Grasping the t-test: Designing an Interactive System to Communicate Statistics

Sowmya Somanath
Interactions Lab, Dept. of Computer Science
University of Calgary
and SMART Technologies, Canada
ssomanat@ucalgary.ca

Kazuki Takashima
Research Institute of Electrical Communication
Tohoku University, Japan
takashima@riec.tohoku.ac.jp

Ian Hargreaves
SMART Technologies
Calgary, Canada
IanHargreaves@smarttech.com

Ehud Sharlin
Interactions Lab, Dept. of Computer Science
University of Calgary, Canada
ehud@ucalgary.ca

Edward Tse
SMART Technologies
Calgary, Canada
EdwardTse@smarttech.com

Abstract
Increasingly, decision making is being informed by access to large amounts of data and statistical analysis. However, many decision makers don't have formal statistical training. In this paper we propose an interactive system that uses touch and visualizations to accurately communicate statistical concepts to novice audiences. Specifically, we report on the challenges of designing a system to communicate the results of a common statistical comparison (t-tests) to business audiences at a technology company. Our visualization attempts to clarify data anomalies that are often neglected while presenting t-test results (i.e. bimodal data, low sample size and outliers) via data behaviour and inclusion of common physical metaphors associated with communication of statistics (i.e. 'pulling out outliers'). Our iterative design process ultimately led us to draw upon data physicalization techniques such as constructive visualization in order to inform our solution.

Author Keywords
Statistics, physics-based visualization, touch

ACM Classification Keywords
H.5.m [Information interfaces and presentation (e.g., HCI)]: Miscellaneous.

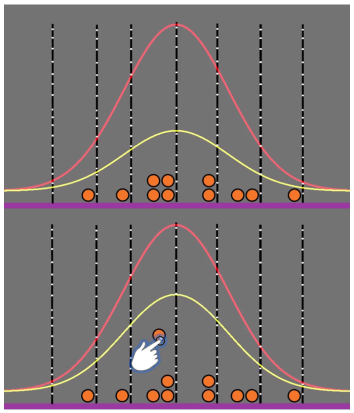

Figure 1: Design 1: Low sample size demonstrated via interaction with the yellow envelope curve. The data points cannot be pushed far due to the resistance presented by the curve.

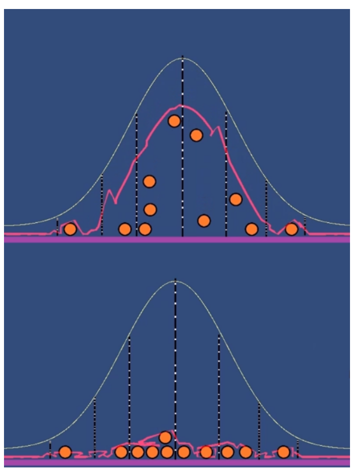

Figure 2: Design 2: Low sample size demonstrated via crumpling envelope curve.

Introduction

With the proliferation of available database and analytic tools (e.g., Google Analytics, Salesforce), companies of all sizes are beginning to track their data, and to view data as a key asset. Gartner has issued a report covering hundreds of use-cases for corporate data ranging from sales to web analytics [1]. Fueling this interest is the assumption that connecting data with decision-making mechanisms will lead to better decisions, and ultimately, to a competitive advantage [4, 12]. This connection of data to decision is crucial.

Access to data and statistical tools in the absence of formal statistical training can lead to misrepresentation, misinterpretation, and error. This raises a troubling question; when it comes to interpreting the results of data, what is the cost of a bad decision? In business, the answer could range from lost earnings to bankruptcy, but it's important to note that this question is not solely restricted to business contexts. Indeed, researchers from disciplines such as HCI [3] and psychology [2, 5] are also struggling with the implications (and hidden costs) of supporting inferential reasoning using null-hypothesis testing (e.g., the debate on p-hacking, [9]). How can we empower decision-makers by making null-hypothesis significance testing more transparent?

Expressions such as "massaging data", "pulling out outliers" are common to conversations about statistical analysis. Inspired by these colloquialisms, we believe that a physics based system utilizing similar embodied metaphors (e.g., grasping, pulling, etc.) could help to free up cognitive resources and provide novice users with a deeper understanding of significance testing. We hope to highlight common misconceptions when reasoning from group averages (e.g., bimodality, non-normal data, outliers, low-n). Moreover, we hope to apply this tool within a business context, where decisions from business presentations are often made using a "good enough" interpretation of the data. Our goal is to support novice decision makers by using gesture and physicality to provide a scaffold to support their reasoning about abstract processes such as *t*-tests.

Related Work

Visualizations seek to offload cognition to external representations such as drawings or physical representations and reduce the cognitive load that users carry when dealing with abstract concepts [8]. Though it is difficult to balance the ease-of-use afforded by infovis with the specificity that statistical graphics require [4], commercial products such as Tableau provide users with visualizations that are appealing and informative. However, Tableau fails to take full advantage of touch-centric interactivity. This is problematic, since a key benefit of using representations are the embodied interactions that they afford [8] and the support for complex cognition about abstract domains [13]. Gestures can also carry additional communicative meaning and reduces the cognitive effort required to learn abstract operations [5].

Data visualizations inspired by physical processes frame abstract processes in familiar terms (e.g., aggregation, decay rate; [7]). Kinetica allows users to explore multidimensional datasets using physics-based visualizations [11], Hans Roslings stage presentations with Gapminder also used gesture to great effect [10], and physicalization makes abstract trends in data into tangible entities [6]. Though the approaches vary, all of these researchers were forced to find an appropriate metaphor that could map abstract relationships found in data to a

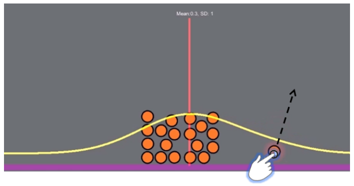

Figure 3: Design 3: Pulling out outlier.

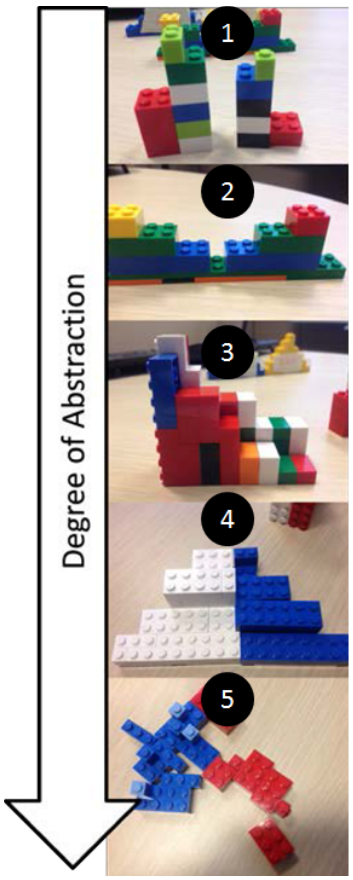

Figure 4: Results of Lego activity organized by degree of abstraction from the data.

concrete representation. In our work, we sought to do the same, but for a system designed to communicate the results of hypothesis testing using a common statistical contrast (Students *t*-test).

Tying Physicality to Statistical Data Representation

Challenge #1
Our first challenge was to identify a physical metaphor that can consistently be used to explain the difference between a significant and not-significant *t*-test.

Challenge #2
How can people identify when to question the validity of a statistical comparison? Such as when distributions are bimodal, when outliers influence measures of central tendency, and when the number of observations is too low to support valid inference. In all of these cases, the intent was for individual data point dynamics to communicate whether a *t*-test was appropriate or not. So the challenge was to identify what dynamic behaviours to design (example: Figures 1, 2, 3) such that viewers could easily identify all of these scenarios and understand how they relate to the idea of means difference testing.

Challenge #3
How do we design interactions for an audience that may have varying levels of statistical expertise?

Preliminary Lego Study

To explore aspects of our three challenges, we conducted a preliminary study using Lego's (similar to the approach used in [6]). The Lego's were presented as simple manipulatives that the participants could use to express their understanding of statistical concepts. We began the study by presenting a brief vignette that detailed a fictitious study (with accompanying hypothesis and *t*-test). We presented the participants with the original data (tabular data with 8 test score values of two groups) and asked them to represent this comparison using Lego. Participants were 5 members of the research group, and had varying degrees of familiarity with statistics. The results were quite diverse, but seemed to align upon a continuum of abstractness (Figure 4):

1. Represented the numeric values of the two group means to signify a *t*-test.
2. Used Legos to create an example of a probability density function that described the comparison of two groups in a general sense (i.e., does not represent this specific data).
3. Grouped by colour, used the X-axis to create a histogram (ordered from highest to lowest) and the Y-axis to indicate score. Disparity in shapes could be used to assess group differences.
4. Grouped by colour, height captures the number of people with a given score, width capture the score, ordinal ranking. Uses shape to identify outliers.
5. Grouped by colour. Individual block constructions represent individual people and their score. Suggested that blocks be randomly draw and scored to test for differences.

We did not observe the use of interactions by our participants, but based on the physical realizations, we recognised that it will be important for our system to show both abstract (numerical mean and standard deviation) and distributional characteristics (number, spread, outliers, etc.) of the data.

Pebble Visualization

Inspired by the Lego study our pebble visualization implementation (Figure's 5, 6, 7, 8) shows both mean

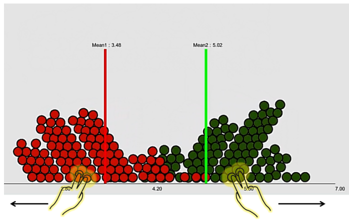

Figure 5: Significance testing: significant difference.

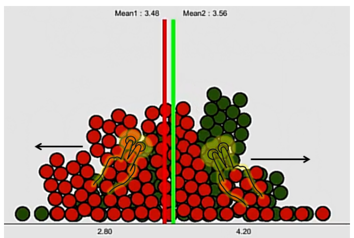

Figure 6: Significance testing: no significant difference.

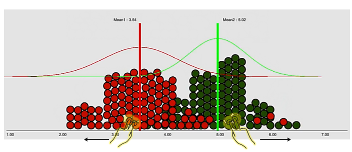

Figure 7: Outlier in data.

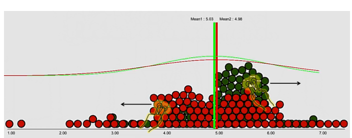

Figure 8: Bimodal data.

and distribution interacting. In our current implementation each pebble is representative of an individual data point and the pebbles are binned according to the probability density function (pdf). We take advantage of physics (gravity, volume packing, contained within boundaries), with pebbles exerting an influence on one-another. Through pebble behaviour we demonstrate concepts such as: large sample size by tightly packed groups, low sample size by loosely packed bins, and outliers as disturbing pebbles with high energy. Significance testing (group differences) is demonstrated by gesturing 'pulling-apart' the two groups. The bin lines are connected to a spring system, and therefore can be physically and visually bend on a pulling gesture. Because of physics, pebbles react to these changing arrangements and look lively visually. If the groups can be separated to a pre-calculated distance, they are significantly different and vice versa (Figure's 5, 6, 7, 8).

Future work and conclusion

In future work we will conduct user-testing with corporate decision-makers. Our study will compare decisions made when statistics are presented using static tables compared with our pebble visualization system. Based on the current literature, we might not see an improvement in decision accuracy, however, we hope to see some evidence of improved depth of understanding or reasoning about null-hypothesis significance testing. Our study will measure behavior, participants' metacognition about their reasoning, and their overall experience with the systems. Our current design is an attempt to answer the challenges of tying physicality to the comprehension of statistics. Like other designers in this area, we face the challenge of reifying abstract concepts, expressing them using new metaphors in a manner that is easily approachable but true to the complexity they stand for.

References

[1] Big Data. *http://www.gartner.com/technology/test/big-data.jsp*.

[2] Cohen, J. The earth is round (p<. 05). *American psychologist 49* (1994), 997.

[3] Dragicevic, P., Chevalier, F., and Huot, S. Running an hci experiment in multiple parallel universes. In *CHI'14 Extended Abstracts* (2014), 607–618.

[4] Gelman, A., and Unwin, A. Infovis and statistical graphics: different goals, different looks. *Journal of Computational and Graphical Statistics 22* (2013), 2–28.

[5] Goldin-Meadow, S., Nusbaum, H., Kelly, S. D., and Wagner, S. Explaining math: Gesturing lightens the load. *Psychological Science 12* (2001), 516–522.

[6] Huron, S., Jansen, Y., and Carpendale, S. Constructing visual representations: Investigating the use of tangible tokens.

[7] Huron, S., Vuillemot, R., and Fekete, J.-D. Visual sedimentation. *Visualization and Computer Graphics, IEEE Transactions on 19* (2013), 2446–2455.

[8] Kirsh, D. Thinking with external representations. In *Cognition Beyond the Brain*. Springer, 2013, 171–194.

[9] Nuzzo, R. Stastical errors. *Nature 50* (2014), 150–152.

[10] Rosling, H. Gapminder. *GapMinder Foundation http://www. gapminder. org* (2009).

[11] Rzeszotarski, J. M., and Kittur, A. Kinetica: naturalistic multi-touch data visualization. In *Proc. of SIGCHI* (2014), 897–906.

[12] Silver, N. *The signal and the noise: Why so many predictions fail-but some don't*. Penguin, 2012.

[13] Wilson, M. Six views of embodied cognition. *Psychonomic bulletin & review 9* (2002), 625–636.