Making Sense of Statistics in HCI: From P to Bayes and Beyond

Alan Dix
HCI Centre
School of Computer Science
University of Birmingham
Birmingham, B15 2TT, UK &
Talis
Birmingham, B1 3HN, UK
alan@hcibook.com
http://alandix.com/

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the Owner/Author.
Copyright is held by the owner/author(s).
CHI'17 Extended Abstracts, May 06-11, 2017, Denver, CO, USA
ACM 978-1-4503-4656-6/17/05.
http://dx.doi.org/10.1145/3027063.3027109

Abstract
Many find statistics confusing, and perhaps more so given recent publicity of problems with traditional p-values and alternative statistical techniques including confidence intervals and Bayesian statistics. This course aims to help attendees navigate this morass: to understand the debates and more importantly make appropriate choices when designing and analysing experiments, empirical studies and other forms of quantitative data.

Author Keywords
Statistics; human–computer interaction; evaluation, hypothesis testing; Bayesian statistics

ACM Classification Keywords
H.5.2. Information interfaces and presentation, User Interfaces: Evaluation/methodology, Theory and methods; G.3 PROBABILITY AND STATISTICS

Benefits
This course is intended to fill the gap between the 'how to' knowledge in basic statistics courses and having a real understanding of what those statistics mean. It will also help attendees to make sense of the various alternative approaches presented in recent articles in HCI and wider scientific literature [e.g. 1,2,3,4].

At the end of the course attendees will have a richer understanding of: the nature of random phenomena and different kinds of uncertainty; the different options for analysing data and their strengths and weaknesses; ways to design studies and experiments to increase 'power' – the likelihood of successfully uncovering real effects; and the pitfalls to avoid and issues to consider when dealing with empirical data.

Attendees will leave better able to design studies that efficiently use resources available and appropriately, effectively and reliably analyse the results

Intended Audience(s)

The course is intended for both experienced researchers and students who have already, or intend to engage in quantitative analysis of empirical data or other forms of statistical analysis. It will also be of value to practitioners using quantitative evaluation.

Prerequisites

The course will assume some familiarity with statistical concepts theoretical or practical, for example, the use of t-tests or similar techniques. There will be occasional formulae, but the focus of the course is on conceptual understanding not mathematical skills.

Content

The course is divided into four main parts:

Wild and wide – concerning randomness and distributions

The course will begin with some exercises and demonstrations of the unexpected wildness of random phenomena including the effects of bias and non-independence (when one result affects others).

We will discuss different kinds of distribution and the reasons why the normal distribution (classic hat shape), on which so many statistical tests are based, is so common. In particular we will look at some of the ways in which the effects we see in HCI may not satisfy the assumptions behind the normal distribution.

Most will be aware of the use of non-parametric statistics for discrete data such as Likert scales, but there are other ways in which non-normal distributions arise. Positive feedback effects, which give rise to the beauty of a snowflake, also create effects such as the bi-modal distribution of student marks in certain kinds of university courses (don't believe those who say marks should be normally distributed!). This can become more complex if feedback processes include some form of threshold or other non-linear effect (e.g. when the rate of a task just gets too much for a user).

All of these effects are found in the processes that give rise to social networks both online and offline and other forms of network phenomena, which are often far better described by a long-tailed 'power law'.

Doing it – if not p then what

In this part we will look at the major kinds of statistical analysis methods:

- Hypothesis testing (the dreaded p!) – robust but confusing
- Confidence intervals – powerful but underused
- Bayesian stats – mathematically clean but fragile
- Simulation based – rare but useful

None of these is a magic bullet; all need care and a level of statistical understanding to apply.

Outline

Wild and wide
Exploring the nature of randomness, uncertainty and 'distributions'

Doing it
Deciding between alternative statistical analyses: the ubiquitous 'p' to Bayesian methods.

Gaining power
Learning how to design studies to avoid the dreaded 'too few participants' problem!

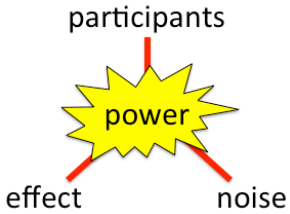
So what?
Making sense of the data you get from your studies and avoiding the pitfalls.

We will discuss how these are related including the relationship between 'likelihood' in hypothesis testing and conditional probability as used in Bayesian analysis. There are common issues including the need to clearly report numbers and tests/distributions used, avoiding cherry picking, dealing with outliers, non-independent effects and correlated features. However, there are also specific issues for each method.

Classic statistical methods used in hypothesis testing and confidence intervals depend on ideas of 'worse' for measures, which are sometimes obvious, sometimes need thought (one vs. two tailed test), and sometimes outright confusing. In addition, care is needed in hypothesis testing to avoid classic fails such as treating non-significant as no-effect and inflated effect sizes.

In Bayesian statistics different problems arise including the need to be able to decide in a robust and defensible manner what are the expected prior probabilities of different hypotheses before an experiment; the closeness of replication; and the danger of common causes leading to inflated probability estimates due to a single initial fluke event or optimistic prior.

Crucially, while all methods have problems that need to be avoided, we will see how not using statistics at all can be far worse.



Gaining power – the dreaded 'too few participants'
Statistical power is about whether an experiment or study is likely to reveal an effect if it is present. Without a sufficiently 'powerful' study, you risk being in the middle ground of 'not proven', not being able to make a strong statement either for or against whatever effect, system, or theory you are testing.

In HCI studies the greatest problem is often finding sufficient participants to do meaningful statistics. For professional practice we hear that 'five users are enough', but less often that this figure was based on particular historical contingencies and in the context of single iterations [5], *not* summative evaluations, which still need the equivalent of 'power' to be reliable.

However, power arises from a combination of the size of the effect you are trying to detect, the size of the study (number of trials/participants) and the size of the 'noise' (the random or uncontrolled factors).

Increasing number of participants is not the only way to increase power and we will discuss various ways in which careful design and the selection of subjects and tasks can increase the power of your study albeit sometimes requiring care in interpreting results. For example, using a very narrow user group can reduce individual differences in knowledge and skill (reduce noise) and make it easier to see the effect of a novel interaction technique, but also reduces generalisation beyond that group. In another example, we will also see how careful choice of a task can even be used to deal with infrequent expert slips.

So what? – making sense of results
You have done your experiment or study and have your data – what next, how do you make sense of the results? In fact one of the best ways to design a study is to imagine this situation *before* you start!

This part will address a number of questions to think about during analysis (or design) including: Whether your work is to test an existing hypothesis (validation) or to find out what you should be looking for

(exploration)? Whether it is a one-off study, or part of a process (e.g. '5 users' for iterative development)? How to make sure your results and data can be used by others (e.g. repeatability, meta analysis)? Looking at the data, and asking if it makes sense given your assumptions (e.g. Fitts' Law experiments that assume index of difficulty is all that matters). Thinking about the conditions – what have you *really* shown – some general result or simply that one system or group of users is better than another?

Practical work

There will be occasional practical exercises, for example, coin tossing experiments … but no complex numerical calculations!

Instructor background

Alan Dix is currently Professor at the University of Birmingham and Senior Researcher at Talis. He is well known for his textbook and research in HCI including CSCW, mobile interfaces, technical creativity, and some of the earliest work on privacy and ethical implications of intelligent data processing. More recent work includes community engagement especially in rural areas and his one thousand mile research walk around Wales, which generated substantial quantitative and qualitative open research data from blogs to biodata.

However, before he was in HCI, Alan was a mathematician, including representing the UK in the International Mathematical Olympiad. He has practised as a professional statistician and applied mathematician including work on modelling agricultural crop sprays, medical statistics and undersea cable detection. Within HCI these skills have been applied in his foundational work on formal methods for interactive systems, the use of Bayesian techniques in education, random sampling for visualisation of big data and uncertainty, and analysis of potential bias against human/applied areas in REF, the UK research assessment exercise.

This unusual combination of skills and experience gives Alan a unique insight into the issues and problems when applying statistics to HCI data.

Resources

A website will accompany this course including full materials and also short videos:

http://alandix.com/statistics

Alan also maintains a number of websites relating to issues of HCI and data, these are accessible from his personal website at:

http://alandix.com/

References

1. Baker, M. (2016). Statisticians issue warning over misuse of P values. *Nature*, 531, 151 (10 March 2016) doi:10.1038/nature.2016.19503
2. Ioannidis, J. (2005). Why Most Published Research Findings Are False. *PLOS*, doi: 10.1371/journal.pmed.0020124
3. Kaptein, M. and Robertson, J. (2012). Rethinking statistical analysis methods for CHI. *Proc. CHI 2012*. ACM, pp. 1105-1114.
4. Kay, M., Nelson, G., and Hekler, E. 2016. Researcher-Centered Design of Statistics: Why Bayesian Statistics Better Fit the Culture and Incentives of HCI. CHI 2016, ACM, pp. 4521-4532.
5. Nielsen, J. and Landauer, T. (1993). A mathematical model of the finding of usability problems. INTERACT/CHI '93. ACM, 206–213.