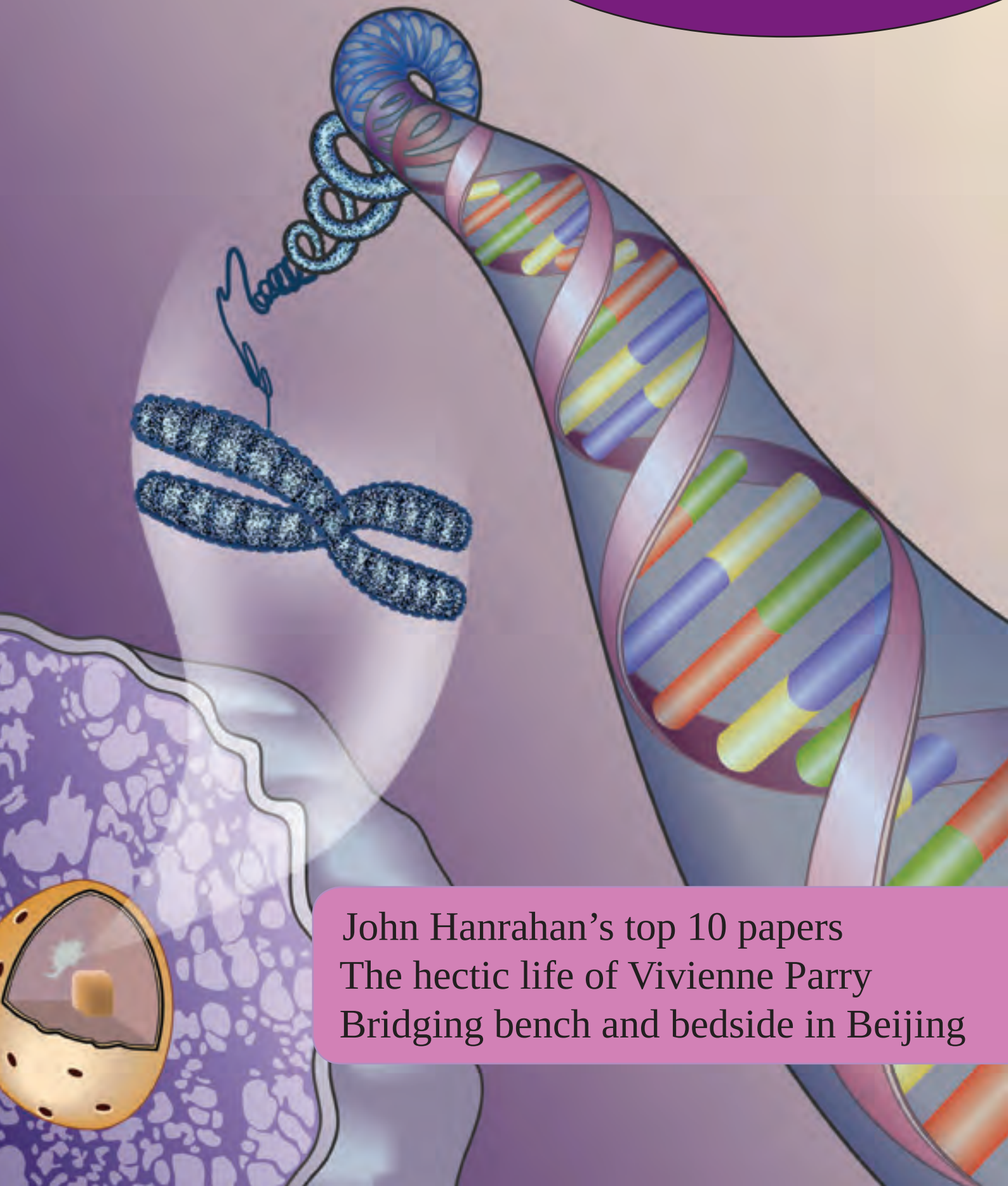
PHYSIOLOGY NEWS

spring 2009 I number 74



John Hanrahan's top 10 papers

The hectic life of Vivienne Parry

Bridging bench and bedside in Beijing

The Society's dog. 'Rudolf Magnus gave me to Charles Sherrington, who gave me to Henry Dale, who gave me to The Physiological Society in October 1942'

Published quarterly by The Physiological Society

Contributions and queries
Senior Publications Executive
Linda Rimmer
The Physiological Society Publications Office
PO Box 502, Cambridge CB1 0AL, UK
Tel: +44 (0)1223 400180
Fax: +44 (0)1223 246858
Email: lrimmer@physoc.org
Website: http://www.physoc.org

ISSN 1476-7996

The Physiological Society is registered in England as a company limited by guarantee: No 323575. Registered office: PO Box 11319, London WC1X 8WQ.
Registered Charity: No 211585.

Printed by The Lavenham Press Ltd

Advancing the science of life

Cover image from *The Journal of Physiology* Symposium *Physiological regulation linked with physical activity and health*

PHYSIOLOGY NEWS

molecular/biochemical events that lead to excessive hepatic fat accumulation, and thus future time course studies are warranted.

R Scott Rector[1]
Jamal A Ibdah[1,2,3]

[1]Division of Gastroenterology and Hepatology, [2]Harry S. Truman Memorial Veterans Medical Center, and [3]Department of Medical Pharmacology and Physiology, University of Missouri,Columbia, MO 65212, USA.

References

Hannukainen JC, Nuutila P, Borra R, Kaprio J, Kujala UM, Janatuinen T, Heinonen OJ, Kapanen J, Viljanen T, Haaparanta M, Rönnemaa T, Parkkola R, Knuuti J & Kalliokoski KK (2007). Increased physical activity decreases hepatic free fatty acid uptake: a study in human monozygotic twins. *J Physiol* **578**, 347–358.

Mokdad AH, Marks JS, Stroup DF & Gerberding JL (2004). Actual causes of death in the United States, 2000. *JAMA* **291**, 1238–1245.

Perseghin G, Lattuada G, De Cobelli F, Ragogna F, Ntali G, Esposito A, Belloni E, Canu T, Terruzzi I, Scifo P, Del Maschio A & Luzi L (2007). Habitual physical activity is associated with intrahepatic fat content in humans. *Diabetes Care* **30**, 683–688.

Rector RS, Thyfault JP, Laye MJ, Morris RT, Borengasser SJ, Uptergrove GM, Chakravarthy MV, Booth FW & Ibdah JA (2008*a*). Cessation of daily exercise dramatically alters precursors of hepatic steatosis in Otsuka Long-Evans Tokushima Fatty (OLETF) rats. *J Physiol* **586**, 4241–4249.

Rector RS, Thyfault JP, Morris RT, Laye MJ, Borengasser SJ, Booth FW & Ibdah JA (2008*b*). Daily exercise increases hepatic fatty acid oxidation and prevents steatosis in Otsuka Long-Evans Tokushima Fatty rats. *Am J Physiol Gastrointest Liver Physiol* **294**, G619–G626.

Rector RS, Thyfault JP, Wei Y & Ibdah JA (2008*c*). Non-alcoholic fatty liver disease and the metabolic syndrome: An update. *World J Gastroenterol* **14**, 185–192.

Shima K, Shi K, Sano T, Iwami T, Mizuno A & Noma Y (1993). Is exercise training effective in preventing diabetes mellitus in the Otsuka-Long-Evans-Tokushima fatty rat, a model of spontaneous non-insulin-dependent diabetes mellitus? *Metabolism* **42**, 971–977.

Shojaee-Moradie F, Baynes KC, Pentecost C, Bell JD, Thomas EL, Jackson NC, Stolinski M, Whyte M, Lovell D, Bowes SB, Gibney J, Jones RH & Umpleby AM (2007). Exercise training reduces fatty acid availability and improves the insulin sensitivity of glucose metabolism. *Diabetologia* **50**, 404–413.

Statistical methodology and reporting – the case for confidence intervals

Peter Cahusac.

Jack just finished collecting some more data on the effects of rhubarb extract #654 (RE654) on the membrane potential of spinal neurones. He passed the new data over to Olivia who had already started up the statistics package on her computer. So far the results hadn't quite reached statistical significance $P < 0.05$; adding the new data would hopefully change that – fingers crossed. Expectantly they both awaited the output of the t test. Yes! Significant! The P value was 0.033, that's good enough. They quickly entered this last bit of data analysis into their now overdue manuscript: '...RE654 also depolarised the resting membrane potential, from –64.7±2.0 to –57.6±1.9 mV ($P < 0.05$, $n = 20$)', and clicked the submit button to the online journal. Right, job done, off to the pub.

Despite the caricature, most people will identify with aspects of the above scenario. Some will also note statistical inadequacies. Recently I spent a couple of days reviewing the statistical methodology and reporting used by *The Journal of Physiology* research papers that appeared in the last four issues of 2008. Among other things I looked at what statistical tests were used, how results were reported, paying particular attention to whether confidence intervals were used. I also noted whether the standard deviation (S.D.) or standard error (S.E.) was preferred, and whether which was used was clearly stated.

As expected, the *t* test was the most popular test, being used in no fewer than 44 of the 60 papers. Last year, readers were treated to a fascinating history of this test (Brown, 2008). In doing a *t* test a statistical package typically churns out the means, each given with their standard error of the mean (S.E.M.) and S.D. Also given are the *t* value, sample size or degrees of freedom, and often the confidence interval. In reporting results, most research papers in my sample chose to give their calculated statistic (e.g. mean) ± some measure of variability. The large majority (46 of 58 papers which gave such measures) preferred to use the S.E., with only seven papers using exclusively the S.D. In most papers, which one was used was clearly stated in the Methods section. However, in six papers it was unclear or one had to look in figure or table legends to find which was used. In a further five papers (9%) there was no mention of what measure of variability was used. One of these was an Open Access paper which gave 57 summary statistics using ± yet nowhere indicated what measure of variability (S.E. or S.D.) was used. My survey results for *The Journal of Physiology* were similar to those for another study which examined 88 research papers in the medical journal *Infection and Immunity* (Olsen, 2003). There, 12 (14%) failed to identify the measure of variability. Why the sloppy reporting? Part of the reason may be that the difference between the S.E. and S.D. is not fully understood by researchers, and that these terms may be used interchangeably (Altman & Bland, 2005). Both S.D. and S.E. are measures of variability, and are related. The S.D. is an estimate of the variability of data points within a population, based upon a sample drawn from that population. In contrast, the S.E. is an estimate of the variability of a sample statistic (such as the mean) obtained by sampling from a population. Hence, the S.E. is also a standard deviation – but of the sampling distribution. The fact that the S.D. and S.E. are both standard deviations (but of different things) and are related, no doubt causes confusion. The persistence of the ± sign in papers is sometimes merely due to it being demanded by journal editors and reviewers. However, many journals, including the *British Medical Journal*, no longer allow the use of the ± sign, and request

		Effect size	
		Large	Small
Statistical significance	Small *P* value < 0.05	No problem	Mistaking statistical significance for scientific importance
	Large *P* value > 0.05	Failure to detect a scientifically important effect	No problem

Table 1. Using *P* values can be misleading when there is a small effect size and small *P* value, and when there is a large effect size and a large *P* value. Adapted from Rosenthal *et al*. (2000).

authors to clearly state whether the S.E. or S.D. is quoted.

In my survey I looked at how *P* values were reported. Giving the actual *P* values obtained by statistical tests is useful in that it indicates how significant the result is. Giving $P = 0.048$ or $P = 0.052$ indicates marginally significant and marginally non-significant results, respectively. Giving $P = 0.002$ and $P = 0.65$ indicates highly statistically and clearly non-statistically significant results. Usually extremely small *P* values can be expressed as $P < 0.001$, even though a computer output gives 0.000. Communicating information via the *P* value was something that the statistician and biologist R. A. Fisher encouraged as "...in doing this we have a genuine measure of the confidence with which any particular opinion may be held, in view of our particular data" (Fisher, 1955)[1]. If values are reported merely qualitatively as $P < 0.05$ or $P > 0.05$, then quantitative information from the *P* value is lost. Unfortunately, such reporting is common practice in *The Journal of Physiology*.

I was keen to determine how many papers reported 95% confidence intervals. Only 3 of the relevant 58 papers did so. Incidentally, one paper (not one of the 3) stated in its Methods section: "Significance was defined by a *P*-value less than 0.05 (95% confidence)." That "95% confidence" was not what I was looking for, and further implies a misunderstanding of what a *P* value represents (but more on that in a moment). So what's the fuss? Readers will justifiably wonder how a confidence interval can materially add to a reported mean, its standard error, sample size and *P* value. Where to begin? Well, the procedure of null hypothesis significance testing (NHST), which is how we normally decide whether an intervention has had an effect or not, is undeniably useful and prevents us from over-interpreting results. However, it has attracted a steady stream of criticism over the years (Cohen, 1994; Sterne & Smith, 2001), including comments from some very distinguished quarters (Cox, 1982). A couple of contributions to this literature make entertaining reading, particularly (Salsburg, 1985) 'The Religion of Statistics as Practiced in Medical Journals'. Another (Gigerenzer, 1993) (in jest) reduces statistical testing to a Freudian ritual. The *P* value that we obtain in a statistical test is the probability of obtaining data as extreme, or more extreme as our sample, assuming a true null hypothesis (typically this is that there is no effect). Although often misunderstood, even by leading textbooks (Bland & Altman, 1988), this *P* value is not the probability of the null hypothesis being true. Nor can we claim, should we for example obtain a statistically significant difference in means ($P < 0.05$), that there is a 95% chance that there is a difference between these means (or the '95% confidence' stated above). Such claims commit the so-called inverse probability error (Fisher, 1947; Cohen, 1994), attributing a Bayesian-like probability to whether hypotheses are true or not (NHST only gives us *P*(Data|Hypothesis), while Bayes' theorem gives *P*(Hypothesis|Data)). One difficulty is that the null hypothesis is rarely true anyway (Chew, 1977; Cohen, 1994). If you get enough data then you will almost always obtain a statistically significant result ($P < 0.05$ or $P < 0.01$ etc) – but the size of the effect may be extremely small and inconsequential, of little scientific interest. The statistical significance which we obtain from a NHST is routinely confused with the scientific importance and even the magnitude of the effect (size of the effect or effect size[2]). Often, in the Discussion section of a paper, much is made of the star-studded Results section (figures and tables emblazoned with *, **, ***). Then, sometimes it's difficult to publish without a '$P < 0.05$' appearing somewhere in the manuscript. However, relying on P values alone can be misleading, as can be seen in Table 1. If the size of the effect is so small as to be unimportant scientifically then its association with statistical significance (even $P < 0.001$) does not necessarily mean that the result is scientifically important. For example, not a week goes by without epidemiologists informing the public (exacerbated by media reporting (Blastland & Dilnot, 2008)) that a dietary component is statistically associated with either benefit or harm – typically such studies involve 1000's of participants, and we are rarely properly informed about the size of the effect. Conversely, when there is a large effect which fails to reach statistical significance, this is often reported as unimportant (Altman & Bland, 1995), and yet it may be very clear from a confidence interval that not enough data were collected.

[1]Fisher's paper is also of interest because it contains a polemic against the works of J. Neyman and E.S. Pearson who suggested the use of a fixed significance level of 0.05. Their proposals for Type II errors and confidence intervals were also attacked – but both these ideas are now accepted by mainstream statisticians.

[2]There is a distinction between these terms. The *size of effect* is given in the original units of measure (e.g. mmHg) and may be, for example, a mean difference. The *effect size* is a dimensionless but standardized quantity, examples being Cohen's *d* (the mean difference divided by σ), a correlation, or an odds ratio (Rosenthal *et al*. 2000).

How can one see this from a confidence interval? As an example, consider the effects of four different interventions on the blood pressure of patients with hypertension, where the horizontal axis represents the mean difference from pre-intervention (see Fig. 1). We assume (for the sake of argument) that any intervention that reduces blood pressure by 5 mmHg or more is worthwhile and clinically (or scientifically) important. In Fig. 1, 95% confidence intervals are plotted for each intervention. In each case, such an interval calculated from sample data will 95% of the time (in the long run) contain the population mean difference value for that intervention. Values bracketed within the interval are consistent with the sample mean difference, while those outside are not ($P < 0.05$). For interventions A and B, the midpoint (sample mean difference) in each interval is 6 S.E.s away from 0 mmHg, and they therefore have identical *t* and *P* values, and are clearly statistically significant (since these confidence intervals do not contain 0 mmHg). Intervention A, though statistically significant, is of little clinical (scientific) interest as its confidence interval lies close to 0, and does not contain or is not less than –5 mmHg. Intervention B is of much more interest as the confidence interval spans a range of values much lower than –5 mmHg (approx. –18 to –9 mmHg). For interventions C and D, the midpoint of each interval is situated only 1 S.E. away from 0, which gives them identical *t* and *P* values. Moreover their intervals span 0, so neither is statistically significant (i.e. each is $P > 0.05$). The interval in C does not include clinically important values (below –5 mmHg), and therefore should be of little further interest. The interval is narrow enough to indicate that we have collected enough data. In contrast, intervention D has a very wide interval that almost reaches to –14 mmHg. So although intervention D is not statistically significant (just as C) it is of much more interest, and indicates that we have not collected enough data. In the circumstances it would be premature to exclude D as a useful intervention. This result could represent that indicated at the bottom left cell in Table 1 (large *P* value and large effect size).

Typically, the 95% confidence interval is reported but others, including 90% and 99% intervals, are also used. Confidence intervals immediately indicate (i) statistical significance (Fig. 1*A* and *B* statistically significant), (ii) the size of the effect (B > D > A > C), (iii) the sensitivity of the study (A, B and C have enough power, D does not), (iv) the precision of the statistic (A and C have greater precision, their intervals are narrower, than B and D respectively). Providing the *P* value only gives us (i). Providing the mean ± S.E. gives us very limited information about (ii)–(iv). Although very rough 95% confidence intervals for mean differences may be mentally calculated quite quickly by the average reader of *The Journal of Physiology*, the same confidence interval calculations will probably not be so easy for other (e.g. non-parametric) statistics. It should be noted that the ± S.E. values given in the opening paragraph '...–64.7±2.0 to −57.6±1.9 mV...' cannot be used to calculate the confidence interval for the difference in means – which is what we are really interested in and to which the *P* value refers (this style of reporting individual means ± S.E. is common in *The Journal of Physiology* papers).

Let us return again to our opening scenario featuring Jack and Olivia, where Jack has just collected some more data. Strictly speaking, if they have already tested their data for an effect using a significance level of 0.05 and it fails to reach significance then that's it – however much more data they collect it is not possible to perform another significance test and claim a statistically significant effect (even if subsequently they obtain $P < 0.000001$). This is an issue about multiple testing and stopping rules. If Jack and Olivia had decided before collecting any data that they would periodically test for statistical significance, that would be fine, but they would need to adjust their significance level accordingly, for example using Bonferroni. So, if they had actually decided to test twice after collecting sets of data (as they actually did), then they would need to use 0.05/2 = 0.025 as their significance level, which with their *P* value of 0.033 would mean that they still could not claim a statistically significant result. It is a fact that, even if the null hypothesis is completely true, you are guaranteed to obtain a statistically significant result, at whatever level you choose,

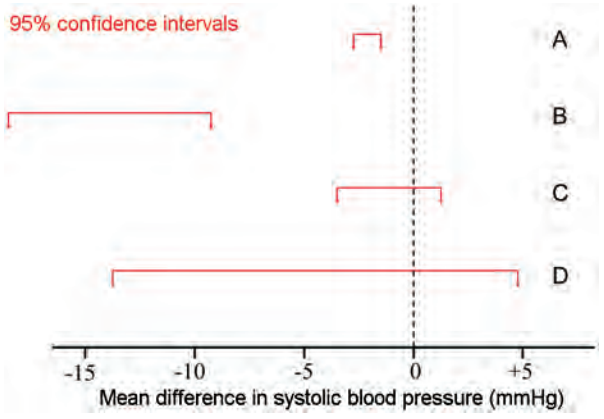


Figure 1. The results of 4 different interventions A – D on the blood pressure of patients with hypertension

The results from each intervention are plotted as 95% confidence intervals along the horizontal axis, in mmHg. For interventions A and B the midpoint (sample mean difference) in each interval is 6 S.E.s away from the 0 mmHg difference, and therefore have identical *t* and *P* values, and clearly statistically significant (actually $P < 0.001$). For interventions C and D, the midpoint of each interval is situated only 1 S.E. away from 0 mmHg, which gives them identical t and *P* values. Adapted from Reichardt & Gollob (1997).

if you continue to add more data to existing data and repeatedly test for significance. Problems like this have prompted some to use the Likelihood approach (Edwards, 1992; Royall, 2004).

Actually, confidence intervals can help us out here too. Continuing with the example illustrated in Fig. 1, and values given in mmHg. If we have collected some data and the 95% confidence interval for the difference in means includes both 0 and –5, then we know immediately that the study was not sensitive enough (i.e. power was too low, as seen in D). The S.E. is too large. Typically this can be reduced by increasing the sample size, that shrinks the confidence interval, until it is acceptably small. How small? Well, a nice stopping rule is suggested by Armitage *et al.* (2002) (p. 615). We should continue collecting data until our 95% confidence interval is just less than 5 units wide. In this way, if the interval includes 0 then it will exclude –5, and if it includes –5 then it will exclude 0. (If it just happens to fall in between 0 and –5 then we would claim a statistically significant effect but it would not be scientifically important.) Note that we are using the width of the confidence interval, not the smallness of the *P* value, to decide whether we have enough data. This procedure can be elaborated. For example, if we were not interested in interventions that reduce blood pressure by up to 4 mmHg, but were interested in interventions that reduce pressure by at least 10 mmHg, then we could continue collecting data until our 95% confidence interval was just less than 6 units wide. Few researchers appear to be aware of this useful stopping rule that allows us to collect data until the required precision is obtained. It forces us to explicitly recognise a size of effect which we believe to be scientifically important, and has the surprising advantage that it does not fall foul of the loss of power due to multiple testing protocols (e.g. Bonferroni (Perneger, 1998)).

For some years now statisticians have been placing greater emphasis on reporting confidence intervals rather than just *P* values (Altman *et al*. 2000). Many medical journals (e.g. the *British Medical Journal*) insist on them – where appropriate. Only limited use is made of them in physiology, and the Instructions for Authors for *The Journal of Physiology* makes no mention of them. In a review of 370 papers published in journals under the auspices of the American Physiological Society in 1996, only two papers reported confidence intervals (Curran-Everett *et al*. 1998). This review, which appeared in the *Journal of Applied Physiology*, highlighted inadequacies of statistical reporting and made a strong case for using confidence intervals, rather than just null hypothesis testing. Ten years on, the December 2008 issue of the same journal finds little improvement, with just 4/35 papers reporting confidence intervals (a proportion similar to my 3/58 for *The Journal of Physiology*).

Why aren't confidence intervals used more widely? Perhaps they are considered superfluous to a results summary (as I used to think). Maybe they are just not understood, and there is a failure to realise what information they carry. Sometimes they can be embarrassingly wide! Whatever the reason, I hope that they will be better appreciated and appear more often in future issues of *The Journal of Physiology* and related journals.

Peter Cahusac

University of Stirling, Stirling, Scotland, UK.

References

Altman DG & Bland JM (1995). Absence of evidence is not evidence of absence. *BMJ* **311**, 485.

Altman DG & Bland JM (2005). Standard deviations and standard errors. *BMJ* **331**, 903.

Altman DG, Machin D, Bryant TN & Gardner MJ (2000). Statistics with Confidence. BMJ Books.

Armitage P, Berry G & Matthews JNS (2002). *Statistical Methods in Medical Research*. WileyBlackwell.

Bland JM & Altman DG (1988). Misleading statistics: errors in textbooks, software and manuals. *Int J Epidemiol* **17**, 245–247.

Blastland M & Dilnot A (2008). *The Tiger That Isn't: Seeing Through a World of Numbers*. Profile Books, London.

Brown A (2008). The strange origins of the Student's t-test. *Physiology News* **71**, 13–16.

Chew V (1977). *Comparisons Among Treatment Means in an Analysis of Variance*. Agricultural Research Service, Washington DC.

Cohen J (1994). The earth is round (p < .05). *American Psychologist* **49**, 997–1003.

Cox DR (1982). Statistical significance tests. *Br J Clin Pharmacol* **14**, 325–331.

Curran-Everett D, Taylor S & Kafadar K (1998). Fundamental concepts in statistics: elucidation and illustration. *J Appl Physiol* **85**, 775–786.

Edwards AWF (1992). *Likelihood*. John Hopkins University Press, Baltimore.

Fisher R (1955). Statistical methods and scientific induction. Journal of the Royal Statistical Society, Series B, Methodological **17**, 69–78.

Fisher RA (1947). *The Design of Experiments*. Oliver & Boyd, Edinburgh.

Gigerenzer G (1993). The Superego, the ego, and the Id in statistical reasoning. In *A Handbook for Data Analysis in the Behavioral Sciences*: *Methodological Issues*, ed. Keren G & Lewis C, pp. 311–339. Lawrence Erlbaum Associates, Hillsdale, New Jersey.

Olsen CH (2003). Review of the use of statistics in infection and immunity. *Infect Immun* **71**, 6689–6692.

Perneger TV (1998). What's wrong with Bonferroni adjustments. *BMJ* **316**, 1236–1238.

Reichardt CS & Gollob HF (1997). When confidence intervals should be used instead of statistical tests, and vice versa. In *What If There Were No Significance Tests*? ed. Harlow LL, Mulaik SA & Steiger JH, pp. 259–284. Lawrence Erlbaum Associates, London.

Rosenthal R, Rosnow RL & Rubin DB (2000). *Contrasts and Effect Sizes in Behavioral Research: a Correlational Approach*. Cambridge University Press, Cambridge.

Royall R (2004). The likelihood paradigm for statistical evidence. In *The Nature of Scientific Evidence*. ed. Taper ML & Lele SR, pp. 119–152. University of Chicago, Chicago.

Salsburg DS (1985). The religion of statistics as practiced in medical journals. *Am Stat* **39**, 220–223.

Sterne JAC & Smith GD (2001). Sifting the evidence - what's wrong with significance tests? *BMJ* **322**, 226–231.