The Effects of Interaction on Bayesian Reasoning

Saahil Claypool, Alex Shoop, Claire Danaher

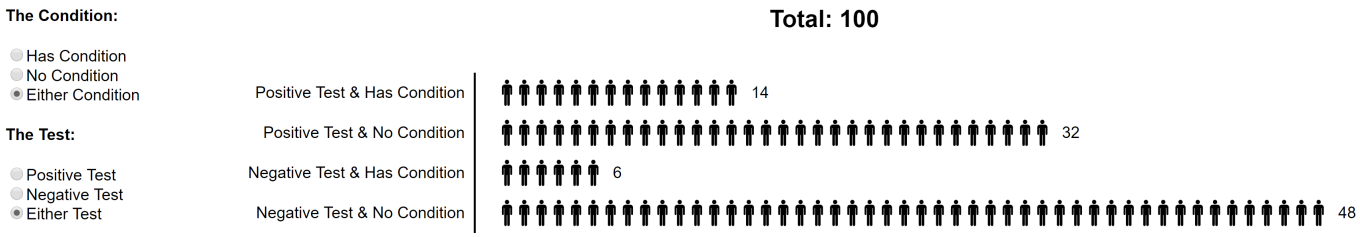

Fig. 1: An Interactive Visualization for Bayesian Statistics

Abstract— Your abstract is great.

Index Terms—Perception, Visualization, Evaluation.

1 Introduction

Advances in technology in the past decade have resulted in an exponential growth in the amount of data being collected on a variety of topics. This has lead to a movement towards using data driven decision making in increasingly diverse arenas. Data is being used to make decisions ranging from the potentially mundane, such as which individuals are qualified for an auto loan, to the life altering, such as medical test results.

Past research has shown that even highly trained individuals are likely to misinterpret statistical results. Consider the following standardized statement about mammography screening results [9]:

> The probability of breast cancer is 1in routine screening. If a woman has breast cancer, the probability is 80that she will get a positive mammography. If a woman does not have breast cancer, the probability is 9.6mammography. A woman in this age group had a positive mammography in a routine screening. What is the probability that she actually has breast cancer?

Despite the seemingly routine nature of this question, research has shown that many medical professional misinterpret these results leading to potentially serious consequences such as over diagnosis [6, 11]. While medical professionals are used in the previous scenario, research has shown that most individuals lack a robust understanding of conditional probabilities which impacts their ability to interpret results.

In the past, researchers have proposed ways to facilitate reasoning by changing the presentation of the results. In research completed by Hoffrage and Giegrenzer [8], the authors studied the comprehension benefits of using frequency formats. They postulate that, when provided frequency data rather than probabilities, medical professionals are better able to interpret conditional probabilities. Further research by Harrison et. al. supports that the presentation of information can have a significant positive effect on the ability for people to reason about conditional probabilities [9]. Even with these improvements, people still make many mistakes while reasoning about these problems.

	Our Results		Their Results	
Measures	**Exp**	**Control**	**Control**	**Exp**
measure A	54.8	48.6	44	33
measure B	32.2	28.8	8	6
measure C	22.8	19.8	35	26

Table 1: As you can see, we are the best.

Similar to previous work, the aim of our research is to further explore techniques for improving the presentation of Bayesian information to improve reasoning. We postulate that, as suggested by previous work, an interactive design will help problem solvers reason about conditional probabilities by providing a 'guide' for reasoning about the problem [10]. Specifically, we

propose that an interactive frequency chart, where users can select or view different conditions will reduce the information presented to a user and thus help them complete a provided Bayesian reasoning problem.

To test this hypothesis, we conducted an experiment asking mechanical turk participants to answer a Bayesian inference problem, similar to the one stated above. We tested three different presentations of the conditional probabilities: plain text, a static visualization, and an interactive visualization. From this experiment, we were able to draw the following conclusions:

- As prior research indicates, people perform poorly on this seemingly simple Bayesian reasoning task. This is even true for users that claim a 'high' statistical familiarity.
- Interaction did not seem to have a positive effective on reasoning. In fact, there is some indication it caused users to perform worse on their reasoning task. We hypothesize this could be because it made the information more confusing or distracted the user.

2 Background

Over the past three decades, research focused on understanding the cause of inaccuracy for both low and high numeracy individuals' ability to interpret Bayesian Statistics. Lines of research include studies to

- *Saahil Claypool is a student at Worcester Polytechnic Institute. Email: [smclaypool]@cs.wpi.edu.*
- *Alex*
- *Claire*

Manuscript received 31 March 2014; accepted 1 August 2014; posted online 13 October 2014; mailed on 4 October 2014.
For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org.

Fig. 2: Please see §6.

investigate the cognitive challenges associated with accurately interpreting these types of statistics as well as whether or not visualization can help overcome these challenges.

Research conducted in the 1990s focused around the hypothesis that inclusion of diagrams would help to improve users mental models whereby improving accuracy. Research completed by Cole in 1989, suggested that the inclusions of visualizations could improve interpretation [4]. The study showed that tendency towards over reliance on sensitivity (whereby influencing the conclusions drawn) persisted despite verbal intervention and training for participants tasked with interpreting written probabilistic outcomes. However, this tendency decreased with the inclusion of a frequentist visual representation. Cole attributes the difficulties of interpretation without graphics to the lack of a mental model to support understanding.

A study completed by Gigerenzer and Hoffrage [8] illustrated the underlying cognitive logic used by individuals to interpret Bayesian statistics. This study found that calculations, when provided asked a problem about Bayesian reasoning, people were better able to interpret frequencies rather than probabilities. The researchers hypothesized that this was a result of reduced cognitive load. The study showed that users were approximately 30% more likely to use Bayesian logic when presented with frequency as opposed to probability information. Furthermore, the study showed that with the use of frequency formats, the use of cognitive algorithms became more consistent over time. Whereas there was little or no improvement over time with the use of probability formats.

These results by Gigerenzer and Hoffrage [8] align with the previously completed work by Cole [4]; participants were found to perform better when they were provided with frequency formats over probabilities. While Cole did not explicitly test frequencies against probabilities hypothesis, numerical frequencies were presented visually for 3 out of the 4 diagrams used in Cole's study. Participants showed improvements with training with the use of visuals(frequencies) but found no improvement with the use of standard probabilities. This improvement provided by frequency formats has been observed and confirmed in other studies in areas such as public health, psychology, and data visualization [5, 7, 3, 2, 1].

A number of studies have test the hypothesis that the inclusion of visualization improves participant accuracy [2, 6, 3]. The results of this work have been inconclusive. In a paper by Ottley et al. [9], the researchers parse through discrepancies of past research by testing finely defined hypotheses pertaining to both linguistic and visual representation of the information. For a detailed discussion of discrepancies in past work please reference Ottley et al [9]. Ottley's findings supported the assertion that using frequencies as opposed to probabilities improved accuracy. However, participant accuracy rates using text only or visualization only outperformed those using text plus visualization.

The focus of this new research is to determine whether the inclusion of interaction can improve upon past accuracy rates. Evaluating the use of interaction is a burgeoning topic in the data visualization community. Bahador et al completed a paper in early 2018 focused on explore the effects on accuracy of different interactive visual encodings. While this work is somewhat tangential, the authors found that the use of position showed the highest level of accuracy. Research completed by Tsai et al [10] failed to produce statistically significant results however findings showed improvement by users who used an interactive tool when compared to frequency and probability only text.

3 METHOD

The methodology used for this research builds upon the methodology used by Ottley et al.[9]. The test completed as part of this work linguistic comparison of text containing probabilities versus frequencies. The findings were consistent with past studies and therefore this research does not delve further on this topic. The second test completed focused on the comparison text only, visualization only(vis only) and visualization and text(vis+text). The results of this study showed that improvement was not achieved through the inclusion of vis+ text when compared to vis only or text only. Similar results have been found in other studies. This study builds upon this past research by testing whether improvements can be achieved through the use of interaction.

We recruited 80 participants using Amazon's Mechanical Turk (MTurk) service and supplemented these participants with our colleagues. Participants were directed to our survey via an external link and results were recorded on a server set up for this experiment. Each participant was given a unique verification code upon the completion of their task.

Tree test cases were used in this study: text-only, vis-only, and interactive vis. The background text for all cases is as follows:

> There is a newly discovered disease, Disease X, which is transmitted by a bacterial infection found in the population. There is a test to detect whether or not a person has the disease, but it is not perfect. Here is some information about the current research on Disease X and efforts to test for the infection that causes it.

The user is then presented with one of the three test cases as described in greater detail below. Users were selected at random as to which test case they were assigned. The numerical specifics that a user was presented with were randomly selected from a set of pre-selected test cases. All participants are then asked to answer the following questions:

'100 people are tested for Disease X':

1. How many people do you think will test positive?

2. Out of those people, how many will have the disease?

3.1 Design 1: Text-Vis

As previously mentioned, the baseline case was designed to match the language used by Ottley et al. [9]. A sample of the textual statement is as follows:

> There is a total of 100 in the population. Out of the 100 people in the population,20 people actually have the disease. Out of these 20 people, 14 will receive a positive test result and 6 will receive a negative test result. On the other hand, 80 people do not have the disease (i.e., they are perfectly healthy). Out of these 80 people, 8 will receive a positive test result and 72 will receive a negative test result.

This description is a combination of the main components of a good text representation, those being framing, narrative, and probing.

3.2 Design 2: Vis-Only

For the vis-only test case, pictorial representation (human-like figures) were used to show the frequency count of tested/conditioned members of the population. Figures were shown along a horizontal axis, grouped along the vertical axis by the specific attribute: 'positive test and has condition', 'positive test and no condition', 'negative test and has condition', and 'negative test and no condition.' An example arrangement can be seen in Figure X.

TODO figures

3.3 Design 3: Interactive-Vis

We implemented interactivity using the same pictorial representation as the vis-only representation. The interactive vis tool includes selectable buttons on the left-hand-side of the chart. Participants can use the selectable buttons to toggle through different scenarios. For example if a user toggles on the 'Has Condition' button, the pictorial representation changes to only show the people icons associated with 'positive test and has condition' and 'negative test and has condition.' Transitions and animation were included for inverse selection. For example, if a user toggled of the 'No Condition' button, icons were removed from the screen. The figures below show a static view of interactive states. Figure X shows the starting screen for the interactive visualization tool. Figure Y is an example of toggling the button to specify the selection to answer the second lab question.

4 Results

TODO fix figures

A total number of 104 individuals participated in the experiment. The distribution of participants by group are 30, 38 and 36 completes for text-only, vis-only and interactive responses respectively. Demographic information was collected as part of the experiment. A summary of demographic information of our experiment participants can be found in Table X.

For the purposes of this analysis, responses were characterized using a binary approach of either correct or incorrect. So for our overall experiment, only 50 of the 104 participants were correct (48% accuracy). The overall accuracy across each visualization type was also observed.The logarithmic (base 2) error score was calculated for both of the two experiment questions, as seen in Figure X.

We also observed how the results looked when separated by the educational background and statistical understanding for each of the participants. The results of which can be found in Figure X and Figure X.

A chi-squared test was conducted across all participants and discovered a not-so-significant value when trying to see if there were significant differences between the accuracy rates and visualization type (χ^2(df = 2, N = 100)=2.1991, p = 0.333). Furthermore, performing pairwise chi-square tests on each pairing also did not show any significance; between vis-only and text, (χ^2(df = 2, N = 68)=0.26845, p = 0.8744), between interactive and text (χ^2(df = 2, N = 66)=1.9683, p = 0.3738), and between interactive and vis-only (χ^2(df = 2, N = 74)=2.9178, p = 0.2525.

5 Discussion

The purpose of this experiment was to test whether participants would more accurately interpret statistics when the results were presented as an interactive visualization tool. The baseline method for this experiment was to ask participants to interpret questions phrased as frequencies. This baseline was adapted from the text used by Ottley. We did not observe statistically significant results for the differences between the text-only, vis-only and interactive vis test cases. Therefore,we cannot conclude that interaction had any positive effect on the accuracy of the participants Bayesian reasoning. In fact, the group that received the interactive visualization performed slightly worse (although not a statistically significant amount). This may indicate that the interaction added an additional layer of complexity and hindered the user's performance. These trends were echoed verbally by users included in an initial pilot study. But, a follow up study with more users would be needed to confirm this.

The results produced by this study align with findings by Ottley where the vis-only and text-only results produced statistically similar accuracy rates. The accuracy rates of approximately 50% for the vis-only case aligned are slightly higher rate of 40% as observed by Metcalfe Et Al. The 40% accuracy rate for the text-only case are lower than those observe by Ottley. However, the sample size for this group was slightly smaller and the difference was not observed to be statistically significant. Therefore, the difference is likely attributable to aberrations due to the smaller sample size.

Interestingly, accuracy rates for those who self identified as having low numeracy skills increased between the text-only, vis-only and interactive vis respectively. This same trend was not observed for medium and high numeracy participants. An area for potential future research could be further exploring the relationship between numerical literacy and accuracy rates between cases.

6 Conclusion

As statistics become increasingly pervasive as a tool in decision making for people of all numeracy skill levels, the need for effectively expressing this information grows. The focus of this research was to determine whether the inclusion of interaction could be used to improve interpretation accuracy rates. The experiment did not produced statistically significant difference between text-only, vis-only and interactive-vis test cases. However, our findings aligned with past studies and can serve as a foundation for future work on the effects of interaction and statistical reasoning.

Acknowledgments

Text.

References

[1] G. L. Brase. Pictorial representations in statistical reasoning. *Applied Cognitive Psychology*, 2009.

[2] E. T. Brown, A. Ottley, H. Zhao, Q. Lin, R. Souvenir, A. Endert, and R. Chang. Finding waldo: Learning about users from their interactions. *IEEE Transactions on Visualization and Computer Graphics*, 2014.

[3] C. A. COHEN and M. HEGARTY. Individual Differences in Use of External Visualisations to Perform an Internal Visualisation Task. *Applied Cognitive Psychology*, 2007.

[4] W. G. Cole. Understanding bayesian reasoning via graphical displays. *SIGCHI Bull.*, 20(SI):381–386, Mar. 1989.

[5] D. M. Eddy. Probabilistic reasoning in clinical medicine: Problems and opportunities. In *Judgment under uncertainty: Heuristics and biases*. 1982.

[6] H. Friederichs, S. Ligges, and A. Weissenstein. Using tree diagrams without numerical values in addition to relative numbers improves students' numeracy skills: A randomized study in medical education. *Medical Decision Making*, 2014.

[7] M. Galesic, R. Garcia-Retamero, and G. Gigerenzer. Using Icon Arrays to Communicate Medical Risks: Overcoming Low Numeracy. *Health Psychology*, 2009.

[8] G. Gigerenzer and U. Hoffrage. How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review*, 1995.

[9] A. Ottley, E. M. Peck, L. T. Harrison, D. Afergan, C. Ziemkiewicz, H. A. Taylor, P. K. Han, and R. Chang. Improving Bayesian Reasoning: The Effects of Phrasing, Visualization, and Spatial Ability. *IEEE Transactions on Visualization and Computer Graphics*, 2016.

[10] J. Tsai, S. Miller, and A. Kirlik. Interactive visualizations to improve Bayesian reasoning. In *Proceedings of the Human Factors and Ergonomics Society*, 2011.

[11] H. G. Welch and W. C. Black. Overdiagnosis in cancer. *JNCI: Journal of the National Cancer Institute*, 102(9):605–613, 2010.