STATISTICS IN MEDICINE
Statist. Med. 2009; **28**:205–220
Published online 7 November 2008 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.3471

Inference by eye: Reading the overlap of independent confidence intervals‡

Geoff Cumming*, †

School of Psychological Science, La Trobe University, Melbourne, Victoria 3086, Australia

SUMMARY

When 95 per cent confidence intervals (CIs) on independent means do not overlap, the two-tailed p-value is less than 0.05 and there is a statistically significant difference between the means. However, p for non-overlapping 95 per cent CIs is actually considerably smaller than 0.05: If the two CIs just touch, p is about 0.01, and the intervals can overlap by as much as about half the length of one CI arm before p becomes as large as 0.05. Keeping in mind this rule—that overlap of half the length of one arm corresponds approximately to statistical significance at $p = 0.05$—can be helpful for a quick appreciation of figures that display CIs, especially if precise p-values are not reported. The author investigated the robustness of this and similar rules, and found them sufficiently accurate when sample sizes are at least 10, and the two intervals do not differ in width by more than a factor of 2. The author reviewed previous discussions of CI overlap and extended the investigation to p-values other than 0.05 and 0.01. He also studied 95 per cent CIs on two proportions, and on two Pearson correlations, and found similar rules apply to overlap of these asymmetric CIs, for a very broad range of cases. Wider use of figures with 95 per cent CIs is desirable, and these rules may assist easy and appropriate understanding of such figures.

KEY WORDS: confidence intervals; overlap; inference by eye; p-values; error bars; statistical graphics

INTRODUCTION

Well-designed figures can in many cases give a quick and valuable overall appreciation of experimental results. Confidence intervals (CIs) provide inferential information and therefore guide the drawing of conclusions from data. Drawing conclusions from figures is *inference by eye*. My

*Correspondence to: Geoff Cumming, School of Psychological Science, La Trobe University, Melbourne, Victoria 3086, Australia.
†E-mail: g.cumming@latrobe.edu.au
‡Overlap and p-values as shown in Figure 1 can be explored using interactive software ESCI ('ess-key'; Exploratory Software for Confidence Intervals), which runs under Microsoft Excel. The component *ESCI Inference by eye* may be downloaded from www.latrobe.edu.au/psy/esci.

Contract/grant sponsor: Australian Research Council

Received 18 June 2008
Accepted 12 September 2008

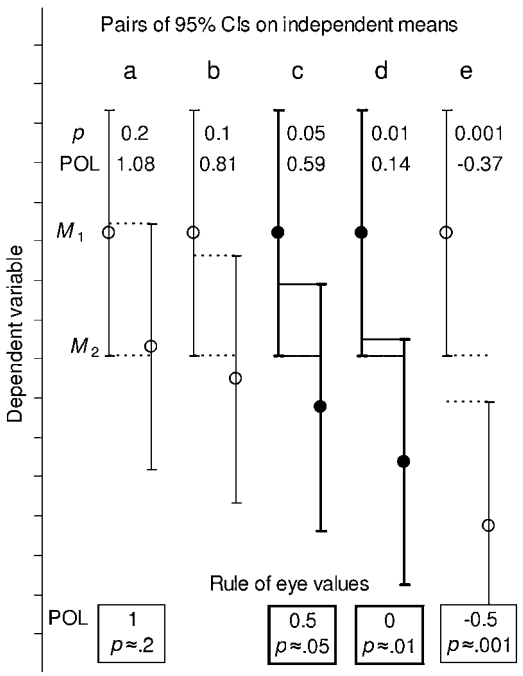

Figure 1. Means M_1 and M_2, and 95 per cent confidence intervals (CIs) for pairs of independent samples. In each pair the two samples have the same size n, and are assumed to come from normal populations having the same variance σ. The CIs are calculated using z; hence, it is assumed σ is known, or n is large. Proportion overlap (POL) is the overlap expressed as a proportion of the length of a single arm of a CI, and POL values are shown near the top. A gap between intervals is signalled by a negative POL value, as for Pair **e**. M_2 is varied to achieve selected values of two-tailed p shown near the top. As overlap progressively decreases from Pair **a** to Pair **e**, the p-value corresponding decreases. The rule of eye p and POL value pairs shown in the lower boxes provide approximate benchmarks for estimating the p-value for any observed amount of overlap of two independent 95 per cent CIs, and show that the rule is slightly conservative for the base case illustrated. For example, Pair **c** shows that POL$=0.59$ gives $p=0.05$, but the rule of eye specifies a POL of approximately 0.5 or less, for $p\leqslant 0.05$.

purpose is to discuss some simple rules to support inference by eye from figures that include CIs or other error bars.

Consider a figure that shows CIs on the means of two independent groups, for example any of the pairs of 95 per cent CIs in Figure 1. It is easy to notice whether the intervals overlap, and the extent of overlap intuitively seems to signal how confident we can be that the two underlying population means differ. Figure 1 illustrates a range of amounts of overlap, and for each pair of means it shows near the top the two-tailed p-value. If intervals overlap considerably, as in Figures 1(a) and (b), the p-value for comparison of the means is large. If they overlap by only a small proportion of a CI arm length, as in Figures 1(c) and (d), the p-value is small and if as in Figure 1(e), they do not overlap at all—there is a gap between the intervals—the p-value is very small. My purpose is to investigate how p-values relate to interval overlap, and to extend previous discussions of rules intended to guide inference by eye with overlapping intervals. I will throughout use two-sided CIs and, correspondingly, two-tailed p-values.

DOI: 10.1002/sim

There has been a widespread belief, especially in medicine [1, 2], that 95 per cent CIs just touching end to end is equivalent to statistical significance, $p<0.05$. However, there have now been a number of explanations [2–5] that prove that this belief is incorrect. The accurate relation is illustrated in Figure 1(c): when $p=0.05$ for the difference between the two means M_1 and M_2, the 95 per cent CIs overlap a little more than half the length of one arm of either CI.

When means are independent, the CIs on those means do indeed contain the information needed to calculate the p-value. However, is the relation between overlap and p sufficiently regular to be useful for inference by eye? The rules I discuss are based on *proportion overlap* (POL), which in Figure 1(c), for example, is the vertical distance between the thin horizontals joining the two 95 per cent CIs, expressed as a proportion of the *margin of error*, where margin of error is the length of one arm of a CI. I use w to refer to the length of one arm of a CI. For symmetric CIs, such as those in Figure 1, the total width of the interval is twice the margin of error, or $2w$. If the two CIs have different widths, and thus different values of w, POL is the proportion of the average of the two margins of error. Looking ahead to later discussion of non-symmetric CIs, it is useful to define POL most generally as the proportion of the average length of the two CI arms that do the overlapping—one arm from each CI.

Correspondingly, *proportion gap* is the gap between the closest ends of non-overlapping intervals expressed as a proportion of the average length of the two arms that are closest. Proportion gap may be expressed as a negative POL value: two intervals that do not overlap have, in a sense, 'negative overlap', as in Figure 1(e). POL values are reported near the top in Figure 1, just below the p-values.

Cumming and Finch [5] investigated the overlap of 95 per cent CIs on independent means for random samples from normal populations and proposed the following *rule of eye*:

Two independent means, **95 per cent CIs**: *For a comparison of two independent means*, *two-tailed* $p\leqslant 0.05$ *when the overlap of the* 95 *per cent CIs is no more than about half the average margin of error*, *that is when POL is about* 0.5 *or less*. (*See Figure* 1(*c*), *and the box just below that pair of means*.)

In addition, $p\leqslant 0.01$ *when the two CIs do not overlap*, *that is when POL is about* 0 *or there is a positive gap*. (*See Figure* 1(*d*), *and the box just below*.)

These relationships are sufficiently accurate when both samples sizes are at least 10, *and the margins of error do not differ by more than a factor of* 2.

The rules of eye are intended [5] to give easily remembered, pragmatically useful guidance for anyone inspecting a figure that presents data. Rules are not intended to give p-values that are precise, and are not intended to replace statistical calculations: If an accurate p-value is desired it should where possible be calculated and reported. The investigation [5] used the method of Welch [6] and Satterthwaite [7], which pools error variances for the denominator of an independent-groups t statistic without requiring the assumption of equal variance in the two underlying populations.

In this article I review previous discussions of interval overlap, then mention a rule for bars that are $\pm$SE, which I refer to as *SE bars*. For the case of sample means from two normal populations, I investigate robustness of the rules to variation in sample sizes and interval arm lengths, and thus seek to justify the statements above for the conditions under which the rule is sufficiently accurate. I then extend the 95 per cent CI rule to include overlap benchmarks for several further p-values. Finally, I investigate CIs on proportions and correlations, cases in which intervals are in general not symmetric.

DOI: 10.1002/sim

INTERVAL OVERLAP: PREVIOUS ANALYSES FOR TWO INDEPENDENT GROUPS

In this section I briefly review previous discussions of the relation between interval overlap and two-tailed p-value or statistical significance. Note that, other things being equal, a greater difference between two independent means implies a smaller overlap or greater gap, and a smaller p-value for the t-test comparison of the means.

Browne [8] studied bar overlap and statistical significance in great detail and provided extensive tables of particular ratios that correspond to statistical significance for different pairs of sample sizes and various comparative lengths of the two intervals. No simple way to summarize the relationships was identified. Most other discussions on overlap and statistical significance have been relatively brief. Some writers have simply commented that the relationships are complex, even prompting Simpson *et al.* [9] to state that 'There is no general correspondence between overlap of these confidence intervals and the t-test for the difference of two means' (p. 352).

Many writers considered only very conservative rules. For example, Browne [8] stated that 'The only truly universal rule found was that when SE intervals overlap, the means are never significantly different [with $\alpha = 0.05$]' (p. 664). This is true but, for zero overlap of SE bars, in the base case (meaning n large, sample sizes equal, and SEs equal), the p-value is actually 0.16 [10, 11]. Similarly, Bulpitt [12] (see also [9]) stated that 'If the [95 per cent] CIs do not overlap then the means are significantly different [with $\alpha = 0.05$]' (p. 496). Again, this is true but very conservative, as noted also by [13, 14]: for zero overlap in the base case, the p-value is 0.006. In their statistics textbook, Shaughnessy *et al.* [15] advised that 'When the [95 per cent] CIs do not overlap, we can be confident that the population means for the two groups differ' (p. 247). They also stated that 'If intervals overlap slightly, then we must acknowledge our uncertainty about the true mean difference and postpone judgment' (p. 247). The rule that 95 per cent CIs just touching is equivalent to statistical significance was discussed by [16–18], and also by Schenker and Gentleman [2], who reported finding more than 60 articles in health sciences journals that had used the rule. Several writers [2–4, 19, 20] have pointed out that considerable overlap can be compatible with a significant difference, $\alpha = 0.05$. It seems that confusion about overlap and statistical significance testing has, at least in the past, been widespread. Recently, Belia *et al.* [1] found direct evidence that a large proportion of leading researchers in medicine, behavioural neuroscience, and psychology have several severe misconceptions about interpretation of overlap of 95 per cent CIs, and SE bars.

Instead of asking what overlap of familiar 95 per cent CIs gives $p = 0.05$, an alternative approach is to ask what intervals would give $p = 0.05$ when they just touch. Bars with $w = 1.39$SE [21], $\sqrt{2}$SE [22], or 1.5 to 1.6SE [23] give $p = 0.05$ or a little less, for zero overlap. For given data, increasing the level of confidence, for example from 95 to 99 per cent, lengthens the CI—by about 31 per cent. Reducing it from 95 to 90 per cent shortens the CI by about 16 per cent. Reducing CI arm length in Figure 1(c) by a little over 25 per cent would give intervals that just touch, for $p = 0.05$; the corresponding level of confidence for the base case is 83.4 per cent. A number of researchers [10, 11, 24–27] have noted that approximately 84 per cent CIs with zero overlap give $p = 0.05$. Goldstein and Healy [21] discussed more generally how intervals can be calculated so that two intervals just touching corresponds to $p = 0.05$, for a set of any number of comparisons.

Sall [28] described an ingenious variation of overlap: Around any mean, draw a circle with radius equal to the margin of error of the 95 per cent CI. Sall showed that if two such circles overlap so that they intersect at right angles, $p = 0.05$ for the comparison of the two means. The

DOI: 10.1002/sim

angle of intersection of circles is the exterior angle between the tangents to the two circles at either point where the circles cross. If this angle is greater than a right angle, the means are too close together—the circles overlap too much—for statistical significance, and $p>0.05$. If the intersection angle is less than a right angle, or the circles do not overlap, $p<0.05$. Sall claimed that with a little practice it is easy to judge whether intersection angles are more or less than right angles. One beauty of this method is that it works for any sample sizes and any margins of error: The circles may be of different sizes, but the intersecting angle rule is still accurate. In addition, with multiple means, any method for protecting against an inflated Type 1 error rate can be chosen and used to calculate circles of increased radius, and then the same intersecting angle rule gives the result of applying the chosen multiple comparison method. Sall even described an extension of the method for correlated means. Circle displays are provided in the JMP statistical software [29], and also in SAS (www.sas.com).

Tryon [26] proposed interesting ways to use CIs to assess statistical significance, statistical equivalence, and neither (statistical indeterminacy), with the aim of avoiding common problems of null hypothesis significance testing. His method uses CIs whose level of confidence, and thus length, is adjusted so the intervals just touching indicates $p=0.05$ for a comparison. As mentioned above, for two independent means in the base case the adjusted intervals are 83.4 per cent CIs. Tryon extended the analysis of [21] to explain how to calculate adjusted intervals for any sample sizes and margins of error, for sets of multiple comparisons and even for correlated means. However, the convenience of zero overlap giving $p=0.05$ comes at the enormous cost of using intervals having various unfamiliar levels of confidence. The familiar error-bar graphic, as used in Figure 1, is highly likely to elicit interpretation as a 95 per cent CI, and perhaps only an alert and strong-willed reader who grasps fully the logic of Tryon's method is likely to interpret the error bars accurately as 83.4 per cent CIs—or intervals with whatever level of confidence is needed for a particular comparison. Tryon's method thus carries the danger that intervals will be interpreted in a strongly anti-conservative way. In practice, a pair of 83.4 per cent intervals is likely to be difficult to interpret in any way other than as a signal of $p=0.05$ for a particular comparison. By contrast, the rules of eye I discuss apply to familiar 95 per cent CIs, which can also be given substantive interpretation in several other ways [5].

An advantage of Sall's circles method and Tryon's approach is that they each cover a wide range of situations, including multiple comparisons and correlated means. A disadvantage is that they focus primarily on dichotomous decision making based on the $p=0.05$ criterion, whereas the rules I discuss refer to $p=0.01$ and other values, in addition to $p=0.05$, and thus they encourage a more graded assessment of a comparison.

SE BARS

The International Committee of Medical Journal Editors [30] recommends CIs rather than SE bars, and most medical journals expect CIs to be routinely reported. Cumming and Finch [5] gave reasons for preferring 95 per cent CIs to SE bars. However, SE bars still appear fairly often in figures, and some disciplines routinely use SE bars. If sample size is at least 10, SE bars are close to half the width of the corresponding 95 per cent CI, and are approximately equivalent to 68 per cent CIs [5]. It is a major problem that the same graphic is used to represent SE bars and 95 per cent CIs. Authors must always be scrupulous to state what error bars in figures represent, and readers must always be certain which intervals are depicted before they make any interpretation.

DOI: 10.1002/sim

SE bars in figures are, unfortunately, sufficiently common that it may be worth mentioning one rule for figures with SE bars. Cumming and Finch [5] proposed the following rule of eye for SE bars:

Two independent means, **SE bars**: *For a comparison of two independent means*, *two-tailed* $p \leqslant 0.05$ *when the gap between the SE bars is at least about the size of the average SE*, *that is when the proportion gap is about* 1 *or greater*.

In addition, $p \leqslant 0.01$ *when the proportion gap is about* 2 *or more*.

These relationships are sufficiently accurate when both samples sizes are at least 10, *and the SEs of the two groups do not differ by more than a factor of* 2.

ROBUSTNESS OF OVERLAP RULES OF EYE

Browne's [8] detailed exploration of overlap led him to conclude that 'there are very few universal rules for visually assessing significance of a mean difference', and 'heteroscedasticity plays a major role in limiting rules of thumb' (p. 663). He also emphasized the central role of sample size and any difference between the sizes of the two samples. In their search for rules of eye with a useful amount of generality, Cumming and Finch [5] investigated a number of ways of expressing overlap, including overlap as a proportion of the larger margin of error, and as a proportion of the margin of error of the larger group. They also investigated the consequences of assuming, or not assuming, equality of population variances. They were pleased to find that expressing POL and gap simply in terms of average arm length gives rules that are surprisingly robust to differences between groups in both sample sizes and margins of error. POL, it seems, can bring order to the complexity described by Browne [8] and others. Cumming and Finch found that overlap rules do apply if homogeneity is assumed, but apply in a wider range of cases if homogeneity of variance is not assumed. As mentioned earlier their reported analyses used the Welch–Satterthwaite method, which does not assume homogeneity of variance, to calculate p-values.

For maximum scope, I also did not assume homogeneity of variance and used Welch–Satterthwaite calculations to investigate robustness of the rules. Figure 2 illustrates robustness by showing how p varies with the size of the smaller sample (which we can label Group 1, of size n_1) for 95 per cent CI POL of 0.5 (solid curves), and for SE proportion gap of 1.0 (dotted curves). The heavy curves apply to the base case of groups of equal size with bars of equal width. The light curves show the relationship for three combinations of sample size and bar width differences. Figure 2 illustrates my general conclusion about robustness, which is that in the great majority of cases that meet the conditions for the rules—sample sizes of at least 10, and margins of error (or SEs) not differing by more than a factor of 2—the p-value is close to and a little below 0.05. It is striking that the rules are reasonable even when Group 2 is five times the size of Group 1 and, in addition, has an interval either twice or half the width of the Group 1 interval. In fact the curves hardly change for even greater differences in group size, but note that in such cases, for interval widths to be within a factor of 2, the standard deviations in the two groups would need to be quite different.

A corresponding figure (not shown) gives a similar justification for the choice of 95 per cent CI POL of 0 and SE proportion gap of 2.0, for p approximately 0.01, and for the statement of the conditions under which the rules for $p = 0.01$ are acceptably accurate. The figures for the 0.05 and 0.01 rules illustrate that, for the base case, the rules are conservative: For large and equal n, and equal interval widths, 95 per cent CI POL of 0.5 gives $p = 0.038$ rather than 0.05, and zero overlap

DOI: 10.1002/sim

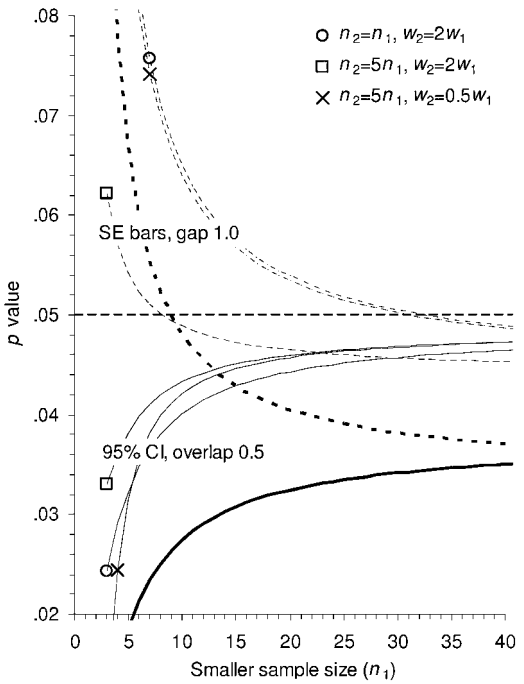

Figure 2. Curves to illustrate the robustness of rules of eye for 95 per cent CIs and SE bars. The two-tailed p-value calculated by the Welch–Satterthwaite method is plotted as a function of sample size, for 95 per cent CI proportion overlap of 0.5 (the criterion for the 95 per cent CI rule of eye, solid curves), and for proportion SE gap of 1.0 (the criterion for the SE bars rule of eye, dotted curves). The horizontal dashed line marks 0.05, the p-value specified approximately in the rules of eye. Groups 1 and 2 have sizes n_1 and n_2, respectively, where $n_1 \leqslant n_2$, and margins of error (or SEs) of w_1 and w_2, respectively. The two heavy curves apply for the base case of groups of equal size having margins of error (or SEs) that are equal. The light curves show the relationship for three combinations of sample size and interval width differences. Curves marked with a circle illustrate equal group sizes ($n_1=n_2$) and one margin of error (or SE) twice the other ($w_2=2w_1$). For curves with a square or a cross, Group 2 is five times the size of Group 1 ($n_2=5n_1$), and has margin of error (or SE) twice (square; $w_2=2w_1$) or half (cross; $w_2=0.5w_1$) that of Group 1. Note that in the great majority of cases the p-value is close to and a little below 0.05, provided that both groups have size at least 10.

gives $p=0.006$ rather than 0.01. This conservatism of the rules for the base case is also illustrated in Figure 1, which shows near the top the exact POL required for p exactly 0.05 (Figure 1(c)) or 0.01 (Figure 1(d)): A little more overlap is permitted, or a smaller gap required, than is stated in the rules. Figure 1 also shows, in small boxes at the bottom, the pairings of POL and approximate p-values stated in the rules of eye.

Relations among C, p, and overlap

In this section I discuss how changes in level of confidence C, the p-value, and extent of overlap are all related. First consider p and overlap, for a fixed C. Figure 1 illustrates how, for 95 per cent CIs, the p-value shown at the top decreases as overlap changes from about 1 in Figure 1(a) to a

DOI: 10.1002/sim

gap in Figure 1(e). Statistical reformers, for example [5], advocate reduced focus on the traditional criterion p-values of 0.05 and 0.01, and instead, reporting of exact p-values as a more informative basis for inference. Bearing in mind the patterns of Figure 1 allow easy estimation of the p-value for any amount of overlap (or gap) of two independent 95 per cent CIs: inference by eye need not be restricted to p-values of 0.05 and 0.01.

The first rule of eye and the bold boxes at the bottom in Figure 1 give the two basic benchmarks that POL$=0.5$ corresponds to about $p=0.05$, and POL$=0$ to about $p=0.01$. Two additional benchmarks worth remembering are shown in the other boxes in Figure 1: An overlap of one (each mean is at a limit of the other interval) gives about $p=0.2$, and a gap of 0.5 gives about $p=0.001$. It is notable that such a relatively small gap between 95 per cent CIs is needed to give a very low p-value. All these benchmarks are a little conservative for the base case, and allow sufficient leeway to meet the standard robustness requirements that the rule is adequate whenever both sample sizes are at least 10, and interval widths do not differ by more than a factor of 2.

Earlier I mentioned that for a given data, and therefore a given p-value, changing C will change interval length and thus the extent of overlap. More generally, it is possible to plot families of curves showing the relation between any two of C, p, and POL. It is also possible to state rules of eye for CIs with any chosen C, although 95 per cent CIs should usually be preferred because they are most familiar, and consistency of practice assists interpretation.

Proliferation of rules and benchmarks to remember is unlikely to be useful for the busy researcher, but I will briefly mention two cases that may be of interest to some groups of researchers. First, 99 per cent CIs are sometimes reported in medicine. For 99 per cent CIs, POL of 0.5 indicates $p=0.01$ approximately. As for the other rules, this 0.5 benchmark allows leeway for the usual robustness conditions (sample sizes at least 10, and margins of error differing by no more than a factor of 2) to suffice.

Second, researchers using structural equation modelling (SEM) often present 90 per cent CIs on root mean square error of approximation estimates. For 90 per cent CIs, the rule identifies overlap of one quarter for $p=0.05$, and a gap of one third for $p=0.01$. Remembering those two values may be useful for quick inference by eye when reading SEM articles. Again the usual robustness conditions apply.

CIs for two proportions

The discussion so far has referred to normal populations and symmetric CIs on sample means. As a first step of generalization, I studied overlap of 95 per cent CIs for two independent proportions $x_1=k_1/n_1$ and $x_2=k_2/n_2$, where k_1, k_2, n_1, and n_2 are all integers, and $0\leqslant k_i\leqslant n_i$. Because the proportions are bounded by 0 and 1, such CIs are in general asymmetric, meaning that the lower and upper arms are of unequal length. To calculate CIs on proportions, I used the approximate method recommended in [31], based on extensive comparisons in [32] that included proportions from the full range from 0 to 1 inclusive, and values of n_1 and n_2 from 5 to 100 000. Figure 3 shows two examples of 95 per cent CIs on two independent proportions and illustrates that the arm closer to 0.5 is the longer arm. POL is the vertical distance between the dotted horizontals, as in Figure 3, divided by the average length of the two arms that overlap—one from each CI.

Incidentally, Figure 3 is the one figure I include that illustrates the overlap of CIs whose margins of error (values of w) differ. Often in practice the two overlapping arms do not differ greatly in length, and it is sufficient to estimate visually the overlap as a proportion of either overlapping

DOI: 10.1002/sim

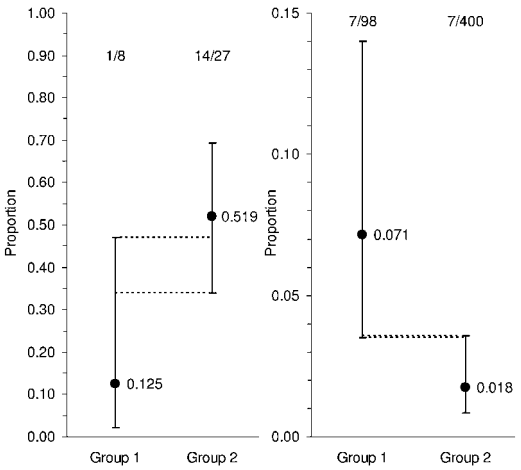

Figure 3. Two examples of 95 per cent CIs for two independent proportions. Note that the panels have different vertical scales. The left panel shows proportions $\frac{1}{8}$ (Group 1) and $\frac{14}{27}$ (Group 2). The proportion overlap is POL$=0.50$, meaning that the vertical distance between the two dotted lines is 0.50 of the average of the two overlapping arms, which are the upper arm of the left CI and the lower of the right. The ratio of the lengths of these two arms is 1.94, and the two-tailed p-value is 0.048. The right panel shows proportions $\frac{7}{98}$ (Group 1) and $\frac{7}{400}$ (Group 2). The proportion overlap is close to zero (it is 0.024), the arm ratio is 2.00, and the p-value is 0.005. These configurations illustrate cases close to the boundaries of the rule of eye for two independent proportions.

arm. Applying the rule in the more challenging case of the left panel of Figure 3, requires first deciding that the longer of the two overlapping arms is not appreciably more than twice the length of the shorter, and so the rule is applicable. Then, the overlap needs to be assessed against the average of the two overlapping arms. One way to do this is to transfer visually the shorter arm alongside the longer, with the two upper ends aligned, and to estimate the point half way between the lower ends of the two—which defines the desired average of the two overlapping arms. Then, assess the extent of overlap against this average. Bear in mind that inference by eye is not intended to be precise, and that the rules are a little conservative and therefore allow a little leeway.

To calculate the p-value for the comparison of the two proportions, I used the method recommended in [31], which was based on extensive evaluations in [33]. I used that method to calculate the 95 per cent CI on the difference between the proportions, then adjusted the confidence level C of that CI until one limit equalled zero. The p-value was then $(100-C)/100$.

These investigations justify the following rule of eye:

Two independent proportions, **95 per cent CIs**: *For a comparison of two independent proportions*, *two-tailed* $p\leqslant 0.05$ *when POL is about* 0.5 *or less—in other words the overlap of the* 95 *per cent CIs is no more than about half the average arm length*, *meaning the average of the two arms that overlap* (*Figure* 3, *left panel*).

In addition, $p\leqslant 0.01$ *when the two CIs do not overlap*, *that is when* POL$\leqslant 0$ *approximately*, *thus overlap is about* 0 *or there is a positive gap* (*Figure* 3, *right panel*).

These relationships are sufficiently accurate when the arm ratio (*the ratio*, *greater than or equal to one*, *of the lengths of the two arms that overlap*) *is no more than about* 2.

DOI: 10.1002/sim

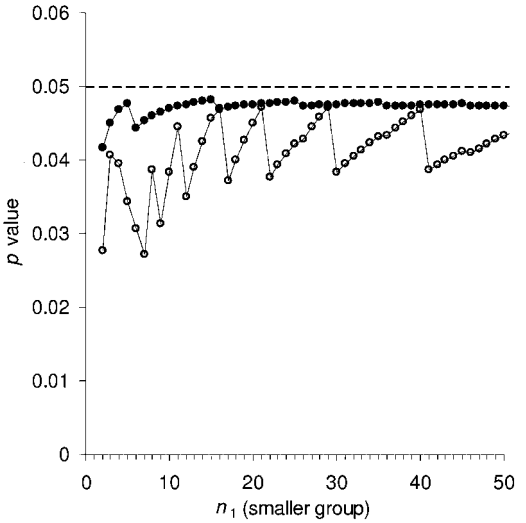

Figure 4. Curves to illustrate the robustness of the rule of eye for two independent proportions. The two-tailed p-value is plotted against n_1, which is the size of the smaller group. The larger group has size $3n_1$. The proportion in the smaller group is $x_1 = k_1/n_1$, and is set as close as possible to 0.9 subject to k_1 being an integer. The proportion in the larger group is $x_2 = k_2/n_2$, which is first set to the value less than x_1 that gives proportion overlap of the two 95 per cent CIs of exactly 0.50. The filled dots show the p-values for these cases. Then, x_2 is rounded down until k_2 is an integer. The open dots mark the p-values for these cases. The steps in the upper curve (filled dots) reflect rounding of x_1 to ensure k_1 is integral, and the steps in the lower curve reflect also the rounding of x_2 so k_2 is integral. The horizontal dashed line marks 0.05, the p-value specified approximately in the rule of eye. The maximum arm ratio for the cases illustrated is 1.94.

To assess the accuracy and robustness of this rule, I investigated numerous cases with n_1 and n_2 ranging from 2 to 100 000, and x_1 and x_2 ranging from 0 to 1. As an example, Figure 4 shows the p-value as a function of n_1 for cases in which $n_2 = 3n_1$, and when x_1 is set as close as possible to 0.9, subject to k_1 being an integer. The procedure was, first, to determine for each n_1 the x_2 that was less than x_1 and gave a 95 per cent CI, having POL of exactly 0.5 with the CI on x_1; the p-values for those cases are marked by the closed dots in Figure 4. I then reduced x_2 by as little as possible until k_2 was an integer; this reduced the overlap and thus the p-values, which are marked by the open dots. The jaggedness in the curves reflects the requirements that k_1 be integral for the upper curve, and both k_1 and k_2 integral for the lower curve. For the example cases reported in Figure 4, the arm ratio lay between 1.66 and 1.94, and all p-values are close to, but below 0.05.

Examination of numerous diagrams like Figure 4 indicated that, when overlap is 0.5, the p-value is in the great majority of cases between 0.035 and 0.05 (including the example shown in Figure 3, left panel), and in almost every case between 0.025 and 0.05. The most common p-value, when x_1 and x_2 are not extreme and n_1 and n_2 do not differ greatly, is around 0.04. This is similar to the p-value of 0.038 for the normal distribution base case.

Similar investigations with zero overlap indicate that the p-value is in the great majority of cases between 0.005 and 0.008 (including the example shown in Figure 3, right panel) and in almost every case between 0.004 and 0.01. The most common p-value, when x_1 and x_2 are not

DOI: 10.1002/sim

extreme and n_1 and n_2 do not differ greatly, is around 0.006. This is also the p-value for the normal distribution base case.

Measuring overlap in terms of the average of the lengths of the two arms that overlap seems appropriate, and only when the arm ratio is greater than 2 do the rules in many cases break down. This limitation on arm ratio means that the n_2/n_1 ratio (or the n_1/n_2 ratio) can be no more than about 4, and somewhat less if an x_i is close to 0 or 1. It is striking that the rules hold even for very small n_1 and/or n_2, and for x_1 and/or x_2 close to or equal to 0 or 1, provided only that the arm ratio is no more than 2. The rules for two independent proportions are, like the corresponding rules for normal populations, a little conservative, in that the p-value is almost always a little smaller than the 0.05 and 0.01 values stated in the rule.

Finally, note that, although the approximate method used here has been extensively tested and is recommended in [31], other methods do exist for calculating CIs for proportions and the p-value for the difference. Investigation and comparison of such methods continues in the literature. Should another method find favour, the rule of eye should be tested with the CIs and p-values it gives, although I would expect the rule to hold for any method that gives CIs with accurate coverage probabilities and little bias.

CIs for two correlations

I took a similar approach to investigate the overlap of 95 per cent CIs on two independent Pearson correlations, r_1 and r_2, in groups of size n_1 and n_2, respectively. I used Fisher's r to z transformation [34] to calculate 95 per cent CIs on each r_i, and the p-value for the comparison of the two correlations. The two underlying populations are assumed bivariate normal. Because correlations are bounded by -1 and 1, the CIs are in general asymmetric, with the arm closer to 0 being the longer. Diagrams showing the overlap of CIs for correlations appear not unlike Figure 3, except that the vertical axis extends from 1.0 down to -1.0, and thus CIs can extend beyond zero.

These investigations justify the following rule of eye:

Two independent correlations, 95 per cent CIs: *For a comparison of two independent Pearson correlations, two-tailed* $p \leqslant 0.05$ *when POL is about* 0.5 *or less—in other words the overlap of the* 95 *per cent CIs is no more than about half the average arm length, meaning the average of the two arms that overlap.*

In addition, $p \leqslant 0.01$ *when the two CIs do not overlap, that is when* POL $\leqslant 0$ *approximately, and thus overlap is about* 0 *or there is a positive gap.*

These relationships are sufficiently accurate when both group sizes are at least 30, *and the arm ratio* (*the ratio, greater than or equal to one, of the lengths of the two arms that overlap*) *is no more than about* 2.

Figure 5 shows the p-value as a function of n_1, the smaller group size, for selected values of r_1 and n_2/n_1. The procedure was to choose values of r_1 and the n_2/n_1 ratio, and then, for each n_1 value, calculate r_2 so the overlap of the 95 per cent CIs on r_1 and r_2 is exactly 0.50, then calculate the p-value for the difference. The heavy curve (1) is for the base case when $r_1 = 0$ and $n_2/n_1 = 1$. The other curves are for the r_1 and n_2/n_1 values indicated, and for r_2 constrained to be greater or less than r_1, as indicated. The upper three curves (4, 5, and 6) were selected so that the arm ratios for points on these curves are at or just below the 2.0 limit specified by the rule. The arm ratios for points on Curves 1–3 are generally in the range 1.0–1.3.

Examination of numerous diagrams like Figure 5 indicated that, when overlap is 0.5, and n_1 is at least 30 and the arm ratio no more than 2, the p-value is in the great majority of cases between

DOI: 10.1002/sim

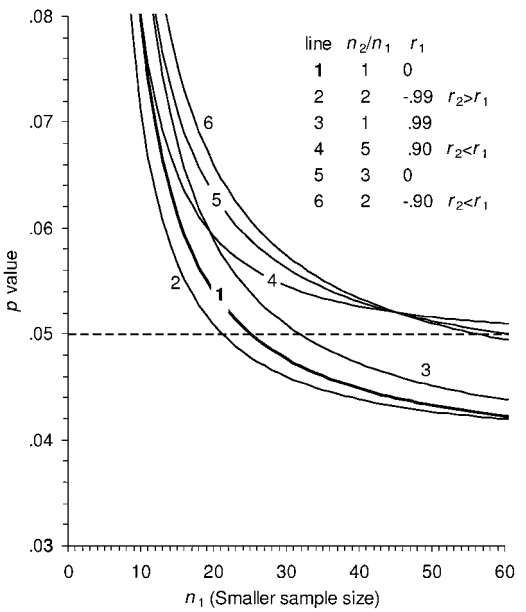

Figure 5. Curves to illustrate the robustness of the rule of eye for two independent Pearson correlations, r_1 and r_2, in groups of size n_1 and n_2, respectively, where $n_1 \leqslant n_2$. The two-tailed p-value is plotted against n_1, the size of the smaller group, and the horizontal dashed line marks 0.05, the p-value specified approximately in the rule of eye. For a chosen r_1 and n_2/n_1 ratio, for each n_1 value r_2 is set so the overlap of the 95 per cent CIs on r_1 and r_2 is exactly 0.50, then the p-value is calculated. The heavy curve (1) is for the base case when $r_1 = 0$ and $n_2/n_1 = 1$. The other curves are for the r_1 and n_2/n_1 values indicated. For curve 2, r_2 was constrained to be greater than r_1, and for curves 4 and 6 to be less than r_1. (Curves 1 and 3 are symmetric with respect to the two correlations, and curve 5 symmetric as to the sign of r_2.) The upper three curves (4, 5, and 6) were selected so the arm ratios for points on these curves are at or just below the 2.0 limit specified by the rule of eye. The arm ratios for points on curves 1–3 are generally in the range 1.0–1.3.

0.04 and 0.055, and in almost every case between 0.035 and 0.06. When n_1 and n_2 are large and similar the p-value is, as expected, close to 0.038, the value for the normal distribution base case. Once again, measuring overlap in terms of the average of the lengths of the two arms that overlap seems appropriate, and only when the arm ratio is greater than 2 or group sizes are small does the rule in many cases break down. It is notable that the rule holds even for r_1 or r_2 close to -1 or 1, provided only that group sizes are at least 30 and the arm ratio is no more than 2.

Similar investigations with zero overlap indicated that the p-value is, in the great majority of cases that satisfy the rule, between 0.006 and 0.011 and in almost every case between 0.006 and 0.015. When n_1 and n_2 are large and similar the p-value is, as expected, close to 0.006, the value for the normal distribution base case. Once again it is notable that the rule holds in such a wide range of cases, subject only to simple conditions. The rule for two independent correlations is, like the earlier rules, a little conservative when groups sizes are large and similar but, as Figure 5 illustrates for $p = 0.05$, for cases close to the boundaries (for example $n_1 = 30$, arm ratio near 2, a correlation close to -1 or 1, as in Curves 4, 5, and 6) p-values are sometimes a little above the 0.05 and 0.01 values stated in the rule.

DOI: 10.1002/sim

CONCLUSION

When inspecting error bars on two means, or other point estimates, it is essential first to be sure what the bars represent: CIs, SE, SD, or some other quantity? If they are CIs, what is the level of confidence? Having established that they are, for example, 95 per cent CIs, it is next essential to be sure that the two samples are independent. If a repeated measure is involved, or the two means are in some other way correlated, the two CIs may not be used to assess the difference, because they do not reflect the correlation. For pre-test and post-test means, for example, the CI on the paired differences is needed [5].

Having established that the 95 per cent CIs are for independent samples, inference by eye can be based on one of the rules stated above, and in the Appendix, provided the conditions in the rule are met. However, bear in mind that the p-value relates to a single comparison of two means, and therefore, using the rules is equivalent to regarding each comparison as a separate decision—in other words using a decisionwise error rate. Saville [20] defended the use of a decisionwise error rate, no matter how many comparisons are made, provided that any differences identified as statistically significant should be considered substantively on their merits, and regarded as possible effects for further investigation, rather than established findings. Alternatively, if many comparisons are made, or if the two means to be compared are chosen *post hoc* from a large set, then a setwise, or experimentwise error rate may be preferred. An informal adjustment towards conservatism, to protect against inflated Type 1 error rates [5], may be to use $p<0.01$, for example, in place of $p<0.05$ as a general inference-by-eye benchmark. More formal adjustments of p, using Bonferroni or some other procedure, would require calculation of p-values rather than eye balling of figures with error bars.

Judgments based on CI overlap can be particularly useful in situations where CIs are readily calculated, but a test for statistical significance is not known, or is difficult to compute [10]. The example given in [10] was a comparison of the coefficients of variation of two independent samples. However, although the above results for proportions and correlations encourage a belief that the CI overlap rules have wide generality, their accuracy should really be examined for each different type of comparison.

It is timely to discuss inference based on CIs because statistical reform, including widespread use of CIs, continues to advance. CIs came in to routine use in medicine during the 1980s, but changes in other disciplines have been more recent. In psychology, the influential *Publication Manual* of the American Psychological Association in 2001 recommended CIs [35], and a similar recommendation was made in 2006 in educational research [36]. In economics the reform debate and advocacy of CIs continues [37].

I have two further caveats: First, exact p-values should not be taken as a precise measure of the strength of evidence given by a set of data because, if you repeat the experiment exactly but with a new sample of subjects, you are likely to obtain a quite different p-value [38]. In this article I have followed convention by calculating p-values precisely, but inference by eye, and indeed any use of p-values, should recognize that they give only a very rough indication of strength of evidence.

Second, discussing these rules of eye may suggest that estimating p-values should be the primary way to interpret CIs. By contrast, Cumming and Finch [5] recommended a range of ways to think about CIs without invoking p-values. Interpretation of any CI should be primarily in terms of point and interval estimates, and interpretation of these in the research context. Consideration of p-values may be helpful, but as statistical reform progresses and we become more familiar with interval estimation [39], the focus on p-values should reduce. In any case, I hope these rules will

DOI: 10.1002/sim

encourage researchers to publish figures showing 95 per cent CIs, and help readers appreciate such figures more readily.

APPENDIX A: RULES OF EYE FOR INTERPRETATION OF CONFIDENCE INTERVALS

The first four rules, for means, assume normally distributed populations, and refer to p-values calculated using Welch–Satterthwaite methods, which do not require the assumption of equal population variances.

Two independent means, 95 per cent CIs [5]. *For a comparison of two independent means, two-tailed* $p \leqslant 0.05$ *when the overlap of the* 95 *per cent CIs is no more than about half the average margin of error, that is when proportion overlap* (*POL*) *is about* 0.5 *or less.* (*See Figure* 1(*c*), *and the box just below that pair of means.*)

In addition, $p \leqslant 0.01$ *when the two CIs do not overlap, that is when proportion overlap is about* 0 *or there is a positive gap.* (*See Figure* 1(*d*), *and the box just below.*)

These relationships are sufficiently accurate when both samples sizes are at least 10, *and the margins of error do not differ by more than a factor of* 2.

Two independent means, SE bars [5]. *For a comparison of two independent means, two-tailed* $p \leqslant 0.05$ *when the gap between the SE bars is at least about the size of the average SE, that is when the proportion gap is about* 1 *or greater.*

In addition, $p \leqslant 0.01$ *when the proportion gap is about two or more.*

These relationships are sufficiently accurate when both samples sizes are at least 10, *and the SEs of the two groups do not differ by more than a factor of* 2.

Estimation of p for two independent 95 per cent CIs. *For a comparison of two independent means, p can be estimated for any observed overlap or gap of the* 95 *per cent CIs by using as approximate benchmarks*: POL = 1 (*one full arm overlap*) *gives two-tailed* $p = 0.2$ (*see Figure* 1(*a*)); POL = 0.5 *gives* $p = 0.05$ (*Figure* 1(*c*)); POL = 0 (*intervals just touching*) *gives* $p = 0.01$ (*Figure* 1(*d*)); *and* $p = 0.001$ *when* POL = −0.5, *meaning a gap of half the average margin of error* (*Figure* 1(*e*)).

These benchmarks are sufficiently accurate when both samples sizes are at least 10, *and the margins of error do not differ by more than a factor of* 2.

Two independent means, 99 per cent and 90 per cent CIs. *For a comparison of two independent means, when the overlap of* 99 *per cent CIs is* POL = 0.5, *p is about* 0.01. *When the overlap of* 90 *per cent CIs is* POL = 0.25, *two-tailed p is about* 0.05; *and p is about* 0.01 *when* POL = −0.33, *meaning a gap of about one third of the average margin of error.*

These relationships are sufficiently accurate when both samples sizes are at least 10, *and the margins of error do not differ by more than a factor of* 2.

Two independent proportions, 95 per cent CIs. *For a comparison of two independent proportions, two-tailed* $p \leqslant 0.05$ *when POL is about* 0.5 *or less—in other words the overlap of the* 95 *per cent CIs is no more than about half the average arm length, meaning the average of the two arms that overlap* (*Figure* 3, *left panel*).

In addition, $p \leqslant 0.01$ *when the two CIs do not overlap, that is when* POL $\leqslant 0$ *approximately, thus the overlap is about* 0 *or there is a positive gap* (*Figure* 3, *right panel*).

These relationships are sufficiently accurate when the arm ratio (*the ratio, greater than or equal to one, of the lengths of the two arms that overlap*) *is no more than about* 2.

DOI: 10.1002/sim

Two independent correlations, 95 per cent CIs. *For a comparison of two independent Pearson correlations, two-tailed* $p \leqslant 0.05$ *when POL is about* 0.5 *or less—in other words the overlap of the* 95 *per cent CIs is no more than about half the average arm length, meaning the average of the two arms that overlap.*

In addition, $p \leqslant 0.01$ *when the two CIs do not overlap, that is when* $\mathrm{POL} \leqslant 0$ *approximately, and thus the overlap is about* 0 *or there is a positive gap.*

Calculation of p-values is based on Fisher's r to z transformation, and the two underlying populations are assumed bivariate normal. The relationships are sufficiently accurate when both group sizes are at least 30, *and the arm ratio* (*the ratio, greater than or equal to one, of the lengths of the two arms that overlap*) *is no more than about* 2.

ACKNOWLEDGEMENTS

This research was supported by the Australian Research Council. I thank Toby Cumming, Cathy Faulkner, Fiona Fidler, and Warren Tryon for valuable comments on drafts.

REFERENCES

1. Belia S, Fidler F, Williams J, Cumming G. Researchers misunderstand confidence intervals and standard error bars. *Psychological Methods* 2005; **10**:389–396. DOI: 10.1037/1082-989X.10.4.389.
2. Schenker N, Gentleman JF. On judging the significance of differences by examining the overlap between confidence intervals. *The American Statistician* 2001; **55**:182–186. DOI: 10.1198/000313001317097960.
3. Austin PC, Hux JE. A brief note on overlapping confidence intervals. *Journal of Vascular Surgery* 2002; **36**:194–195. DOI: 10.1067/mva.2002.125015.
4. Wolfe R, Hanley J. If we're so different why do we keep overlapping? When 1 plus 1 doesn't make 2. *Canadian Medical Association Journal* 2002; **166**:65–66.
5. Cumming G, Finch S. Inference by eye: confidence intervals and how to read pictures of data. *American Psychologist* 2005; **60**:170–180. DOI: 10.1037/0003-066X.60.2.170.
6. Welch BL. The significance of the difference between two means when the population variances are unequal. *Biometrika* 1938; **29**:350–362. DOI: 10.1093/biomet/29.3-4.350.
7. Satterthwaite FE. An approximate distribution of estimates of variance components. *Biometrics Bulletin* 1946; **2**:110–114. DOI: 10.2307/3002019.
8. Browne RH. On visual assessment of the significance of a mean difference. *Biometrics* 1979; **35**:657–665. DOI: 10.2307/2530259.
9. Simpson GG, Roe A, Lewontin RC. *Quantitative Zoology* (revised edn). Harcourt, Brace & World: New York, 1960.
10. Payton ME, Greenstone MH, Schenker N. Overlapping confidence intervals or standard error intervals: what do they mean in terms of statistical significance? *Journal of Insect Science* 2003; **3**:34–39. Available at insectscience.org/3.34.
11. Payton ME, Miller AE, Raun WR. Testing statistical hypotheses using standard error bars and confidence intervals. *Communications in Soil Science and Plant Analysis* 2000; **31**:547–551.
12. Bulpitt CJ. Confidence intervals. *The Lancet* 1987; **i**:494–497. DOI: 10.1016/S0140-6736(87)92100-3.
13. Cole SR, Blair RC. Overlapping confidence intervals. *Journal of the American Academy of Dermatology* 1999; **41**:1051–1052.
14. Wheeler MW, Park RM, Bailer AJ. Comparing median lethal concentration values using confidence interval overlap or ratio tests. *Environmental Toxicology and Chemistry* 2006; **25**:1441–1444. DOI: 10.1897/05-320R.1.
15. Shaughnessy JJ, Zechmeister EB, Zechmeister JS. *Research Methods in Psychology* (7th edn). McGraw-Hill: New York, 2006.
16. Bigby M, Gadenne A. Understanding and evaluating clinical trials. *Journal of the American Academy of Dermatology* 1996; **34**:555–590. DOI: 10.1016/S0190-9622(96)80053-3.
17. Bigby M, Gadenne A. Reply. *Journal of the American Academy of Dermatology* 1997; **37**:804–805. DOI: 10.1016/S0190-9622(97)70127-0.

18. Rahlfs VW. Understanding and evaluating clinical trials. *Journal of the American Academy of Dermatology* 1997; **37**:803–804. DOI: 10.1016/S0190-9622(97)70126-9.
19. Ryan GW, Leadbetter SD. On the misuse of confidence intervals for two means in testing for the significance of the difference between the means. *Journal of Modern Applied Statistical Methods* 2002; **1**:473–478.
20. Saville DJ. Basic statistics and the inconsistency of multiple comparison procedures. *Canadian Journal of Experimental Psychology* 2003; **57**:167–175.
21. Goldstein H, Healy MJR. The graphical presentation of a collection of means. *Journal of the Royal Statistical Society A* 1995; **158**:175–177. DOI: 10.2307/2983411.
22. Burghardt GM. Comparative prey-attack studies in newborn snakes of the genus *Thamnophis*. *Behaviour* 1969; **33**:77–114. DOI: 10.1163/156853969X00332.
23. Moses LE. Graphical methods in statistical analysis. *Annual Review of Public Health* 1987; **8**:309–353. DOI: 10.1146/annurev.pu.08.050187.001521.
24. Barr DR. Using confidence intervals to test hypotheses. *Journal of Quality Technology* 1969; **1**:256–258.
25. Nelson LS. Evaluating overlapping confidence intervals. *Journal of Quality Technology* 1989; **21**:140–141.
26. Tryon WW. Evaluating statistical difference, equivalence, and indeterminacy using inferential confidence intervals: an integrated alternative method of conducting null hypothesis statistical tests. *Psychological Methods* 2001; **6**:371–386. DOI: 10.1037/1082-989X.6.4.371.
27. Julious SA. Using confidence intervals around individual means to assess statistical significance between two means. *Pharmaceutical Statistics* 2004; **3**:217–222. DOI: 10.1002/pst.126.
28. Sall J. Graphical comparison of means. *Statistical Computing and Graphics Newsletter of the American Statistical Association* 1992; **3**(1):27–32.
29. Sall J, Creighton L, Lehman A. *JMP Start Statistics*: *A Guide to Statistics and Data Analysis Using JMP and JMP IN Software* (3rd edn). Thomson Brooks/Cole: Pacific Grove, CA, 2005.
30. International Committee of Medical Journal Editors. Uniform requirements for manuscripts submitted to biomedical journals. *Annals of Internal Medicine* 1997; **126**:36–47.
31. Newcombe RG, Altman DG. Proportions and their differences. In *Statistics with Confidence*: *Confidence Intervals and Statistical Guidelines* (2nd edn), Altman DG, Machin D, Bryant TN, Gardner MJ (eds). British Medical Journal Books: London, 2000; 45–56.
32. Newcombe RG. Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in Medicine* 1998; **17**:857–872. DOI: 10.1002/(SICI)1097-0258(19980430)17:8<857::AID-SIM777>3.0.CO;2-E.
33. Newcombe RG. Interval estimation for the difference between independent proportions: comparison of eleven methods. *Statistics in Medicine* 1998; **17**:873–890. DOI: 10.1002/(SICI)1097-0258(19980430)17:8<873::AID-SIM779>3.0.CO;2-I.
34. Howell DC. *Statistical Methods for Psychology* (5th edn). Duxbury: Pacific Grove, CA, 2002.
35. American Psychological Association. *Publication Manual of the American Psychological Association* (5th edn). Author: Washington, DC, 2001.
36. American Educational Research Association. Standards for reporting on empirical social science research in AERA publications. *Educational Researcher* 2006; **35**:33–40. DOI: 10.3102/0013189X035006033.
37. Fidler F, Cumming G, Burgman M, Thomason N. Statistical reform in medicine, psychology and ecology. *Journal of Socio-Economics* 2004; **33**:615–630. DOI: 10.1016/j.socec.2004.09.035.
38. Cumming G. Replication and *p* intervals: *p* values predict the future only vaguely, but confidence intervals do much better. *Perspectives on Psychological Science* 2008; **3**:286–300. DOI: 10.1111/j.1745-6924.2008.00079.x.
39. Cumming G, Fidler F, Leonard M, Kalinowski P, Christiansen A, Kleinig A, Lo J, McMenamin N, Wilson S. Statistical reform in psychology: is anything changing? *Psychological Science* 2007; **18**:230–232. DOI: 10.1111/j.1467-9280.2007.01881.x.