Investigating the Effect of the Multiple Comparisons Problem in Visual Analysis

Investigating the Effect of the Multiple Comparisons Problem in Visual Analysis

Emanuel Zgraggen[1] **Zheguang Zhao**[1] **Robert Zeleznik**[1] **Tim Kraska**[1, 2]

Brown University[1]
Providence, RI, United States
{ez, sam, bcz}@cs.brown.edu

Massachusetts Institute of Technology[2]
Cambridge, MA, United States
kraska@mit.edu

ABSTRACT

The goal of a visualization system is to facilitate data-driven insight discovery. But what if the insights are spurious? Features or patterns in visualizations can be perceived as relevant insights, even though they may arise from noise. We often compare visualizations to a mental image of what we are interested in: a particular trend, distribution or an unusual pattern. As more visualizations are examined and more comparisons are made, the probability of discovering spurious insights increases. This problem is well-known in Statistics as the multiple comparisons problem (MCP) but overlooked in visual analysis. We present a way to evaluate MCP in visualization tools by measuring the accuracy of user reported insights on synthetic datasets with known ground truth labels. In our experiment, over 60% of user insights were false. We show how a confirmatory analysis approach that accounts for all visual comparisons, insights and non-insights, can achieve similar results as one that requires a validation dataset.

ACM Classification Keywords

H.5.m. Information Interfaces and Presentation (e.g., HCI): Miscellaneous

Author Keywords

Multiple Comparisons Problem; Visual Analysis; Visualization; Statistics; Experiment.

INTRODUCTION

Here is a thought experiment. Imagine a game where you roll a pair of dice and win if you get two sixes. The probability of winning is $1/36$. Now let's change the game mechanics. Instead of just rolling once, you can continue rolling the dice. You might get a three and a four in your first roll. You did not win, but you keep going. On your 100^{th} try you get two sixes and win. Everyone will win this game eventually. The probability of winning after an infinite number of rolls is 1. Even after just 100 rolls the chances of winning are over 94%.

CHI 2018, April 21–26, 2018, Montréal, QC, Canada.

ACM ISBN 978-1-4503-5620-6/18/04 ...$15.00.
http://dx.doi.org/10.1145/3173574.3174053

Similar to others [9], we argue that the same thing happens when performing visual comparisons, but instead of winning, an analyst "loses" when they observe an interesting-looking random event (e.g., two sixes). Instead of being rewarded for persistence, an analyst increases their chances of losing by viewing more data visualizations. This concept is formally known as the *multiple comparisons problem (MCP)* [5]. Consider an analyst looking at a completely random dataset. As more comparisons are made, the probability rapidly increases of encountering interesting-looking (e.g., data trend, unexpected distribution, etc.), but still random events. Treating such inevitable patterns as insights is a *false discovery* (Type I error) and the analyst "loses" if they act on such false insights.

Unlike the above random data example, empirical datasets are an unknown weighted product of noise and signal and thus not totally random. The data analyst's "game" is to detect data patterns that are real and ignore those that are spurious. Unfortunately, the difference between the two may be small or non-existent. Thus an analyst also "loses" by ignoring a real pattern because it looks uninteresting. This is known as a *false omission* (Type II error). False discoveries and omissions might be rare, but, due to the MCP, they are increasingly likely to occur as analysts look at more visualizations.

To further demonstrate MCP in exploratory visualization, we present a representative random data scenario; we consider non-random data in our experiments (§Experimental Method).

Jean works at a small non profit organization. Every year they send their donors a small thank-you gift and want to repeat that this year. From past experience, the organization knows that only half of all new donors become recurring donors. Jean suspects there might be a relationship between retention rate and thank-you gift type. Maybe better-liked gifts trigger repeat donations. Jean uses his favorite visualization tool to explore data from the last 10 years. He first looks at the 2006 data and sees that slightly less than half of the new donors donated again (Figure 1 (a)). Then Jean inspects the 2007 data and sees the same result (Figure 1 (b)). After scanning through all the other years and coming to similar conclusions, Jean looks at 2016 (Figure 1 (c)). Instantly he sees that this visualization is much different than the others, depicting a noticeable shift toward more repeat-donors. Jean gets excited. He believes he has figured out a way to improve the donor retention rate. People liked the USB-drive his company sent out that year so much that they stayed loyal to the organization. Even though

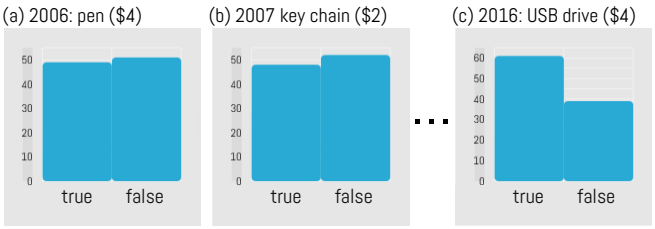

Figure 1. A user inspects several graphs and wrongly flags (c) as an insight because it looks different than (a) and (b). All were generated from the same uniform distribution and are the "same". By viewing lots of visualizations, the chances increase of seeing an apparent insight that is actually the product of random noise.

this gift is the most expensive one, it is worth it to send it again this year. Long term donors bring in a lot of money.

Is Jean's insight correct? It is not. The dataset Jean looked at was sampled from a uniform distribution. It was completely random. We controlled the process that generated this dataset and there was no signal that related gifts to donor retention rate in any form. Jean's false discovery led to a spurious insight. By doing ten comparisons he increased the likelihood of finding a seemingly interesting pattern in random data.

There are various common approaches for following up Jean's exploratory analysis with statistical analysis, each with different trade-offs. We introduce these via co-workers with whom Jean shares his insight: Noemi (*confirmation; same dataset*), Hasan (*confirmation; validation dataset*) and Kendra (*mixing exploration and confirmation*).

Noemi transforms Jean's insight into something statistically testable. She defines a null hypothesis: becoming a repeat-donor is just as likely as being a onetime-donor. She tests if the 2016 data rejects this. The p-value turns out to be 0.028 indicating a significant effect (for a significance level of 0.05). Noemi arrives at the same, wrong, conclusion as Jean. By confirming a hypothesis on the same dataset that has informed that hypothesis, she introduced systemic bias.

Like Noemi, Hasan converts Jean's insight into a statistical hypothesis but tells Jean it's unsafe to test on the same dataset. They agree to send the USB-drive again this year and re-run the test after obtaining new retention data. The test comes out as not significant, refuting Jean's initial insight. Hasan got it right. Running confirmatory analysis on a validation dataset is statistically sound. However, obtaining new data is, in practice, often expensive, time-consuming or even impossible.

Kendra takes yet a different approach. She suspects that Jean probably did several visual comparisons prior to reporting his insight, all of which need to be incorporated in any confirmatory analysis done on the same dataset to avoid multiple hypotheses errors. She asks Jean to meticulously recount what he did and maps all of Jean's visual comparisons to equivalent statistical hypotheses. Jean made ten comparisons: one explicit for 2016, and nine implicit, unreported, ones for the years 2006 - 2015. Kendra runs statistical tests on the original dataset using the Benjamini-Hochberg [6] procedure to control for such a multiple comparisons scenario. The corrected p-value for 2016 equals 0.306. Kendra deems the test insignificant and informs Jean that his insight is likely due to random noise in the data. While Kendra's approach (*mixing exploration and confirmation*) requires Jean to remember his visual analysis session in detail, it also allows for statistically valid confirmation of his findings using the same dataset.

This paper presents an experiment that quantifies the accuracy of user reported insights, where we define insights as observations, hypotheses and generalizations directly extracted from data. We acknowledge this definition is narrow. Insights from visualizations can be much broader and multifaceted. Visualizations help users gain a deep understanding of a problem domain. However, we purposefully limit insights to this subset because there is no ambiguity of what correctness means. Using synthetic datasets with known ground truth labels, we can assign a binary score to each insight: true or false. We then compute an accuracy score for each participant by dividing the count of correct insights by the number of all insights.

We follow up by manually mapping insights to corresponding statistical tests and evaluate the three different confirmatory approaches just illustrated. We discuss how an approach that validates user insights on the same dataset as used during exploration inflates the false discovery rate due to the MCP. We show that these errors are dramatically reduced by validating on a separate dataset. Finally, we demonstrate that by accounting for all visual comparisons done by a user during exploration, the approach of *mixing exploration and confirmation*, can achieve similar results to using a separate dataset.

WHY THE VISUALIZATION COMMUNITY SHOULD CARE

In theory, there is a clear divide between exploratory and confirmatory data analysis methods [47]. The goal of the former is to browse through data letting visualizations trigger potential insights and hypotheses. The latter extends this process with a directed search intended to confirm or reject insights and hypotheses given a priori [46]. Within this realm, mistakes are acceptable in the exploratory phase because it is expected that the confirmatory phase will correct them.

In practice, however, the distinction between the two methods can be blurry. Oftentimes users will unwittingly switch between the two and "convert results from investigative analysis into evidence" [31]. There are also pitfalls associated with this approach that are unobvious to non-statisticians; for example doing confirmatory data analysis on the same dataset as the exploratory analysis introduces systemic bias known as data dredging or p-hacking [28]. While splitting a dataset into exploratory and confirmatory parts gets around that problem, it significantly lowers the power of any test due to smaller sample sizes. And without using advanced controlling procedures for multiple hypotheses error that allow for incremental testing [54], iterative switching between exploration and confirmation can not be done. Standard procedures such as Bonferroni [16] can only be applied once per dataset.

The blurring of the lines between exploratory and confirmatory analysis is arguably magnified by how both commercial visualization systems and research prototypes are advertised: "...uncover hidden insights on the fly...", "...harnesses people's natural ability to spot visual patterns quickly, revealing everyday opportunities and eureka moments alike..." [44], "...no

expertise required.." [45], "...an interactive data exploration system tailored towards "fast-forwarding" to desired trends, patterns, or insights, without much effort from the user..." [42]. We believe that such statements instill a false sense of confidence and reliability in insights derived from exploratory visual analysis. This might be especially true for tools that target *data enthusiasts* - people who are "not mathematicians or programmers, and only know a bit of statistics" [27].

RELATED WORK

We relate and compare our work to prior art in the areas of *Insight-based Evaluation*, *Visual Inference and Randomness* and *Multiple Comparisons Problem in Statistics*.

Insight-based Evaluation

Many have argued that information visualization's primary goal is to provide insights [10, 37, 11], and, unsurprisingly, the visualization community has increasingly adopted insight-based evaluation methods [35, 23, 26, 52, 40]. Beyond measuring directly how well systems achieve this main goal, these methods also allow for ecologically valid design comparisons. Plaisant argues [38] that evaluation methods should estimate not only how efficiently a new technique reveals trends or phenomena from data, but also the potential risk for errors. Insight-based evaluation methods clearly address the former but largely ignore the latter. Among other risks for misinterpretations [8], visualizations are subjective and can be misleading [48]. Users may perceive a visual feature even though it arises from random data noise; the risk for this increases proportionately with the number of visualizations inspected.

Van Wijk [48] introduces an economic model that equates the investments associated with a visualization (e.g., initial cost to create a visualization, perception and exploration costs) with the return on those investments (i.e., the total knowledge gained by a user). We want techniques that optimize this model for low cost and high investment return. Usability studies and controlled experiments on benchmark tasks help us understand the cost of a particular design, and insight-based evaluations attempt to assess the other side of this equation. However, measuring the number of insights without any quality weighting paints an incomplete picture and has been mentioned specifically as a limitation of such study designs [52].

Several proxy metrics have been proposed, for example, insight "value" as assessed by domain experts [40] or "originalty" scores (how often the same insight was reported). We augment this work with an experimental method based on synthetic datasets that assigns each insight a binary quality score: true or false. Our definition of insights is comparatively narrow [37, 11] and only encompasses observations, hypotheses and generalizations directly related to the data and not on any other sources such as prior knowledge or domain expertise.

Visual Inference and Randomness

Buja et al. [9] outline the parallelism between quantitative testing and visual analysis and argue that the term "discovery" (or insight) in visual analysis can often be equated to "rejection of a null hypothesis". Seeing an interesting upward trend in a visualization, for example, can be taken as a rejection of uniformity. Our study leverages this notion as we manually extract null hypotheses from user reported insights. Follow-up work [36] indicates visual inference can perform comparably to quantitative testing under certain protocols. Similarly, there is a large body of work in visualization and graphical perception covering individual value estimation [12], correlation [39, 34], searching [24] or regression [13]. The consensus is that user inferences are accurate given a task appropriate visualization. However, none of this work analyzes if or how MCP affects visual inference when comparisons are done in series.

People are known to judge randomness inaccurately [33, 25] and see patterns where there are none [29]. Well-known illustrations include: *gambler's fallacy* [43], *Hot-hand fallacy* [3] or *Birthday paradox*. The "Rorschach protocol" [9, 50] can be used to test people's tendency to see patterns in random data.

We consider the interplay between visual inference and judgment of randomness. *Gambler's fallacy* is famously known in the game of roulette where people might misinterpret a streak of 20 reds in a row as a pattern. Humans are good at spotting such patterns but are bad at judging that this outcome is not more or less uncommon than any other red and black sequence of the same length. Data exploration allows users to browse a large number of visualizations quickly especially when using automatic visualization recommendation systems [49, 41]. While scanning through lots of visualizations we might find one with an interesting pattern without considering it could be the artifact of random noise.

Multiple Comparisons Problem in Statistics

Formally, MCP occurs when the risk of observing a falsely significant result increases as more than one hypothesis is considered at once. This phenomenon is also known as the *multiple hypotheses error*, data dredging or *p*-hacking [5]. Suppose we are looking for indicators in a census dataset that affects salary distribution. To examine factors such as "age" or "education", we set up the corresponding *null hypothesis* that states the proposed attribute has no correlation with the salary distribution. We then use a statistical test to infer the likelihood of observing a likewise spurious correlation under the null hypothesis. If this likelihood, commonly referred to as the *p*-value, is lower than the chosen significance level such as 0.05, then the null hypothesis is rejected, and the *alternative hypothesis* that the proposed attribute is correlated with salary is deemed statistically significant.

However, if we keep searching through different indicators in the dataset, we are almost guaranteed to find a statistically significant correlation. For example, choosing a significance level for each test of 0.05 means that statistically we have a 5% chance of falsely rejecting a given null hypothesis; even if the dataset contains completely random data, we would, on average, falsely discover a spurious correlation that passes our significance level after only 20 hypothesis tests.

Several techniques exist to control for multiple hypotheses error. Procedures such as Bonferroni [16] control for *family-wise error rate* (FWER), which is the probability of incurring *any* false discovery given the hypotheses. For the previous example, Bonferroni can reduce the FWER from 100% to just 5%. FWER procedures in general provide the safest control.

The downside of FWER procedures is their statistical power decreases as the number of hypotheses increase to compensate for the risk of making any error. In situations where false discovery is costly, such as medical trials, FWER may be appropriate. However, common data science applications are often more concerned with the accuracy of observed significant results than the possibility of an error. Thus *false discovery rate* (FDR) is proposed as an alternative control target which specifies the expected proportion of false discoveries among only the discoveries made (i.e. the rejected null hypotheses), instead of all the hypotheses examined. An FDR procedure such as Benjamini-Hochberg [6] bounds the ratio of false rejections among only the rejected tests to be say 5%. In general FDR is a weaker guarantee than FWER but has shown tremendous benefit for discovering truly significant insights. Recent work on FDR and its variants such as the marginal FDR (mFDR) improves over Benjamini-Hochberg for dynamic settings [19] and specifically interactive data exploration [54].

MCP is also manifest as overfitting in data mining and machine learning. Common practice is to train models on one dataset and then evaluate them on an independent validation dataset. However the validation dataset is not easily reusable because of conflation with multiple hypotheses error. Recent work borrowing from Differential Privacy [17] reuses the hold-out dataset in adaptive data analysis.

In summary, MCP is well covered in statistics but very much overlooked in the visualization community. Oftentimes we compare a mental model of what interests us against the visualizations we observe. If these models align, we notice it. Perhaps unconsciously, we are doing comparisons. While these comparisons are not well-expressed mathematically, they still are tests and subject to the same MCP as statistical tests.

EXPERIMENTAL METHOD

The aim of this work is to investigate the effect of MCP in visual analysis. To achieve this, we need to evaluate the accuracy of user reported insights. We designed an experiment where an insight is an observation, hypothesis or generalization that could be directly extracted from the data and that did not require prior knowledge or domain expertise. Our study followed an open-ended protocol where we let participants explore a synthetic dataset and instructed them to report their reliable insights by writing them down. We believe this workflow models common real-world data exploration scenarios where users report interesting observations to discuss later with colleagues, to draw a conclusion or take an action, to analyze further or to convert into a publication. By using synthetic datasets generated from mathematical models we control, we can match ground truth labels to user observations and label each reported insight as being true or false. We use a full-factorial 2 dataset-types (*shopping* and *sleep*) $\times$ 2 dataset-sizes (300 and 1000 records) experiment. The remainder of this section describes the details of our experimental design.

System

The data exploration tool used in our study derives from Vizdom [14] and PanoramicData [53]. Figure 2 shows a screenshot. As with Tableau [44], users access dataset attributes from a list (a) and use drag and drop to create and modify visualizations (b, d). Our system supports two binned visualizations types: heat-maps (c) and bar-charts (d). Selecting bars or bins in visualizations reveals their raw data values. Brushing (e), filtering (f) and note-taking (g) operations are also available.

Datasets

Our experiment uses three dataset-types from different domains. A dataset-type specifies the attributes and value ranges the data consist of. The first dataset-type (*shopping*) contained customer information from a fictional shopping website. This dataset-type contained 12 attributes (4 quantitative, 3 nominal, 5 ordinal), and included information like ages of customers, incomes, region customers are from, average purchase amount per month and average minutes they spend on the site. The second one (*sleep*) consisted of data from a fictional study about people's sleep patterns and contains 10 attributes (5 quantitative, 1 nominal, 4 ordinal). Some example attributes include average hours of sleep, time to fall asleep, sleep tracker usage, fitness and stress levels. The third dataset-type (*restaurants*), used only to introduce participants to the system, contained ratings and attributes from restaurants of four different cities.

From these dataset-types we generated actual datasets of two different sizes: 1000 and 300 entries. The first size is derived from the fact that roughly half the datasets found on a popular website collecting machine learning-related datasets [30] have less than a 1000 records. The second size originates from Anthoine et al. [2] which analyzed 114 recent medical publications and found that the median sample size of studies in those publications was 207. While there are many others, these models, we believe, represent two common scenarios for visual data exploration and analysis: a user wants to draw conclusions from a study or a user wants to explore a dataset to inform feature selection for a machine learning task.

Our scheme for synthetic data generation is similar to [1] and addresses several challenges. First, in order to convey a realistic context to the user, the generated data should retain the domain-specific properties of the empirical distributions. For example, the synthetic sample of "age" should not be negative, and the sample mean should be within reasonable range. To this end, we extract the domain-specific parameters from empirical sample datasets to generate synthetic data.

Second, to assess insight accuracy, we need to embed ground truth labels in the dataset. To simulate real-world scenarios, our generated datasets must be a mix of signal and noise; however, we need to know definitively if a user insight is correct. To inform how to construct such datasets, we ran a six participant pilot study using the same tool on real-world datasets. Analyzing user-recorded insights, we found that most concerned distribution characteristics and relationships among attributes. Distribution characteristics include "the mean of age distribution is between 20 and 30" whereas attribute relationships range from "the income and the age are correlated" to "the people of age between 40 to 50 work the most".

To create a synthetic dataset we construct a model based on multiple bivariate normal random variables where the ground truth of both types of insights can be determined. For an *n*-

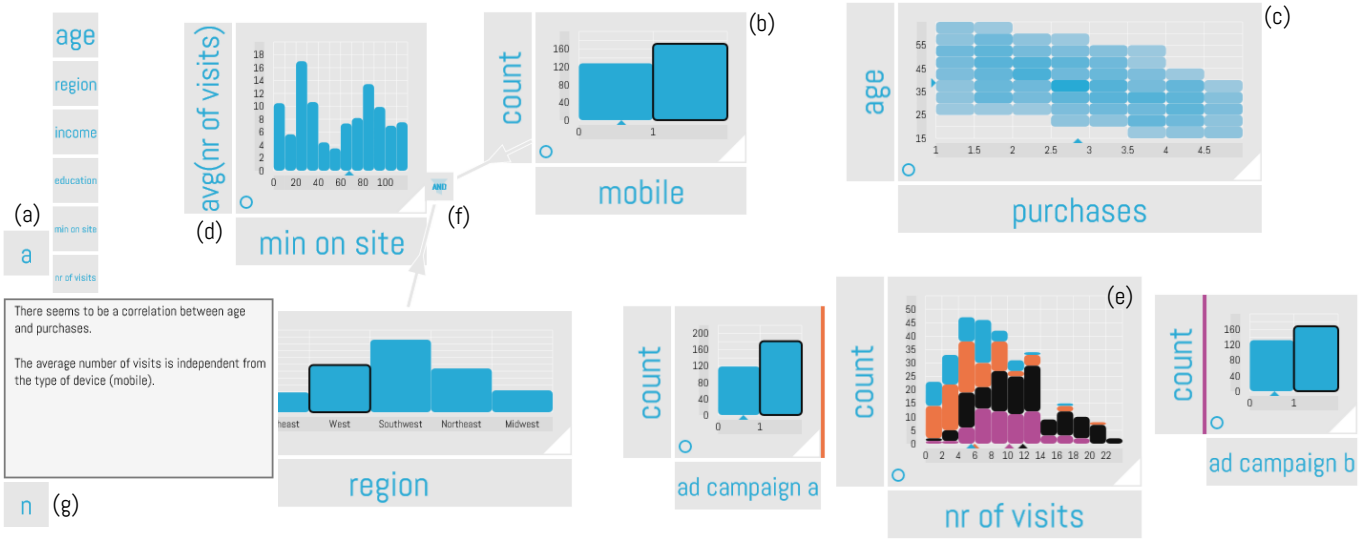

Figure 2. Screenshot of the visual analysis tool used in our study. The tool features an unbounded, pannable 2D canvas where visualizations can be laid out freely. The lefthand side of the screen gives access to all the attributes of the dataset (a). These attributes can be dragged and dropped on the canvas to create visualizations such as (b). Users can modify visualizations by dropping additional attributes onto an axis (c) or by clicking on axis labels to change aggregation functions (d). The tool supports brushing (e) and filtering (f) operations. Where filtering operations can be arbitrarily long chains of Boolean operators (AND, OR, NOT). The system offers a simple textual note-taking tool (g).

attribute dataset, we embed $n/2$ true relationships as correlated attribute pairs. Pairs are chosen randomly and given a randomized, non-zero correlation coefficient between -1 and 1. We then generate data by sampling from bivariate normal random variables, parameterized by these correlation coefficients and with means and variances extracted from empirical datasets. This process is repeated for each participant in our study.

If two attributes are sampled from independent normal random variables, then any insight involving the relationship between these two attributes is false. For two correlated variables, any simple insight is true. For more complex insights (e.g., statements about a sub-population mean like "the average age of married people is 35 to 40"), the ground truth can be calculated either analytically or computationally. Due to the truncation of the random variables for empirical domains, the analytical computation of ground truths for correlated attributes is complicated [51]. Instead, we generate datasets with 100M records from the same model and extract ground truth labels using hypothesis testing with Bonferroni correction [16]; labels are generated with 95% confidence.[1]

Procedure

We recruited 28 participants from a research university in the US. All participants were students (25 undergraduate, 3 graduate), all of whom had some experience with data exploration or analysis tools (e.g., Tableau, Pandas, R) and have taken at least introductory college-level statistics and probability classes. Our experiment included dataset-type (*shopping* and *sleep*) and dataset-size (300 and 1000 records) as between-subject factors. Each participant got to see one pairing of dataset-type and dataset- size. The study design was fully balanced - each unique combination of dataset-type and dataset-size was given to 7 participants. The actual dataset records and correlations were generated uniquely for each participant using different random seeds and according to the method outlined in §Datasets. Even if two users saw the same combination of dataset-type and dataset-size they still got a unique dataset in terms of the individual values of attributes and ground truths.

Each participant session had three parts. The first consisted of a 15 minute tutorial on how to interact with the system using a separate third dataset-type (*restaurant*). We showed participants all relevant tool features and answered questions.

In the second part, participants read a handout describing the dataset and instructions about the scenario. These instructions mentioned that the datasets were "a small but representative sample" and that they should find and report "any reliable observations" that could be used to understand the "site's customer population" or "patient's sleeping patterns" or could be used to improve "customer growth" or provide "sleep recommendations". The handout stated that participants should write down textual descriptions (using the system's note-taking tool) about observations they want to report. After clearing up potential questions about the instructions, participants were given up to 15 minutes to analyze the dataset at their own pace. At any time, participants who felt they exhausted the use case could stop. During this second part, we instructed users to think-aloud [18] and we captured screen and audio-recordings and eye-tracking data. An experimenter was present throughout all of the session, and users were encouraged to ask technical questions or questions about the definition of dataset attributes.

In the third part, the experimenter and participant re-watched video recordings of the session together (with overlaid eye-tracking data). This was an involved process where the examiner paused playback at every interaction, instructed users to explain their thought process, re-wound if necessary and let

[1]The code used in this study to generate synthetic datasets and run empirical tests can be found at https://github.com/zheguang/macau.

participants recount which parts of visualizations they were observing (reinforced by the eye-tracking data) and what visual features they had been looking for. The examiner asked detailed questions about the above points if needed. In a post-session questionnaire, participants ranked their background in statistics, familiarity with statistical hypothesis testing, and experience interpreting visualizations on a 5-point Likert scale.

ACCURACY OF USER INSIGHTS

For our analysis, we considered all insights reported by users through the tool's note-taking feature with a few noted exceptions. We excluded insights that were based on prior knowledge or personal experience, that were not directly observable from the dataset, that were based on reading numbers and in no way made a broader statement applicable to a larger population or that misintepreted the visual display. Examples of excluded insights include: "Users wealthy on average compared to median income in the US", "design: 399 red, 329 green, 195 yellow, the rest blue" or "Between stress level and average age, the people with stress level 5 tend to be the oldest at 40 among females" (the participant thought they were filtering to only females but in fact did not). In total, we excluded six insights from five participants. We were left with an average of 5.536 ± 2.742 insights[2] per participant ($n = 28$).

Since the datasets used in this study were generated synthetically, each user-reported insight had a known boolean ground truth label. For example, we know whether the insight "age is correlated with purchase amount" is true since the model generating the dataset contains this information. If age and purchase amount are sampled from independent normal random variables, this insight is false. If the two variables are correlated, it is true. For roughly 70% of the insights, ground truth labels were extracted analytically by inspecting variable relationship in the dataset-models. For the remainder (e.g., statements about means like "the average age of people is between 35 and 40"), we used empirical methods as described (§Datasets). For example, if a user made 10 observations, on average, we generated ground truth labels empirically for 3 of those under Bonferroni correction (over only those 3 insights) on larger datasets.

We modeled this experiment as a binary classification problem and borrow standard techniques and metrics from the machine learning community to evaluate the results. For each insight, a user, depending on how it was reported, either made a positive observation ("Age and purchases look correlated") or a negative one ("There is no correlation between age and purchases"). We summarize the accuracy of a user's insights in a confusion matrix where an insight falls into one of four categories: *True positive* (TP): the user insight is a positive observation and the ground truth agrees, *false positive* (Type I error, FP): user insight is positive but the ground truth says otherwise, *true negative* (TN): user insight is negative and the ground truth agrees and finally *false negative* (Type II error, FN): user insight is negative and the ground truth disagrees. The insight "There is no correlation between age and purchase amount," for example, would fall into the FN category if in our dataset-model the age and purchase amount values were sampled from two correlated random variables.

We report the following averages: $TP = 1.000 \pm 1.217$, $FP = 3.250 \pm 2.287$, $TN = 1.250 \pm 1.404$ and $FN = 0.036 \pm 0.189$. Additionally, we computed the following per-user metrics:

$$\text{Accuracy (ACC)} = (TP+TN)/(TP+TN+FP+FN)$$
$$\text{False discovery rate (FDR)} = FP/(TP+FP)$$
$$\text{False omission rate (FOR)} = FN/(TN+FN)$$

Where ACC measures the overall accuracy (the percentage of times users' insights were correct), FDR the percentage of times users reported an insight as positive ("age and purchase amount is correlated") but it turned out not to be true and FOR the percentage of times users reported an insight as negative ("there is no relation between age and purchase amount") but it turned out not to be true. ACC summarizes overall performance, whereas FDR and FOR give a more detailed view about where mistakes where made. We found that the average ACC across all users is 0.375 ± 0.297, the average FDR is 0.738 ± 0.296 and the average FOR is 0.018 ± 0.094.

Our study featured a full-factorial 2 dataset-types (*shopping* and *sleep*) $\times$ 2 dataset-sizes (300 and 1000 records) study design. We applied an analysis of variance test (ANOVA) with dataset-type and dataset-size as the between-subject factors. We found that dataset-type as well as dataset-size had no significant effect on accuracy ($p = 0.792, \eta^2 = 0.003$ and $p = 0.091, \eta^2 = 0.109$ respectively, $\alpha = 0.05$).

CONFIRMATORY STATISTICAL HYPOTHESIS TESTING

The results of our study show that for our synthetic datasets and the particular tool we used, over 60% of user reported insights were wrong. Results from visual analysis are often considered exploratory. They need to be confirmed in a second phase. In fact, roughly a fourth of our study participants mentioned at one point or another that they would want to verify a reported insight through statistical testing. In this section, we report on an experiment where we validated user insights through different confirmatory analysis approaches.

The approaches we used included *confirmation; same dataset* and *confirmation; validation dataset*. *Confirmation; same dataset* models an approach where statisical analysis is done on the same dataset used in the exploration phase. *Confirmation; validation dataset* follows Tukey's model [47] of conducting exploratory and confirmatory data analysis on two separate datasets; for this we generated a new dataset of the same size and with the same parameters (i.e., random variables have the same mean and variance and the correlation coefficients between variables are the same) as used during exploration. For both approaches, we use the Benjamini and Hochberg procedure [6] to correct for multiple hypotheses.

From Insights to Statistical Tests

To perform this experiment, we converted user insights into testable statistical hypotheses via multiple steps. We first created an encoding scheme based on insight classes. Insights were coded as instances of these classes. An insight became an object where its type and necessary properties were defined

[2] Averages appear with the standard deviation as the second number.

by its class. We then defined null hypotheses and testing procedures for each insight class. To transform insights into testable hypotheses: the insight class indicates the statistical test to use and the properties of the encoding inform that test's input parameters. The following sections explain these steps.

Insight Classes

Our goal was to find a classification model with the fewest classes that could still accurately describe all insights gathered in our study. We employed a process where we first generated candidate classes which we then iteratively refined. In this process we considered all available data from our user study. This includes the video, audio and eye-tracking recordings, the textual description of insights provided users, as well as participant commentary that we gathered when re-watching session videos with the participants.

We arrived at a system that encompasses five insight classes: *shape*, *mean*, *variance*, *correlation* and *ranking*. *Mean* and *variance* described insights with direct statements about the means or variances of distributions. *Correlation* considered all insights where a relationship between two variables was established. *Shape* covered observations about the shape of one or more distributions, and, finally, *ranking* included observations about sub-population rankings or orderings. Each class of insight defined several properties that fully described its class instances, such as which attributes were involved, which sub-populations were getting compared, whether parts of the data were filtered out and what comparison were being made (e.g., is something smaller or bigger than something else).

Coding

We describe our 155 insights as instances of their corresponding classes. On average per participant we encoded 1.250 ± 1.404 *correlation*, 2.786 ± 2.347 *mean*, 1.143 ± 1.860 *ranking*, 0.821 ± 1.517 *shape* and 0.107 ± 0.315 *variance* insights. Following Liu et al. [35], the first author did the majority of the coding, revising it with co-authors to reduce bias.

Figure 3 illustrates examples of user reported insights from our study. It shows the visual display that triggered an insight alongside the textual description provided by participants and the corresponding insight class, with its properties, that encoded the insight. Note that we again relied heavily on the commentaries made in our post-session video review with the participants, as well as our recorded eye-tracking data. Figure 3 (c) depicts an instance where the user's statement alone did not make the mapping to an insight class obvious; however, post-session commentary and eye-tracking data resolved this.

Mapping Insight Classes to Null Hypotheses

We now need to convert insights we encoded as class instances into testable hypotheses. For each insight class, we define a general null hypothesis pattern that we can fill out with specifics from the actual insight. For example, the null hypothesis pattern for the *mean* class is $E[X] = E[Y]$, so an insight that "the average age is 50" would populate this pattern as: $H_0 : E[age] = 50$. Table 1 shows all null hypotheses patterns.

There are certain ambiguities in these translations from insights to hypothesis. For instance in the above example, the histogram bin for age 50 was the highest but the bin width was 5. So the user was more likely to imply that the population mean was around but not exactly 50. We modified null hypotheses in such cases to account for this level of ambiguity by adding an interval of 10% around the hypothesized mean. Another drawback of this null hypothesis mapping is that we need the null hypothesis to be testable. For example, if the user insight specifies a certain order of the age groups, then conceptually we should test against all other possible orders. However, this null hypothesis is very hard to test statistically. Thus, we chose uniformity as the null hypothesis, meaning no particular order in all the age groups. In general, we resolve ambiguities by erring on the side of users by choosing more relaxed null hypotheses where statistical results are more likely to "agree" with user judgment.

For each null hypothesis pattern we define a corresponding Monte Carlo permutation or bootstrap test. We chose resampling for hypothesis testing because it offers several advantages over the parametric testing such as the t-test and the χ^2-test. First, randomization tests do not assume the distributions of the test statistics [15]. Moreover, some parametric tests such as the χ^2-test require samples to be large enough to make accurate inferences [7]. However, many user insights were based on skewed data or highly selective filters, and hence might not always meet the sample size requirement.

In general, the Monte Carlo permutation or bootstrap tests share a common computational form of resampling [15]. First a test statistic is determined based on the hypothesis. Then the data is permuted or bootstrapped to obtain the distribution of the test statistic under the null hypothesis. Finally the proportion of the test statistics in permutations that are more extreme than in the user observation forms the estimated p-value, $\hat{p}$. To determine how many permutations, n, we needed for a sufficiently accurate estimate, we used the Central Limit Theorem to derive the 95% confidence interval [7]:

$$\hat{p} \pm 1.96 \sqrt{\hat{p}(1-\hat{p})/n}$$

With enough permutations, we get a non-overlapping interval $\hat{p}$ for significance α, following the decision rule to reject the null hypothesis if it is not greater than α. We summarize the details of the randomization tests in Table 1.

Analysis

Using the above procedure, we mapped all insights to hypotheses tests. Depending on the confirmatory approach, we computed test results on either the same dataset shown to users (*confirmation; same dataset*) or a newly generated validation dataset (*confirmation; validation dataset*). We again modeled this experiment as a binary classification problem. Statistical significance ($\alpha = 0.05$) provided positive or negative insight predictions which were then evaluated against ground truth labels. Figure 4 reports experimental results including individual datapoints, means and 95% confidence intervals.

MIXING EXPLORATION AND CONFIRMATION

The two confirmatory analysis approaches outlined and compared in the previous section have their drawbacks. For example, while *confirmation; validation dataset* is statistically

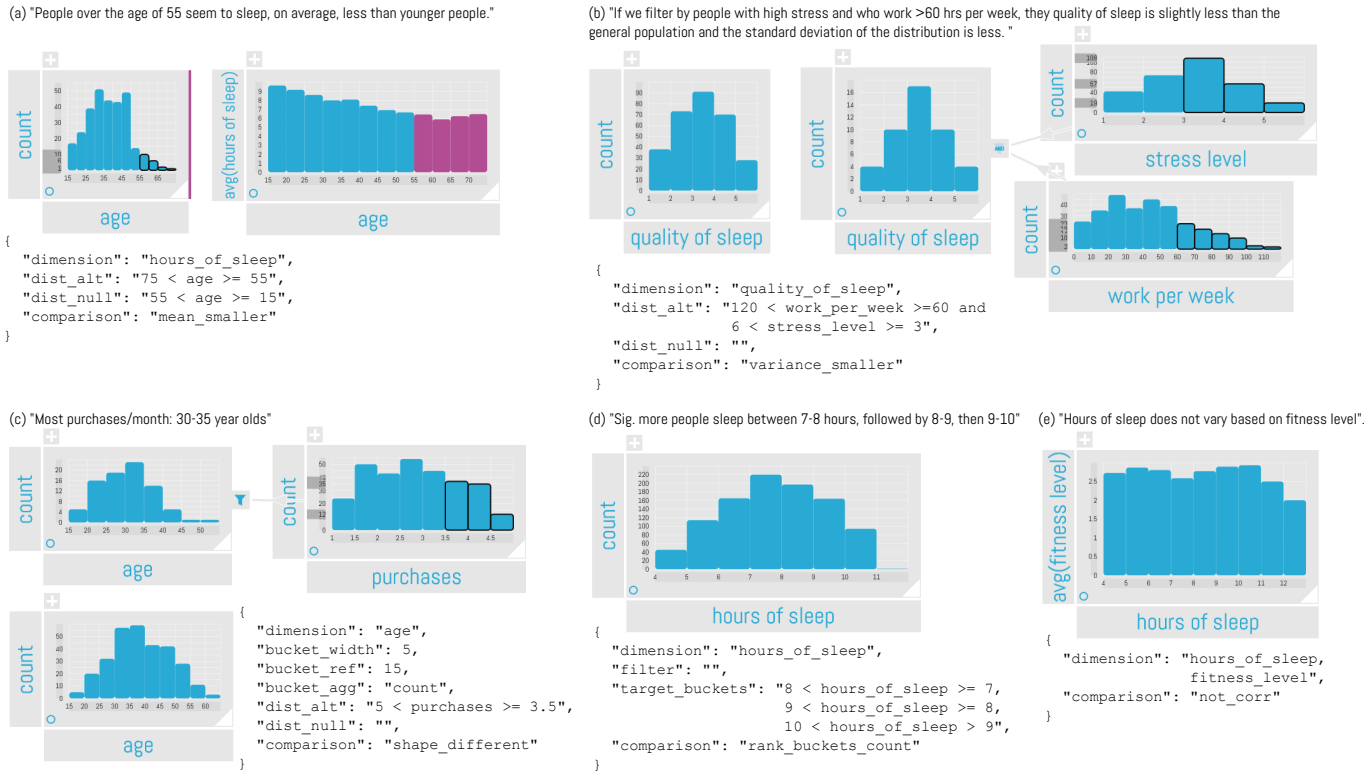

Figure 3. Examples of user reported insights from our study. The figure shows the visual display that triggered an insight alongside the textual description participants reported and the corresponding insight class with its properties we encoded it to. (a) An example of a *mean* insight. The user directly mentions that he is making a statement about averages. We encode the dimension that we are comparing across ("hours of sleep"), the two sub-populations that are being compared ("75 < age >= 55" and "55 < age >= 15") as well as the type of comparison ("mean_smaller"). (b) Our user compares the standard deviation of the two "quality of sleep" charts. We encode this as a *variance* insight. We again describe this instance fully by recording the dimension involved, the sub-populations compared and the type of comparison being made. (c) Example of a *shape* class insight. From the user statement alone, it was not obvious to which class this insight corresponded. However, in the post-session video review, the participant mentioned that she was "looking for changes in age distribution for different purchases" and that she observed a change in the shape of the age distribution when filtering down to high purchase numbers. This was reinforced by analyzing the eye-tracking data of the session. The participant selected bars in the "purchases" histogram and then scanned back and forth along the distribution of the filtered "age" visualization and the unfiltered one. (d) An example of an insight where a *ranking* among parts of the visualization was established. (e) The user created a visualization with two attributes. The y-axis was mapped to display the average "fitness level". Our user notes report insights that discuss the relationship of the two attributes. We classified this as a *correlation* insight.

Insight Class	Null Hypothesis	Permutation π	Test Statistic
Mean	$E[X] = E[Y]$	$X \cup Y$	$\lvert\mu_X - \mu_Y\rvert$
Variance	$var(X) = var(Y)$	$X \cup Y$	$\lvert\sigma_X^2 - \sigma_Y^2\rvert$
Shape	$P(X\lvert Y = y_1) = P(Z\lvert Y = y_2)$	Y	$\lVert P(X\lvert Y = y_1) - P(Z\lvert Y = y_2)\rVert$
Correlation	$X \perp Y$	X	$\lvert\rho(X,Y)\rvert$
Ranking	$X \sim Unif(a,b)$	$\pi \sim Unif(a,b)$	$\begin{cases} 1 & rank(X_\pi) = rank(X_{obs}) \\ 0 & \text{else.} \end{cases}$

Table 1. Summary of randomization hypothesis tests to which insights are mapped to for confirmatory analysis. The random variables represent attributes with arbitrary conditions from the dataset.

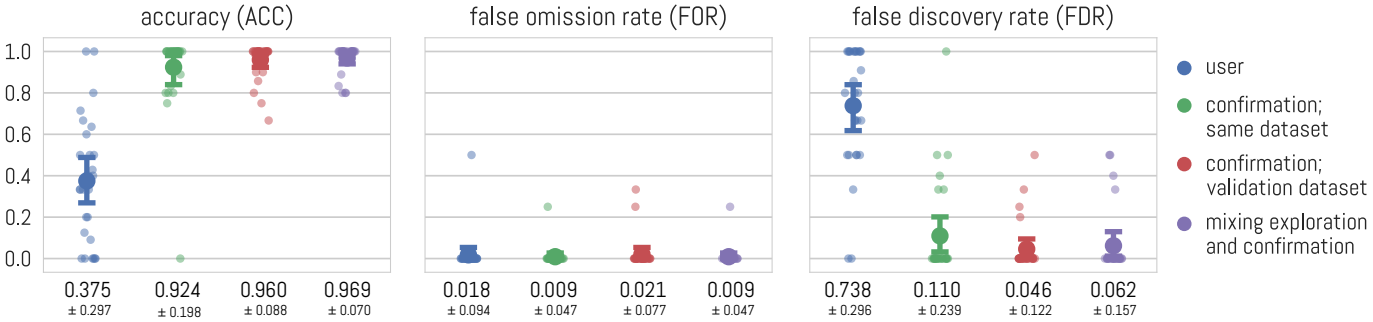

Figure 4. Plot with average scores, where the second number is the standard deviation, and rendered 95% confidence intervals for accuracy (ACC), false omission rate (FOR) and false discovery rate (FDR) for users and different confirmatory approaches. Overlaid are all individual datapoints ($n = 28$).

sound, it requires users to collect additional data which is often unfeasible in practice. Data collection might be expensive, as is the case for user studies, medical trials or crowd-sourcing, or might be done by an outside provider over which the user has no control. Splitting a dataset into exploratory and confirmatory parts significantly lowers the power of comparison done on either part due to smaller sample sizes. Perhaps more problematic, *confirmation; same dataset* can lead to inherent systematic bias because we are statistically testing insights on the same data that initially informed them.

We want to compare these approaches to one we call *mixing exploration and confirmation*. So far we have only analyzed users reported insights. We call these *explicit* insights. However, during an exploration session, users might have made a significant number of along-the-way comparisons that did not trigger insights. The reasons for these could be manifold. The comparisons may have involved uninteresting visualizations, confirmed an assumption the user already held or just been inconclusive. Regardless, the result of such a comparison is still a type of insight: that the result of the comparison was not interesting enough to be reported. We call these *implicit* insights. The previous confirmatory approaches ignore *implicit* insights completely. With *mixing exploration and confirmation* we simulate an approach where comparisons that resulted in either type of insight, *explicit* or *implicit*, were incorporated.

Coding

We again considered all data collected from our study: video, audio and eye-tracking recording and commentary from participants. This time we focused on *implicit* insights where users made an unreported comparison, likely because it was uninteresting. We encoded such instances with the same insight classes as before. For example, a user might view "age" for two different sub-populations; eye-tracking data indicates visual scanning between the two; the post-session commentary reveals the user compared the two trends but did not see a difference. Our coding marks this as an *implicit shape* insight.

Analysis

Overall we found 620 *implicit* insights, with 22.143 ± 12.183 average *implicit* insights per user. Following the previous procedure, we converted these *implicit* insights into statistical tests. We conducted statistical analysis as with the other confirmatory approaches, but added tests based on *implicit* insights to the multiple hypotheses correction procedure (Benjamini and Hochberg [6]). We used the same p-value cutoff as before and report the same metrics (see Figure 4).

Note we did not report *implicit* test accuracy. All metrics were solely based on *explicit* insights since we only cared about their correctness. Consider again the example from the introduction. Jean made nine *implicit* and one *explicit* insights but only shared the *explicit* one. Only the accuracy of the explicit one matters since only it will be acted upon; yet its accuracy depends on the set of *implicit* insights.

DISCUSSION

Real-world datasets are weighted compositions of noise and signal. One goal of visual data analysis is to assist users at efficiently separating the two. We want visualization systems where users can maximize their total number of insights while minimizing false insights. Insight-based methods only compare systems based the former. Analyzing errors requires quantification of the correctness of insights. For real-world datasets this is frequently not possible because there is no way to know which effects are true or not.

In this paper we use a method based on synthetic datasets where we can classify the correctness of an insight as either true or false. Our notion of insight is limited to observations, hypotheses and generalizations directly extracted from the data. If a user tells us "age" and "hours of sleep" are correlated we know if that statement is true or not.

For the visualization tool used in our study, over 60% of user reported insights were found incorrect. However, this error needs to be interpreted anecdotally. Results may vary greatly between users, visualizations, tools and datasets. The high error rate is perhaps unsurprising. Mathematical procedures might be better suited to make such data-driven inferences than humans. More surprisingly, when following up user generated insights with statistical tests on the same dataset, we are still left with 11% false discoveries (*confirmation; same dataset*, Figure 4). Double of what statistics promise when using a significance level of 5%. This is because we introduced systemic bias by testing hypotheses on the same dataset that informed them and hence FDR was inflated due to MCP.

The key takeaway here is that without either confirming user insights on a validation dataset (*confirmation; validation dataset*) or accounting for all comparisons made by users during exploration (*mixing exploration and confirmation*) we have no guarantees on the bounds of the expected number of false discoveries. This is true regardless which specific visual analysis tool or visualization is used. Taking action, making decisions or publishing findings this way becomes risky.

Validating user generated insights with the *confirmation; same dataset* approach is not statistically sound and *confirmation; validation dataset* requires additional data, which in practice is often hard to acquire. In our experiments we manually coded *explicit* and *implicit* insights to show the benefits of *mixing exploration and confirmation*: it guarantees the same FDR bounds as confirmation on a validation dataset. However, burdening users to remember and code all of their insights during an exploration session is unfeasible. Could we create tools that automatically do this encoding while users explore a dataset? We believe that augmenting visual analysis systems to do this is a promising direction for future research.

User Performance

We next examine several questions about user insights and performance. Is user accuracy correlated to self-reported expertise levels in statistical hypothesis testing, interpreting visualization and general statistical background? Is accuracy influenced by individual insight support size? Does the study design pressure users near the end of a session?

We correlated participant accuracy scores with their self-reported expertise levels. We found that neither background in

statistics ($r = -0.258, p = 0.223$), their familiarity with statistical hypothesis testing ($r = -0.328, p = 0.117$) or their experience with interpreting visualizations ($r = 0.005, p = 0.982$) had a significant correlation with accuracy percentages. In general we focused on novice participants since we believe that they are a large part of the target audience for visualization tools. More experiments are needed to assess if our results generalize to other populations (e.g., statistical experts).

The average normalized support (i.e., the percentage of data-records involved in an insight) per user was 0.590 ± 0.270 for correct insights and 0.463 ± 0.282 for incorrect ones. While the difference is not statistically significant ($p = 0.118, d = 0.471$) examining this trend in more detail is warranted.

We extracted timestamps of insights and normalized them by session length. The average normalized time for incorrect and correct insights is 0.585 ± 0.271 and 0.587 ± 0.248 which is not a statistically significant difference ($p = 0.957, d = 0.009$).

Relating to Statistical Theory

Without any hypothesis testing, the false discovery rate averages over 73% (*user*, Figure 4). With hypothesis testing on the same dataset (*confirmatory; same dataset*) the false discovery rate reduced somewhat to 11% (Figure 4). With multiple hypotheses control only on the *explicit* insights, the average FDR inflated above the theoretical bound of 5%. By not controlling for *implicit* tests, we are exposed to the multiple hypotheses error as described in section Multiple Comparisons Problem in Statistics. Essentially this is a misuse of the control procedure.

With proper multiple hypotheses correction on both *implicit* and *explicit* hypotheses *mixing exploration and confirmation* achieved average false discovery rates around 5% (Figure 4). This can be seen from the theoretical perspective where the Benjamini and Hochberg procedure [6] guarantees the expected proportion of false discoveries V among all discoveries R is upper bounded by a given significance level $\alpha = 0.05$:

$$E[\|V\| / \|R\|] \leq 0.05$$

We achieved a similar false discovery rate with *confirmation; validation dataset* which tested hypotheses on a separate dataset. This is akin to replication studies in science.

Statistical procedures only provide bounds on the expected number of false discoveries. Tightening these bounds will automatically result in higher false omission rates. Balancing this trade-off is highly domain specific. In drug trials, false discoveries must be avoided, whereas, in security related scenarios, false omissions can have disastrous effects. Sound statistical methods, like *confirmatory; validation dataset* and *mixing exploration and confirmation*, facilitate these trade-offs.

Larger Datasets

Theoretically, having complete knowledge of the population distribution, or infinite resource to sample from it to appeal to the Law of Large Numbers, could eliminate the risk of false discovery. However, in many practical cases approximating this theory is difficult. Fundamentally, many factors affect the uncertainty of the statistical inference on the data, including sample sizes, variations, effect sizes, and measurement errors [21]. Thus it requires significant upfront effort to determine the sample size with enough statistical power by controlling the other factors. Some factors may be hard to compute. For example, the space of possible hypotheses on the data may not be known; users may select and compare many different data subsets. These aspects also complicate the notion of having a single sufficiently large data size for all possible analysis. Yet studies are warranted that explore in detail the relationship between user accuracy and data size.

Base Rate Fallacy and Other Errors

The psychological tendency to neglect the global base rate while overestimating by the local, more specific likelihood has been well studied in Pscyhology and Behavioral and Brain Sciences [4, 32]. In the context of visual analysis, it would be interesting to see how user performance changes if users were informed about the noisiness of the underlying dataset. Interestingly, some statistical testing procedures automatically approximate data randomness to improve their performance [54]. Our study however excludes this variable by fixing the base rate of ground truths and not disclosing it to the participants.

Beyond Type I and Type II, other error types have been proposed to quantify the likelihood of mistaking the sign (Type S) or overestimating the magnitude (Type M) [22, 20]. However in our study user observations were often vague regarding these effects. For example, instead of saying "mean A is 45 and less than mean B" participants would typically say "mean A is different than mean B". Furthermore, sign and magnitude do not apply to several types of user insights (e.g., when comparing shapes or rankings). For the subset of insights where users stated directions, we setup one-sided null-hypotheses which captured sign errors in our FDR calculation. Based on our experience, a more detailed study of Type S and Type M errors would likely involve a new study design that invites the users to be more conscious about making observations on signs and magnitudes.

CONCLUSION

Comparing a visualization to a mental image is akin to performing a statistical test, thus repeated interpretation of visualizations is susceptible to the MCP. In this work we attempted to empirically characterize this. We presented an experiment based on synthetically generated datasets that enabled us to assess the correctness of user reported insigths. We showed that by not accounting for all visual comparisons made during visual data exploration, false discovery rates will be inflated even after validating user insights with further statistical testing. We demonstrated that a confirmatory approach that addresses this can provide similar statistical guarentees to one that uses a validation dataset.

ACKNOWLEDGEMENTS

This research is funded in part by DARPA Award 16-43-D3M-FP-040, NSF Award IIS-1514491 and IIS-1562657, the Intel Science and Technology Center for Big Data and gifts from Adobe, Google, Mellanox, Microsoft, Oracle and VMware. Thanks also to Professors van Dam, Binnig and LaViola for their support and guidance.

REFERENCES

1. Georgia Albuquerque, Thomas Lowe, and Marcus Magnor. 2011. Synthetic generation of high-dimensional datasets. *IEEE transactions on visualization and computer graphics* 17, 12 (2011), 2317–2324.
2. Emmanuelle Anthoine, Leïla Moret, Antoine Regnault, Véronique Sébille, and Jean-Benoit Hardouin. 2014. Sample size used to validate a scale: a review of publications on newly-developed patient reported outcomes measures. *Health and quality of life outcomes* 12, 1 (2014), 2.
3. Peter Ayton and Ilan Fischer. 2004. The hot hand fallacy and the gambler's fallacy: Two faces of subjective randomness? *Memory & cognition* 32, 8 (2004), 1369–1378.
4. Maya Bar-Hillel. 1980. The base-rate fallacy in probability judgments. *Acta Psychologica* 44, 3 (1980), 211–233.
5. Yoav Benjamini. 2010. Simultaneous and selective inference: current successes and future challenges. *Biometrical Journal* 52, 6 (2010), 708–721.
6. Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)* (1995), 289–300.
7. Dimitri P Bertsekas and John N Tsitsiklis. 2002. *Introduction to probability*. Vol. 1. Athena Scientific Belmont, MA.
8. Sabrina Bresciani and Martin J Eppler. 2009. The risks of visualization. *Identität und Vielfalt der Kommunikations-wissenschaft* (2009), 165–178.
9. Andreas Buja, Dianne Cook, Heike Hofmann, Michael Lawrence, Eun-Kyung Lee, Deborah F Swayne, and Hadley Wickham. 2009. Statistical inference for exploratory data analysis and model diagnostics. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* 367, 1906 (2009), 4361–4383.
10. Stuart K Card, Jock D Mackinlay, and Ben Shneiderman. 1999. *Readings in information visualization: using vision to think*. Morgan Kaufmann.
11. Remco Chang, Caroline Ziemkiewicz, Tera Marie Green, and William Ribarsky. 2009. Defining insight for visual analytics. *IEEE Computer Graphics and Applications* 29, 2 (2009), 14–17.
12. William S Cleveland and Robert McGill. 1984. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American statistical association* 79, 387 (1984), 531–554.
13. Michael Correll and Jeffrey Heer. 2017. Regression by Eye: Estimating Trends in Bivariate Visualizations. In *ACM Human Factors in Computing Systems (CHI)*. `http://idl.cs.washington.edu/papers/regression-by-eye`
14. Andrew Crotty, Alex Galakatos, Emanuel Zgraggen, Carsten Binnig, and Tim Kraska. 2015. Vizdom: interactive analytics through pen and touch. *Proceedings of the VLDB Endowment* 8, 12 (2015), 2024–2027.
15. Bernd Droge. 2006. Phillip Good: Permutation, parametric, and bootstrap tests of hypotheses. (2006).
16. Olive Jean Dunn. 1961. Multiple comparisons among means. *J. Amer. Statist. Assoc.* 56, 293 (1961), 52–64.
17. Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Leon Roth. 2015. Preserving statistical validity in adaptive data analysis. In *STOC*. ACM, 117–126.
18. Karl Anders Ericsson and Herbert Alexander Simon. 1993. *Protocol analysis*. MIT press Cambridge, MA.
19. Dean P Foster and Robert A Stine. 2008. α-investing: a procedure for sequential control of expected false discoveries. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 70, 2 (2008), 429–444.
20. Andrew Gelman and John Carlin. 2014. Beyond power calculations: Assessing Type S (sign) and Type M (magnitude) errors. *Perspectives on Psychological Science* 9, 6 (2014), 641–651.
21. Andrew Gelman and Eric Loken. 2013. The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking"and the research hypothesis was posited ahead of time. *Department of Statistics, Columbia University* (2013).
22. Andrew Gelman and Francis Tuerlinckx. 2000. Type S error rates for classical and Bayesian single and multiple comparison procedures. *Computational Statistics* 15, 3 (2000), 373–390.
23. Steven R Gomez, Hua Guo, Caroline Ziemkiewicz, and David H Laidlaw. 2014. An insight-and task-based methodology for evaluating spatiotemporal visual analytics. In *Visual Analytics Science and Technology (VAST), 2014 IEEE Conference on*. IEEE, 63–72.
24. Connor C Gramazio, Karen B Schloss, and David H Laidlaw. 2014. The relation between visualization size, grouping, and user performance. *IEEE transactions on visualization and computer graphics* 20, 12 (2014), 1953–1962.
25. Thomas L Griffiths and Joshua B Tenenbaum. 2001. Randomness and coincidences: Reconciling intuition and probability theory. In *Proceedings of the 23rd annual conference of the cognitive science society*. University of Edinburgh Edinburgh, 370–375.
26. Hua Guo, Steven R Gomez, Caroline Ziemkiewicz, and David H Laidlaw. 2016. A case study using visualization interaction logs and insight metrics to understand how analysts arrive at insights. *IEEE transactions on visualization and computer graphics* 22, 1 (2016), 51–60.

27. Pat Hanrahan. 2012. Analytic database technologies for a new kind of user: the data enthusiast. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*. ACM, 577–578.
28. Megan L Head, Luke Holman, Rob Lanfear, Andrew T Kahn, and Michael D Jennions. 2015. The extent and consequences of p-hacking in science. *PLoS Biol* 13, 3 (2015), e1002106.
29. Sandra L Hubscher and August Strindberg. 2007. Apophenia: Definition and analysis. *Digital Bits Skeptic* (2007).
30. UC Irvine. 2017. UC Irvine Machine Learning Repository. (2017). `http://archive.ics.uci.edu/ml/`.
31. Youn-ah Kang and John Stasko. 2012. Examining the use of a visual analytics system for sensemaking tasks: Case studies with domain experts. *IEEE Transactions on Visualization and Computer Graphics* 18, 12 (2012), 2869–2878.
32. Jonathan J Koehler. 1996. The base rate fallacy reconsidered: Descriptive, normative, and methodological challenges. *Behavioral and brain sciences* 19, 1 (1996), 1–17.
33. Robert Ladouceur, Claude Paquet, and Dominique Dubé. 1996. Erroneous Perceptions in Generating Sequences of Random Events. *Journal of Applied Social Psychology* 26, 24 (1996), 2157–2166.
34. Jing Li, Jean-Bernard Martens, and Jarke J Van Wijk. 2010. Judging correlation from scatterplots and parallel coordinate plots. *Information Visualization* 9, 1 (2010), 13–30.
35. Zhicheng Liu and Jeffrey Heer. 2014. The effects of interactive latency on exploratory visual analysis. *IEEE transactions on visualization and computer graphics* 20, 12 (2014), 2122–2131.
36. Mahbubul Majumder, Heike Hofmann, and Dianne Cook. 2013. Validation of visual statistical inference, applied to linear models. *J. Amer. Statist. Assoc.* 108, 503 (2013), 942–956.
37. Chris North. 2006. Toward measuring visualization insight. *IEEE computer graphics and applications* 26, 3 (2006), 6–9.
38. Catherine Plaisant. 2004. The challenge of information visualization evaluation. In *Proceedings of the working conference on Advanced visual interfaces*. ACM, 109–116.
39. Ronald A Rensink and Gideon Baldridge. 2010. The perception of correlation in scatterplots. In *Computer Graphics Forum*, Vol. 29. Wiley Online Library, 1203–1210.
40. Purvi Saraiya, Chris North, and Karen Duca. 2005. An insight-based methodology for evaluating bioinformatics visualizations. *IEEE transactions on visualization and computer graphics* 11, 4 (2005), 443–456.
41. Tarique Siddiqui, Albert Kim, John Lee, Karrie Karahalios, and Aditya Parameswaran. 2016. Effortless data exploration with zenvisage: an expressive and interactive visual analytics system. *Proceedings of the VLDB Endowment* 10, 4 (2016), 457–468.
42. Tarique Siddiqui, John Lee, Albert Kim, Edward Xue, Xiaofo Yu, Sean Zou, Lijin Guo, Changfeng Liu, Chaoran Wang, Karrie Karahalios, and others. 2017. Fast-Forwarding to Desired Visualizations with Zenvisage. In *CIDR*.
43. James Sundali and Rachel Croson. 2006. Biases in casino betting: The hot hand and the gambler's fallacy. *Judgment and Decision Making* 1, 1 (2006), 1.
44. Tableau. 2017. Tableau Product Description. (2017). `https://www.tableau.com/products/desktop`.
45. TIBCO. 2017. TIBCO Spotfire Product Description. (2017). `http://spotfire.tibco.com/data-discovery`.
46. Christian Tominski. 2006. *Event based visualization for user centered visual analysis.* Ph.D. Dissertation.
47. John W Tukey. 1977. Exploratory data analysis. *Addison-Wesley Series in Behavioral Science: Quantitative Methods, Reading, Mass.: Addison-Wesley, 1977* (1977).
48. Jarke J Van Wijk. 2006. Views on visualization. *IEEE transactions on visualization and computer graphics* 12, 4 (2006), 421–432.
49. Manasi Vartak and others. 2015. SEEDB: Efficient Data-Driven Visualization Recommendations to Support Visual Analytics. *PVLDB* 8, 13 (2015).
50. Hadley Wickham, Dianne Cook, Heike Hofmann, and Andreas Buja. 2010. Graphical inference for infovis. *IEEE Transactions on Visualization and Computer Graphics* 16, 6 (2010), 973–979.
51. Stefan Wilhelm and others. 2012. Moments calculation for the doubly truncated multivariate normal density. *arXiv preprint arXiv:1206.5387* (2012).
52. Emanuel Zgraggen, Alex Galakatos, Andrew Crotty, Jean-Daniel Fekete, and Tim Kraska. 2016. How Progressive Visualizations Affect Exploratory Analysis. *IEEE Transactions on Visualization and Computer Graphics* (2016).
53. Emanuel Zgraggen, Robert Zeleznik, and Steven M Drucker. 2014. Panoramicdata: Data analysis through pen & touch. *IEEE transactions on visualization and computer graphics* 20, 12 (2014), 2112–2121.
54. Zheguang Zhao, Lorenzo De Stefani, Emanuel Zgraggen, Carsten Binnig, Eli Upfal, and Tim Kraska. 2016. Controlling False Discoveries During Interactive Data Exploration. *arXiv preprint arXiv:1612.01040* (2016).