Visually Communicating Bayesian Statistics to Laypersons

Alvitta Ottley
Tufts University
alvittao@cs.tufts.edu

Blossom Metevier
University of Maryland Baltimore County
bloss1@umbc.edu

Paul K. J. Han
Maine Medical Center
hanp@mmc.org

Remco Chang
Tufts University
remco@cs.tufts.edu

December 12, 2012

Abstract

Effectively communicating Bayesian statistics to laypersons has been an open challenge for many years. Recent research in psychology proposed that there is a direct correlation between comprehension and representation. Specifically, a series of studies suggests that pictorial representations with icon arrays may be better suited for communicating Bayesian statistics than Euler diagrams. Though these results are compelling, the experiments were conducted in controlled lab settings and with limited samples. In this paper, we extend the previous research by expanding the sample to a more diverse population through crowdsourcing. We conducted a user study that compares three different pictorial representations of Bayesian statistics icon arrays, Euler diagrams and discretized Euler diagrams. Our findings fail to replicate previous results and demonstrate no significant difference between the three representations. We discuss possible explanations for these findings and propose directions for future investigations.

1 Introduction

As the medical field transitions towards evidence-based and shared decision making, effectively communicating risks to patients has emerged as common challenge. In recent years, researchers and practitioners have worked at developing novel techniques for risk communication, however conveying these risks to laypersons remains a difficult task. This is largely due to the deficits of numeracy among the general population and as a result, research suggests that current risk communication methods often lead to patients' confusion and sometimes "ambiguity aversion" – the avoidance of decision making due to uncertainty regarding the reliability, accuracy or credibility of information about risk and the potential consequences of decisions [7, 10, 11].

At the heart of many of these decisions is Bayesian inference. Deciding between two tests based on their reported specificity and sensitivity, or deciding whether to take a potentially harmful drug after receiving a positive test requires some understanding of conditional probabilities. For example, the classic mammography problem below can be reduced to a Bayesian inference problem:

The probability of a woman over 40 having cancer is 1%. However, 80% of women who are tested will receive a positive result on their mammography. Additionally, the probability of a false positive is 9.6%.
What is the probability of getting a positive result given that you actually have the disease?

Studies show that less than 20% of typical laboratory participants were able to correctly calculate the accuracy of the mammography [3, 8, 9] and even more alarming, only 5 out of 100 trained physicians [6]. In an effort to improve these statistics, Gigerenzer and Hoffrage proposed the use of natural frequencies instead of probabilities and found that this significantly improved performance [9]. Follow up studies also investigated the use of visualizations to facilitate comprehension and reported significant effects [1, 16].

These studies suggest that there is a direct correlation between comprehension and representation. Brase [1] investigated this notion by conducting a comparative study and showed that pictorial representations with icon arrays may be better suited for communicating Bayesian statistics than Euler diagrams or text alone. Though these results are compelling, the experiments were conducted in controlled lab settings and with limited samples.

We propose that the results of traditional laboratory studies may be not representative of the general population.

In this paper, we extend the previous research by expanding the sample to a more diverse population through crowdsourcing. We conducted a user study using Amazon's Mechanical Turk and replicated the experiment design by Brase [1] which compared three different pictorial representations of Bayesian statistics: icon arrays, Euler diagrams and discretized Euler diagrams. Our findings fail to replicate previous results and demonstrate no significant difference between the three representations. In the final section, we discuss possible explanations for these findings and propose directions for future investigations.

2 Related Work

There is a substantial body of work that has been aimed at developing novel, more effective methods of communicating Bayesian statistics [2, 5, 4, 9]. The current work is aimed at understanding which visualizations are better suited for communicating Bayesian statistics to layperson. To our knowledge, there have only been two studies with this agenda [1, 14]. We summarize these below.

2.1 Visualizing Bayes Reasoning

Brase [1] first attempted to identify visualizations that facilitate Bayesian reasoning. In a comparative study, he compared participants' accuracy using three different visualizations: icon arrays, Euler diagrams and discretized Euler diagrams. Discrete items represented by the icon array were expected to elicit a frequentist representation, thus improving Bayesian reasoning while Euler diagrams were expected to enhance the perception of the nested-set relations which is inherent in Bayesian inference problems. The discretized Euler diagram was design to represent a hybrid of the two.

A study was conducted with 412 participants who were recruited from a pool of university undergraduates. There were four groups of participants: (1) the control -participants were presented with only a textual representation of the problem, (2) textual representation was accompanied by an icon array, (3)textual representation was accompanied by an Euler diagram and (4)textual representation was accompanied by a discretized Euler diagram.

Each participant was presented with a Bayesian inference problem and were ask the calculate the overall chance of having the disease and the hit rate. They found that participants who were given an icon array had the best accuracy rate overall and their performance was significantly better than the control group and those who performed the task with the Euler diagram. Table 1 below summarizes his results.

Type of Visualization	Overall Accuracy(%)
Control	35.4
Euler diagram	34.7
Discretized Euler diagram	41.7
Icon Array	48.4

Table 1: The summarized results of Brase [1]

These results suggest that visualizations with discrete items may be best suited for facilitating Bayesian reasoning. However there are several flaws in the experiment design:

- the Euler diagram was not area-proportional
- the number of discrete items in the discretized Euler diagrams did not match the problem posed
- only 2 of the 3 diagrams were labeled (the icon array had no labels)
- the labels were inconsistent with the text used for the problem
- the glyphs used for the icon array were inconsistent with the ones used for the discretized Euler diagram (dots vs anthropomorphic)

In addition to these, the results are not easily generalizable as the study sample does not represent the general population of laypersons. The current work extends this study by addressing these issues and expanding the subject pool to a more diverse population through crowdsourcing. Recent research performed concurrently with the current work [14] also had a similar agenda, we discussing their and our findings in the Section 6.

3 Hypothesis

Research suggests that discretized visualizations may be best for facilitating Bayesian reasoning. While these results are intriguing, there were several flaws in the experiment design and the limited sample incites questions of generalizability. In this paper, we extend the previous study by improving the experiment design and expanding the subject pool to a more diverse population. We hypothesize that the overall accuracy will be lower and but we anticipate the same performance trend.

4 Experiment

To test our hypothesis, we replicated the user study performed by Brase [1] in an online environment. Participants were asked to solve the same Bayesian inference problem and were given the same visualizations (minor adjustments were made to address the aforementioned concerns).

4.1 Participants

We recruited 194 participants over Amazon's Mechanical Turk service. Mechanical Turk is a service offered by Amazon that allows computers to harness human skills for jobs that require human intelligence. This site is essentially a virtual market place that allows individuals or companies to post small jobs for completion and where persons can easily "work" for a small remuneration. There are over 100,000 workers from around the world registered with this service, who are able to choose jobs from a pool and are paid small amounts upon completion [15].

Mechanical Turk is becoming increasingly attractive to Human-Computer Interaction and Visualization researchers as it facilitates the recruitment of a more diverse study population [12, 17]. One of the biggest concerns when using this tool is participants randomly clicking through the study only to be paid. Researchers have compensated for this by providing bonuses for correct responses and participants are paid only upon completing the study [13]. This has been proven effective and has been adapted for our study.

Of the 193 participants, only 1 did not report their age. Of the rest, there were 123 males and 69 females. Their self-reported age ranged from 18 to 66, with a mean of $32.31(\sigma = 10.61)$. Additionally, majority reported to have at least a bachelor's degree (Table 2).

Highest Level of Education	Frequency
High School	30
Associates	15
Bachelors	105
Masters	40
PhD	3
Other	1

Table 2: Participants' self-reported education level

4.2 Materials

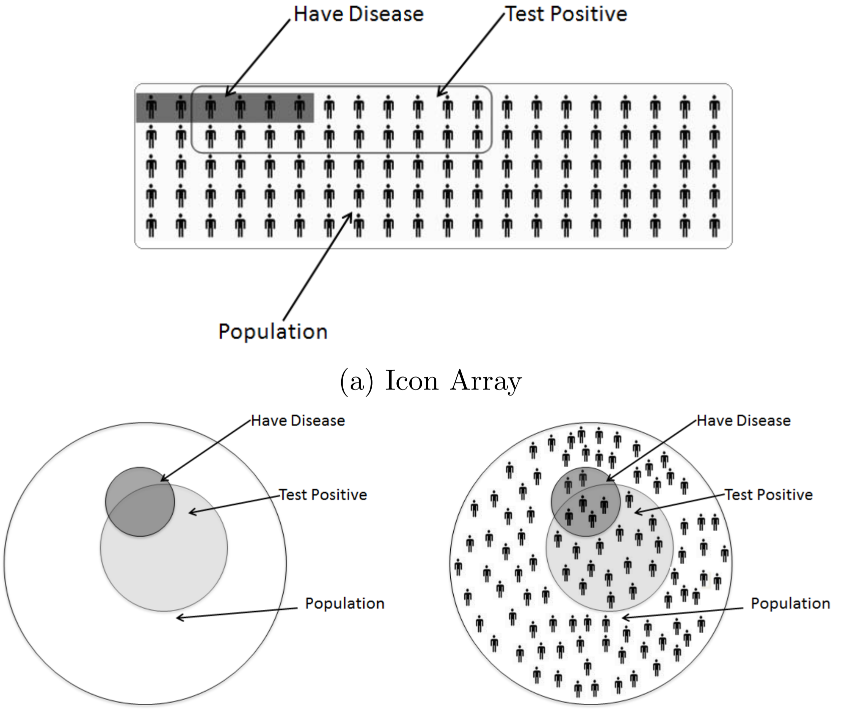
Each participant was presented with the following information:

There is a newly discovered disease, Disease X, which is transmitted by a bacterial infection. Here is some information about the current research on Disease X and efforts to test for the infection that causes it.

A person has 6 chances out of 100 of having the infection. There is a test to detect whether or not a person has this infection, but it is not perfect. Specifically, only 4 of the 6 chances of having the infection were associated with a positive reaction from the test. On the other hand, 16 of the remaining 94 chances of not having the infection (that is, being perfectly healthy) were also associated with a positive reaction from the test.

This was then following by one of 4 visualizations: (1) the control - no visual representation, (2) an icon array, (3) a Euler diagram and (4) a discretized Euler diagram (Figure 1). Participants were then asked to solve the following problem:

Imagine Michael is tested now. Out of a total of 100 chances, Michael has _____ chance(s) of positive reaction from the test, _____ of which will be associated with actually having the infection.



(a) Icon Array

(b) Euler Diagram

(c) Discretized Euler Diagram

Figure 1: The three visualizations used in the study

4.3 Procedure

After selecting the task from the Mechanical Turk website and informed consent, participants were presented with the experiment instructions. Once they selected to continue, they were presented the Bayesian inference problem described above and one of the four visual conditions. To separate the time spent reading the question from the time spent actually solving the problem, the question was not visible until they clicked the “Next” button. They were instructed to take as much time as needed and when they were ready to answer, they clicked the “Ready to Answer” button and keyed in their responses in the text fields provided.

Once they were done, they given a short demographic questionnaire. We measured their time spent performing the task.

5 Results

Our study yielded an overall accuracy of 30% which is slightly lower than the accuracy reported by Brase [1]. However, an analysis across all participants revealed no significant difference between the participants' accuracy for the four conditions (F(3,190) = .172, p = .915). Table 3 below summarizes our results.

Type of Visualization	Overall Accuracy(%)
Control	33
Euler diagram	28
Discretized Euler diagram	27
Icon Array	32

Table 3: Percentage of participants who accurately calculated the exact answers for both parts of the problem.

6 Discussion and Future Work

Our results showed no significant difference between the four conditions. Although, this does not reflect the results reported by Brase [1] it is consistence with similar work which also compared visualizations of Bayesian inference problems through crowdsourcing [14]. Our results suggest that simply adding a visualization yields no measurable benefits. We hypothesize that participants may have neglected the visualizations in an attempt to optimize their time. A similar study by Micallef et al. [14] supports this notion. They found that by simply removing the numbers in the text, there was a significant improvement in participants' accuracy as they were now forced to utilize the visualization. Taken together, these results are quite telling for future studies on crowdsourcing platforms. One solution may be to factor in incentivization for using the visualizations.

Though there are indeed some important considerations when conducting studies in a crowdsourcing environment, once done right, it still remains a valuable avenue for research. Researchers may also be able to improve future results by adapting techniques such as storyboarding or highlighting to better enhance the link between the text and visualizations. By uniting the two, we hypothesize that the user will be less likely to ignore the visualizations and this can be an avenue for more effective comparisons.

Future work can also look at isolating visual elements that facilitate Bayesian reasoning. The results of Brase [1] suggests that visualizations with discretized items may be best suited for facilitating Bayesian reasoning. However, the types of glyphs and the number of glyphs used for the two visualizations differed significantly. We hypothesize that anthropomorphic glyphs encourage the user to adapting an egocentric view of the problem. This may be a better fit to the user's mental model and better aid understanding over simple dot glyphs.

7 Conclusion

In this paper, we used crowdsourcing to replicate previous studies comparing visualizations of Bayesian statistics. Our results were inconsistent with traditional laboratory experiments and highlighted the sensitivity of the crowd. We found no significant difference between the visualizations, and our result showed no significant improvement when visualizations are added. We hypothesize that this may be reducible to an incentivization problem and highlights differences between experiments conducted in traditional laboratory settings and those conducted using crowdsourcing platforms.

References

[1] G.L. Brase, *Pictorial representations in statistical reasoning*, Applied Cognitive Psychology **23** (2008), no. 3, 369–381.

[2] K. Burns, *Painting pictures to augment advice*, Proceedings of the working conference on Advanced visual interfaces, ACM, 2004, pp. 344–349.

[3] W. Casscells, A. Schoenberger, T.B. Graboys, et al., *Interpretation by physicians of clinical laboratory results.*, The New England Journal of Medicine **299** (1978), no. 18, 999.

[4] W.G. Cole, *Understanding bayesian reasoning via graphical displays*, ACM SIGCHI Bulletin, vol. 20, ACM, 1989, pp. 381–386.

[5] W.G. Cole and J.E. Davidson, *Graphic representation can lead to fast and accurate bayesian reasoning*, Symp Computer Application in Medical Care, 1989, pp. 227–231.

[6] D.M. Eddy, *Probabilistic reasoning in clinical medicine: Problems and opportunities*, Judgment under uncertainty: Heuristics and biases (1982), 249–267.

[7] D. Ellsberg, *Risk, ambiguity, and the savage axioms*, The Quarterly Journal of Economics (1961), 643–669.

[8] J.S.B.T. Evans, S.J. Handley, N. Perham, D.E. Over, V.A. Thompson, et al., *Frequency versus probability formats in statistical word problems*, Cognition **77** (2000), no. 3, 197–213.

[9] G. Gigerenzer and U. Hoffrage, *How to improve bayesian reasoning without instruction: Frequency formats.*, Psychological review **102** (1995), no. 4, 684.

[10] P.K.J. Han, W.M.P. Klein, T. Lehman, B. Killam, H. Massett, and A.N. Freedman, *Communication of uncertainty regarding individualized cancer risk estimates effects and influential factors*, Medical Decision Making **31** (2011), no. 2, 354–366.

[11] P.K.J. Han, B.B. Reeve, R.P. Moser, and W.M.P. Klein, *Aversion to ambiguity regarding medical tests and treatments: measurement, prevalence, and relationship to sociodemographic factors*, Journal of health communication **14** (2009), no. 6, 556–572.

[12] J. Heer and M. Bostock, *Crowdsourcing graphical perception: using mechanical turk to assess visualization design*, Proceedings of the 28th international conference on Human factors in computing systems, ACM, 2010, pp. 203–212.

[13] R. Kosara and C. Ziemkiewicz, *Do mechanical turks dream of square pie charts?*, Proceedings of the 3rd BELIV'10 Workshop: BEyond time and errors: novel evaLuation methods for Information Visualization, ACM, 2010, pp. 63–70.

[14] L. Micallef, P. Dragicevic, J.D. Fekete, et al., *Assessing the effect of visualizations on bayesian reasoning through crowdsourcing*, IEEE Transactions on Visualization and Computer Graphics (2012).

[15] J. Ross, L. Irani, M. Silberman, A. Zaldivar, and B. Tomlinson, *Who are the crowdworkers?: shifting demographics in mechanical turk*, Proceedings of the 28th of the international conference extended abstracts on Human factors in computing systems, ACM, 2010, pp. 2863–2872.

[16] P. Sedlmeier and G. Gigerenzer, *Teaching bayesian reasoning in less than two hours.*, Journal of Experimental Psychology: General **130** (2001), no. 3, 380.

[17] C. Ziemkiewicz and R. Kosara, *The shaping of information by visual metaphors*, Visualization and Computer Graphics, IEEE Transactions on **14** (2008), no. 6, 1269–1276.