Judging significance from error bars

Thomas P. Minka

November 6, 2002

Abstract

This paper explores and evaluates different ways to judge statistical significance from a graph of estimates with error bars. The traditional bar-overlaps-bar test for judging significance is found to have low power. Better power is obtained from the bar-ovarlaps-point test, or a combination of tests. Finally, it is shown that an optimal graphical test is possible by using error *circles* instead of error bars.

1 Introduction

An important problem in statistical graphics is indicating the reliability of statistical estimates. The conventional approach to this is error bars, which are often chosen to indicate a 68% or 95% confidence interval on each estimate (Schmid, 1983). However, it has not been made clear how confidence intervals of this type should be used to judge significance between two estimates, what significance level is achieved by different methods, and what is their power. In this paper, these questions are studied in detail. The results show, among other things, that putting 95% confidence intervals on each estimate is overly conservative: the error bars will overlap unless the difference between estimates is significant at the *98%* level.

The fundamental question is: given a plot of two estimates, $\hat{x}$ and $\hat{y}$, with error bars, is their ordering significant? In this paper, we will focus is on symmetric error bars: $\hat{x} \pm e_x$. A more formal statement of the problem is that there are unknown quantities x and y about which we have some information, and we want to test the hypothesis that $x \leq y$ (assuming $\hat{x} < \hat{y}$ on the graph). Note that we are not testing whether $x = y$, only whether the ordering of $\hat{x}$ and $\hat{y}$ seen on the graph should be believed.

If our estimates are $\hat{x}$ and $\hat{y}$, with standard errors e_x and e_y respectively, then the classical formula for this is the z-test:

$$z = \frac{\hat{y} - \hat{x}}{\sqrt{e_x^2 + e_y^2}} \tag{1}$$

The normal cumulative distribution function ϕ converts from z into a significance level, e.g. $z = 2$ corresponds to $\phi(2) = 98\%$ significance, and $z = 1.64$ corresponds to 95% significance. Note that we are using a z-test instead of a t-test because the standard errors are assumed known, otherwise we would need error bars for the error bars, or somehow include sample sizes on the plot.

The above procedure has a simple Bayesian interpretation. Let the posterior for x be normal with mean $\hat{x}$ and standard deviation e_x, and likewise for y. We want to compute the probability that $x < y$, where x and y are regarded as independent normal random variables. This probability is $\phi(z)$.

Taking the z-test as a gold standard, how closely can we approximate it by graphical operations? Let's start with the commonly-used bar-overlaps-bar test.

2 Bar-overlaps-bar test

In a bar-overlaps-bar test, you check if x's error bar overlaps y's error bar. Let the error bars be $\hat{x} \pm ce_x$ and $\hat{y} \pm ce_y$ for some scale factor c. We want to reject the null hypothesis that $\hat{y} \leq \hat{x}$, i.e. that the ordering on the graph is wrong.

Theorem If the error bars do not overlap, then the ordering is significant at a level exceeding $\phi(c)$.

Proof If the error bars do not overlap, then

$$\begin{aligned} \hat{y} - \hat{x} &> c(e_x + e_y) && (2) \\ &> c\sqrt{e_x^2 + e_y^2} && (3) \end{aligned}$$

because

$$e_x + e_y \geq \sqrt{e_x^2 + e_y^2} \tag{4}$$

(This is easy to check by squaring both sides.) Therefore

$$z = \frac{\hat{y} - \hat{x}}{\sqrt{e_x^2 + e_y^2}} \geq \frac{\hat{y} - \hat{x}}{e_x + e_y} > c \tag{5}$$

For example, if $c = 1.64$ then when the bars don't overlap $z \geq 1.64$ and the test has $\geq 95\%$ significance (in the Bayesian case, $Pr(x < y) \geq 0.95$). Sometimes you find graphs where $c = 1$. In this case, the test has $\geq 84\%$ significance. The test is conservative because it might not identify the ordering as significant even though a z-test would. The situation is illustrated in figure 1.

The degree of conservatism depends on the similarity of e_x and e_y, as given in the following theorem:

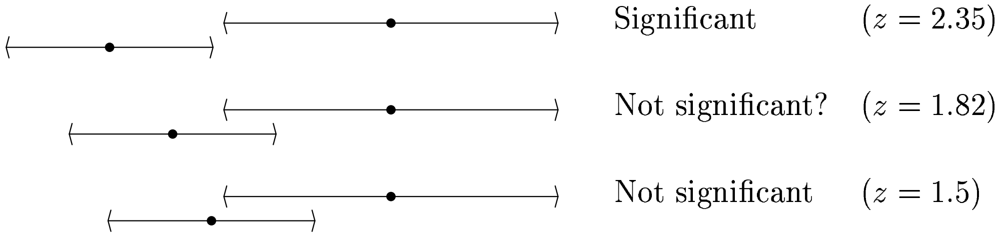


Figure 1: The weakness of the bar-overlaps-bar test when $e_x = 1.3$ and $e_y = 2.1$. Bars are scaled by $c = 1.64$, corresponding to the assumption that $z > 1.64$ is significant. The bars have similar length, so the test is overly conservative in declaring an ordering significant.

Theorem The bar-overlaps-bar test has greatest power when $e_x = 0$ or $e_y = 0$, and least power when $e_x = e_y$.

Proof When $e_x = 0$ or $e_y = 0$, (4) is an equality, therefore the bar-overlaps-bar test always identifies the ordering as significant when a z-test would, i.e. it has maximal power. When $e_x = e_y$, then $(e_x + e_y) = \sqrt{2}\sqrt{e_x^2 + e_y^2}$, so the bars overlap until $z > \sqrt{2}c$, which is very conservative.

Since $e_x \approx e_y$ is a common situation, it is difficult to recommend this test in general.

3 Reverse bar-overlaps-bar test

The poor performance of the bar-overlaps-bar test can be salvaged by applying it a reverse direction. That is, we use it to determine if $z < c$ (lack of significance) instead of $z > c$. In this direction, the test has high power.

Theorem If the error bars do overlap, then the ordering lacks significance at all levels exceeding $\phi(c\sqrt{2})$.

Proof If the error bars do overlap, then

$$\hat{y} - \hat{x} < c(e_x + e_y) < c\sqrt{2}\sqrt{e_x^2 + e_y^2} \tag{6}$$

because

$$e_x + e_y \leq \sqrt{2}\sqrt{e_x^2 + e_y^2} \tag{7}$$

(This is easy to check by squaring both sides.) Therefore

$$z = \frac{\hat{y} - \hat{x}}{\sqrt{e_x^2 + e_y^2}} \quad \leq \quad \frac{\hat{y} - \hat{x}}{(e_x + e_y)/\sqrt{2}} < c\sqrt{2} \tag{8}$$

For example, if the error bars have length ce_x and ce_y where $c = 1.64/\sqrt{2} = 1.16$ then when the bars overlap $z < 1.64$ and the ordering lacks significance at the 95% level.

Theorem The reverse bar-overlaps-bar test has greatest power when $e_x = e_y$, and least power when $e_x = 0$ or $e_y = 0$.

Proof When $e_x = e_y$, (7) is an equality, therefore the reverse bar-overlaps-bar test always identifies the ordering as lacking significance when a z-test would. When $e_x = 0$ or $e_y = 0$, then $e_x + e_y = \sqrt{e_x^2 + e_y^2}$, so the bars don't overlap until $z < c$, which is very conservative.

4 Bar-overlaps-point test

In a bar-overlaps-point test, you check if $\hat{x}$ is outside y's error bar *and* $\hat{y}$ is outside x's error bar. Let the error bars be $\hat{x} \pm ce_x$ and $\hat{y} \pm ce_y$ for some z.

Theorem If the bar-overlaps-point test is satisfied, the ordering is significant at a level exceeding $\phi(c/\sqrt{2})$.

Proof The bar-overlaps-point test is satisfied when $\hat{y} - \hat{x} > ce_x$ and $> ce_y$, or equivalently

$$\begin{aligned} \hat{y} - \hat{x} &> c \max(e_x, e_y) \qquad (9) \\ &> c\sqrt{e_x^2 + e_y^2}/\sqrt{2} \qquad (10) \end{aligned}$$

because

$$\max(e_x, e_y) > \sqrt{e_x^2 + e_y^2}/\sqrt{2} \tag{11}$$

(This is easy to check by squaring both sides.) Therefore

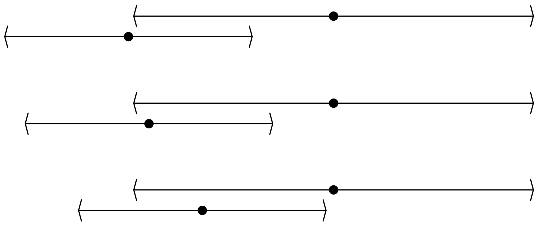
$$z = \frac{\hat{y} - \hat{x}}{\sqrt{e_x^2 + e_y^2}} \quad \geq \quad \frac{\hat{y} - \hat{x}}{\sqrt{2}\max(e_x, e_y)} > c/\sqrt{2} \tag{12}$$

For example, if $c = 1.64\sqrt{2} = 2.32$ then $z \geq 1.64$ and the test has $\geq 95\%$ significance.

Theorem The bar-overlaps-point test has greatest power when $e_x = e_y$ and least power when $e_x = 0$ or $e_y = 0$.

Proof When $e_x = e_y$, (11) is an equality, therefore the bar-overlaps-point test always identifies the ordering as significant when a z-test would, i.e. it has maximal power. When $e_x = 0$ or $e_y = 0$, then $\max(e_x, e_y) = \sqrt{e_x^2 + e_y^2}$, so the bars overlap a point until $z > c$, which is very conservative.

Since $e_x \approx e_y$ is a common situation, this recommends the bar-overlaps-point test over bar-overlaps-bar. Figure 1 illustrates the increased power of the bar-overlaps-point test. The bar-overlaps-point test can also be reversed like the bar-overlaps-bar test, but this is less useful.



Significant $(z = 2.02)$

Not significant? $(z = 1.82)$

Not significant $(z = 1.30)$

Figure 2: The power of the bar-overlaps-point test when $e_x = 1.3$ and $e_y = 2.1$. Similar to figure 1 except the bars are $\sqrt{2}$ times longer. This test is less conservative in declaring an ordering significant.

5 Combination test

Because the bar-overlaps-bar and bar-overlaps-point tests are strong in different situations, we can increase their power by combining them. Draw the usual set of error bars of length c, plus outer bars of length $\sqrt{2}c$. The ordering between $\hat{x}$ and $\hat{y}$ is significant if *either* the inner bars do not overlap *or* none of the outer bars overlap the other estimate. This test is less conservative, i.e. has more power, than either of the individual tests, and by the arguments above has significance level $\phi(c)$. Furthermore, according to the last section, a bar-overlaps-point test can be used on the inner bars to determining if the ordering lacks significance. See figure 3 for examples.

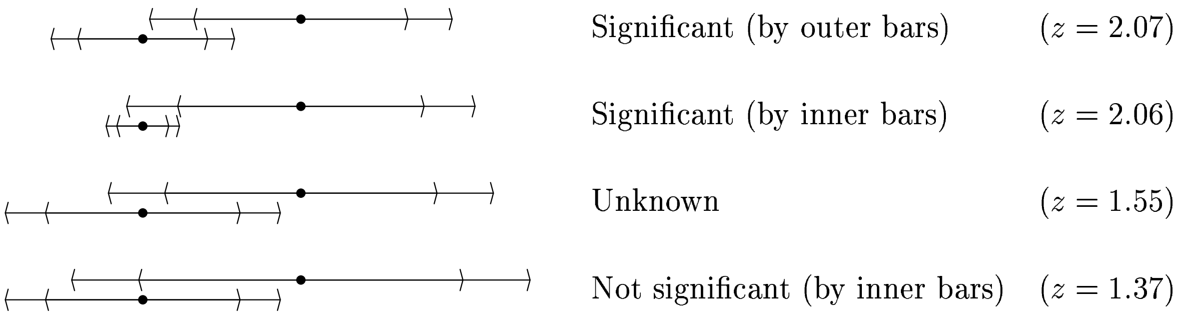

Significant (by outer bars) $(z = 2.07)$
Significant (by inner bars) $(z = 2.06)$
Unknown $(z = 1.55)$
Not significant (by inner bars) $(z = 1.37)$

Figure 3: A combination bar-overlaps-bar and bar-overlaps-point test. The combination catches more significant cases than either individual test.

6 A graphical z-test

So far we have been talking about graphical tests which approximate a z-test. But it is also possible to a make an exact graphical z-test. The idea is to draw an error *circle* instead of an error bar.

Draw a circle of radius ce_x centered on $\hat{x}$, and a circle of radius ce_y centered on $\hat{y}$. If the circles don't overlap, then the ordering is significant at level $\geq \phi(c)$ (the usual bar-overlaps-bar test). If one estimate is contained in the other circle, then the ordering is not significant (a reverse bar-overlaps-point test).

When neither is true, things get interesting. The circles intersect at two symmetric points. Call one of these points P. Consider the triangle $(\hat{x}, \hat{y}, P)$, whose sides are $(\hat{y} - \hat{x}, ce_x, ce_y)$. If the angle at P is 90°, then by Pythagoras $\hat{y} - \hat{x} = c\sqrt{e_x^2 + e_y^2}$ so $z = c$. If the angle is greater than 90°, then $z > c$, and if the angle is less than 90°, then $z < c$. So we have an exact z-test.

To judge the angle, it is sufficient to look at how the circles meet at their intersection point. If they form a symmetric "X", then the angle is 90°. If the "X" is elongated in the direction of $\hat{x} - \hat{y}$, then the angle is less than 90° (not significant). If the "X" is narrow in the direction of $\hat{x} - \hat{y}$, then the angle is greater than 90° (significant). Figure 4 illustrates the idea. Note that in the first example, a bar-overlaps-bar test would fail to identify the ordering as significant. You might argue that this method makes it difficult to discriminate cases near the borderline, but this could also be construed as a *feature* of this method over error bar tests, since it is fairly pointless to discriminate 94% significance from 95% significance.

References

Schmid, C. F. (1983). *Statistical graphics: Design principles and practices*, chapter 10. John Wiley & Sons.

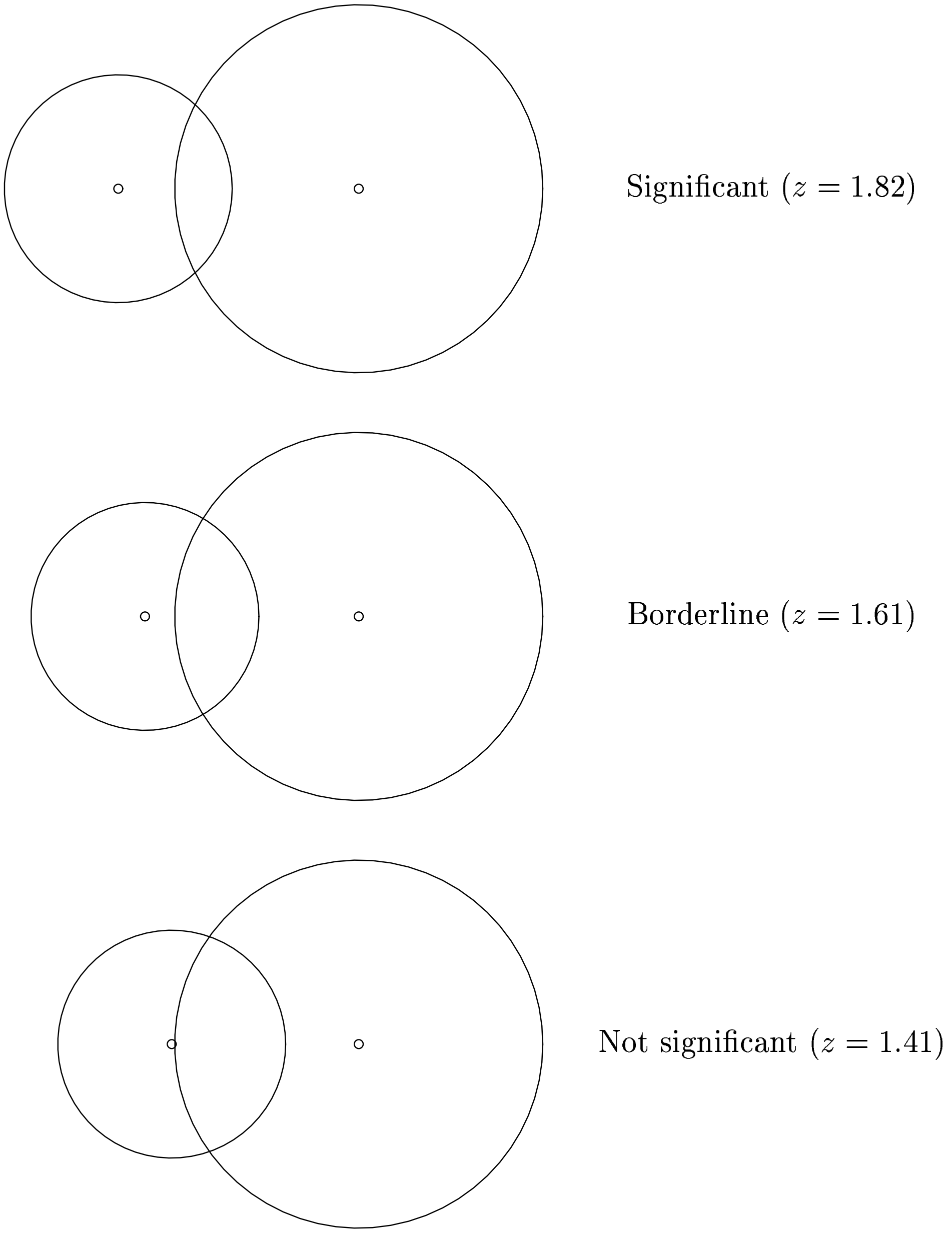


Figure 4: A graphical z-test