#103 Describing Subjective Experiment Consistency by $p$-Value P–P Plot

Jakub Nawała <jnawala@agh.edu.pl>, Lucjan Janowski, Bogdan Ćmiel, Krzysztof Rusek, Marc A. Kastner, Jan Zahálka

Background: Reproducibility

» The Reproducibility track @ ACM MM'21
» Authors of papers from ACM MM'20 & '19 only
» Interactive & open review process
» The Results Reproduced badge

“The main results of the paper have been obtained in a subsequent study by a person or team other than the authors, using, in part, artifacts provided by the author.”
Background: Original Paper


» A software tool assessing subjective experiment consistency
» Pinpoints potentially problematic stimuli
» Data & code openly available
Background: Original Paper

- Consistent = consisting of mostly typical stimuli.
- Typical vs atypical
- GSD(ψ, θ)
Background: Original Paper

![Graph showing empirical cumulative distribution function (ecdf) of p-values against theoretical uniform cdf]

- The graph illustrates the ecdf of p-values against the theoretical uniform cdf.
- The ecdf is represented by red dots, and the theoretical uniform cdf is shown as a black line.
- The arrow points to the point on the x-axis with a value of 0.043, indicating a specific p-value.

**Key Points**
- The graph helps to assess the distribution of p-values and their alignment with the theoretical uniform distribution.
- The p-value 0.043 is highlighted to indicate a significant finding or threshold for statistical significance.
**Background: Original Paper**

- Source code openly available
  - [https://github.com/Qub3k/subjective-exp-consistency-check](https://github.com/Qub3k/subjective-exp-consistency-check)

- Subjective responses also available (in the *tidy data* format)
  - six studies
  - 21 experiments
  - almost 100,000 responses

- [https://grouplens.org/datasets/movielens/1m/](https://grouplens.org/datasets/movielens/1m/)
**Results:** Successful Reproduction

» What our software is for?
  - assessing consistency of a subjective experiment
  - pinpointing potentially problematic stimuli
  - complementing indications of consistency yielded by other methods*
Results: How to Run the Code?

$ python3 friendly_gsd.py
   hdtv1_exp1_scores_pp_plot_ready.csv

» Assumptions:
   – tidy data as input,
   – required Python packages and Python itself installed.

» In case of doubts, refer to the README.md file on GitHub (or to the new paper).

» Caveat: computations take a lot of time*
Extending Our Work

» Franz Hahn (VQA Group at Universität Konstanz) has already contributed 💌

» You can create an issue on GitHub.
  – Propose new functionality.
  – Report bugs.

» You can test the framework using your own model.
Thank You

» Zhi Li.
» Netflix, Inc.
» PL-Grid Infrastructure.
» Polish Ministry of Science and Higher Education.

Jakub Nawała
jnawala@agh.edu.pl
AGH University of Science and Technology
Department of Telecommunications

www.agh.edu.pl