TOWARDS 2000: REFORM IN RESEARCH PRACTICE AND STATISTICAL EDUCATION

Neil Thomason, Department of History and Philosophy of Science,
The University of Melbourne, Australia
Jan Les, School of Psychological Science, La Trobe University, Australia

In psychology, there are serious flaws in practices for publishing statistical results and in textbook presentation of concepts, reflecting poor statistical understanding. Despite repeated calls from many, including disciplinary leaders, for reform of research practice and standards, only recently has there been substantial movement for change. The failure to reform research practices comes largely from the failure of much statistics education to emphasize conceptual understanding. With computers, students can for the first time explore vividly such key concepts as the Central Limit Theorem and statistical power. The central aim of the StatPlay project is to provide tools and activities to help learners understand statistics' fundamental concepts, and thereby, in the long run, help reform research practice.

CURRENT RESEARCH PRACTICE AND THE NEED FOR REFORM

The dominant analysis strategy in psychology is null hypothesis significance testing (NHST; Hubbard, Parsa and Luthy, 1997), including analysis of variance, *t*-tests and many other procedures. Yet for decades well-credentialed psychologists and statisticians have published powerful critiques of NHST, which have identified:

1. major problems with the theoretical basis of NHST practices,
2. widespread misconceptions held by researchers as to what NHST means and can tell us, and
3. damaging effects on research decision-making, and research progress arising from use of NHST.

The debate about NHST and the need for reform of analysis practices in psychology has now become mainstream, with for example publication of the powerful papers in *Psychological Science* (1997, *8*(1)). Reform has also received strong advocacy in Australia (Hammond, 1996).

We give here some brief background for the claims of the above paragraph:

1. The logical coherence of current NHST practices is widely challenged:
   'What has become institutionalized as *inferential statistics* in psychology is not Fisherian statistics. It is an incoherent mishmash of some of Fisher's ideas on one hand, and some of the ideas of Neyman and E. S. Pearson on the other. ... it is

time to go beyond this institutionalized illusion. We must write new textbooks and change editorial policies.' (Gigerenzer, 1993, p. 314)

2. Misconceptions of statistics are widespread, persistent, and strongly resistant to conventional education (Thomason, Cumming and Zangari, 1994):

   "... a whole collection of misguided and influential conceptions are simultaneously held by what appears to be the vast majority of academic psychologists." (Oakes, 1986, p. 82)

Crucial NHST misconceptions include confusion between statistical and substantive significance, the misinterpretation of the *P* value (exact significance level) as the probability the null hypothesis is true, and the unjustified acceptance of a null hypothesis that is not rejected. A central problem is 'The Law of Small Numbers' (Tversky and Kahneman, 1971): the severe underestimation of sampling variability that underlies many misconceptions about power, confidence intervals and the replicability of results.

3. NHST has caused great damage to the whole research enterprise:

   "... use of significance test results leads to terrible errors in review studies. Most review studies falsely conclude that further research is needed to resolve the 'conflicting results' in the literature. These errors ... can only be eliminated if errors in the interpretation of significance tests can be eliminated. ... maybe it is time to abandon the significance test." (Hunter and Schmidt, 1990, p. 31)

The few published defenses of NHST (e.g. Chow, 1996; Cortina and Dunlap, 1997) concede much of the reform agenda, including the need for better understanding of NHST and more complete reporting of statistical information. The reforms will require major changes to editorial policies, researchers' thinking, and statistics education.

REFORMS IN DATA ANALYSIS AND RESEARCH DECISION MAKING PRACTICES

Despite these criticisms, until recently researchers and statistics teachers have by-and-large ignored the evidence showing the poor statistical intuitions and understanding in lay-people and the statistically educated alike; in particular, psychology research practice appears little informed by these findings. Happily, the indications are that

psychology's slow response to these issues has begun to change. The American Psychological Association's Task Force on Statistical Inference, formed in 1996, is considering the role of NHST and modification of current practices in quantitative treatment of data. Some of the draft recommendations of the APA Task Force (APA, 1996) are that 'more extensive descriptions of the data be provided to reviewers and readers. This should include means, standard deviations, sample sizes, five-point summaries, box-and-whisker plots, other graphics, and descriptions related to missing data as appropriate.' They also recommend that 'size of effect... and their confidence intervals should be provided routinely...'. Although '...the task force does not support ... banning the use of null hypothesis significance testing...', the effect of its recommendations would be that NHST would have a greatly diminished role. Our recent contact with the Task Force confirms that the recommendations are current and that extensive across-discipline consultation is occurring so that reform can extend broadly to other disciplines, including sociology and anthropology.

A PART OF THE EXPLANATION FOR THE PRESENT STATE OF AFFAIRS

Why have statistics practices in psychology remained so resistant to change despite the powerful indictment of them and the strong evidence that these practices were seriously retarding progress in psychological research?

Whatever the answer to this fascinating question proves to be, one key element will surely be that many highly influential research psychologists and statistics teachers simply did not (do not) understand the basic statistics that they so fervently believed (believe) in and taught to generation after generation of students. As we indicated above, the evidence for this is overwhelming. This brings us to the crux of this talk: the role of future statistics education in the modification of deeply flawed statistical practices in psychology. Most present statistics textbooks do little to undermine either teachers' or students' flawed intuitions.

STATISTICS REFORM, STATISTICS EDUCATION AND STATPLAY

The proposed reforms of statistical practice, without dramatic reforms of statistical education, will only partly succeed. Without conceptual understanding, the faulty intuitions and formula-driven mis-applications of statistics will remain. Psychologists will replace one set of arbitrary and ill-understood procedures with another.

They will undoubtedly calculate an experiment's statistical power if reporting that statistic is required to get their articles published. But, it does not follow that they will understand what they have published. Nor does it follow that they will understand the implications of low statistical power when they read it in a journal article. Without increased conceptual understanding, the reforms won't stop poorly designed experiments or seriously faulty readings of articles.

To intelligently evaluate one's experimental results or read the research literature, a psychologist must intuitively understand many things: What can be said about the likelihood of replication? What does a significant result mean? What does a non-significant result mean? What is the relationship between confidence intervals and replications?

A reformed statistics education should explicitly address common misconceptions, raise issues of interpretation and debate about statistical methods, and reflect recommendations for the use of a range of descriptive, inferential and modeling tools. Implementation of reform presents a considerable challenge to teachers of statistics. It requires new teaching tools and approaches, as well as moves to more inclusive curricula. Both students and teachers need to better understand the central concepts of statistics, many of which have been identified as problematic. These include many aspects relating to sampling and probability as they apply to interpretation of research results.

Until recently, a major (perhaps the major) difficulty in teaching the central concepts of inferential statistics is their essentially stochastic nature. Textbooks are limited to presenting one image at a time. But, these concepts deal with patterns that emerge when the "same" process is repeated many many times. They depend on concepts such as the distribution of the results of indefinitely many random samples.

The personal computer has freed statistics education from the limitation of textbooks. They can easily present the results of many experiments and can let the student explore and experiment with these results. Students can test the applicability of the Central Limit Theorem to their own curves. They can explore the relationship between sample size and standard error, which underlies the calculation of, *inter alia*, confidence intervals. They can discover for themselves what statistical power is and why it matters. We would like to report our personal experience here. Although in some abstract way, we knew what statistical power is and why it matters, it was only in seeing it

in action on the computer screen that we really came to have a deep intuitive understanding of what power means and its crucial importance.

We don't have enough pages to explain StatPlay. There are several other papers at this conference about StatPlay and its effectiveness (Cumming and Thomason, 1998; Finch and Cumming, 1998). In summary, StatPlay is advanced multimedia educational software that is designed to enable students to understand the basic statistical concepts. Guided by the considerations above, StatPlay provides tools and activities to help learners master fundamental statistical concepts. StatPlay does this by providing vivid graphical presentations of these concepts, challenging the faulty intuitions that nature seems to have so deeply implanted in our psyches. Our preliminary results indicate that it works remarkably well. We deeply hope that it will contribute to important reform of statistical education but also to the quality of psychological research practice.

ACKNOWLEDGEMENTS

StatPlay has been supported by the Australian Research Council (ARC), the Committee for the Advancement of University Teaching (CAUT), the Committee for University Teaching and Staff Development (CUTSD) and The University of Melbourne.

REFERENCES

American Psychological Association, Board of Scientific Affairs Task Force on Statistical Inference (1996). Initial Report (Draft) [On-line]. www.apa.org/science/tfsi.html

Chow, S. L. (1996). *Statistical Inference: Rationale, Validity and Utility*, London: Sage.

Cortina, J. M., and Dunlap, W. P. (1997). On the logic and purpose of significance testing, *Psychological Methods, 2,* 161-172.

Cumming, G. and Thomason, N. (1998). Statplay: Multimedia for statistical understanding. Proceedings of ICOTS-5, The Fifth International Conference on Teaching Statistics, Singapore, June.

Finch, S., and Cumming, G. (1998). Assessing conceptual change in learning statistics. Proceedings of ICOTS-5, The Fifth International Conference on Teaching Statistics, Singapore, June.

Gigerenzer, G. (1993). The superego, the ego and the id in statistical reasoning. In Keren, G. and Lewis, C. (Eds.) *A handbook for data analysis in the behavioural sciences: Methodological issues*, Hillsdale, New Jersey: Lawrence Erlbaum.

Hammond, G. (1996). The objections to Null Hypothesis Testing as a means of analysing psychological data, *Australian Journal of Psychology, 48*, 104-106.

Hubbard, R., Parsa, R. A., and Luthy, M. R. (1997). The spread of statistical significance testing in psychology, *Theory and Psychology, 7(4),* 545-554.

Hunter, J. E., and F. L. Schmidt (1990). *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings*, Newbury Park, CA: Sage.

Light, R. J., Singer, J. D., and Willett, J. B. (1994). The visual presentation and interpretation of meta-analyses. In Cooper, H. and Hedges, L. V. (Eds.), *The handbook of research synthesis*, New York: Russell Sage Foundation.

Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioural sciences*, Chichester: Wiley.

Thomason, N., Cumming, G., and Zangari, M. (1994). Understanding central concepts of statistics and experimental design in the social sciences. In Beattie, K., McNaught, C. and Wills, S. (Eds), *Interactive multimedia in University education: Designing for change in teaching and learning*, Amsterdam: North-Holland.

Tversky, A., and Kahneman, D. (1971). Belief in the law of small numbers, *Psychological Bulletin*, *76*, 105-110.