

Some Problems with Confidence Intervals

Juliet Popper Shaffer

March 21, 2006

- Invited paper “Confidence intervals on subsets may be misleading”
- In online (free) Journal of Modern Applied Statistical Methods (2004), Vol. 3, No. 2, 261-270.
- <http://tbf.coe.wayne.edu/jmasm>
- Click on an issue to download the whole issue (wait until downloaded before scrolling down).

References and Errata

- I found some errors in the paper, and have added an errata on the back of the reference sheet, in case you look at the paper.

- The idea behind the talk is that although a confidence interval may have the correct overall coverage (i.e. 95% confidence intervals cover the true value 95% of the time), if you look at confidence intervals only for selected results, that coverage no longer holds.

Outline

- The relation between significance and confidence interval coverage
- The difference between overall confidence and confidence for selected subsets
- Relation between true coverage and effect size
- Relation between true coverage and length of confidence interval
- Different publication policies and effects on confidence coverage
- If there is time and interest, I'll describe another approach.

The model

- I'm going to talk about problems with confidence intervals in general, but I'm going to illustrate them quantitatively using a very simple model: A comparison of the means of two distributions, where the observations are assumed normal with equal variance. We test the hypothesis that the difference between the means is zero (any other value would work as well) and/or we calculate a confidence interval for the difference.

Significance and confidence

- For simplicity, assume equal sample size n for each group. If we're given a confidence interval, it's easy to convert it to an estimated mean difference and a p -value, and vice versa.

Confidence interval and significance test

I'll use δ for the difference $\mu_1 - \mu_2$,

D for its estimate $M_1 - M_2$,

σ_D for the standard deviation $\sigma\sqrt{(2/n)}$ of D , and

s_D for the estimated standard deviation $s\sqrt{(2/n)}$ of D ,
where s is the unbiased estimate of σ .

- The standard confidence interval is
- $D \pm t(2n-2, .05) s_D$, where $t(2n-2, .05)$ is the .05 critical value of t with $2n-2$ degrees of freedom.
- Suppose the 95 % confidence interval is 5 to 11, or 8 ± 3 ,
- If $n = 10$, the .05 level critical t with 18 df is 2.10.
- Then d is 8 and $s_D = 3 / t(2*n-2, .05) = 3/2.10 = 1.43$.
- The observed t is $D/s_D = 8/1.43 = 5.59$.
- The p -value is the probability of obtaining a value of $t \geq$ the observed value which is smaller than .0001.

- So the (sample mean and p-value) and the $(1-\alpha)$ confidence interval) give the same information, in different forms.
- It is often claimed that the confidence interval form is more useful, but that can be debated.

- The advice often given is to first test the difference for significance, and if it is significant, calculate a confidence interval.
- Equivalently, we can calculate the confidence interval but consider it only if it doesn't include zero.
- Then what is the probability that the interval covers the true value *given that the hypothesis has been rejected?*

Conditional probability

If an interval includes the true value with probability .95 we can write this as:

The probability that the interval covers the true value given rejection, times the probability of rejection, plus the probability that the interval covers the true value given acceptance, times the probability of acceptance, equals .95.

The implications of this equation will be described.

Variation in confidence coverage:

Conditional coverage

- Suppose we test the hypothesis that the difference between the means is zero at the 5 % level. Then there is a 5% chance of concluding the means are different, i.e. making a Type I error. If we consider a 95% confidence interval in that 5% of cases, the probability that the confidence interval covers the true value (which is zero) *in those cases* is zero.

- In testing, we either accept or reject a hypothesis at a given level. If the true value is not zero and we reject, we've made a correct decision. But confidence interval coverage varies depending on the true value if the interval is considered only when the hypothesis is rejected.
- Intuition: If a confidence interval does not cover the true value when that value is zero, it is less likely to cover the true value if it is very close to zero.

- Thus, if we pay special attention to confidence intervals that don't include zero, these may have confidence coverage less than the nominal $1-\alpha$.
- On the other hand, the confidence intervals that do include zero are likely to have coverage probability greater than $1-\alpha$.
- From now on, I'll assume $\alpha = .05$, although the qualitative issues are completely general.

- Some history:
- There was some earlier literature, but the most relevant is a 1973 article by Olshen.
- It deals with the Scheffé method of multiple comparisons using the F distribution, but applies in this simpler case.
- Olshen's talk.

- In 1977 Scheffé answered Olshen's criticism, Olshen commented and Scheffé gave a rejoinder.

- We can't know the exact coverage probability, because it depends on both the sample size and the effect size, and the effect size is unknown, but we can consider some typical situations and see what we might expect.

Coverage probability in relation to effect size

- I'll use e.s. for effect size, which equals
- δ/σ and might be estimated by d/s .

The coverage probabilities of the confidence intervals that do or do not include zero depend on the noncentrality parameter, which equals $(e.s)(\sqrt{(n/2)})$.

- The following graph shows the true confidence levels of those intervals that include zero (acceptance of null) and those that don't include zero (rejection of null). This graph assumes σ is known. That doesn't make much difference unless degrees of freedom are small, in which case another issue comes up.

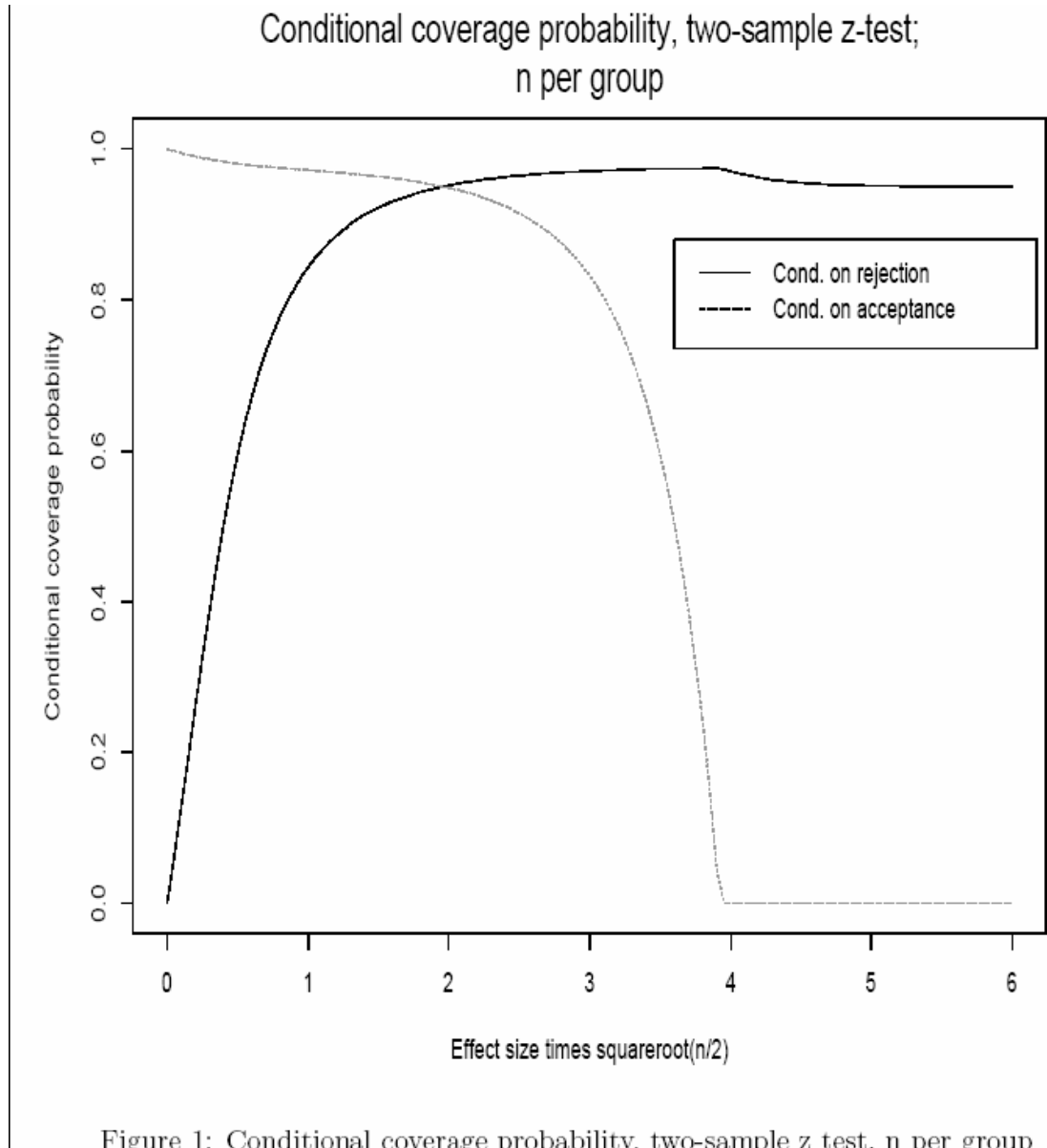


Figure 1: Conditional coverage probability, two-sample z test, n per group

- What effect might this difference in coverage probability have on typical studies in education and social sciences?
- Cohen (1962) gave as guidance for typical effect sizes in social-behavioral research the quantities .2 (small), .5 (medium) and .8 (large), and stated
- “Many effects sought in personality, social, and clinical-psychological research are likely to be small effects as here defined...”
- Certainly that’s true in educational research.

- I just casually picked out a couple of educational studies with many reported effect sizes, and found many smaller than .2, a few between .2 and .3, and very few greater than .4. Furthermore, these were meta-analyses, and undoubtedly missed some studies, so the reported effect sizes were probably overestimates.

- The following table gives some combinations of sample sizes (per group) and effect sizes with probabilities of rejecting hypotheses (i.e. power) and true confidence interval coverage probabilities, when nominal .95 level confidence intervals are reported.

Sample size	Effect size				
	.1	.2	.3	.4	.5
5	.20(.05)	.41(.06)	.57 (.07)	.69(.10)	.77 (.12)
10	.29(.05)	.55(.07)	.72(.10)	.81(.15)	.87(.20)
20	.41(.06)	.69(.09)	.83(.16)	.90(.24)	.93(.35)
30	.49(.07)	.77(.11)	.88(.21)	.93(.34)	.95(.48)
40	.55(.07)	.81(.14)	.91(.26)	.94(.43)	.96(.60)
50	.60(.08)	.84(.17)	.92(.32)	.95(.51)	.96(.70)

- When the nominal probabilities don't cover the true difference, they are likely to cover larger differences for the sample size-effect size combinations in the table. Thus, they are likely to give overoptimistic ideas of the size of the difference.

True coverage and length of interval

- Another complicating factor becomes somewhat important with small sample sizes, involving a different type of conditional coverage.
- Results so far have been under the assumption that the true standard error of the difference, σ_D is known.

- With known σ_D the length of the confidence interval is fixed; for a 95% interval it is
- $2 (1.96) (\sigma_D)$. However, usually the standard error has to be estimated, so the interval length then is $2 (t(2n-2, .05)) (s_D)$.
- Since s_D is a random variable, the length of the interval is random.

- With small sample sizes, s_D can be quite variable, so the interval lengths vary considerably.
- The true coverage of the confidence intervals depends on (i.e. is conditional on) the length: the shorter the interval, the less likely it is to contain the true value.

- There's no way we can have any idea whether we've obtained an unusually small or unusually large estimated standard error, so there's no way of evaluating this possible effect.
- I've looked into the magnitude of the effect, and fortunately it is relatively small unless the sample sizes are smaller than the usual ones in practice.

Publication policies and confidence interval coverage

- All the effects we've covered are operating even in the ideal case in which all results are reported. But we know that is often not the case.
- Papers with statistical significance are more likely to be published.
- Some companies suppress results that are contrary to their interests.

Publication policies and confidence coverage

- It's now well known that some companies have suppressed experimental results that cast doubt on their products. There are two types of suppression that may occur.
- 1. Studies with significant results in a direction bad for their product are suppressed.
- 2. Studies that do not significantly support their product are suppressed.

- 1. Either neutral or favorable studies are reported, while unfavorable studies are suppressed.
- If the product really is neutral or favorable, the confidence intervals will have at least the nominal coverage, varying with the noncentrality parameter.
- If the product really is unfavorable, coverage probability again varies with the noncentrality parameter, and is given in a table in the paper.

- 2. Only significantly favorable studies are reported.
- If the product really is favorable, the coverage probabilities will depend on the noncentrality parameter, and will be somewhat higher than in the table I showed before.
- If the product really is unfavorable, the coverage probability will be zero.

Summary so far

- Given that model assumptions are satisfied, confidence intervals with nominal $1 - \alpha$ probability of including the true parameter value will have that probability. But if interest is confined to confidence intervals that meet some criterion depending on the observed data, that nominal coverage no longer applies.

- If the set of confidence intervals that don't include the hypothesized value (usually zero) are the only ones calculated or considered, the true confidence coverage can be substantially less than the nominal coverage.

- There is no obvious solution to this problem, but sensitivity checks might give some idea of how bad it could be.
- A different approach to the subject is described in a 2005 article (with discussion) by Benjamini and Yekutieli.

A different approach

- Instead of considering the conditional coverage probability, consider the *unconditional noncoverage* probability; i.e. the probability of making an error. We make an error, in this formulation, if we calculate the confidence interval and it doesn't include the true value.

- Suppose the null hypothesis is true. Then if we calculate the confidence interval only if the result is significant, the probability that we calculate it is .05. The probability that the interval is wrong (doesn't include the true value) in that case is 1. The probability of calculating the interval and making a mistake is therefore $.05 \times 1 = .05$.

- When the null hypothesis is not true, the probability of rejecting it increases but the probability of making an error (a confidence interval that doesn't include the true value), given that rejection, decreases. Benjamini and Yekutieli proved that the product is always smaller than the significance level, .05 in our examples.

- Benjamini and Yekutieli apply this result and extensions in a multiple testing context in this 2005 journal article. It's more meaningful in that context. There are many discussants, and the article gives some guidance on how to use the authors' approach in a multiple testing context.