This is a late draft of the paper - please see journal for final version
International Journal of Mobile Human Computer Interaction. 1 (3). 2009

Paper Rejected (p>0.05): An introduction to the debate on appropriateness of null-hypothesis testing

M. D. Dunlop* and M. Baillie

Computer and Information Sciences, University of Strathclyde,
Richmond St, Glasgow G1 1XH, Scotland
[Mark.Dunlop/Mark Baillie]@cis.strath.ac.uk

Abstract. Null-hypothesis statistical testing has been seriously criticised in other domains, to the extent of some advocating a complete ban on publishing p-values. This short position paper aims to introduce the argument to the mobile-HCI research community, who make extensive use of the controversial testing methods.

Keywords. Systems Evaluation; Human-Machine Systems; Mobile Technologies

INTRODUCTION

The approach of statistical analysis using a null-hypothesis testing has been heavily criticised in other domains. Reinvigorated by Cohen and Meehl's seminal papers (Cohen, 1994) (Meehl, 1990) there has been a long running debate in experimental psychology that has led to The American Psychological Association considering, but not going so far as, a complete ban on reporting of p-values (Wilkinson, 1999). While this debate has reached medicine (e.g. (Ioannidis, 2005)), education (e.g. (Cliner, Leech, & Morgan, 2002)), political science (e.g. (Gill, 1999)) and other branches of computer science (e.g. (Demsar, 2008)) for example, it has yet to take root in HCI, despite our inheritance of many methods from experimental psychology. Nor has it had much effect on text books (Cliner, Leech, & Morgan, 2002) from which we and our students typically learn our statistics – a serious problem because many researchers have to teach themselves statistics and one that is compounded because many of the expectations for good practice are "actually implicit and arise from the culture of practising statistics rather than being found in books" (Cairns, 2007). This short position paper does not directly contribute anything new to this long running debate as there have been several very eloquent essays within other domains, our aim is to introduce the mobile-HCI community to this

discussion and raise awareness of some key papers that discuss the limitations of p-based null-hypothesis statistical testing.

The paper starts with an introduction to the key problems raised in the long discussion in the statistics and experimental psychology domains and moves on to discuss key suggested alternatives - throughout we will make reference to the common use of statistics in mobile HCI work. We feel these issues are relevant to all HCI work but especially relevant to mobile-HCI. Mobiles are used in noisy and complex environments in which the user is often mobile. Experimental design is now, more often than not, reflecting this complex environment to some extent – this makes the studies more complex but also introduces many more potential compounding variables that might bias or simply confuse our results. So the magic formulae of p-testing and ANOVA give us "some degree of reassurance that we are following good scientific practices" (Drummond, 2008). But is this reassurance misplaced or, worse, distorting the investigative nature of science?

KEY PROBLEMS WITH P-BASED STATISTICS

The debate on null-hypothesis testing has identified many "sins" of null-hypothesis significance testing (NHST) and the way that it is normally used in scientific work. Here we look at them as we perceive the severity of the problem in mobile-HCI:

1. Treating NHT as a binary approval of result validity;
2. Confusing strength of p-value results with effect size;
3. Abusing the statistical tests themselves;
4. Making conclusions from non-significant results;
5. Making illogical arguments based on results.

Reviewing recent proceedings of MobileHCI, we are not as guilty as other domains in which null-hypothesis testing has been criticised. However, we tend to be guilty of the first three sins quite widely and we perceive a risk that as publication becomes more competitive, reviewers might push us further along the route of inappropriate statistics.

1: One of the key problems with NHST that has been identified in other domains is the binary treatment of results. The focus on pre-set levels of statistical significance, usually $p<0.05$, leads to simplistic analysis of results: if this level of significance is reached authors tend not to probe deeper as to the reasons and reviewers tend to accept the claims

as valid. On the other hand both authors and reviewers are often much more critical of papers where the results do not reach this level of significance, sometimes without probing deeper into the reasons. However, there is nothing magical about 0.05, indeed the fixed level was originally introduced only for convenience so that back-of-the-book tables could be produced in days before computerised stats packages. In mobile-HCI, as with many domains, we very rarely consider what level of significance is required before an experimental result is meaningful: do we need 95% confidence in rejecting the null hypothesis or would 90% do, or do we really need 99.97 for this kind of result? Reviewing recent mobile-HCI papers about half of them do not report the actual p value confirming the binary treatment of this value, our experiments have either achieved this magical number and thus our results are important or they have not and are thus no better than random: clearly a gross simplification.

2: Most statistics books, and people who use statistics for experimental analysis, know quite clearly that you are more likely to get a statistically significant result with more people. What is less clear is that, if there is a statistically significant result in there, then the value of p is inversely related to the number of subjects – the more people you study the smaller the p value will become. What is not strictly related to the p-value is the size of the effect. A study with a large number of users will most likely find a statistically significant effect but that does not mean that the effect is meaningful, large or *scientifically significant* – it may be a trivial difference that would never be noticed in real use never mind have a commercial benefit. However, the smaller the sample the less likely the sample is to be representative of the real population and, thus, "true" (Ioannidis, 2005). Not reporting effect size in some form becomes especially dangerous when linked with our binary thinking of probability.

3: Most statistical procedures (including the standard t-tests and ANOVA) make strong assumptions about the underlying data and are invalid if these assumptions are not met. In particular, they assume the data is taken from an underlying population that is normally distributed. In many psychological tests, e.g. reaction time, it is assumed that the whole population will follow a normal distribution – this is not true for many mobile tasks and experiments. For example, in text entry there is a very wide range of abilities and it is hard to assess the underlying population spread – there are many people with high performance but a long and important tail. There are techniques to overcome this problem (either use of non-parametric tests or adjusting the data, say by using log values for times) but the discussion of parametric checking rarely happens in experimental

papers. Furthermore, the distribution in "the population" also differs greatly depending on what the underlying population is expected to be – and we rarely report what underlying population we are studying: again for text entry, is it all mobile users, regular 12-key users, teenagers, twin-thumbers, ...? Cairns discusses these and other statistical problems in a review of the use of statistics in British HCI Conference papers (Cairns, 2007).

4: While other domains are more guilty of this than HCI, there is still sometimes a tendency to want to spin a *non-significant result* into a *significant non-result*. This spin negates the whole point of null-hypothesis testing: the authors are trying to use NHST to argue exactly what it is meant to prevent. When we use NHST we are trying to say "the chances of this happening randomly are very low so we have a meaningful difference", the negation is "the chances of this happening randomly are not very low so we have no clear result" and not "the chances of this happening by chance are high, therefore there is no real difference".

5: NHST tests tell us the probability that the observed data occurs by chance given the null-hypothesis is true, usually the probability that we would observe this data given that there is no difference in performance of two systems on a certain measure. This is not the same as the probability that they are the same, nor is 1-p the same as the probability of there being a difference. This is a fairly complex argument involving Baysian probability and modus-tollens validity, we direct the reader to Cohen (Cohen, 1994) for discussion and examples.

KEY SUGGESTED SOLUTIONS

If there is a single lesson from the discussion of null-hypothesis testing in other domains it is that the size of the effect should be reported in some way – usually along with the p-value results. Effect size tells us how big the observed differences were while p-values indicate how much confidence we should attribute to the basic result. There are two ways of presenting effect size: graphing the results, which to a large extent is normal practice in (mobile-)HCI but could still be standardised somewhat, and using measures of effect size, which are rare in mobile-HCI papers (but also probably less informative than graphs). See (Denis, 2003) for a discussion of this point and an extensive and balanced review of alternatives to null hypothesis testing.

Graphing results is standard procedure in HCI papers and typically shows much more information than straight p-value results (Loftus, 1993) (Wilkinson, 1999): good graphs show trends over time/practice and the size of the difference as well as the range of results. This is good practice and a subject in which the HCI community deserves praise over other domains. However, we are not perfect and the display of error bars on graphs is not as consistent as it should be: sometimes they are not present, sometimes they report a standard deviation, sometimes a standard error or 95% confidence interval, and sometimes the absolute range. By graphing suitable confidence intervals and stating the confidence level of the estimate, alongside point estimates of the population parameter(s), we illustrate visually both the differences between groups and the reliability of the estimates made (i.e. the experimental mean for system A is x and we are 95% confident that the true mean lies between x-d_1 and x+d_2). As well as reflecting the range of values, confidence intervals also provide an indication of the sample size as larger samples will tend to result in tighter intervals. Figure 1 shows three graphs of the same data: an artificial experiment comparing two systems over six experimental tasks. The first graph shows the simplest plot of only means, this plot gives the impression that one system is better at the beginning but that performance swaps over around task 4. The second plot adds error bars showing the 95% confidence range and shows clearly that the data overlaps massively at the beginning and is only likely to be conclusive at the right-hand side of the graph. Finally, the third plot replaces the error bars with scatter bars of the actual data – highlighting the inconclusive nature of tasks 1 through 5 and that even in task 6 we do not have perfect separation between the two systems.

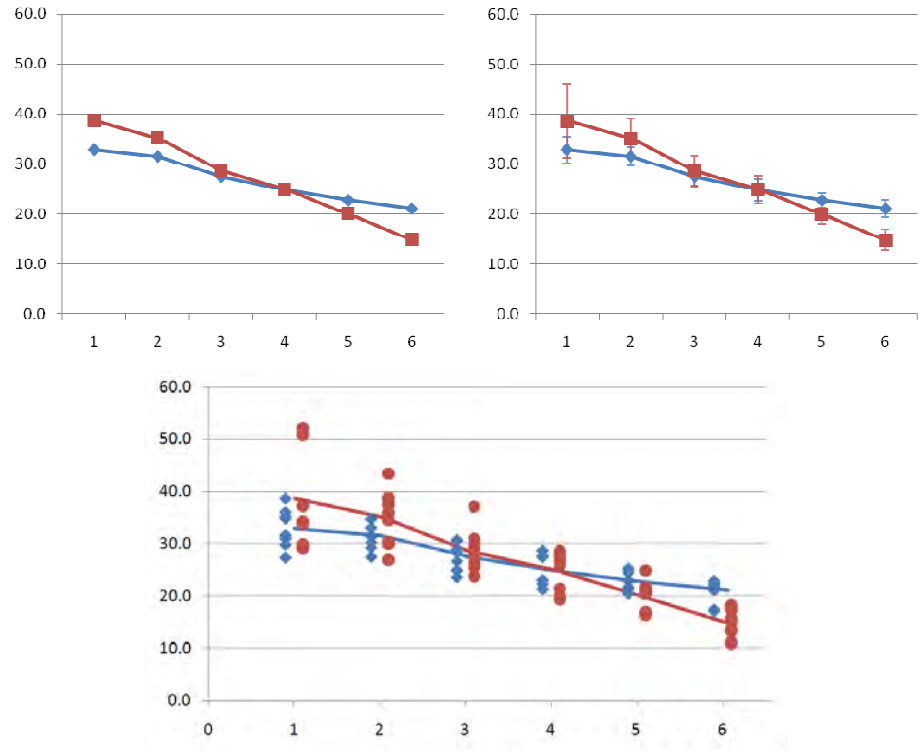

Figure 1: Three plots of the same data – six tasks on two systems (means only; means plus 95% confidence interval error bars; and means plus scatter plots)

Alongside the display of confidence intervals it would be desirable to report the effect size: a scaled estimate of the difference between groups. Reporting the effect size allows for the practical importance of a result to be determined which cannot be conveyed through statistical significance alone. Encouraging both confidence intervals and effect sizes to be reported enables the reader / reviewer to evaluate the results of an experiment more effectively than a p-value alone, regardless of whether statistical significance was achieved. Also, by reporting a standardised effect size opens up the potential for future meta-analysis of related studies through the use of pooled samples. Another criticism of HCI is the lack of replication: other domains base their science on publishing results that others then replicate to further understand and to confirm (or refute) the original. Ioannidis motives his criticism by highlighting the "high rate of nonreplication (lack of confirmation) of research discoveries is a consequence of the convenience, yet, ill-founded strategy of claiming conclusive research findings solely on the basis fo a single study assessed by formal statistical significance..." (Ioannidis, 2005). In a domain that does not

attempt, nor support publication of, replicated results – we don't know how bad our non-replication problem is.

CONCLUSION

This short paper has aimed to raise awareness in the mobile-HCI community, and perhaps the wider HCI community, of the fundamental concerns that other disciplines have raised over null-hypothesis testing. While we do not appear to be as bad at "statistical sinning" as other domains, we cannot afford to be complacent – particularly as conferences and journals tend to become monotonically harder to publish in, poor understanding and treatment of null-hypothesis testing may seriously affect the types of papers and results that make it through to publication. We therefore encourage anyone involved in writing up or reviewing experimental work in mobile-HCI to read the papers in our, deliberately short, bibliography.

REFERENCES[1]

Cairns, P. (2007). HCI... not as it should be: Inferential Statistics in HCI Research. *Proceedings of HCI 2007 (People and Computers XXI)*. Lancaster, UK.

Cliner, J. A., Leech, N. L., & Morgan, G. A. (2002). Problems with null hypothesis significance testing (NHST): What do the textbooks say? *The Journal of Experimental Education , 71* (1), 83-92.

Cohen, J. (1994). The Earth is Round (p<.05). *American Psychologist , 49* (12), 997-1003.

Demsar, J. (2008). On the Appropriateness of Statistical Tests in Machine Learning. *Proceedings of The 3rd workshop on Evaluation Methods for Machine Learning at ICML 2008*. Helsinki, Finland.

Denis, D. (2003). Alternatives to Null Hypothesis Significance Testing. *Theory & Science , 4* (1).

Drummond, C. (2008). Finding a Balance between Anarchy and Orthodoxy. *Proceedings of The 3rd workshop on Evaluation Methods for Machine Learning at ICML 2008*. Helsinki, Finland.

Gill, J. (1999). The insignificance of null hypothesis significance testing. *Political Research Quarterly , 52* (3), 647-674.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Med , 2* (8).

[1] All papers are available on-line by searching for author's surname and title.

Loftus, G. (1993). A picture is worth a thousand p-values: On the irrelevance of hypothesis testing in the computer age. *Behavior Research Methods, Instrumentation and Computers , 25*, 250-256.

Meehl, P. (1990). Why Summaries of Research on Psychological Theories Are Often Uninterpretable. *Psychological Reports , 66*, 195-244.

Wilkinson, L. (1999). Statistical Methods in Psychology Journals: Guildelines and Explanations. *American Psychologist , 54* (8), 594-604.

ABOUT THE AUTHORS

Dr Mark Dunlop is a senior lecturer in Computer Science at The University of Strathclyde in Glasgow, Scotland. His research focuses on improving the usability of mobile systems through new interaction techniques - in particular, he has investigated text entry methods and visualisation methods for mobiles (usually using null-hypothesis based analysis!). He was involved in the first MobileHCI conference back in 1998 and has been heavily involved in the series since, jointly running conferences in 1999, 2001 and 2004 as well as chairing the steering committee. He is on the editorial board for the International Journal of Mobile Human Computer Interaction, Advances in Human-Computer Interaction, and Personal & Ubiquitous Computing.

Dr Mark Baillie is a research fellow in the Computer and Information Sciences department at the University of Strathclyde, Glasgow, Scotland. His research spans a range of research areas such as applied statistical modelling, epidemiology, knowledge discovery and information access, with particular focus on aspects such as data-mining of large databases, content-based indexing, and the integration of user context in information retrieval and management systems.