, Windschitl 2008). Recognizing that we work from a model allows us to improve that model, for example, by seeking

to add new, more informative data, or better interpreting data on hand (Borgman 2015, Wilkerson 2017).

Based upon their reading of the cognitive literature, Scaife and Rogers propose that we conduct formal analysis by selectively abstracting our perceptual models, until we can organize key data points into a cohesive whole which presents a simplified, yet still useful version of reality (Scaife & Rogers 1996). They perceive data visualizations as an attempt to externalize these internal, analytic models. By externalizing analytic models, we can manipulate them, interact with them, increase their complexity by adding data via computer memory; in other words, we can use them for external cognition.

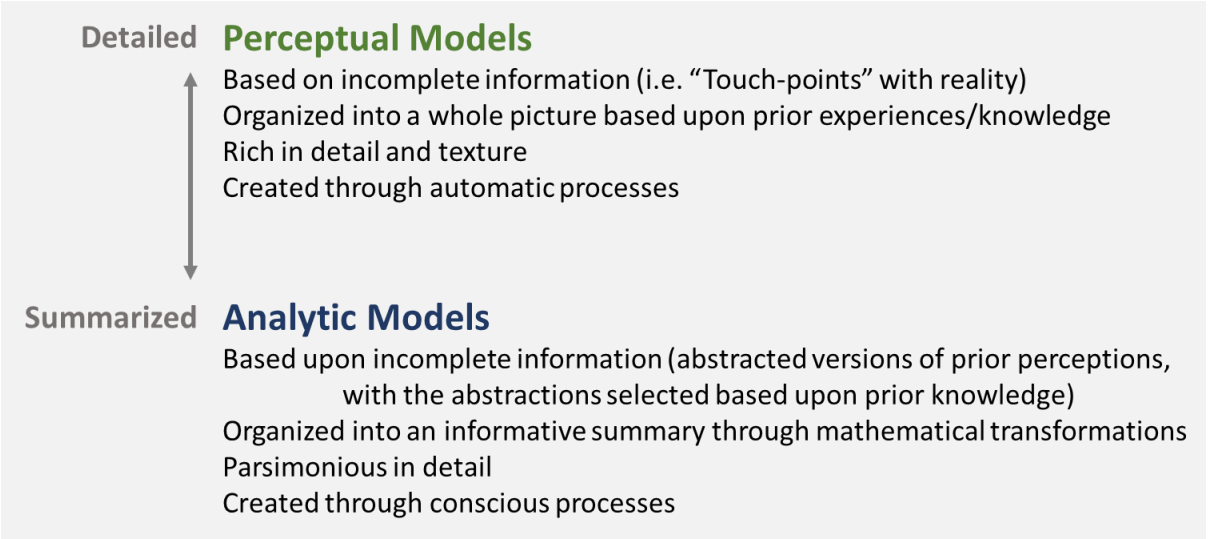
Equation-based statistical models can perform similar functions – simplifying reality to an externalizable model to join with external tools, such as computers, to add data that outstrips our memory and native calculative power. The link to our internal models is perhaps less obvious, since data visualizations employ visual metaphors which suggest physical objects in a virtual space, while equation-based models present their results in more abstract forms (Scaife 1996, Few 2009).

It appears that internal perceptual models, and formal analytic models derive from processes that are analogous to one another (Nisbett 1983). In each, our cognitive faculties piece together a best-guess cohesive whole from partial data. Both transform data into processed information to help us make decisions (Cosmides 1996). Both detailed, perceptual models and analytic models rely upon prior knowledge for their formation. Yet perceptual models are automatic, while analytic models require conscious effort.

1.3 Four Kinds of Analytic Models

Models created through deliberate, conscious processes, as compared to those built automatically, may be analogous, but the resulting models have different characteristics. We can classify models along two axes already discussed, Detail (detailed vs. summarized) and Location (Internal to our minds vs. external).

Fig. 1.1: The Detail Axis

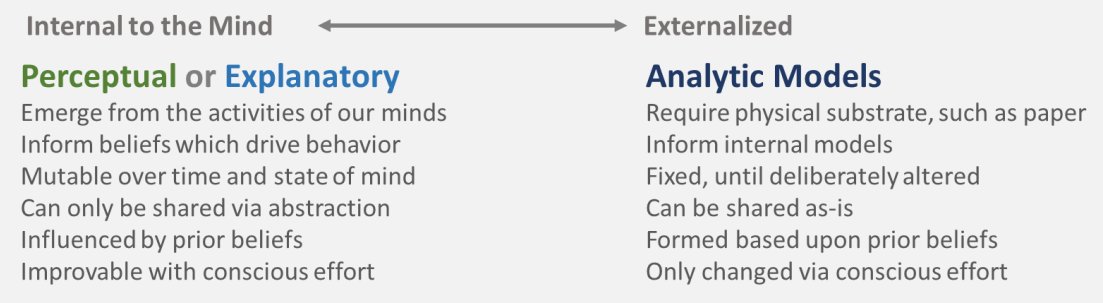


Analytic models are necessarily far simpler (figure 1.1). We read and study analytic models, but we experience perceptual models, which requires a richness of detail analytic models can neither match nor profit from. Indeed, analytic models make a virtue of their simplicity. When an analytic model makes accurate predictions about the real world based upon only a few assumptions and minimal data, we presume the elements of the model have expressed some simple, hidden truth which lays beneath the apparent complexity we see around us; this intuition motivates scientific reductionism, and it makes analysis a powerful tool for gaining understanding (Sawyer 2002 lays out advantages and limits of reductionism in

contrasting it to "emergentism," which holds some phenomena can only be studied holistically).

Much of the model-creation behind both scholarship and scientific advancement during this information age takes place in the external world via tools of External Cognition (Hutchins 1995, Webster 2014, Chang 2009, Fekete 2008, Duguid 2015, Jones 2002, Scaife 1996, Hollan 2000, Ainsworth 2011). This defines the Location axis (Fig. 1.2). Models may be Internal, that is wholly contained within our person, and External, that is, contained principally on some combination of external cognitive tools. This is my interpretation of distinctions principally drawn by Scaife and Rogers in their "External Cognition", along with Hutchins in his Cognition in the Wild. I expand on my interpretation in the remainder of this section and the next, and propose a unified schema for considering model type, function, and their inter-relationships.

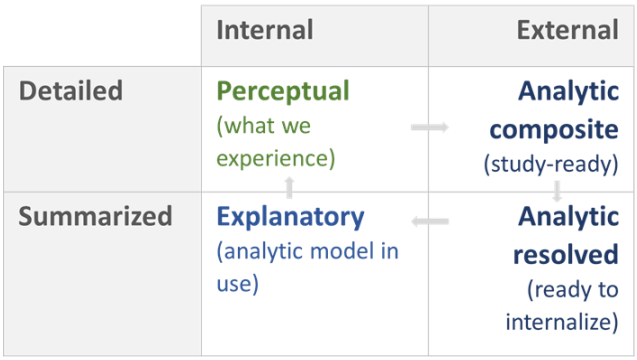
Fig. 1.2: The Location Axis



Internal models may be perceptual, dealing with what we experience, or they may be explanatory, that is, an internal model of the causes of what we experience. For the purposes of this research all external models are analytic, being used to inform our internal, explanatory models.

Data Visualization as a form of external cognition well illustrates the distinctions between different kinds of models, and how one kind of model informs the next in a cycle of generation (Ainsworth 2011, Fekete 2008, Scaife 1996, Friendly 2005) (Figure 1.3).

Fig. 1.3: Four kinds of models, and the cycle that connects them



Data visualizations, as well as traditional equation-based statistical methods, seek to create new explanatory models by pairing externalized versions of perceptual models (coded as data) with externalized versions of existing, insufficient explanatory models (reformed as hypotheses). This results in a composite model suitable for analytic study, that being typically accomplished through various mathematical transformations (Friendly 2005, Wilkerson 2017; Moore 1998 also describes this while cautioning against mistaking data analytic processes for pure math). Once the analytic model resolves into a final form that provides a satisfactory explanation for the phenomena at hand, the user can re-internalize the results.

The process of externalizing and then re-internalizing models sets limits on the form models can take (Jones 2002, also Bing 2009 whose "Evidentiary warrants" and "Framings" are examples of attempts by students to externalize their explanatory models for physics problems).

Full perceptual models appear to defy expression into an externalized medium, as we presently lack any way to communicate the depth of detail perceptual models encompass. We only externalize perceptual models after simplifying them, hopefully into a form that preserves enough salient features to still yield utility after transmission to others or to our own future selves (Jamrozik 2016, Hutchins 1995). For example, while pursuing statistical analyses, we typically abstract our perceptions into codable data points, that is, countable instances of well-defined events. While potentially useful, these data points always sacrifice some degree of detail (Strasser 2017, Borgman 2015). For example, in studying our eating habits, we may code foods by calories, fat content, protein, or other features. But we won't, and indeed, can't record the full experience of eating – taste, smell, the texture of ice cream as it melts on our tongue. Transferring a model from internal to external, and then back, is necessarily a loss function. But it also enforces a discipline on the analyst, who must balance the practical limits of perceptual data transmission against their analytic needs (Hutchins 1995).

The situation is reversed for external analytic models, which relying upon external cognition tools can include thousands, millions, or even billions of data points. Here, the detail is too great to take in without some kind of synthesis (Tukey 1977). We typically use such external models by creating statistical summaries, and it is these summaries we internalize – means, medians, standard deviations, correlations, regression coefficients, clusters, maps, and the like – whether in the form of equations or their visual analogues (Tukey 1977, Tufte 1997 and 2006). Thus,

analytic models go through simplification/summarization as part of their resolution, and then transformation into an internal, explanatory model.

1.4 Data Visualization and Statistics for All

In this light, data visualization can be seen as an attempt to use visual metaphors of physical objects as a linking mechanism between internal and external models, to move simplified explanatory models closer to being perceptual models, and thus tap some of the cognitive mechanisms of perception in the brain to do analytic work. It enfolds both conscious and automatic processes. The main goal of this research thus becomes exploring whether it is possible to use visual metaphors to embed enough of the intellectual foundations of quantitative analysis into a data visualization tool to relieve users of some part of the substantial burden of training which performing good quantitative analysis traditionally imposes.

1.4.1 Study 1

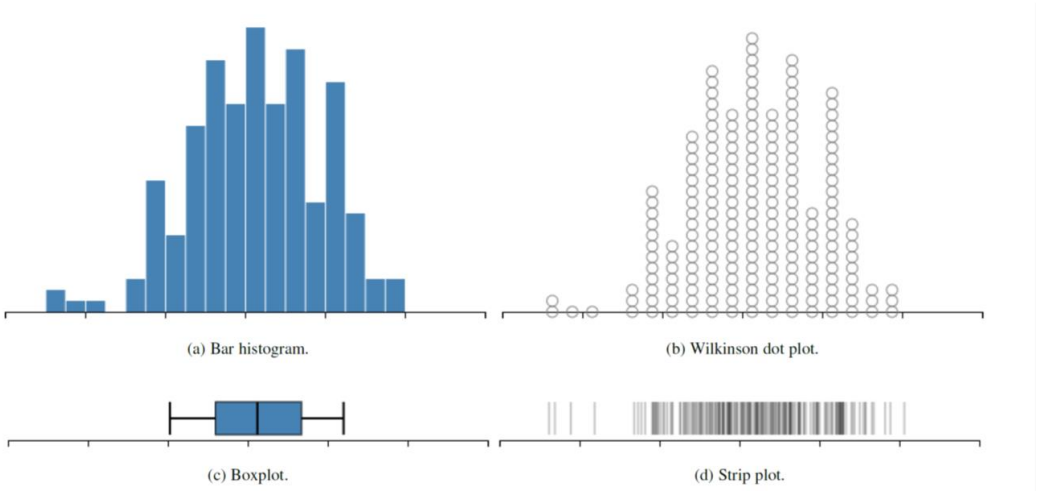
Building a statistics for all system requires tapping visual intuitions about statistical processes, if they exist. The first experiment in this research sought to test as directly as possible whether people have visual intuitions around the normal data distribution. Understanding data distributions in general, and the normal distribution in particular, is a fundamental skill in statistical practice. Many confirmatory statistical tests such as the Student t-test (Student1908), analysis of variance (ANOVA), and logistic regression are ***parametric***: that is, they rely on a known and finitely parameterized data distribution (Lehmann2005, Schervish1995). Being able to apply these methods thus requires the ability to characterize distributions from a

given sample; i.e., to assume the relationship between actual data distributions and the idealized curves parametric tests employ.

However, there currently exists little work investigating the abilities of non-experts to perceive how such curves fit real data, let alone how these curves can be used to make inferences about the underlying populations.

In chapter 3, I 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 center-point (mean) and width (spread or standard deviation) of a Gaussian curve overlaying the sample to create the best possible fit. Visualizations studied included bar histograms, Wilkinson dotplot histograms, strip plots, and boxplots (Fig. 1.4).

Figure 1.4: Distribution visualization in Experiment 1



The results show that participants in the aggregate were fairly reliable in fitting curves on data, with some variations among the four graphic representations.

Errors in locating the center-point were, with some exceptions, unbiased and small; absolute errors typically ranged from about 12% to 18% across visualization types.

Errors in finding curve width were somewhat larger, and biased for some visualization types, with absolute error ranging from about 22% to 33% across visualization types.

During the experiment, a survey of the prior statistics experience of each participant found no clear influence of this on participant performance.

I interpret the lack of bias in errors of finding the curve center point, and moderate absolute magnitudes of average errors for both center point and spread fitting, as indicators that the human visual system provides intuitions that may make manually fitting bell curves, with appropriate assistance, a feasible approach for supporting graphical inference.

This would either be as an adjunct to, or a replacement for, traditional equation-based inferential methods.

1.4.2 Study 2

Central to much confirmatory data analysis is the task of determining whether two data samples are drawn from the same or different populations. For the simplest of such inferential statistical tests---the Z family of tests---this essentially amounts to fitting a normal distribution to each sample and then determining the overlap between them, adjusting expectations for the amount of overlap in light of the n-sizes of the two samples.

However, even Z-tests, or the more common T-test, are at best poorly understood by the general population, and this is doubly true for more sophisticated statistical tests.

Yet, most of us routinely do this kind of task in our daily lives, such as when determining whether a specific credit card bill is out of the ordinary---potentially indicating credit card fraud---or when assessing and comparing a child's, employee's, or public figure's performance in school, at work, or in the public eye. These tasks boil down to detecting whether a measure is so out of range with some expected distribution as to warrant further examination.

This second experiment studies how graphical formulations of t-tests might support users who have no specialized statistical training in assessing the differences between two or more data samples. Assuming normally distributed data, a straightforward way to achieve this is to fit normal distributions to each data sample, and then visualize the samples as overlapping bell curves. A user can then manually determine whether they think the two curves represent samples drawn from the same or different underlying populations. Such a graphical formulation corresponds to a classic Z-test, or a Student t-test (Student1908).

However, normal curves fitted to samples erase potentially important details of those samples. Other graphic forms afford some of that detail to the user, with the potential of better supporting both user performance as well as user understanding of these statistical tests. Therefore, this investigation also include alternative visual representations: bar histograms (overlapped and stacked), Wilkinson dot plots, strip plots, and Tukey boxplots.

The experiment was crowdsourced, studying whether people with no statistical training can use graphical formulations to perform a T-test; i.e., to determine whether or not two samples are drawn from the same or different populations. A total of N = 212 "participants" were given a sequence of trials where they were asked this question under varying conditions: different visualizations (overlapping bell curves, stacked bell curves, bar histograms, dot plots, strip plots, and boxplots) and data sizes (36, 44, or 1,000 items per dataset). Different data sizes in this experiment provided variations in both statistical noise and effect size. A single trial typically lasted less than 10 seconds, allowing collection of a large number of trials per participant.

The results indicate, perhaps unsurprisingly, that this is a difficult task, and that people regularly overestimate how much divergence is needed for two samples to be different.

Furthermore, the task gets particularly difficult---approaching the random 50% accuracy---when the difference is small, and easier for very different samples.

What is more surprising, however, is that complex visualizations that include more detail and less aggregation yield less accuracy than the highly aggregated boxplot and idealized bell curve representations, particularly for higher data sample sizes. It appears graphical inferences of T-tests can be effective in some circumstances, but that abstracting and aggregating the data generally yields better results. In other words, humans have visual intuitions which align with this task, but may align imperfectly. These findings contribute insights into the design requirements for the next experiment.

1.4.3 Study 3

Having established that humans possess intuitions which can, with support, pair visualizations with statistically informative summaries of distributions, and make comparisons between summaries of distributions, this research turns to the work of assembling the specific visualizations that might form components of a statistics-for-all visualization toolkit, focusing on inferential statistics.

Central to inferential statistics are null hypothesis tests. Yet despite their widespread use, some studies have suggested that few practitioners have a real understanding of the subtleties of the several statistical assumptions Null Hypothesis Statistical Testing (NHST) makes (Hoekstra 2014). These same authors discuss the dangers of such misunderstanding, and tie them to the 'replicability crisis.' Yet at the same time, they document that NHST remains foundational to the modern scientific endeavor, with p-values and related statistical accoutrement ubiquitous across any research literature which relies upon quantitative analyses.

If a statistics-for-all system is to gain credibility in the scientific community, it must provision its users with methods of inferential statistics. Moreover, the resulting tool must include features which support how experienced statisticians actually approach NHST – including related analyses they perform or information resources they rely upon as they contend with the challenges of NHST. And since the proposed statistics-for-all system would rely heavily upon visualization, acceptance of such a tool by the scientific community would also hinge upon whether experienced statisticians accept visualization as a legitimate analytic tool in general.

In sum, the visual tools hypothesized by this research program must include the real understanding of these methods – both explicit and tacit – which practitioners bring to their work. They must also include features which match the expectations practitioners have for outputs from using NSHT methods, and must present these features credibly.

Study 3 is an in-depth interview study with professional statisticians to understand their visualization practices for understanding and making decisions about data. Recruiting a total of 18 statisticians with a combined 350 years of professional experience (average 19.7 years), the study focuses on three research questions:

RQ1 How do statisticians use visualization in their daily analytic work?

RQ2 What mental models of inferential statistics do statisticians have?

RQ3 What designs for visually representing statistical inference might build on current practice while enabling novices to also benefit?

This interview study was conducted as semi-structured interviews over Zoom videoconference. Each session involved three phases: (1) statistical practice; (2) graphical elicitation of their internal understanding of inferential statistics; and (3) design review of a design probe (Gaver 1999): prototype visual representations for statistical inference which incorporate the design lessons learned during Study 1 and Study 2. All sessions were professionally transcribed. Transcripts were coded using an open-coding approach (Lazar 2017). Findings were derived using thematic analysis.

At a high level, the study found that visualization tends to be a key activity in most statistician's daily workflow, and not just during presentation. Furthermore,

participants mostly reported mental models for inferential statistics that are themselves visual. Participants tend to abhor dichotomous thinking and distrust insights lacking multiple evidence. Several design recommendations emerged from these overarching themes, along with more specific, low level recommendations.

Chapter 2: Prior work

Inherent within the search for an alternative method of quantitative analysis is an assumption that such analyses are worth pursuing. Indeed, this research assumes the value of quantitative analyses, and the epistemology they rely upon, as a given. This work thus becomes an investigation into the potential for finding a method of sharing that value with a broader audience. In design terms, I am looking for a quantitative analysis method which provides affordances that will aid novice users over and above what equation-based methods provide. Many authors have contributed work relevant to this search.

2.1 Comparing Affordances: Equation-based Math vs. Visualization

Arithmetic, that is, the mathematics of common, measurable experience, both reflects reality and is a template for understanding it, having at its core rules of continuity similar to the Law of Conservation of Energy and Matter (Lakoff 1997, and Erlwanger 1973 who offers an example of a student failing when they saw math only as arbitrary rules to follow, rather than as a model for analyzing real world problems). Objects are what they are; groups of a certain size remain that way until we add or subtract objects; you can't turn $1 into $20 just by wishing.

Counting also affords benefits to our cognition; it is an abstraction that acts as data reduction, with all the potential benefits (and risks) to the analyst associated with that practice (Boyd 2012). We reduce objects in the world, sometimes of extreme complexity and individual variation, to one dimensional numbers paired with nominal labels. We reduce memory load, and eliminate distracting details, while hopefully

preserving information relevant to making a decision or a discovery (Hutchins 1995, Jamrozik 2016). For example, the statement, "5 tigers escaped from the zoo, and we recaptured 5 tigers," has more power than, "Some tigers escaped, and we recaptured some."

Counting becomes even more powerful when paired with writing, at which point we begin to externalize our cognition, offloading memory and calculative tasks (Hutchins 1995, Ainsworth 2011). Indeed, abstractions like counting are vital to the external cognition process, since we have only narrow channels (like words or geometries) for enunciating an internal model into the external world where it might be manipulated, shared, and, possibly, improved. Afterward, we have only those same narrow channels for re-internalizing the new, improved model.

Mathematical functions more complex than simple arithmetic operations can represent more complex real-world phenomena, with potentially even greater benefits in both data reduction and the capacity for external cognition. Mathematical expressions following rules which correspond to real world phenomena can allow us to make inferences about underlying causes of those phenomena, and, critically, to make predictions about what would happen were values in the model to return to some prior measured level, or even, with additional uncertainty, some as yet unencountered level (Cosmides 1996, Windschitl 2008).

For example, humans have a difficult time imagining the consequence of exponential growth – we find some of its results counterintuitive (Stango 2009). Yet modeling such growth with calculations which predict population X at time Y provides an antidote. Biologists, engineers, or economists making frequent use of

such equations (and graphics) can even develop intuitions around exponential phenomena (Stango 2009); in other words, they internalize the exponential model.

We can conceive of performing calculations upon mathematical representations of phenomena as a form of thought experiment, one which pairs mathematical rigor with the cognitive benefits of data reduction. We sometimes discover new expectations for real world phenomena, even where our unaided intuitions might fail us (Friendly 2005, Wilkerson 2017). For example, much of climate change science relies upon the exploration of mathematical models resolved by computers (United Nations 2020). From the paper and pencil era, we have the example of Gregor Mendel conceiving his Laws of Genetics by assuming measured relationships (a series of consistent ratios) in traits of successive pea plant generations represented some then unknown physical feature of cells (now understood to be chromosomes composed of dual-stranded DNA) (Sandler 2000).

The act of measurement (an activity closely allied with counting) appears to engage a mental discipline that produces benefits for explanatory model building (Serin 2001, Fekete 2008, Strasser 2017). The process of choosing what to measure means creating a data model, a mathematical framework linking abstract math and concrete experiment, that is, between hypotheses and the reality.

On the practical side, measurement tools – scales, rulers, and other instruments which increase the precision with which we perceive the world – provide additional touchpoints with reality to improve our perceptual models (Benjamin 2003, Hughes 1987, Borgman 2015, Van House 2004). Moreover, shifting our senses to external tools also provides standard measurements we can share from person to

person, when otherwise we might have only our individual impressions, which can disagree (Hutchins 1995, Fekete 2008, Hollan 2000, Ainsworth 2011, Van House 2004). Measurement is thus a cornerstone of collaborative knowledge building (Strasser 2017, Argote 1999, Nelson 1982). It is also foundational to external cognition aimed at knowledge building through quantitative analyses. External cognition amplifies human intellect (Hutchins 1995, Clark 2008).

Offloading memory and calculative tasks onto some external platform – whether a sheaf of papers or a computer mainframe – allows us to focus on thinking about what the data mean, rather than struggling to work out/recall what they are (Ainsworth 2011, Fekete 2008, Scaife 1996). The researcher can focus on finding patterns in the data, some intricate or subtle, which point to behaviors of real-world phenomena. Offloading memory also means the researcher can examine thousands or millions of data points instead of a few dozen.

Data visualization, as an alternative form of mathematical expression, incorporates all the affordances of measurement, data models, and arithmetic processes, and also supports external cognition (Scaife 1996). In Scaife and Rogers, "How data visualizations work", the authors argue that data visualization suggests a limited set of comparisons between represented values (Saife 1996). When users take in the graphic forms, they conceive of them as physical objects operating within a space, and like all physical things, obey rules of behavior. For example, a bar chart might be conceived of as a stack of objects, where the taller stack contains more objects. This seems natural to us – it reflects our experience of real-world stacks. This model fits well with the common rule in visualization practice that bar charts

must start at zero, or risk conveying a wrong impression to users, because, in the real world, stacks of objects start on the floor. We have an intuitive physics, based upon gravity, which we automatically draw upon in considering the bar chart.

By this model of operation, data visualization can be conceived of as math without formal notation, yet still usefully constrained by the intuitions of the user. These constraints will tend to limit comparisons of displayed values to a relatively narrow set, which, if they are the correct set, can become an affordance that directs lay users to making valid and informative choices in their analyses.

2.2 Graphics for Data Analysis, Exploratory and Otherwise

Much prior work has explored the particulars of using data visualization for select phases of quantitative analysis. For example, visualization is often touted as a tool primarily for exploratory data analysis (Tukey 1977) due to its bottom-up and data-driven nature. Recent work has explored this practice more closely, for example, by comparing the relative efficiency of different visual representations for exploratory tasks (Nguyen 2020, Correll 2019). These authors found that the choice of visualization can make a difference to both the ability of untrained users to notice important features in data, and even their confidence in their own findings.

Notably, these studies have not directly addressed whether users can accurately estimate basic summary measures like means and standard deviations from the most common distribution visualization methods, nor whether different visual forms produce different estimates of these bedrock statistics.

At the same time, mounting evidence suggests that visualization can also be used for confirmatory data analysis (Lehmann 2005, Schervish 1995).

2.2.1 Graphical Inference in Statistics

Confirmatory analysis when framed as inferential statistics involves using statistical methods to make inferences about a population based on a sample of data (Casella 2001). The field can be traced back to the work of Ronald Fisher, who is considered one of the founding fathers of modern statistics, (Fisher 1922, Hald 1998).

The growth of visualization as a discipline separate from statistics is a relatively new development. Creating graphical representations of data remains a common and natural part of statistical workflows (Cleveland 1993), and even central to some, such as exploratory data analysis (Tukey 1977). Accordingly, making inferences from these graphical representations – i.e. graphical inference – is as commonplace and unremarkable as making statistical inferences from algorithmic representations; for example, rather than invoking the full machinery of a formal test of normality, such as Shapiro-Wilkes, many practicing statisticians will instead eyeball the sample in a Q-Q plot (quantile-quantile plot) against the normal distribution to ensure that the resulting plot falls on the 1:1 line (indicating high positive correlation between the sample's data distribution and the standard normal distribution).

Early examples of such practice date back to Scott et al.'s seminal work from 1954 (Scott 1954) on validating astronomical models by generating artificial star charts using the model parameters and then asking people to compare them to real charts.

Similarly, bootstrapped (Efron 1992) confidence intervals, such as for user study results (Dragicevic 2016), are often reported using graphs (typically some form of error bar around a central dot).

Work by Correll and Gleicher (Correll 2014) lends at least partial empirical support to this practice, with the caveat that user assessments of uncertainty distributions vary depending upon the specific visualization method used. Hullman et al. (Hullman 2018) also suggest that visualization choice impacts user understanding of distributions, specifically differences between discrete and continuous visualization forms.

2.2.2 Graphical Analysis in Modern Visualization

With the increased availability of automatic computation and the crystallization of visualization as a discipline in its own right, graphical inference has become somehow less valued, at least outside the statistics community.

For example, with the exception of the aforementioned bootstrapped results, few visualization research papers rely on pure graphical inference for its validation, such as when comparing completion times or accuracy measures for competing visualization techniques. This remains paradoxically true despite the fact that making inferences from data is a central tenet of data visualization (Card 1999).

Much research has gone into designing visual representations that make effective use of visual channels (Cleveland 1984, Munzner 2014) so that important data features become pre-attentive and ``pop out" (Ariely 2001, Healey 1998, Healey 2012) with less mental effort than that required by tabular presentations.

Another strong appeal of a visual representation is that it can support more exploratory data analysis than dedicated statistical tests, which tend to be more tied to preconceived questions or prior knowledge about the dataset (Tukey1977).

However, it is only recently that the statistics and visualization communities have begun to ask how graphical representations of data can support aggregate or higher-order tasks beyond merely reading values, trends, and outliers. Buja et al. (Buja 2009) propose frameworks for visual statistics, where multiple visual representations form an analogue to a test statistic, and human judgement serves as the critical value for comparison. They demonstrate this approach using a ``Rohrschach" test of random data, as well as a “lineup” of small multiples, only one of which uses the real data. Note that these approaches put the impetus of decision making back on the individual judgement of the user, and as such, take a step away from the kind of (apparent) objectivity provided by purely calculative statistical tests (where the significant/not decision arises from essentially blind calculations).

In follow-up work, Wickham et al. (Wickham 2010) adapt the idea to the visualization community, describing how these protocols can be used with common visualizations to uncover new findings while avoiding false positives. Beecham et. al. (Beecham 2017) applied this "lineup" protocol for graphical inference to geographic clustering visualizations. Correll et. al. (Correll2019) used the lineup protocol to investigate whether common distribution graphics (histograms, density plots, strip plots) were effective in displaying features of datasets, such as outliers or gaps.

More centrally connected to this dissertation, the visualization community has also investigated how well people can estimate aggregate statistics from

visualizations. Correll et al. (Correll 2012) studied how line graphs can be designed to enable accurate comparisons of averages in time-series data. Albers et al. (Albers 2014) generalized this idea to six aggregate tasks for eight different time-series visualizations.

Other time-series examples include that of Aigner et al. (Aigner 2012), who enriched line graphs with color to better support visual statistics, and Fuchs et al. (Fuchs 2013), who derived line glyphs to support higher-level aggregate tasks.

In 2015, Correll (Correll 2015) summarized a suite of techniques for improving visual statistics as it applies to visualization. Two years later, Correll and Heer (Correll 2017) studied people's ability to fit trend lines to bivariate visualizations in a crowdsourced experiment, in essence testing the crowd's ability to perform regression analysis. In 2020, Nguyen et. al. (Nguyen 2020) used a crowdsourced experiment to explore how different visual aggregations might impact users' perception of summary statements about sample populations. Gleicher et al. (Gleicher 2013) study our ability to judge means in multi-class scatterplots, finding that performance is reliably high, independent of the number of points and conflicting encodings.

Fouriezos et al. (Fouriezos 2008) asked participants to compare the average height of two groups of bar charts, finding generally high accuracy, which improved with the number of bars, but degraded with higher by variance datasets.

Based on results from crowdsourced experiments, Correll and Gleicher (Correll 2014) propose redesigns of error bars in bar charts, showing how violin or gradient plots produce insights more aligned with statistical inference.

Several of these studies were inspirational for this work.

2.2.3 Distributions and Uncertainty

While descriptive statistics are often used to summarize a dataset, Anscombe's Quartet (Anscombe 1973) shows us that aggregate measures of central tendency or variability can often be insufficient for making deeper inferences about the data.

Similar to how scatterplots can show detailed relationships between two variables, data visualizations used to show single-variable distributions allow viewers to inspect the data for flaws, missing values, or noise. Such visualizations typically show more detail -- involve less aggregation – than descriptive statistics like means, medians, or standard deviations.

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. Histograms are subject to distortions, depending upon the aggregation scheme used (Correll2019). Choosing the right number of bins, and thus bin size, is a primary concern to which the visualization community has responded with several rules of thumb. This work will employ Sturges's rule (Scott 2009), which is based on the assumption that the distribution to be binned is Gaussian (this is appropriate, since all trial datasets for these experiments are drawn from essentially normal distributions).

Distribution visualizations remain an active area of visualization research, producing such alternate representations as strip plots, density plots, violin plots, and gradient plots (Kay 2016}. These newer forms offer variety in their level of abstraction vs. detail, but 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. A compromise may be found in Wilkinson dot-histograms (Moon 2016), where each discrete item in a bin is represented as a circle in a stack of circles. This may provide users with a familiar overall design, while providing more detail about sample size (also suggested by the work of Kale et. al. (Kale 2021), Hullman et. al. (Hullman 2018, Correll et. al. (Correll2019).

Another active area of research which offers potential design implications for distribution visualization is statistical uncertainty visualization, since we often calculate uncertainty as an idealized distribution of potential variance around a point estimate. 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 (Kay 2016).

Hypothetical outcome plots (HOPs) (Hullman 2015) use animated draws to illustrate uncertainty, and have shown superior performance compared to violin plots and error bars.

Of particular interest to experiment 1 is work by Hullman et al. (Hullman 2018), where participants were asked to sketch their predictions of uncertainty distributions using both continuous and discrete representations prior to seeing the

actual distributions. In a similar fashion, this study asks participants to fit a continuous distribution to discrete distribution graphics.

2.3 Visual Intuitions For Statistical Inference

John Tukey's Exploratory Data Analysis (Tukey 1977) helped launch the modern era of data visualization with its publication. It argued that data visualization provided researchers powerful means for detecting outliers or errors, identifying a multitude of patterns, and recognizing structures within their data, all of which made it an ideal analytic tool for hypothesis formation. Given the flexibility visualization affords researchers in finding the unexpected within data, he argued that visualization tools are particularly well suited for an exploratory---rather than confirmatory---role. Tukey went on to recommend the use of traditional, equation-based statistical methods for testing and confirming hypotheses. He believed that different methods should be used for hypothesis formation and confirmation, to ensure the analyst avoids the trap of circular reasoning.

Since Tukey endorsed this use of data visualization, it has gained growing acceptance, not just for communicating findings, but also as an analytic method in its own right, at least for early stage analysis. However, while visualization may be seen as useful by the research community, traditional confirmatory statistics retain a status as essential.

Tests of statistical significance are a cornerstone of statistical and scientific practice because, while still including elements of subjective, arbitrary assumptions, such as the choice of a significance level (alpha value), statistical tests nonetheless provide standards for decision making that push interpretations of experimental

results closer to objectivity. Ideally, they act as a check when our enthusiasm for a favored hypothesis might otherwise have motivated us to make claims based upon results that could just as easily have been explained by random noise in the data. This indispensable discipline comes in myriad forms, such as Z-tests and t-tests, where each is suited to different analytic methods but all are founded upon a five-step process:

1. The researcher forms an expected value based upon their null hypothesis;

2. The researcher selects a level of improbability that they consider statistically significant (for example, traditionally alpha = 5% in the social sciences);

3. The researcher extracts an actual value from the data and compares it to the value expected under the null hypothesis, thus creating a test statistic;

4. The researcher applies appropriate probability calculations (which account for a set of assumptions tailored to the particular data collection in question) to determine how unlikely it was that random noise would result in a test statistic as large as the one they found;

5. If the result is so unlikely to have happened by chance that it breaches the pre-chosen significance limit, the researcher rejects the null hypothesis and declares that they have at least some support for their favored hypothesis.

This process affords the researcher a sense that their results are more than mere anecdote, but instead, form hard evidence. Yet, applying statistical tests correctly can be a challenge.

Equation-based statistics include subjective elements at nearly every stage, from the selection of an initial hypothesis and its mathematical expression, to the choice of a p-value limit, to the creation of experimental design, to the choice of the right probability assumptions to form a test statistic. These elements make it possible for people to, for example, ``p-hack," either knowingly or by mistake (Cumming 2014).

Ironically, the wide availability of automated computation tools, such as spreadsheets and other statistical software packages, may compound the problem.

Automated statistical software turns statistical procedures into magic black boxes, which consume data and spit out results without requiring users to understand all the assumptions which undergird those results. Moreover, it becomes simple to churn out multiple analytic variants on a dataset, until one pops out that looks promising. When conducting 100 experiments, we should expect through random chance to find five results that are significant at the 5% level. Reporting these hits without the misses gives a false impression of evidence supporting a finding when there is none. This kind of error---which can be entirely unintentional---may be at the root of the “replicability crisis”, which has plagued many scientific fields, including visualization (Kosara2018, Echtler 2018). Researchers have attempted to address this crisis through pre-registration, registration reports, replications of past work, etc.

Despite these issues, the application of equation-based statistical methods to hypothesis confirmation has become a linchpin in advancing our understanding of the world. Tables of results with p-values are ubiquitous within published scientific work. This elevates equation-based statistics in the scientific community, while leaving visual analyses in a secondary role.

For example, during their undergraduate studies, most science majors will take an equation-based statistics course by requirement, but outside of a few select fields (such as data science), they are unlikely to learn visualization in any formal way.

This situation suggests two gaps in the literature which this research will attempt to fill.

First, we can attempt to demonstrate that visual analyses can act in the hypothesis confirmation role. This would at least bring visualization on par with purely calculative methods, though it would in itself provide no reason for an analyst to switch.

Second, we can document affordances provided to researchers through visual confirmation analyses which differ from those provided by equation-based methods. Different affordances would create the potential for visualization, in at least some circumstances, to provide a subset of users an advantage over purely calculative methods in pursuing evidence-based decision making when confronted with data that include inherent uncertainty.

There is some research which suggests working with graphic forms of mathematical representation improves user understanding, including statistical

understanding, such as recent work in uncertainty visualization that uses graphic 'elicitation' (Crilly 2006). In elicitation, a researcher gathers information about what respondents are thinking by asking them to draw pictures of mathematical relationships, rather than forcing respondents to attempt to describe the same relationships with words or equations. While this is often used as a data gathering technique, by asking participants to draw visual representations of data as part of the evaluation, respondents improve recall and comprehension (Crilly 2006). Kim et al. (Kim 2017) found that graphically eliciting participant's prior knowledge and data they had observed helped them to both reason about, and remember, findings. In follow-up work, Hullman et al. (Hullman 2018) asked participants to sketch their predictions of uncertainty distributions using both continuous and discrete representations prior to seeing the actual distributions. Participant predictions of expected outcomes improved. These suggest that actively engaging users in creating or manipulating visualizations can serve to improve their analytic understanding.

A visual approach to statistics can help novice users perform advanced statistical tests (Grammel 2010, Huron 2014, Pousman 2007). As a case in point, recent work has shown how even novice users can compare averages in time-series data (Albers 2014, Correll 2012), fit trend lines to point clouds (Correll 2017), and make mean value judgments in multi-class scatterplots (Gleicher 2013) without specialized training or knowledge.

2.3.1 Visual Confirmatory Analysis

Tukey's reasoning on using one set of methods for hypothesis formation and a second set for hypothesis confirmation, comes with no requirement that visualization fill the first role. Should we develop sufficiently flexible equation-based methods for hypothesis generation, we could conceivably pair them with visual analyses for hypothesis confirmation. Or we might create two different kinds of visual analyses, one set for exploration, the other for hypothesis testing.

But what would a data visualization designed for hypothesis confirmation look like?

Just as we can describe statistical tests in five steps, we can describe what goes into those steps as four intellectual products:

1. **Precise measures of some phenomena** (precise relative to the effect we want to explore), including an observed value of interest. These measures emerge from our experimental designs, and comprise the data;

2. **Distributions of idealized data populations** from which the data could have been drawn given the observed data distributions and sample scheme. These reflect our assumptions about the true population, where, for example, in the case of data believed to come from a normally distributed population, we use our observed data to calculate a mean and standard deviation. We treat these derived values as estimates of the unseen, true values for the total population, that is, we assume they describe the central tendency and spread of that larger population;

3. **An expected value.** This derives from the null hypothesis, based usually upon an assumption of random behavior as opposed to behavior driven by some combination of underlying forces we hope to detect. We typically calculate this by considering what would happen should we run our experiment in an environment in which the effect size for the forces we wish to explore were set at zero;

4. **A table of probabilities** expressing how unlikely we are to measure results that deviate by this or that amount from the expected value, given the sample size. This table derives from a function built with the same assumptions that underlie the idealized data distributions, paired with a model to account for our sampling processes.

All these products are fundamentally quantitative, and thus, could be expressed with appropriate geometries. At a minimum, we might create visualizations of hypothesis tests by performing a one-to-one mapping of the math onto shapes and lines, just as we can, for example, represent regression equations with straight lines drawn through a data field on a scatterplot.

But it still remains to be seen whether we, in fact, perform a regression with our eyes, or whether we are instead performing some visual proxy for the regression or other mathematical operations. Yuan et al. (Yuan 2019) found that when faced with a complex visual task, respondents may reduce their cognitive load by relying upon more primitive perceptual cues, or proxies. These provide shortcuts but may reduce accuracy (Yuan 2019), (Ondov 2021).

If users rely on mere proxies when faced a visual form of a statistical test, then being able to perform a literal translation of quantities into geometries tells us little about whether we might usefully decode such images with our eyes, nor what affordances (if any) such an approach might provide to the prospective user.

In light of these questions, this research will attempt to explore the potential for graphical formulations of a classic statistical test to:

1) Allow users to perform tests accurately;

2) Provide affordances which novice users will find helpful over and above those they would find in more traditional, purely calculative methods.

Chapter 3: Experiment 1: Fitting Bell Curves

3.1 Study: Fitting Bell Curves

The goal of this first study was to understand how well, or even if, lay users would be able to fit normal curves to a data sample drawn from a normal distribution. If so, this would suggest that people have visual intuitions that would support them in linking real data samples with the idealized distributions upon which inferential statistics ultimately depend.

This was a crowdsourced user study, run via Amazon Mechanical Turk, where participants were asked to fit a normal curve on a data sample through an interactive interface which allowed them to control the position (mean, or ``center point'') and width (standard deviation, or ``spread'') of a Gaussian curve. Sample distributions appeared on screen in one of four different visualizations: a bar histogram, dot histogram, strip plot, or 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 in future experiments.

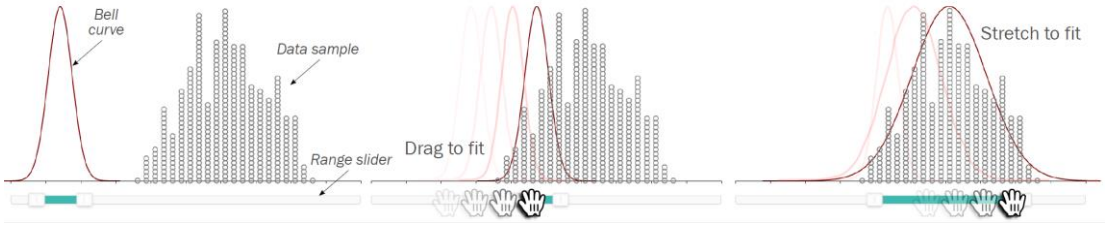
Along with the visual representation of the data samples, the experimental design varied the size of the random sample (n = 50 or 200), and the noise in the data (coefficients of variation = 0.2 or 0.4).

To minimize the possibility of a “left-to-right bias,” (Spalek2005) 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 included only Wilkinson dot histograms, and was pre-registered (anonymously) on OSF.{https://osf.io/behwz/?view_only=96de7ce0e71146c9abd8d1f79c46915e}. The current version of this study expanded to include the three additional graphic forms. The new anonymous pre-registration can be found on OSF.{https://osf.io/a9b48/?view_only=70bbe4653a7d41ed8649abf42400583a}. The discussion below only concerns the new experiment; the original data for 200 participants are not included in this study and are thus not reported here.

Figure 3.1 – Experimental task



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.

3.1.1 Participants

This study focused on low-level perceptual tasks that require no specific training or prior data visualization expertise, and thus was a good candidate for crowd-sourcing via Amazon Mechanical Turk. Prior work has shown that simple visual tasks like these are particularly amenable to this kind of study (Heer2010). The essential task here is ‘curve fitting’, that is, matching the curve to the

distribution. While this results in more than one distinct measure (mean and standard deviation), this experiment treats it as a single task, that is, fitting. Should future work by other authors suggest this assumption is unwarranted, the results of this experiment might require revisiting.

The use of Mechanical Turk (MTurk) affords little control over participant demographics and expertise, or their computer hardware. This study used a demographic survey and exclusion rules for certain kinds of devices in an attempt to account for these factors.

All experimental factors were within-participants. The experimental plan called for recruiting a total of 100 participants. Participation was limited to people living in the United States due to tax and compensation restrictions imposed by the study IRB.

Participants were screened to ensure at least a working knowledge of English; this was required to follow the instructions and task descriptions in the 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). The actual 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.

3.1.2 Apparatus

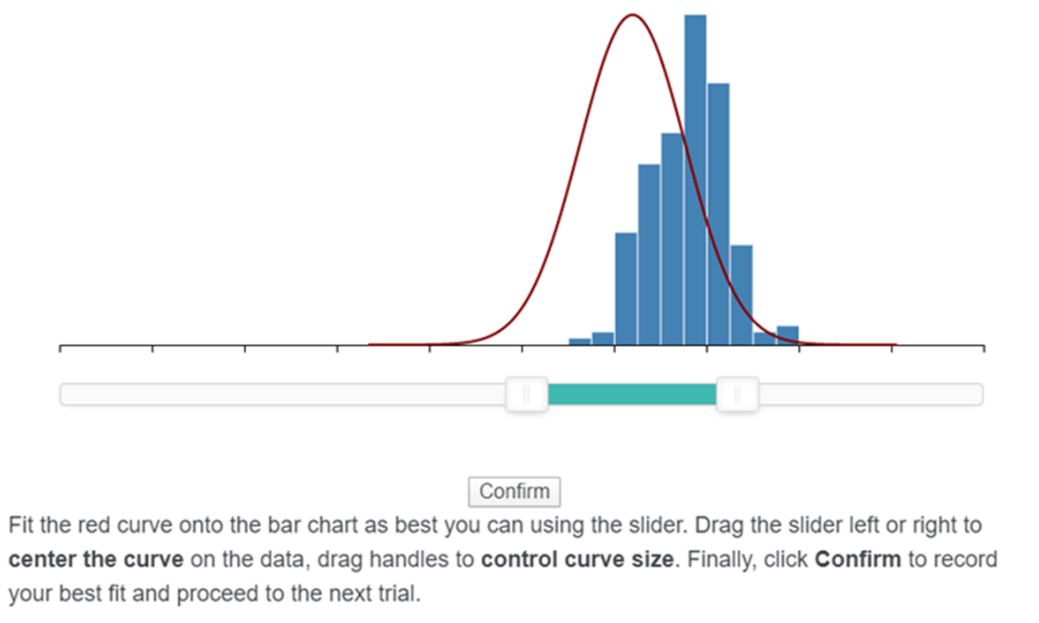
A crowdsourced setting makes it difficult to ensure respondents will have consistent computer equipment, a concern since the study apparatus was distributed through the user's web browser. Therefore, a screening question required that all

devices used by respondents were personal computers (laptop or desktop) or touch tablets; smartphones were disallowed due to the limited screen space available. Browser windows were required to be at least 1280 x 800 pixels.

3.1.3 Task and Training

The tasks consisted of fitting a normal (Gaussian) curve onto a data visualization using a range slider (Ahlberg1992) that controlled the spread (i.e., standard deviation) of the curve using the width of the interval and its center point (i.e., mean) by moving the position of the interval on the slider.

Figure 3.2 Screenshot of a typical task.



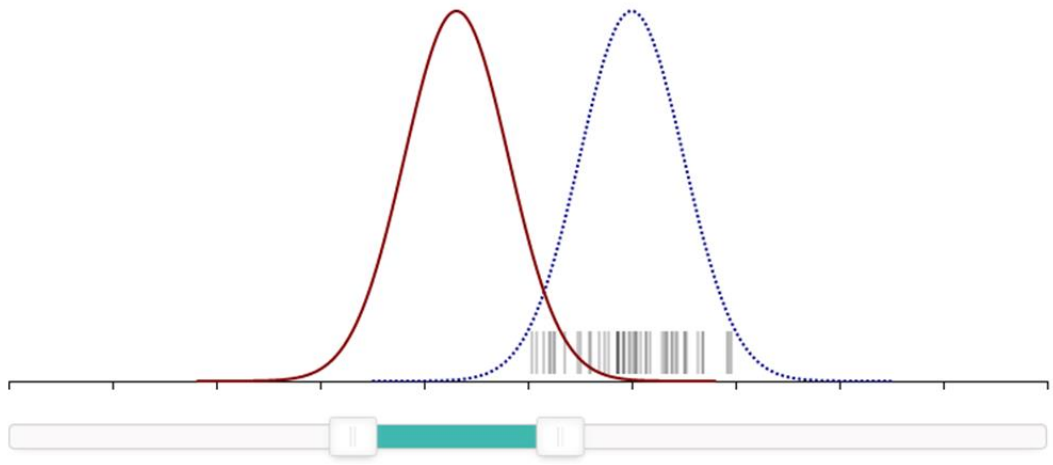
Curve fitting task (bar histogram)
Screenshot of the testing platform showing a red Gaussian curve controlled by the user. The bar histogram represents the sample to fit. The range slider below the horizontal axis controls the spread using the width of the range, and the centerpoint using its position on the slider. Once the participant is satisfied with the fit, they click the ``submit'' button to finish and then proceed to the next trial.

Participants were instructed to find the ``best fit" between the curve and the visualized data.

In a training trial (Fig. 3.3) for each visualization block (which also served 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.

The testing platform was implemented in JavaScript using D3 (Bostock2011) and embedded into a Qualtrics survey accessed using the participant's web browser. 'NOUSlider' (https://refreshless.com/nouislider/) served as the range slider implementation.

Figure 3.3 Training and Attention Trial



Training and attention trials
Participants were asked to match their controlled curve (red line) with a correctly fitted curve (blue line). This example shows a strip plot; we included training trials for all visual representations. These trials also served as attention trials for our experiment.

Participants achieved a fit by moving a slider, which adjusted the curve. Participants could either drag the end points of the slider, adjusting both spread and center point, or they could drag the whole slider from the middle, changing only the

center of the curve without effecting spread. Participants were unable to drag the curve itself to move its position or spread, instead having to make all moves via the slider.

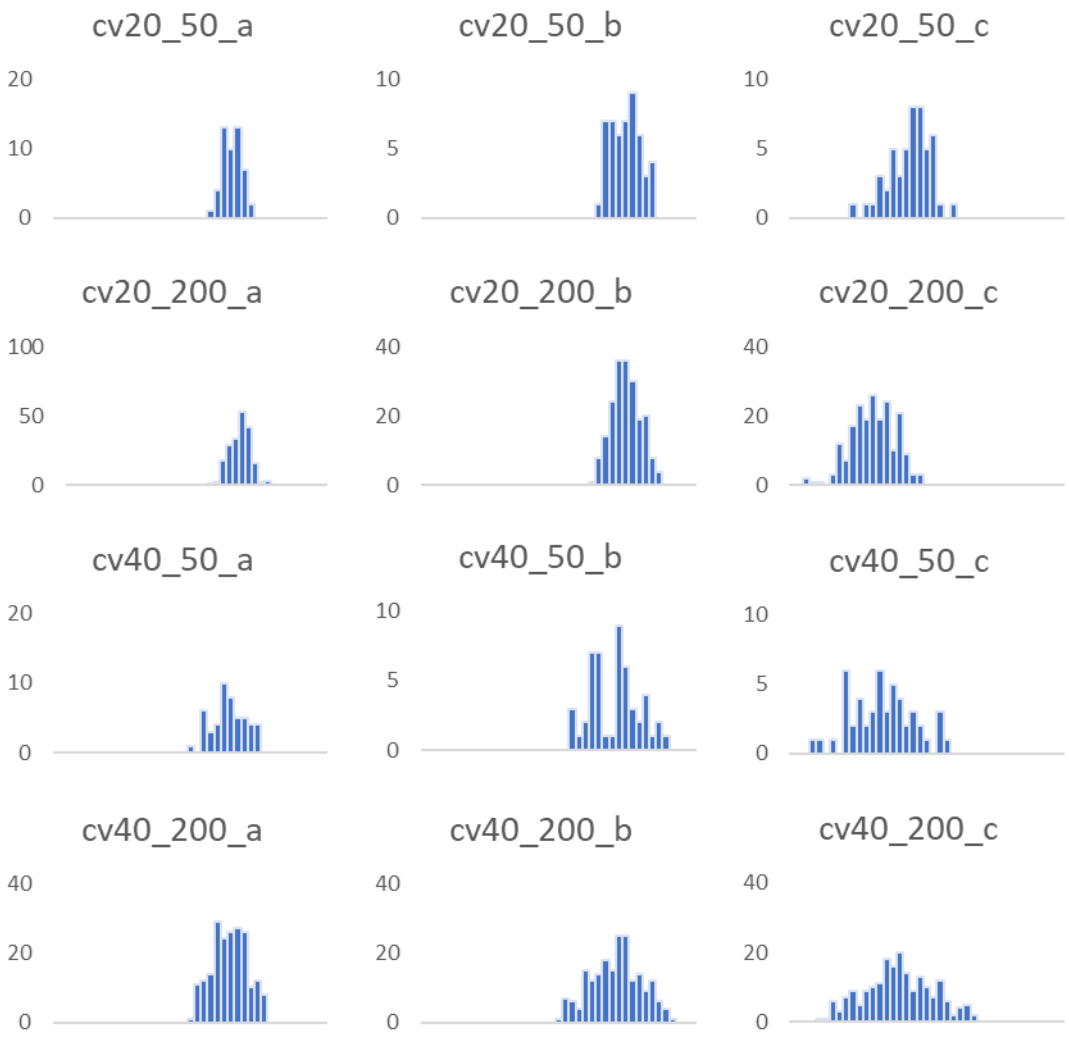
3.2 Dataset Generation

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 (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 only a fraction of these, based on spread.

All trial datasets were drawn from a normal distribution, as the purpose with this experiment is to fit normal curves rather than having participants detect the optimal distribution to use. Despite this, by varying the experimental factors (below), trial distributions arose with sufficient "noisiness" in appearance to present a challenge to respondents. This does mean that our datasets were all more or less symmetric (Fig. 3.4).

Figure 3.4 Experimental stimulus datasets, as histograms



3.2.1 Experimental Factors

The experiment varies three factors:

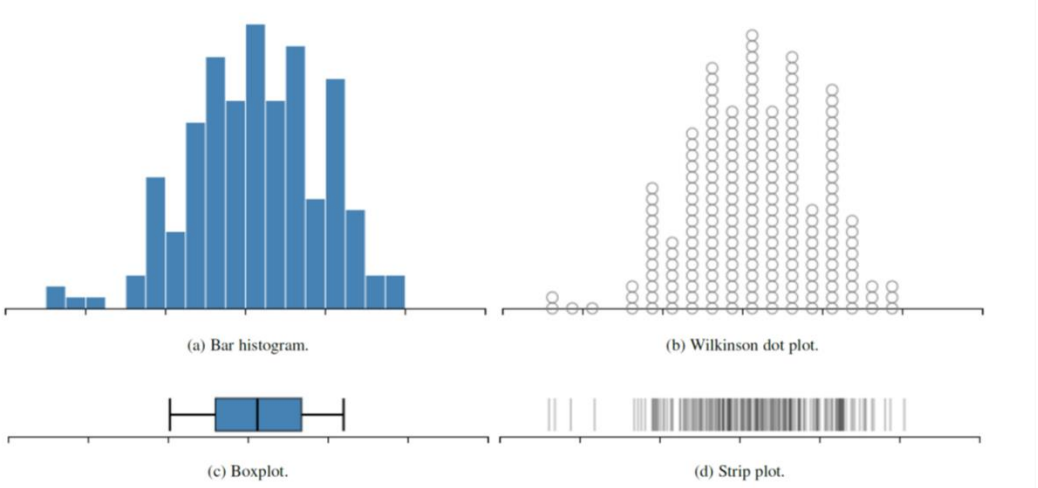
Data Size (D): The number of cases in the data sample to fit. In general, larger samples, when drawn stochastically, tend toward greater regularity and adherence to the underlying distribution from which they derive. This experiment employed two levels of data size: 50 and 200 cases. At 50 cases, samples have

sufficient size to reliably subject them to statistical tests, for example the T-test, or Cohen's D without correction for small sample size; this serves as a marker of the experimental condition having sufficient regularity to create a "fair" test of respondent ability.

Spread (S): The standard deviation of the sample being fitted. The experiment employs two levels for this factor, expressed as the coefficient of variation (CV) (or relative standard deviation), i.e., as a ratio between the standard deviation σ and the mean μ (σ/μ): 20% and 40%. These values reflect our prior experience, and the results of the first version of this experiment (note that this was inaccurately described as a "pilot test" in our preregistration). These values ensured test samples provided both relatively noisy distributions (for high values of S) and relatively regular distributions (for low values of S). Data samples are within 1% of these target levels.

Visualization (V): The visualization type used to represent the data samples. Drawing from the literature, this experiment employs four distinct distribution graphics (Figure 3.5).

Figure 3.5 The Four Distribution Graphics



Bar histogram: A "classic" histogram where the aggregated number of data items for each bin is represented using a bar of uniform width (Figure 3.5a).

Wilkinson dot histogram: A variant of histograms initially proposed by Wilkinson (Wilkenson 1999), dot histograms are unit visualizations (Park 2018) that organize individual dots (circles) for each item into bars for each bin (Figure 3.5b).

Boxplot: A box-and-whisker plot as pioneered by John W. Tukey (Wickham 2011), where a central rectangle contains the middle half of the data (from the 25th to the 75th percentile), the median (50th 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 10th and 90th percentiles) (Figure 3.5c). Boxplots in this study were not augmented with icons like dots or stars to indicate outliers. This sacrificed some information while increasing the simplicity of the graphic.

Strip plot: A unit visualization (Park 2018) 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 3.5d).

This selection of graphic types provides variation in the level of abstraction vs. detail. For example, the strip plot provides a view of each data point, while the box plot includes no individual data, but only statistical averages. The two histograms represent a points between these two extremes.

Forming histograms requires the creation of "bins", that is, the uniform ranges within which data points are enumerated. The choice of bin size is a key parameter in forming histograms (Correll 2019). Narrower, more numerous bins create more detailed, 'spikier' displays; wider bins give a more aggregated, blockier appearance. This experiment did not directly model variations in histogram bin width, but instead used a consistent number of bins (50) across the horizontal axis. The number of bins with data in them thus related to spread, with greater S values resulting in more bars per histogram.

3.2.2. Experimental Design

This experiment employed a within-participant design, where each participant saw all data sizes, spreads, and visualizations. The relatively small total number of conditions kept typical sessions shorter than 20 minutes, minimizing fatigue and maximizing alert attention for crowd workers. The order of trials was randomized for each individual participant. This yielded the following design (Table 3.1):

Table 3.1 Experimental Design

2	**Data Size D** (50, 200 samples)
2	**Spread S** (0.2, 0.4)
4	**Visualization V** (bar, dot, box, strip)
3	**repetitions** (labeled a,b,c)
48	**trials per participant**

The pre-registered experimental plan called for 100 participants and a total of 4,800 trials. However, due to a miscommunication within the research team, 150 respondents were recruited (discussed later in the section on deviations from preregistration). During each trial, the instrument captured accuracy of fit and completion time. Accuracy was based on two metrics:

Mean error: If μc is the mean of the normal curve fitted by the participant and μs is the actual mean of the sample, mean error is calculated **|(μc - μs)|**. Most analysis below take the absolute value of this metric.

Standard deviation error (%): If σc is the standard deviation of the normal curve fitted by the participant and σs is the actual standard deviation of the sample, standard deviation error is calculated as **|(σc - σs)/σs|**, expressed as a percentage. Expressing this measure as a percent error provides a more comparable scale across trials with different S values. Most analysis below take the absolute value of this metric.

Completion time: was measured from when the trial was displayed to the participant until the participant submitted an answer. This metric did not form part of measuring task accuracy, but rather, served as a rough control to eliminate crowd workers who may have become distracted during administration of the instrument.

3.2.3 Procedure

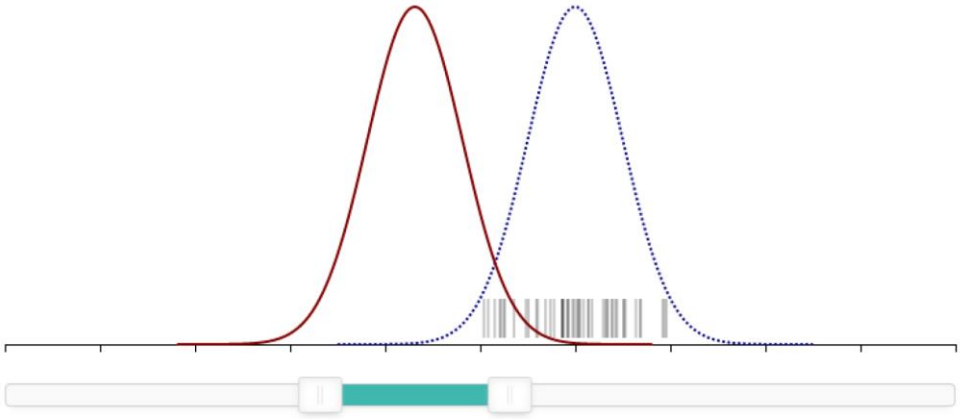
All recruitment was conducted via Amazon Mechanical Turk. Participants that fit the eligibility criteria opened the survey in a separate browser window. At the end of their participation, they copied a unique completion code back into the Mechanical Turk interface, and were later paid as their work was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

Each session started with a consent form. Failing to give consent terminated the experiment. Participants were instructed that they could abandon their session at any point in time. Unfortunately, it was only possible to pay participants who completed a full session. The consent form informed participants of this fact at the start of the session.

After consenting, participants were asked their age, education level, and knowledge of statistical concepts. Participants had to reaffirm that they were using a tablet or computer to participate.

Participants were shown practice trials for each visualization type, including 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 (Fig 3.6).

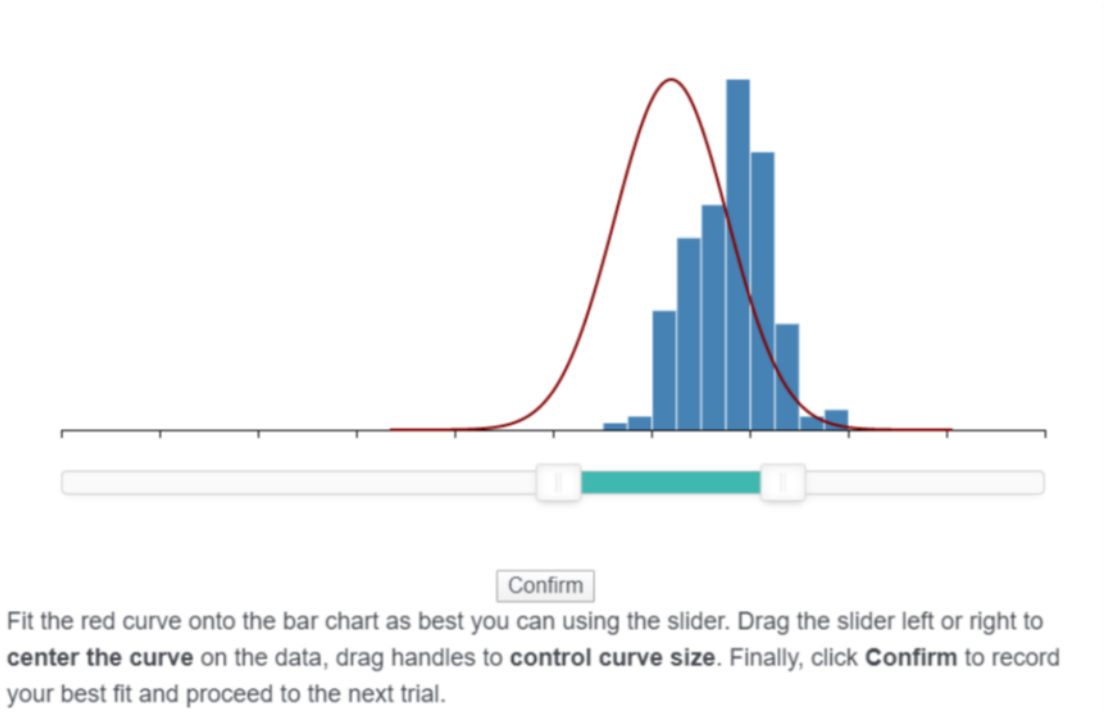
Figure 3.6 Training and attention trial



These practice trials also served as “attention trials.” The purpose of these attention trials was to eliminate responses from crowd workers 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. The consent form disclosed this fact.

Each individual trial started with the display of the dataset and the curve and ended when participants clicked the “confirm” button (Fig 3.7).

Figure 3.7 – Experimental instrument



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 the Qualtrics logs indicate that these participants took long breaks between trials (presumably due to interruptions).

3.3 Hypotheses

Estimation of means will be more accurate than estimation of spread. Intuition, prior experience, and the literature all suggested that people are able to

visually determine averages with high accuracy (Gleicher 2013), but fitting the curve to the sample will be less accurate.

Participants will be more accurate at estimating both mean and spread as the number of data cases 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.

Performance will vary with visualization type. In particular:

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 normally distributed data samples.

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 underestimate the spread of distribution when using boxplots, due to their excluding the tail ends of the distributions.

3.4 Results

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, only move the curve's center point. We believe this was a mistake, leading some participants to believe moving the center point constituted the full task. Since we are unable to assess the impact of this apparently misleading training, we opted to remove those 10 participants from our analysis. Based on the preregistration, we eliminated outlier trials (not participants) with a completion time higher than 3σ ; 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 (Heer 2010), and

(b) none of our hypotheses are based on completion times.

The final dataset, after removing outliers, had 5,528 trials. 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%).

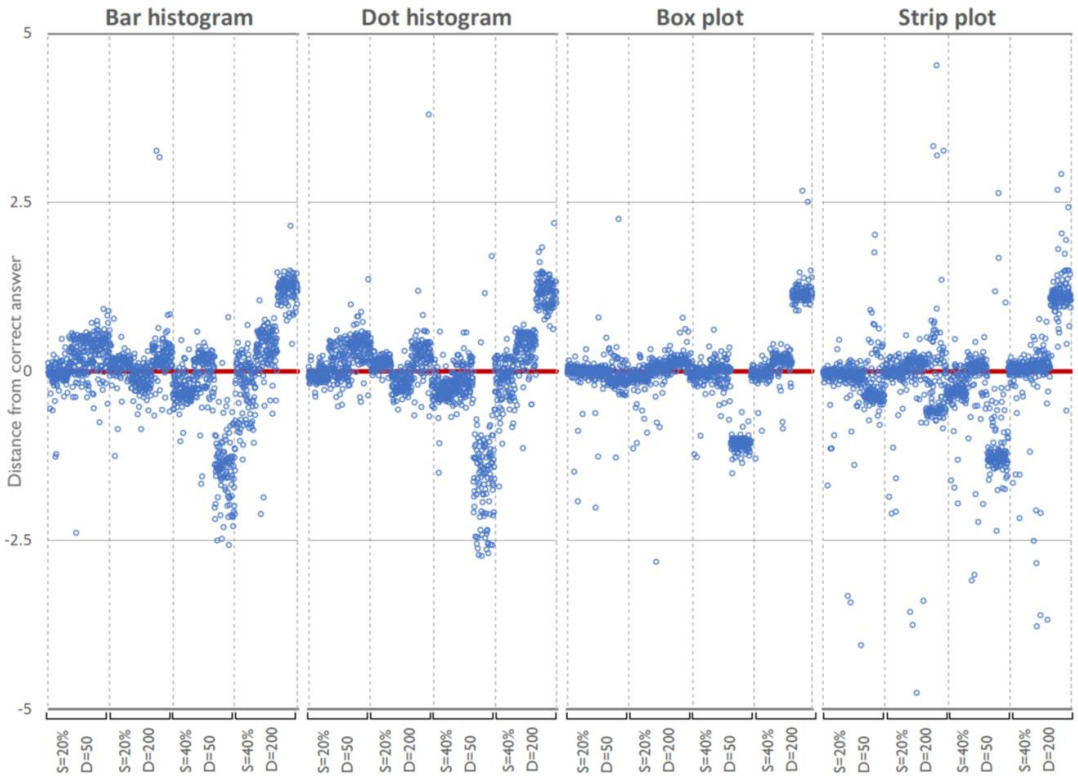
Below we analyze participant performance and then go into detail on the characteristics of the different factors.

3.4.1 Averages and Individual Analysis

Figure 3.8 summarizes all trials for means (blue dots) and Figure 3.9 does the same for standard deviations (red dots).

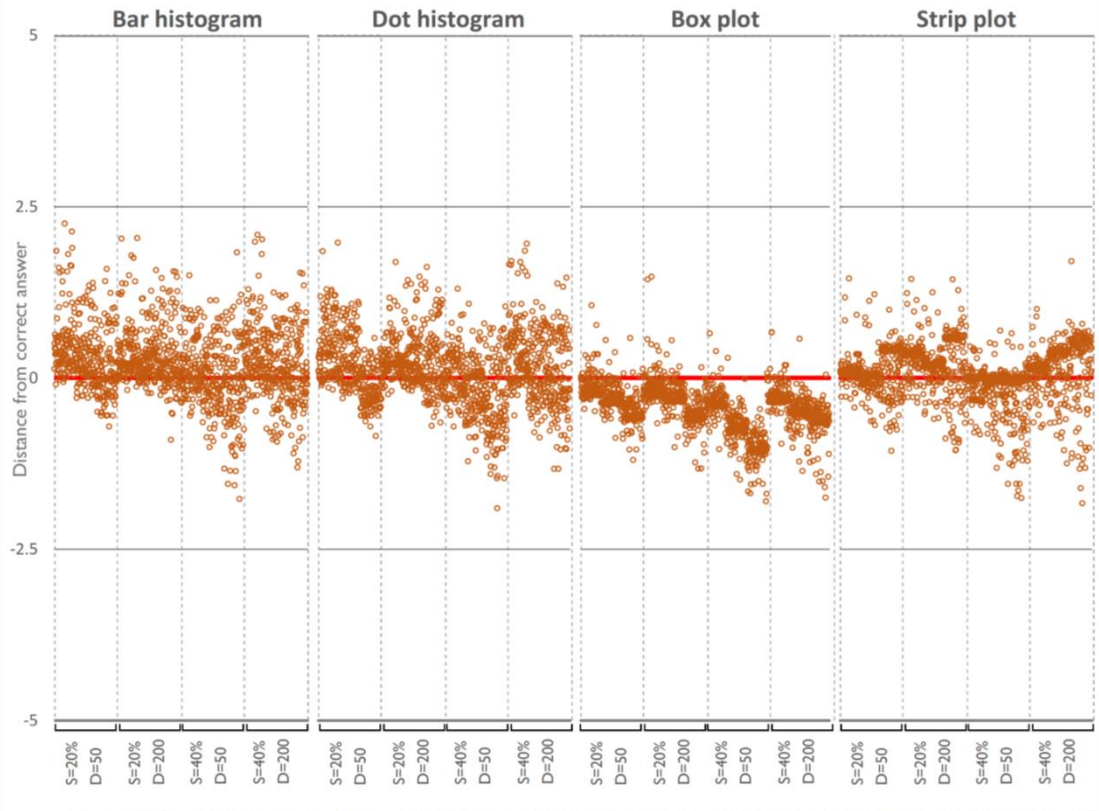
Estimates of the mean are more or less centered around the baseline, with errors equally distributed above and below zero. This suggests no systemic trend for over- or underestimating the mean. Boxplots appear to have the smallest magnitude errors.

Figure 3.8 Errors in estimating means



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).

Figure 3.9 Errors in estimating standard deviations

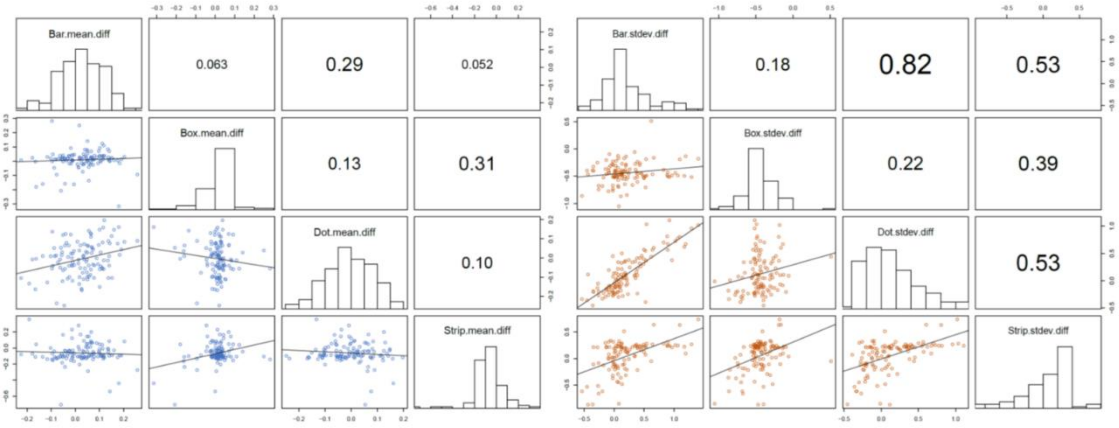


Standard deviation error performance. Distance between true values and individual participant estimates of dataset standard deviations. The red line denotes zero distance (correct answer). Each region for a specific spread *S* and dataset size *D* represent sets of trials, with the same datasets used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

Individual estimates of standard deviations appear larger, with typically greater spread around the zero-line. This, however, may only reflect the different scale of the two measures (absolute error vs. percent error). More importantly, errors in standard deviation estimation are unbalanced, biased upward for both histogram types, and downward for box plots, despite participants being somewhat more consistent in their standard deviation estimates when using boxplots. Strip plots at times yielded significant underestimation (potentially because of overplotting), whereas performance was more consistent for other trials.

Figure 3.10 considers whether respondent performance on one kind of graph is predictive of how they will perform on another. Each dot represents a respondent's average percent error at the task across all trials for that graph type. As stated earlier, outliers larger than 3σ have been removed from this data. However, respondent performance in estimating standard deviations shows considerable variation.

Figure 3.10 Comparing performance between visualization types



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.

Absolute values of Pearson r correlation on performance between pairs of visual representations also appear 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). However, performance on histograms does not appear to correlate well with performance on strip plots or boxplots (r < 0.13). This inconsistency between graph types suggests that variations in performance on this task derive more from the design characteristics of specific distribution visualizations

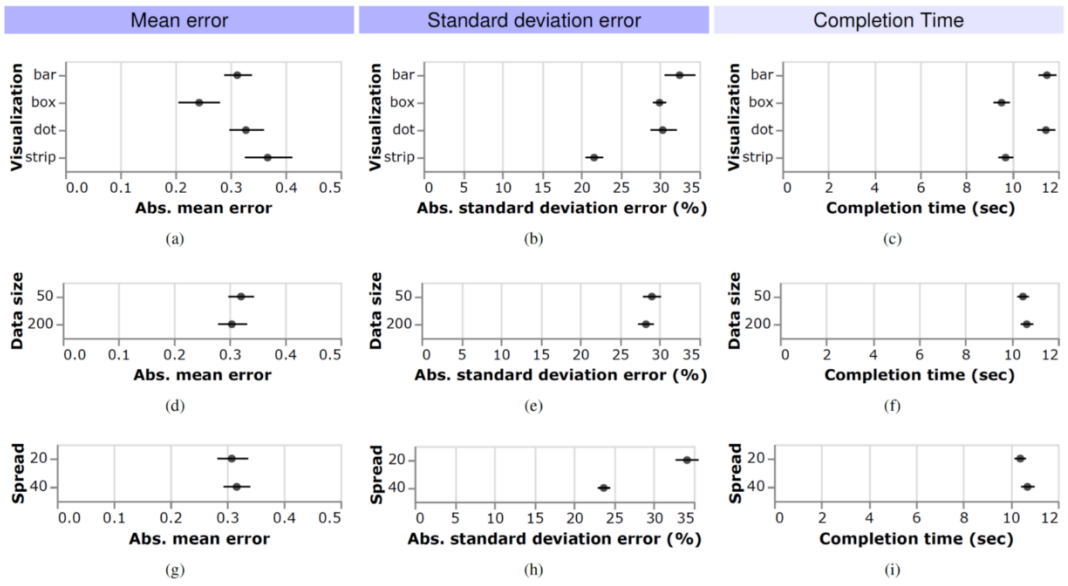
respondents were presented with that from any variations in the native skill or prior experience of individual respondents.

For standard deviation error, absolute values of the 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).

3.4.2 Analysis by Characteristics of Factors

We analyzed the results from the study using bootstrapping (Efron 1992) (N = 1, 000 repetitions) to compute 95% confidence intervals (CIs) (Dragicevic 2016). We also report effect sizes based on these intervals.

Figure 3.11 Bootstrap results of analysis by experimental factor



Overall performance. Effect of Visualization *V*, Data Size *D*, and Spread *S* on mean error, standard deviation error, and completion time. (Note that completion time is not a significant part our analysis, and is only included for the sake of completeness.)

The first three rows of Figure 3.11 summarize 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). Completion time is included for completeness only, and does not contribute to the main body of this analysis, or later discussion.

The second row in Fig. 3.11 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 3.11, 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).

Table 3.2 Effect sizes for absolute mean error and absolute standard deviation error(%) for the four Visualization types *V*.

Visualization *V*	**mean**	**s.d.**	**Cohen's** *d*
Absolute mean error:			
– Bar histogram	0.312	0.480	−0.001
– Boxplot	0.242	0.713	−0.140
– Dotplot	0.328	0.581	0.031
– Strip plot	0.367	0.833	0.110
Absolute s.d. error (%):			
– Bar histogram	32.2	39.3	0.148
– Boxplot	30.7	17.1	0.082
– Dotplot	30.6	34.6	0.078
– Strip plot	21.9	24.6	−0.308

Table 3.2 summarizes the effect sizes for absolute mean error and standard deviation error (%). For Visualization (the first row), error in estimating the mean was

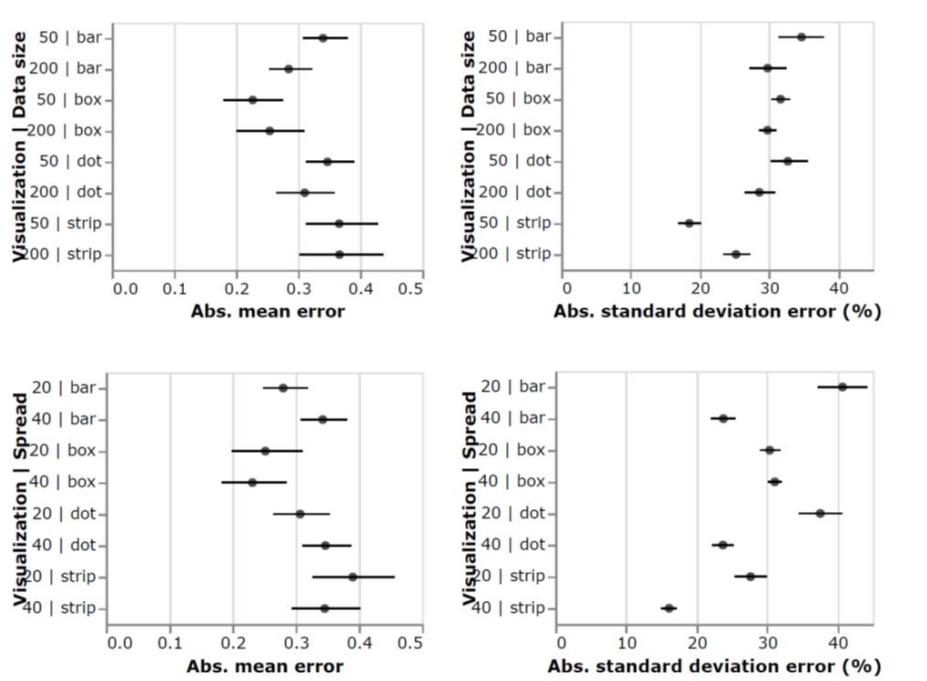
lower for box plots (absolute mean error=0.24, Cohen's d = 0.14) compared to the other chart types. Strip plots had slightly worse performance (absolute mean error=0.37, d = 0.11), but were comparable in performance to bar and dotplots. Participants were considerably more accurate in estimating standard deviation with

strip plots compared to the other visualizations (absolute percentage s.d. error=0.21, Cohen's d = 0.31). Bar histograms were the least accurate for estimating standard deviation (absolute percentage s.d. error=0.32, Cohen's d = −0.15), but the difference was lower.

3.4.3 Interactions between experimental factors

Figure 3.12 looks for potential interaction effects between experimental factors and respondent task performance.

Figure 3.12 Interactions among experimental factors



Interactions between factors. Effect of Visualization *V* with Data Size *D* and Spread *S* on absolute mean and standard deviation error.

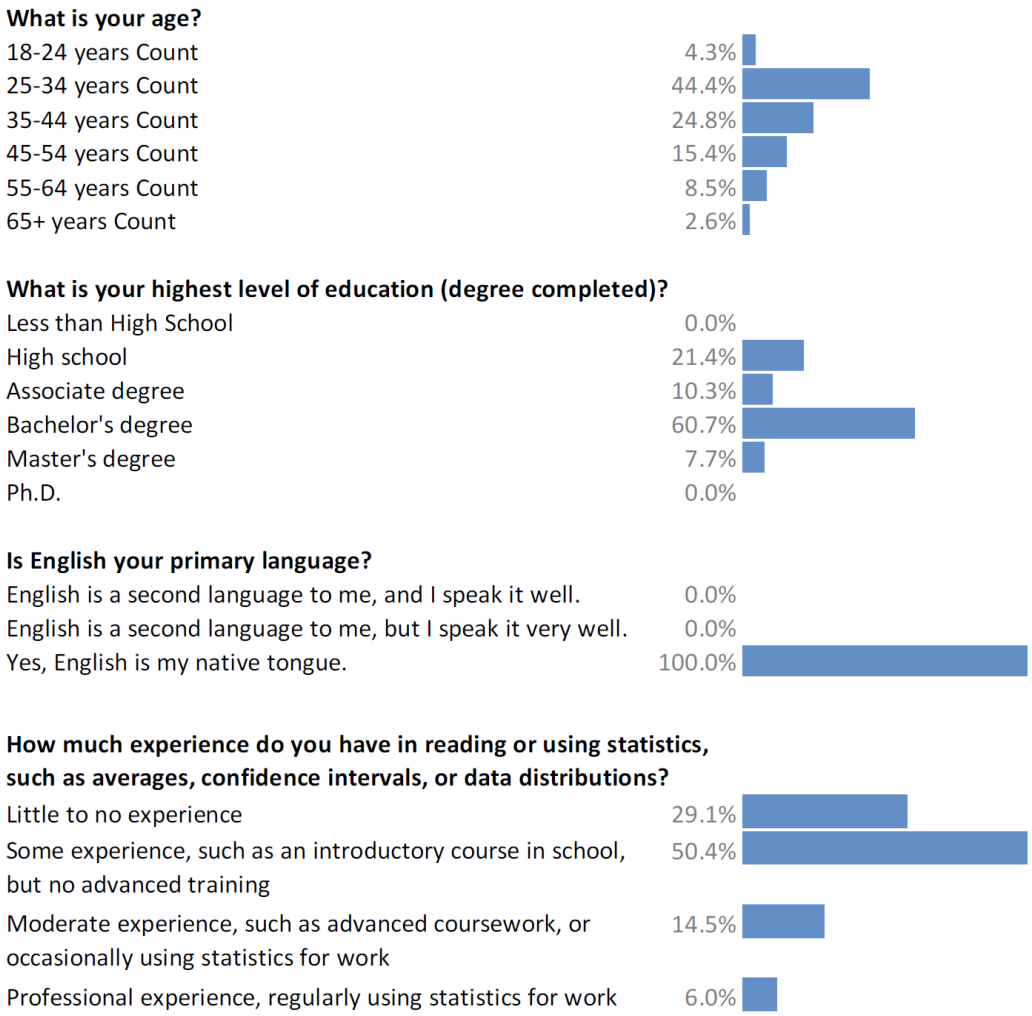
For interactions between **data size and visualization**, absolute mean error appears to decrease with larger data size for histograms. Strip plots and boxplots, however, show little to no effect of sample size; box plots might even show evidence of worse performance at larger sample sizes, though the effect is so small compared to the confidence intervals that it may be illusory. As for absolute standard deviation error, errors appear to decrease with higher data sizes for most visualization types we tested. However, strip plots exhibit different behavior: respondent estimates of

standard deviation appear to worsen with increased sample size (perhaps due to data occlusion), yet at both sample size levels, exhibit consistently lower error than the other techniques.

For interactions between **spread and visualization**, the absolute mean error appears to increase with higher spread for both histograms; however, this effect does not persist for boxplots or strip plots, and may even reverse. 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.

3.4.4 Demographics and Participant Feedback

Figure 3.13 Demographic survey of participants



The demographic characteristics of our respondents (Figure 3.13) are more or less consistent with prior attempts at understanding the Amazon Mechanical Turk population (Ross 2010). At 60.7%, the proportion of people with a Bachelor's degree is nearly twice the national average (37.5% in 2020, Census Bureau -- https://www.census.gov/data/tables/2020/demo/educational-attainment/cps-detailed-tables.html), which point to a more affluent, more educated population than the general public. However, very few respondents (8) had professional statistics

experience, and only about 1 in 7 had moderate statistics experience. This suggests that our results should be at least somewhat generalizable to the broader population, assuming a general education may be less influential on task performance than specific statistical training.

Thirty-seven of 117 participants provided free-form feedback. The feedback was generally positive, with several participants noting that the task was fun and engaging.

Participant P21812 wrote, *"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.

3.5 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 possibly misunderstood the curve fitting task. We classified such misunderstanding when at least one full block of 12 trials (out of 48) 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 inaccurate—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.

3.6 Discussion

This study gave rise to several interesting findings while confirming some basic intuitions about this task.

3.6.1 Reviewing Hypotheses

We find strong evidence in our results that the **estimation of means is more accurate than estimation of spread**. Overall, estimates of means were less biased, and more consistent across visualization types, while estimates of standard deviations where typically biased. This supports our first hypothesis.

Furthermore, we see **no evidence that the errors in estimating means or standard deviation decrease as the data size increases**. Thus the second hypothesis is rejected.

We do find that **performance varies with visualization type**. Boxplots yielded the lowest error for estimating the mean, and strip plots the highest, while strip plots outperformed the other visualization types with regard to standard deviation estimation.

Our fourth hypothesis – which emerged out of our early work – 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, resulting in a positive bias in standard deviation estimates. 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. This pattern of positive bias (Figure 3.9) in the two histograms **supports the "umbrella effect"**, the fourth hypothesis. An inverse version of this effect may also be at play for the boxplot, as the underestimation may indicate participants are fitting the whiskers, which would exclude some data points, or even the central rectangle of the boxplot, which only contains 50% of the data.

3.6.2 Explaining the Results

Our results indicate the effect of visualization type is larger than the impact of sample size or spread. Indeed, there appear to be interactions between specific visualization types and these experimental factors, and, therefore, no consistent pattern of behavior change with either spread or sample size. 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 above, 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 for 3 of 4 visualizations as the spread S increases. This is counterintuitive, but may possibly be explained by the aforementioned "umbrella effect." This is 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.

All techniques except strip plots aggregate data, whereas strip plots merely draw a line for each exact value in the sample. Our results suggest that aggregation is beneficial; for strip plots, multiple similar values in the data sample will yield overplotting, making it difficult to see high concentrations of data. This could explain the high variation in standard deviation in Figure 3.9.

A normal curve goes even further, being a completely noiseless and idealized representation of a data sample. Hindsight being 20/20, 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. This may only be counterintuitive to us as data visualization researchers, since 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 an imperfect 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 which can precisely identify (often critical) measures of central tendency, since noise in the data may distract our visual capacity.

Finally, we believe our "umbrella effect" observations may be explained by prior work on ensemble processing (Alvarez 2011), which tells us how the visual system will perceptually average all the points in a complex scene to yield a manageable abstraction. When estimating the spread, such averaging will essentially yield a convex hull of all of the points in the visualization. In contrast, when calculating a standard deviation of a dataset, several outlier points will fall outside of the envelope of the fitted bell curve. This means that the perceptually averaged representation will be more generous than the actual calculated spread.

3.6.3 Generalizing the Results, Design implications, and Limitations

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 (Student 1908). However, at n = 50, the two curves are similar, and we expect that the challenge-to-fit presented by the extra noise 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 low aggregation (dot histograms), moderate aggregation (bar histograms), to total aggregation (boxplots). We believe that other distribution visualizations, 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, in application, it may be that other graphics may exhibit entirely different behavior. Hypothetical outcome plots (HOPs) (Hullman 2015), for example, use animation to convey uncertainty, which may reduce the observed umbrella effect.

The overarching vision motivating this work is to lower the threshold of using real statistical methods so that people with little mathematical or statistical background might, with a reduced training burden, use them. Literature in perception and visualization (e.g., Albers 2014, Gleicher 2013, Correll 2014) support this goal, at least indirectly, by showing how the appropriate visual representations can enable people to manually derive even sophisticated aggregated statistics.

This experiment suggests that people with minimal training can gain statistically useful information from visual displays. Moreover, this result points to the presence of visual intuitions that allow users to see the connection between real

data samples, and the idealized curves of inferential statistics. Even if these intuitions are biased, they are still there and can be designed around using results from this experiment, along with others.

Limitations

Even though results presented here represent the second incarnation of this study, the experiment still has some weaknesses. We made several deviations from our preregistration, and a few of our analyses were also not explicitly detailed in the preregistration.

Training trials primed participants to answer our questions by showing a prefitted idealized curve on top of a data distribution for each visualization type. It is conceivable that such training teaches people less about fitting curves to data than fitting curves to a specific visualization. If this proves to be the case, subsequent experiments which make the assumptions of this one might yield null results. 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 (Correll 2019), 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.

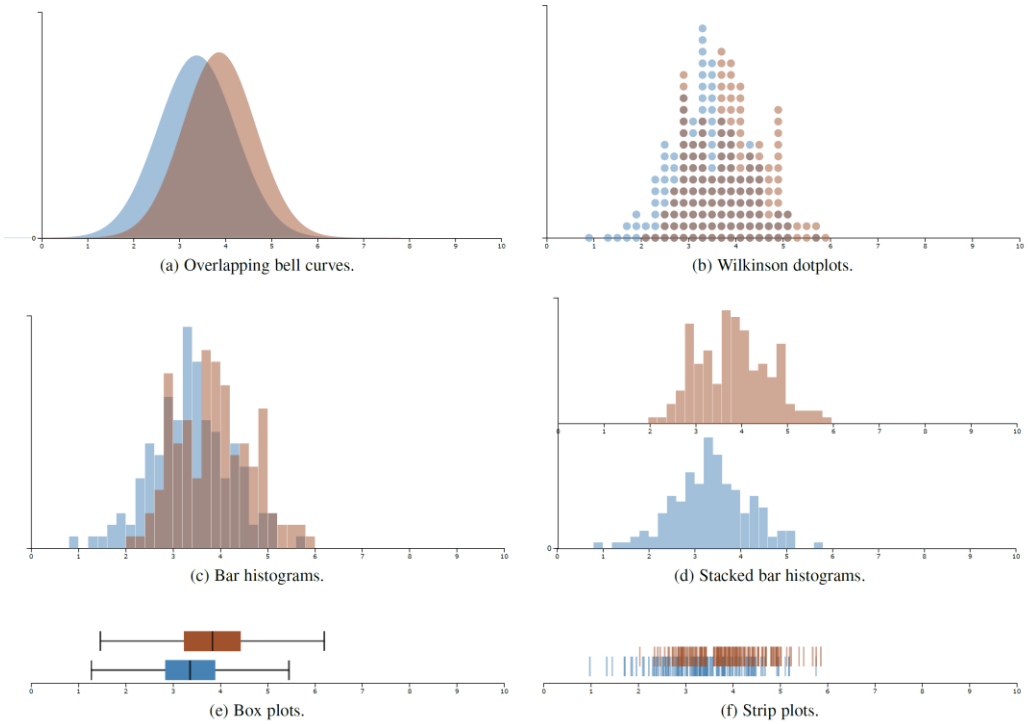
Chapter 4: Testing visual analogues to a T-test

Having demonstrated that people possess visual intuitions (however biased) for connecting idealized curves and real life data samples, testing a visual analogue to a traditional statistical test was the next step toward investigating the possibilities of visual confirmatory analysis.

4.1 Study: Overlapping Bell Curves

The fundamental research question tackled here is how well people can determine whether two data samples are drawn from the same or different populations, and whether their ability depends upon the visual representation used. To answer this question, we conducted a crowdsourced experiment where participants were asked to make a decision about whether two samples visualized using idealized bell curves, boxplots, overlapping bar histograms, stacked bar histograms, Wilkinson dotplots, or strip plots (Fig 4.1), represented a statistically significant difference (T-test at the $p = 0.05$ level).

Figure 4.1 T-tests built with six different graphic forms



Since statistical significance is heavily dependent on the size of each sample, we grouped trials by the sample size of test datasets, and included an instruction page about that specific sample size, followed by two examples, one significantly different and one not. Where possible, trials included visual representations to convey the scale of the underlying data, whether graphically, as in the case of strip plots or Wilkinson dotplots, or as a labels on the vertical axis for histograms and bell curves. Boxplots included no sample size indicators on specific trials.

We preregistered this study on OSF (https://osf.io/r3jpn/). Below we review our methods, followed by our results in the next section.

4.1.1 Participants

We designed this study to engage members of the general population in graphical inference tasks. For reasons similar to those described in the last chapter, we opted to conduct our study using Amazon Mechanical Turk (MTurk).

We originally planned to recruit a total of 500 participants (Turkers), from each of whom we would solicit answers to a bare minimum of trials. However, our approach evolved as we chose Turk "Master Workers" to help ensure high quality responses; we reduced the number of respondents, increasing the number of trials assigned to each. In the end, we recruited a total of 212 responders, limiting participation of Turkers to those with a proven track record on the site, that is, workers with 50 or more prior tasks done for other employers on the site, and with a 95 % acceptance rate for their work. We also limited participation to people from the United States due to tax and compensation restrictions imposed by our IRB. We screened participants for standard color vision (to not be color-blind, self- reported), as color perception might impact results given the experimental apparatus.

Data collection for the six visualization types took place through six separate surveys, each identical in question order and instructions, varying only in the visualization type presented. Individuals were prevented from participating in the same survey multiple times, however, it was possible for an individual to complete surveys for more than one visualization type. For the purpose of these analyses, we treat such cases as separate responses. All participants were ethically compensated at a rate chosen to be consistent with an hourly wage of at least $15/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 545 seconds, this yielded an hourly wage of approximately $16.50/hour.

4.1.2 Experimental apparatus and task

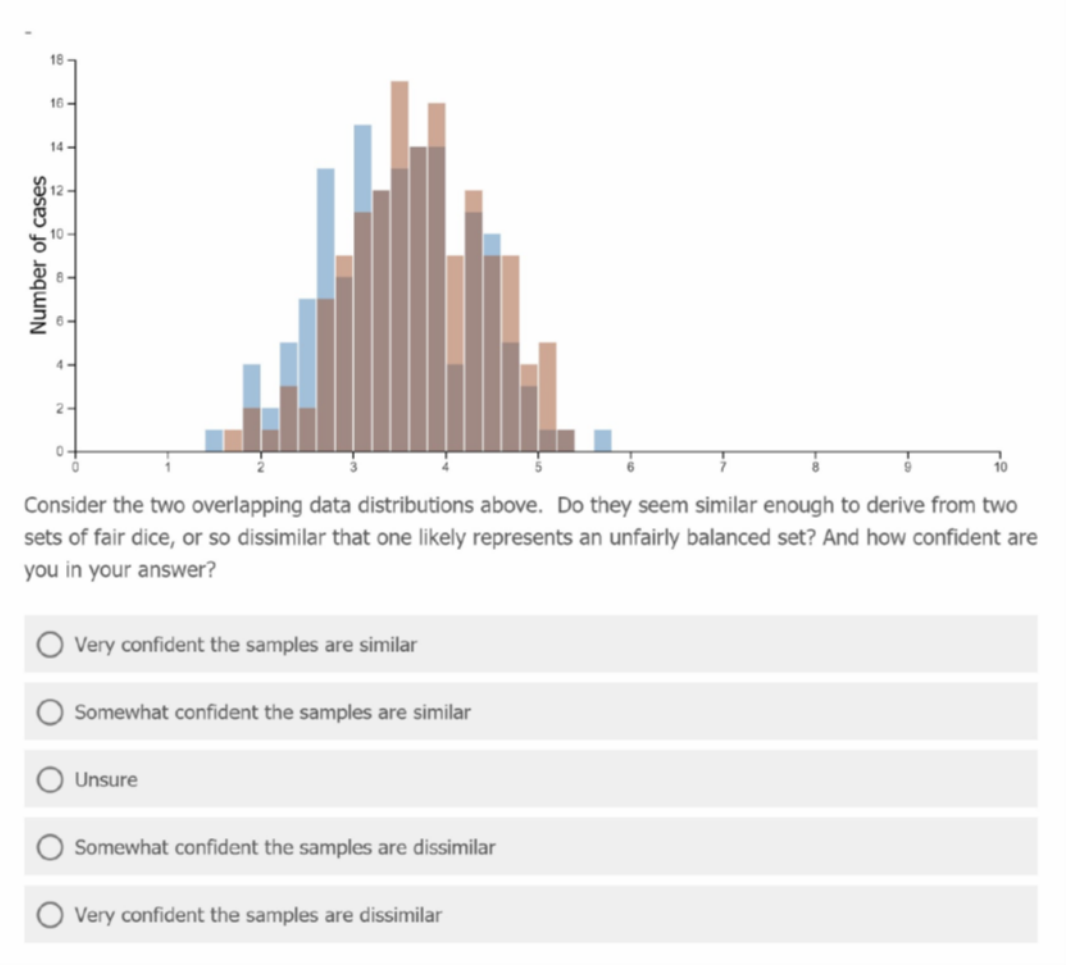
The study was distributed remotely through the user's web browser. This also meant that we were not able to control the specific computer equipment that the participants used. We required all devices to be personal computers (laptop or desktop); mobile devices were disallowed due to the limited screen space available.

Our study consisted of a sequence of trials involving a single task: determining whether two data samples visualized in a non-interactive chart in the user's browser represented the same or different source populations. The participants were given detailed instructions prior to beginning these trials which put the task in terms of comparing different sets of dice, one known to be fair and another uncertain. In each trial, they were shown the chart as well as the following prompt (Figure 4.2):

> *Consider the two overlapping data distributions above. Do they seem similar enough to derive from two sets of fair dice, or so dissimilar that one likely represents an unfairly balanced set? And how confident are you in your answer?*

They were provided with five potential answers, ranging from, "Very confident the samples are similar" to "Very confident the samples are dissimilar." The testing platform was implemented in JavaScript using D3 (Morgan 1991) and embedded into a Qualtrics survey accessed using the participant's web browser.

Fig. 4.2: Example of a single trial in our crowdsourced experiment. In this example, the sample size S is 144 cases and the visual representation V is bar histograms. The blue bars represent sample A and the tan bars represent sample B; the brown color is their intersection. The vertical axis conveys the scale of each data sample.



We generated a collection of 1,800 pairs of data samples, 600 pairs per each sample size (see below). All datasets were drawn from an essentially normal distribution, since our focus was on the type of normally distributed data that are typical for parametric statistical testing. Data were generated through a stochastic process. The different pairs had the same number of items; 36, 144, or 1,000. This also creates distributions with some variation in noise level and occasional irregular

features. To generate interesting pairs of samples, where the samples in each pair are referred to as A and B, we added a small constant to raise the mean for sample B such that approximately 50% of the pairs were significantly different at the p = 0.05 level using a t-test. We used the t-test because for samples of n = 36 there is small difference in the critical value compared to a Z statistic, which might impact scores for borderline cases. In addition to the actual data, we calculated the mean, standard deviation, and t-statistic for each of the pairs of samples.

We opted not to manipulate the standard deviation for the samples in each pair; the standard deviation ranges from approximately 0.6 to 0.8 for both samples in a pair. Thus, we left investigation of the impact of standard deviation on user performance in assessing significant differences for future work.

4.1.3 Experimental Factors and Design

We modeled two factors in our experiment:

Sample Size (S): The number of items in the two samples being visualized. As the number of items increases, the samples will begin to approach the idealized distribution. We chose three levels: 36, 144, and 1,000 items. The first two levels represent typical dataset sizes that the general population may encounter in their daily life, whereas the third represents a large dataset where only small changes in the distribution will typically yield statistically significant differences. Also, since the 'statistical power' of a data sample is approximately equal to the square root of sample size, samples of 36 are about half the power of samples of 144, and 1/5th Samples of 1,000.

Visualization (V): The visualization type used to represent the two data samples. Based on our review of the literature, we chose six distinctive visualization techniques (Figure 1):

– **Overlapping bell curves**: Two superimposed continuous filled-area charts visualizing fitted normal distributions of the underlying data samples (Figure 1a).

– **Wilkinson dot plots (dot histograms)**: Dot histograms (Ariely 1999) are unit visualization (Melcher 1999) versions of histograms where individual dots (circles) are stacked to represent each bin (Figure 1b).

– **Bar histogram:** Two "classic" histogram where the aggregated number of data items for each bin is represented using a bar of uniform width, both drawn in the same visual space so that they overlapped (Figure 1c).

– **Stacked bar charts:** Two "classic" bar histograms as above, juxtaposed one over another with no overlap (stacked), and with each chart receiving half of the available vertical display space (Figure 1d).

– **Boxplot:** The box-and-whisker plot as pioneered by John W. Tukey (Alvarez 2011, Tukey 1977) (Figure 1e). The central rectangle contains the middle half of the data (from the 25th to the 75th percentile), the median (50th 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 10th and 90th percentiles). We did not visualize outlier data in our representation.

– **Strip plot:** A unit visualization (Melcher 1999 where each item is drawn as a short vertical line with opacity on the horizontal axes, i.e., with no vertical data encoding (Figure 1f).

Since our task requires visualizing two samples (A and B) to allow comparisons, we opted to draw both overlapping bell curves, histograms, and dotplots in the same visual space using steelblue and sienna colors at 50% transparency. This gives rise to a special overlapping color (brown in Figures 1 and 2) when visual marks for the samples overlap. For strip plots, we also used 50% transparency, but overlapped the two plots only halfway, preventing overplotting (Figure 1f). Finally, for boxplots, we separated the two plots entirely (Figure 1e).

The number of bins is a significant parameter for histograms (Gleicher 2013). 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).

Experimental Design

We used a mixed design, where each participant saw all data sizes but only one of the six available visualizations. This allowed us to minimize the amount of training which would otherwise be required to instruct participants in the use of each visualization type. The small total number of conditions enabled us to keep sessions shorter than 10 minutes in duration to minimize fatigue and maximize attention for crowd workers. Each trial pulled at random one of the pre-computed dataset pairs.

Trials for each dataset size were grouped to provide respondents the maximum opportunity to learn during the experiment. They were presented with 10 trials of sample size 36 first, then 10 of 144, then 10 of 1000. Within each group, there were presented with two training trials, one which showed an example of a significant difference, the other not. It yielded the following design:

	3	**Sample Size** *S* (36, 44, 1000 samples)
×	1	**Visualizations** *V* (bell, stacked, bar, dot, box, strip)
×	10	repetitions + 2 training repetitions
	30	live trials per participant, plus 6 training trials

We ended up with 6,360 trials. For each trial, we captured the correctness. Correctness was defined as whether (1.0) or not (0.0) the participant correctly assessed a set of samples to be significantly different or not (based on a t-test at $p = 0.05$). We also captured statistics of the actual datasets participants saw, such as the p-value of the corresponding t-statistic for the two samples. A short set of optional text questions followed the trials, as well as a few demographic questions about respondents.

We recruited participants through 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. Failure to give consent terminated the experiment. Participants were asked to confirm that

they had no color vision deficiency. Participants were allowed to 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 before starting the session.

After each trial, the participant was given the correct answer after deciding whether or not a specific pair of samples were drawn from different populations or not. After all trials, participants were asked demographic questions about their age, education level, and knowledge of statistical concepts.

Each individual trial started with the display of the two samples and ended when participants clicked the one of the five answer buttons. A progress bar at the top of the screen showed the study progress. Amazon provides the option of rejecting specific respondents who failed to fully participate in a task, for example, by repeatedly clicking the same answer button to rush through the trials. However, variability checks after collection showed no such cases in the Master Turker population we drew from.

Typical sessions lasted between 5 and 6 minutes in duration. A few participants used significantly longer to complete their sessions, but our logs indicate that these participants took significant breaks between trials (presumably due to real-world interruptions). We believe that the effective time spent on the experiment was no more than 10 minutes.

4.2 Hypotheses

We preregistered the following hypotheses about our experiment:

High-fidelity vs. low-fidelity visualization. Participants will be more accurate at assessing significant differences ($p = 0.05$) between samples when using a high-fidelity visualization than when using a lower-fidelity visualization.

Bars and dots. Bar histograms and Wilkinson dot plots have higher fidelity than all other representations tested, and will thus yield higher accuracy than all those other representations.

Strip plots. Strip plots are high-fidelity visualizations that will yield better accuracy than boxplots and idealized bell curves, but occlusion will result in lower accuracy than for bar histograms and dotplots, especially for large sample sizes.

Boxplots. Boxplots are an intermediate-fidelity visualization that will yield better accuracy than idealized bell curves.

Individual differences. There will be non-uniformity in performance across individual participants, that is, we anticipate that there will exist cohorts of participants with qualitatively different patterns of performance.

4.3 Results

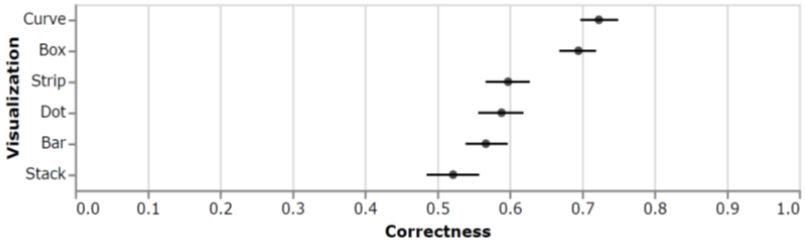
We analyzed the results primarily by computing correctness scores within analytic groups (based on Visualization Type, Sample Size of datasets, and P-value of the specific dataset pairs from each trial). We judged a trial correct if its associated p value was < .05 and the respondent selected one of the two “dissimilar” categorical answers, or if the p value was >=.05 and the respondent selected one of the two

"similar" answers. We then used bootstrapping [25] (N = 1, 000 repetitions) on aggregated trials to compute 95% confidence intervals. We plotted these confidence intervals and used graphical inference to compare the different conditions.

4.3.1 Overall Correctness Analysis

As any given trial produces a dichotomous result, the 50% cutoff is a useful comparison in the following to assess a visualization technique. The nearer to 50% correctness a technique produces in respondents, the nearer the results for that technique come to those we could expect through random guessing (and so the utility of that technique for statistical testing is diminished).

Fig 4.3 Correctness by visualization type

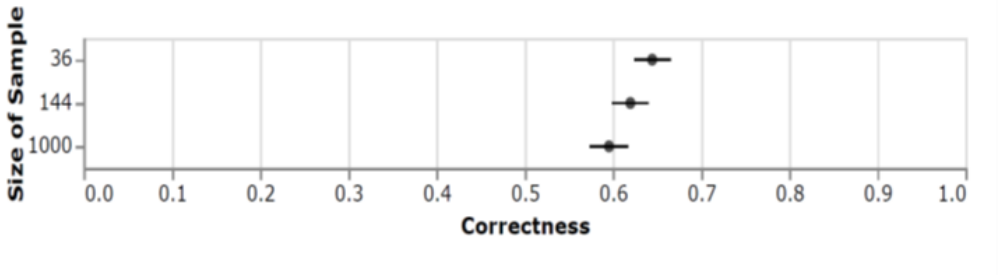


Beginning with visualization type (Figure 4.3), we find that respondents did appreciably better with the idealized bell Curve plots and Boxplots than with any other. Respondents were correct about 70% of the time with boxplots, and a bit more with idealized bell Curves. This not only fails to support our first hypothesis, it appears to directly contradict it. These low-fidelity visualizations appear to allow respondents the best performance on this task, with the highest performance appearing in the lowest-fidelity visualization (the idealized bell Curves). Dot plots, Bar histograms, and Stacked bar histograms showed the worst performance, with Stacked being essentially indistinguishable from chance. Strip plots performed worse

than Normal curve and Box plots, but at least as well as Dot plots & Bar histograms, and better than Stacked bar histograms.

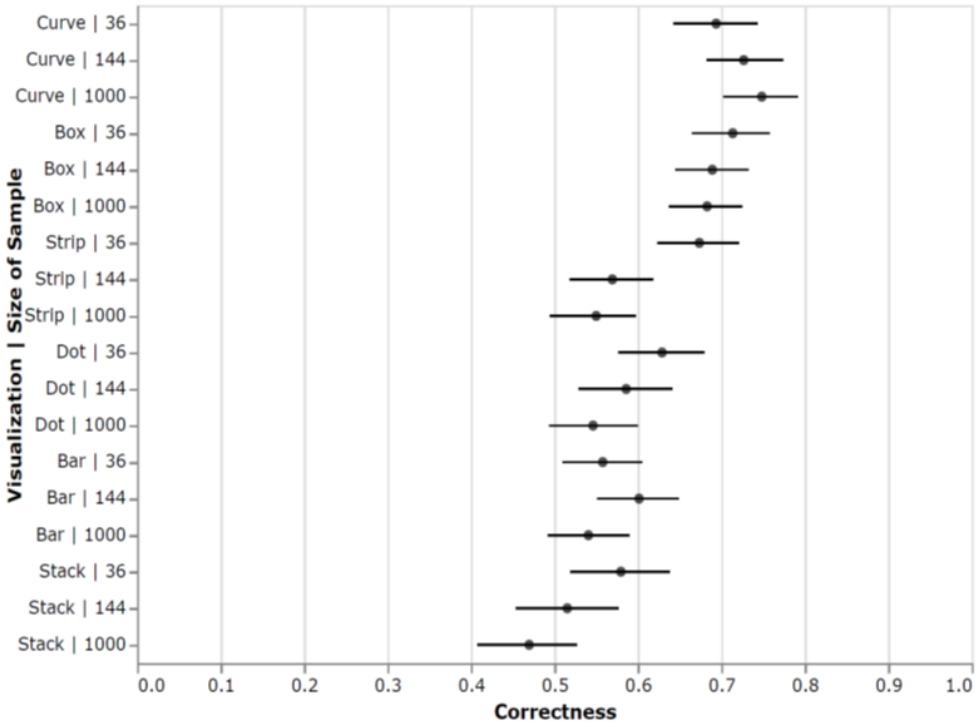
We find that the size of datasets being compared had only a small impact on average correctness when considered as a stand-alone factor (Figure 4.4), certainly much smaller than the differences across visualization type.

Fig 4.4 Correctness by visualization type



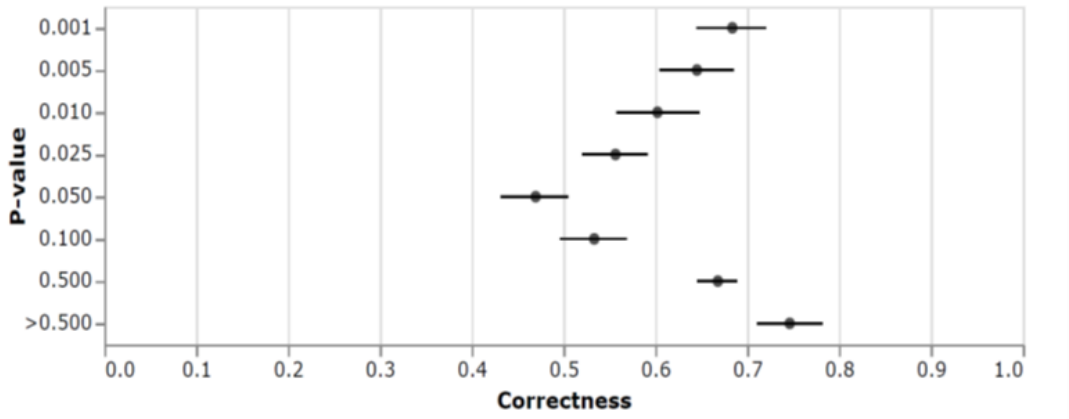
There may be some degree of interaction between the size of compared datasets and specific visualization type (Figure 4.5). Normal bell curves may have a tendency to perform better at higher sample sizes. In comparison, Strip plots and Dotplots did worse with higher dataset sizes; Strip plots in particular appear to decline in effectiveness for this kind of comparison once the smallest sample size is exceeded. This supports one part of our hypothesis regarding Strip plots, namely, that at higher dataset sizes occlusion might become a problem. However, Dotplots may also suffer from occlusion. The Stacked bar histograms also appear to decline in effectiveness as sample sizes increase.

Fig. 4.5 Correctness by visualization by sample size.



We also examined the impact of the degree of difference between particular datasets respondents had to compare, expressed as p-values (Figure 4.6). This should serve as a sort of measure of difficulty of the individual trial, where extremely low p-values will indicate dataset pairs with large mean differences, while p-values closer to 1 should indicate datasets with nearly total overlap.

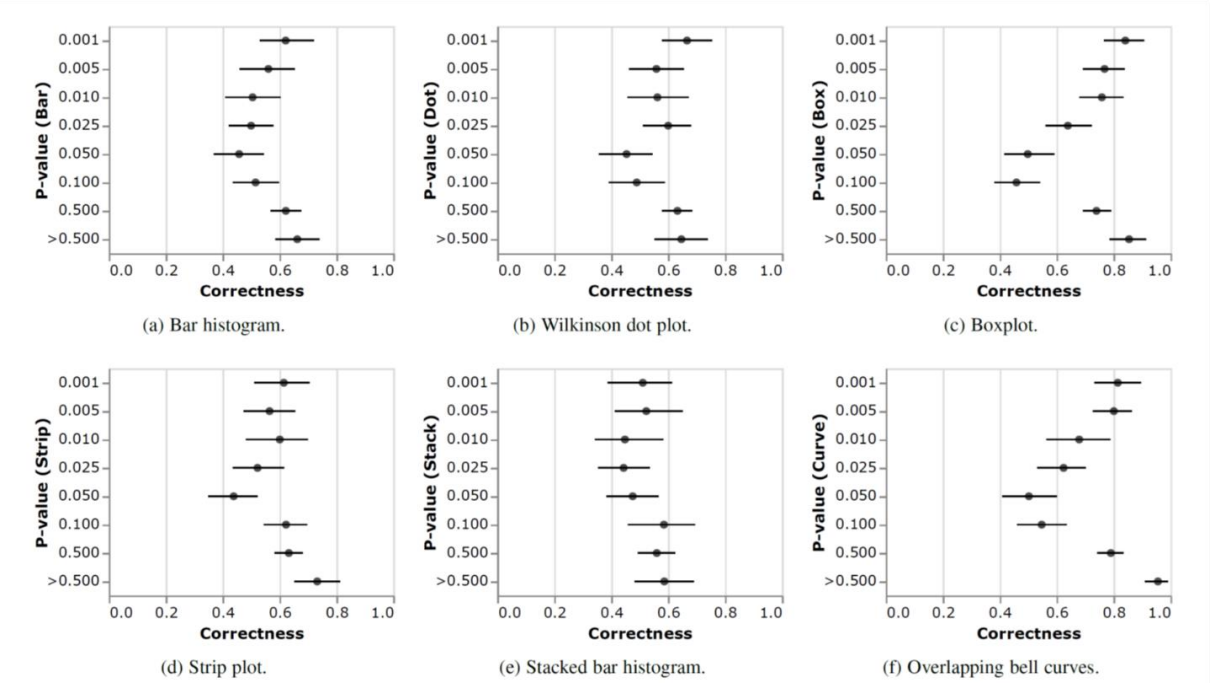
Fig 4.6 Correctness by P-value



Indeed, there are large differences in correctness by p-value. Respondents were far more likely to correctly identify datasets as dissimilar when the p-values were lowest (less than 0.001). At the high end (p-values between .500 and 1), respondents were even more likely to make a correct choice (that datasets were "similar"). Trials with more intermediate p-values, those between p = .250 and p = .500, gave respondents the most trouble; they did slightly worse than random chance, misidentifying dataset pairs in this p-value range as "similar." Dataset pairs with p-

values that approached significance (0.050 < p < .010) also presented a challenge, with respondents doing little better than chance on these.

Fig 4.7 Correctness by visualization type and P-value



(a) Bar histogram. (b) Wilkinson dot plot. (c) Boxplot.

(d) Strip plot. (e) Stacked bar histogram. (f) Overlapping bell curves.

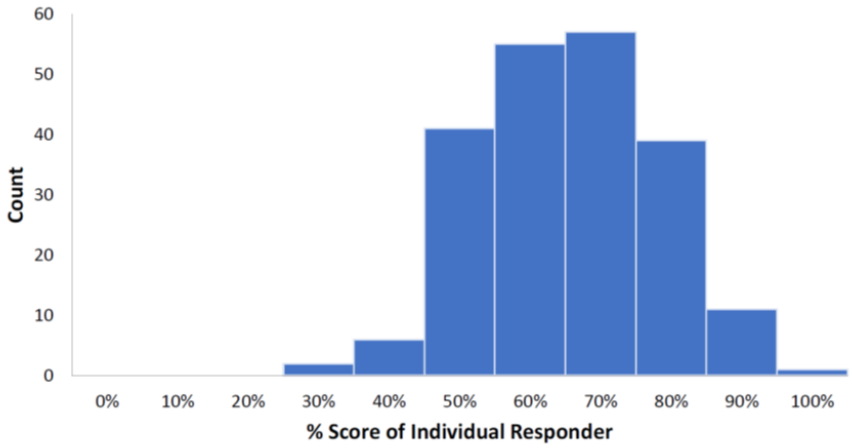
The interaction between visualization type and p-value confirms the general pattern of both individual factors, while also providing details that might be informative (Figures 4.7a–4.7f). The high average correctness respondents achieve with normal Curve and Boxplots comes from the very high percentage correctness at high and low p-values; respondents judged Curves with p-values greater than .500 correctly more than 90% of the time (Figure 7f). Yet with intermediate values using these visualizations, respondents did essentially no better than chance guessing. This in part reflects the nature of classic statistical tests, which in their simplest interpretation require a dichotomous response (significant or not), yet it also suggests a shortcoming of this approach, where even the most effective visualizations

become ineffective when faced with difficult cases. Strip plots (Figure 7d), with which respondents performed at least as well or better than the different kinds of histograms, appear to get a boost from higher scores on the high p-value trials, but on other trials show quite modest scores.

4.3.2 Individual Analysis

In our final hypothesis, we proposed that some group of individuals would show a propensity for higher performance on this task, that is, some people would have an "eye" for this kind of comparison. However, our results do not support this contention (Figure 4.8).

Fig. 4.8 Distribution of individual respondent correctness.



The scores across participants appears to be normally distributed, with no clusters or modal humps that might suggest any structure other than that which random noise can explain.

It is possible that this normal distribution also represents a gradient in innate ability, which, like height or birth weight, has variation around some norm. However,

we would require repeated tests over time on the same respondents to confirm this interpretation.

4.3.3 Demographics and Participant Feedback

Clusters in performance associated with demographic characteristics would also point to the possibility of consistent differences in performance among some group of individuals. However, there is little evidence for such differences in these data (Table 4.1).

Table 4.1 Percent correct by demographic characteristics

	Responders		Percent
	Count	%	Correct
Total	**212**	100.0%	62.0%
18-24 years	1	0.5%	NA
25-34 years	58	27.4%	
35-44 years	72	34.0%	
45-54 years	39	18.4%	
55-64 years	34	16.0%	
65+ years	8	3.8%	
Yes, English is my native tongue.	211	99.5%	NA
English is a second language to me, but I speak it very well.	1	0.5%	
High school	64	30.2%	62.1%
Associate degree	32	15.1%	60.2%
Bachelor's degree	92	43.4%	62.5%
Master's degree	20	9.4%	60.8%
Ph.D.	4	1.9%	68.3%
Little to no experience	59	27.8%	60.2%
Some experience, such as an introductory course in school, but no advanced training	100	47.2%	62.3%
Moderate experience, such as advanced coursework, or occasionally using statistics for work	50	23.6%	63.6%
Professional experience, regularly using statistics for work	3	1.4%	63.3%

Average percent correct achieved by respondents varied little by either education or prior experience with statistics (self reported).

4.4 Deviations from the Preregistration

This experiment was preregistered in September 2020, but data collection only commenced in Summer 2021 due to what can only be expressed as pandemic fatigue. We made the following deviations from the original study plans:

Added a visualization technique: After feedback from colleagues, we added stacked bar histograms to the lineup of visualizations tested, bringing it up from 5 (as named in the preregistration) to 6. The benefit of this change was to add a familiar and commonly encountered visual representation to the study, one which addressed the perhaps unfamiliar overlapping of bar histograms we also employ. We do not anticipate that this had any ill effect on the validity or results of the experiment.

Fewer participants: We had originally aimed for 500 participants, with approximately 100 per visualization type. We ended up with only approximately 35 participants per visualization type because we raised our recruitment qualification to Turk Master Workers, which were both more expensive and more difficult to recruit. However, we believe that the increased quality arising from these highly rated workers made this deviation worthwhile.

Increased compensation: Recruiting Turk Master Workers meant a necessary increase in compensation from the $1 listed in the preregistration. Again, we believe this should have no detrimental effects on the collected data.

4.5 Discussion

We address our hypotheses as follows:

- Overall, we find that what we call **"high-fidelity" visualizations**—bar histograms, dotplots, and strip plots— **yielded lower accuracy for this task** than "lower-fidelity" visualizations—idealized bell curves and boxplots. This is contrary to our hypothesis, where we postulated that the increased fidelity would yield better accuracy. **(Rejected)**
- **Bars** (overlapping and stacked) **as well as dotplots did not yield the highest accurac**y; in fact, they arguably performed the worst. **(Rejected)**
- **Strip plots performed better than expected**—certainly better than bar histograms and Wilkinsonian dotplots—but still yielded lower accuracy than bell curves and boxplots. **(Rejected)**
- Boxplots, which we name "intermediate-fidelity" visualizations, **did not yield better accuracy than idealized bell curves**; there is little evidence for any difference in accuracy between the two techniques. **(Rejected)**
- We find non-uniformity in performance across individual participants; as our individual analysis showed, there are some participants who were able to complete this task much more accurately than others. However, without additional rounds of data collection, **we are unable to confirm that this variation reflects the innate or learned ability of particular individuals rather than some other source of random variation. (Inconclusive)**

With so many of our original hypotheses rejected, these results constitute a set of (to us) new observations which we find worthy of further investigation.

4.5.1 Explaining the Results

Our results contradict or fail to support our major hypotheses, and so surprised us. Rather than confirming the utility of detailed visualizations like dot histograms and strip plots, they suggest that aggregate visualizations are more appropriate for looking for differences between sample datasets. This goes against our instincts as visualization researchers and practitioners; our bias is toward more detailed views rather than less. But should we have been so surprised?

Calculating Z or T statistics for comparing two samples requires only the mean and standard deviation of each sample, and the sample sizes. All the information going into the calculation of either statistic is aggregate—just like the normal curves from our trials. The curves are drawn by inputting a mean and standard deviation, and assuming a normal distribution. T and z tests also assume normal distributions in sampled populations. Thus, in a very real sense, the normal curves presented respondents with the most direct visual analog to the t-test we used to judge their answers. Boxplots are the next most aggregate visualization, and respondents using them performed nearly as well as those using normal curves. Boxplots provided respondents with no visual reference for estimating sample size, a vital consideration in statistical testing. We attribute our respondents' success with this form at least in part to our clustering trials by sample size, and preceding each group of trials with worked examples to give them a "feel" for the critical degree of overlap. Respondents did worse with all the less aggregated/more detailed visualization forms we tested.

We speculate that this additional detail may distract respondents from correctly identifying the critical degree of overlap, particularly when faced with borderline cases.

The poor performance of our respondents when facing border- line cases, regardless of visualization type, may suggest another possibility. We speculate that statistical significance, measured as a p-value of .05, may not align well with our intuitions for what constitutes a difference between two distributions. This would potentially present a stumbling block to using visual methods for statistical testing where borderline cases are a possibility. At the very least, it suggests that for borderline cases, some kind of visual cue should be provided which establishes the scale required difference for users.

4.5.2 Generalizing the Results

As in any crowdsourced study conducted via online tools, the participant pool sets limits on the applicability of results to the broader population. All our participants had internet access, a computer, and access to some form of electronic banking. All participants were U.S. residents, and all but one spoke English as their native language. However, participants in our study came from a broad range of age groups, education levels, and prior experience with statistics. We believe these results may have modest general application.

We tested only some of the visual methods for displaying distributions. However, the methods we tested vary in both degree of aggregation, detail, and visual complexity. Due to this variability, we believe it is possible to use our results to

gauge the likely performance of other visualization types. For example, density plots are somewhere between bar histograms and normal curves in their degree of detail and visual complexity. We speculate that their performance in tasks like those in this paper would reflect this intermediate position.

Finally, we believe the three dataset sizes we tested are representative of dataset sizes in a diversity of fields, from education, to opinion polls, to product acceptance testing. However, these results may be unhelpful to those studying "big data" visualization problems, or other fields where data sizes are typically several orders of magnitude larger. Similarly, while normally distributed data are common, many other distributions (e.g. bimodal or highly skewed), find uses in a multitude of fields. Few of the results presented here may generalize to these. Indeed, it may be that creating visual aids for inferential statistics that hinge upon more complex distributions requires the kinds of detailed representations that performed poorly in our experiment.

4.5.3 Implications for Design

Our study has direct implications for people tasked with displaying multiple overlapping data distributions, but also those seeking better distribution graphics in general. Our central finding suggests that some tasks suffer from additional detail, and this observation alone is worth investigating in light of future design efforts.

In this example, where the center and degree of spread around that center formed the primary basis for a decision, more detail impeded a good decision. However, the responsible designer will have to weigh carefully whether they are

working on such a problem, as the loss of detail prevents a user from making any other discoveries in the “extraneous” information.

Our findings also suggest that comparing distributions visually may be most reliable where differences are either large or very small. Where differences are only moderate, judging the degree of difference by eye may be a challenge. The visual designer might be able to meet that challenge with forms which magnify visual differences—a design calculation which would require further testing. However, the fact that humans may have a low ability to judge the degree of separation where p values are intermediate may itself have important implications for communicating statistical information to a general audience. Our findings about stacked bar histograms suggest that overlapping bar histograms, despite the interference the overlap can cause, are a better choice when considering two distributions. It appears that strip plots may be a good choice for displaying very small datasets, but other forms may be better when dataset size increases. It also appears that boxplots remain a powerful tool for comparing distributions. Tukey invention remains relevant.

4.6 Conclusion and Future Work

This chapter presents results from a crowdsourced evaluation investigating how well people can perform graphical inference of what essentially amounts to a Z or T-test: comparing visual representations of two data samples to determine whether they are drawn from the same or different populations. Contrary to our expectations and professional biases, we found that the more abstracted visual representations yielded more accurate user performance this task than visualizations which showed

an unaggregated version of the data: idealized bell curves and boxplots yielded better accuracy than histograms, dot histograms, and strip plots. Furthermore, we found that these abstracted representations were unaffected (or even improved) by increasing sample sizes, something which was not true for the other representations.

However, upon further reflection, we note that this is perhaps not so surprising since the t-test we use as ground truth is based on idealized representations in the first place. In other words, this study's main finding may be that graphical formulations of statistical tests can be powerful where differences in sample means are large (corresponding to P values less than .01), even to the point where they can stand in for traditional statistical tests for user populations that are not trained in inferential statistics. Overall, we view these observations as a victory for data visualization, but also a caution about the biases we hold that led to our original speculations that more detailed visualization techniques would prove superior.

We think that our study suggests a host of visual statistics work in the future: work that explores borderline cases, work that seeks to identify situations where detail helps and where it hurts, and work that explores which equation-based statistical methods may be effectively transformed into intuitive visual analogues, and which require more complex analogues that support greater degrees of embedded calculation.

The next chapter attempts to leverage these ideas with professional statisticians to uncover design principles for visual tools of statistical inference acceptable to the scientific community.

Chapter 5: Features of a Visual Inferential Statistics Tool for Novices Emerging from the Experience of Professional Statisticians

5.1 Overview

I conducted semi-structured interviews with experienced statisticians, focusing on aspects of their relationship with visualization, and how they understand statistical inference, with an emphasis on frequentist, parametric statistical tests. However, within the open-ended questions which formed part of the interview, many participants expressed their understandings of non-parametric and/or Bayesian approaches as well. Data were captured and coded according to a composite scheme which mixed a priori and emergent codes.

5.1.1 Data

How Data were Captured

All interviews were conducted via Zoom. Sixteen of 18 Interviews were recorded, both video and audio, with the remaining two captured via researcher notes only. Video allowed for screen-captures of sketches drawn by participants. Audio recordings were submitted to a transcription service (rev.com) to provide accurate text for coding. The researcher took notes during the sessions, including sketches of participant-described visualizations which the researcher then showed to the participants via the video feed for their approval of the sketches. The researcher also

took notes immediately after each session to record general impressions, and recall details not captured in the moment.

Exceptions During Data Collection

The first participant used their cell phone to enter the Zoom room, which resulted in lower fidelity transmission of graphics for the third section of the interview. All subsequent interviewees were required to use a laptop or other full-size screen device.

Two of the 18 valid interviews failed to record, and, therefore, only the researcher's notes captured the outcomes of these two sessions.

One interview skipped the graphic elicitation section of the interview, while completing the other two sections in full.

All other interviews followed the outlined collection procedure.

Participants

Criteria for Participation

Participation was limited to adults 18+ with a combination of education and work experience that would qualify them as an "experienced statistician".

Table 5.1 – Participant Characteristics

	Industry	#years work	Publications	Highest degree	Field of Degree	Significant teaching experience?	Age	Gender	Self-Described Race
1	Academia	10 to 14	75 to 99	PhD	Statistics	yes	40 to 44	male	White
2	Academia	15 to 19	150+	PhD	Biostatistics	yes	40 to 44	female	Black
3	Academia	15 to 19	10 to 24	Masters	Data Science	yes	35 to 39	female	Asian
4	Academia	15 to 19	50 to 74	PhD	Statistics	yes	45 to 49	female	Asian
5	Academia	20 to 24	150+	PhD	Public health	yes	55 to 59	female	Asian
6	Academia	20 to 24	50 to 74	PhD	Statistics	yes	50 to 54	female	Hispanic/Latin American
7	Academia	25 to 29	100 to 149	PhD	Statistics	yes	50 to 54	female	White
8	Government	10 to 14	10 to 24	Masters	Statistics	yes	35 to 39	male	White
9	Government	15 to 19	25 to 49	PhD	Statistics	no	40 to 44	female	Hispanic/Latin American
10	Government	20 to 24	10 to 24	Masters	Sociology	no	50 to 54	female	White
11$_{notes}$	Government	25 to 29	0*	Masters	Data Science	yes	60 to 64	male	White
12	Government	25 to 29	10 to 24	Masters	Statistics	yes	50 to 54	male	White
13$_{notes}$	Government	25 to 29	26 to 49	PhD	Statistics	yes	55 to 59	male	Middle Eastern/North African
14	Government	25 to 29	10 to 24	Masters	Survey methodology	no	60 to 64	female	White
15	Private industry	5 to 9	10 to 24	Masters	Data Science	no	30 to 34	male	White
16	Private industry	5 to 9	5 to 9	Masters	Survey methodology	no	25 to 29	female	Asian
17	Private industry	25 to 29	25 to 49	PhD	Statistics	yes	55 to 59	female	White
18	Private industry	35 to 39	25 to 49	PhD	Statistics	yes	60 to 64	male	White

Participant numbering is unrelated to the order interviews were taken in.
"notes" = interview recorded only through interviewer notes.
* = All interviewee's publications were strictly for internal agency use, and, therefore, not subject to peer-review processes.
Field of Degree summarized in some instances to protect the identity of interview subjects.

All participants had to have at least one degree in an applicable field. This could be an undergraduate degree, however, all recruited participants had a masters degree or more in a field focused on quantitative analysis. More than half had one or more PhDs.

All participants had to be statistical professionals with 5 or more years of experience after the completion of their education. The minimum work experience among recruited participants was 7 years post-graduation, the maximum 35, with an average of 19.7 years. Several volunteers were excluded due to this criteria.

All participants had to regularly work with methods of statistical inference, such as statistical tests or confidence intervals. One volunteer was excluded due to this criterion.

Publication counts serve as another indicator of the depth of professional experience possessed by the participants. While not required for entry into the study,

participants were asked to provide the number of professional publications they have authored or co-authored. These could be papers peer reviewed through a traditional academic process, or official government reports which must undergo a rigorous agency edit and review process. Only two had fewer than 10 publications to their names: one, the youngest participant, reported 5-9 publications, while a second participant (who had 25+ years professional experience) had only created research products for internal use within their organization. Three participants reported more than 100 research publications each; with 51 being the overall average for all participants.

Recruitment

Volunteer participants were recruited via direct email request from the researchers. Participants were offered $25 compensation in the form of a gift card.

Two thirds of participants were selected from among members of the American Statistical Association, with guidance from the organization's leadership. The selection process helped to ensure participants had the requisite professional background, and came from a variety of industries. With one exception, these participants had no prior professional contact with the study authors.

The remaining one third of participants were reached through the authors' professional networks.

5.1.2 Diversity of Participants

Having a diverse panel of participants is vital to ensuring a study of this sort generates a diversity of ideas, broadening the applicability of any findings (Lazar 2017). In the case of this study, diversity of statistical training and experience were the goals. Our recruitment efforts explicitly targeted statisticians working in three broad industries – government statistical agencies, academia, and private industry. Table 5.2 demonstrates that these efforts bore fruit.

Table 5.2 – Participant Summary

Professional experience	**Industry**	**Professional title**	**#years work**	**Publications**
	Academia - 7	Professor	Sum: 350+	Sum: 915+
	Government - 7	Biostatistician	Mean: 19.7	Mean: 51
	Private Industry - 4	Data Scientist		
		Statistician		
		Survey methodologist		
Educational Experience	**Highest degree**	**What degree?**	**Teaching?**	
	PhD - 10	Biostatistics	Yes - 13	
	Masters - 8	Data Science	No - 5	
		Public Health		
		Sociology		
		Statistics		
		Survey Methodology		
Demography	**Age**	**Gender**	**Race**	
	25 to 34 2	male - 7	White - 10	
	35 to 44 5	Female - 11	Hispanic or Latin American - 2	
	45 to 54 5		Black - 1	
	55 to 64 6		Asian - 4	
			Middle Eastern/North African - 1	

Seven of the participants worked in government, 7 in academia, and 4 in private industry. Ten participants had a PhD, while 8 had founded their careers upon a masters degree. Degrees were in a variety of fields – statistics, of course, but also Biostatistics, Sociology, Public Health, and others. Similarly, the job titles of

participants varied (see table). Years of work experience varied somewhat, as discussed above. Five of the participants had no significant teaching experience, while 13 had taught at least one statistics course of a semester in length (or an equivalent outside an academic setting).

Participants varied demographically as well. While they tended to be older (a natural consequence of targeting people with a lengthy professional experience) they varied in age, with two as young as 25-34, and the remainder nearly evenly distributed across the next three 10-year age groups, up to 55 to 64. Seven participants identified as male, eleven as female. Ten participants identified as white, 2 as Hispanic or Latin American, 4 as Asian, 1 as Black, 1 as North African/Middle Eastern.

5.1.3 Positionality Statement:

With more than three decades of professional statistical experience across government, academia, and private industry, I am aware of my positionality as an insider to the community being studied. I believe this insider status may provide me with an informative perspective, as well as unique access to the community. But at the same time, it may bias the data I collect, and influence my interpretation of results. As the lead researcher, I have done my best to keep this awareness in the forefront during all research phases, and have implemented strategies to mitigate the impact of these potential biases.

During participant recruitment, I worked to draw from a mix of analytic communities to ensure the best possible diversity of analytic viewpoints, but also to

reach statisticians with experiences separate from my own. For example, while several participants were former colleagues of mine at the Census Bureau, most have moved on to other Federal agencies with which I have at most a consulting relationship. Among the academic statisticians in this study, all have far more depth of experience in that world than I can claim. Some of the participants from private industry have experience with high tech manufacturing, which is very different from the work I have done during my years in private industry.

While creating the collection instrument, I mixed three different collection modes (open ended questions, graphic elicitation, observations on strawmen designs). This diversity of data-types hopefully gives the collection some resilience against potential biases. For example, open ended questions expanded the data collection on the front end, giving respondents space to offer potentially surprising answers which might exceed limits otherwise imposed by possibly biased or leading shorter-answer questions. More structured parts of the collection may be less subject to bias in coding, the back end of the data collection process.

Similarly, I used a combination of open and closed codes. Closed coding required me to take a disciplined approach to some of the results, with the potential to directly falsify my initial hypotheses. While open coding encouraged me to think beyond these initial hypotheses.

In conducting interviews, I endeavored to create an environment which would provide participants the comfort of feeling they were having a conversation with an interested and supportive colleague: a colleague, to create a space in which they could openly discuss the most technical aspects of their work without fear of alienating their

listener; and a supportive one to create a safe space to discuss their private thoughts about their work. There is evidence this environment succeeded, as more than one participant expressed their relief that the study's anonymity precautions would ensure none of their employers would know what they had been saying.

Throughout this study, I have actively sought alternative interpretations to my own, as well as seeking disconfirming evidence to challenge my findings. My goal has been to foment a genuine discussion with each of the research subjects, and to do my best to hear what they have said, rather than merely what I was listening for.

One additional goal may be relevant, namely, my intention to find a position as a professional track lecturer upon completion of my degree. This makes me especially interested in the areas of overlap between this research effort and statistical pedagogy. Creating visual aids for teaching statistics to novice analysts and making visual tools for use by novice analysts both require the embedding of an experienced statistician's understanding within the visualization. However, a tool designed for teaching and one designed for use in the field may have important differences in features. Thus, the results from this design study should be verified through practical field tests.

5.2 Development of Interview Procedure

Prior to data collection, the researchers conducted two practice interviews using early versions of the script, and a sketch-version of the first strawman graphic. Scripts and strawman #1 were refined based upon feedback during these practice runs, and a second strawman graphic was added.

5.2.1 Interview Script

Interviews followed a script (Table 5.3 pt. 1&2). Questions in **bold** were asked word for word of all participants, optional follow-up questions are indented.

Table 5.3 – Interview Script pt 1

Minutes	Section	#	Script
1	Intro and permisssions	1	*"I am Researcher Name, conducting research into experts' understanding of statistical methods. I would like to record this interview. All your responses will be anonymized. Any voice recordings or images captured during this interview process will be held in strict confidence, available only to the researchers, and will be erased upon completion of the research. Portions of interview transcripts, once stripped of all personally identifiable information, may be made available as part of the publication process. Do I have your permission to record this interview?"*
4	Intake questions	2	**Do you work (mostly) in Academia, government, or private industry?**
		3	**What is your professional title?**
		4	**What is your highest relevant degree?** (in what field?)
		5	**Approximately how many publications in peer-reviewed journals, or other professional venues, have you written that included quantitative analyses?** (none, 1-4, 5 - 10, 11-20, 20+)
		6	**Do you regularly work with analyses which require statistical testing, confidence intervals, or similar analytic tools?**
		7	**Do you have teaching experience related to data analysis** (such as statistics or econometrics)**?**
		8	**What is your age** (18-24, 25-29, 30-34, 35-39, 40-44, 45-49,50-54,55-59,60-64, 65+)**?**
		9	**What is your prefered gender identification?**
		10	**What is your prefered racial identification?**
25	Analytic Process	11	**Can you describe how you approach a typical analytic problem in your work?**
		12	How do you approach a new dataset?
		13	What steps do you take?
		14	Can you tell me what that step accomplishes?
		15	What does that look like?
		16	Do you have a mental image of step 1 (2, etc.)
		17	Can you draw a picture of that?
		18	How would you describe this to someone you needed to train to do your job?
		19	Can you draw a picture of that?
		20	**Do you regularly use visualization methods during the analytic phase of your work?**
		21	For your own analytic purposes?
		22	**Which viz do you use?**
		23	What does such a viz do for you?
5	Graphic Elicitation	24	**When it comes to using statstical testing in your work, how do you think about them?**
		25	What steps do you take?
		26	Can you tell me what that step accomplishes?
		27	What does that look like?
		23	**Can you draw a picture of that?** (if expresses inability to draw: "How would you describe this to someone you needed to train to do your job?")
		24	**What does a stat test mean to you?**

Table 5.3 – Interview Script pt 2

20	Strawmen graphics	25 *(Present sketches to respondent using 'share-screen' feature)* 26 Strawman Graphic #1 - **Are you comfortable based upon what you see on the screen that orange and blue represent different sub-populations at the 1:10 level, or might they come from the same population?** 27 Strawman Graphic #1 - **Are you comfortable based upon what you see on the screen that orange and blue represent different sub-populations at the 1:20 level, or might they come from the same population?** 28 Strawman Graphic #1 - **Are you comfortable based upon what you see on the screen that orange and blue represent different sub-populations at the 1:100 level, or might they come from the same population?** 29 Strawman Graphic #1 - **Are you comfortable based upon what you see on the screen that orange and blue represent different sub-populations at the 1:1000 level, or might they come from the same population?** 30 Strawman Graphic #2 - **Are you comfortable based upon what you see on the screen that orange and blue represent different sub-populations at the 1:10 level, or might they come from the same population?** 31 Strawman Graphic #2 - **Are you comfortable based upon what you see on the screen that orange and blue represent different sub-populations at the 1:20 level, or might they come from the same population?** 32 Strawman Graphic #2 - **Are you comfortable based upon what you see on the screen that orange and blue represent different sub-populations at the 1:100 level, or might they come from the same population?** 33 Strawman Graphic #2 - **Are you comfortable based upon what you see on the screen that orange and blue represent different sub-populations at the 1:1000 level, or might they come from the same population?** 34 Strawman Graphic #3 - **Are you comfortable based upon what you see on the screen that orange and blue represent different sub-populations at the 1:10 level, or might they come from the same population?** 35 Strawman Graphic #3 - **Are you comfortable based upon what you see on the screen that orange and blue represent different sub-populations at the 1:20 level, or might they come from the same population?** 36 Strawman Graphic #3 - **Are you comfortable based upon what you see on the screen that orange and blue represent different sub-populations at the 1:100 level, or might they come from the same population?** 37 Strawman Graphic #3 - **Are you comfortable based upon what you see on the screen that orange and blue represent different sub-populations at the 1:1000 level, or might they come from the same population?** 38 **Which of these graphics do you prefer, if any?** 39 **Do these sketches match your image of a statsitical test?** 40 **How do they differ?** 41 **What is present that misrepresents your understanding?** 42 **What is missing?**
5	Thank you	43 **Thank you for your time today. I am going to turn off the camera in a minute, to give you a chance to aske me anytying you like without being recorded. But do you any questions for me right now?** 44 **Is there a question I should have asked which I didn't?** 45 Is there something you have been waiting to tell me when I asked about it, but I never did? 46 **I will now end recording.**

5.2.2 Data Collection Procedures

Pre-interview

Participants were asked to fill out a consent form before their interview. This included a small number of demographic and qualifying questions. Participants were told to expect an hour-long interview, at a time convenient to their schedule.

Before recording began, participants were given the opportunity to ask any questions they had. Some had questions about the purpose of the research, and these were answered with the long term goal of the primary author's research program, that is, the creation of statistical tools for novice analysts. With these preliminaries cleared, recording began.

Recorded Interview

Re-affirming Consent

Participants were asked on-camera for their verbal consent, and taken through a demographic and qualifying survey similar to what they had already filled out online, though with some additional details (*Introduction and Consent, Intake Questions*). These early questions provided time to establish a rapport, while also making doubly sure of participant consent.

Analytic Process

The first substantive questions asked participants to describe the steps of their analytic process. This framed subsequent discussion as focused on their day-today work process. Probes into these processes sought to capture how visualization played

a part (or didn't) in the various work phases. Additional probes sought to uncover the statisticians' tacit understanding of their research processes.

Graphic Elicitation

The next section used any references to inferential statistics from the prior interview section as a bridge to turn the conversation toward participants' understanding of inferential statistics in general, and t-tests in particular. Participants were asked whether they had a picture in their heads of what a two-mean t-test looked like. They were then asked to draw a picture of that image using paper and pencil on hand, and walk the interviewer through their sketch.

Strawmen Graphics

During next section, the interviewer presented participants with graphic versions of a t-test constructed by the researchers as "design probes" (Gaver 1999, Wallace 2013) to elicit feedback from participants. Participants were walked through each version (2 versions for the first three interviews, with strawman #3 added to the rest based upon the early results).

All strawmen were based upon a single pair of samples, denoted "orange" and "blue". Samples were generated through processes which would tend to create approximately normal distributions (the summing of multiple random variables in excel). The samples each had an n of 36, a standard deviation of approximately .2821, and means which differed by just enough to meet statistical significance according to a two tailed t-test ($p=.0498$). Choosing a borderline case was meant to

stimulate conversation. For this same reason, the researchers selected one sample (orange) which appears to deviate from normality (it is somewhat bi-modal).

Observing Participant Performance Using Strawmen Graphics

Within this section of the interview, participants were asked to use a series of graphic t-tests

Participants were first asked to use the graphics to answer whether the displayed example met a selected significance level (p = [0.1, 0.05, 0.01, 0.001] for strawman #1 and #2, only p=.05 for strawman #3.). The same underlying data were displayed with all three graphics, so that differences in response would reflect the graphic rather than a difference in data samples. The three graphics were always presented in the same order, and this may have also had an impact. After capturing participant performance in using each graphic, 'correct' answers were provided before moving onto the next strawman. Afterward, participants answered a series of questions designed to probe for the basis of their understanding or initial misunderstanding, what was missing or wrong about the graphics, or other design recommendations.

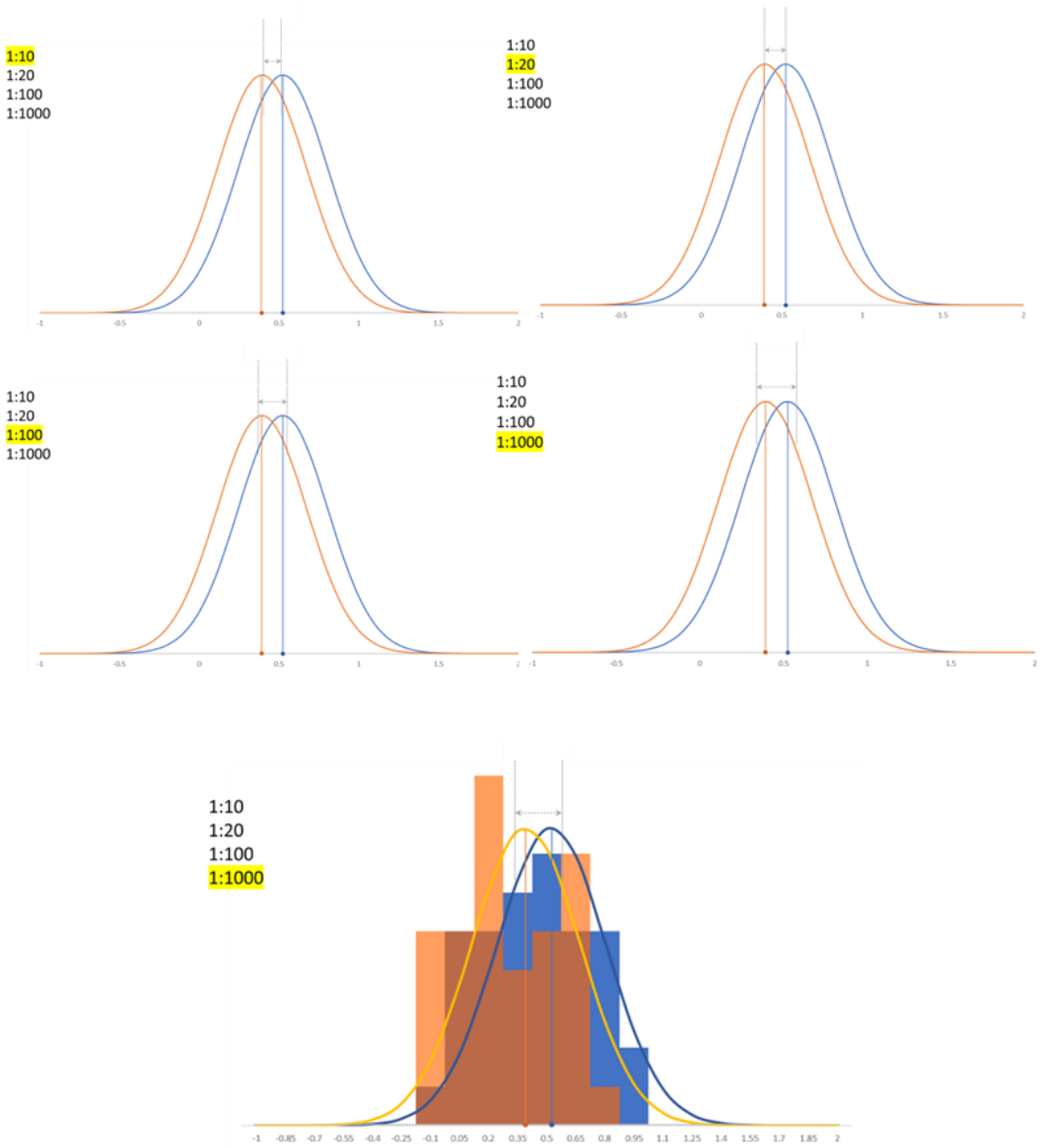
Asking for Other Feedback

The interviews ended in two phases. Participants were first thanked, and asked on camera whether they had additional questions or reactions to share. They were also told that they would have an opportunity to share unrecorded feedback should they wish.

5.2.3 Sources of Strawmen Graphics

The three strawmen graphics used as objects of discussion during this research each reflected a different source.

Fig 5.1
Strawman #1 – Overlapping Normal Distributions



Strawman #1

Strawman #1 illustrates the overlap of two sample distributions – the blue and orange samples – represented both as a pair of histograms, and as a pair of normal curves which had been fit to the samples. The goal for users was to determine whether the two samples were similar enough in their means that they were likely drawn from a common population, or whether they were so unlikely to have been drawn from a single population that they probably represented sub-populations. The graphic included an aid for users to make this determination in the form of a 'difference ruler', that is, a pair of grey lines connected by a double headed arrow meant to signify the amount of separation two sample means needed to show to represent a statistically significant difference for a given confidence level (alphas corresponding to a p-value of .1, .05, .01. or .001).

Studies 1 & 2 inspired the use of overlapping histograms with fitted bell curves to represent a statistical test of two samples (Newburger 2023). The use of the difference ruler was a novel design element added for this study.

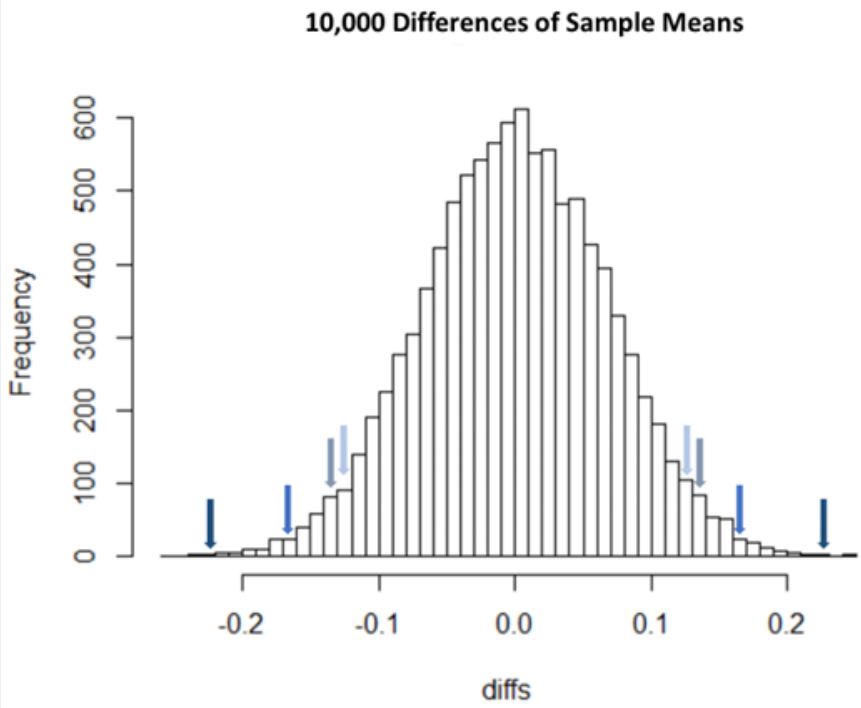
Strawman #2

Strawman #2 was suggested by the subject during the one of the pre-interviews. It attempts to represent the distribution of the test statistic for the two sample means, with arrows to indicate how far out on the distribution a test statistic needed to fall to indicate statistical significance at a given alpha level. The graphic was created using a simulation process, in which pairs of samples were drawn from a

hypothetical single population which complied with assumptions similar to those required of a t-test:

1. Using R (appendix), samples were drawn from a normally distributed population;
2. The population had a mean that was the mean of the two original samples (orange and blue);
3. The population had a variability (standard deviation) that was the joint variability of the two original samples (orange and blue);
4. 10,000 pairs of samples were drawn from this hypothetical population;
5. For each pair of samples, a difference of means was calculated;
6. These mean differences were plotted in a histogram with sufficient bins (50) to the appearance of a fairly smooth distribution, while still showing minor deviations from a perfect curve to indicate the data were discrete, rather than a mathematical representation of an actual normal distribution;
7. Arrows were then added to indicate the bar in which the mean-difference resided representing the minimum difference required for statistical significance at a given confidence level.

Fig 5.2
Strawman #2 – Distribution of test statistic



Strawman #3

Strawman #3 is a pair of overlapping 95% confidence intervals, and was inspired by the first group of interviews. Three of the first four participants indicated that some version of overlapping confidence intervals was their internal model of a t-test. Therefore, the researcher included strawman #3 in all subsequent interviews.

Figure 5.3
Strawman #3 – Overlapping Confidence Intervals

5.3 Coding Interview Recordings

The three parts of the interview process were coded by separate processes, with some overlap, using a combination of closed and open codes (Lazar 2017).

5.3.1 Analytic Process Coding

The main body of the interview received the most coding attention of any section, and the most complex coding process.

1. Researchers reviewed their notes, looking for themes;
2. Based upon the themes, researchers created coding schema (Table 5.4) combining a priori codes with open spaces for capturing emergent themes;
3. Researchers coded transcripts;
    a. Texts with sufficient content to be coded were summarized;
    b. A priori codes were captured;
        i. Counts of the number of participants who mentioned a priori codes at least once were enumerated;
4. Open spaces for themes were given a Primary Code;
5. Primary codes were reviewed, and grouped by Detailed theme;
6. Detailed Thematic Codes were then grouped into Broad Themes.
        i. Counts of the number of participants who mentioned Detailed Thematic Codes, and Broad Themes, at least once, were enumerated.

7. A random sample of codes were verified through parallel coding by a second researcher.

Table 5.4 - Coding Schema

Dimension of analysis	Theoretical Framework		Code group	Summary codes
Mindset What are they thinking?	**Expertise**	Novice	Understanding of visual tool	NA
			Understanding of inferential statistics	
			Use of visual statistics	
		Expert	Epistemic warrants	Building validity -- (list)
				Violations to validity -- (list)
			Tacit understanding of statistics	Tacit understanding -- (list)
Tasks What are they doing?	**Analytic Process**	Pre-collection	Understand the question - clarify or reform	Gather Background
			Understand available data - sources and collection methods	
		Data collection/cleaning	Ordering data for use	Structuring data
		Exploratory analysis	Quantification	Estimation
				Projection
				Distribution
			Outliers/Gaps	Data discontinuities
			Comparison	Correlations
			Hypothesis generation	Predict pattern/model
		Confirmatory analysis	Hypothesis testing	Confirmatory stat testing
				Outlier/gap analysis
				Consistency over time/space
		Communicate findings	Reports	Make findings interpretable (list)
				Overcoming misunderstanding (list)
Tools With what are they doing it?	**Quantitative analysis as external cognition**	Interface between human and machine	Calculative, language or symbol based metaphor	Using numbers (list)
			Visual metaphors	Multi-use viz analytic tools (list)
				Single purpose viz analytic tools (list)
				Eye-brain system limitations (list)

Coding Schema

Researchers applied three parallel coding schema to the main body of the interviews based upon three theoretical frameworks – two selected prior to work, and one selected after an initial review of researcher interview notes. Each framework answers a different fundamental question about the participants' expectations from inferential statistical methods, potential visualizations of inferential statistics, visualizations in general. As such, the three frameworks constitute what amount to

three different dimensions of analysis, orthogonal to one another, any or all of which might be indicated by a single participant statement.

For example, participant #7 described part of their analytic process as:

> *So then I fully understand then I receive the data. And then once we got the data, then I think half of our research time, I mean, the data half the time is understanding the data. So using descriptive statistics, a graphical illustration, and then also some missing data pattern, or some unusual outliers.*

This response generated the following **Summary**:

From the Mindset framework – "Tacit: Half the job is prepping and understanding the data"

From the Tasks framework – "Structure, Background, Distribution, Estimates, Discontinuities"

From the Tools framework – "Multitool: descriptive graphics"

The parts of the Summary were then reviewed to produce **Primary Codes:**

Mindset – "Data prep is much of stat work"

Tasks – task names are the a priori codes, thus "Structure, Background, Distribution, Estimates, Discontinuities" are enumerated.

Tools – no specific tool is mentioned, however, "descriptive graphics" refer to a broad-use class of visualizations. This is enumerated. Had a specific tool (such as Histogram or Scatterplot) been mentioned, these would also have been enumerated.

Primary codes from the Mindset framework are further grouped into a detailed

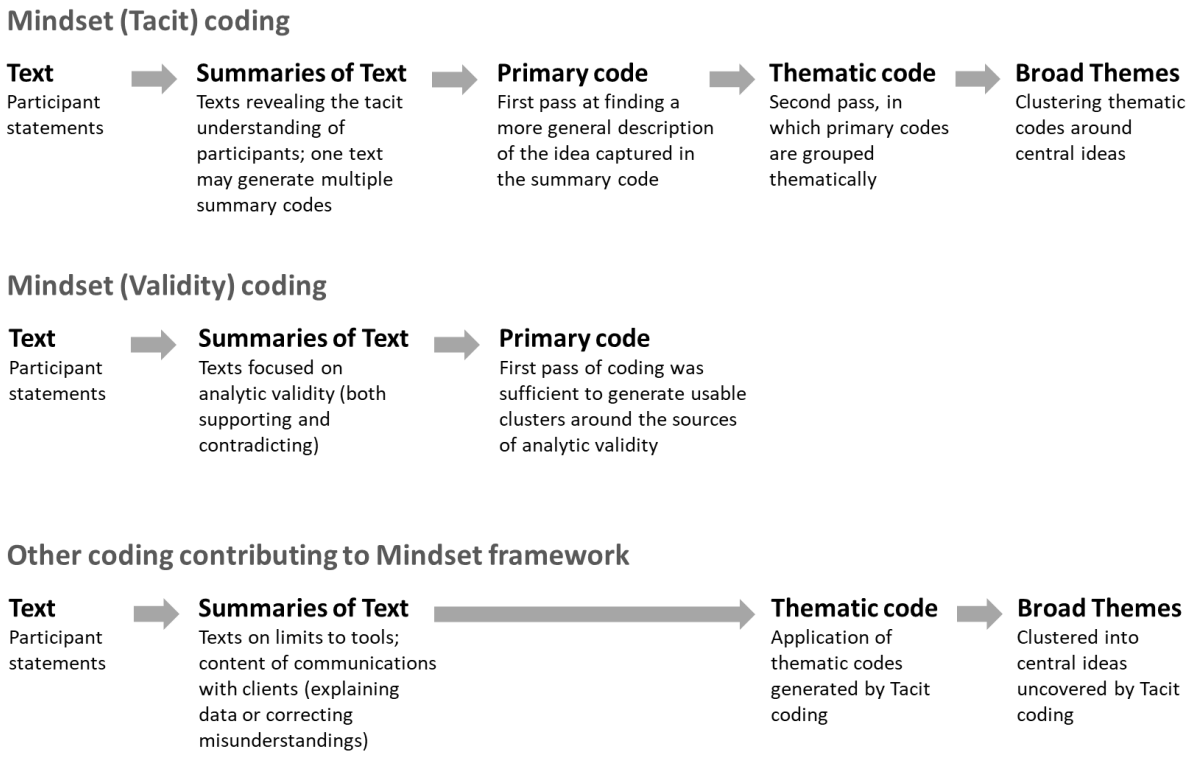
Thematic Code:

Mindset – "Reality is the authority". This code is enumerated.

Detailed Thematic codes are further grouped into **Broad Themes**:

Mindset – "Warrant". This theme is enumerated.

Figure 5.4 – Coding Process for Mindset Codes



5.3.2 Theoretical Frameworks

The Expertise Framework (Mindset)

The first theoretical framework – that of "Expertise" – is fundamental to the conception of this study, which assumes that experts working in an occupation have tacit knowledge of their field. Such tacit knowledge is typically acquired only after significant work experience (Wood 2009), and, therefore, is not explicitly taught as part of formal training prior to entering the occupation. This study assumes that experienced statisticians have tacit knowledge about inferential statistics, and, furthermore, that analytic tools which have that tacit knowledge intuitively embedded within them might raise novice analysts' level of performance to something closer to that of experts.

The Expertise framework contrasts the mindset of experts with that of novices. Using this framework helped us to form requirements for study participation. During interviews, the framework was still active, summarized by the question, "What are participants thinking during their interactions with inferential statistics?" The study-relevant answers to this question come under two broad code grouping.

The first group of codes captured the Epistemic Warrants experts attribute to different inferential methods, that is, the degree to which a particular analytic method provides support for or against some claim. Two emergent codes made up this group.

'Build validity' captures those elements of analysis which shore up a reasoned line of evidence.

'Violations to Validity' captures those elements of analysis which tend to undercut evidence.

The second code group includes only a single emergent code.

'Tacit' captures what experienced statisticians have, through work experience, come to believe about their methods.

The Tasks Framework

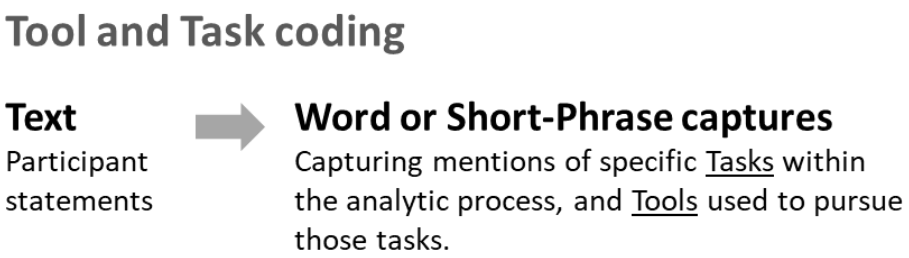
The "Tasks" framework is composed primarily of a priori codes. It emerged out of initial reviews of the interview results, and presumes that there are common steps in the analytic process all statisticians pursue, regardless of the sub-field in which they work.

Identifying the core tasks of analysis as perceived by experienced statisticians provides a context in which to place the kinds of inferential statistics tools this research effort hopes to design, and may suggest features the design should incorporate. More broadly, Task sets the context for understanding the other two dimensions of analysis.

Initial review of interviews indicated five analytic phases within Task (see table), each associated with a number of code groups, and resultant summary codes. As the focus of this research is inferential statistics, a later stage part of the analytic process, some detail was sacrificed during coding for the earliest work phase (the code groups of understanding the research questions, and understanding data sources, were consolidated into a single code, "Background").

While the researchers initially expected this framework to consist of a priori codes, it was found during early review that the Communicate Findings codes (labeled “Myth” and “Comm”) often included details of what needed to be communicated that gave further insight into the mindset of participants. Thus, these were recoded as Emergent codes, with details captured and summarized for thematic coding using the Mindset Thematic Codes list. Thematic codes were applied directly to the Summaries, without the need for an intermediary Primary coding step, due to the narrow focus of these texts (i.e. an explanation of findings or processes offered to a client, or a correction to a client’s misunderstanding).

Figure 5.5 – Coding Process for Mindset Codes



The Tools Framework (a.k.a. External Cognition)

Data visualization serves as a form of external cognition, with the visuals acting as an interface between human and computer, to create a composite mind with advantages from both (Scaife 1996). The computer provides a place to offload memory and calculative tasks, leaving the human mind freer to recognize patterns, form and test hypotheses, and creatively explore venues for further discover.

Modern statistical analysis relies heavily upon the calculative power and memory of computers, making it an ideal candidate for tools of external cognition. However, visualization is only one form external cognition can take. For example,

SQL accessible databases represent a powerful memory aid based upon symbolic logic (programming), or a statistician might employ the vast calculation power of a mainframe through a short SAS script.

The Tools Framework captures the specific interfacing-tools statisticians choose to employ during their work. Broadly speaking, these tools can either be visual metaphors (data visualization) or symbolic (such as code or written equations).

The Visual Metaphor code group includes three summary codes – **Broad focus** visualizations that may have many uses (such as histograms, from which an analyst might discern the mean, median, variance, skewness, range, location and number of modes, gaps in the distribution, outliers, and shape), **Narrow-focus** visualizations (such as QQnorm plots, which are used almost exclusively to test the normality of a distribution), and observations on **Limits** to the eye brain system offered by participants. Broad focus and Narrow focus are emergent codes, however, they require minimal additional coding – the capture of the tool by name is informative on its own. However, the Limits codes are similar to Myth and Comm codes, in that they provide insight into the mindset of participants, specifically, what they think about the use of visualization. Thus, several Limits codes have been thematically coded.

The Symbolic Tools code group was captured with only one summary code, since visualization is the focus of this research. Similar to the Broad and Narrow focus summary codes from the Visual Metaphor code group, Symbolic Tools were merely captured by name.

5.3.3 Graphic Elicitation Coding

Participant graphics were screen-captured, along with participant descriptions of the images. These were labeled by type, then grouped according to their analytic focus.

5.3.4 Strawman Graphics

Coding of the strawman graphics section was a combination of a priori and emergent codes.

Table 5.5 Strawman Graphics Reaction Coding

Reactions to Strawman Graphics		
Testing strawman graphics	Understanding	Understood
		Misunderstood
	Affinity	Liked (list)
		Disliked (list)
	Recommendations	Recommendations positive (list)
		Recommendation negative (list)

For all graphics presented to participants, codes captured:

1. **Participant Understanding** – Whether the participant correctly assessed the statistical significance at the given alpha value using the graphic – this was an a priori code;
    a. Correct assessment of statistical significance was taken as an indicator of understanding;
    b. Incorrect assessment was taken as an indicator of a lack of understanding;
    c. Delayed understanding was noted, along with what additional information supplied by the researcher had brought the participant to understanding;

2. **Participant Affinity for the Strawman Graphic** – Whether participants found the strawman graphic useful, informative, or otherwise saw it is a positive/negative light. – These were emergent codes;
    a. Positive comments were noted and summarized;
    b. Negative comments were noted and summarized;
3. **Participant Design Recommendations –** Any directly stated or implied design recommendations to improve the strawman graphic. – These were emergent codes;
    a. Affirmative suggestions were noted and summarized;
    b. Negative suggestions were noted and summarized.

The intention of this section was to generate informative conversation about, and observations on, potential elements for graphic inference tools. The hope was that confronting participants with the experience of using a graphic tool to perform a statistical inference would generate deeper insights than a simple presentation of the graphics with a request for comments. This worked to such an extent that, in addition to revealing several design recommendations, this section of the interviews generated responses that were informative of expert statisticians understanding of inferential statistics in general. These more general comments were mixed with texts from the Analytic Process section of the interview, and coded by that three-part schema.

5.3.5 Code Reliability

The above processes resulted in four sets of codes, two simple, and two more complex. We tested the later for inter-coder reliability, reasoning as follows:

1. Word and short-phrase captures required minimal subjective judgement to assign, and thus did not require a reliability test. These included:
    a. lists of tools named by analysts, for example, “scatterplots”, “Bar charts”, or “regression”, which were simple word captures;
    b. tasks named through short phrases, for example, “missing-data pattern” and “unusual outliers” directly describe tasks represented by the code *discontinuities* - the identification of gaps and outliers (discontinuities) in a data distribution. Similarly, the phrase, “descriptive statistics” denotes tasks represented by the codes, *Distribution,* and *Estimates*.;
2. Categorization of the participant-drawn graphics: these also require little subjective judgement, as the participants themselves provided detailed descriptions of their understanding of their drawings. In addition, it is possible to publish the actual graphics without risking a breach of anonymity;
3. Participants’ descriptions of factors which either add to, or detract from, analytic validity were expressed as longer phrases more subject to interpretation. These called for inter-coder reliability measures.
4. Inter-coder reliability measures were also required for all participant expressions which the researchers coded using the thematic codes developed for parsing participant tacit understanding.

Independent Coding

The technical nature of the interviews meant that an outside coder would have to be a statistician of similar experience to the researchers and participants. After

finding a qualified volunteer, they were provided with a 10% random sample of texts (with codes stripped), along with the full list of Primary Codes applied the researchers had devised to parse statements of 'analytic validity'. Their choices were compared to the researchers'. The volunteer was next provided with a 10% random sample of texts expressing 'tacit understanding' (with codes stripped), along with the full list of 'Thematic' codes researchers had devised to parse statements of 'tacit understanding.' Their choices were compared to the researchers'. Measures of intercoder reliability were calculated using Cohen's Kappa.

5.4 Results

In all, 313 texts were assigned codes associated with the 'mindset' theoretical framework (Table 5.6). The majority of texts were captured during participants' descriptions of their analytic process, however, applicable texts could come from any phase of the interview.

Table 5.6 – Coded Texts by Code Type

	Coded texts	Summary of texts	Primary codes	Thematic codes	Broad Themes
All Mindset texts	**313**				
Tacit		260	107	44	6
Validity		61	15		
All Tool texts	**149**				
Multi-focus		115			
Narrow-focus		38			
Numbers		33			
Limits		7		4	3
All Task texts	**383**				
Myths		16		13	6
Comm		14		5	4

Note: Texts may generate multiple codes

Two hundred sixty texts received 'Tacit' understanding codes, resulting in 107 primary codes, 44 thematic codes, organized into 6 broad themes. Sixty one texts were deemed informative about sources of analytic validity (building or reducing validity), which were given 15 primary codes. These 15 codes were considered sufficiently detailed that further refinement was unnecessary (Table 5.7)

Table 5.7 'Validity' Codes

Participants reporting	Two-word Code	Definition
10	**assume false**	We assume our findings reflect an error until we exhaust every possible mistake we could have made.
7	**check assumptions**	We need to make sure all our mathematical assumptions actually match our analyses before we can believe.
5	**avoid rationalization**	We should make predictions/plan ahead/discipline our analyses to avoid fooling ourselves after the fact.
5	**data model**	It's critical to make sure we have a link between our measures and reality. It's easy to lose that link.
4	**parallel analyses**	We can check our analytic steps by doing them more than one way -- whether that's parallel coding or using more than one method (viz and numeric) to get at the same thing.
3	**check expectations**	We need to compare findings with expected values (whatever the source of those) as a check against unknown forces that could make our findings a lie. It's a sort of ground truth test.
3	**consider mistakes**	We should always consider stuff we might have done wrong -- programming, data collection, etc.
3	**SM experts**	Subject matter experts provide the data model, and provide expected values, both of which are vital to building validity.
3	**trust data**	We should always believe the data more than our hypotheses.
2	**admit uncertainty**	We need to consider uncertainty measures when we think we have findings.
2	**sample size**	Sample size is vital to believing our results.
1	**close study**	It's easy to make analytic errors, because this stuff is subtle and quick thinking tends to result in mistakes.
1	**consider luck**	We need to consider whether our results could be due to random chance -- we can quantify this null hypothesis.
1	**distrust eyes**	We can fool ourselves -- that's why we want to externalize the decision making.
1	**narratives convince**	We can fool ourselves and others with a good story. We must be careful with this.

One hundred forty nine texts were parsed for mentions of specific tools, including 115 listing multi-focus tools, 38 narrow-focus, and 33 purely numeric or equation-based. Some mentions of specific tools were also collected from researcher notes.

Seven texts included mentions by participants of perceived limits on visual tools. These were assigned thematic codes (4), which came from 3 of the six broad themes.

Texts which mentioned analytic tasks were the most common (383), which likely reflects the structure of the interviews. These texts were parsed for short phrases which identified some specific instance of a step in the analytic process. For example, the phrase, *"...have to calculate confidence intervals and then ...I have to see if things overlap,"* resulted in the task code, "Test", signifying confirmatory analysis, in this case, overlapping confidence intervals as an approximation of a t-test.

Some mentions of specific analytic tasks were also captured from researcher notes.

A subset of these texts (14) included descriptions of communications participants felt were critical to deliver along with their analyses, to aid their clients' understanding. Another subset of 16 texts included misunderstandings held by clients the participants often had to correct. Both of these subsets were coded as insights into participants' tacit understanding of statistics. They were assigned thematic codes from among the 44 codes specified during "Tacit" coding, fitting into the 6 broad themes.

5.4.1 Intercoder reliability

Coding was tested for intercoder reliability via a 10% sample of texts. Since coding was done on individual texts, the context of the full interview – such as

participant comments just prior to and following each text in sample – are missing. This may have reduced intercoder reliability measures.

"Validity" code agreement between coders was 71%. Cohen's Kappa was found to be .588, near the top of the "Moderate agreement" range.

"Tacit" coding was also tested for intercoder reliability via a 10% sample at the Broad Themes and Thematic code levels. Codes were tested independently, that is, as two separate code lists rather than as a unified coding schema in which one set was series of subsets of the other. In reality, a missed Broad code necessarily meant a missed Thematic code, naturally resulting in lower scores for the latter (percent agreement for Thematic codes within the subset of cases where Broad code matched was 75%, considerably higher than the independent estimate reported below).

Among Broad codes, there was 67.7% agreement, with a Cohen's Kappa of .611, just at the bottom of the "Substantial Agreement" range. Agreement between coders for Thematic Codes was 51.7%, with Cohen's Kappa at .495, the middle of the "Moderate Agreement" range.

5.4.2 Six Broad Themes

The six Broad Themes have wide support from participant interviews. A majority of participants made comments coded into all 6, and no participant made comments coded into fewer than 3 (Table 5.8).

Table 5.8 Expressed and Coded Broad Themes of Tacit Understanding

	Caution	Expertise	Limits	Planning	Viz	Warrant
Participant 1	●	●	●	●	●	●
Participant 2	●	●		●	●	●
Participant 3	●	●	●		●	●
Participant 4	●	●	●	●	●	●
Participant 5	●	●	●	●	●	●
Participant 6	●	●	●	●	●	●
Participant 7	●	●			●	●
Participant 8		●	●		●	
Participant 9	●	●	●	●	●	●
Participant 10	●	●	●	●		●
Participant 11	●	●	●	●	●	●
Participant 12	●	●	●	●	●	●
Participant 13	●	●	●	●	●	●
Participant 14	●	●	●	●	●	●
Participant 15	●		●		●	●
Participant 17	●	●	●	●	●	●

Caution

The "Caution" grouping expressed the multifaceted concerns statisticians have in pursuing their work, and their approaches to addressing those concerns. It included thematic codes such as, "People act upon our results," an admonition to remember that people trust statistical work, take action based upon it, and, therefore, it is a statistical professional's responsibility to put in whatever time and effort is required to always provide the best possible advice. The group also includes, "Distrust findings," which captured the many ways participants remind themselves to always check and recheck their work, since statistical analysis is a fundamentally complex

endeavor which allows for many points of failure. The Caution theme also touched upon statisticians' relationship to inferential statistical methods, with, "P-values encourage bad thinking," referring to concerns in the statistical community about the tendency for statistical testing to foment dichotomous thinking about complex realities, and, "Stat tests required for publication," expressing participants' belief that whatever risks parametric statistical tests entail, they are nonetheless required for acceptance within many scientific communities, and thus must be used as best they can.

Expertise

The "Expertise" group applies to the several aspects of acquired analytic understanding which, as a body, represent a divide between statistical professionals and people outside the field. It includes codes such as, "Analysis takes a statistician," capturing participants' expressions of their belief that people outside the statistical field typically misunderstand at least some aspects of quantitative work. It also included, "Stat testing is hard for statisticians, too," which captured participants' expressions that statistical inference is a subject so complex that they don't trust their own knowledge without the use of references.

Limits

The "Limits" group captured several observations from participants describing ways that details of a data collection can limit the range of statistical tools available to apply, but also the ways in which the choice of statistical methods can limit the scope of analytic research. These codes are not specifics about the various limitations discussed, but, rather, the awareness among participants that quantitative work does entail limitations. Example codes include, "Sample size important," expressing the multiple dependencies between sample size and the validity of inferences made about populations from which those samples were drawn, and, "Stat methods define scope," capturing expressions of the ways in which the tools of statistics define the kinds of questions statistical research can address.

Planning

"Planning" codes capture the importance of planning in quantitative work: its utility, costs, and pitfalls. For example, the code, "Predict to escape rationalization," captures the participants' understanding that post hoc rationalization (sometimes called the, "Texas Sharpshooter" fallacy in statistics) is a constant temptation during analytic work which threatens results validity, and that the way to avoid this through planning ahead; they plan the analyses they will run, the test statistics they will accept, etc. "Plan defines scope," captures the participants' awareness of how the planning process, while vital to the work, once entered, limits possible discoveries the work may yield.

Viz

"Viz" captures participants' understanding of data visualization as a tool in their analytic work.

Warrant

"Warrant" captures participants' sometimes contradictory understandings of what elements within, or conditions are required by, their quantitative work to give the power to make statements about the world. These codes can be broad, such as, "Reality is the authority", expressing participants' focus on always connecting their computations as directly as possible to the subject of their study, or checking results against expected values extracted from 'facts on the ground' sources, such as news reports. Some are more specific, such as, "Effect size >= p value," which expressed the common feeling among participants that statistical significance was less important, or at least no more important, than the practical significance in their results. For example, a trial on a cholesterol drug with a large enough sample size might show a statistically significant reduction in blood cholesterol levels, but that reduction could still be so small as to have no expected effect on clinical outcomes for patients.

5.4.3 Analytic Process

The Framework of Analytic Tasks derived from the initial review of researcher notes proved to be well supported by subsequent formal coding processes. Thirteen of 18 participants reported performing work steps which fell within all 5 of the proposed tasks, and no participant reported fewer than 3 (Table 5.9).

Table 5.9 Statistician's Analytic Process

		Interviewee 1	Interviewee 2	Interviewee 3	Interviewee 4	Interviewee 5	Interviewee 6	Interviewee 7	Interviewee 8	Interviewee 9	Interviewee 10	Interviewee 11	Interviewee 12	Interviewee 13	Interviewee 14	Interviewee 15	Interviewee 17	Interviewee A	Interviewee B
Pre-collection	Gather Background	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
Data collection/cleaning	Structuring data	●		●	●	●	●	●	●	●	●	●	●	●	●		●	●	●
Exploratory analysis	Estimation	●		●			●	●		●	●	●	●		●	●	●		●
	Distribution	●	●	●	●	●	●	●	●	●	●	●	●		●	●	●	●	●
	Data discontinuities	●	●		●	●	●	●	●	●	●	●	●		●	●	●	●	
	Correlations	●		●	●	●	●		●		●	●	●		●	●	●		●
	Predict pattern/model	●		●	●	●	●		●	●	●	●	●		●	●	●		●
Confirmatory analysis	Confirmatory stat testing	●		●		●	●	●		●	●	●	●		●		●	●	●
	Consistency over time/space				●	●	●		●	●	●						●		
Communicate findings	Make findings interpretable	●	●		●	●	●			●		●	●	●	●	●	●	●	●
	Overcoming misunderstanding	●			●		●	●	●	●			●	●	●		●		

	Number of interviewees																		
All 5 task groups	13	●			●	●	●	●	●	●		●	●		●		●	●	●
Pre-collection	18	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
Data collection/cleaning	16	●		●	●	●	●	●	●	●	●	●	●	●	●		●	●	●
Exploratory analysis	17	●	●	●	●	●	●	●	●	●	●	●	●		●	●	●	●	●
Confirmatory analysis	15	●		●	●	●	●	●	●	●	●	●	●		●		●	●	●
Communicate findings	16	●	●		●	●	●	●	●	●		●	●	●	●	●	●	●	●

Pre-collection activities, such as meeting with clients to determine their needs, gathering background information on available datasets, or proposing analytic methods, were universally reported among participants, with other steps nearly so. Confirmatory analysis was the least reported step, with 15/18 participants calling it out during descriptions of their analytic work.

Note that to qualify for the study, all participants began by confirming that they performed statistical testing as part of their regular work process, or had at some point. Similarly, 17 of 18 participants reported publishing their work publicly, and the 18th reported sharing their work internally within their organization, all of which constitutes communicating results. Therefore, it is likely that while these steps were not universally captured during the coding exercise, all participants did, in fact,

perform these steps during their work. This may further substantiate the Tasks framework.

5.4.4 Tools of External Cognition

Parsing texts for mentions of specific analytic tools (named methods, procedures, or statistical routines encapsulated within software packages) resulted in three lists, following the External Cognition theoretical framework: Broad use visualization tools, Narrow use tools, and Purely numeric or equation-based approaches. Unsurprisingly given the visual-analytic focus of these conversations, mentions of visual tools outnumbered purely numeric ones. However, the diversity of Narrow use tools did surprise the authors (Table 5.10).

Table 5.10 Tools of External Cognition

Broad-use Visualization tools	**Narrow-use Visualization tools**	**Purely Numeric tools**
Animation	Bland-Altman plot/Means difference plot	ANOVA
Bar chart	Classification trees	Bayesian modeling
Box plots	Confidence intervals	Confidence intervals
Business intelligence plots (bar chart etc.)	convergence checks on Bayesian models	hazard ratio
Butterfly chart	Decision matrix	Key Performance Indicators/Indexes
Choropleth	Dendrograms	Machine learning (in general)
decision trees	Genome graphics	Machine learning (random forests)
density plots	Kaplan-Meier plots	natural language processing
Dot plots	Model parameter plots	odds ratios
Heat map	Multiple-model plot to show cone of uncertainty	parameters
Histograms	Qqplots of normality	Probit
Icon map	Residual plot	P-values
Line diagrams	Spatial Model plot	Regression (linear)
Maps	Survival curve (non-parametric)	Regression (logistic)
number line	Variogram	sample size
pie charts	Venn diagram	Standard error
Pre-packaged descriptive viz collections	volcano graph	statistical testing
Process model		Summary statistics
ranked lists		
scatterplot with trend line		
Scatterplots		
Sorted lists		
spaghetti plots		
Stem and Leaf plot		
Tables of descriptive statistics for exploration		
Tabular interactive (Spreadsheet)		
Tree maps		
Violin plot		

Standard Graphic Tools

Some broad-use statistical graphics have a long history, wide availability in software packages, and ubiquitous appearances in literature with statistical content (Shneiderman 1996, Grammel 2010, Huron 2014, Pousman 2007). We can interpret use of such tools by participants as an indicator of the degree to which they fold visualization into their work.

Table 5.11 focuses on seven graphic forms: Scatter plot, Histogram, Boxplot, Line Diagram, Bar Chart, Table, Pie chart. Note that tables, while numeric in content, make use of a visual organizational schema (Bartram 2021).

Table 5.11 Use of Standard Broad-use Visualizations

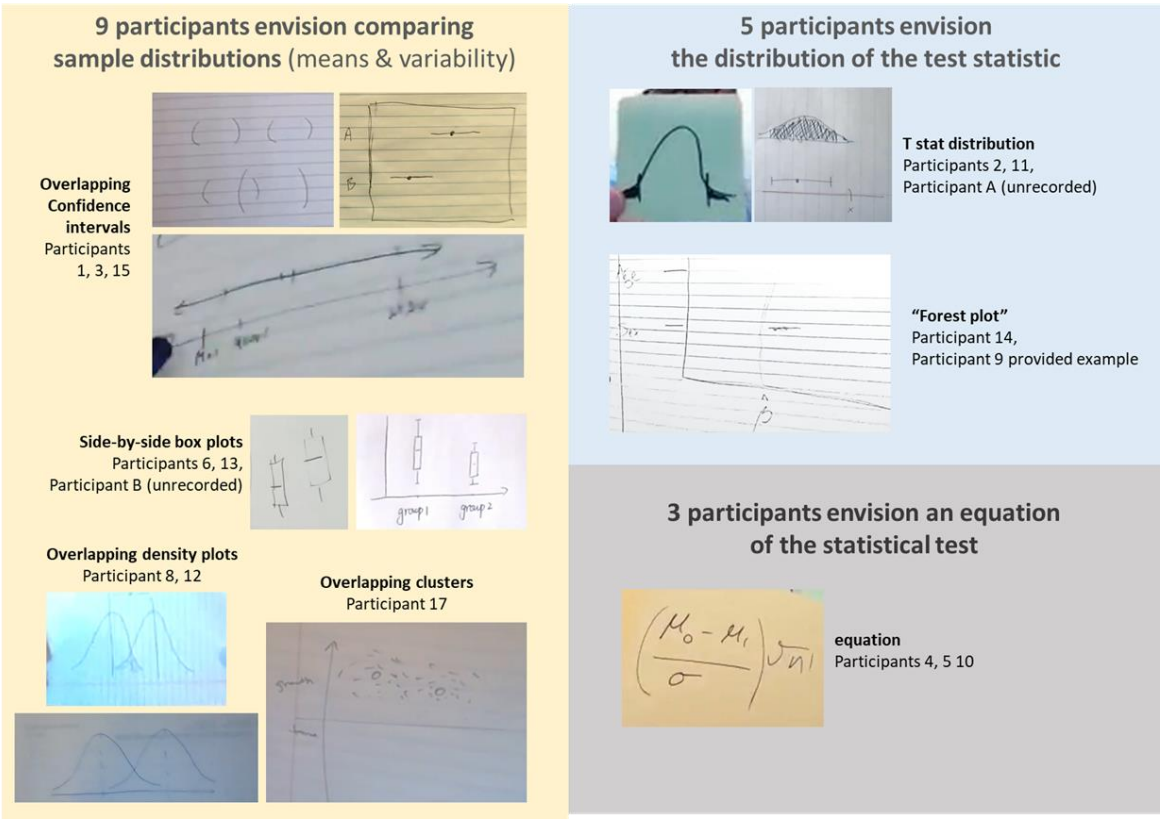
	Any Standard Graphic	Scatter Plot	Histogram	Box Plot	Line diagram	Bar Chart	Table	Pie chart
Total with 1+ mentions	3.8	13	12	11	11	10	9	2
Participant 17	7	✓	✓	✓	✓	✓	✓	✓
Participant 3	5	✓	✓	✓	✓	✓		
Participant 10	5	✓	✓		✓	✓	✓	
Participant 13	5	✓	✓	✓	✓	✓		
Participant B	5	✓	✓	✓		✓	✓	
Participant 2	4	✓	✓	✓			✓	
Participant 4	4		✓		✓	✓	✓	
Participant 5	4	✓			✓	✓	✓	
Participant 8	4	✓	✓	✓	✓			
Participant 9	4	✓		✓	✓		✓	
Participant 15	4	✓		✓	✓	✓		
Participant A	4			✓		✓	✓	✓
Participant 14	3		✓		✓		✓	
Participant 6	3	✓		✓	✓			
Participant 12	3	✓	✓			✓		
Participant 7	2		✓	✓				
Participant 1	1		✓					
Participant 11	1	✓						

All 18 participants used some visualization from this list, likely with analytic intent. Scatterplots were the most widely named graphic form, mentioned by 13/18 participants. Histograms followed closely, with 12/18. Both of these forms are explicitly analytic in function compared with other forms (pie charts) which are less useful for analysis but are sometimes favored in communicating findings. Among participants reporting only 1 tool used, it was either a Scatterplot or Histogram. All graphic forms on this list were widely used with the exception of the Pie Chart (2/18). On average, participants reported use 3.8 out of these 7 tools.

5.4.5 Graphic Elicitation

Twelve of 18 participants provided sketches representing what a statistical test looks like to them, which were grabbed through screen captures (Figure 5.6). Of the remaining 6, two participants' sketches (Participants A and B) were captured via researcher notes, and confirmed by the participant (the researcher displayed their notes though the video conference screen for approval). These are described below. Two participants expressed that their mental image of a statistical test was that of an equation and provided no sketch (a third drew an equation). One participant named a Forrest plot as their internal image of a statistical test, and provided an example from the literature. One participant didn't participate in the graphic elicitation portion of the interview.

Figure 5.6 – Graphic Elicitation



All participants who provided a sketch were asked to walk the researcher through their work. In summary, 9 participants communicated that their vision of a statistical test was a pair of sample center (mean or median) displayed side by side, with some indicator of sample variability around those centers (such as overlapping confidence intervals or side-by-side boxplots). Five participants indicated they envisioned the distribution of the relevant test statistic, with the location of the realized test statistic noted, and with an indicator of the likelihood of being at that location on the distribution. Three more participants indicated their internal model of a statistical test is purely equation-based.

5.4.5 Participant Reactions to Strawmen Graphics

Participant reactions to strawmen graphics provided a basis for several design recommendations. Liking/disliking spoke to whether participants' would accept graphic features of the strawmen as legitimate analytic tools. Understanding/misunderstanding spoke to intuitiveness of graphic features. Table 5.12 summarizes these reactions.

Table 5.12 Reaction to Strawmen Graphics

	Positives about Strawman Graphic	Negatives about Strawman Graphic
Strawman #1	P1 - seems complete	P3 - showing different sig levels invites rationalization/p-hacking
	P2 - Converting the critical value from the book tables into an actual difference on screen solidifies the math	P4 - Disliked
	P3 - All the stuff they want to see in one spot	P9 - #1 missing information about sample size and thus standard error
	P4 - Liked histograms at least	P13 - without effect sizes, why bother?
	P5 - Underlying sample distributions from #1	P15 - Noise of #1
	P6 - histograms provide better understanding of distribution than box plots	P17 - #1, #2 because of assumption violations.
	P7 - Like graph #1 for showing the distributions	P-A - Disliked -- "very few phenomena are normal"
	P11 - #1 because of the raw distribution display	P-B - Struggled at first, but eventually understood
	P14 - Animation of the grey bars highly informative| Both #1 and #3 show the actual difference, and that matters for effect size|First one for showing histograms/underlying data to assess assumptions of the test	
Strawman #2	P2 - he liked #2 best, but thought #1 would work better for other people just starting out. |linking measures to distribution to concretize test	P1 - Preferred the completeness of #1
	P4 - Liked	P3 - showing different sig levels invites rationalization/p-hacking
	P6 - Liked #2 best	P5 - #2 doesn't show the underlying sample distributions, which #1 did.
	P9 - Liked #2 best	P17 - #1, #2 because of assumption violations.
	P12 - Liked graph #2 best	
	P13 - Favorite=#2 Liked it a lot	
	P14 - #2 best	
	P15 - #2 and #3 best	
	P-A - Preferred approach as less parametric	
	P-B - More comfortable with distribution of probability	
Strawman #3	P5 - #3 is clean, though less informative, but is the one they use often	P9 - #3 too inflexible
	P6 - Liked #3 a lot; would want to see both #2 and #3 together	P12 - Disliked #3
	P10 - #3 best	P14 - Dislike: Confidence intervals appear to have less information than #2
	P13 - The changing size of CIs with alpha changes -- really connects the test to the evidence.	
	P14 - #3 best if there are consequences to the decision|Confidence intervals emphasize uncertainty	
	P15 - #2 and #3 best	
	P17 - #3, Liked the Simplicity	

5.4.6 Participant Design Suggestions

Some participants contributed observations on the strawmen graphics which constituted design recommendations. These were parsed by researchers during coding.

Table 5.13-
Participant Recommendations Inspired by Strawmen Graphics

6	numbers
5	sample size
4	Effect size
4	size CIs
1	Clearer distinction between data and rulers
1	Include #2
1	include all three
1	Include visual of test's assumptions
1	include data view
1	Include instructions
1	Include labels
1	avoid stat verbiage like 'simulation'

5.5 Findings

5.5.1 Statisticians' Relationship to Visualization

Statisticians make extensive use of visualization in their analytic work. Every participant in this study employed visualization as a regular part of their analytic process. All 18 made use of one or more of the common broad-use visualizations. Most used several. Most also used narrow-use visualizations; 14 of 18 participants reported making use of at least one specialized visualization with a narrow analytic focus. Indeed, participants reporting using the fewest of the 'common broad-use'

visualizations all reported making use of a specialized narrow-use visualization. A third of participants (6/18) shared observations on the limits of data visualizations while still making use of them. This suggests they use these tools enough to weigh their plusses and minuses.

Table 5.14 Participant Visualization Use

	Any Standard Graphic	Broad-use visualization	Narrow-use visualization	Display raw numbers or calculations	Expressing limits to visualization	Participants' mental conceptions of a T-test
Total with 1+ mentions	3.8	18	14	16	6	
Participant 17	7	✓		✓		Sample centers with variability measures
Participant 3	5	✓	✓	✓		Sample centers with variability measures
Participant 10	5	✓		✓		Equation
Participant 13	5	✓	✓	✓		Sample centers with variability measures
Participant B	5	✓		✓		Sample centers with variability measures
Participant 2	4	✓	✓			Test statistics distribution
Participant 4	4	✓		✓	✓	Equation
Participant 5	4	✓	✓	✓		Equation
Participant 8	4	✓	✓	✓	✓	Sample centers with variability measure
Participant 9	4	✓	✓	✓		Test statistics distributions
Participant 15	4	✓	✓	✓		Sample centers with variability measures
Participant A	4	✓	✓	✓	✓	Test statistics distribution
Participant 14	3	✓	✓	✓		Test statistics distribution
Participant 6	3	✓	✓	✓		Sample centers with variability measures
Participant 12	3	✓	✓	✓	✓	Sample centers with variability measures
Participant 7	2	✓	✓			NA
Participant 1	1	✓	✓	✓	✓	Sample centers with variability measures
Participant 11	1	✓	✓	✓	✓	Test statistics distribution

Beyond mere use, it appears that statisticians conceptualize at least some parts of their work in visual terms. Fourteen of 17 participants indicated that their internal conception of at least one inferential statistical method is visual (Table 5.12).

Figure 5.7 - Tacit coding

Planning	10	**Plan defines scope**	P1 P2 P4 P5 P9 P11 P12 P13 P14 P17
	6	**Predict to escape rationalization**	P2 P5 P9 P11 P14 P17
	5	**Analysis is iterative**	P6 P9 P11 P12 P14
	3	**Software defaults can bias work**	P5 P10 P13
Viz	11	**Viz is analysis**	P1 P3 P4 P6 P7 P8 P9 P11 P12 P14 P15
	10	**Viz communicates stories**	P1 P2 P4 P5 P6 P9 P12 P13 P14 P17
	2	**Visualization is art**	P4 P13
	1	**Designer's dilemma**	P12
	1	**Viz as analysis isn't "Viz"**	P13
	1	**Viz of test changed understanding**	P3
Warrant	14	**Reality is the authority**	P1 P2 P4 P5 P6 P7 P9 P10 P11 P12 P13 P14 P15 P17
	9	**Effect size >= p value**	P1 P2 P3 P7 P9 P10 P11 P13 P14
	7	**Reality makes the math understandable**	P1 P2 P5 P6 P10 P12 P13
	5	**p=.05 weak**	P5 P9 P12 P13 P15
	5	**Test as evidentiary warrant**	P1 P5 P14 P15 P17
	5	**Viz > math**	P3 P4 P6 P12 P15
	4	**Math > Viz**	P3 P5 P11 P14
	4	**Viz too subjective**	P3 P5 P11 P14
	2	**Analytic assumptions can distract from the data**	P10 P13
	1	**Math for objectivity**	P5
Caution	10	**Distrust findings**	P1 P4 P5 P6 P9 P10 P12 P13 P14 P15
	10	**People act upon our results**	P4 P5 P7 P9 P10 P11 P12 P13 P14 P15
	7	**People hijack stats**	P4 P9 P10 P11 P13 P15 P17
	6	**Must match uncertainty to consequences**	P7 P9 P10 P12 P6 P14
	6	**Rely upon published authorities**	P1 P2 P3 P5 P6 P10
	4	**Informed guesses are better than none**	P10 P11 P14 P15
	4	**P values encourage bad thinking**	P7 P10 P13 P15
	2	**Stat tests required for publication**	P10 P13
	1	**Privacy requirements can interfere with our work**	P5
Expertise	14	**Analysis takes a statistician**	P2 P3 P4 P5 P6 P7 P8 P9 P10 P11 P12 P13 P14 P17
	8	**Stat testing is hard for statisticians, too**	P1 P2 P3 P4 P5 P6 P8 P10
	3	**Analytic cultures differ**	P7 P11 P13
	3	**Data can be involving**	P4 P12 P14
	3	**Rely upon collaborators**	P1 P2 P10
	1	**Stat experience teaches expertise beyond training**	P11
	1	**Statisticians lean on habits**	P6
Limits	9	**Sample size important**	P1 P4 P5 P6 P9 P10 P11 P12 P13
	4	**Analytic comparisons**	P1 P3 P8 P17
	4	**Stat methods define scope**	P1 P6 P9 P17
	3	**Comparing quantitates is universal**	P5 P9 P11
	3	**Small samples problematic**	P5 P6 P9
	3	**Statistical assumptions matter**	P6 P12 P15
	2	**T-test compares distributions**	P6 P11
	2	**Uncertainty measures are tightly bound to stat tests**	P6 P14

Eleven of 16 participants expressed the idea that visualization IS analysis, describing how it is fully integrated into their quantitative work (Figure 5.7). For example:

> P6 - *No, not exactly. Maybe I just take that for granted, the image thing. I think once we have data and the first thing probably is to plot the data and see what the data looks like. I think understanding data, to understand the data, visualization is a very important tool. I once advised a student and who is quite mathematical, and he's very [inaudible 00:25:26] resistant to plot the data, because he thinks can give you a Y and X dataset, and he can just fit linear regression using R and output the coefficient and the confidence interval for all the coefficients. And he can do hypothesis test for the parameter, get P value. That's it. I have to talk with him for a very long time to tell him you first need at least to plot X and a Y data and see what's their relationship.*

Or…

> **P11** - *I know people have different approaches to this, but I tend to just dive in and start working on, if it's something about [data] relationships, then I start examining the relationships, whether that's with correlations or just plots. I think that was something that I learned. I think teachers throughout my career have emphasized the value of plotting things. I probably don't do it as much as they would like me to…*

Yet despite their frequent use of visualization and their understanding of their analyses in visual terms, it is not necessarily the visualization work they share with others.

> P13 - *Oh yes. I use line graphs. I use bar graphs. I use scatter plots. Absolutely. For me, to visualize sort of how the data hang together, and then, are there aspects of a model that don't fit? Where are the outliers? I use, to see how the variance, is it consistent or is it a trumpet? How are the errors distributed? I use that myself. That's not something I would share typically…*

Indeed, some participants described using visualization, but didn't think that counted as visualization. Two (P4 and P13) felt that artifacts which rose to the level of deserving the label, "visualization," had to exceeded mere utilitarian analysis and achieve aesthetic value, something they had little confidence they could themselves create:

> P13 - *I am really bad at visualizing things. The latent variable modeling version of visualizing helps me to share what would otherwise be numeric or otherwise would be sort of abstract relationships among variables for people, but I'm really bad at visualizing things. Another thing, I use Venn diagrams for relations, but it almost always ends up being a function of arrows, circles, and squares. I have used workflows. In fact, now I created a workflow diagram to reflect a procedure, which was, so you can see the logic, how it flows? That's for the [client], but that's based on a qualitative analysis of how [characteristics] do and do not group together, so it's all qualitative data that we're analyzing in those figures.*

Statisticians find visualization is an indispensable tool for analysis, but also a tool for the essential last stage of their work process, communicating results. Ten analysts talked about visualization's ability to communicate stories:

> P12 - *Let me put it this way, I'm not going to put, well, I have and I will in the future, I'm sure. But you have your analysis, your P-value, your confidence intervals, your hypothesis test, what decision did you make, all of that. That's not what I'm going to put in a presentation typically. It's going to be the graph, and that would be either supporting it on the same page or in the backup. It [numeric representation of results] is important, I'm not trying to minimize the importance of it, but it's not what people understand when you're trying to communicate a result, typically -- it's -- unless you're in a room with statisticians.*

For some, visualization may be preferable to purely calculative methods. Five participants reported that visualization provides a more powerful evidentiary warrant for claiming findings than purely numeric methods:

> *P12 - Yeah. Okay, so I'll tell you what I tell every class I teach when we go through inference is, do the graphs first. And if the inference that you do doesn't support what you saw in the graph, something's terribly wrong. I see statistical inferences as kind of like it's absolutely important and it makes decisions consistent if we all are using the same inference. But it's a formalizing of the graph what you're seeing in the graph. I see two-sample tests, yes, we do them all the time. It's a way to make things consistent, make consistent decisions, but in the end it's supporting the graph really more than anything…*

Yet nearly as many (4) reported the opposite, that math is a greater evidentiary warrant, and gave a reason, namely, that visualization relies too much upon the judgement of the analyst. As participant 5 said, *"Yeah, you don't want to rely on your eyes. Having a number is better."*

Other strikes against visualization include the primacy of calculative methods in achieving publication. Two participants reported some version of being wary of aspects of p-value based statistical tests, but using them because publication required it:

> *P13 - That tradeoff is in my mind all the time. … A lot of the work that I do that's funded requires a T-test. A lot of the work that I do that is unfunded is geared towards coming up with something that reflects the outcome of interest for the scientist or science, but is compatible with a T-test, so it needs to be usable but also compatible with a T-test. For me, that's a frontloading of the design…*

Visualization can also result in a lot of work for the statistician, without necessarily earning them concomitant rewards. Intuitive design is critical for visualizations meant to communicate findings, but intuitive design tends to disappear

from view (Norman 2013). Thus, a statistician who works hours creating a display which their client can understand with a glance is, in effect, hiding their work. Labeled the "Designer's dilemma" in this research, it describes a situation where the more successful the work, the less it might be noticed (or appreciated) by its consumers:

> *P12 - Oh, nice, nice. So, you get where I'm going with that, right? I try to make it no chart junk, how about that? Or very little, minimal chart junk. But it's never just I sit down and I proceed down a linear path and get an analysis. I shouldn't say never. There are some that are very straightforward I can do that. But more complicated problems, it's not a linear process, it's a scattered process. And then you come back onto the line and you communicate. So, [clients] think maybe you hand them a graph and the analysis results and, "Oh, that probably took them 20 minutes to do that." Because if you just did that, it would take you 20 minutes. But they miss all the hours and hours of work that go into that final result.*

Thus, the relationship between statisticians and visualization is complex, despite data visualization apparently having been integrated into every phase of a statistician's workflow. Different practitioners give more or less emphasis to its use depending upon their inclinations, though all the participants in this study report using visualization for at least some phases of their work. They variously reported using process diagrams for planning analyses, visual exploratory data analysis for both data cleaning and hypothesis formation, specialized narrow-use visualizations during confirmatory analysis either as a reasonableness check or a source of primary findings, and, finally, communication of results. Further, most think about their analyses in visual terms, and find in visualization a powerful tool for supporting evidence-based results.

Participant 3 may provide the clearest example of the conflicted relationship statisticians appear to have with visualization. This person reported both a preference for calculative math over visualization, and for visualization over calculative math. On the one hand, in referring to one of the strawmen graphics:

> *P3 - ...So it seems like it's more subjective in a way. And I don't think that a new statistician or anyone who's been through the school that I went through, and maybe it's different now, has seen this often enough to make the best decisions with it.*

But on the other hand, when describing the process of understanding a hypothetical regression output during their workflow:

> *P3 - But then I see where's the plot? How does it look here? What's the shape? So all those things go on in my head. And I guess knowing all this stuff through school and then through work, I probably don't go in the right order when I do this stuff, but I just do it-*

5.5.2 Evidentiary Warrants

Figure 5.8 Validity Coding

10	**assume false**	p1	p4	p5	p6	p8	p11	p12	p13	p15	p17
7	**check assumptions**	p4	p5	p6	p8	p12	p13	p17			
5	**avoid rationalization**	p5	p8	p9	p13	p17					
5	**data model**	p1	p5	p6	p8	p13					
4	**parallel analyses**	p4	p8	p11	p12						
3	**check expectations**	p1	p4	p8							
3	**consider mistakes**	p6	p12	p15							
3	**sample size**	p4	p5	p9							
3	**SM experts**	p5	p6	p12							
3	**trust data**	p6	p8	p11							
2	**admit uncertainty**	p6	p8								
1	**close study**	p8									
1	**consider luck**	p6									
1	**distrust eyes**	p4									
1	**narratives convince**	p4									

If visualization acts as interface between human mind and computer brain in the model of extended cognition (Scaife 1996), then the resulting compound system hinges upon human judgement – human visual intuitions, human visual acuity, human subject matter pre-knowledge, and human imagination. It is thus subject to human biases, mental blind spots, and the various shortcomings of our eyes, as pointed out by 5/16 participants (4 describing viz as too subjective, and a fifth making much the same point when suggesting reliance upon visual systems can reduce analytic validity).

By contrast, traditional, equation-based statistical tests fully externalize the critical decision point, ostensibly removing the human element. Provided results criteria are selected in advance (as pointed out by 6/16 participants), and all test assumptions are met (3/16 participants), results from statistical tests FEEL more

objective. Five of 16 participants described relying upon statistical tests to provide evidentiary warrants for their findings, despite nearly as many warning that p values encourage bad thinking (4/16).

For example, participant 15 expressed both ideas. On the one hand, statistical tests encourage dichotomous thinking about situations which can be more textured:

> *P15 - …aspects of understanding uncertainty. And there's nothing wrong with frequentist methods [such as parametric statistical tests], but it's kind of black and white.*

Yet at the same time:

> *P15 - …[valid statistical testing] tells me whether my results should be something that are actionable, meaningful to someone…the actual [sample] estimates themselves being different does not necessarily mean that things truly are different out in the population.*

This reliance upon calculative statistical tests, and distrust of visualization, appears to conflict with visualization's wide use as an analytic method by participants; it is in direct conflict with participants expressions that visualization is an important check on the statistical tests themselves.

5.5.3 A Resolution of Apparent Conflicts

A central thread may explain this seeming conflict: experienced statisticians argue against their own judgements wherever they can, constantly seeking to verify their findings with multiple independent methods. In short, it is not that statisticians trust this or that method and distrust another – they distrust them all, or more precisely, never trust any one method by itself.

10 of 16 participants expressed the idea that they should always distrust their own findings. For example:

> *P6 - So, you want to use simulation to test that and also compare to the previous methods and say if your new method indeed works better for this case. And then you apply your methods to the data and try to answer the question, an original question you had. And at this time you also need to talk with your collaborators and say, because they are the people who understand science, understand the problem. And then they will help you to evaluate whether the results make sense or not, and how to interpret that and what kind of new discoveries. Maybe sometime there are some surprising results for them. And then you want to analyze whether it is because something that there is a bug in your method, there's a bug in your code or something like that. If not, and then that's more interesting...So, that's the usual process.*

Or:

> *P12 - There's always a reason. There's always, …It sometimes will come down to sample size or the overall variation if it's a two-sample test of maybe one group has larger variation than the other. But mathematically you can go dig into that. There's always a reason. That's in my head, maybe I'm wrong, but in my head there's always a reason for that.*

When numbers would seem to support their favored hypotheses, statisticians look to visualization to see whether unknown outliers, gaps, or the like, explain away their findings. But if they see a pattern in a visualization, they look to calculative methods to act as a check on their eyes. Presented with an apparent difference between sample means, they test whether there is any likelihood that random chance can explain that away. They check their assumptions (7/16 participants) and conduct parallel analyses (4/16). With already checked and tested findings in hand, they ask subject matter experts whether the results make sense to them (3/16 participants).

In each of these cases, the statisticians are seeking any way they can to undercut apparent findings. They pit one analytic method against another, and all must line up for the experienced analyst to accept a finding as probably, or even possibly, true.

This mindset has strong implications for the design of new statistical tools.

5.5.4 Connecting to Reality: A Preference for Effect Sizes over Math

Among the most frequently expressed understanding captured during this study is the constant effort by experienced statisticians to link their results back to the reality they are meant to describe. Fourteen of 16 participants discussed this during their interviews. For example:

> *P9 - ... if necessary, go back and revise the statistical analysis plan in light of the reality of the situation as we've discovered it.*
>
> *P11 - ... I think part of statistics is, in more a meta general sense, is you should trust the data and not come in with those...strong priors can be a problem.*
>
> *P17 - And I find generally data will tell the truth.*

Five participants talked about the validity of their results hinging upon the quality of their data model, that is, a clear and well understood link between their measures and their phenomena of interest (Figure 5.8). Every statistical model is a simplification of reality built upon abstractions of phenomena, so maintaining a strong link between data and study subject can be non-trivial. Indeed, seven

participants described how thinking through facts-on-the-ground was key to their being able to understand the math they had to use for their analyses. For example:

> *P1 - Well, that's the thing, is I don't understand statistics. I will freely admit I don't understand it, purely mathematically. I understand it if I'm looking at something tangible that has meaning to me…*

It appears that a key outgrowth of this reality-focus is a preference among many statisticians for privileging effect sizes over p-values. Nine of 16 participants expressed this idea.

Effect sizes are about the practical significance of a finding, rather than statistical (or probabilistic) significance. In the current climate of concern over p-value statistical testing, focusing on effect sizes is one potential answer. But apart from these statistical concerns, focusing on effect sizes necessarily means focusing on the reality of the study subject.

5.5.5 A preference for Confidence Intervals – Uncertainty Representation as an Analog to Physical Measurement Error

Nine of the participants expressed their understanding of a t-test as some version of a comparison of overlapping confidence intervals, i.e. sample means with some indicator of variability. This approach emphasizes the samples themselves, rather than the unseen population they are meant to measure. The metaphor of the confidence interval is that of a measurement tool with an allowance for error – a very physical conception, different in kind from the mathematical abstraction of a probability distribution.

The t-test and p-value do focus on the population, by posing the likelihood that the population looks a certain way given the samples. Yet the majority of participants in this study keep their focus on the samples, while asking whether their results achieve statistical significance, and thus provide an evidentiary warrant to make statements about the population. The practical result is logically equivalent, but speaks to the difficulty of statistical inference. In many cases, even the experts don't fully embrace the meaning of the mathematics they rely upon, instead mentally falling back on simpler heuristics.

If the experts face challenges with statistical inference (as 8 of 16 participants reported), novices must face greater challenges.

5.5.6 Design Recommendations

Design recommendations from the work in this chapter emerge both from its findings on statisticians' tacit understandings of visualization & inferential statistics, and from participant reactions to the strawman graphics. These recommendations, along with those from prior chapters, form a unified set of recommendations for potential visual tools of statistical inference, and form the bulk of chapter 6.

Chapter 6: Conclusion – A Prototype Inferential Statistics Visualization

6.1The Key Affordance of Purely Calculative Statistics Tests – Blindness

Traditional statistics tests provide users with the appearance of objectivity because the calculations show no details of the data. This pairs with reliance upon probability distributions (such as the Student's T, the Chi-Square family of distributions, the F-family, and others) of such complexity that memorizing them presents a challenge to the most experienced data analyst. In a very real sense, this lack of detail and complexity make a traditional statistical test blind, that is, something outside the user's ability to monitor or control. This blindness pushes analysis toward the objective; answers come out of the presumably impartial and objective machinery of the test itself.

This is an affordance visualization-based methods can't match. Visualizations are the opposite of blind. Even aggregate visualizations, such as box plots or normal distribution curves, rely ultimately upon the judgement of the user to interpret (derive meaning from) the analytic display.

Yet blindness also comes with a cost. A blind test can become a black box, inviting users to accept resulting answers without fully understanding them. This may be related to the wide misunderstanding among published authors of the assumptions of statistical tests they themselves have used (Hoekstra 2014), and, perhaps, has contributed to the current, "Replication Crisis" (Cumming 2014).

Such black box thinking may contribute to the "cliff effect" described by Helske et. al. (Helske et. al. 2021): the tendency for researchers to interpret small differences near the p=.05 line as large differences in the confidence they should have in their experimental results. Researchers following this kind of thinking look at the output of a statistical test as a declaration of experimental success or failure, with essentially nothing uncertain on either side of the line. Helske and co-authors found that switching to more detailed visual forms helped to a alleviate this effect. Though the particular forms they chose for their tests (violin plots) resulted in some confusion among their test participants, overall, the authors found that the negative effects were offset by improved statistical understanding.

Helske's work suggests that visual presentations of inferential statistics can make a difference in user understanding. This research has attempted to identify design features for visualizations of inferential statistics which would produce the right understanding in users, particularly novice users.

6.2 Designing Recommendations for Inferential Visualization

Table 6.1 summarizes the most relevant findings from studies 1 & 2. These combine with the findings from Chapter 5 to produce the following list of design recommendations.

Table 6.1

Summary of Relevant Findings/Observations

Experiment 1	Experiment 2
1 People have intuitions for pairing detailed sample distributions with the idealized distributions we often use to summarize them.	1 People have the ability to judge the amount of overlap between a pair of distributions, and relate it to a question of statistical likelihood.
2 Their intuitions are biased towards a too-broad spread ("Umbrella effect").	2 This ability varies with graph type.
3 People fit curves more accurately to less detailed visualizations -- the box plot for example, which provides a guide post to the mean	3 In general, people perform better at this task when using LESS detailed, more abstract graphic forms (the normal curve itself worked best).
4 The two histograms had nearly identical results, therefore, even within the detailed graphics -- the two histograms -- people didn't use the additional detail of the dot version.	4 A minimum of training is enough to allow people to perform this task.
5 The shape of the dataset being fit matters. This suggests further studies.	5 People perform better when judging the overlap for distributions with larger deviations from p=.05, either high or low, and perform worse on borderline cases.
	6 People treated this as a dichotomous task, despite the graphic approach and the answer format. This may reflect the framing of the question, or their prior experience with statistics, or a natural inclination.

6.2.1 Validity of visual analysis – in parallel with calculative methods

Given statisticians' use of visualization in every aspect of their analytic process, it appears that, as a general approach, visualizing inferential statistics may be acceptable to the statistical community. However, it also appears that no single approach to analysis will have the confidence of that community. Rather, **visualization should be paired with the equivalent numeric output**. This will provide the analyst both an intuitive understanding of the data (visual), and the more 'objective' numbers (where the decision point is determined by the calculation rather than the user's eye).

6.2.2 A Multi-facet Display

Nine of 17 participants began the graphic elicitation portion of the interview thinking of statistical tests as some version of overlapping confidence intervals (equivalent to strawman #3 or #1). However, once shown the Strawman graphics, 10 of 18 participants preferred strawman #2, which considers the distribution of some test statistic. The latter more directly represents what a statistical test actually does, while the former is a mere heuristic. This mixed result supports use of both graphic types. In combination with the participants' multi-faceted approach to analysis, it may suggest that a tool for novice analysts **should present multiple facets of the statistical testing process in a compound display**. This recommendation echoes findings by Shneiderman, et. al. in their work on data exploration, in which they found that several simple graphics, each presenting a different facet of a dataset, and linked by interaction, could work better than a single, complex visualization.

A multi-display provides space for including graphic representations along with numeric results, and for providing background information, such as raw distributions of samples with curve fits, to check testing assumptions. Figure 6.1 A-C provides mock-ups of such a display ("Inferential Quartet").

Figure 6.1 A

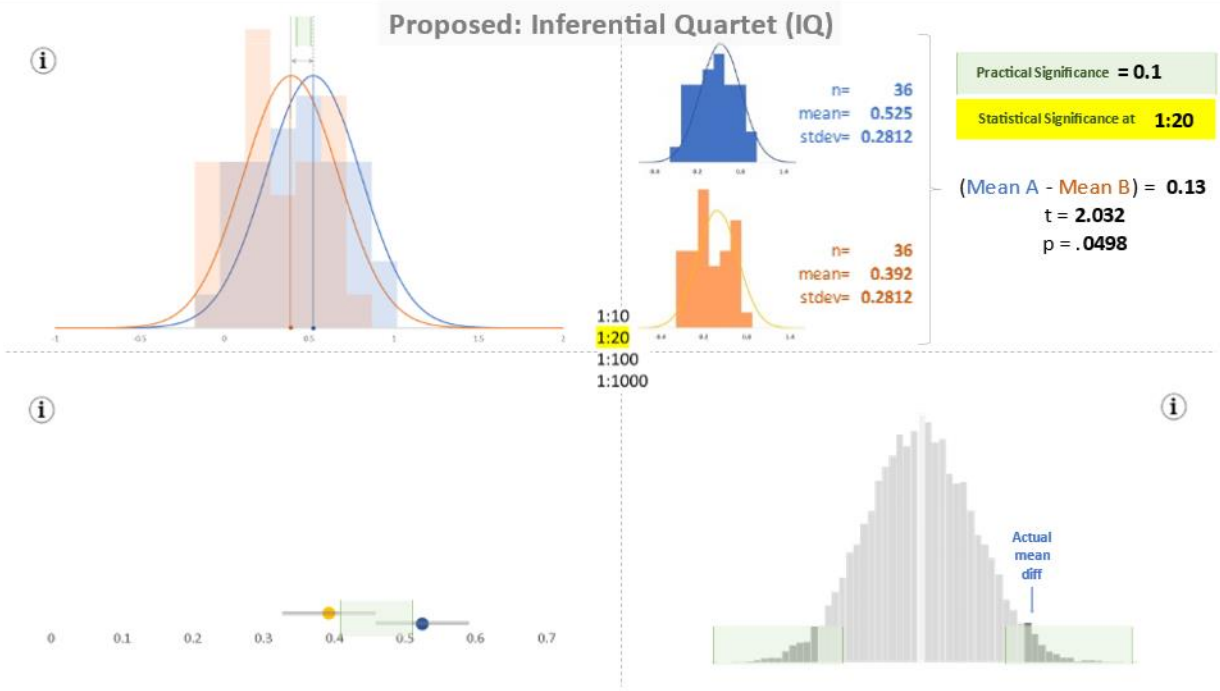


Figure 6.1 B

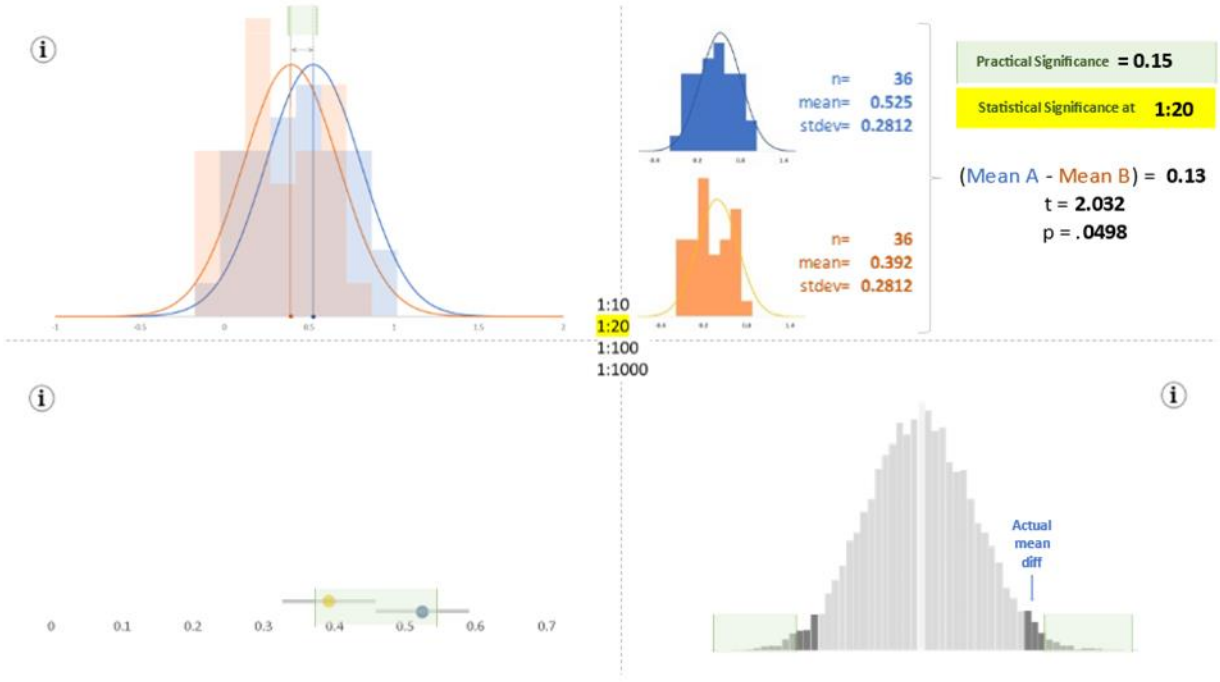
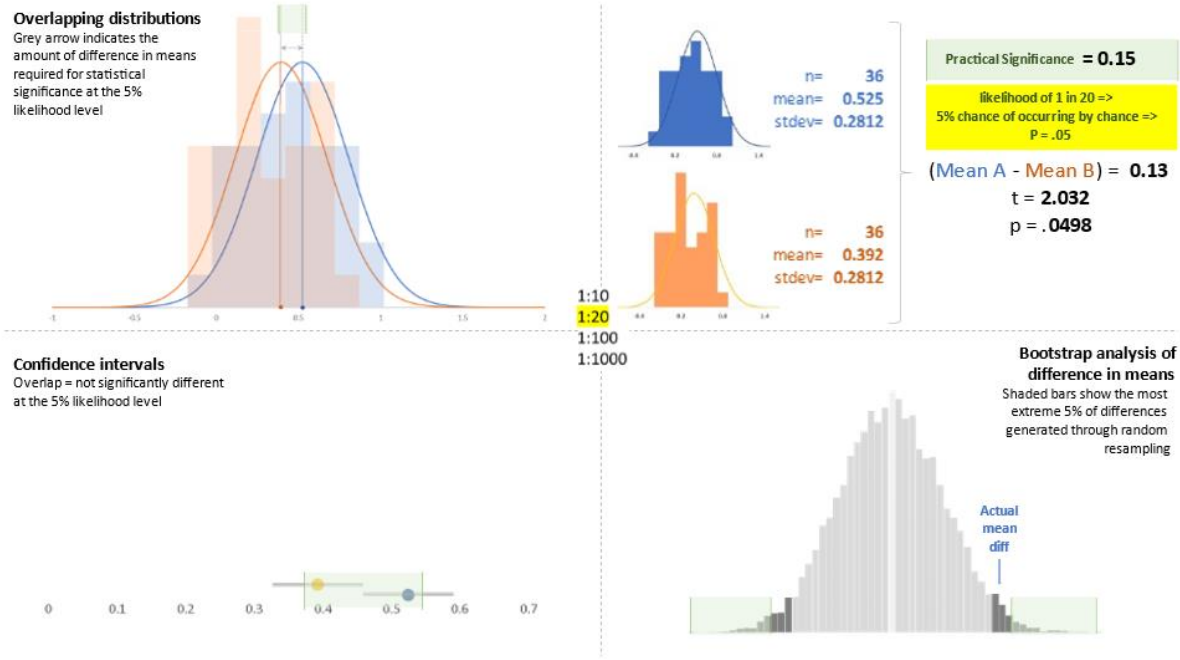


Figure 6.1 C



6.2.3 Keep an Eye on Reality – Include Effect Sizes

Following participants frequent preference for effect sizes over p values, and their emphasis on keeping the reality of the study situation in focus, **visual inferential tools for novices should include an indicator of analytically meaningful effect size**. It should be declared in advance of seeing the data, just like Alpha (minimum acceptable p-value) to avoid post-hoc rationalization. Selecting effect sizes requires users to understand their data, usually through contact with subject matter experts. It automatically combats dichotomous thinking, as having two measures to choose (alpha and effect size) turns the significance decision multi-dimensional.

A simple version of an effect-size visualization has been included in Fig 6.1A-C.

The effect size measure provides different information in each side of the quartet. In the confidence interval displays, it provides information similar to Cohen's D. In the test statistic distribution, it provides a rough guide to the statistical power represented by the samples in hand.

6.2.4 Re-size confidence intervals for readability

Most people (including many participants in this study) read common CIs (95%) as indicating statistical significance if they don't overlap, but in fact, they can overlap considerably and still show significance at the traditional p=.05 level. Therefore, we recommend **resizing the confidence intervals such that a pair which don't overlap can be directly read as representing a statistically significant difference at the chosen p-value.**

Confidence intervals have traditional been built upon arbitrary confidence limits, such as 95%. This is convenient for the analyst as they calculate CIs, as it obviates the need for them to explicitly consider degrees of freedom (which p-values for t-tests are linked to). Creating a 95% confidence interval is as simple as multiplying the standard error by 1.645, and adding and subtracting the resulting margin of error from the relevant statistic (such as the mean) (Table 6.2).

Table 6.2
Sizing Confidence Interval to Test P-value Testing

		SE Multiplier	CI	
P value of two-tailed t-test w/70 d.f.	p = .10	**1.181**	76.2%	
	p = .05	**1.411**	84.2%	
	p = .01	**1.874**	93.9%	
	p = .001	**2.438**	98.5%	
	NA	**1.645**	90.0%	Typical CIs
	NA	**1.960**	95.0%	
	NA	**2.330**	98.0%	

Confidence intervals linked to a p-value from a particular statistical test must be individually calculated, however, the calculations can be automated based upon degrees of freedom.

The confidence interval display in Fig. 6.1A-C has been so modified.

6.2.5 Favoring Familiar Formula? Bootstrapping in the Inferential Quartet

Strawman #2, while favored by many participants in the third study, did raise concerns in a few of them due to its resembling an established method of inference – "Bootstrapping" – while departing from that method in execution. Strawman #2 is a repeated sampling from a simulate population, rather than a resampling method in the strict sense. The simulation approach more closely resembles the assumptions of the null hypothesis from a t-test. It produces results very similar to bootstrapping, but the variance from established procedure bothered some participants.

Referring back to coding results, many participants expressed distrust for their recollections for the details of some of their analyses, leaning on published resources and their peer collaborators. It may not be surprising, therefore, that a variance from established procedure would cause concern. The suggestion is that professional statisticians prefer to follow established analytic procedures where possible. This is only a suggestion, as this was not a dimension explicitly tracked during thematic coding.

The probability distribution panel in the quartet in Fig. 6.1 A-C reflects formal bootstrapping, rather than the simulated approach taking in the strawmen graphics. Given the similarity in the final graphic forms of the two approaches, it seems unlikely this will make an important difference for novices. However, **building tools for visual inference using familiar analytic components wherever possible** may increase the likelihood experienced statisticians will trust the results a novice analyst shares with them having used such a tool.

Note that while the difference between formal bootstrapping and the simulation method from strawman #2 resulted in only minor differences given the symmetrical probability distribution under study, future work, particularly on asymmetrical probability distributions, should consider explore this difference more thoroughly.

6.2.5 Extending Designs to Other Statistical Tests

The inferential quartet prototype developed in this chapter presents a two-tailed, two sample t-test. However, the design principles used to create likely apply to other tests. Table 6.3 lists out suggestions for design changes which would make this quartet approach applicable to other common statistics tests.

Table 6.3

Suggested Design Variations on the Inferential Quartet

		Panel A (Overlapping histograms)	Panel B (Overlapping confidence intervals	Panel C (Bootstrap)	Panel D (Numeric panel)
Symmetrical test distributions	One Sample vs. standard value test	Replace sample distribution with indicator of standard value (such as a dot on the X axis)	Remove whiskers from first confidence interval, leaving dot to indicate standard value	No change required	Replace first distribution graphic/descriptive statistics with definition of standard value
	One tailed tests	No design changes required	No design changes required	Shading only one side of distribution	Define test as single-tailed positive or negative
	Z test	No design changes required	No design changes required	No design changes required	No design changes required
Asymmetrical test distributions	Chi-square Goodness of Fit test	Eliminate normal curves, replace first histogram with expected distribution	Overlap confidence interval for each category with expected value from that category	Distribution will be Chi-square	Remove normal curves from histograms, replace appropriate statistics
	Chi-square test of independence	Eliminate normal curves	Overlap confidence intervals -- one per category	Distribution will be Chi-square	Remove normal curves from histograms, replace appropriate statistics
	F-test of Variance	Center both distributions on a single central point. Grey separation arrow now indicates horizontal gap between distribution spreads at vertical midpoint of graph	No design changes required	Distribution will be F-distribution	Replace appropriate statistics, remove mean indicator from graphics, alight graphics one above the other on a center line

As illustrated by this table, some design features of the quartet are specific to t-tests, while others are more generalizable. A call for parallel analyses, the use of both numeric and visual displays in tandem, and the inclusion of effect sizes, all have broad application to inferential statistics. Including confidence intervals as a visual metaphor for physical measurement, and sizing those intervals to represent the test in

question rather than an arbitrary confidence level, is broadly applicable. Providing raw sample distributions in the same display as summary statistics, the test, and the probability distribution the test is based upon, is also broadly applicable.

I believe the inferential quartet represents a model of a visual tool for inferential statistics, one based upon several principles that are generalizable beyond the t-test prototype of Fig. 6A-C.

6.3 Limitations and Future Research

6.3.1 Limitations of Study 3

Study 3 involved a total of 18 professional statisticians, but while we took care to choose participants from many different fields, educational backgrounds, and demographics, there are certainly several threats to generalizing these results too widely. Given the qualitative and highly personal nature of these practices, it can be hard to draw conclusive findings from this work. Furthermore, as discussed in my positionality statement (5.1.3), as a researcher working with collaborators, we represent only a small fraction of the worldwide statistician population. While design probes have been proven effective because they provide a common ground for discussion (Wallace 2013), which can be helpful, they may also constrain ideation. Furthermore, the strawman graphics were not radical or even particularly novel, and perhaps a more radical set of design ideas could have sparked more innovation. However, the interview protocol took care to ask for graphic elicitation (Phase II) prior to showing the strawmen (Phase III) to avoid biasing participants. The purpose

of this study, and thus the strawmen, was mostly to understand the mindsets and practices of professional statisticians, and more effort will be needed to develop prototype graphics in the future.

6.3.2 Limitations of this Research – Parametric Statistics

This work has been largely focused on parametric inferential statistics. Statistical inference is obviously a much larger field, and includes topics such as non-parametric and Bayesian methods. This may limit the generality of design recommendations and suggests avenues for future research.

For example, the graphics developed during Study 2 and Study 3 apply to a two-sample t-test – a parametric statistical method. Even the bootstrapping of Panel C in the quartet – a hypothetically non-parametric method – nonetheless employs parametric assumptions of normality.

It may be that the design recommendations resulting from this research apply best to parametric methods. This would be especially true if non-parametric methods, even those like Bootstrapping which include some parametric assumptions, include cognitive or calculative aspects which prove less amenable to visual representation.

This limitation suggests at least one direction for further research, namely, that of identifying and testing visualizations for non-parametric methods.

6.3.3 Limitations of this Research – Symmetry

In addition to being parametric, the visualizations used as prompts during all three studies also reflect symmetrical distribution curves, either normal curves or

those of the t-distribution. Thus, these results can not be directly generalized to tests which rely upon non-symmetrical distributions, such as the Chi-Square or F-distributions. As argued above, some design principles may carry over, such as the use of multi-faceted displays showing parallel analyses and raw data distributions, the pairing of visualization with numeric output, and the recommendation to size confidence intervals to match statistical significance rather than arbitrary confidence levels. However, further experimentation should be undertaken to establish whether the human eye-brain system comes equipped with visual intuitions for linking asymmetric distributions with the data they reflect.

6.4 Suggestions for Further Research

The limitations of these studies suggest future research already mentioned above.

Further studies should test whether people have the visual intuitions to connect asymmetric probability distributions with underlying raw data.

More generally, further work should test whether the design recommendations put forward in this dissertation can be adapted to non-parametric statistical tests.

The strawmen graphics used during study 3 generated highly informative reactions from participants. However, they also represent a limitation on creating ideation. The next logical step would be to revisit a group of similarly experienced statisticians with more radical prototypes, especially those based upon non-parametric methods.

But perhaps the most immediate recommendation is for testing versions of the inferential quartet with novice users. The quartet embodies the tacit understanding of inferential statistic held by experienced statisticians, and studies 1 and 2 suggest that novices have the visual intuitions to take advantage of the elements within the quartet. However, in practice, there may be interactions with those elements sharing the same screen which strike novice users differently than expected. The next study would test this open question.

6.5 Developing the Quartet Design for Field Use by Novice Analysts

As described in this chapter, the present design of the quartet reflects the limitations of the research that went into its creation. It is best thought of as a prototype at this stage, a model for creating a variety of inferential statistics visualizations that might gain acceptance in the broader statistical community while also providing an on-ramp to inferential methods for novice analysts. Developing the prototype into something with maximum practical utility for novice users will require further design research, particularly focused on adding features that allow for contextualizing the analysis in a specific domain.

Many (7/16) of the experienced statisticians in this study rely upon context clues to understand the mathematics of their analyses. They find they best understand particular analytic methods by considering those methods in the more concrete terms offered by specific examples of phenomena under study, rather than as abstract variables in a mathematical procedure. It is reasonable to expect that many novices would benefit from similar context focus (it is also common wisdom that concrete

examples are useful in learning). Additionally, many novices first approaching data analysis may only do so because of an interest in a specific subject matter amenable to data-focused research, rather than seeking to understand statistical methods for their own sake.

Thus, future design work on the quartet will benefit from, and likely requires, partnerships between researchers who possess a knowledge of data visualization methods, and those with specific subject matter expertise in a variety of fields which may be of interest to groups of novice analysts. Experts and researchers from several disciplines working in concert should be able to create, and test, fully contextualized versions of the quartet in a variety of settings.

For example, a potential application for a fully contextualized quartet might be its use in high school science classrooms, providing students the ability to make discoveries with sample data prior to their acquiring extensive statistical training. However, designing for this population would require the collaboration of science educators, curriculum authors, data visualization and design specialists, school administrators, and even the students themselves. Conducting the research would require the buy-in of parents and local school officials.

Should several such efforts prove successful, a review of the contextualized quartets thus produced might provide additional insights, as it may uncover generalizable principles of how design features of such statistical graphics can be adjusted to maximize the connection for users between the math and the reality under study.

Appendix

R Code

```
#create 10,000 pairs of samples with a stdev of .2821 and mean of .458597 and n of 36
#then take the differece of each pair and graph as histogram


d <- mean(rnorm(36, mean=.458597,sd=.2821)) - mean(rnorm(36,
mean=.458597,sd=.2821))

x<-0
diffs<-NULL
repeat{
d <- mean(rnorm(36, mean=.458597,sd=.2821)) - mean(rnorm(36,
mean=.458597,sd=.2821))
diffs<-append(diffs,d,after=length(diffs))
x=1+x
if(x==10000){break}
}
x
hist(diffs,breaks=50)
```

Bibliography

[1]
C. Ahlberg, C. Williamson, and B. Shneiderman, "Dynamic queries for information exploration: An implementation and evaluation," in *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, 1992, pp. 619–626.

[2]
W. Aigner, A. Rind, and S. Hoffmann, "Comparative evaluation of an interactive time-series visualization that combines quantitative data with qualitative abstractions," in *Computer Graphics Forum*, 2012, vol. 31, no. 3pt2, pp. 995–1004.

[3]
S. Ainsworth, V. Prain, and R. Tytler, "Drawing to Learn in Science," *Science*, vol. 333, no. 6046, pp. 1096–1097, Aug. 2011, doi: 10.1126/science.1204153.

[4]
D. Albers, M. Correll, and M. Gleicher, "Task-Driven Evaluation of Aggregation in Time Series Visualization," *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, vol. 2014, pp. 551–560, 2014, doi: 10.1145/2556288.2557200.

[5]
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, Apr. 2010, doi: 10.1177/0956797610363543.

[6]
G. A. Alvarez, "Representing multiple objects as an ensemble enhances visual cognition," *Trends in Cognitive Sciences*, vol. 15, no. 3, pp. 122–131, 2011.

[7]
F. J. Anscombe, "Graphs in statistical analysis," *The American Statistician*, vol. 27, no. 1, pp. 17–21, 1973.

[8]
L. Argote, "Organizational Learning: Creating, Retaining, and Transferring Knowledge." Kluwer Academic Publishers, 1999.

[9]
D. Ariely, "Seeing Sets: Representation by Statistical Properties," *Psychological Science*, vol. 12, no. 2, pp. 157–162, Mar. 2001, doi: 10.1111/1467-9280.00327.

[10]
H. Arkin and R. R. Colton, *An outline of statistical methods, as applied to economics, business, education, social and physical sciences, etc*. Barnes & Noble, 1938.

[11]
P. W. Atkins, "The limitless power of science," *Nature's Imagination: The Frontiers of Scientific Vision*, pp. 122–132, 1995.

[12]
L. J. Bain and M. Englehardt, "Introduction to Probability and Mathematical Statistics, 1987," *Duxbury Press*.

[13]
L. Bartram, M. Correll, and M. Tory, "Untidy Data: The Unreasonable Effectiveness of Tables," arXiv, arXiv:2106.15005, Jun. 2021. doi: 10.48550/arXiv.2106.15005.

[14]
M. J. Bates, “The invisible substrate of information science,” *Journal of the American Society for Information Science*, vol. 50, no. 12, pp. 1043–1050, 1999.

[15]
M. I. Bauer and P. N. Johnson-Laird, “How Diagrams Can Improve Reasoning,” *Psychol Sci*, vol. 4, no. 6, pp. 372–378, Nov. 1993, doi: 10.1111/j.1467-9280.1993.tb00584.x.

[16]
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, Jan. 2017, doi: 10.1109/TVCG.2016.2598862.

[17]
S. Belia, F. Fidler, J. Williams, and G. Cumming, “Researchers Misunderstand Confidence Intervals and Standard Error Bars.,” *Psychological Methods*, vol. 10, no. 4, pp. 389–396, 2005, doi: 10.1037/1082-989X.10.4.389.

[18]
W. Benjamin, “Extracts from the work of art in the age of mechanical reproduction,” *The Photography Reader*, pp. 42–52, 2003.

[19]
L. Besançon and P. Dragicevic, “The continued prevalence of dichotomous inferences at CHI,” in *Extended Abstracts of the ACM CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–11.

[20]
T. J. Bing and E. F. Redish, “Analyzing problem solving using math in physics: Epistemological framing via warrants,” *Physical Review Special Topics - Physics Education Research*, vol. 5, no. 2, Dec. 2009, doi: 10.1103/PhysRevSTPER.5.020108.

[21]
S. Blenkinsop, P. Fisher, L. Bastin, and J. Wood, “Evaluating the Perception of Uncertainty in Alternative Visualization Strategies,” *Cartographica: The International Journal for Geographic Information and Geovisualization*, Sep. 2006, doi: 10.3138/3645-4V22-0M23-3T52.

[22]
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, 2011.

[23]
N. Boukhelifa, A. Bezerianos, T. Isenberg, and J. Fekete, “Evaluating Sketchiness as a Visual Variable for the Depiction of Qualitative Uncertainty,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 12, pp. 2769–2778, Dec. 2012, doi: 10.1109/TVCG.2012.220.

[24]
G. C. Bowker and S. L. Star, *Sorting Things Out: Classification and its Consequences*. MIT press, 2000.

[25]
S. Bresciani and M. J. Eppler, “The risks of visualization,” *Identität und Vielfalt der Kommunikations-wissenschaft (2009)*, pp. 165–178, 2009.

[26]
J. K. Brueckner, P. S. Calem, and L. I. Nakamura, “Subprime mortgages and the housing bubble,” *Journal of Urban Economics*, vol. 71, no. 2, pp. 230–243, Mar. 2012, doi: 10.1016/j.jue.2011.09.002.

[27]
B. Bryson, *A Short History of Nearly Everything*. Broadway Books, 2004.

[28]
A. Buja *et al.*, “Statistical inference for exploratory data analysis and model diagnostics,” *Phil. Trans. R. Soc. A.*, vol. 367, no. 1906, pp. 4361–4383, Nov. 2009, doi: 10.1098/rsta.2009.0120.

[29]
J. Burrell and K. Toyama, “What Constitutes Good ICTD Research?,” *Information Technologies & International Development*, vol. 5, no. 3, Art. no. 3, Oct. 2009, Accessed: May 09, 2020. [Online]. Available: https://itidjournal.org/index.php/itid/article/view/382

[30]
A. Cairo, *The Functional Art: An introduction to information graphics and visualization*. New Riders, 2012.

[31]
A. Cairo, *The Truthful Art: Data, Charts, and Maps for Communication*. New Riders, 2016.

[32]
M. Card, *Readings in Information Visualization: Using Vision to Think*. Morgan Kaufmann, 1999.

[33]
R. Chang, C. Ziemkiewicz, T. M. Green, and W. Ribarsky, “Defining insight for visual analytics,” *IEEE Computer Graphics and Applications*, 2009.

[34]
Y. Chen, A. Johri, and H. Rangwala, “Running out of STEM: a comparative study across STEM majors of college students at-risk of dropping out early,” in *Proceedings of the International Conference on Learning Analytics and Knowledge*, New York, NY, USA, Mar. 2018, pp. 270–279. doi: 10.1145/3170358.3170410.

[35]
I. K. Choi, T. Childers, N. K. Raveendranath, S. Mishra, K. Harris, and K. Reda, “Concept-Driven Visual Analytics: an Exploratory Study of Model- and Hypothesis-Based Reasoning with Visualizations,” in *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems*,,, pp. 1–14. doi: 10.1145/3290605.3300298.

[36]
S. C. Chong and A. Treisman, “Representation of statistical properties,” *Vision Research*, vol. 43, no. 4, pp. 393–404, 2003.

[37]
L. Ciccione and S. Dehaene, “Can humans perform mental regression on a graph? Accuracy and bias in the perception of scatterplots,” *Cognitive Psychology*, vol. 128, p. 101406, Aug. 2021, doi: 10.1016/j.cogpsych.2021.101406.

[38]
A. Clark and P. of P. and D. of P. N. P. P. A. Clark, *Supersizing the Mind: Embodiment, Action, and Cognitive Extension*. Oxford University Press, USA, 2008.

[39]
Clement, John. "Misconceptions in graphing." Proceedings of the ninth international conference for the psychology of mathematics education. Vol. 1. Utrecht,, The Netherlands: Utrecht University, 1985.

[40]
W. S. Cleveland, *Visualizing Data*. Hobart Press, 1993.

[41]
W. S. Cleveland and R. McGill, “Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods,” *Journal of the American Statistical Association*, vol. 79, no. 387, pp. 531–554, Sep. 1984, doi: 10.1080/01621459.1984.10478080.

[42]
M. A. Correll, “Improving Visual Statistics,” PhD Thesis, The University of Wisconsin-Madison, 2015.

[43]
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*, Austin, Texas, USA, 2012, p. 1095. doi: 10.1145/2207676.2208556.

[44]
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, Dec. 2014, doi: 10.1109/TVCG.2014.2346298.

[45]
M. Correll and J. Heer, “Regression by eye: Estimating trends in bivariate visualizations,” in *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems*, 2017, pp. 1387–1396.

[46]
M. Correll, M. Li, G. 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, Jan. 2019, doi: 10.1109/TVCG.2018.2864907.

[47]
L. Cosmides and J. Tooby, “Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty,” *Cognition*, vol. 58, no. 1, pp. 1–73, Jan. 1996, doi: 10.1016/0010-0277(95)00664-8.

[48]
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.

[49]
G. Cumming, “The New Statistics: Why and How,” *Psychological Science*, vol. 25, no. 1, pp. 7–29, Jan. 2014, doi: 10.1177/0956797613504966.

[50]
G. Cumming and S. Finch, “Inference by Eye: Confidence Intervals and How to Read Pictures of Data.,” *American Psychologist*, vol. 60, no. 2, pp. 170–180, 2005, doi: 10.1037/0003-066X.60.2.170.

[51]
R. Davidson and J. G. MacKinnon, “Graphical Methods for Investigating the Size and Power of Hypothesis Tests,” *Manchester School*, vol. 66, no. 1, pp. 1–26, Jan. 1998, doi: 10.1111/1467-9957.00086.

[52]
E. Dimara, S. Franconeri, C. Plaisant, A. Bezerianos, and P. Dragicevic, “A Task-Based Taxonomy of Cognitive Biases for Information Visualization,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 2, pp. 1413–1432, Feb. 2020, doi: 10.1109/TVCG.2018.2872577.

[53]
A. A. diSessa and P. Cobb, “Ontological Innovation and the Role of Theory in Design Experiments,” *The Journal of the Learning Sciences*, vol. 13, no. 1, pp. 77–103, 2004, Accessed: May 09, 2020. [Online]. Available: https://www.jstor.org/stable/1466933

[54]
P. Dragicevic, “Fair statistical communication in HCI,” in *Modern statistical methods for HCI*, Springer, 2016, pp. 291–330.

[55]
P. Duguid, “The ageing of information: From Particular to particulate,” *Journal of the History of Ideas*, vol. 76, no. 3, pp. 347–368, 2015.

[56]
J. Eatwell, M. Milgate, and P. Newman, *The New Palgrave: Utility and Probability*. W. W. Norton & Company, 1990.

[57]
F. Echtler and M. Häußler, “Open source, open science, and the replication crisis in HCI,” in *Extended Abstracts of the ACM CHI Conference on Human Factors in Computing Systems*, 2018, pp. 1–8.

[58]
B. Efron, “Bootstrap methods: another look at the jackknife,” in *Breakthroughs in Statistics*, Springer, 1992, pp. 569–593.

[59]
B. Efron and R. Tibshirani, *An Introduction to the Bootstrap*. New York: Chapman & Hall, 1993.

[60]
L. Endrenyl, “Statistics. by David Freedman, Robert Pisani, Roger Purves. WW Norton & Co., Inc., New York, 1978. xv+ 506+ A83 pp., US $13.95. ISBN 0-393-09076-0.[Instructor’s Manual, 135 pp. ISBN 0-393-09041-8.].” Wiley Online Library, 1978.

[61]
S. H. Erlwanger, “BENNY’S CONCEPTION OF RULES AND ANSWERS IN LPI MATHEMATICS’,” p. 21, 1973.

[62]
D. Eyisi, “The Usefulness of Qualitative and Quantitative Approaches and Methods in Researching Problem-Solving Ability in Science Education Curriculum,” *Journal of Education and Practice*, vol. 7, no. 15, pp. 91–100, 2016, Accessed: May 09, 2020. [Online]. Available: https://eric.ed.gov/?id=EJ1103224

[63]
J.-D. Fekete, J. J. van Wijk, J. T. Stasko, and C. North, “The Value of Information Visualization,” in *Information Visualization: Human-Centered Issues and Perspectives*, A. Kerren, J. T. Stasko, J.-D. Fekete, and C. North, Eds. Berlin, Heidelberg: Springer, 2008, pp. 1–18. doi: 10.1007/978-3-540-70956-5_1.

[64]
S. Few, *Now You See It: Simple Visualization Techniques for Quantitative Analysis*, 1st ed. USA: Analytics Press, 2009.

[65]
M. A. Fisherkeller, J. H. Friedman, and J. W. Tukey, “PRIM-9: An interactive multi-dimensional data display and analysis system,” 1988.

[66]
R. Forrester, “Empowerment: Rejuvenating a Potent Idea,” *The Academy of Management Executive (1993-2005)*, vol. 14, no. 3, pp. 67–80, 2000, Accessed: May 14, 2020. [Online]. Available: https://www.jstor.org/stable/4165660

[67]
M. Friendly and D. Denis, “The early origins and development of the scatterplot,” *J. Hist. Behav. Sci.*, vol. 41, no. 2, pp. 103–130, 2005, doi: 10.1002/jhbs.20078.

[68]
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 SIGCHI Conference on Human Factors in Computing Systems*, 2013, p. 3237. doi: 10.1145/2470654.2466443.

[69]
S. Gattei, “Karl Popper’s philosophical breakthrough,” *Philosophy of Science*, vol. 71, no. 4, pp. 448–466, 2004.

[70]
E. L. Gettier, “Chapter 6. IS JUSTIFIED TRUE BELIEF KNOWLEDGE?,” in *Causal Theories of Mind*, De Gruyter, 2012, pp. 135–137.

[71]
V. Giardino, “Intuition and Visualization in Mathematical Problem Solving,” *Topoi*, vol. 29, no. 1, pp. 29–39, Apr. 2010, doi: 10.1007/s11245-009-9064-5.

[72]
G. Gigerenzer and U. Hoffrage, “How to improve Bayesian reasoning without instruction: Frequency formats,” *Psychological Review*, vol. 102, no. 4, pp. 684–704, 1995, doi: 10.1037/0033-295X.102.4.684.

[73]
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, Dec. 2013, doi: 10.1109/TVCG.2013.183.

[74]
D. G. Goldstein and D. Rothschild, “Lay understanding of probability distributions,” *Judgment and Decision Making*, vol. 9, no. 1, p. 14, 2014.

[75]
L. G. Halsey, “The reign of the *p* -value is over: what alternative analyses could we employ to fill the power vacuum?,” *Biol. Lett.*, vol. 15, no. 5, p. 20190174, May 2019, doi: 10.1098/rsbl.2019.0174.

[76]
J. Heer and M. Bostock, “Crowdsourcing Graphical Perception: Using Mechanical Turk to Assess Visualization Design,” p. 10, 2010.

[77]
J. Heer and G. Robertson, “Animated Transitions in Statistical Data Graphics,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 13, no. 6, pp. 1240–1247, Nov. 2007, doi: 10.1109/TVCG.2007.70539.

[78]
J. Helske, S. Helske, M. Cooper, A. Ynnerman, and L. Besançon, “Can Visualization Alleviate Dichotomous Thinking? Effects of Visual Representations on the Cliff Effect,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 27, no. 8, pp. 3397–3409, Aug. 2021, doi: 10.1109/TVCG.2021.3073466.

[79]
R. Hoekstra, R. D. Morey, J. N. Rouder, and E.-J. Wagenmakers, “Robust misinterpretation of confidence intervals,” *Psychon Bull Rev*, vol. 21, no. 5, pp. 1157–1164, Oct. 2014, doi: 10.3758/s13423-013-0572-3.

[80]
U. Hoffrage and G. Gigerenzer, “Using natural frequencies to improve diagnostic inferences,” *Academic medicine*, vol. 73, no. 5, pp. 538–540, 1998.

[81]
D. R. Hofstadter, *Gödel, escher, bach*. Harvester Press London, 1979.

[82]
J. Hollan, E. Hutchins, and D. Kirsh, “Distributed cognition: toward a new foundation for human-computer interaction research,” *ACM Transactions on Computer-Human Interaction*, vol. 7, no. 2, pp. 174–196, 2000.

[83]
T. P. Hughes and W. Bijker, “The evolution of large technological systems: The social construction of technological systems,” *The social construction of technological systems*, pp. 49–82, 1987.

[84]
J. Hullman, M. Kay, Y.-S. 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, Jan. 2018, doi: 10.1109/TVCG.2017.2743898.

[85]
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, Nov. 2015, doi: 10.1371/journal.pone.0142444.

[86]
T. M. Hutchins, Edwin, “Cognition in the Wild,” *The MIT Press*, Feb. 1995. https://mitpress.mit.edu/books/cognition-wild (accessed Oct. 30, 2018).

[87]
A. Jamrozik, M. McQuire, E. R. Cardillo, and A. Chatterjee, “Metaphor: Bridging embodiment to abstraction,” *Psychon Bull Rev*, vol. 23, no. 4, pp. 1080–1089, Aug. 2016, doi: 10.3758/s13423-015-0861-0.

[88]
N. Jardine, B. D. Ondov, N. Elmqvist, and S. Franconeri, “The Perceptual Proxies of Visual Comparison,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 1, pp. 1012–1021, Jan. 2020, doi: 10.1109/TVCG.2019.2934786.

[89]
B. D. Jones, “Bounded rationality and public policy: Herbert A. Simon and the decisional foundation of collective choice,” *Policy Sciences*, vol. 35, no. 3, pp. 269–284, 2002.

[90]
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, Feb. 2021, doi: 10.1109/TVCG.2020.3030335.

[91]
P. Kalinowski and F. Fidler, “Interpreting significance: the differences between statistical significance, effect size, and practical importance,” *Newborn and Infant Nursing Reviews*, vol. 10, no. 1, pp. 50–54, 2010.

[92]
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 CHI Conference on Human Factors in Computing Systems*, 2016, pp. 5092–5103. doi: 10.1145/2858036.2858558.

[93]
Y.-S. Kim, K. Reinecke, and J. Hullman, “Explaining the Gap: Visualizing One’s Predictions Improves Recall and Comprehension of Data,” in *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 2017, pp. 1375–1386. doi: 10.1145/3025453.3025592.

[94]
Y.-S. Kim, L. A. Walls, P. Krafft, and J. Hullman, “A bayesian cognition approach to improve data visualization,” in *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–14.

[95]
G. Klein, B. Moon, and R. R. Hoffman, “Making Sense of Sensemaking 1: Alternative Perspectives,” *IEEE Intelligent Systems*, vol. 21, no. 4, pp. 70–73, Jul. 2006, doi: 10.1109/MIS.2006.75.

[96]
K. H. Knuth, “Optimal Data-Based Binning for Histograms,” *arXiv:physics/0605197*, May 2006, Accessed: Aug. 15, 2019. [Online]. Available: http://arxiv.org/abs/physics/0605197

[97]
S. Koppman, C. L. Cain, and E. Leahey, “The Joy of Science: Disciplinary Diversity in Emotional Accounts,” *Science, Technology, & Human Values*, vol. 40, no. 1, pp. 30–70, Jan. 2015, doi: 10.1177/0162243914537527.

[98]
R. Kosara and S. Haroz, “Skipping the replication crisis in visualization: Threats to study validity and how to address them: Position paper,” in *Proceedings of the Workshop on Evaluation and Beyond-Methodological Approaches for Visualization*, 2018, pp. 102–107.

[99]
S. M. Kosslyn and S. M. Kosslyn, *Graph Design for the Eye and Mind*. Oxford University Press, USA, 2006.

[100]
T. Kremer, “The Significance of Solipsism,” in *Proceedings of the Aristotelian Society*, 1959, pp. 35–60.

[101]
E. Kuo, M. M. Hull, A. Elby, and A. Gupta, “Mathematical Sensemaking as Seeking Coherence between Calculations and Concepts: Instruction and Assessments for Introductory Physics,” *arXiv:1903.05596 [physics]*, Mar. 2019, Accessed: May 22, 2020. [Online]. Available: http://arxiv.org/abs/1903.05596

[102]
G. Lakoff, “Metaphorical Structure of Mathematics,” 1997.

[103]
C. Lambdin, “Significance tests as sorcery: Science is empirical—significance tests are not,” *Theory & Psychology*, vol. 22, no. 1, pp. 67–90, Feb. 2012, doi: 10.1177/0959354311429854.

[104]
P.-M. Law, A. Endert, and J. Stasko, “What are Data Insights to Professional Visualization Users?,” *arXiv:2008.13057 [cs]*, Oct. 2020, Accessed: Aug. 06, 2021. [Online]. Available: http://arxiv.org/abs/2008.13057

[105]
E. L. Lehmann, J. P. Romano, and G. Casella, *Testing statistical hypotheses*, vol. 3. Springer, 2005.

[106]
J. Litman, “Curiosity and the pleasures of learning: Wanting and liking new information,” *Cognition and Emotion*, vol. 19, no. 6, pp. 793–814, Sep. 2005, doi: 10.1080/02699930541000101.

[107]
Lyn D. English and Jane M. Watson, “Development of Probabilistic Understanding in Fourth Grade,” *Journal for Research in Mathematics Education*, vol. 47, no. 1, p. 28, 2016, doi: 10.5951/jresematheduc.47.1.0028.

[108]
F. Machlup and U. Mansfield, “Cultural diversity in studies of information,” *The study of information: Interdisciplinary messages. New York: Wiley*, pp. 3–56, 1983.
[109]
C. Mackay, *Extraordinary popular delusions and the madness of crowds*. Simon and Schuster, 2012.
[110]
J. I. Marden, “Positions and QQ Plots,” *Statistical Science*, vol. 19, no. 4, pp. 606–614, 2004, Accessed: Aug. 18, 2021. [Online]. Available: https://www.jstor.org/stable/4144431
[111]
L. Martignon and C. Wassner, “Teaching decision making and statistical thinking with natural frequencies,” 2002.
[112]
D. Melcher and E. Kowler, “Shapes, surfaces and saccades,” *Vision research*, vol. 39, no. 17, pp. 2929–2946, 1999.
[113]
J. Middendorf and D. Pace, “Decoding the disciplines: A model for helping students learn disciplinary ways of thinking,” *New Directions for Teaching and Learning*, vol. 2004, no. 98, pp. 1–12, 2004, doi: 10.1002/tl.142.
[114]
F. C. Mills, *Statistical Methods Applied to Economics and Business*. Holt, 1938.
[115]
D. S. Moore, “Should Mathematicians Teach Statistics?,” *College Mathematics Journal*, vol. 19, no. 1, pp. 3–7, 1988.
[116]
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.
[117]
D. Moritz *et al.*, “Formalizing visualization design knowledge as constraints: Actionable and extensible models in draco,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 25, no. 1, pp. 438–448, 2018.
[118]
H. M. Natter and D. C. Berry, “Effects of active information processing on the understanding of risk information,” *Applied Cognitive Psychology*, vol. 19, no. 1, pp. 123–135, Jan. 2005, doi: 10.1002/acp.1068.
[119]
R. R. Nelson, *An evolutionary theory of economic change*. Harvard University Press, 1985.
[120]
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, doi: 10.1111/cgf.13902.

[121]
R. E. Nisbett, D. H. Krantz, C. Jepson, and Z. Kunda, “The use of statistical heuristics in everyday inductive reasoning,” *Psychological Review*, vol. 90, no. 4, pp. 339–363, 1983, doi: 10.1037/0033-295X.90.4.339.

[122]
D. A. Norman, *The design of everyday things*, Revised and Expanded edition. New York, NY: Basic Books, 2013.

[123]
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, 2020.

[124]
S. S. Pak, J. B. Hutchinson, and N. B. Turk-Browne, “Intuitive statistics from graphical representations of data,” *Journal of Vision*, vol. 14, no. 10, pp. 1361–1361, Aug. 2014, doi: 10.1167/14.10.1361.

[125]
G. Paolacci and J. Chandler, “Inside the Turk: Understanding Mechanical Turk as a Participant Pool,” *Current Directions in Psychological Science*, vol. 23, no. 3, pp. 184–188, Jun. 2014, doi: 10.1177/0963721414531598.

[126]
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, 2017.

[127]
K. Potter, J. Kniss, R. Riesenfeld, and C. R. Johnson, “Visualizing Summary Statistics and Uncertainty,” *Computer Graphics Forum*, vol. 29, no. 3, pp. 823–832, Aug. 2010, doi: 10.1111/j.1467-8659.2009.01677.x.

[128]
J. Ross, L. Irani, M. S. Silberman, A. Zaldivar, and B. Tomlinson, Who are the crowdworkers? Shifting demographics in Mechanical Turk.In Extended Abstracts *on Human Factors in Computing Systems*, 2010, pp. 2863–2872.

[129]
D. H. Rouvray, “Elements in the history of the Periodic Table,” *Endeavour*, vol. 28, no. 2, pp. 69–74, Jun. 2004, doi: 10.1016/j.endeavour.2004.04.006.

[130]
A. Rubin, “Interactive visualizations of statistical relationships: What do we gain,” 2002.

[131]
R. Ruggles and H. Brodie, “An Empirical Approach to Economic Intelligence in World War II,” *Journal of the American Statistical Association*, vol. 42, no. 237, pp. 72–91, Mar. 1947, doi: 10.1080/01621459.1947.10501915.

[132]
I. Sandler, “Development: Mendel’s Legacy to Genetics,” *Genetics*, vol. 154, no. 1, pp. 7–11, Jan. 2000, Accessed: May 14, 2020. [Online]. Available: https://www.genetics.org/content/154/1/7

[133]
R. K. Sawyer, “Emergence in psychology: Lessons from the history of non-reductionist science,” *Human development*, vol. 45, no. 1, pp. 2–28, 2002.

[134]
M. Scaife and Y. Rogers, “External cognition: how do graphical representations work?,” *International Journal of Human-Computer Studies*, vol. 45, no. 2, pp. 185–213, Aug. 1996, doi: 10.1006/ijhc.1996.0048.

[135]
M. J. Schervish, “Theory of statistics,” 1995.

[136]
D. W. Scott, “Sturges’ rule,” *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 1, no. 3, pp. 303–306, 2009.

[137]
B. L. Sherin, “A Comparison of Programming Languages and Algebraic Notation as Expressive Languages for Physics,” *International Journal of Computers for Mathematical Learning*, vol. 6, no. 1, pp. 1–61, May 2001, doi: 10.1023/A:1011434026437.

[138]
M. Shermer, *Why people believe weird things: pseudoscience, superstition, and other confusions of our time*, Rev. and Expanded. New York: A.W.H. Freeman/Owl Book, 2002.

[139]
M. Shermer, *The believing brain: From spiritual faiths to political convictions–How we construct beliefs and reinforce them as truths*. Hachette UK, 2012.

[140]
B. Shneiderman, “The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations,” in *The Craft of Information Visualization*, B. B. Bederson and B. Shneiderman, Eds. San Francisco: Morgan Kaufmann, 2003, pp. 364–371. doi: 10.1016/B978-155860915-0/50046-9.

[141]
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.

[142]
V. Stango and J. Zinman, “Exponential Growth Bias and Household Finance,” *The Journal of Finance*, vol. 64, no. 6, pp. 2807–2849, 2009, doi: 10.1111/j.1540-6261.2009.01518.x.

[143]
S. L. Star and J. R. Griesemer, “Institutional ecology,translations’ and boundary objects: Amateurs and professionals in Berkeley’s Museum of Vertebrate Zoology, 1907-39,” *Social studies of science*, vol. 19, no. 3, pp. 387–420, 1989.

[144]
K. Stenning and J. Oberlander, “A cognitive theory of graphical and linguistic reasoning: Logic and implementation,” *Cognitive Science*, vol. 19, no. 1, pp. 97–140, Jan. 1995, doi: 10.1016/0364-0213(95)90005-5.

[145]
B. J. Strasser and P. N. Edwards, “Big data is the answer… But what is the question?,” *Osiris*, vol. 32, no. 1, pp. 328–345, 2017.

[146]
Student, “The probable error of a mean,” *Biometrika*, pp. 1–25, 1908.

[147]
G. M. Sullivan and R. Feinn, “Using Effect Size—or Why the P Value Is Not Enough,” *J Grad Med Educ*, vol. 4, no. 3, pp. 279–282, Sep. 2012, doi: 10.4300/JGME-D-12-00156.1.

[148]
S. Tak, A. Toet, and J. van Erp, “The Perception of Visual UncertaintyRepresentation by Non-Experts,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 6, pp. 935–943, Jun. 2014, doi: 10.1109/TVCG.2013.247.

[149]
M. Tory and T. Möller, “Evaluating Visualizations: Do Expert Reviews Work?,” *IEEE computer graphics and applications*, vol. 25, pp. 8–11, Sep. 2005, doi: 10.1109/MCG.2005.102.

[150]
E. R. Tufte, *Beautiful evidence*. Graphis Press, Cheshire, MA, USA, 2006.

[151]
E. R. Tufte, S. R. McKay, W. Christian, and J. R. Matey, “Visual explanations: Images and quantities, evidence and narrative.” American Institute of Physics, 1998.

[152]
J. W. Tukey, *Exploratory data analysis*. Reading, MA, USA: Addison-Wesley, 1977.

[153]
J. W. Tukey, “We Need Both Exploratory and Confirmatory,” *The American Statistician*, vol. 34, no. 1, pp. 23–25, 1980, doi: 10.2307/2682991.

[154]
R. Tuomela, “Methodological solipsism and explanation in psychology,” *Philosophy of Science*, vol. 56, no. 1, pp. 23–47, 1989.

[155]
N. A. Van House, “Science and technology studies and information studies,” *Annual review of information science and technology*, vol. 38, no. 1, pp. 1–86, 2004.

[156]
J. P. Vandenbroucke, “Is ‘the causes of cancer’ a miasma theory for the end of the twentieth century?,” *International journal of epidemiology*, vol. 17, no. 4, pp. 708–709, 1988.

[157]
H. WAINER, *Graphic Discovery: A Trout in the Milk and Other Visual Adventures*. Princeton University Press, 2005. doi: 10.2307/j.ctt4cgc63.

[158]
H. Wainer, *Visual Revelations: Graphical Tales of Fate and Deception From Napoleon Bonaparte To Ross Perot*. Psychology Press, 2013.

[159]
C. Ware, *Visual Thinking for Design*. Elsevier, 2010.

[160]
C. Ware, *Information Visualization: Perception for Design*. Morgan Kaufmann, 2019.
[161]
W. WEAVER, "RECENT CONTRIBUTIONS TO THE MATHEMATICAL THEORY OF COMMUNICATION," *ETC: A Review of General Semantics*, vol. 10, no. 4, pp. 261–281, 1953, Accessed: May 21, 2020. [Online]. Available: https://www.jstor.org/stable/42581364
[162]
F. Webster, *Theories of the information society*. Routledge, 2014.
[163]
N. H. T. E. R. West, "Effects of Data Noise on Statistical Judgement," *Thinking & Reasoning*, vol. 3, no. 2, pp. 111–132, Apr. 1997, doi: 10.1080/135467897394383.
[164]
D. Whitaker and H. Walker, "Centroid evaluation in the vernier alignment of random dot clusters," *Vision Research*, vol. 28, no. 7, pp. 777–784, Jan. 1988, doi: 10.1016/0042-6989(88)90024-7.
[165]
H. Wickham, "A Layered Grammar of Graphics," *Journal of Computational and Graphical Statistics*, vol. 19, no. 1, pp. 3–28, Jan. 2010, doi: 10.1198/jcgs.2009.07098.
[166]
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, Nov. 2010, doi: 10.1109/TVCG.2010.161.
[167]
H. Wickham and L. Stryjewski, "40 years of boxplots," *Am. Statistician*, p. 2011, 2011.
[168]
M. H. Wilkerson and M. Fenwick, "Using mathematics and computational thinking," *Helping students make sense of the world using next generation science and engineering practices*, pp. 181–204, 2017.
[169]
L. Wilkinson, "Dot Plots," *The American Statistician*, vol. 53, no. 3, pp. 276–281, Aug. 1999, doi: 10.1080/00031305.1999.10474474.
[170]
L. Wilkinson, "The Grammar of Graphics," in *Handbook of Computational Statistics: Concepts and Methods*, J. E. Gentle, W. K. Härdle, and Y. Mori, Eds. Berlin, Heidelberg: Springer, 2012, pp. 375–414. doi: 10.1007/978-3-642-21551-3_13.
[171]
L. Wilkinson, A. Anand, and R. Grossman, "Graph-theoretic scagnostics," in *Information Visualization, IEEE Symposium on*, 2005, pp. 21–21.
[172]
M. Windschitl, J. Thompson, and M. Braaten, "Beyond the scientific method: Model-based inquiry as a new paradigm of preference for school science investigations," *Science education*, vol. 92, no. 5, pp. 941–967, 2008.

[173]
N.Wood, C. Rust, G. Horne, “A Tacit Understanding: The Designer’s Role in Capturing and Passing on the Skilled Knowledge of Master Craftsmen,” *International Journal of Design*. 2009 http://www.ijdesign.org/index.php/IJDesign/article/view/559 (accessed Apr. 03, 2023).

[174]
T. Yarkoni, “The Generalizability Crisis,” PsyArXiv, preprint, Nov. 2019. doi: 10.31234/osf.io/jqw35.

[175]
J. S. Yi, Y. Kang, J. T. Stasko, and J. A. Jacko, “Understanding and characterizing insights: how do people gain insights using information visualization?,” in *Proceedings of the Workshop on BEyond time and errors: novel evaLuation methods for Information Visualization*, 2008, pp. 1–6.

[176]
L. Yuan, S. Haroz, and S. Franconeri, “Perceptual proxies for extracting averages in data visualizations,” *Psychon Bull Rev*, vol. 26, no. 2, pp. 669–676, Apr. 2019, doi: 10.3758/s13423-018-1525-7.

[177]
A. Zaidan, Z. Ismail, Y. M. Yusof, and H. Kashefi, “Misconceptions in descriptive statistics among postgraduates in social sciences,” *Procedia-Social and Behavioral Sciences*, vol. 46, pp. 3535–3540, 2012.

[178]
J. Zhang and D. A. Norman, “Representations in Distributed Cognitive Tasks,” *Cognitive Science*, vol. 18, no. 1, pp. 87–122, 1994, doi: 10.1207/s15516709cog1801_3.

[179]
J. Zhang, “The nature of external representations in problem solving,” *Cognitive Science*, vol. 21, no. 2, pp. 179–217, Apr. 1997, doi: 10.1016/S0364-0213(99)80022-6.

[180]
“UNITED NATIONS Climate Change Summit,” 2020. https://www.un.org/en/climatechange/reports.shtml (accessed May 14, 2020).