How should we design violin plots?

E. Molina, L. Viale & P. Vázquez

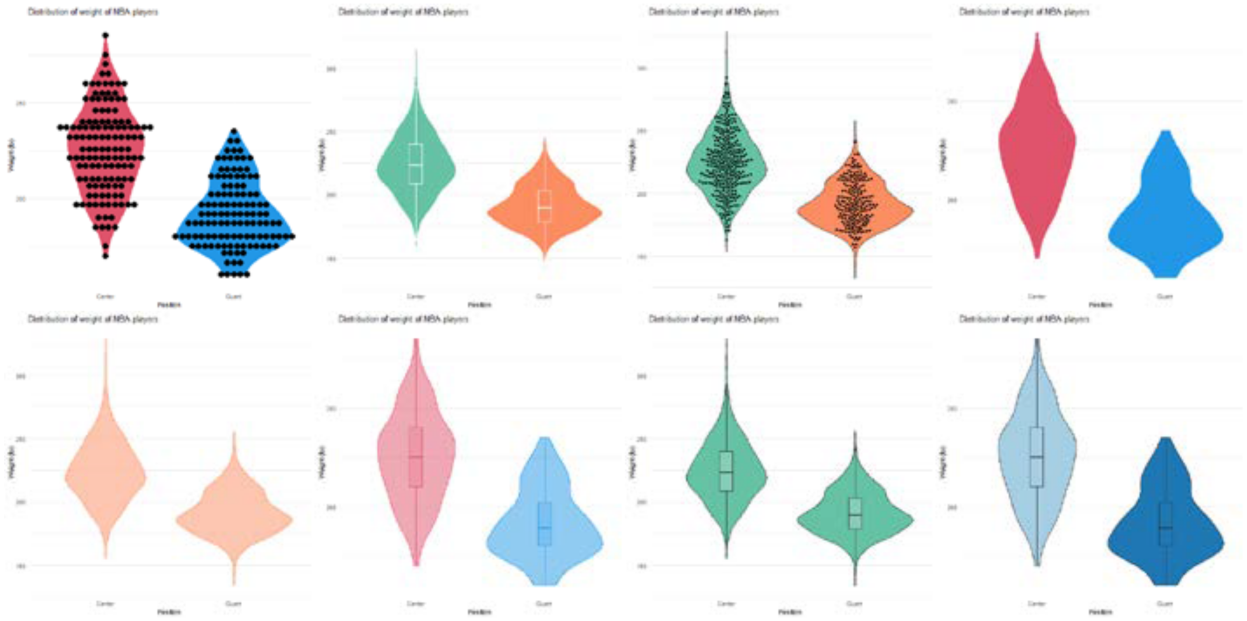

Fig. 1. Some of the designs of violin plots we can find in the literature. Though the original violin plot design has boxplots, the ones used in literature sometimes are enriched with beeswarms, or directly only show the density plot mirrored around an axis.

Abstract—One way to illustrate distributions of samples is through the use of violin plots. The original design is a combination of boxplot and density plot mirrored and plot around the boxplot. However, there are other designs in literature. Although they seem a powerful way to illustrate distributions, the fact that they encode distributions makes them difficult to read. Users have problems comparing two different distributions, and certain basic statistics such as the mean can be difficult to estimate properly. To get more insights on how people interprets violin plots, we have carried out an experiment to analyze how the different configurations affect judgments over values encoded in those plots.

Index Terms—Violin plot, visualization, computing, perception, distribution, perceptual study, modeling.

- *E. Molina is a second year PhD student in Universitat Politècnica de Catalunya.*
  E-mail: elena.molina.lopez@upc.edu.
- *L. Viale is a student in Universitat Politècnica de Catalunya.*
  E-mail: laia.viale@estudiantat.upc.edu.
- *P. Vázquez is an associate professor in Universitat Politècnica de Catalunya.*
  E-mail: pere.pau.vazquez@upc.edu.

1 INTRODUCTION

Violin plots are designed to provide information about the distribution of a set of items. Their original design [12] extended the boxplot design with a "density trace", the smooth version of histograms, also known as density plots (or ridgeline plots), with the goal of comparing multiple distributions. Actually, the authors added the density plots around the boxplot twice, to facilitate comparisons. Violin plots are adequate when we are interested in how the values distribute, and the individual data is not so relevant. Many variations of violin plots have been proposed in the literature. From simple versions that do not include the boxplot (actually, are the most widespread version, if we search for images in the web), to other versions that include beeswarms. The most common representations, as well as some style variations, are shown in Figure 1. Despite violin plots may do a good job to communicate the shape of a distribution, they are difficult to interpret for most of the people, especially when comparing different probability distributions [13]. Obviously, the higher the exposure to this kind of charts, the better should be at interpreting values. Unfortunately, violin plots are not very popular and, though they may be suited for certain tasks, one might wonder how accurate are judgments based on violin plots. To cast more light on those issues, and complement previous research, we ran a user study on the perception of different visual configurations of violin plots for several tasks. In this paper, we reflect on the results obtained. Our main contribution is a user study that covers different violin plot variations and the analysis on how these configurations may affect the estimation of mean, the maximum density, and the comparison of different distributions, including bimodal distributions.

The rest of the paper is organized as follows. Section 2 reviews previous work. Section 3 describes the hypotheses and the design of the experiment. Section 4 presents our analysis of the gathered data. The results obtained are summarized in Section 5. We analyze the study with the opinions of the participants in section 6. Finally, we conclude with potential lines for future research in Section 7.

2 PREVIOUS WORK

Data visualization practitioners have written all sorts of guidelines to help other people to create their designs effectively. However, these guidelines, though very useful, are commonly quite general [20], and do not help to understand how well do people interpret concrete charts. In the same line, some researchers have also created general guidelines (e.g., [22]). Moreover, newer research has shown that one cannot blindly adhere to those general principles. For example, using more ink than the *strictly necessary* can help to communicate context about the data [2], or to give a takeaway/actionable message [19]. Therefore, it is common to design charts that include redundant information (such as an overlaid beeswarm in a violin plot) to facilitate the data interpretation. To entirely understand how well people will interpret charts, user studies using crowdsourcing platforms such as Amazon Mechanical Turk and Prolific have been demonstrated to be useful to run these experiments, when taking enough precautions to avoid careless users [10].

A recent survey by Franconeri et al. [7] overviews what has been discovered in several research fields such that require visual communication of data. However, experimental data must be handled cautiously, and not all the studies can obtain definitive conclusions. For example, the controversial technique of pie charts have lengthily been criticized because humans do not read angles very well. But new research has shown that angle may not be the cue that humans use to estimate quantities in pie charts [21]. However, though the experiments were intended to discriminate among angle, arc length or area, it is still unclear whether we do estimate values by measuring the area or the length of the arcs in these charts.

Regarding distribution data, there are many ways to create comparative visualizations, as surveyed by Blumenschein et al. [**?**]. They also analyzed the utility given by experts to many of those, including violin plots. Some representations have been studied from the perceptual perspective. For example, vIbrekk and Morgan [14] studied the effect of explanations and the lack of in the interpretation of many distribution representations (including box plots and symmetric density plots), showing that they are difficult to read by either laymen and educated users. Other distribution representations such as quantile dot plots [23] and cumulative distribution plots [14] have also been successfully tested for certain tasks that involve estimation of values under uncertainty conditions [6, 17].

Hullman et al. [13] compared error bars, violin plots, and hypothetical outcome plots (HOPs), an animated version of the samples of a distribution. They found that error bars and violin plots were better than HOPs at estimating individual values in one variable setups, but HOPs seem superior as the number of variables to compare grows to two or three. There is, however, a larger space for experimentation. For example, we are not aware of perceptual studies evaluating the performance when violin plots have different modes, such as tasks requiring to compare two distributions, one being unimodal, and the other bimodal. Or how the different designs (such as including a beeswarm) may affect the estimation of values. Therefore, our purpose was to build upon this kind of work to better understand how people read violin plots of different kinds.

Though our work is centered in violin plots, it is also worth mentioning that other researchers have proposed alternative representations, such as the beanplots [16]. Beanplots include the individual samples as lines. Thus, they are in some sense similar to violin plots with a beeswarm. Beanplots can be suitable for a low number of observations, but will generate clutter if the samples are big. Moreover, to facilitate a more direct comparison of two distributions, beanplots, can be designed to hold two distributions. This is commonly handled by avoiding the mirroring part, and showing a density plot in each part of the vertical axis. The same strategy has also been used with violin plots. However, this only scales for two distributions, and our goal was to deal with the typical violin plots we find more commonly in literature. Other extensions to violin plots are the so-called violin superplots [18], where the authors include several distributions inside a violin, each one with its shape, similar to a streamgraph, but in vertical.

In previous works, researchers have studied how different auxiliary elements or configurations of the same plots may influence the accuracy in the estimation of values. For example, Díaz et al. found that a gridline on top of bar charts was related to a better reading of values in bar charts [3]. Fuchs et al. [8] found that Star glyphs were easier to read under certain configurations, such as using outlines. Our work is similar to these two approaches in that we examine how different configurations of the same base plot (we call them auxiliary elements) affect the accurate estimation of the values encoded. However, the fact that we evaluate charts that represent distributions leads us to ask questions that go beyond simply estimating values. We believe that our results can be considered a step forward to get a proper guideline [4].

3 Experiment design

As already noted, violin plots depict distributions, and therefore, are suitable to address tasks related to distribution characterization and comparison [1]. We were especially interested in how people read certain values, such as the mean, as well as how univariate and bivariate distributions can be compared using violin plots. Moreover, we wanted to explore how the addition of elements such as the boxplot (present in the original violin plot design [12]) or beeswarm can help interpret those values.

3.1 Research questions and hypotheses

After analyzing a set of designs found on the internet (see the following subsection), we observed that there were many questions worth to analyze. To define a finite set of objectives, we ended up restricting the goals for the project. These are summarized in the following research questions:

- Which of the auxiliary elements of the violin plot can help users to be more accurate?
- Do bimodal distributions pose problems at perceiving violin plots?
- Do violin plots at similar height facilitate the comparison of distribution widths? And, as an extension, would different heights make width perception more difficult?

To answer those questions, we need to define a concrete set of hypotheses that can be individually tested through experimentation. The list of hypotheses we defined is:

- **H1**: Mean estimation is easier with a boxplot as auxiliary element than using a beeswarm.
- **H2**: Density comparison is easier when evaluating unimodal versus bimodal than with two unimodal distributions.
- **H3**: Density comparison at a fixed point is easier when both plots are at different heights (that is, distributions are quite different).
- **H4**: Probability (percentage) estimation in a bimodal distribution generates more error than an unimodal distribution.
- **H5**: Beeswarms and boxplots facilitate the estimation of probabilities (percentages).
- **H6**: It is easier to compare maximum density of violin plots when they are placed at similar heights.

Note that, in an effort to make use of a simple language, we wrote the tasks related to probabilities as percentages, since we expected them to be more accessible. We also selected a dataset and tasks to make them relatable, such as in the paper of Kay et al. [17] that presents their research around the question of when is a bus arriving.

3.2 Violin plot designs

To determine what representations of violin plots were common in literature, we first looked at research papers from EuroVis and IEEE Vis conferences. However, the sample was minimal. Therefore, we changed our strategy and performed a Google image search with the term "violin plot". After discarding the images that were not violin plots, we selected the first 100 and gathered the following information:

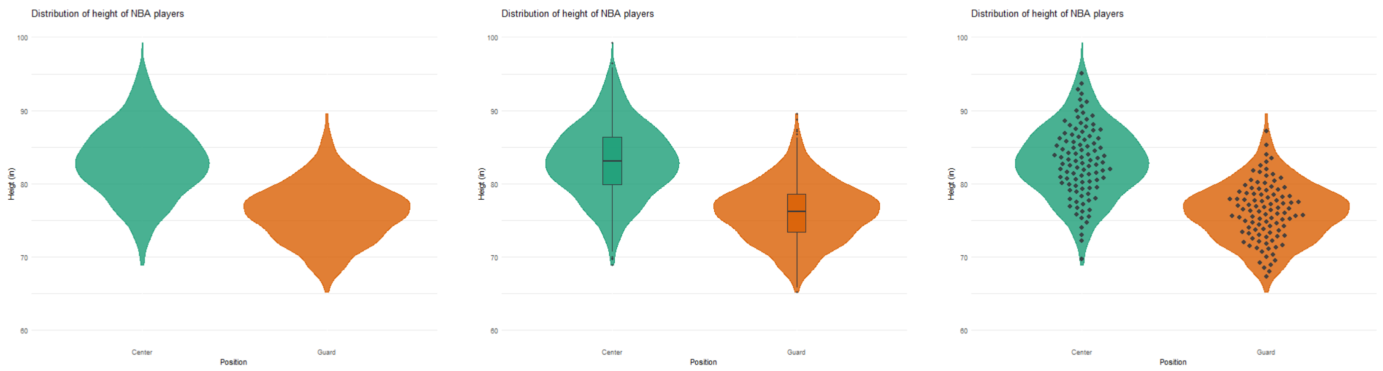

Fig. 2. The design of the violin plots used for the experiment. Left image shows the simple violin plot, center shows the original violin plot, with a boxplot, while the rightmost figure shows the violin plot with a beeswarm.

i) Single or multiple plots, and number of graphical representations, *ii)* Symmetry and whether a symmetry axis was present, *iii)* Whether multiple plots were drawn with a single or multiple colors and the contour color, *iv)* Unimodal or unimodal and bimodal distributions, *v)*, Explicit encoding of outliers, *vi)* Orientation, *vii)* Auxiliary elements: boxplot, beeswarm, or nothing.

We then proceeded to analyze the data and found that most of the charts were vertical, symmetric, and had no symmetry axis. In addition, the majority also consisted of at least two plots, filled with different colors, and an outline. When extra elements were present, the most common was the boxplot, and some cases showed a distribution of samples in the form of points. On the downside, the search showed many webpages that explained what violin plots are, and thus, they are probably not highly representative on how they are used in a professional environment, for example. We observed that Google Images results exhibit a higher prevalence of violin plots with a boxplot than what appears to be common in research articles. After discussing what we found, we decided to test the following tentative configurations:

- Number of charts: 2.
- Orientation: vertical.
- Same color and different colors for the two charts.
- Color palettes: continuous and categorical.
- Outline: no outline, darker, and lighter than the area color.
- Background: white.
- Grid: Horizontal (with a low number of grid lines).

These different variations are the ones represented in Figure 1. To further reduce the configuration space, the three authors analyzed samples of those variations individually, and agreed on a subset that was not too large to perform a crowdsourced study. The set of variations includes parameters in the following categories:

- Overall chart: we chose to include two vertical violin plots, to provide comparison tasks.
- Graph elements: violin plots without boxplot (we call them simple), with boxplot, and beeswarm. All of them without explicit ouliers.
- Chart design: categorical colors for the fill (the most common configuration found), no outline (since it overlapped with the beeswarm).

None of these decisions was taken arbitrarily. First, we needed to reduce the experimental space. But we also wanted to avoid the collision of different design elements. Therefore, when the authors thought that the addition of some graphical elements might generate unnecessary clutter, we chose the simpler (visually) configuration. For example, we skipped charts with outliers because they could be confused with beeswarm points. Furthermore, when there were several options (e.g., on the concrete hues to use), we selected well-known sources. For example, the colors of the violin plots were chosen from a categorical ColorBrewer [9] palette. The resulting set of configurations used in the experiment is depicted in Figure 2.

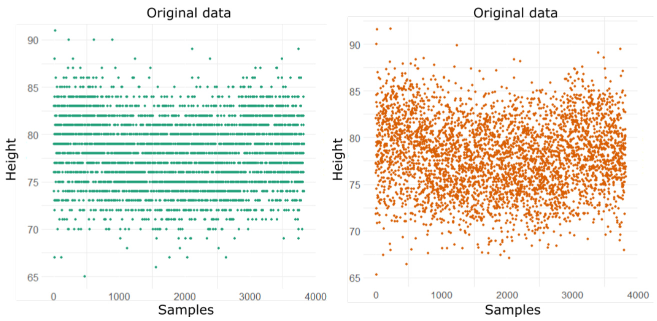

Fig. 3. Original (left) and data with a random decimal to get a smoother distribution.

3.3 Data acquisition and processing

To generate the charts, we used data from Kaggle, more concretely, a dataset of NBA players stats since 1950[1]. We chose the height and the position of the players as the variables to analyze. To get two different distributions, we reduced the players positions from 7 (center, center-forward, forward, forward-center, forward guard, guard, guard-forward) to 2 (center, guard) by rewriting the first four to *center*, and the last three to *guard*. As a result, we have 2466 samples in the *center* position, and 1931 for *guard*. We also further analyzed the data to check whether rounding to the closer value was present, which sometimes happens with human-measured data [15]. Since there was some evident rounding when plotting the values(see 3-left), we added a random decimal to all values. As a result, the distribution looks more natural (see 3-right).

3.4 Tasks

When designing the experiment, we tried to make tasks and data understandable and relatable. That is why we chose a dataset of NBA players. Then, over different tasks that we initially considered, we selected some that could be expressed with simple language. Thus, we ended up defining the following tasks:

[1] https://www.kaggle.com/datasets/drgilermo/nba-players-stats

- Task A: Choose the player position in which there is a larger density of players.
- Task B: Guess the percentage of players above a given height on the left plot.
- Task C: Guess in which player position is more likely to belong to a player with a given height.
- Task D: Determine the average height of players on the left plot (closer to the axis).

3.5 Data gathering

Once the tasks were designed, we prepared an initial tutorial, that describes what a violin plot is, shows the different configurations used in the study and describes the whole process. We also prepared an initial set of three charts, where the users had to solve the 4 tasks, so that participants got familiar with the interface. In this case, for numerical answers, we provided a set of options through radio buttons, instead of leaving the user with a text box for a free answer. Each participant was shown 24 charts (including control charts) sorted randomly, in which they had to solve the 4 tasks.

A pilot study determined that the average time of completion would be below 20 minutes. We assigned a payment of the minimum wage estimated by Prolific[2] (a service designed for crowdsourcing experiments) per hour for 22 minutes. We configured the Prolific study to have a 50% gender balance, and made it available for all countries in the world where Prolific is deployed, but requiring English knowledge to perform the study. Moreover, we started the project around 2 pm in Central European Summer Time (UTC +2) so that people both in Europe, Africa, and America could answer the questionnaire.

Without considering a few participants who skipped the study or timed-out (Prolific set a maximum amount of time for the study above 50 minutes to detect participants who did not continue), we got 48 answers on Prolific. We also passed the link to the experiments to our friends and colleagues and got 23 extra answers.

4 Data Analysis

4.1 User filtering

To guarantee the ecological validity of the samples, researchers often perform a set of control tests before and during the experiments. The first one is intended to determine that participants have understood their tasks. The second one tries to avoid users that randomly answer to get the money in the minimum amount of time. Our initial control did not work due to a bug, so we got users that would not get in other circumstances. Luckily, this helped us to see that we had been very strict in the control tests. We also realized that the control tests during the experiment (that consisted of showing the same chart more than once, swapping the plots) were not completely trustworthy. The reason behind is that, except for the qualitative questions, the quantitative ones may be slightly different, not necessarily due to a careless behavior of the participant, but because some questions were difficult to answer properly. Therefore, removing all participants that did not answer a control chart properly would leave us with an unrepresentative set of the whole population.

The errors of the different introductory tasks using the unfiltered sample are very high, especially for some charts. In light of such results, we proceeded to examine those tasks more carefully and decided a new strategy to filter users based on the perceived difficulty. Hence, we relaxed our initial restrictions in a validity test that filtered out participants. To this end, we defined a task as correct when the following conditions were satisfied: **Task A:** Maximum one error in task A, **Task B:** Up to one error (answers with absolute error smaller or equal to 5 units count as correct), **Task C:** Up to one error, **Task D:** No error (imprecision of up to 5 units).

Once we established the correctness factor for each task in the introductory questionnaire, we considered that users failing more than two of these tasks should be discarded, either because they had not understood the problem, or because they were clicking through. This reduced the available sample set from the initial 71 valid users to 46. This way, the errors in the tutorial diminished substantially. For example, the errors in the qualitative questions reduce to a quarter (task A) or roughly half (task C).

Once the users were accepted to the experiment, it is known that we may check whether some of them might be doing it as fast as possible, just to get the money, and without taking care on the instructions [10]. As a result, we analyzed the answers to try to avoid careless users. To address this, we introduced the so-called *control* charts. These are plots that had appeared previously, to check whether the participants were consistent with their answers. Ideally, users would respond the same. Unfortunately, there were some large unbalances because the tasks were complex.When analyzing the charts with higher number of errors, we found that these were difficult to respond. This happens in three cases: unimodal figures with similar widths in task A, plots at similar heights in task C, and average estimation in bimodal figures (task D). We therefore classified the control charts among easy and difficult and allow the users to err in the difficult ones but not in the easy ones.

After both filtering processes, we ended with 39 users (14 female) of ages 20-64 (average 31). Although we had selected a gender-balanced set on the Prolific platform, adding the other participants among friends and colleagues resulted in a non-balanced sample (27 female out of 71). The filtering process left the percentages almost equal.

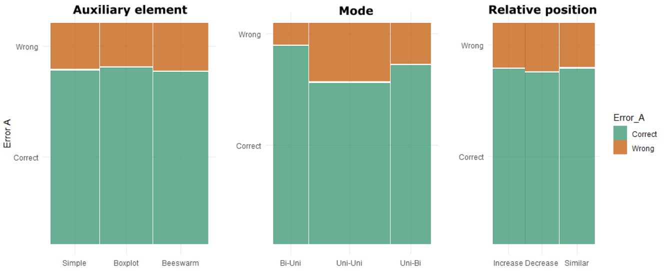

Fig. 4. Error distribution for task A for the different factors. From these plots, it seems that the variable mode may be relevant.

4.2 Data analysis techniques

Before starting with our data analysis, we checked for data normality, we found that none of the answers had a normal distribution. As a result, we could not analyze the data using ANOVA as usual. To facilitate the reading, we are going to first explain the modelling process of each task, and then will summarize the conclusions for the whole set of hypotheses.

4.2.1 Modeling of task A: Larger density detection

Task A is qualitative. Therefore, we can calculate the error as a binary variable. To evaluate the task, we adjusted a generalized linear model of binary response against the explaining variables. We consider the explaining factors to be: *a)* the auxiliary element (simple, boxplot, or beeswarm), *b)* the mode of the chart (bimodal-unimodal, unimodal-unimodal, and unimodal-bimodal), and *c)* the relative position of the charts in the Y dimension (increasing, decreasing, or similar height). We can see the distributions of errors in Task A in Figure 4.

To determine which factor categories are related with the response variable, we applied a chi-square test. The result for task A is that mode is the only relevant factor (p-value below 0.05). After this initial test, we analyzed the data using a complete model with interactions, and then without interactions. The data was analyzed using a binomial distribution and logit link [11]. The result of this analysis is that the parameters unimodal-unimodal and bimodal-unimodal are relevant (p-values of 0.00 and 0.01, respectively). The coefficients in log-odds terms indicate that these have an error probability of 0.25 and 0.17, respectively.

[2] https://app.prolific.co/

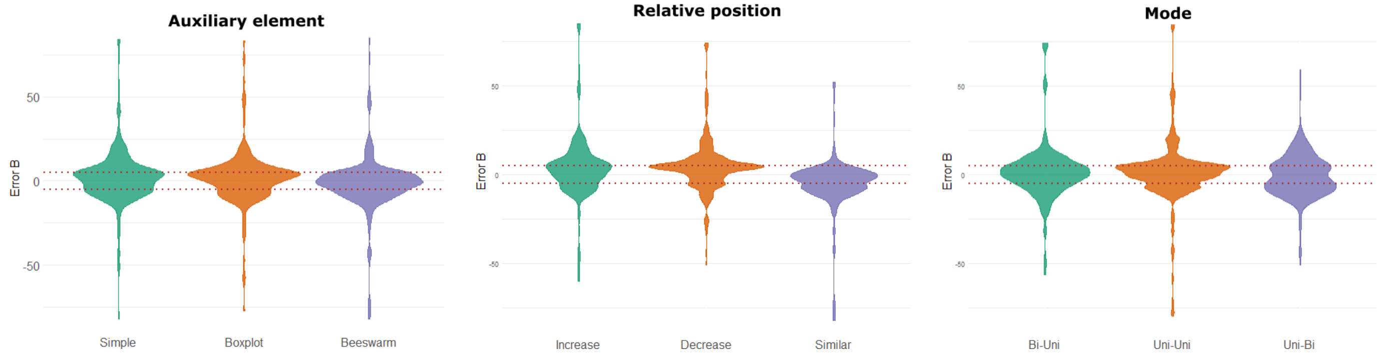

Fig. 5. Error distribution for task B for the different configurations. The dashed lines indicate the $[-5,+5]$ error range already mentioned. While configurations seem not influence the error (left), the heights distribution of the charts look slightly different (center). And crearly, the distributions of unimodal-unimodal answers are clearly above the others (right).

4.2.2 Modeling of task B: Players above a given height

Since task B has continuous response, we analyzed the data using a normal lineal model, where the response variable is the difference between the correct answer and the one by the user in this task. An initial descriptive analysis is shown in Figure 5, where we use violin plots to show the error distributions of the answers in task B for the three different factors. Most of the answers fall in the acceptable error range (indicated by the dashed lines) for the different configurations (left). If we analyze considering the different heights, the answers paint a different picture: we see multimodalities, and the distributions are not so smooth. A larger range in the responses appears when the plots are in increasing heights (from left to right). On the other hand, when the plots are in decreasing height, the errors exhibit less range, but with a positive bias, which seems to indicate that users tend to underestimate the percentage of players over a certain value. Finally, when both plots are at similar height, there appears to be an overestimation of values above a certain point in the Y axis. But the significant differences appear on the mode plot (right), where the error distribution of the unimodal-unimodal configuration is clearly from the others, with larger error ranges. After this descriptive analysis, we analyzed the results with an additive model with three explaining factors. The results showed that when the heights are similar, the mode variable explains most of the error in task B. To confirm this result, we analyzed the data using a stepwise model ([5]) with interactions. The results showed that the mode variable is not significant by itself, but it is when interacting with the relative position.

4.2.3 Modeling of task C: players positions and heights

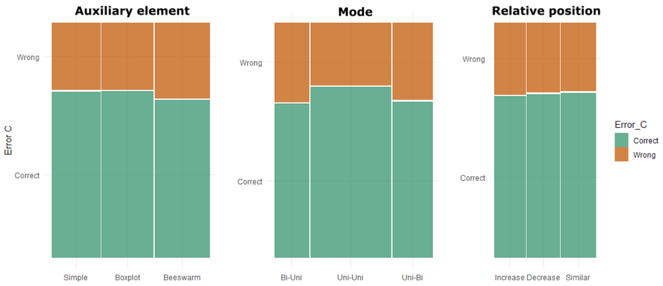

Fig. 6. Error distribution for task C for the different factors. None of the factors seem to have an influence in this task.

Like in the case of task A, we begin with a mosaic plot of the errors to understand their distribution in Figure 6. For the case of the auxiliary element, the configuration with larger relative error is beeswarm, while boxplot and simple violin plots have a similar number of errors. Regarding the mode (center), the one that exhibits a smaller number of errors is the unimodal-unimodal configuration. Finally, the relative positions of the figures do not show large differences in the responses. We then proceeded to analyze the different categories using a chi-square test, and none of them showed significant results. The next analysis was a logit link. With a complete model without interactions, no variable seems significant. However, the model with interactions showed a significant interaction between the mode at unimodal-unimodal level and the relative position at similar heights. The posterior stepwise model did not show significance for any variable. Finally, we validated the stepwise model using an analysis of Pearson residues versus the linear predictor. We found that residues had concrete values, which is anomalous and therefore prevents us to validate this model.

4.2.4 Modeling of task D: Average heights

The task D can be modeled analogously to task B, since both result in continuous answers. The error distributions, faceted by the different factors, is shown in Figure 7. For the auxiliary element, we see distributions with characteristic variations that seem to indicate that errors concentrate mostly in certain values. Note that, however, most of the errors are concentrated in the acceptable intervals. The relative height (center) exhibits the same behavior. On the contrary, the mode factor (right) is different. The unimodal-bimodal distribution (rightmost chart) is not as smooth as the others, though its values are closer to zero. Bimodal-unimodal responses show some degree of polarization, with a second mode larger than zero. This seems to indicate an underestimation of the average in the bimodal distribution. After this initial evaluation, we proceeded to analyze the data using an additive model without interactions. The result is that mode variable in the value unimodal-bimodal is significant to explain the error in task D (p-value 0.00). This was followed by a stepwise modeling, that shows that the only variable included is the mode. This means that the errors seem not to be influenced neither by the auxiliary element nor the relative position of the figures.

5 RESULTS

After modeling all variables, we are going to see what is their meaning regarding the hypotheses we defined.

H1: Estimating averages in a chart with boxplot is easier than without auxiliary element, or with a beeswarm. We expected boxplots to help determine average values because they explicitly encode the median, which, in the representations used in the experiment, were quite similar to the mean. However, the stepwise model that describes the behavior of the error in task D shows that the average estimation model does not include the boxplot variable. Descriptively, we can see that beeswarms seem to produce more biased error estimations. This is confirmed by the coefficients from the additive model, but the numbers are not significant. As such, we cannot accept this hypothesis.

H2: Comparing widths between unimodal and bimodal plots is simpler than with two unimodal plots. The goal of this analysis

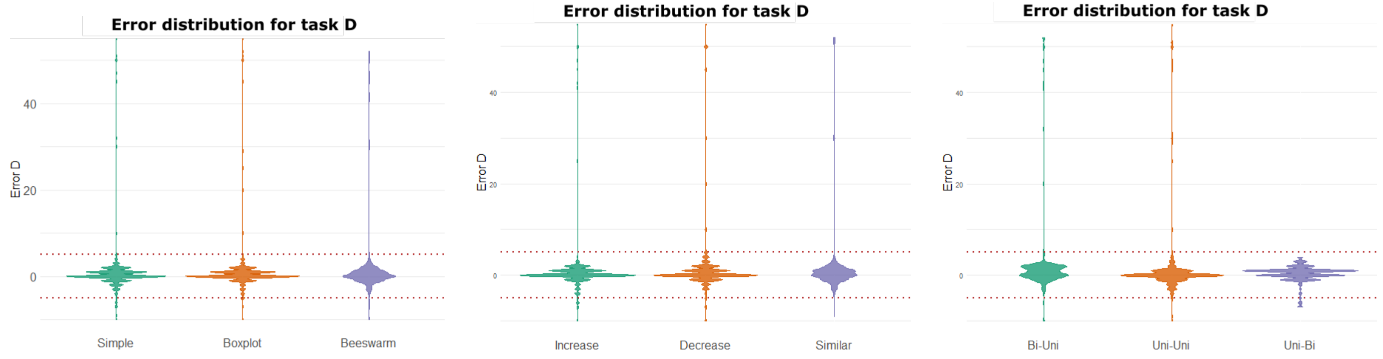

Fig. 7. Error distribution for task D for the different factors: auxiliary elements (left), relative height (center), and mode (right).

was to determine whether a bimodal distribution is more difficult to read than an unimodal distribution. We realize this by analyzing the estimation of average or percentage in these two distributions. This is achieved both in Task A and in Task C. In A, the widths to estimate where at different heights, while in C, we were asking the width for a certain, fixed Y point. Since in our samples the widths were clearly different, the results seemed trivial, but we analyzed them thoroughly. We performed a lineal modelling of the errors in tasks A and C. Afterwards, we applied the stepwise methodology to both complete models. The result is that the coefficients associated to the mode variable are significant in both models. This makes us to accept this hypothesis for our case.

H3: Density comparison at a certain Y point is simpler when figures are placed at different heights In this case, we expected that height would be a significant variable. Unfortunately, the results for tasks A and C are not conclusive, so we cannot accept this hypothesis.

H4: Estimating percentages or averages in bimodal plots is more difficult and generates larger errors This task is related to errors in tasks B and D. In both cases, the modeling procedures found that the mode variable is relevant. Thus, we can accept the hypothesis that there is a difference in percentage and average estimation between unimodal and bimodal figures. To determine whether this is a positive or negative difference, we need to get deeper into the analysis. For model B, the error distribution of bimodal plots seems to be centered in 0, while unimodal charts show the errors centered around the absolute value of 5. But stepwise modeling shows that this error depends on the relative position of the figures. However, the analysis showed that we cannot conclude this. On the other hand, mean estimation (task D) also exhibits larger errors in bimodal distributions. The error seems to be centered at 3 units (absolute value) while for unimodal distributions the error is centered at 0. In this case, it is clear that the estimation error is larger for bimodal distributions than for unimodal ones. As a result, we can accept hypothesis H4 for the average estimation, but it is unclear the concrete percentage.

H5: Addition of a beeswarm or boxplot improves the accuracy in density/percentage estimation This hypothesis is analyzed through task C. But we showed previously that this has not led to conclusions because the modeling did not present a proper validation. Therefore, we can only establish that beeswarm charts produced a larger error in this task than the other auxiliary elements. Note that we also found the model was not acceptable, so it is difficult to extract relevant conclusions.

H6: Similar height in two unimodal figures facilitates the maximum density comparison This hypothesis refers to the error in task A. Here, we found that the interaction between mode and relative position is not significant. But when estimating the maximum density, we found that the error depends on the auxiliary element. More concretely, a violin plot without auxiliary elements was difficult, as well as charts with boxplots. On the contrary, beeswarms showed more accuracy.

6 Study Analysis

After the tasks were answered, we also posed a questionnaire to the users to evaluate their degree of understanding and satisfaction on the study. This was designed as a regular form with answers in a 1-5 Likert scale, where 1 means strongly disagree and 5 means strongly agree. The questions asked were: **Q1:** I have understood the tasks to perform, **Q2:** It has been easy to complete the tasks, **Q3:** I am satisfied with the amount of time spent to complete the tasks, and **Q4:** Open comments.

Most of the people declared they understood the tasks (strongly agree: 20.51%, agree: 64.10%, neutral: 12.82%). Only a tiny bit recognized not proper understanding (disagree: 2.56%, strongly disagree: 0.00%). Participants also found them easy to solve (strongly agree: 7.69%, agree: 51.28%, neutral: 17.97%), but around one quarter found them complex (disagree: 23.08%, strongly disagree: 0.00%) The majority of the participants were also satisfied (strongly agree: 20.51%, agree: 53.85%, neutral: 15.38%), and only a tiny amount declared not having understood them properly (disagree: 10.26%, strongly disagree: 0.00%). Unfiltered sample showed roughly the same percentages.

7 Conclusions and Future Work

After the experiment and the evaluation of the results, it appears that there are more open questions than before. Although the study revealed several insights, some of them opposite to our beliefs, there is still a lot to learn about how people interpret violin charts. We found that mixing unimodal and bimodal charts leads to more difficult tasks than similar unimodal distributions, which are more common. We also found that the relative height seems not relevant to determine the maximum density point, and that boxplots do not seem to help in estimating averages. Thus, our advice (that still requires further experimentation) would be not to use violin plots for bimodal distributions, and keep the design simpler, without boxplots.

Our research questions revolved about what components of violin plots are more useful to interpret distributions. However, we were unable to answer most of the questions through our study. Though the initial set of 71 participants seems big enough, the user filtering process let us with a total of 39 participants. One notable result is that the distribution of the answers was not normal. This prevented us to use a typical ANOVA modeling and had to turn to more sophisticated statistics such as the logit link and the stepwise models.

We think that progress can be made in two different directions. First, we could get more participants into the study to solve some partial results (e.g., beeswarms seem to produce more errors). And we could generate new studies with the more promising configurations and get bigger sample sizes.

Acknowledgments

This project has been supported by project TIN2017-88515-C2-1-R(GEN3DLIVE), from the Spanish Ministerio de Economía y Competitividad, by 839 FEDER (EU) funds.

References

[1] R. Amar, J. Eagan, and J. Stasko. Low-level components of analytic activity in information visualization. In *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.*, pp. 111–117. IEEE, 2005.

[2] M. Correll and M. Gleicher. Bad for data, good for the brain: Knowledge-first axioms for visualization design. In *IEEE VIS 2014*, 2014.

[3] J. Díaz, O. Meruvia-Pastor, and P.-P. Vázquez. Improving perception accuracy in bar charts with internal contrast and framing enhancements. In *2018 22nd International Conference Information Visualisation (IV)*, pp. 159–168. IEEE, 2018.

[4] A. Diehl, M. Kraus, A. Abdul-Rahman, M. El-Assady, B. Bach, R. S. Laramee, D. Keim, and M. Chen. Studying visualization guidelines according to grounded theory. *arXiv preprint arXiv:2010.09040*, 2020.

[5] N. R. Draper and H. Smith. *Applied regression analysis*, vol. 326. John Wiley & Sons, 1998.

[6] M. Fernandes, L. Walls, S. Munson, J. Hullman, and M. Kay. Uncertainty displays using quantile dotplots or cdfs improve transit decision-making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pp. 1–12, 2018.

[7] S. L. Franconeri, L. M. Padilla, P. Shah, J. M. Zacks, and J. Hullman. The science of visual data communication: What works. *Psychological Science in the public interest*, 22(3):110–161, 2021.

[8] J. Fuchs, P. Isenberg, A. Bezerianos, F. Fischer, and E. Bertini. The influence of contour on similarity perception of star glyphs. *IEEE transactions on visualization and computer graphics*, 20(12):2251–2260, 2014.

[9] M. Harrower and C. A. Brewer. Colorbrewer. org: an online tool for selecting colour schemes for maps. *The Cartographic Journal*, 40(1):27–37, 2003.

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

[11] J. M. Hilbe. *Logistic regression models*. Chapman and hall/CRC, 2009.

[12] J. L. Hintze and R. D. Nelson. Violin plots: a box plot-density trace synergism. *The American Statistician*, 52(2):181–184, 1998.

[13] J. Hullman, P. Resnick, and E. Adar. Hypothetical outcome plots outperform error bars and violin plots for inferences about reliability of variable ordering. *PloS one*, 10(11):e0142444, 2015.

[14] H. Ibrekk and M. G. Morgan. Graphical communication of uncertain quantities to nontechnical people. *Risk analysis*, 7(4):519–529, 1987.

[15] B. Jones. *Avoiding data pitfalls: how to steer clear of common blunders when working with data and presenting analysis and visualizations*. John Wiley & Sons, 2019.

[16] P. Kampstra. Beanplot: A boxplot alternative for visual comparison of distributions. *Journal of statistical software*, 28:1–9, 2008.

[17] M. Kay, T. Kola, J. R. Hullman, and S. A. Munson. When (ish) is my bus? user-centered visualizations of uncertainty in everyday, mobile predictive systems. In *Proceedings of the 2016 chi conference on human factors in computing systems*, pp. 5092–5103, 2016.

[18] M. Kenny and I. Schoen. Violin superplots: Visualizing replicate heterogeneity in large data sets. *Molecular Biology of the Cell*, 32(15):1333–1334, 2021.

[19] C. N. Knaflic. *Storytelling with data: A data visualization guide for business professionals*. John Wiley & Sons, 2015.

[20] J. Schwabish. The practice of visual data communication: What works. *Psychological Science in the Public Interest*, 22(3):97–109, 2021.

[21] D. Skau and R. Kosara. Arcs, Angles, or Areas: Individual Data Encodings in Pie and Donut Charts. *Computer Graphics Forum*, 2016. doi: 10.1111/cgf.12888

[22] E. R. Tufte. The visual display of quantitative information. *The Journal for Healthcare Quality (JHQ)*, 7(3):15, 1985.

[23] L. Wilkinson. Dot plots. *The American Statistician*, 53(3):276–281, 1999.