Noncentrality Based Interval Estimation

James H. Steiger

Department of Psychology and Human Development
Vanderbilt University

Noncentrality Based Interval Estimation

1. Introduction
2. Problems with Significance Testing
   - Introduction
   - Two Kinds of Significance Testing
   - Reject-Support Testing
   - Accept-Support Testing
   - Summing Up
3. Some Typical Problems
   - Misleading p-values
   - Illogical Model Testing
4. The Value of Interval Estimates
   - The Right Answer to the Right Question
5. Reasons Why Confidence Intervals Were Reported Infrequently
6. Methods for Interval Estimation
   - Introduction
   - Basic Notation for Interval Estimation
   - Traditional Confidence Intervals
   - A More General Approach
   - Obtaining a Confidence Interval for δ
   - Graphical Construction of a Confidence Interval on δ
   - The Inversion Confidence Interval Principle
   - Confidence Intervals on E_s
7. Speeding Things Up with Software
8. Some Caveats

Introduction

In this module, we review some general principles of confidence interval estimation.

We take the position that confidence intervals generally provide more of the kind of statistical information that social scientists are looking for in their data analyses.

Social science textbooks present only a narrow account of what is available in confidence interval estimation methods.

We demonstrate how the virtues of confidence interval estimation can be extended beyond simple univariate statistics to more complex methods such as ANOVA, multiple regression, structural equation modeling, and multivariate analysis.

Problems with Significance Testing

Introduction

We begin by returning briefly to first principles.

We argue hat significance tests, though almost always reported in the analysis of social science data, are seldom to be preferred, and often simply inappropriate.

Problems with Significance Testing

Two Kinds of Significance Testing

We are performing a simple two-group experiment in which an experimental group is compared to an independently sampled control group.

The theoretical question of interest is frequently phrased as, "Has the experimental treatment made any difference?"

Problems with Significance Testing

Two Kinds of Significance Testing

In this case, the statistical null and alternative hypotheses are

$$H_0: \ \mu_1 = \mu_2 \qquad H_1: \ \mu_1 \neq \mu_2.$$

We test the hypothesis with the standard 2-sample t-test.

If the t is large enough in absolute value, we reject H_0.

Otherwise, loosely speaking, we "accept" H_0.

The possibilities are summarized in the classic 2×2 table.

Problems with Significance Testing

Two Kinds of Significance Testing

2×2 Table for Statistical Decisions

		State of the World	
		H_0	H_1
Decision	H_0	Correct Acceptance	Type II Error β
	H_1	Type I Error α	Correct Rejection

There are two kinds of error, with probabilities α and β.

There is a trade-off between the two probabilities.

The *substantive meaning* of the two kinds of error changes, depending on the kind of statistical testing, so of course, the meaning of the trade-off also changes.

In the context of significance testing, we can define two basic kinds of situations, reject-support (RS) and accept-support (AS).

Problems with Significance Testing

Two Kinds of Significance Testing

In RS testing, the null hypothesis is the opposite of what the researcher actually believes, and rejecting it supports the researcher's theory.

In a two group RS experiment, the experimenter believes the treatment has an effect, and seeks to confirm it through a significance test that rejects the null hypothesis.

In the RS situation, a Type I error represents, in a sense, a "false positive" for the researcher's theory.

From society's standpoint, such false positives are particularly undesirable.

Problems with Significance Testing

Reject-Support Testing

In RS testing, a Type II error is a tragedy from the researcher's standpoint, because a theory that is true is, by mistake, not confirmed.

So, for example, if a drug designed to improve a medical condition is found (incorrectly) not to produce an improvement relative to a control group,

1. A worthwhile therapy will be lost, at least temporarily, and
2. The experimenter's worthwhile idea will be discounted.

As a consequence, in RS testing, society, in the person of journal editors and reviewers, insists on keeping α low.

The statistically well-informed researcher makes it a top priority to keep β low (and power high) as well.

Ultimately, of course, everyone benefits if both error probabilities are kept low, but unfortunately there is usually, in practice, a nontrivial trade-off between the two types of error.

Problems with Significance Testing

Accept-Support Testing

In AS testing, H_0 is what the researcher actually believes, so accepting it supports the researcher's theory.

In this case, a Type I error is a false negative for the researcher's theory, and a Type II error constitutes a false positive.

Consequently, maintaining a very low Type I error rate like .001, is actually "stacking the deck" in favor of the researcher's theory in AS testing.

Problems with Significance Testing

Summing Up

To summarize, in RS research:

1. The researcher wants to reject H_0.
2. Society wants to control Type I error.
3. The researcher must be very concerned about Type II error.
4. High sample size works for the researcher.
5. If there is "too much power," trivial effects become "highly significant."

In AS research, on the other hand:

1. The researcher wants to accept H_0.
2. "Society" should be worrying primarily about controlling Type II error, although it sometimes gets confused and retains the conventions applicable to RS testing.
3. The researcher must be very careful to control Type I error.
4. High sample size works against the researcher.
5. If there is "too much power," the researcher's theory can be "rejected" by a significance test even though it fits the data *almost* perfectly.

Some Typical Problems

Misleading *p*-values

In response to the inherent dissatisfaction with the dichotomous outcomes inherent in Neyman-Pearson hypothesis testing, researchers have looked to *p*-values as a solution.

Probability levels can deceive about the "strength" of a result, especially when presented without supporting information.

For example, a p level of .075 could represent a powerful effect operating with a small sample, or a tiny effect with a huge sample.

Clearly then, we need to be careful when comparing p levels.

Some Typical Problems

Illogical Model Testing

Model testing often involves computation of a "badness of fit" test, often mislabeled as a test of "goodness of fit."

Rejecting the null hypothesis rejects the model, which, typically, the experimenter wants to accept.

Rejection of an "almost true" null hypothesis in such situations frequently has been followed by vague, convoluted statements that the rejection shouldn't be taken too seriously.

On occasion, the testing has been performed with such low precision that accepting the model was inevitable — yet few noticed.

On other occasions, the testing was performed with such high precision that virtually *any* departure from perfect fit would be detected.

The Value of Interval Estimates

The Right Answer to the Right Question

Much psychological research is exploratory. The fundamental questions we are usually asking are:

1. What is our best guess for the size of the population effect?
2. How precisely have we determined the population effect size from our sample data?

Significance testing fails to answer these questions directly.

The Value of Interval Estimates

The Right Answer to the Right Question

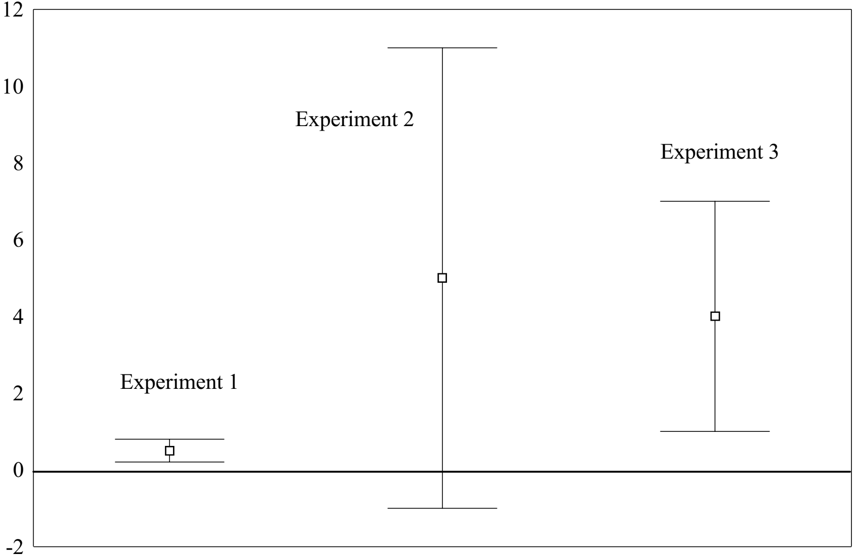
Confidence interval estimation provides a convenient alternative to significance testing in most situations.

In the 2-sample t-test of equal means, if the hypothesis test is 2-sided, the hypothesis is rejected if and only if the confidence interval excludes zero.

Confidence intervals tell us *whether* H_0 was rejected, but also give us a much better idea of *why* it was or wasn't rejected.

On the next slide, we see 3 confidence intervals from 3 experiments. Contrast what the hypothesis test result might convey with what the confidence intervals clearly tell us.

The Right Answer to the Right Question



Reasons Why Confidence Intervals Were Reported Infrequently

Tradition. Traditional approaches to psychological statistics emphasize significance testing much more than interval estimation.

Pragmatism. In RS situations, interval estimates are sometimes embarrassing. When they are narrow but close to zero, they suggest that a "highly significant" result may be statistically significant but trivial. When they are wide, they betray a lack of experimental precision.

Reasons Why Confidence Intervals Were Reported Infrequently

Ignorance. In spite of the enhanced emphasis on confidence interval estimation by journal editors and scientific societies, many people are simply unaware of some of the very valuable interval estimation procedures that are available. For example, the vast majority of psychologists are simply not aware that it is possible to compute a confidence interval on the squared multiple correlation coefficient. The procedure is not discussed in many standard texts, and it is not implemented in major statistical packages.

Lack of availability. Some of the most desirable interval estimation procedures are computer intensive, and are not implemented in major statistical packages like SAS, SPSS, STATISTICA, and so on. This makes it less likely that anyone relying on such software will try the procedure.

Methods for Interval Estimation

Introduction

In this section, we begin by reviewing the basic definition of a confidence interval, and the simple approach used to generate the simple confidence intervals found in most textbooks.

Then we describe the less conventional, more computer-intensive approach which allows much more interesting and useful intervals to be derived.

Here the discussion becomes somewhat more technical, and we employ notations that are common in mathematical statistics texts, but that the typical reader with a basic background in introductory applied statistics texts may find slightly intimidating.

We try to strike a balance that provides sufficient, but not extraneous, detail.

Methods for Interval Estimation

Basic Notation for Interval Estimation

X is a sample of n independent observations from some population.

$A(X)$ is a statistic calculated on X.

θ is a parameter to be estimated.

Methods for Interval Estimation

Basic Notation for Interval Estimation

A *lower confidence limit* (or *lower confidence bound*) is a statistic that is less than or equal to the unknown parameter a certain proportion of the time. A function $A(X)$ of the observed data X is a $1-\alpha$ lower confidence limit for θ if, over repeated samples,

$$\Pr(A(X) \leq \theta) = 1 - \alpha \tag{1}$$

In a similar vein, A function $B(X)$ of the observed data X is a $1-\alpha$ *upper confidence limit* or *upper confidence bound* for θ if, over repeated samples,

$$\Pr(B(X) \geq \theta) = 1 - \alpha \tag{2}$$

Methods for Interval Estimation

Basic Notation for Interval Estimation

Upper and lower confidence limits usually are combined to yield a *confidence interval* $(A(X), B(X))$, whose endpoints surround the parameter θ a certain proportion of the time. We say that $A(X)$ and $B(X)$ bound a $1-\alpha$ *confidence interval* for θ if

$$\Pr(A(X) \leq \theta \leq B(X)) = 1 - \alpha \tag{3}$$

In practice, one usually constructs the confidence interval by choosing $A(X)$ and $B(X)$ to be, respectively, lower and upper $1-\alpha/2$ confidence limits so that the confidence interval is equally likely to be entirely below or above θ.

Methods for Interval Estimation

Traditional Confidence Intervals

Traditional confidence intervals given in most introductory and intermediate texts rely on manipulation of a simple probability statement.

So, for example, we start with a statement about the sampling distribution of the sample mean, i.e.,

$$\Pr\left(-1.96 \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq +1.96\right) = .95$$

After simple manipulation, we obtain

$$\Pr\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right) = .95$$

Methods for Interval Estimation

Traditional Confidence Intervals

A number of simple inequalities can be converted into confidence intervals in this way. Typically, one finds, in elementary to intermediate texts, confidence intervals for:

1. A single mean
2. The difference between two means
3. A single contrast on means
4. A single variance
5. The ratio of two variances
6. A single correlation
7. A single proportion

An element common to the preceding intervals is that an interval statement about the distribution of the null distribution of a test statistic can be manipulated easily to yield the desired confidence interval.

Methods for Interval Estimation

Traditional Confidence Intervals

Situations where

1. The distribution of the test statistic changes as a function of the parameter to be estimated, and
2. Simple interval manipulation does not yield a convenient confidence interval

are generally not discussed.

Methods for Interval Estimation

Traditional Confidence Intervals

Example

Consider the sample squared multiple correlation, whose distribution changes as a function of the population squared multiple correlation. Confidence intervals for the population squared multiple correlation are very informative, yet are not discussed in most standard texts, because a single simple formula for the direct calculation of such an interval cannot be obtained in a manner analogous to the way we obtain a confidence interval for μ.

Methods for Interval Estimation

A More General Approach

A general method for confidence interval construction is available that includes the method discussed earlier as a special case, but also allows confidence limits and confidence intervals to be constructed when the aforementioned method cannot be applied.

This method combines two general principles, which we call the *confidence interval transformation principle* and the *inversion confidence interval principle*.

Methods for Interval Estimation

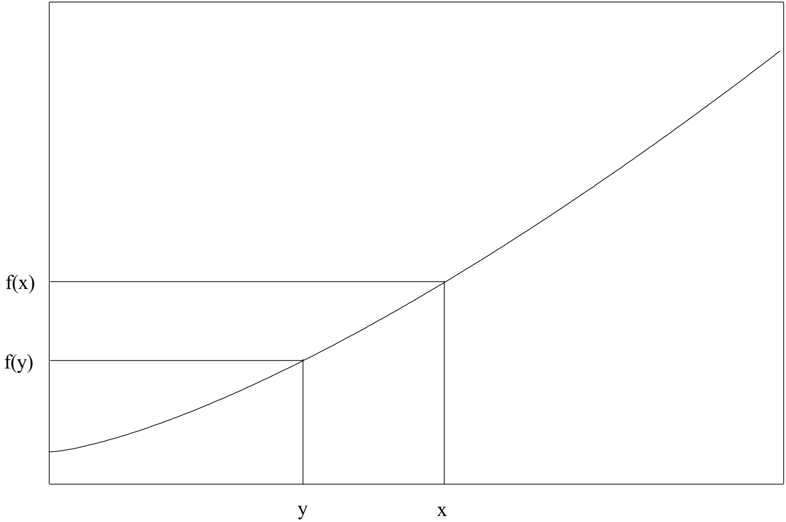
A More General Approach

Confidence Interval Transformation Principle. Let $f(\theta)$ be a monotonic, strictly increasing continuous function of θ. Let a and b be endpoints of a $1-\alpha$ confidence interval on quantity θ. Then $f(a)$ and $f(b)$ are endpoints of a $1-\alpha$ confidence interval on $f(\theta)$.

To prove the proposition, recall that a function is monotonic and strictly in-creasing if, when plotted in the plane, the graph "keeps going up" from left to right, that is, it never flattens out or goes down. A monotonic, strictly increasing function is order preserving. Because the plot never flattens out, if $x > y$, then $f(x) > f(y)$. This can be seen easily by examining the figure on the next slide.

Methods for Interval Estimation

A More General Approach



Methods for Interval Estimation

A More General Approach

If a and b are endpoints of a valid .95 confidence interval on quantity θ, then 95% of the time in the long run,θ is between a and b.

If $f()$ is a monotonic strictly increasing function, b is greater than θ, and θ is greater than a, then it must also be the case that $f(b) > f(\theta)$, and $f(\theta) > f(a)$.

Consequently, if a and b are endpoints of a $1 - \alpha$ confidence interval for parameter θ, then $f(a)$ and $f(b)$ are endpoints of a valid $1 - \alpha$ confidence interval on $f(\theta)$.

Methods for Interval Estimation

A More General Approach

Example (A Confidence Interval for σ)

Suppose you calculate a confidence interval for the population variance σ^2. Such a confidence interval is discussed in many elementary textbooks. You desire a confidence interval for σ. Confidence intervals for σ are seldom discussed in textbooks. However, one may be derived easily. Because σ takes on only nonnegative values, it is a monotonic increasing function of σ^2 over its domain. Hence, the confidence interval for σ is obtained by taking the square root of the endpoints for the corresponding confidence interval for σ^2.

Methods for Interval Estimation

A More General Approach

Example (Confidence Interval for ρ)

The confidence interval transformation principle is most valuable when you have a confidence interval for a quantity that is an uninteresting function of an interesting quantity. Suppose one calculates a confidence interval for $z(\rho)$, the Fisher transform of ρ, the population correlation coefficient. Taking the inverse Fisher transform of the endpoints of this interval will give a confidence interval for ρ. This is, in fact, the method employed to calculate the standard confidence interval for a correlation.

Methods for Interval Estimation

A More General Approach

Example (Confidence Interval for E_s)

Recall from the preceding lecture on the t distribution that the 1-sample t-test has a distribution that is noncentral Student t, with $n-1$ degrees of freedom, and a noncentrality parameter that is equal to

$$\delta = \sqrt{n}E_s = \sqrt{n}\frac{\mu - \mu_0}{\sigma}$$

Since E_s is a monotonic, strictly increasing function of δ, if we can obtain a confidence interval for δ, we can directly transform its endpoints into a confidence interval for E_s.

Methods for Interval Estimation

Obtaining a Confidence Interval for δ

The picture on the next slide contains the information that is key to understanding how to obtain a confidence interval on δ.

Suppose degrees of freedom for the t statistic are fixed at ν.

Consider any value of δ. With degrees of freedom fixed, it completely determines the distribution of t.

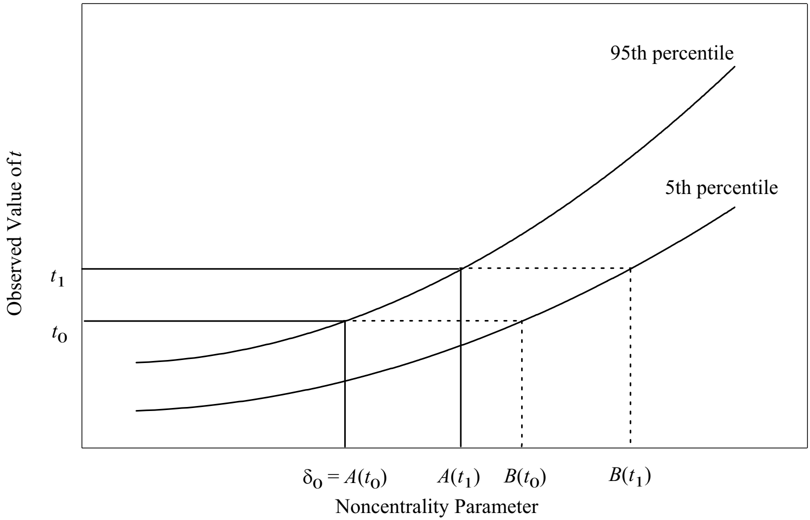
Among other things, δ determines the 95th and 5th percentiles of the distribution of t.

As δ goes up, the distribution of t goes to the right, and the 5th and 95th percentiles are monotonic functions of δ.

Notice that each value of δ is connected with one and only one value of P_{95}. *And*, each 95th percentile is connected with one and only one value of δ.

Methods for Interval Estimation

Obtaining a Confidence Interval for δ



Methods for Interval Estimation

Obtaining a Confidence Interval for δ

Suppose δ_0 is the true value of δ.

Looking at the figure, you can see that corresponding to δ_0 is a value t_0 that represents the 95th percentile of the t_{ν,δ_0} distribution.

What percentage of the time will a value of t exceed t_0? Since it is the 95th percentile, this will happen 5% of the time, i.e. a probability of 0.05.

t_1 is an example of such a value of t.

Methods for Interval Estimation

Obtaining a Confidence Interval for δ

Now here is the key. Notice that, for any *observed* value of t, we can draw a line over to the plot and choose a value of δ. For example, if we were to observe t_0, we would select δ_0.

If δ_0 is the true parameter, what percentage of the time would we index a value larger than δ_0 on the 95th percentile line? And what percentage of the time would we index a value smaller than δ_0?

So, if every time we observe a t, we select the corresponding value on the 95th percentile line, we have a rule for a 0.95 lower bound for δ!

In a similar way, the 5th percentile line selects a 0.95 upper bound for δ.

Taken together, these two bounds produce a 90% confidence interval for δ.

Methods for Interval Estimation

Obtaining a Confidence Interval for δ

In other words, when we observe a value of t, we ask two questions:

1. What value of δ is so small that it places the observed t at the 95th percentile?
2. What value of δ is so large that it places the observed t at the 5th percentile?

Taken together, these values constitute a "zone of reasonableness" for δ. More importantly, they are an exact 90% confidence interval for δ.

Methods for Interval Estimation

Graphical Construction of a Confidence Interval on δ

Suppose we observe a t value of 2.60 with $\nu = 18$.

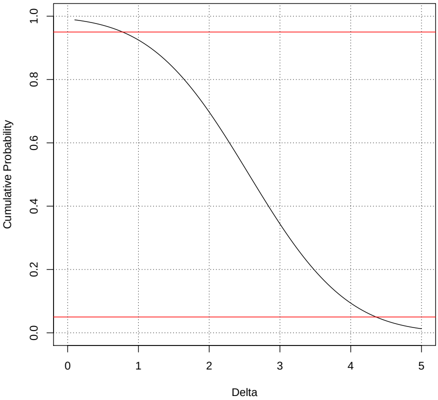
Below, using R, we show a graph of the percentile values of a value of 2.60 in the $t_{\nu,\delta}$ distribution as a function of δ.

We are looking for a value of δ that places the observed value of $t = 2.60$ precisely at the 95th and the 5th percentiles.

Methods for Interval Estimation

Graphical Construction of a Confidence Interval on δ

```
> curve(pt(2.6, 18, x), 0.1, 5, xlim = c(0, 5), ylim = c(0,
+     1), ylab = "Cumulative Probability", xlab = "Delta")
> abline(h = 0.05, col = "red")
> abline(h = 0.95, col = "red")
> grid(col = 1)
```

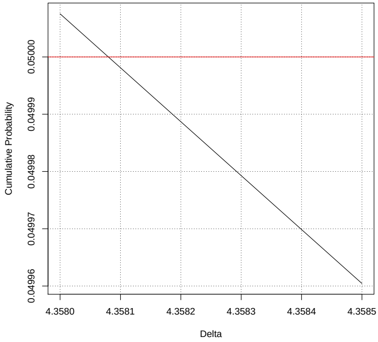


Methods for Interval Estimation

Graphical Construction of a Confidence Interval on δ

Quick Inspection of the plot reveals that the lower confidence limit is between 0.7 and 1.0, and the upper limit is between 4.2 and 4.5. Working on the upper limit, it takes only a minute to home in on the range of the graph a few times and arrive at the following picture, which pinpoints the value at around 4.3581.

```
> curve(pt(2.6, 18, x), 4.358, 4.3585, ylab = "Cumulative Probability",
+     xlab = "Delta")
> abline(h = 0.05, col = "red")
> grid(col = 1)
```



Following the same approach, we establish the lower limit for δ at 0.7768.

Methods for Interval Estimation

Graphical Construction of a Confidence Interval on δ

The identical approach we used here to find a confidence interval for δ in the noncentral t distribution may be used to find a confidence interval for the noncentrality parameter λ in the noncentral F or noncentral χ^2 distributions.

Such confidence intervals are seldom if ever interesting in and of themselves.

However, δ and λ are often monotonically related to statistical quantities of considerable interest.

Consequently, we may employ the confidence interval transformation principle to transform the uninteresting intervals to interesting ones.

Methods for Interval Estimation

The Inversion Confidence Interval Principle

Inversion Confidence Interval Principle. Let ν be the observed value of X, a random variable having a continuous (cumulative) probability distribution expressible in the form

$$F(\nu, \theta) = \Pr(X \leq \nu|\theta)$$

for some numerical parameter θ.

Let $F(\nu, \theta)$ be monotonic, and strictly decreasing in θ, for fixed values of ν. Let l_1 and l_2 be chosen so that

$$\Pr(X \leq \nu|\theta = l_1) = 1 - \alpha/2$$

and

$$\Pr(X \geq \nu|\theta = l_2) = 1 - \alpha/2$$

Then l_1 is a lower $1 - \alpha/2$ confidence limit for θ, l_2 is an upper $1 - \alpha_2$ confidence limit for θ, and the interval with l_1 and l_2 as endpoints is a $1 - \alpha$ confidence interval for θ.

Methods for Interval Estimation

Confidence Intervals on E_s

If this were a test on a single mean, with $n = 19$, we know from the preceding module that $\delta = \sqrt{n}E_s$ tem Using the confidence interval transformation principle, we divide the endpoints of $(0.7768, 4.3581)$ by $\sqrt{19}$, obtaining

```
> c(0.7768, 4.3581)/sqrt(19)

[1] 0.1782 0.9998
```

Our 90% confidence interval for E_s ranges from 0.1782 to 0.9998.

In this case, $E_s = (\mu - \mu_0)/\sigma$ represents the amount by which the null hypothesis is wrong in standard deviations. We see that the confidence interval ranges from an effect that generally would be considered small to one that would almost certainly be considered large.

In practical terms, such a range is very substantial, illustrating that our experimental precision is quite low in this case.

Methods for Interval Estimation

Confidence Intervals on E_s

In the preceding example, we assumed that our t statistic was based on the analysis of a single mean, and we emerged with a confidence interval on $E_s = (\mu - \mu_0)/\sigma$.

Now suppose that our t statistic based on 18 degrees of freedom had instead come from a 2-sample, independent sample design, based on $n_1 = n_2 = 10$ observations from two independent groups, and that the null hypothesis was that $H_0 : \mu_1 = \mu_2$.

In this case, we saw in our lecture on the t distribution that the test statistic has a noncentral t distribution, with 18 degrees of freedom and noncentrality parameter given by

$$\begin{aligned} \delta &= \sqrt{\frac{n_1 n_2}{n_1 + n_2}} E_s \\ &= \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \frac{\mu_1 - \mu_2}{\sigma} \end{aligned} \tag{4}$$

Methods for Interval Estimation

Confidence Intervals on E_s

We see from Equation 4 above that we can obtain a confidence interval for E_s from the confidence interval on δ by multiplying the endpoints of the latter by $\sqrt{(n_1 + n_2)/(n_1 n_2)}$.

We obtain

```
> sqrt((10 + 10)/(10 * 10)) * c(0.7768, 4.3581)

[1] 0.3474 1.9490
```

The width of the confidence interval is extremely large. Clearly the effect is non-zero, but our precision is quite low with only 10 observations per group.

Speeding Things Up with Software

We have seen that, by simply employing R graphics and homing in on an appropriate range, we can deduce a confidence interval on δ to any desired level of precision.

Software is available to automate this process by iteration. For example, RDASA3 cites SAS (commercial software) and SPSS (commercial software) scripts to accomplish the task.

Statistica for Windows (commercial software) includes noncentrality interval estimation calculators for a variety of statistical tests including t-tests, ANOVA, multiple regression, and structural equation modeling.

I have a (freeware) little Windows calculator called NDC that performs various calculations on the noncentral t, noncentral F, and noncentral χ^2 distributions. You can download that from the software section of my website.

There is also a somewhat archaic program R2 that runs in a DOS window under windows and computes a wide variety of tests, confidence intervals, and power and sample size analyses for multiple regression.

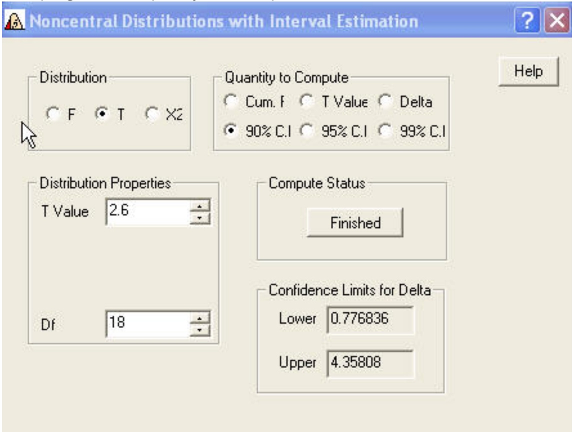
Ken Kelley at Notre Dame has written a very nice R statistics package called MBESS that automates a number of the calculations we have examined here.

Speeding Things Up with Software

In NDC, you simply select the quantity to compute, distribution, and the dialog box will adjust to ask you for the proper unknowns.

Clicking on *Compute* will iterate the answer.

The program will quickly iterate quantiles and confidence intervals.



Speeding Things Up with Software

Here are some sample calculations in MBESS.

There are some aspects of the notation in MBESS that might confuse you. Note that, throughout, Kelley labels the input statistic value with the variable name *ncp*, rather than t or F. Also, in the single sample case, he uses upper-case N for the sample size, while in the 2-sample case he uses n.1 and n.2.

```
> library(MBESS)
> # 90% ci on delta
> conf.limits.nct(2.6, 18, 0.9)

$Lower.Limit
[1] 0.7768

$Prob.Less.Lower
[1] 0.05

$Upper.Limit
[1] 4.358

$Prob.Greater.Upper
[1] 0.05

> # ... on a single standardized mean
> ci.sm(ncp = 2.6, N = 19, conf.level = 0.9)

[1] "The 0.9 confidence limits for the standardized mean are given as:"
$Lower.Conf.Limit.Standardized.Mean
[1] 0.1782

$Standardized.Mean
[1] 0.5965

$Upper.Conf.Limit.Standardized.Mean
[1] 0.9998

> # ... on a standardized mean difference
> ci.smd(ncp = 2.6, n.1 = 10, n.2 = 10, conf.level = 0.9)

$Lower.Conf.Limit.smd
[1] 0.3474

$smd
[1] 1.163

$Upper.Conf.Limit.smd
[1] 1.949
```

Some Caveats

Standardized effect size estimation based on confidence intervals on the noncentrality parameter offer a substantial improvement over the traditional t and F hypothesis test results.

They offer all the information in the hypothesis test result, and a lot more.

However, there are some immediate caveats that we should observe:

1. The same issues of robustness that pertain to the traditional hypothesis tests are also relevant in the interval estimation situation.
2. Generalization of a "standardized effect size" to the case of correlated sample or unequal variances is not straightforward.