Statistics-IDE: Supporting the Design of Empirical Experiments for Non-Experts during Early Stages of Research Projects

Frode Eika Sandnes[1, 2], Evelyn Eika[1]

[1]Faculty of Technology, Art and Design, Oslo and Akershus University College of Applied Sciences, Oslo, Norway
[2]Westerdals Oslo School of Art, Communication and Technology, Oslo, Norway
{frodes, Evelyn.Eika}@hioa.no

Abstract. Empirical experimentation and inferential statistics are often perceived as challenging among students and researchers. Yet, empirical experiments are highly relevant and appropriate methods in several areas of ergonomics and human computer interaction. This paper argues that the traditional research proposal written in prose is an ineffective mechanism for ensuring sound experiments. This paper discusses the idea of tool-based support for the design of quantitative experiments. The objective of an experimental design tool is to help the experimenter make correct and wise decisions early in the process by providing feedback on ideas at early stages. We discuss the benefits of the IDE (Integrated Development Environment) paradigm as a platform for designing quantitative experiments.

Keywords: Inferential Statistics · Design of Experiments · Integrated Development Environment

1 Introduction

The collection of empirical evidence through quantitative experiments is often viewed as challenging for both students and formally trained researchers. Knowledge about statistics is usually acquired during one's education, and the challenges associated with teaching statistics are acknowledged in several studies [1, 2]. Various pedagogical approaches have been explored for improving statistics education such as problem based learning [3].

One of the most common challenges in inferential statistics includes misconceptions about significance and so-called p-values [4]. Another common challenge is selecting the appropriate test for a given problem [5]. These problems are not only common among students but also among researchers as several studies have revealed the incorrect use of statistics in scientific papers written by formally trained researchers [6, 7]. In some cases, parametric tests are used instead of non-parametric tests although the assumptions for the parametric tests are not satisfied. In an attempt to remedy the situation, practical guides have been published such as McCrum-Gardner's useful tables [8] which clearly show which test to use given the datatype of the observations.

Although perceived by many as difficult to apply, quantitative methods are often highly appropriate in research projects. For example, the authors of this article have employed statistical methods for a wide range of problems, including the study of students' coursework practices [9], cross cultural comparisons of students [10], choice of university life [11], university preferences [12], and specific study tool preferences [13]. Other examples include the quantitative study of phonetic perceptions [14], speakers' ability to communicate emotions in a second language [15], and pronunciation traits of language learners [16].

Computer science research often involves the building of proof-of-concept systems [17, 18] where one measures performance or accuracy [19, 20]. Such systems are often deterministic and give identical results if an experiment is repeated. Human computer interaction is considered a subarea of computer science covering further subareas such as accessibility [21]. The trait of human computer interaction research is the study of humans' interacting with systems. Although such systems can be studied using classical mathematical concepts such as graph theory [22], heuristic evaluation [23], and document studies [24, 25], inferential statistics is usually applied as humans are diverse and one rarely obtains the same results when testing with different individuals. For example, our own research includes the study of text entry rhythm [26], menu-driven text input [27], search systems [28], gesture input [29], directional selection biases [30], text input with joysticks [31], and chording [32, 33]. In other words, there is a demand for inferential statistical knowledge among students and researchers.

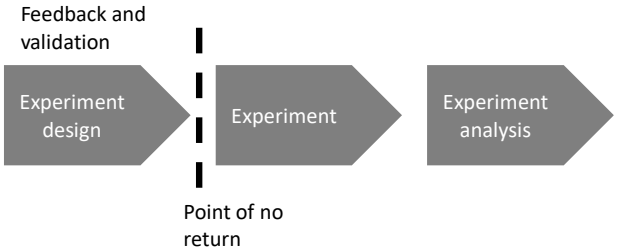

Fig. 1. Experimental design workflow.

2 Weaknesses of Research Proposals

In many study programs, especially at master level, it is common for students to write research proposals which are subject to an approval process before students are allowed to start the research project. The objective of such proposals is to help identify weaknesses so that these can be identified early and reducing the risk of project failure. Research proposals are also common in the research system, for instance, when applying for funding.

Although the intention of such proposals is laudable, it is our opinion that there are several challenges associated with research proposals. First, a proposal is usually written as prose, with verbose descriptions of introduction, literature background, methodological explanation, and timelines. A problem with prose is that it may be hard to interpret and analyze in a structured manner as the information is presented in an ad-

hoc and non-standard way spread throughout the text according to the authors preferences. The reader may therefore risk overlooking certain important details. Moreover, prose can be ambiguous and the reader may perceive something different than what the author intends to communicate, causing potentially problematic issues to go undetected. Moreover, the subjective nature of the textual descriptions means that both the readers and the authors may focus on less relevant passages of the text.

Another problem is the inexperience of some proposal authors. It is easy to elaborate on the easy part of a project and postpone more difficult issues such as details regarding data analysis. It is not uncommon for students to jump into an experiment without a clear idea about how the data is to be analyzed, and some data simply cannot be easily analyzed with the available methods.

It is even more problematic if the reader, or assessor, is inexperienced. Clearly, the statistical knowledge of researchers varies greatly, and it is natural to assume that many proposals have been approved by assessors with insufficient knowledge.

Another issue is the lack of suitable statistical software packages. There are several comprehensive commercial packages such as IBM SPSS, which may seem daunting to use for beginners and inaccessible for many due to its cost. Open source alternatives such as R-Project are quite cumbersome to use for complex mixed analyses of variance. Excel is therefore a popular tool which has several simple statistical procedures built in, but Excel has also received criticisms [34]. More recently, a promising open source alternative JASP [35] has emerged which has an interactive explorative user interface, and it supports commonly needed statistical tests. We have therefore recently started to use JASP in our teaching.

Requirement specifications can be used as an analogy to illustrate the problems with research proposals. A traditional software requirement specification was unstructured and written in prose and very hard for readers to visualize, understand, and assess. Consequently, rapid prototyping emerged as it is much easier for customers to assess a set of specifications expressed as a prototype.

We thus argue for introducing computer assisted tools that allow inexperienced experimenters to assess their project ideas early on before committing to particular design decisions (see Fig. 1).

3 Statistics IDE

Integrated Development Environments (IDEs), such as Eclipse, Netbeans, etc., are the de facto paradigm for developing software. They have also proved useful in teaching programming to students [36]. An IDE revolves around a code editor with syntax highlighting that allows the programmer to easily see if an identifier or variable is spelled incorrectly [37] and proper indention clearly reveals the code structure [38]. Modern development frameworks also strive for reduced coupling and increased coercion by clearly separating different elements such as the model, view and controller [39].

Another powerful feature of IDEs is the integration of compilers and code analyzers. By compiling the program, the programmer receives immediate feedback on the validity and correctness of the program and warnings when the programmer makes decisions that are often associated with bad practices. In this way the IDE contains knowledge

about good software development practices which is shared with the programmer on demand.

The immediate feedback as the program is typed stands in stark contrast to early computing where students had to prepare their programs on punch cards and wait several days before getting the results back from a computer operator. However, practices with research proposals bear similarities to the punch card workflow as the experimenter usually does not receive feedback before it is too late.

We propose the idea of employing the IDE paradigm to the specification of empirical experiments. Key issues that need to be specified are the independent and dependent variables and their datatypes, for example, interval, ordinal, or categorical. The experimenter needs to define if the independent variables are within-groups or between-groups. Another key attribute is to define how many participants one intends to recruit.

Given such a definition, it is possible to automatically detect if the definition is incomplete with information missing, in which case the experimenter can be notified and given a chance to correct the definition. If the user is unfamiliar with terms such as within-group design, the IDE can provide contextual help.

In addition, it is also possible to automatically provide feedback if an experimental design is not feasible, such as a multifactor mixed model design requiring non-parametric tests. Next, the experimenter can be given suggested experimental plans based on the design, that is, randomized treatment orders for within-group designs, and appropriate tests suitable for the design can be suggested. Moreover, the tool can synthesize dummy data which the experimenter can test using one of the many statistical tools available. Thus, an important characteristic of this tool is that it can be disconnected from the actual statistical analysis of later stages which may be performed with state-of-the-art statistical software packages such as R-Project, SPSS, or JASP.

4 Conclusions

The paper argues for employing the IDE paradigm to the design of empirical experiments. The IDE allows experimenters to get early feedback on their ideas based on a knowledge base for experienced statisticians, thereby preventing experimenters from committing to an experiment before all the necessary details of the experiment are specified and validated.

References

1. Phua, K.: How to Make the Learning of Statistics Interesting, Fun and Personally Relevant: Using Progressive Material as Examples for In-class Analysis and to Raise Social Awareness. Radical Statistics 95, 4 (2007)
2. Yilmaz, M. R.: The challenge of teaching statistics to non-specialists. Journal of statistics education 4, 1-9 (1996)
3. Bland, J. M.: Teaching statistics to medical students using problem-based learning: the Australian experience. BMC Medical Education 4, 31 (2004)
4. Haller, H., Krauss, S.: Misinterpretations of significance: A problem students share with their teachers. Methods of Psychological Research 7, 1-20 (2002)

5. Skaik, Y.: The bread and butter of statistical analysis “t-test”: Uses and misuses. Pakistan Journal of Medical Sciences 31, 1558-1559 (2015)
6. Wu, S., Jin, Z., Wei, X., Gao, Q., Lu, J., Ma, X., Wu, C., He, Q., Wu, M., Wang, R., Xu, J.: Misuse of statistical methods in 10 leading Chinese medical journals in 1998 and 2008. The Scientific World Journal 11, 2106-2114 (2011)
7. Yim, K. H., Nahm, F. S., Han, K. A., Park, S. Y.: Analysis of statistical methods and errors in the articles published in the Korean journal of pain. The Korean journal of pain 23, 35-41 (2010)
8. McCrum-Gardner, E.: Which is the correct statistical test to use?. British Journal of Oral and Maxillofacial Surgery 46, 38-41 (2008)
9. Jian, H.-L., Sandnes, F.E., Huang, Y.-P., Cai, L., Law, K.: On students’ strategy-preferences for managing difficult course work. IEEE Transactions on Education 51, 157-165 (2008)
10. Jian, H.-L., Sandnes, F.E., Huang, Y.-P., Huang, Y.-M., Hagen, S.: Towards harmonious East-West educational partnerships: A study of cultural differences between Taiwanese and Norwegian engineering students. Asia Pacific Education Review 11, 585-595 (2010)
11. Jian, H. L., Sandnes, F. E., Huang, Y. P., Huang, Y. M., Hagen, S.: Studies or Leisure?: a Cross-cultural Comparison of Taiwanese and Norwegian Engineering Students' Preferences for University Life. International journal of engineering education 26, 227-235 (2010)
12. Jian, H. L., Sandnes, F. E., Huang, Y. P., Huang, Y. M.: Cultural factors influencing Eastern and Western engineering students’ choice of university. European Journal of Engineering Education 35, 147-160 (2010)
13. Jian, H.-L., Sandnes, F.E., Huang, Y.-P., Law, K., Huang, Y.-M.: The role of electronic pocket dictionaries as an English learning tool among Chinese students. Journal of Computer Assisted Learning 25, 503-514 (2009)
14. Eika, E., Hsieh, Y.: On Taiwanese pupils’ ability to differentiate between English/l/and/r: A study of L1/L2 cross-language effects. First Language 0142723717709106, (2017)
15. Jian, H.-L.: On English speakers’ ability to communicate emotion in Mandarin. Canadian Modern Language Review 71, 78-106 (2015)
16. Jian, H.-L.: Prosodic challenges faced by L1 English speakers reading Mandarin. ACTA Linguistica Hungarica 62, 35-62 (2015)
17. Huang, Y.P., Hsu, L.W., Sandnes, F.E.: An intelligent subtitle detection model for locating television commercials. IEEE Trans. Man Cybern. B 37, 485–492 (2007)
18. Huang, Y. P., Chang, Y. T., Sandnes, F. E.: Ubiquitous information transfer across different platforms by QR codes. Journal of Mobile Multimedia 6, 3-13 (2010)
19. Sandnes, F.E.: Where was that photo taken? Deriving geographical information from image collections based on temporal exposure attributes. Multimedia Syst. 16, 309–318 (2010)
20. Sandnes, F.E.: Determining the geographical location of image scenes based on object shadow lengths. J. Sig. Process. Syst. 65, 35–47 (2011)
21. Whitney, G., Keith, S., Bühler, C., Hewer, S., Lhotska, L., Miesenberger, K., Sandnes, F.E., Stephanidis, C., Velasco, C.A.: Twenty five years of training and educa-tion in ICT Design for All and Assistive Technology. Technology and Disability 3, 163-170 (2011)
22. Sandnes, F. E.: Evaluating mobile text entry strategies with finite state automata. In: Proceedings of the 7th international conference on Human computer interaction with mobile devices & services, pp. 115-121. ACM (2005)
23. Sandnes, F.E., Jian, H.L., Huang, Y.P., Huang, Y.M.: User interface design for public kiosks: an evaluation of the Taiwan high speed rail ticket vending machine. J. Inf. Sci. Eng. 26, 307–321 (2010)
24. Eika, E.: Universally designed text on the web: towards readability criteria based on antipatterns. Stud. Health Technol. Inform. 229, 461–470 (2016)

25. Eika, E., Sandnes, F.E.: Assessing the reading level of web texts for WCAG2.0 compliance—can it be done automatically? In: Di Bucchianico, G., Kercher, P. (eds.) Advances in Design for Inclusion. Advances in Intelligent Systems and Computing, vol. 500, pp. 361–371. Springer, Cham (2016)
26. Sandnes, F.E., Jian, H.L., Pair-wise variability index: Evaluating the cognitive difficulty of using mobile text entry systems. In: International Conference on Mobile Human-Computer Interaction. LNCS, vol. 3160, pp. 347-350. Springer Berlin Heidelberg (2004)
27. Sandnes, F.E., Thorkildssen, H.W., Arvei, A., Buverad, J.O.: Techniques for fast and easy mobile text-entry with three-keys. In: Proceedings of the 37th Annual Hawaii International Conference on System Sciences. IEEE (2004)
28. Berget, G., Sandnes, F.E.: Do autocomplete functions reduce the impact of dyslexia on information searching behaviour? A case of Google. J. Am. Soc. Inf. Sci. Technol. 67, 2320–2328 (2016)
29. Sandnes, F.E., Tan, T.B., Johansen, A., Sulic, E., Vesterhus, E., Iversen, E.R.: Making touch-based kiosks accessible to blind users through simple gestures. Universal Access in the Information Society 11, 421-431 (2012)
30. Sandnes, F.E.: Directional bias in scrolling tasks: a study of users' scrolling behav-iour using a mobile text-entry strategy. Behav. Inf. Technol. 27, 387–393 (2008)
31. Sandnes, F.E., Aubert, A.: Bimanual text entry using game controllers: relying on users' spatial familiarity with QWERTY. Interact. Comput. 19, 140–150 (2007)
32. Sandnes, F. E., Huang, Y. P.: Chording with spatial mnemonics: automatic error correction for eyes-free text entry. Journal of information science and engineering 22, 1015-1031 (2006)
33. Sandnes, F.E.: Can spatial mnemonics accelerate the learning of text input chords? In: Proceedings of the Working Conference on Advanced Visual Interfaces, pp. 245–249. ACM (2006)
34. Mélard, G.: On the accuracy of statistical procedures in Microsoft Excel 2010. Computational statistics 29, 1095-1128 (2014)
35. Marsman, M., Wagenmakers, E. J.: Bayesian benefits with JASP. European Journal of Developmental Psychology 14, 545-555 (2017)
36. Reis, C., Cartwright, R.: Taming a professional IDE for the classroom. ACM SIGCSE Bulletin 36, 156-160 (2004)
37. Sarkar, A.: The impact of syntax colouring on program comprehension. In: Proceedings of the 26th annual conference of the psychology of programming interest group, pp. 49-58, (2015)
38. Kesler, T. E., Uram, R. B., Magareh-Abed, F., Fritzsche, A., Amport, C., Dunsmore, H. E.: The effect of indentation on program comprehension. International Journal of Man-Machine Studies 21, 415-428 (1984)
39. Sandnes, F. E., Eika, E.: A Simple MVC-Framework for Local Management of Online Course Material. In: Uskov V. Howlett R., Jain L. (eds.) International Conference on Smart Education and Smart E-Learning. Smart Innovation, Systems and Technologies, vol. 75, pp. 143-153. Springer, Cham. (2017).