CS 8803-MDS
Human-in-the-loop Data Analytics

Lecture 22

11/09/22

Today's class

Investigating the Effect of the Multiple Comparisons Problem in Visual Analysis

Authors: Tanya, Yanhao

Reviewer: Siddhi, Harshal

Archaeologist: Akshay

Practioner: Cangdi

Investigating the Effect of the Multiple Comparisons Problem in Visual Analysis

Authors: Emanuel Zgraggen, Zheguang Zhao, Robert Zeleznik, Tim Kraska

Tanya Garg, Yanhao Wang

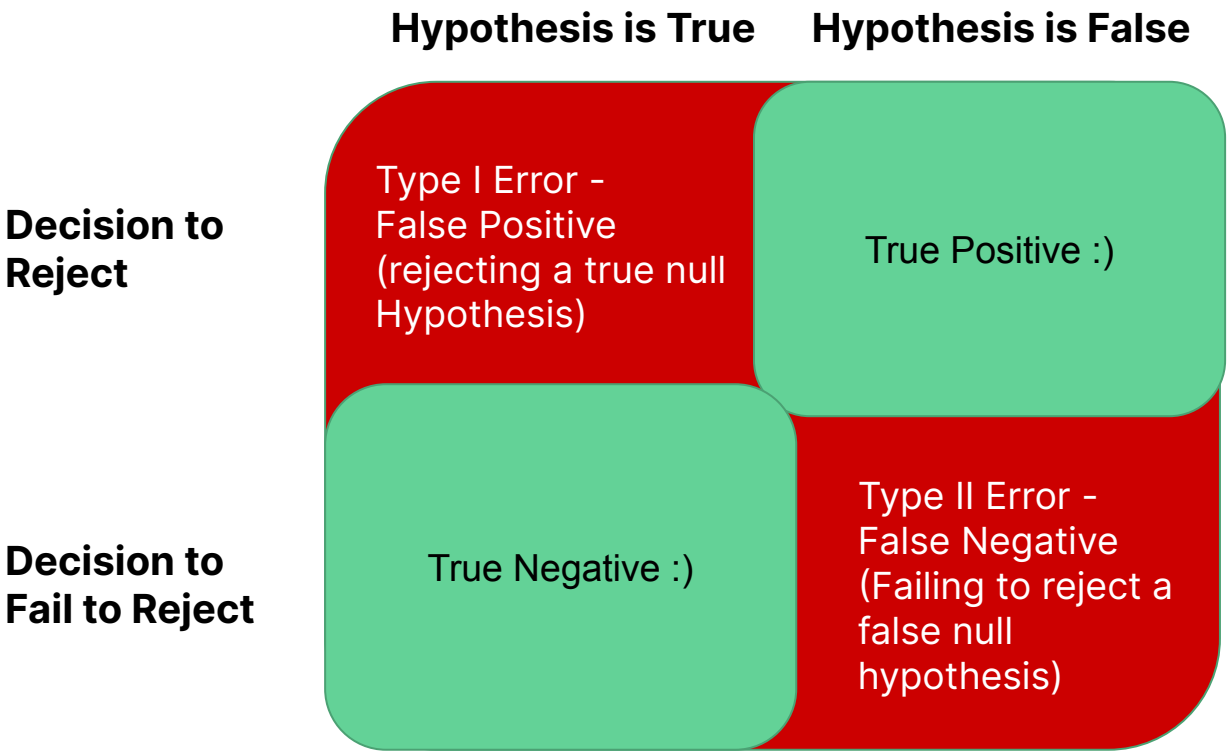
Motivation - Thought Experiment

Task Environment: 1 player with a pair of dice
Winning Condition: Roll a double six

Probability (Win in one trial) = 1/36
Probability (Win in 10 trials) = 0.94
Probability (Win in infinite trials) = 1

What if this becomes a lose condition?

Motivation - Multiple Comparison Problem



Motivation - Random Data Scenario

- **Null Hypothesis:** 2016 saw an improve in donor retention rate due the USB gift.
- Multiple Comparison and unadjusted null hypothesis testing on the same dataset accept the above stated hypothesis.
- Unadjusted null hypothesis testing on the validation dataset and adjusted null hypothesis on the same dataset reject the above stated hypothesis.

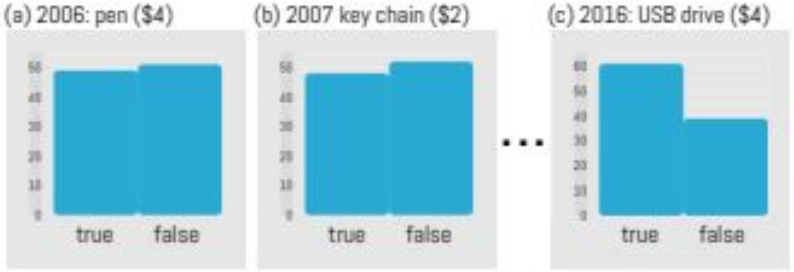
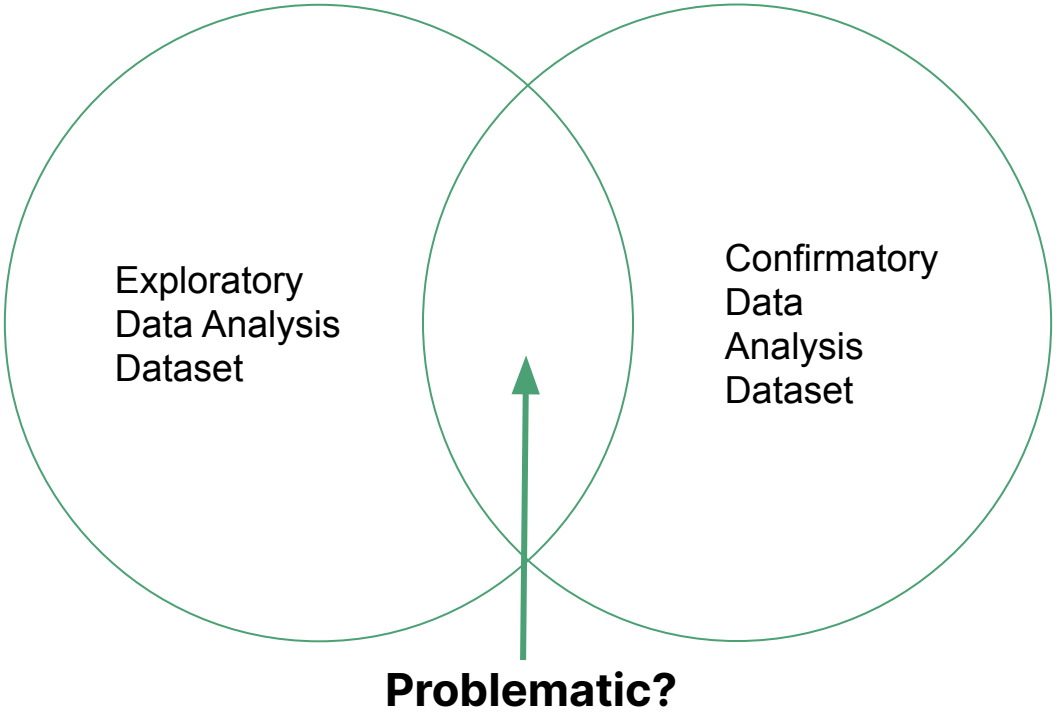


Figure 1. A user inspects several graphs and wrongly flags (c) as an insight because it looks different than (a) and (b). All were generated from the same uniform distribution and are the "same". By viewing lots of visualizations, the chances increase of seeing an apparent insight that is actually the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

Why the visualization community should care?



Related Works

Insight-Based Evaluation

- ➔ Addresses how efficiently a new technique reveals trends or phenomena from data, but **not the potential risk for errors.**
- ➔ Highly subjective and thus misleading
- ➔ Economic model to weigh investments to ROI. Measures the number of insights without any quality weighting ⇒ **Incomplete Picture**
- ➔ Proxy Metrics: Value (domain experts), originality (how often the same insight was reported); We use binary quality score

Visual Inference and Randomness

- ➔ Discovery != Rejection of a null hypothesis
  ⇒ **No results on how MCP affects** visual inference when comparisons are done in series.
  - ➔ People identify random patterns but judge them inaccurately
    ⇒ Can encounter insight due to random noise

Multiple Comparisons Problem in Statistics

- ➔ More than one hypothesis considered at once
  ⇒ risk of observing a falsely significant result increases
  ⇒ multiple hypotheses error
- ➔ **Bonferroni :** Controls family wise error rate (FWER)
  ⇒ statistical power decreases as the number of hypotheses increase
- ➔ **Benjamini-Hochberg** : Controls false discovery rate (FDR), weaker guarantee but better benefits

Synthetic Datasets Experiment

Propose and evaluate strategies for alleviating the concerns brought by MCP

Experimental Method - Visualization Tool

Access to attributes and addition to canvas

Changing Aggregates and Filtering for better analysis

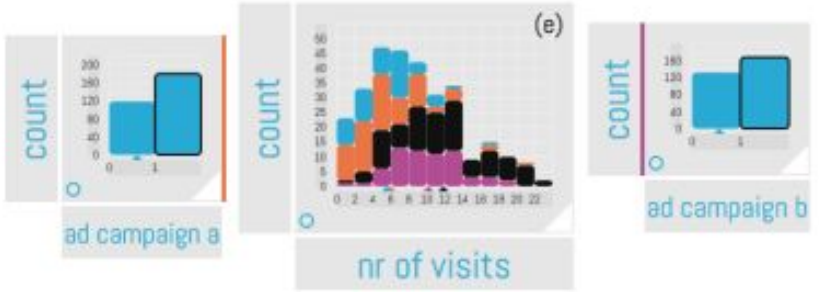
Making textual inference

Experimental Method - Visualization Tool

Access to attributes and addition to canvas

<u>Changing Aggregates and Filtering for better analysis</u>

Making textual inference



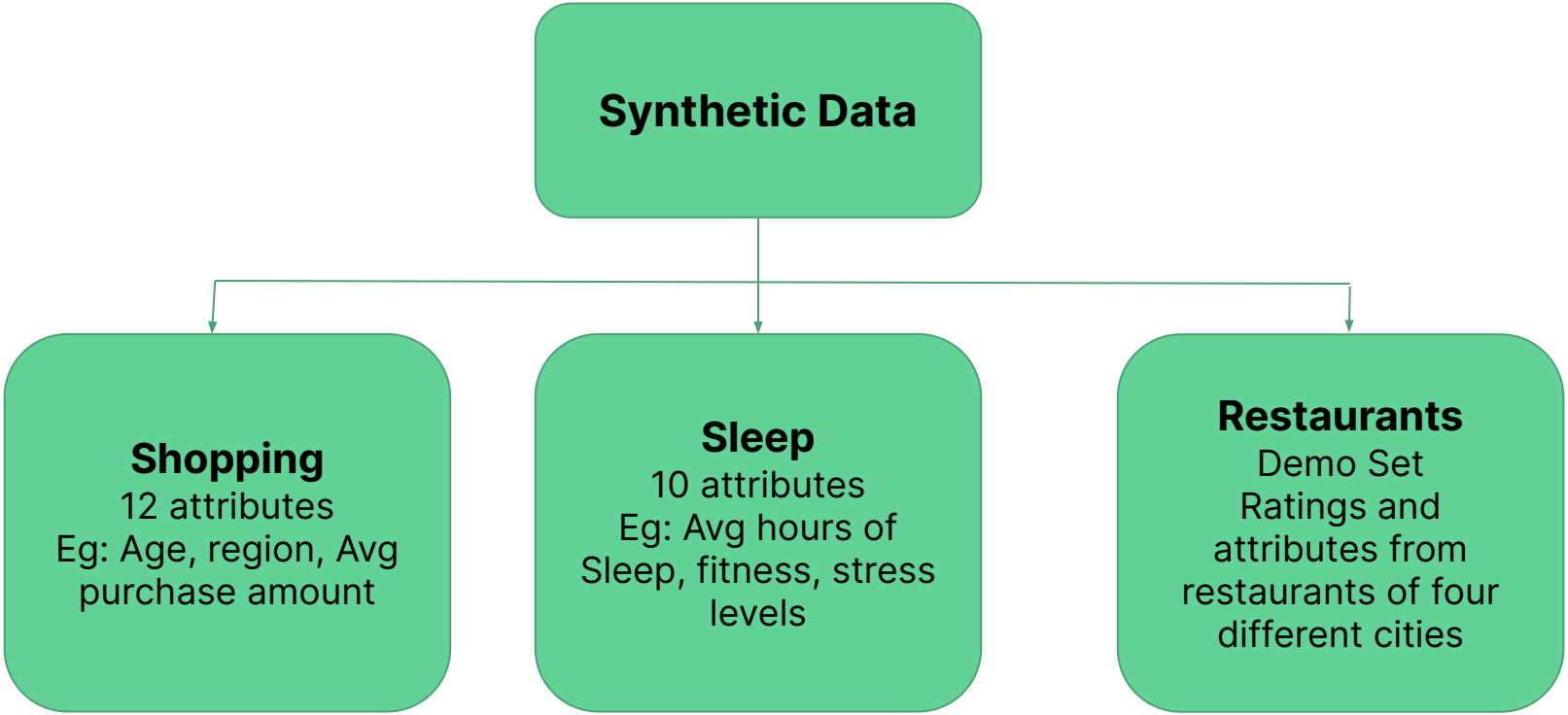
Experimental Method - Visualization Tool

Access to attributes and addition to canvas

Changing Aggregates and Filtering for better analysis

<u>Making textual inference</u>

Experimental Method - Datasets



Experimental Method - Synthetic Data

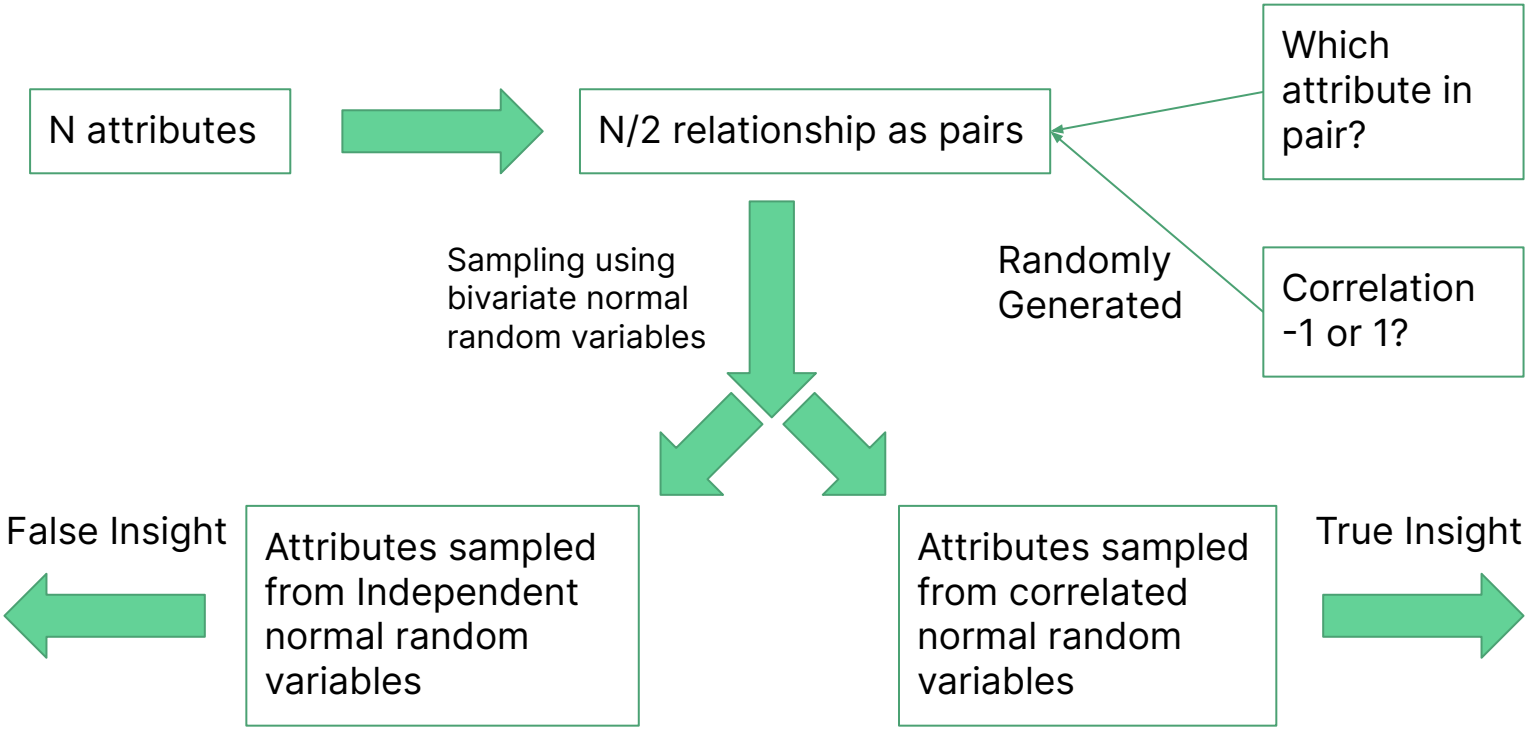
Pre-checks before building a refined synthetic dataset:

✅ Retain domain specific properties : Extracted from Empirical Sample Datasets
✅ Signal + Noise : For real-world replication

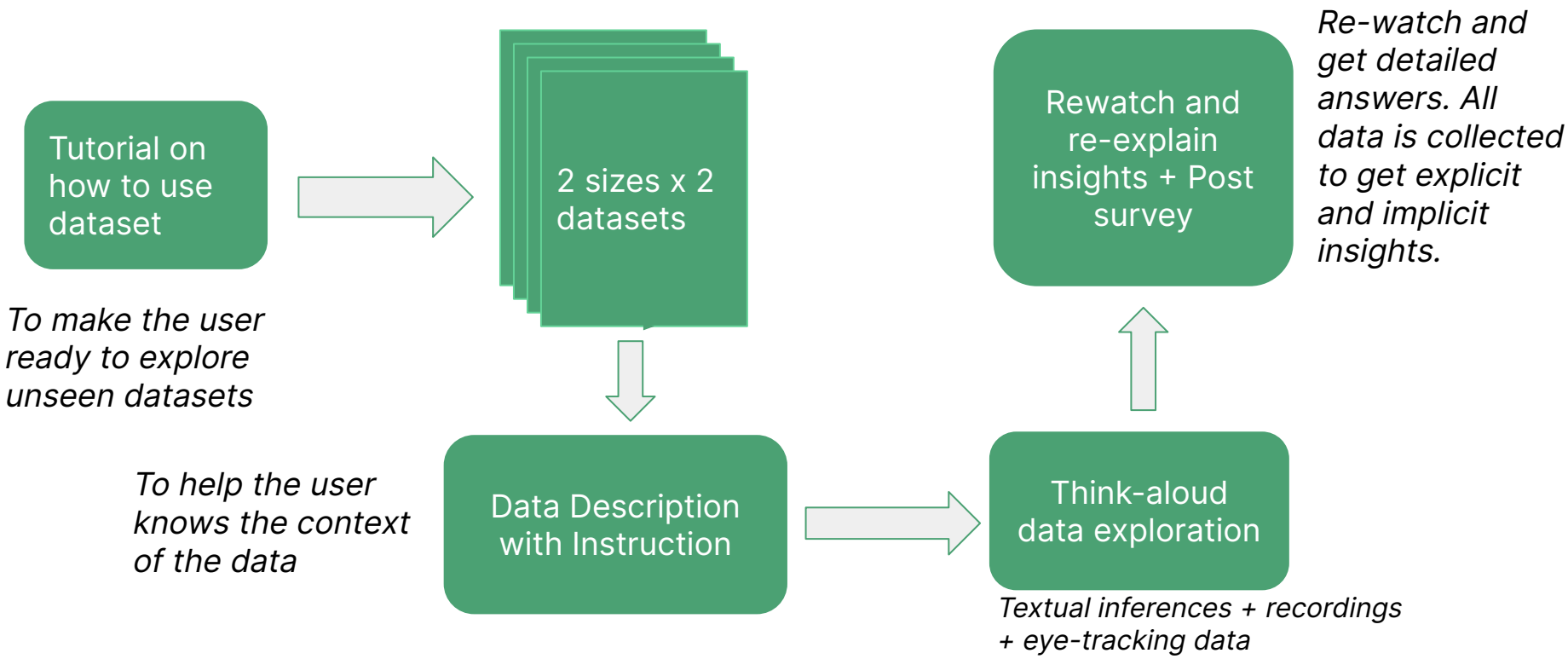
What about Ground Labels?

Six participant pilot study to find most concerned distribution characteristics and relationships among attribute

Experimental Method - Synthetic Data



Experimental Method - Procedure



User Insights

“Observations, hypotheses, and generationlizations **directly extracted from data**”

Explicit Insight

Insights that **user reported** through the system’s note-taking tool

Derived from user’s notes, and consolidated by post-analysis interview and recordings

5.536±2.742 explicit insights per participant

Implicit Insight

Comparisons **made** by users **but unreported**, likely because it was uninteresting

Derived from video, audio, eye-tracking recording, and commentary from participants

22.143±12.183 implicit insights per participant

User Insights

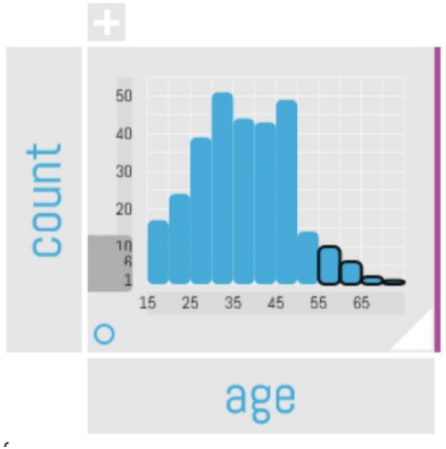
mean

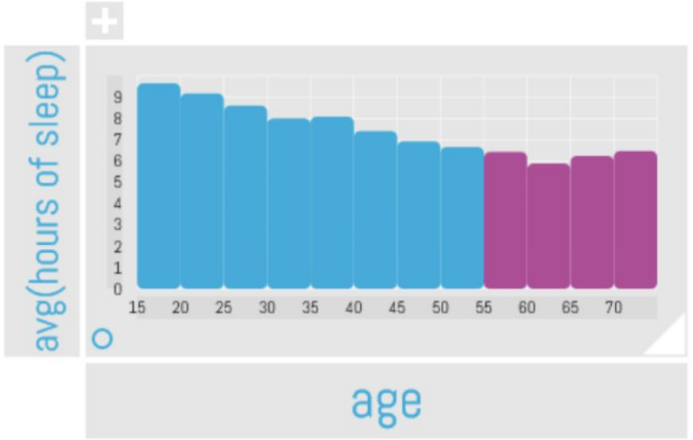
variance

shape

ranking

correlation





"People over the age of 55 seem to sleep, **on average**, less than younger people."

{ "dimension": "hours_of_sleep", "dist_alt": "75 < age >= 55", "dist_null": "55 < age >= 15", "comparison": "mean_smaller" }

User Insights

mean

variance

shape

ranking

correlation

"If we filter by people with high stress and who work >60 hrs per week, they quality of sleep is slightly less than the general population and the **standard deviation** of the distribution is less. "

{ "dimension": "quality_of_sleep",
"dist_alt": "120 < work_per_week >=60 and 6 < stress_level >= 3",
"dist_null": "",
"comparison": "variance_smaller" }

User Insights

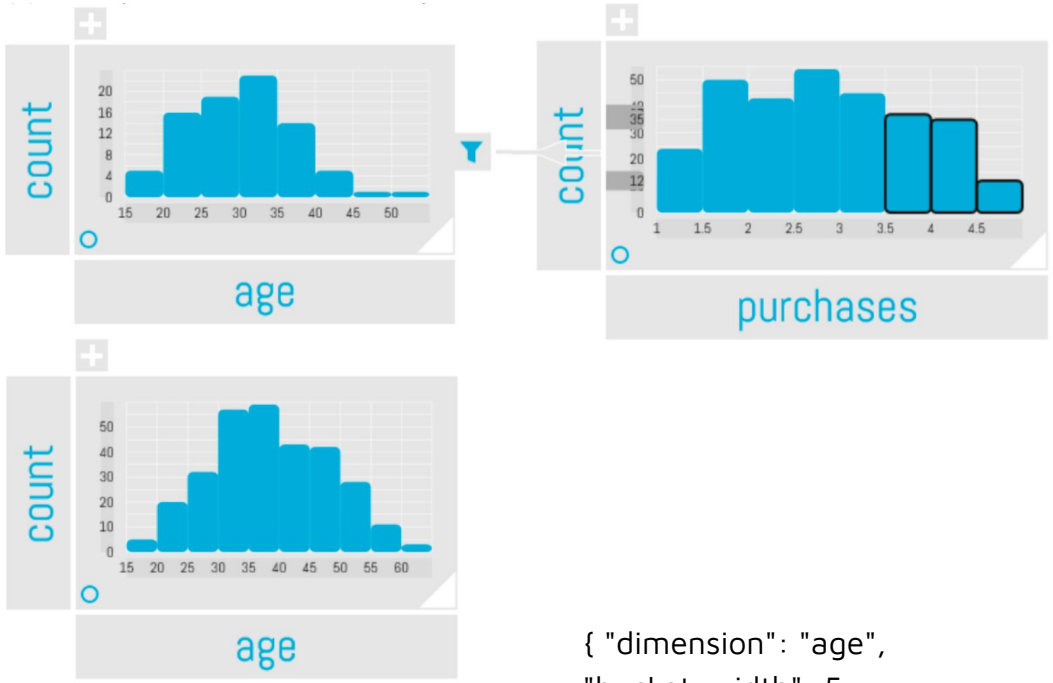
mean

variance

shape

ranking

correlation



"Most purchases/month: 30-35 year olds"
"Looking for changes in age distribution for different purchases"

{ "dimension": "age",
"bucket_width": 5,
"bucket_ref": 15,
"bucket_agg": "count",
"dist_alt": "5 < purchases >= 3.5",
"dist_null": "",
"comparison": "shape_different" }

User Insights

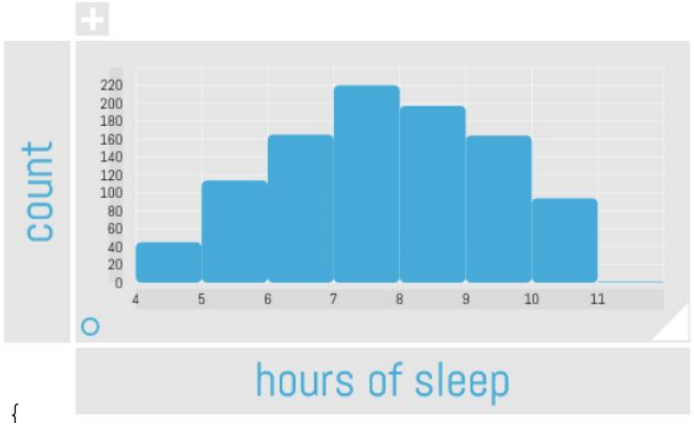
mean

variance

shape

ranking

correlation



{

"Sig. more people sleep between 7-8 hours, followed by 8-9, then 9-10"

{ "dimension": "hours_of_sleep",
"filter": "",
"target_buckets": "8 < hours_of_sleep >= 7, 9 < hours_of_sleep >= 8, 10 < hours_of_sleep > 9",
"comparison": "rank_buckets_count" }

User Insights

mean

variance

shape

ranking

correlation

"Hours of sleep **does not vary** based on fitness level"

{ "dimension": "hours_of_sleep",
"filter": "",
"target_buckets": "8 < hours_of_sleep >= 7, 9 < hours_of_sleep >= 8, 10 < hours_of_sleep > 9",
"comparison": "rank_buckets_count" }

User Insights Evaluation: Ground Truth

mean

variance

shape

ranking

correlation

Generate datasets with 100M record from the same model, and extract ground truth labels using hypothesis testing with Bonferroni correction.

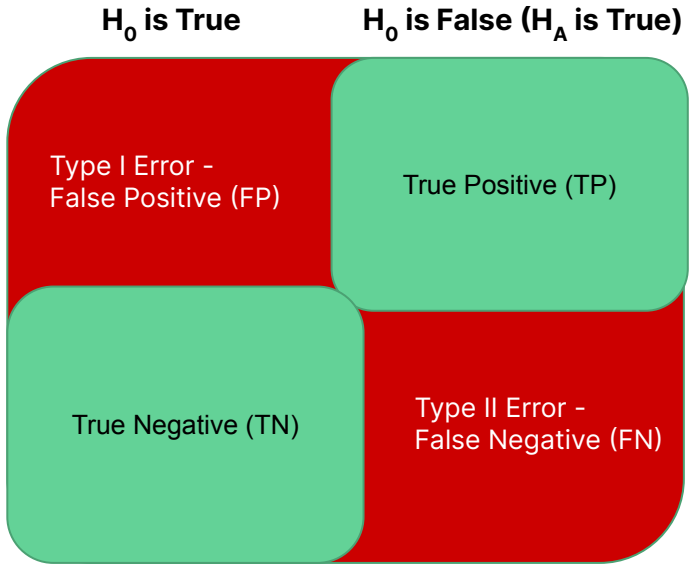
multiple hypothesis correction,
control the probability for at least one Type I error (FWER)

$\alpha_{bon} = \alpha/n$

Directly use the ground truth label from the syntactic dataset.

n/2 true relationships are embedded in an n-attribute dataset

User Insights Evaluation: Metric



Accuracy

$$\frac{TP + TN}{FP + TP + FN + TN}$$

False Discovery Rate

$$\frac{FP}{FP + TP}$$

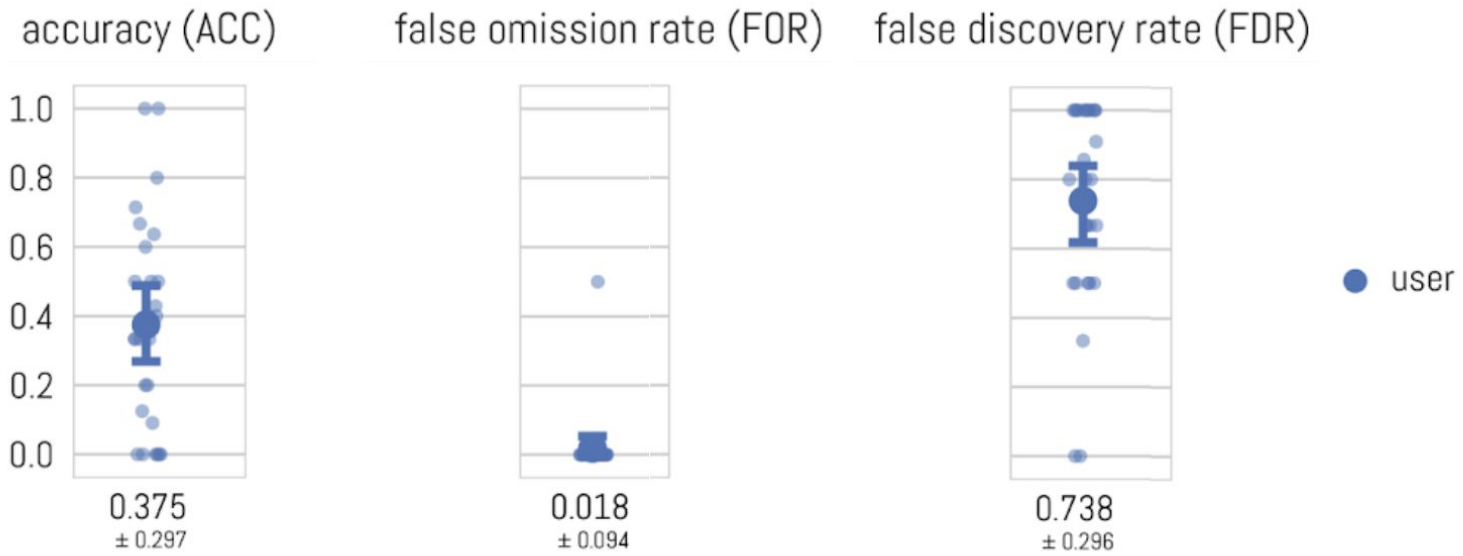
❌ *"age and purchase amount is correlated"*
Reported: -1 or 1, Ground truth: 0

False Emission Rate

$$\frac{FN}{FN + TN}$$

❌ *"there is no relation between age and purchase amount"*
Reported: 0, Ground truth: -1 or 1

User Insights Evaluation



- Over 60% of user reported insights were wrong.
- Low accuracy is mostly contributed by Type I errors (False Positives)!

User Insights Validation

Raw Data

Data Cleaning & Preprocessing

Processed Data

Exploratory Data Analysis

Insights

How convincing?
Need to validate!!

User Insights Validation

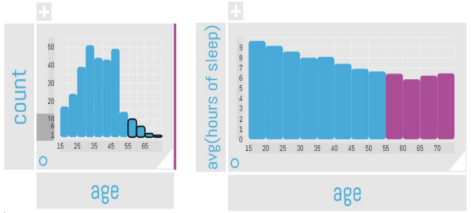
Sleep
Age, average hours of sleep, time to fall asleep, sleep tracker usage...

"People over the age of 55 seem to sleep, on average, less than younger people."

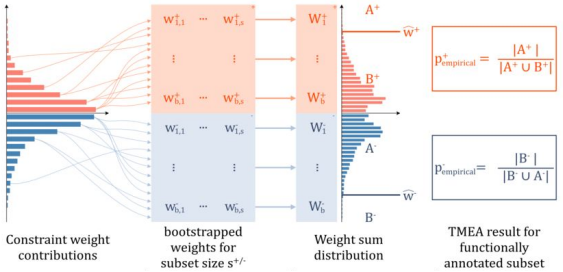
{ "dimension": "hours_of_sleep",
"dist_alt": "75 < age >= 55",
"dist_null": "55 < age >= 15",
"comparison": "mean_smaller"}

H_0:
$E[hours\_of\_sleep_{55<=age<75}] = E[hours\_of\_sleep_{15<=age<55}]$

Data → Insights → Coding → Encoding → Hypothesis



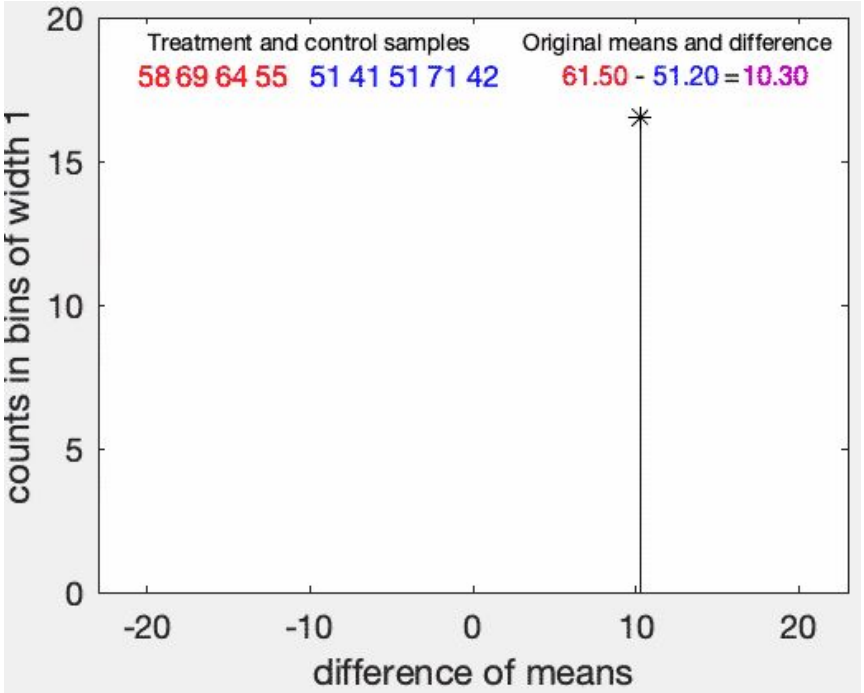
Exploratory Data Analysis



Confirmatory Hypothesis Testing

Confirmatory Hypothesis Testing

- Parametric test
  - Z-test, T-test, Chi-square test, F-test
  - Assume the distribution of test statistics
  - Users sometimes place highly selective filter that skews the data
- Permutation test / Randomization test
  - Main idea: under H_0, shuffling the label and recalculating the test statistics won't matter
  - Generate enough permutations, recalculate the test statistic distribution, and observe where the initial test statistic falls within this distribution



Monte-Carlo Permutation Testing

- Instead of generating all permutations, it uses Monte-Carlo sampling to resample from the subpopulation to build an approximation to the test statistic distribution

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	$X \sim Unif(a,b)$	$\pi \sim Unif(a,b)$	$\begin{cases} 1 & rank(X_\pi) = rank(X_{obs}) \\ 0 & \text{else.} \end{cases}$

Where should we conduct the testing?

- The same dataset?
  - Systemic bias (data dredging / p-hacking)
- Validation dataset?
  - Collect more dataset using the same approach (same size, same parameters)?
    - Statistically sound, but requires additional data collection.
    - Expensive or even infeasible in some scenarios.
  - Split the dataset into exploratory and confirmatory parts?
    - Lowers the power of any tests, because of a smaller sample size
- Or...

Mixing exploration and confirmation

- Taking **implicit insights**, i.e. history of comparing and exploring the dataset, into consideration
    - incorporates along-the-way comparisons into the confirmatory testing
- Conduct similar Monte-Carlo Permutation Testing for confirmatory testing as before
- Add **multiple hypotheses correction** on the **implicit insights**
    - Benjamini-Hochberg correction
        - Designed to control the FDR
        - The α level correction is not uniform for each hypothesis testing (unlike Bonferroni Correction)
        - Correction is varied depending on the P-value ranking $P_k = \frac{k}{n}\alpha$

User insight validation w/ confirmatory analysis

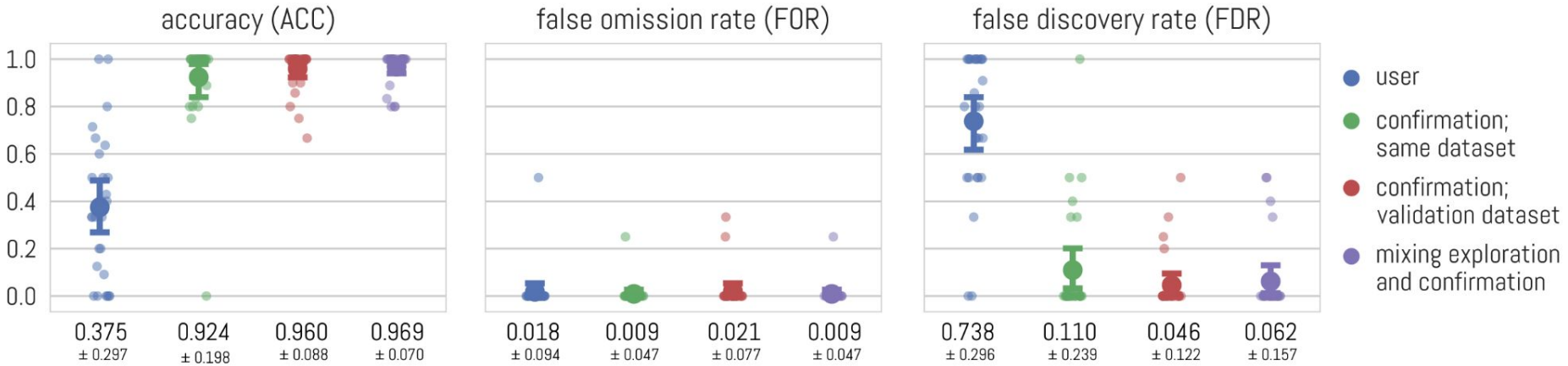


Figure 4. Plot with average scores, where the second number is the standard deviation, and rendered 95% confidence intervals for accuracy (ACC), false omission rate (FOR) and false discovery rate (FDR) for users and different confirmatory approaches. Overlaid are all individual datapoints ($n = 28$).

- Confirming hypotheses on the same dataset reduces FDR down to 11%, but still double of the significance level (5%): FDR inflated due to MCP.
- Confirming on a new data set or mixing exploration and confirmation further reduces the FDR to 4.6% or 6%, respectively, which is similar to the given significance level.

Conclusion & Contribution

- MCP has been well covered in statistics but very much overlooked by the visualization community. Visualization systems are designed to facilitate insight discovery, but pay less attention to the potential large number of false discoveries.
- We empirically characterize the MCP problem in visual exploratory analysis by conducting an experiment based on synthetically generated datasets which allows for assessing the correctness of user-reported insights.
- We examined the effectiveness of three confirmatory methods for validating the insights and controlling the FDR.

Takeaways & Implication

- Nearly 75% of all insights produced by EDA process in the experiment are false discoveries.
    - Raises concerns for the design of analysis tools: viz tools should not only maximize insights, but also minimizing false insights
- Without either confirming user insights on a validation dataset, or accounting for all comparisons made by users during exploration, we have no guarantees on the bounds of the expected number of false discoveries.
    - Taking actions or publishing EDA findings without MCP adjustment can be risky
- Mixing exploration and confirmation produces guarantees the same FDR bounds as confirmation on a validation dataset.
    - Tool that automates the insight encoding procedure will augment visual analysis systems

Thank you!

Tanya Garg, Yanhao Wang

Discussion

Which visualizations are considered hypothesis?

B: is the gender distribution different given salary > 50k?

C: is the gender distribution for salary over and under 50k different?

Discussion

Which visualizations are considered hypothesis?

- Visualization without any filter conditions is nota hypothesis, unless the user makes it one.
- Visualization with a filter condition is a hypothesis with the null hypothesis that the filter condition makes no difference compared to the distribution of the whole dataset.
- If two visualization with the same but some negated filter conditions are put next to each other, it is a test with the null hypothesis that there is no difference between the two visualized distributions, which supersedes the previous hypothesis

Discussion

How to design an interface to mitigate MCP?

- Allow users to see the hypotheses the system assumes
- Hypothesis rejection decisions should never be changed by future interactions
- Users can provide feedback by bookmarking hypothesis

Additional reading

Integrating the control of multiple hypothesis testing into interactive data exploration systems

Controlling False Discoveries During Interactive Data Exploration

Zheguang Zhao Lorenzo De Stefani Emanuel Zgraggen Carsten Binnig
Eli Upfal Tim Kraska
Department of Computer Science, Brown University
{firstname_lastname}@brown.edu

Investigating the Effect of the Multiple Comparisons Problem in Visual Analysis

Reviewer: Siddhi Pandare

Summary of Contributions

1. The paper proposes a method to evaluate the Multiple comparisons problem in data visualization.
2. They present an experiment which uses synthetically generated datasets. Based on the correctness of the insights from the participants of the experiment they substantiated their argument.
3. They illustrated a confirmatory data analysis approach without using a holdout dataset and reported that it can provide similar statistical guarantees.

Strengths

1. Transforms the problem of data comparisons/insights to statistical hypothesis tests.
2. Poses questions about reliability of the user in finding insights.
3. The experiments are thorough and detailed. Addresses several challenges in generating synthetic data for the study like retaining domain specific properties of empirical distributions.
4. They incorporated implicit insights in the confirmatory statistical hypothesis testing.

Weakness

1. The paper assumes that any non-correlated data is false discovery. However other correlations like nonlinear, nonparametrically, potentially highly complex are not mentioned.
2. Are false insights really harmful? The analyst can observe the deviation in the graph but what matters is the cost of action.
3. ‘Don’t forget the priors’ .Prior + test/likelihood -> posterior probability

Accept

Multiple Comparisons Problem (MCP) in Visual Analytics

Review

Harshal Gajjar

8803-MDS Fall 22

Prof. Kexin Rong

Summary of contribution

- **Noticing the issue of multiple comparisons problem (MCP)** in visual analytics
- Experiment to **find how widespread the problem can be** (i.e. impacting what fraction of insights derived from visualizations)
- **Insight classes** to convert from words to testable hypothesis (perhaps the future of symbolic language)
- **Comparing confirmatory approaches** (confirmation with already seen data, confirmation with unseen data, mixing exploration and confirmation)

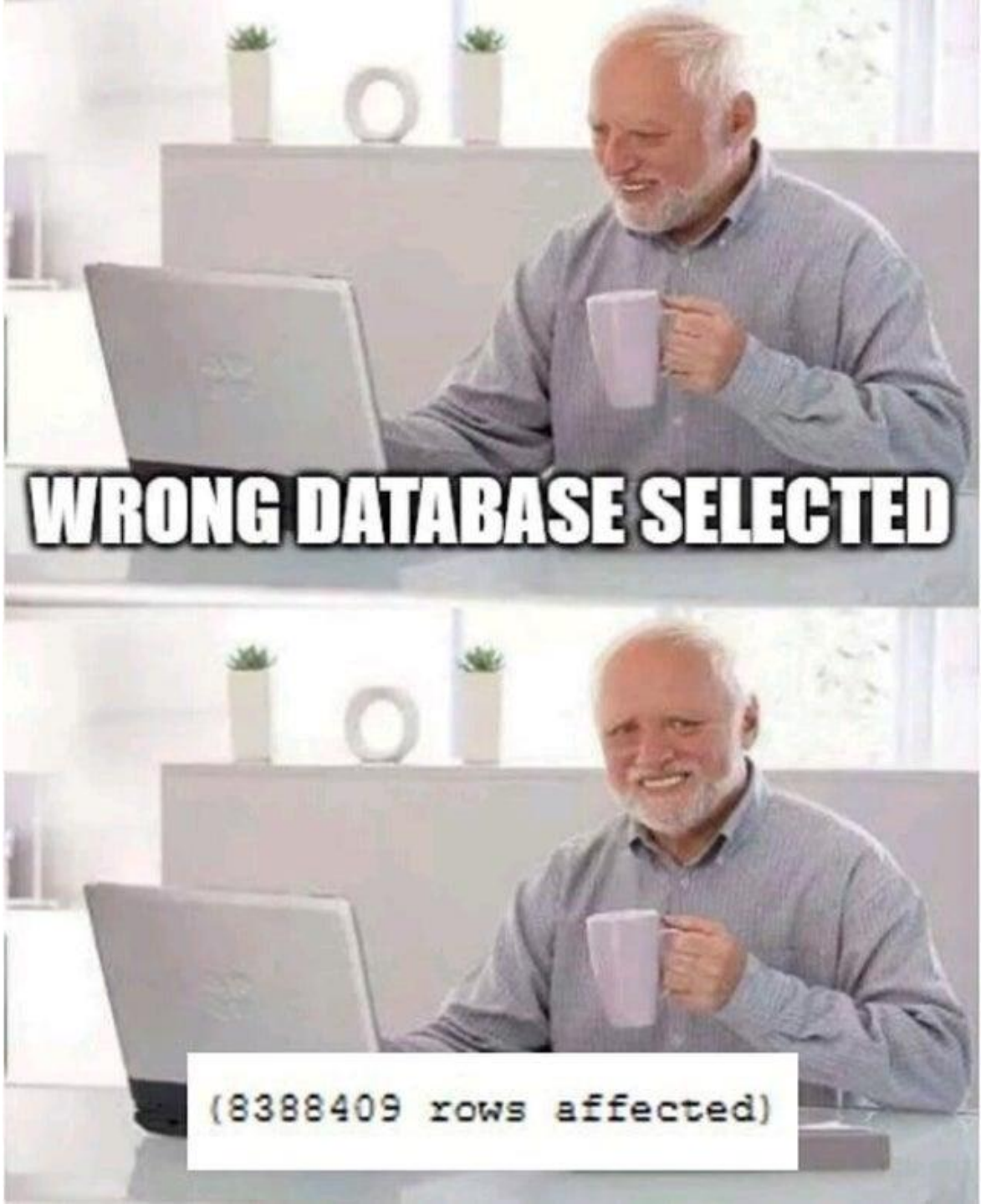
Strong Points

- Design and development of insight classes; and a transformation to null hypothesis from the class
- Thorough user study that (attempted to) capture all the implicit insights and remove ambiguity by tracking eyes and conversations
- Smartest use of synthetic data that I have seen so far (using data generation information in the test)
- Very surprising (and hence likely influential) results:
  - 60% of user insights are wrong, worse than tossing a coin
  - Neither background in stats or familiarity with hypothesis testing or experience with interpreting visualizations had a significant correlation with accuracy (at least for novice participants)

Improvements

- A concrete example explaining the process of “mixing exploration and confirmation” would have been helpful.
- Authors say nothing about the influence of type of visualization, scatter plots (or 2d histograms) might be better to compare multiple dimensions; does the accuracy of insights increase?
- I want to see how expert with a background in stats or familiarity with hypothesis testing or experience with interpreting visualizations impact the accuracy; and the methods they use to mitigate MCP. Test subjects can be diversified.

Reject



Investigating the Effect of the Multiple Comparisons Problem in Visual Analysis

Practitioner: Cangdi Li

Nov 9th

Recall

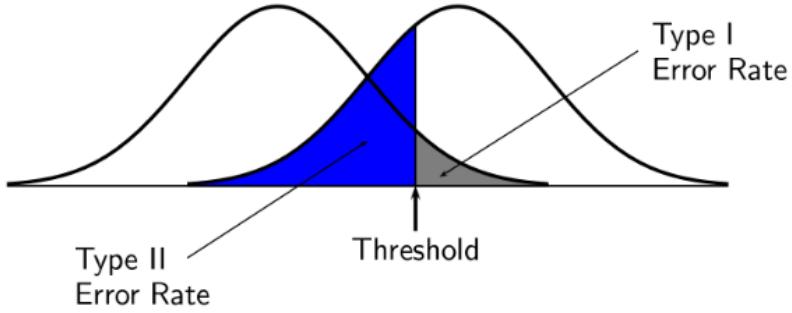
- **multiple comparisons problem:** While the chance of noise affecting one result may be small, the more measurements we make, the larger the probability that a **random** fluctuation is mis-classified as a meaningful result.
- **MCP in Visualization:** We often compare visualizations to a **mental image** of what we are interested in, but as more visualizations are examined and more comparisons are made, the probability of **discovering spurious insights increases**.
- **This paper** shows that a confirmatory approach of mixing exploration and confirmation can achieve similar results to using a separate validation dataset.

Type I & II error

- **Type I error:** As more visual comparisons are made, the probability of encountering false discovery rises.
- **Type II error:** An analyst might ignore a real pattern because it looks uninteresting, this is false omission.
- Setting a lower significance level decreases a Type I error risk, but increases a Type II error risk.
- Increasing the power of a test decreases a Type II error risk, but increases a Type I error risk.

One Test, One Threshold

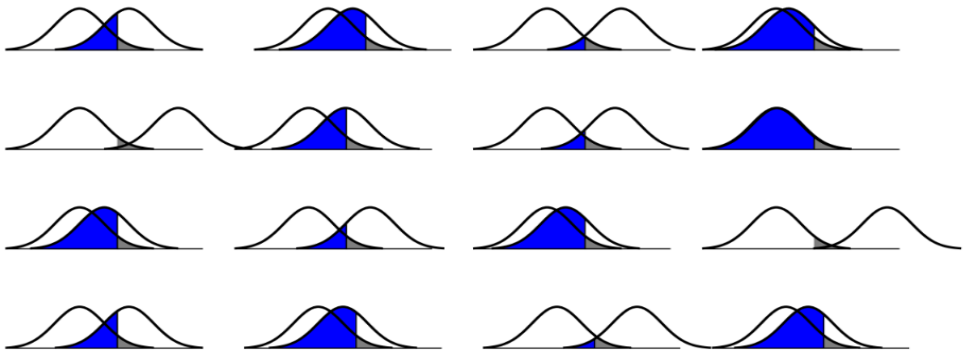
With a single hypothesis test, we choose a rejection threshold to control the Type I error rate,



while achieving a desirable Type II error rate for relevant alternatives.

Many Tests, One Threshold

With multiple tests, the problem is more complicated



Scope

- Data Analyst, Data Scientist and Business Intelligent Developer who need to deal with datasets and create reports with visualization to help with business decision making.
- Choosing the right tradeoff between type I error and type II error might be vary from team to team. In drug trials, false discoveries must be avoided, whereas, in security related scenarios, false omissions can have disastrous effects.
- For this study, we focus on improving the quality of overall visualization report for a generic engineering team.

- Pros:
  - We can ask Analysts to always including a validation process in their report to avoid some false findings in the visualizations, either by splitting the origin dataset, or by gathering some new data.
  - Analysts will need to save more intermediate comparisons to record implicit insights, and we might be able to make some of them useful.
  - The way we suggest most teams to balance the trade-off between type I and II errors is to have a report which divided into 2 parts, promising findings and implicit insights that might needs future data to check on.
  - This paper shows that participants who lacked domain knowledge could have large FDR, thus we will have another round for senior analyst with more experience help validate the report if it’ s done by junior analyst or marketing people.

- Cons:
  - It’ s time-consuming and expensive to gather new data.
  - It’ s lowers the power of comparison when splitting some of the current data just for validation purpose, especially when the data are sparse.
  - It’ s possible that we ends up dropping some important visualization findings because they are not promising enough.

Multiple Comparisons Problem in Visual Analysis: Archaeologist

Akshay Iyer

Georgia Tech

Paper Summary

- When more data is viewed/explored the probability of encountering interesting but insignificant results increases
- Idea of MCP: Likelihood of inferring a falsely significant result when multiple hypotheses are considered
- This idea is important for visualizations because the line between exploratory and confirmatory data analysis is often blurry in practice
- Authors ran an experiment where University students had to describe patterns they observed on shopping/sleep datasets
- 60% of user reported insights were wrong, but a fair number of students wanted to verify their findings
- Authors state that if users don't have a way to confirm their insights on a validation dataset the number of false discoveries can be very high
- Remedy that authors propose is hypothesis testing on the same dataset or multiple hypotheses control on the explicit insights

Previous Paper: The hot hand fallacy and the gambler's fallacy: Two faces of subjective randomness?

- MCP paper cites how people are known to be bad at judging randomness
- Two opposing ideas: hot hand fallacy vs gambler's fallacy
  - Hot hand fallacy—same outcome will continue.
  - Gambler's fallacy—Outcomes will be balanced out in the short term

- User study trials
  - "While subjects' predictions show negative recency with respect to the sequence of outcomes of the roulette wheel (the gambler's fallacy), their beliefs in the sequence of success and failure of their predictions show positive recency (the hot hand fallacy)" (5).

  - People are statistically more likely to link hot hand fallacy when human behavior is involved. They are also more likely to attribute streaking results to people instead of random events like a Roulette table

Georgia Tech

Previous Paper: The hot hand fallacy and the gambler's fallacy: Two faces of subjective randomness?

Big misconception: small chunks of random sequences are representative of long-term mathematical/statistical results

Authors speculate on why people struggle with understanding randomness. They list:

- Experience with phenomena like weather
- False equivalency of randomness with ease of memorization
- Ideas about luck

Overall, human tendency is such that both opposing viewpoints are in the mind when evaluating probabilistic decisions

Future Paper: Ethical Dimensions of Visualization Research

- The paper focuses on ethical aspects of visualization, argues that all visualization has some underlying moral characteristics
- Data is not neutral
  - Collection of data especially about people often has political implication
  - E.g. IBM subsidiary developing computing machines that expedited the Final Solution
  - On the other hand, not collecting data has consequences(as seen with facial recognition)
  - All data collected is from a biased perspective, through the eyes of an individual/group
- Visualization is also not neutral
  - Visualizations can present themselves as matter-of-fact and leave out opposing viewpoints
  - They can also conceal various types suffering, by shifting focus on numbers

Design Dilemmas in “Ethical Dimensions of Visualization Research”

Automated Analysis

- Many analytics systems can display conclusions are not fully supported by the data. (Multiple Comparisons Paper cited here)
- Access to analytics vs making accurate data-based claims

Machine Learning

- “We are therefore empowering the creators of ML models, but are not empowering the people affected by these models” (6).
- Simple models are often more explainable but less accurate. Similar idea holds for data exploration tools

To address visualization concerns, the paper gives 3 broad guidelines.

1.) Make the Invisible Visible (Address limitations and focus on marginalized groups/activities)

2.) Collect Data with Empathy (Limit data collected to protect individual’s privacy)

3.) Challenge Structures of Power (Push back against large institutions and challenge their practices. Also, point out unethical practices and denounce misinformation)

Thank you!

Georgia Tech

Next class

Vega-lite: A grammar of interactive graphics

Authors: Yanhao, Yiheng

Reviewer: Qiandong

Archaeologist: Haotian

Practioner: Aniruddha