Knowing What You Don't Know: Choosing the Right Chart to Show Data Distributions to Non-Expert Users

Arkaitz Zubiaga
Dublin Institute of Technology, Ireland
arkaitz@zubiaga.org

Brian Mac Namee
University College Dublin, Ireland
brian.macnamee@ucd.ie

ABSTRACT

An ability to understand the outputs of data analysis is a key characteristic of data literacy and the inclusion of data visualisations is ubiquitous in the output of modern data analysis. Several aspects still remain unresolved, however, on the question of choosing data visualisations that lead viewers to an optimal interpretation of data, especially when audiences have differing degrees of data literacy. In this paper we describe a user study on perception from data visualisations, in which we measured the ability of participants to validate statements about the distributions of data samples visualised using different chart types. We find that histograms are the most suitable chart type for illustrating the distribution of values for a variable. We contrast our findings with previous research in the field, and posit three main issues identified from the study. Most notably, however, we show that viewers struggle to identify scenarios in which a chart simply does not contain enough information to validate a statement about the data that it represents. The results of our study emphasise the importance of using an understanding of the limits of viewers' data literacy to design charts effectively, and we discuss factors that are crucial to this end.

Categories and Subject Descriptors

H.5.2 [**Information Interfaces and Presentation**]: User Interfaces User-centered design

1. INTRODUCTION

Although the definition of data literacy remains somewhat fluid [10, 2], most definitions include an ability to interpret the outputs from data analysis. For example, Harris [7] defines data literacy as *"competence in finding, manipulating, managing, and interpreting data, including not just numbers but also text and images"*, Beauchamp [1] defines it as *"the ability to interpret, evaluate, and communicate statistical information"*, and Schield [14] as the ability *"to access, assess, manipulate, summarize, and present data"*. Many of the outputs from data analysis referred to in these definitions take the form of data visualisations. In fact the importance of data visualisation is included in a number of discussions on the characteristics of data literacy [10, 17, 16].

Although data visualisations are ubiquitous in everyday publications [9], the level of literacy that the general public brings to different chart types is not always clear. The choice of an appropriate chart type for a particular dataset can, in fact, condition subsequent interpretation by viewers. Using the chart type that most effectively conveys insight is especially important when viewers have differing levels of data literacy.

In this work we evaluate the effectiveness of different chart types for visualising the distribution of a variable to enable average, non-expert users to accurately extract basic insight. To this end, we conduct a user study through Mechanical Turk[1], where participants validate the veracity of statements about the distributions of variables shown alongside different visualisations of these distributions. We measure performance on a set of 160 different tasks for different combinations of chart type, statement, and variable distribution for which we collected a total of 8,000 assessments.

We find that, among the five types of chart under study, histograms are not only the most complete in terms of details given, but also the chart type that leads viewers to the most accurate understanding of the underlying data. We also find, however, that viewers are not good at determining the limits of what can be understood about data from different chart types, i.e. *they don't know what they don't know.*

2. RELATED WORK

One of the best known studies on chart perception is by Cleveland and McGill [3], who define a theory to examine the elementary perceptual tasks that viewers perform when looking at charts, as well as the extent to which they lead viewers to accurate understanding. More recently both Shah and Hoeffner [15] and Glazer [6] summarise three major factors that influence viewers' interpretations of data visualisations: (i) the visual characteristics of a chart, (ii) a viewer's knowledge about charts, and (iii) a viewer's background and expectations of the content in the chart. The authors highlight, however, that no single chart type is necessarily better overall than any other, and new tasks might require careful studies to choose a suitable chart. In general, re-

[1]Mechanical Turk: `http://www.mturk.com/`

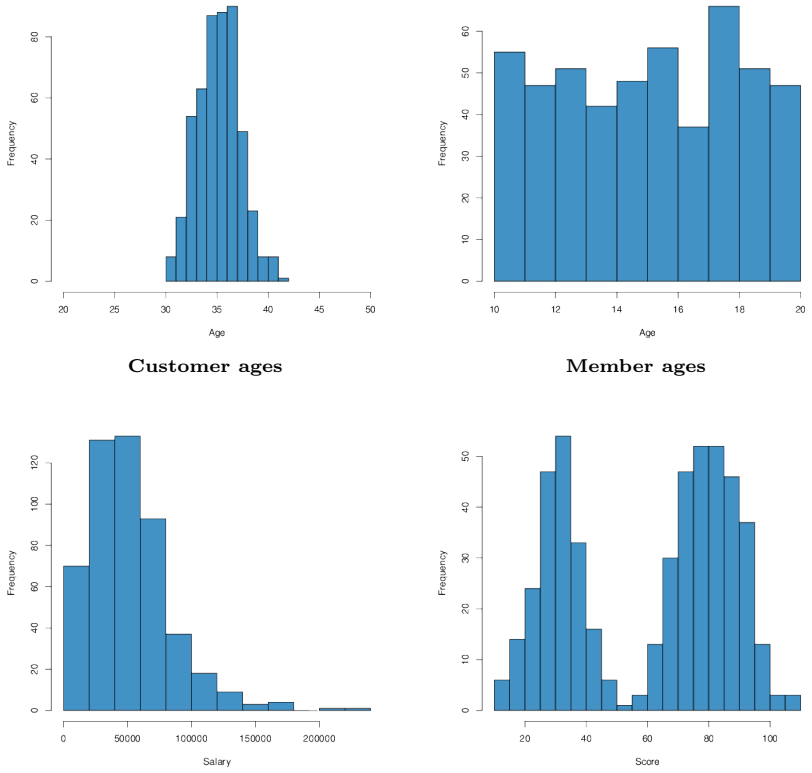

Figure 1: Histograms depicting the shapes of the data distributions under study.

searchers have pointed out that creating appropriate charts so that viewers perceive the intended message is harder than it might at first seem, and that detailed study of the effectiveness of different chart types for different tasks is required [5, 15]. Furthermore, the literature does not contain extensive studies of how well viewers can interpret charts showing the distribution of a variable.

3. EXPERIMENTAL METHOD

In the basic unit in our experimental method a chart is shown to a participant along with an associated textual statement about the distribution of the variable represented in that chart. Participants rate how well they think the statement corresponds to the data represented in the chart. This is repeated for different combinations of three different factors: (i) the underlying distribution of the variable, (ii) the type of chart used to show the data, and (iii) the type of statement made about the data. Overall, we include four different variables, five distinct chart types, and four different types of statements.

The four variables used were: (i) ages of customers of an online movie service, (ii) ages of members of a youth sports centre, (iii) salaries of a city's residents, and (iv) scores of students in an exam. Each variable exhibited a different distribution (see Figure 1). Five commonly used chart types were selected for this study: (i) bar charts showing the average value of the distribution, (ii) bee swarms, (iii) boxplots, (iv) stacked bar charts, and (v) histograms. Figure 2 shows an example of each.

The textual statements were of the following four types: (i) "*the data ranges from* $\boldsymbol{X}$ *to* $\boldsymbol{Y}$", (ii) "*most data points fall around* $\boldsymbol{X}$", (iii) "*most data points fall under/over* $\boldsymbol{X}$", and (iv) "*data points are clustered to either side of* $\boldsymbol{X}$". For each chart type and variable combination, two versions of each statement type were presented to participants: one that was true and one that was false. For example, for the data shown in the first histogram in Figure 1 the true statement "*The data ranges from 30 to 42*", and the false statement "*The data ranges from 25 to 45*" were used.

The combinations of variables (4), chart types (5), and statements (8) amounted to a total of 160 different tasks. Participants in our experiments were shown one task at a time and had to rate the accuracy of the statement shown on a five point Likert scale: *strongly agree*, *agree*, *neutral*, *disagree*, and *strongly disagree*. Additionally, participants could opt for an alternative choice *impossible to tell from this chart*. Tasks were presented in random order to control for learning effects. With 50 ratings collected for each of the 160 tasks, we gathered a total of 8,000 ratings.

We conducted our experiments through Mechanical Turk. The use of a crowdsourcing platform such as Mechanical Turk for this study is motivated by Heer and Bostock [8], who showed that it is an effective and reliable way in which

Statement	T	F	I
Data ranges from X to Y	12	12	16
Points fall around X	9	17	14
Points fall under/over X	11	13	16
Points clustered to either side of X	12	12	18
Chart Type	**T**	**F**	**I**
Bar chart (average)	0	0	32
Bee swarm	14	18	0
Boxplot	9	11	12
Stack chart	5	7	20
Histogram	14	18	0
Variable	**T**	**F**	**I**
Online movie customer ages	12	15	13
Youth sports centre ages	8	13	19
Salaries	16	16	8
Student scores	6	10	24

Table 1: Distribution of correct answers as defined in the ground truth (T: True, F: False, I: Impossible to tell).

Statements	
Data ranges from X to Y	0.416 (moderate)
Points fall around X	0.304 (fair)
Points fall under/over X	0.440 (moderate)
Points clustered to either side of X	0.360 (fair)
Charts	
Bar chart (average)	0.232 (fair)
Bee swarm	0.495 (moderate)
Boxplot	0.313 (fair)
Stack chart	0.211 (fair)
Histogram	0.479 (moderate)
Variables	
Online movie customer ages	0.442 (moderate)
Youth sports centre ages	0.413 (moderate)
Salaries	0.391 (fair)
Student scores	0.288 (fair)
Overall Inter-rater agreement	**0.390 (fair)**

Table 2: Inter-rater agreement values by item, and overall.

to perform graphical perception studies. To take part in the study participants did not need to have any prior expertise in data analytics as we were interested in measuring the ability of average, non-expert viewers to interpret different chart types. We did, however, restrict participation to US-based participants to control for English language capability, and to participants with at least a 95% *HIT acceptance rate*, which is Mechanical Turk's internal measure of how well participants perform tasks on the platform. A high HIT acceptance rate guarantees that participants have been deemed reliable in other experiments and filters *bots*.

The *ground truth* for each task was manually annotated by the experiment designers, with the following distribution of responses: 42 cases were true, 54 were false, and 64 were impossible to tell. Table 1 shows the distribution of the ground truth responses, broken down by variable, chart type and statement type. The most important differences in these distributions relates to the chart types. All of the tasks showing a simple average bar chart fall into the "*impossible to tell*" category as the average bar chart does not provide enough evidence to assess the associated statements. With bee swarm charts and histograms it is possible, in all cases, to assess each statement. With boxplots and stacked bar charts it is possible to assess only some statements.

4. RESULTS

We examine the data collected in these experiments in three ways: (1) **inter-rater agreement** to assess the level of agreement in the responses given by different participants; (2) **accuracy** to assess how well participant responses match the ground truth and (3) a **confusion matrix** to understand the types of errors made by participants.

Inter-rater Agreement. We measure inter-rater agreement using *Krippendorff's alpha coefficient* [11]. Overall, the 8,000 ratings show a *fair* level of inter-rater agreement of 0.39[2]. Table 2 shows inter-rater agreement values for each chart, statement, and variable. We see two major differences here. Firstly, with regard to statement type, participants tend to agree when assessing the ranges of variables and whether variable values are above or below a given threshold; and tend to disagree when asked about values being clustered around a certain value. Secondly, with regard to chart type, participants showed a larger degree of agreement for bee swarms and histograms; and a much lower degree of agreement for the other three chart types. This is likely due to the high number of answers that are *impossible to tell*.

Accuracy. To compute the accuracy values, we rely on majority voting, i.e., the rating that has been chosen by most participants. This allows us to choose a single rating from the 50 provided for each task. For the purposes of computing accuracy, we collapse ratings of *agree* and *strongly agree* to *true*, and ratings of *disagree* and *strongly disagree* to *false*[3]. The final accuracy values reported here refer to the number of cases in which the majority vote of participants coincides with the ground truth. An overall accuracy value, and values broken down by variable, chart type and statement, are shown in Table 3.

On statements of type "*data ranges from X to Y*" and "*points fall under/over X*" participants were substantially more accurate (90% and 75%, respectively) than for the other two types of statements (55% and 52.5%). More specifically we found that participants struggled with bar and stack charts when assessing "*points fall around X*" statements, and with stack charts when assessing "*points clustered to either side of X*" statements.

Regarding chart type, the most accurate answers were those for bee swarms and histograms (both above 90%). This is slightly surprising as these are relatively complex chart types for non-expert viewers. Even though both bee swarm charts and histograms potentially allow viewers to determine the

[2]We report the strength of agreement using the benchmarks suggested by Landis and Koch [12] for interpreting kappa.

[3]In fact, participants seemed reluctant to choose strong judgements, choosing *agree* and *disagree* much more than *strongly agree* and *strongly disagree*.

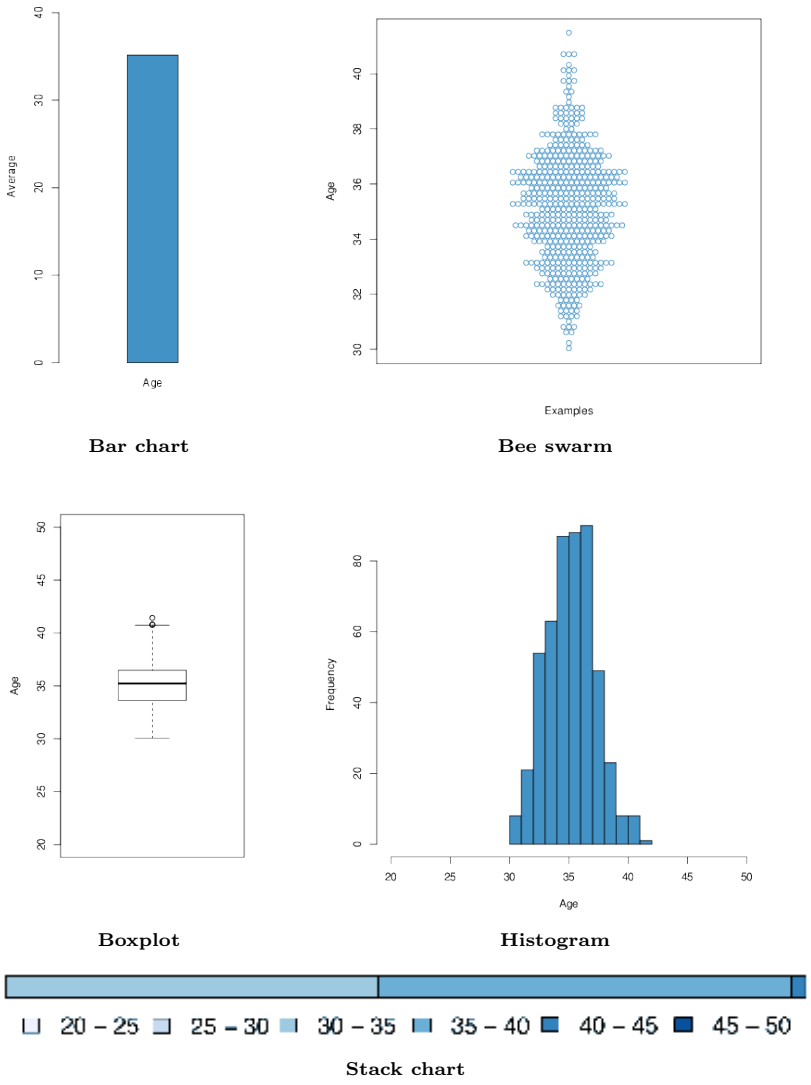

Figure 2: Examples of each chart type used for the user study.

veracity of all of the statements, and provide similar information, viewers seem to find it slightly easier to comprehend values from a histogram.

Finally, participants struggled slightly to answer questions about the student scores data. This data has a bimodal distribution which could be more difficult for viewers to parse.

Confusion Matrix. Table 4 shows a confusion matrix for all tasks (note that *Imp.* refers to responses of *impossible to tell*, that *Neutral* did not occur in the ground truth, and that cells marking correct responses are highlighted in bold). The precision for each category is also included. Most notably here, we observe that when the correct response was *impossible to tell*, participants mostly deemed statements false (45.9% of the time), or even true (24.9% of the time), and only identified 23.8% of the cases correctly. When the correct response was either true or false, participants again rarely chose *impossible to tell* as the answer.

Statements	
Data ranges from X to Y	0.900
Points fall around X	0.550
Points fall under/over X	0.750
Points clustered to either side of X	0.525
Charts	
Bar chart (average)	0.531
Bee swarm	0.906
Boxplot	0.563
Stack chart	0.438
Histogram	0.969
Variables	
Online movie customer ages	0.700
Youth sports centre ages	0.700
Salaries	0.750
Student scores	0.575
Overall accuracy	**0.681**

Table 3: Accuracy values by item, and overall.

		Responses			
		Imp.	**False**	**Neutral**	**True**
Ground Truth	**Imp.**	**23.8**	45.9	5.4	24.9
	False	6.9	**72.1**	6.4	14.6
	True	4.7	14.0	5.5	**75.8**
Precision		67.2	54.6	-	65.7

Table 4: Confusion matrix for all the tasks combined (in %).

Taken altogether we believe that these results indicate that, although participants do well when assessing true cases (accuracy 75.8%) and false cases (72.1%), they have trouble when facing charts that do not enable them to determine the veracity of a statement and do not recognise this shortcoming. We were surprised that participants did not use the *neutral* choice in these cases (the *neutral* response was only used in 6% of cases).

5. DISCUSSION

In this work, we have studied the effectiveness of different chart types in conveying information on the distributions of variables. We have seen that histograms allow the most accurate interpretations—viewers achieved 97% accuracy from histograms, compared to 91% with bee swarms, and lower than 60% for the other charts—and are an appropriate choice of chart type when visualising the distribution of a variable for an average, non-expert audience. This reinforces previous findings from Meyer et al. [13] and Zacks and Tversky [18] concluding that bar charts are a suitable visualisation medium to support reading exact values, identification of maxima, and describing contrasts in data.

More interestingly, this study highlighted a shortcoming in the ability of average, non-expert viewers to recognise the limitations of different chart types—viewers *don't know what they don't know*. This is a significant issue as it means that there is a strong possibility that viewers are likely to make incorrect inferences from charts, or that they can be very easily misled using charts. This finding reinforces the need to carefully design charts for different tasks [15, 6] and highlights a shortcoming in the data literacy of non-experts.

Another interesting point arising from the apparent effectiveness of histograms compared to bee swarms is that it reinforces the finding by Fischer et al. [4] that viewers find it easier to interpret vertical bars (present in histograms) than horizontal bars (present in bee swarms). We also believe that there might be a difference between centring the data points in a bee swarm around a virtual vertical axis in the middle of the chart, and placing the data points upwards starting from the X axis in a histogram. The gap between two bars lying on the same axis can be easily quantified visually, while the gap between two bars centred on an axis is halved on both sides of the bar making it more difficult to quantify. The alignment of the bars with respect to the axis might affect perception—this warrants further study.

The main focus of our future work, however, is to extend the work presented here to examine how effectively different chart types allow viewers to quantify the differences between pairs of distributions. This is part of a broader effort to understand the limits of the general public's data literacy with respect to data visualisations.

6. REFERENCES

[1] A. Beauchamp. What is data literacy?, January 2015.

[2] J. Calzada Prado and M. Á. Marzal. Incorporating data literacy into information literacy programs: Core competencies and contents. *Libri: International Journal of Libraries & Information Services*, 63(2):123 – 134, 2013.

[3] W. S. Cleveland and R. McGill. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *J. American Statistical Association*, 79(387):531–554, 1984.

[4] M. H. Fischer, N. Dewulf, and R. L. Hill. Designing bar graphs: Orientation matters. *Applied Cognitive Psychology*, 19(7):953–962, 2005.

[5] S. N. Friel, F. R. Curcio, and G. W. Bright. Making sense of graphs: Critical factors influencing comprehension and instructional implications. *Journal for Research in Math. Education*, pages 124–158, 2001.

[6] N. Glazer. Challenges with graph interpretation: A review of the literature. *Studies in Science Education*, 47(2):183–210, 2011.

[7] J. Harris. Data is useless without the skills to analyze it. *Harvard Business Review*, 2012.

[8] J. Heer and M. Bostock. Crowdsourcing graphical perception: using mechanical turk to assess visualization design. In *Proceedings of CHI*, pages 203–212. ACM, 2010.

[9] J. Heer, M. Bostock, and V. Ogievetsky. A tour through the visualization zoo. *Communications of the ACM*, 53(6):59–67, 2010.

[10] T. Koltay. Data literacy: in search of a name and identity. *Journal of Documentation*, 71(2):401–415, 2015.

[11] K. Krippendorff. *Content analysis: An introduction to its methodology*. Sage, 2012.

[12] J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, pages 159–174, 1977.

[13] J. Meyer, D. Shinar, and D. Leiser. Multiple factors that determine performance with tables and graphs. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 39(2):268–286, 1997.

[14] M. Schield. Information literacy, statistical literacy and data literacy. *IASSIST Quarterly*, 28(2/3):6–11, 2004.

[15] P. Shah and J. Hoeffner. Review of graph comprehension research: Implications for instruction. *Educational Psychology Review*, 14(1):47–69, 2002.

[16] R. Womack. *Data Visualization and Information Literacy*, volume 38. 2014.

[17] S. Wright, M. Fosmire, J. Jeffryes, M. Stowell Bracke, and B. Westra. A multi-institutional project to develop discipline-specific data literacy instruction for graduate students. Libraries Faculty and Staff Presentations, Paper 10, 2012.

[18] J. Zacks and B. Tversky. Bars and lines: A study of graphic communication. *Memory & Cognition*, 27(6):1073–1079, 1999.