A Comment on "Evaluation of Sampling Methods for Scatterplots"

Steve Haroz

Abstract — "Evaluation of Sampling Methods for Scatterplots" claims that people perform dot-density comparisons best when using random sampling rather than other sampling approaches. This claim and other core conclusions of the article are not supported by the article's empirical evidence. The article's reported results and figures do not meet its own stated threshold of statistical significance, and the analyses are ill-suited for the research questions. Some of these issues are present in the article and could have been spotted by reviewers, whereas other issues were only noticeable because I requested the code and data after publication (which a reviewer would have been prohibited from demanding during the IEEE TVCG review process). A reanalysis calls into question whether any generalizable claims can be made from these results. This comment, its analysis code, and the original article's data are freely available at http://osf.io/hsujr.

1 Internal Contradictions

The figures and numerical results from experiment 1 of "Evaluation of Sampling Methods for Scatterplots" [4] do not support the article's core claims:

- Abstract: "*random sampling is preferred for preserving region density*"
- Section 6.2: "*Random sampling has the highest accuracy (98.63%) and the shortest completion time (2904ms)*"
- Conclusion: "*random sampling is the best in region density preservation in terms of time and accuracy*".

Graphical inference: If random sampling (RS) performed better than the alternatives, there should be a clear difference in the confidence interval of RS compared to the other sampling techniques. But in both the accuracy and response time graphs in figure 7 (recreated and annotated here in fig. 1), the RS confidence interval heavily overlaps with that of at least one other condition.

Numerical results: In section 6.1, the article sets a "standard significance level of $\alpha = 0.05$". But figures 8 and 9 as well as the numerical pairwise comparisons in section 6.2 show that 6 out of 12 comparisons with RS do not meet that threshold. With the article's own results, the claim that RS performs better than all alternatives is not supported.

Takeaway: The results do not match the claims.

2 Mismatched Statistics

Even if the article's results supported the conclusions, the statistical methods are incapable of describing relative performance of RS compared to other techniques.

Between-subject analysis for a within-subject experiment: Each of the 100 subjects in the experiment ran 2 trials for each of the 7 sampling techniques and 8 stimuli datasets, yielding (7 sampling x 8 datasets x 2 repetitions) 112 trials per subject. However, the analyses and visualizations treat the samples as independent. In other words, it is analyzed as though 11,200 different people each ran in one trial.

To compare subject performance by condition, a visualization would need to scale each condition's results relative to a baseline. However, the error bars in the article's visualizations are merely described as "95% confidence intervals" without any indication of such scaling. Confidence intervals that aggregate data irrespective of subject or secondary categories do not communicate which variables explain the differences.

This lack of specificity is apparent in the article, but similar problems with the inferential analysis would be difficult to detect without the analysis code, which was not made available until after the article had been accepted. In that analysis, the data is aggregated by condition before performing a Kruskal-Conover posthoc comparison between each sampling technique. This aggregation inappropriately collapses the subject variable, eliminating information about individual differences. Again, aggregation obfuscates which variables explain the results.

Multiple comparisons: Comparing each combination of the 7 sampling techniques for 2 dependent variables (accuracy and response time) comprises 42 tests. Though the article claims an alpha (false positive rate) of 0.05, conducting many comparisons inflates that rate. A summary of multiple comparison adjustment can be found in Benjamini & Braun 2002 [1]. While the article does not mention any adjustment, the analysis code overrides multiple comparison adjustment via the argument `p_adjust=None`.

Takeaway: The analyses do not match the experiment.

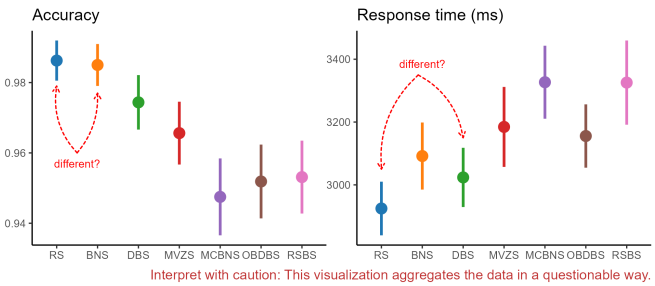

Fig. 1. A recreation of figure 7 shows that the blue RS confidence interval heavily overlaps with at least one other condition in each graph, which should arouse suspicion in a within-subject experiment.

Steve Haroz is with Inria and Université Paris-Saclay.
E-mail: sampling@steveharoz.com

3 Excluded Data

The original analysis code includes this troubling line:

```
if res[k][i]['problem_timespend'] < 30000:
# Filter invalid answers caused by network errors
```

The article does not mention any exclusions. Only 9 trials meet the criterion, so impact is likely minimal. But excluding data without mentioning it in the article is a poor practice.

Takeaway: The article does not report data exclusions.

4 Rethinking the Analysis

Normality: The article uses non-parametric analyses due to the results of a Shapiro-Wilk test. But that test is incompatible with trial counts in the thousands. Instead, I centered and scaled the response time data within each of the 112 conditions and ran t-tests from simulated samples to estimate the false positive rate. The deviations from normality are minor enough to not substantially impact the false-positive rate (see Knief & Forstmeier 2020 [3]).

Here are the results of a within-subjects ANOVA:

Technique (T): $f_{6,594}$=8.5, p=10^{-9}, η_p^2 90% CI [0.041, 0.11]

Dataset (D): $f_{7,693}$=84, p=10^{-88}, η_p^2 90% CI [0.41, 0.49]

Stimulus number (S): $f_{1,99}$=26, p=10^{-6}, η_p^2 90% CI [0.10, 0.32]

T*D: $f_{42,4158}$=4.4, p=10^{-19}, η_p^2 90% CI [0.024, 0.043]

T*S: $f_{6,594}$=3.7, p=0.001, η_p^2 90% CI [0.009, 0.055]

D*S: $f_{7,693}$=27, p=10^{-33}, η_p^2 90% CI [0.17, 0.25]

T*D*S: $f_{42,4158}$=4.4, p=10^{-19}, η_p^2 90% CI [.024, .043]

Categorical variables: The experiment used a limited set of datasets and stimuli rather than generate unique parameterized stimuli per trial. The categorical nature of these variables and their interactions with technique imply that they impact performance differently by technique. There is not enough information to determine the cause of these interactions or how they generalize to a broader population of datasets and stimuli. Consequently, I model them as independent variables rather than pool them.

Two dependent variables: Overall accuracy was 97%, suggesting that manipulations primarily impacted response time. 74% of all errors occurred with the "clothes" dataset, and 100% of remaining errors occurred with stimulus number 0, again implying substantive differences on the impact of dataset and stimulus number.

Relative values: I converted the response times into relative values. For each trial, subtract the subject's response time for the same dataset and same stimulus number but with RS. These relative response times represent how much faster a subject performed relative to RS under similar conditions. Fig. 2 – made with ggdist [2] – shows a subset of conditions with bootstrapped means of these relative values. Many are indistinguishable from noise, and effect size varies by dataset and stimulus number in both magnitude and sign. The categorical conditions appear largely independent of each other, so it would be inappropriate to make strong claims that generalize across these categories.

New conclusion: The inconsistency of performance across categorical unparameterized datasets and stimuli prevents a generalizable claim that one sampling technique yields reliably best performance for other datasets.

5 Discussion

Some of these inaccuracies are apparent even if one only reads the abstract, figures, and conclusion, suggesting a serious failure of the review process. Moreover, because reviewers are prohibited from demanding code or data, less visible omissions and inaccuracies may be impossible to detect if made in other TVCG submissions.

Takeaway: The review process needs to change to facilitate finding similar inaccuracies in the future.

References

[1] Y. Benjamini and H. Braun, "John W. Tukey's contributions to multiple comparisons," The Annals of Statistics, 2002, doi: 10.1214/aos/1043351247.

[2] M. Kay, ggdist: Visualizations of distributions and uncertainty (2.4.0), 2020, doi: 10.5281/zenodo.3879620.

[3] U. Knief and W. Forstmeier, "Violating the normality assumption may be the lesser of two evils," Scientific Communication and Education, preprint, Dec. 2020. doi: 10.1101/498931.

[4] J. Yuan, S. Xiang, J. Xia, L. Yu, and S. Liu, "Evaluation of Sampling Methods for Scatterplots," IEEE TVCG, 2020, doi: 10.1109/TVCG.2020.3030432.

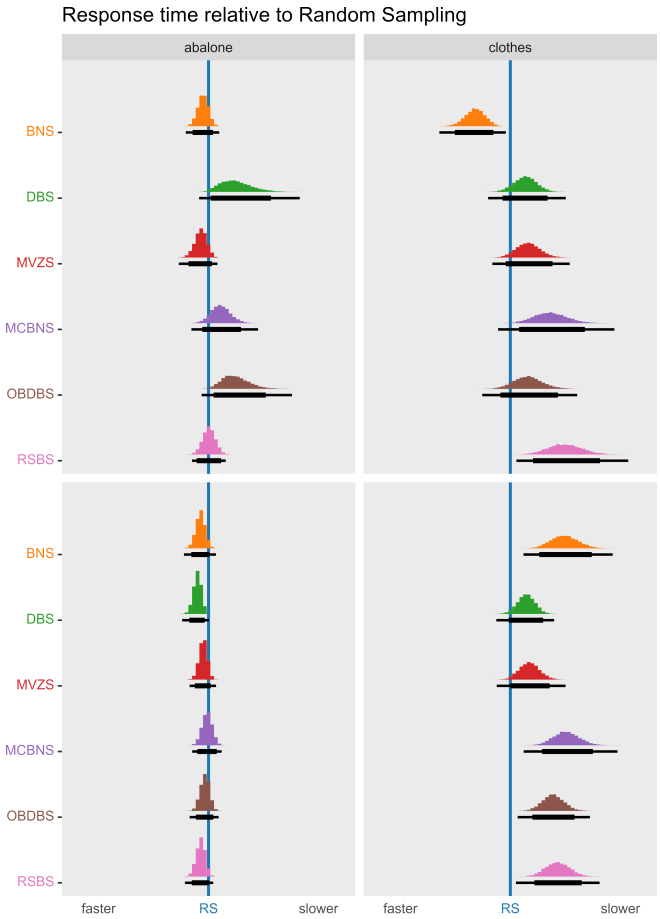

Fig. 2. Response time relative to RS. The thick and thin black bars are the Holm-Bonferroni unadjusted and adjusted 95% confidence intervals of bootstrapped means. Facets: X = Stimulus dataset (2/8 shown, full results at osf.io/hsujr), Y = Repetition number. Does this graph show that all alternative techniques are reliably slower than RS?