DOCUMENT RESUME

ED 462 441 TM 033 705

AUTHOR	Capraro, Mary Margaret
TITLE	An Introduction to Confidence Intervals for Both Statistical Estimates and Effect Sizes.
PUB DATE	2002-02-00
NOTE	40p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (25th, Austin, TX, February 14-16, 2002).
PUB TYPE	Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Effect Size; *Estimation (Mathematics); Research Reports; Statistical Analysis
IDENTIFIERS	*Confidence Intervals (Statistics)

ABSTRACT

This paper summarizes methods of estimating confidence intervals, including classical intervals and intervals for effect sizes. The recent American Psychological Association (APA) Task Force on Statistical Inference report suggested that confidence intervals should always be reported, and the fifth edition of the APA "Publication Manual" (2001) said confidence intervals were "the best" reporting device. The correct use of confidence intervals uses intervals to compare results across prior studies and of prior studies with current studies. Confidence intervals provide a graphical tool to integrate or synthesize results across studies. They invoke two primary concepts, intervals and confidence levels. Intervals are determined by the standard errors of statistics, and confidence levels are chosen by the researcher and given as percentages. In this way, a range null hypothesis is tested rather than a point null hypothesis. New software has reduced the difficulty of establishing confidence intervals. Combining effect size with confidence intervals is the wave of the future in continuing efforts to make research understandable for the reader. Appended are esci (exploratory software for confidence intervals, La Trobe University, Australia)-created figures. (Contains 5 tables, 9 figures, and 30 references.) (Author/SLD)

Running head: CONFIDENCE INTERVALS

An Introduction to Confidence Intervals For Both Statistical Estimates and Effect Sizes

Mary Margaret Capraro

Texas A&M University

Mary Margaret Capraro, mmcapraro@coe.tamu.edu, is an Assistant Clinical Professor at Texas A&M University. Her research interests include teacher beliefs about mathematics, cultural influences on mathematics achievement, and reliability generalization meta-analytic studies. Correspondence can be sent to 4232 TAMU Department of Teaching, Learning, and Culture, College Station, TX 77843-4232

Paper presented at the 2002 Annual Meeting of the Southwest Educational Research Association, Austin, TX, February 14-16, 2002.

Abstract

The paper summarizes methods of estimating confidence intervals, including classical intervals and intervals for effect sizes. The recent APA Task Force on Statistical Inference report suggested that confidence intervals should always be reported, and the 5th edition of the APA Publication Manual (2001) said confidence intervals were "the best" reporting device.

An Introduction to Confidence Intervals

For Both Statistical Estimates and Effect Sizes

The APA Task Force on Statistical Inference recently published its recommendations (Wilkinson & APA Task Force on Statistical Inference, 1999). Among other recommendations, the Task Force suggested that:

> [Confidence] [i] nterval estimates should be given for <u>any</u> effect sizes involving principal outcomes. . . . Comparing confidence intervals from a current study to intervals from previous, related studies helps focus attention on stability across studies. . . . Collecting intervals across studies also helps in constructing plausible regions for population parameters. (p. 599, emphasis added)

The Task Force further stated that "It is hard to imagine a situation in which a dichotomous accept/reject decision is better than reporting an actual p value or, <u>better still</u>, a confidence interval" (Wilkinson & APA Task Force on Statistical Inference, 1999, p. 599).

And the fifth edition of the APA (2001) <u>Publication Manual</u> emphasized

> The reporting of confidence intervals... can be an extremely effective way of reporting results. Because confidence intervals combine information on location

> and precision and can often be directly used to infer significance levels, they are, in general, the best reporting strategy. The use of confidence intervals is therefore strongly recommended. (p. 22, emphasis added)

But confidence intervals (CIs) may be poorly understood, in part because they are so infrequently used. Finch, Cumming, and Thomason (2001) reviewed 60 years of reporting practices in the *Journal of Applied Psychology*. Of the 150 articles studied, only four contained confidence intervals and two used visual displays to report their data. Finch et al. (2001) were not deluded into thinking that CIs would cure all the woes of statistical reporting because even three of these four researchers mentioning CIs in their results failed to use CIs wisely in interpreting their results. Substantive interpretation was used in only one of the four articles. Finch et al. (2001) disappointingly concluded "many important aspects of inference practices and reporting were the same in 1999 as a half century earlier" p. 204. In the same vain, Kieffer, Reese and Thompson (2001) reviewed 756 articles published in *American Education Research Journal* and *Journal of Counseling Psychology* from 1988 to 1997 and found only one article which reported confidence intervals.

Some people wrongly say that confidence intervals involve nothing more than null hypothesis significance tests (NHST) in a different form (cf. Hagen, 1997; Knapp & Sawilowsky, 2001). This is a fairly mindless use of an otherwise powerful analytic tool (Thompson, 2001). As Thompson (1998) explained,

> If we mindlessly interpret a confidence interval with reference to whether the interval subsumes zero, we are doing little more than nil hypothesis statistical testing. But if we interpret the confidence intervals in our study in the context of the intervals in all related previous studies, the true population parameters will eventually be estimated across studies, even if our prior expectations regarding the parameters are widely wrong (Schmidt, 1996). (p. 799)

Thus, the correct use of CIs uses intervals to compare results across prior studies, and of prior studies with current studies (Fidler & Thompson, 2001). This is exactly the particular application that facilitates the "meta-analytic thinking" so critical to informed research practice (Cumming & Finch, 2001). Cahn (2000) stated that "statistical significance is not a 'kosher certificate' for observed effects" (p. 33) and recommended a two-step approach that includes computing confidence intervals when evaluating empirical results. He went on to state that the editorial policies of journals can offer an impetus toward requiring researchers to use CIs rather than just statistical significance when reporting their results. As Finch, et al. (2001) observed,

> Editors of many journal acting in concert would be more likely to achieve substantial change. . . . Everyone (writers of statistics texts and software, statistics teachers, researchers themselves, journal editors, and manuscript reviewers, and participants in APA and other policy-making bodies) . . . . needs to take the responsibility for promoting change. (pp. 206-207)

Similarly, Caruso and Cliff (1997) advocated the use of confidence intervals and moving away from hypothesis testing. Caruso related an experience where he was working with extremely large data sets and thus finding all of his results to be statistically significant. This experience convinced this Caruso that reporting confidence intervals definitely make the results more enlightening and interesting to the reader.

One advantage of more thoughtful use of confidence intervals is that *they provide a graphical tool to integrate* or *synthesize results across studies*. This is important, because such comparisons evaluate result replicability. Too few researchers attend to replicability issues, because many researchers misunderstand what statistical significance tests do (cf. Nelson, Rosenthal & Rosnow, 1986; Oakes, 1986; Rosenthal & Gaito, 1963; Zuckerman, Hodgins, Zuckerman & Rosenthal, 1993), and incorrectly believe that statistical significance tests evaluate result replicability (cf. Cohen, 1884; Thompson, 1996).

Rosenthal and Gaito (1963) had 29 participants (19 faculty members and 10 graduate students) rate their level of belief in a variety of p levels with an n of 10 and an n of 100. Their results indicate that these researchers placed greater confidence in p levels that contained a larger n. Nelson et al. (1986) obtained comparable results when they sent a similar questionnaires to 85 psychologists and found that these psychologist overrelied on p levels less than .05. Zuckerman et al. (1993) surveyed 551 psychologists and found that these researchers had a limited understanding of basic concepts in statistics including "the role of power and effect size as criteria for successful replications" (p. 49).

Thompson (1996) stated the necessity for researchers to report to the reader some techniques that evaluate the replicability of their results. Externally this can be accomplished by actually doing the study again with different participants. Because this is not usually done, internal measures of "cross-validation, the jack-knife, and/or the bootstrap" and confidence intervals can be employed to determine if the results will be consistent across various samples (Thompson, 1996, p. 29).

Cohen (1994) recommended that researchers present effect sizes as CIs. He claimed that "everyone knows" that CIs contain much more information than significance tests. CIs provide information about both the nil hypothesis and also about non-nil null hypotheses. Cohen (1994) felt that the reason for the lack of use of CIs is that often they are so wide or imprecise. He encouraged researchers to improve

> . . . . our measurement by seeking to reduce the unreliable and invalid part of the variance in our measures (as Student himself recommended almost a century ago). . . Larger sample sizes reduce the size of confidence intervals as they increase the statistical power of the null hypothesis significance testing. (p. 1002)

The case for the use of CIs can also be based on their power to evaluate theory, as against the statistical significance test's lack of utility in this regard. As Serlin (1993) explained,

> The point null hypothesis, like any universal theoretical proposition, must always be false. . . . Thus, the point null hypothesis cannot be used to specify a potential falsifier [of a theoretical proposition]; because the point null hypothesis is always false, a test of it would always (in principle) provide support. The appropriate null hypothesis must be derived from the theoretical prediction (fortified with a good-enough belt), which means that we must specify and test a range null hypothesis. (p. 352)

Serlin (1993) noted that too often if a particular theory was supported then researchers were not interested in how large their effect was or if their theory was not supported how close did they come. He reminded us that "the critics of significance testing suggested the use of confidence intervals as a way of improving the scientific utility of statistical methodology" (p. 352).

Purpose

The present paper will summarize various methods of estimating confidence intervals, including classical intervals and intervals for effect sizes. The second application is difficult, because such estimates require (a) the use of special statistical distributions that are called "noncentral" (e.g., "noncentral $\underline{t}$", "noncentral $\underline{F}$"), with which many researchers may be unfamiliar, and (b) the use of computer-intensive estimation procedures, because iterative estimation must be used rather than a computation formula. Fortunately, new software and/or new programming for old software have overcome these two difficulties (Cumming & Finch, 2001; Smithson, 2001).

What are Confidence Intervals?

Confidence intervals are common tools of inference, measuring how sure we are of our results. CIs across studies tell us how accurately and consistently our data operates over time. CIs invoke two primary concepts, intervals and confidence levels. Intervals are determined by the standard errors of statistics. Confidence levels are chosen by the researcher and are given as percentages. Simply put a 95% confidence level says the method used by the researcher gives an interval that covers the true population parameter 95% of the time. For example, by calculating a confidence interval for my cholesterol level taken twenty times (n=20), I can state how confident I am that the sample mean accurately reflects my choloesterol level. A range null hypothesis (160 - 200) is tested rather than a point null hypothesis (180).

There exists a seesaw relationship between confidence levels and intervals: the higher the confidence level the wider the interval or the larger the margin of error. The lower the confidence level, the narrower the interval or the smaller the margin of error. For the CI for the mean the standard deviation also effects the margin or error, as there is more variance in the population, the wider the interval as shown in Figures 1a & 1b. Figure 1c suggests that to make the margin of error smaller, the researcher must collect more data which shrinks the margin of error due to the formula

$$x \pm z^*\left(\frac{\sigma}{\sqrt{n}}\right)$$

where z^* is a z score related to your p value and is a measure of how many standard deviations away from the mean you are. The z^* for .05 is 1.96 equaling a 95% confidence level, z^* for .01 is 2.576 equaling a 99% confidence level (Consortium for Mathematics and Its Applications, 1989).

Bohrnstedt and Knoke (1982) defined confidence intervals as "a range of values constructed around a point estimate which makes it possible to state the probability that the interval contains the population parameter between its upper and lower confidence limits" (p. 144). Thus 95% of the intervals constructed in repeated sampling of the population mean will be contained within the boundaries defined by two standard deviations above and below the mean (Bohrnstedt & Knoke, 1982). According to Finch et al. (2001) a confidence interval "presents an estimate of the true effect and its precision; this alone should encourage substantive interpretation" (p. 203). Smithson (2001) defined a confidence interval for a statistic as a "range of values that contain a specified percentage 100(1 $-\alpha$) of the sampling distribution of that statistic" (p. 607). CIs compliment data given by power analysis in analyzing studies.

Power alone is not enough in determining an effect of a certain size. There is a necessity to understand how spreadout the CIs are for the effect size, given a particular sample and a desired confidence level. As Smithson (2001) noted "CIs are an essential component in the accumulation of scientific knowledge because they avoid the misleading 'vote counting' to which NHST is prone" (p. 626). The alpha level determines the confidence level associated with CIs for a study. Thus a researcher who assigns a .05 significance level to a study will use a 95% confidence level for that same study (Sullivan, 2001).

Advantages of Using Confidence Intervals

There are a variety of reasons why CIs should be used when reporting results:

1. CIs lend themselves to enhanced understanding and are fairly easily obtained using SPSS or the ESCI software developed by Cumming and Finch (2001).
2. CIs and NHST are related. If a value causes a hypothesis to be rejected then that value will be outside the confidence interval.
3. CIs are helpful in compiling studies. They support meta-analysis and thinking.
4. CI width can be figured out *a priori* a study. Width of CI interval can be used to determine study design and sample size (Cumming & Finch, 2001).
5. CIs are easier to present in graphic display and thus easier for readers to interpret (Sullivan, 2001).

Finding Confidence Intervals

Let us suppose that we wanted to determine the probability that I would find pairs of shoes on sale that would be equal to the population mean price of shoes in College Station at $17.99. H_0: [$\mu - \mu_0$; μ_0 = $17.99] ($\underline{n}$=20). If all I did was a NHST then the only thing the reader of my results would know is that \$17.99 was statistically significant (*p* = .0003, *SD* = 12.065). If instead, I present the confidence interval information included in Figure 2 from the ESCI software *CIoriginal* (Cumming & Finch, 2001), the reader can now actually see all the prices of the 20 shoes I bought, which pairs of shoes lie within my 95% confidence interval from \$27.24 to \$38.53 with a standard error of measurement of ±5.64. Thus almost all but two pairs of my purchased shoes were too costly compared to the price of shoes in College Station. I would be considered a poor shopper.

If I do a second study and this time determine that the population mean for shoes in College Station is \$29.99, then this time the reader can easily see, as shown in Figure 3, that even though my results are no longer statistically significant (*p* = .295, *SD* = 12.65), many of the shoes I bought would be a good price and I would be considered a wise shopper.

The next section of the ESCI software package (Cumming & Finch, 2001) demonstates *CIJumping*. This section represents an artificial situation involving many samples and not just a usual study involving one sample where we do not know μ. In Figure 4, one can see how in to make the CI width (range) smaller, sample size needs to get larger, therefore necessitating the use of a larger sample size. A *n* four times larger is needed to cut the margin of error in half. In this example we will imagine that the average score on the midterm is 50 points (*SD*=20). We will assume that the σ for the population of university students is also 50. In Figure 4 an *n* = 15 is used and the CIs are very large. When the population is multiplied by four (*n* = 60) and the μ and σ remain the same in Figure 5, the CIs are much smaller. Note in Figure 5, 24 of the 25 samples were captured (smaller margin of error) in contrast to Figure 4 where 24 of the 24 samples were captured (larger margin of error).

The next section, *NonCentral t* of the ESCI software package (Cumming & Finch, 2001), demonstrates the use of confidence intervals to find noncentral *t* distributions. Central *t* is always used in null hypothesis testing since $\mu = \mu_0$ (null hypothesis always true) causes no shift in distribution. When there are different value for μ (true population mean) and μ_0 (chosen value), two different curves are needed and the difference between the two ($\mu - \mu_0$) (sometimes divided by some SD) is the effect size. This comes from the formula

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

Obtaining a confidence interval for $\underline{\Delta}$ is not easy because Δ is a function of two parameters $\underline{\mu}$ and $\underline{\sigma}$ and these parameters are estimated from the data. We must think of a possible upper and lower limit for $\underline{\Delta}$, calculate these limits, then divide them by the $\underline{\sqrt{n}}$ and this will give us the limits for Cohen's $\underline{\delta}$.

Next using our $\underline{\bar{x}}$ and $\underline{s}$, we can calculate

$$t_{n-1,\Delta} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

for a 95% confidence level, using .025 from the $\underline{z^*}$ table, we have the probability of the upper and lower tails. No formula can give us $\underline{\Delta}$ only statistical software (cf. Smithson, 2001). After obtaining these upper and lower limits for $\underline{\Delta}$ we can use the formula

$$\delta = \mu - \mu_0 / \sigma = \Delta\sqrt{n}$$

And from these equalities we can derive that

$$\Delta = \delta\sqrt{n}$$

Figure 6 displays a noncentral distribution where $\underline{\Delta}$ = 10; when $\underline{\Delta}$ = 2 as on can see in Figure 7 and is getting closer to zero (both curves would be identical = central t) the curves have more and more common area.

Serlin (1993) suggested that CIs should be obtained based on *range* null hypotheses rather than *point* null hypotheses, noting that "In the case of a range null hypothesis... one uses the observed data to test. . . an infinite number of hypotheses H_0: $[\mu - \mu_0] < \Delta_0$ where $\underline{\Delta_0}$ is varied over all possible values" (p. 354). These non-rejected values of $\underline{\Delta_0}$ make up the confidence interval. Hodges and Lehmann (1954) acquainted us with the procedure for testing the range null hypotheses and "obtaining the corresponding confidence interval in a one- or two-sample experiment when the width of the good-enough belt is specified in raw (unstandardized) units" (Serlin, 1993, p. 355). This procedure can be extended to multiple-sample experiments while still considering Type I error rate prior to the experiment and also controlling alpha at the familywise level. Multiple comparisons that will permit the measurement of range-based confidence are the Bonferroni and the Holm method.

Kennedy and Schumacher (1993) recommend using the algorithm of bootstrapping to calculate confidence intervals. One of the advantages in doing this is that the population does not have to be normal. The complete *Minitab* program found in their appendix (pp. 98-99) allows students to easily obtain a bootstrap interval for a population parameter. The subroutine allows students to concretely construct the 95% confidence interval of the population variance. Although labor-intensive, these statistics experts feel that bootstrapping is a worthwhile procedure for students to learn and practice.

Lambert, Wildt, and Durand (1991) also advocated bootstrapping but as a means of approximating confidence intervals for factor pattern coefficients. These researchers were searching for an alternative to the exploratory way that factors are generally retained or excluded. By applying confidence intervals to factor pattern coefficients, a criterion value could be used to make comparisons and thus determine how to treat various factors. They liken this process "to hypothesis testing in which sampling variances are taken into account" (p. 422).

Smith (1982) advocated the use of the jackknife procedure for finding confidence intervals for variance component estimates in generalizability theory. He found that jackknife developed by Mosteller and Tukey in 1968 was beneficial. Intervals could be used *a priori* as an indication of the exactness that these components could be estimated given the obtainable means.

Psychologists are encouraged to use CIs when reporting test scores to clients and school personnel. Relating "confidence bands" assists educators in making decisions based on the fallibility of test result data. Schulte and Borich (1988) suggested using the procedure of reporting test scores along with CIs to be as understandable as possible to receivers of scores. They caution psychologists, however, that some test manuals can be confusing and contain misinterpretations. In spite of this warning Schulte and Borich (1988) strongly recommended that this feedback can be interpreted as a range of scores and thus an individual can be pretty sure that if they took the test again their score would fall in between the lower and upper limits of their score. Or as Sattler (1982) maintains, "If we construct a 95 percent confidence interval, then the chances are only 5 in 100 that a person's true score lies outside the confidence interval" (p. 22). Silver and Clampit (1991) agreed that CIs should be reported in conjunction with a person's IQ. Their article contains 95% and 99% confidence interval tables for the WISC-R that utilize Schulte and Borich (1988) method based on the standard error of estimate or the standard error of prediction.

Two widely used methods for computing confidence limits are based on standard error or measurement and standard error of estimation. Glutting, McDermott, and Stanley (1987) recommended the use of the formula developed by Stanley in 1971 to establish CIs around an estimated true score: $\hat{T} \pm (z)(S_e)(r)$. These researchers "point out that the standard error of measurement is *larger* than the measurement error associated with estimated true score (unless that test is perfectly reliable or the obtained score is the same as the test mean)" (p. 610) and that these intervals are "both sensitive to different reference groups and consistent with classical test-score theory" (p. 614).

Summary

As Cumming & Finch (2001) emphasized, "We strongly support these calls for reform and believe that wider understanding and use of CIs should be a central aspect of changes to statistical practice in psychology, education, and cognate disciplines (p.535). This sentiment was echoed by many other researchers who currently advocate the use of confidence intervals to replace significance testing methods. As Schmidt (1996) so aptly stated "reliance on statistical significance testing in the analysis and interpretation of research data has systematically retarded the growth of cumulative knowledge in psychology". The APA <u>Publication Manual</u> (2001), the APA Task Force, and an increasing number of journal editors have strongly recommended the use of confidence intervals. Fortunately, new software has made the difficulty of identifying confidence intervals a thing of the past (Cumming & Finch, 2001; Smithson, 2001).

Cohen (1994) charged researchers with the task of constructing confidence intervals when he explained,

> As researchers, we have a considerable array of statistical techniques that can help us find our way to theories of some depth, but they must be used sensibly and be heavily informed by informed judgement. Even null hypothesis testing complete with power analysis can be useful if we abandon the rejection of point nil hypotheses and use instead "good-enough" range null hypotheses. (p. 1002)

Combining effect size with confidence intervals is the wave of the future in continuing efforts to make research understandable to the reader. Logically, if effect sizes are <u>good</u>, and confidence intervals are <u>good</u>, then confidence intervals about effect sizes should be darn <u>nifty</u>.

References

American Psychological Association. (2001). *Publication manual of the American Psychological Association* (5th ed). Washington, DC: Author.

Bohrnstedt, G. W., & Knoke, D. (1982). *Statistics for social data analysis*. Chicago: Peacock.

Cahan, S. (2000). Statistical significance is not a "kosher certificate" for observed effects: A critical analysis of the two-step approach to the evaluation of empirical results. *Educational Researcher, 29*(1), 31-34.

Caruso, J., & Cliff, N. (1997). Empirical size, coverage, and power of confidence intervals for spearman's rho. *Educational and Psychological Measurement, 57*, 637-654.

Cohen, J. (1994). The earth is round ($p \leq .05$). *American Psychologist, 49*, 997-1003.

Consortium for Mathematics and Its Application. (1989) *Confidence intervals: Significance tests* [Videorecording]. Part of a series: Against All Odds: Inside Statistics (A production of the Chedd-Angier Production Co, No.19-20). Lexington, MA: COMAP

Cumming, G., & Finch, S. (2001). A primer on the understanding, use and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement, 61*, 532-575.

Finch, S., Cumming, G., & Thomason, N. (2001). Reporting of statistical inference in the Journal of Applied Psychology: Little evidence of reform. *Educational and Psychological Measurement, 61*, 181-210.

Glutting, J., McDermott, P., & Stanley, J. (1987). Resolving differences among methods of establishing confidence limits for test scores. Educational and *Psychological Measurement, 47*, 607-614.

Hagen, R. L. (1997). In praise of the null hypothesis statistical test. *American Psychologist, 52*, 15-24.

Kennedy, K., & Schumacher, P. (1993). The introduction of bootstrapping techniques for finding confidence intervals using simulation. *Journal of Computers in Mathematics and Science Teaching, 12*(1), 83-92.

Kieffer, K. M., Reese, R. J., & Thompson, B. (2001). Statistical techniques employed in AERJ and JCP articles from 1988 to 1997: A methodological review. *Journal of Experimental Education, 69*, 280-309.

Knapp, T., & Sawilowsky, S. (2001). Constructive criticisms of methodological and editorial practices. *Journal of Experimental Education, 69,* 65-79.

Lambert, Z., Wildt, A., & Durand, R. (1991). Approximating confidence intervals for factor loadings. *Multivariate Behavioral Research, 26,* 421-434.

Nelson, N., Rosenthal, R., & Rosnow, R. (1986). Interpretation of significance levels and effect sizes by psychological researchers. *American Psychologist, 41,* 1299-1301.

Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioral sciences.* New York: Wiley.

Rosenthal, R., & Gaito, J. (1963). The interpretation of level of significance by psychological researchers. *Journal of Psychology, 55,* 33-38.

Sattler, J. (1982). *Assessment of children's intelligence and special abilities (2nd ed.).* Boston: Allyn & Bacon.

Silver, S., & Clampit, M. (1991). Corrected confidence intervals for quotients on the WISC-R, by level of quotient. *Psychology in the Schools, 28,* 9-14.

Schmidt, F. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for the training of researchers. *Psychological Methods, 1,* 115-129.

Schulte, A., & Borich, G. (1988). False confidence in intervals: Inaccuracies in reporting confidence intervals. *Psychology in the Schools, 25,* 405-412.

Serlin, R. C. (1993). Confidence intervals and the scientific method: A case for Holm on the range. *Journal of Experimental Education 61,* 350-360.

Smith, P. (1982). A confidence interval approach for variance component estimates in the context of generalizability theory. *Educational and Psychological Measurement, 42,* 459-466.

Smithson, M. (2001). Correct confidence intervals for various regression effect sizes and parameters: The importance of noncentral distributions in computing intervals. *Educational and Psychological Measurement, 61,* 605-632.

Sullivan, J. (2001). The utility of statistical significance testing in psychological and educational research: A review of recent literature and proposed alternatives. *Research in the Schools, 8,* 83-91.

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher, 25*(2), 26-30.

Thompson, B. (1998). In praise of brilliance: Where that praise really belongs. *American Psychologist, 53,* 799-800.

Thompson, B. (2001). Significance, effect sizes, stepwise methods, and other issues: Strong arguments move the field. *Journal of Experimental Education, 69,* 80-93.

Wilkinson, L., & APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 594-604. {reprint available through the APA Home Page: http://www.apa.org/joournals/amp/amp548594.html}

Zuckerman, M., Hodgins, H., Zuckerman, A., & Rosenthal, R. (1993). Contemporary issues in the analysis of data: A survey of 551 Psychologists. *Psychological Science, 4,* 49-53.

Figure 1

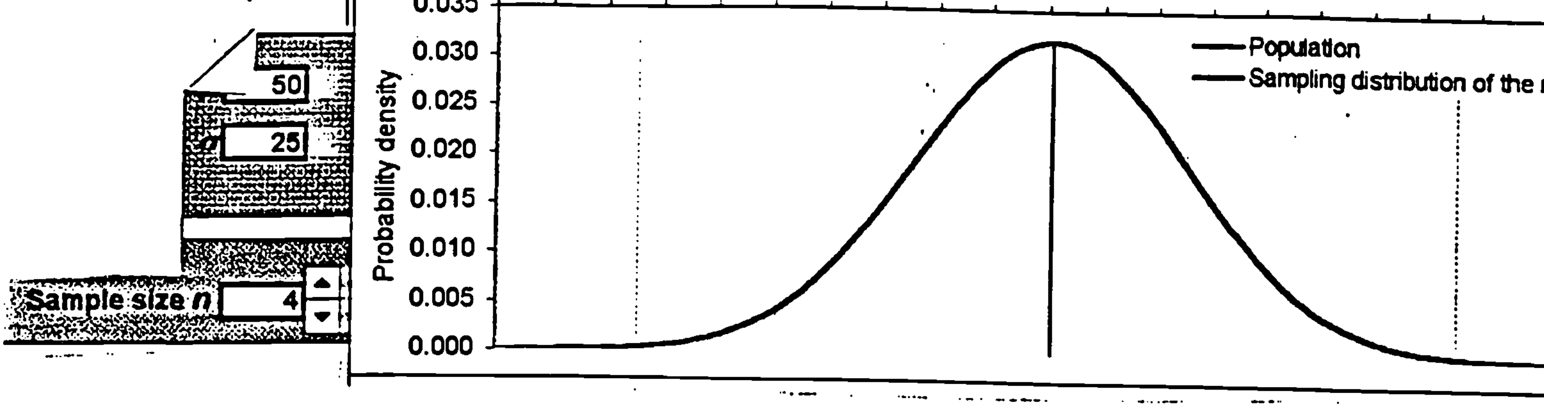


Figure 1b

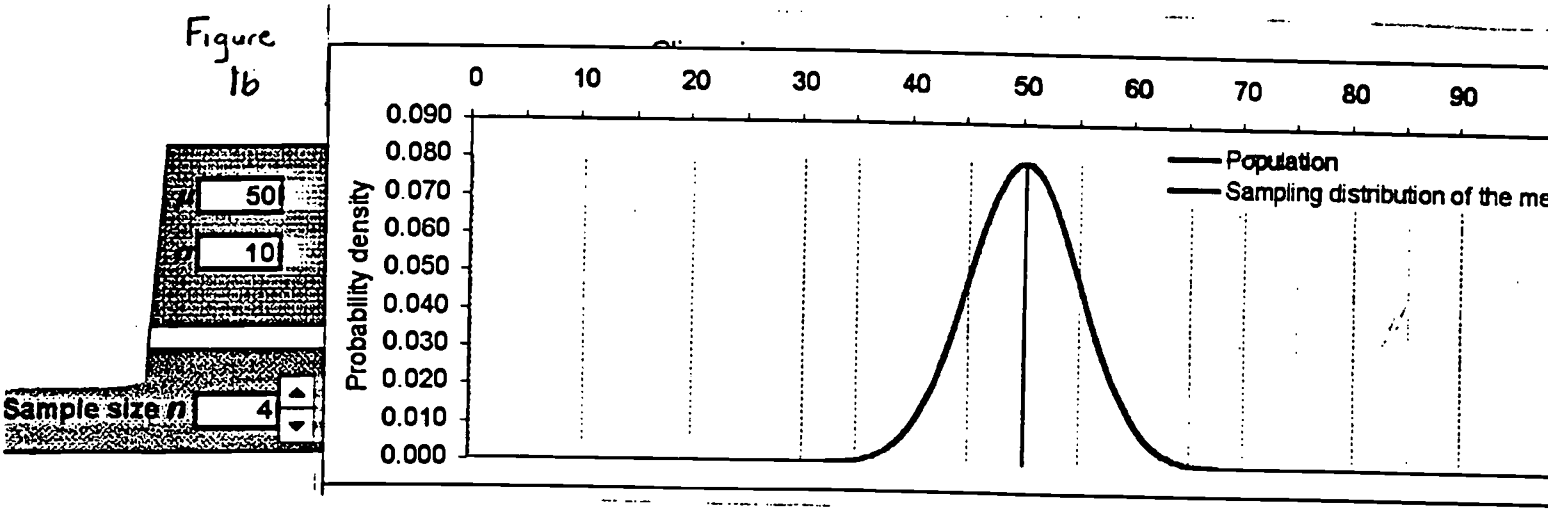


Figure 1c

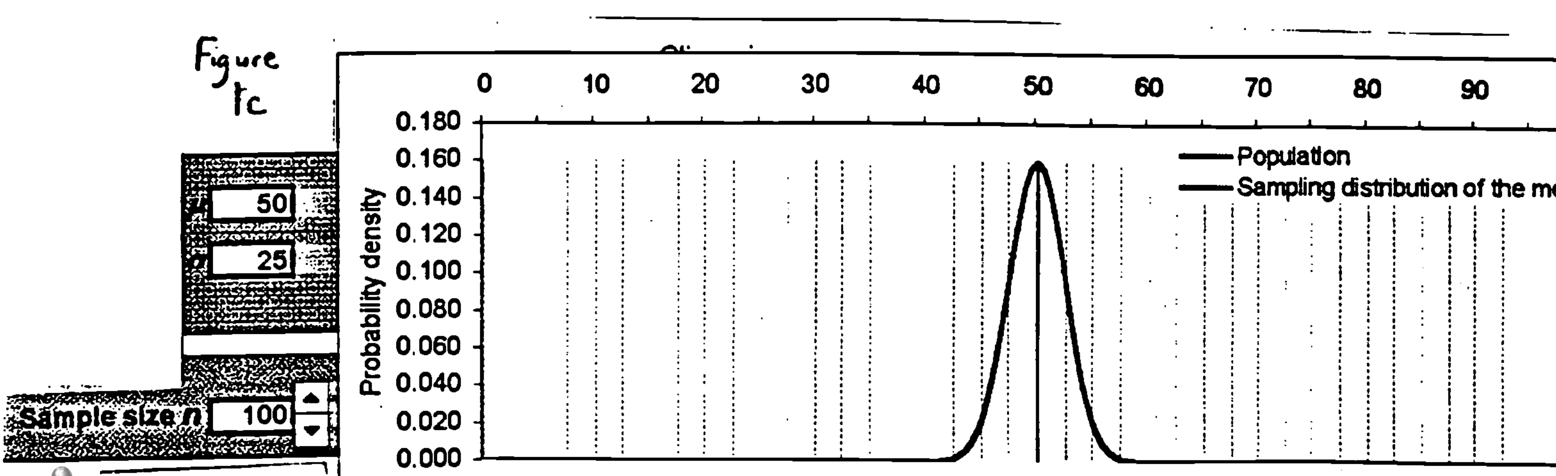
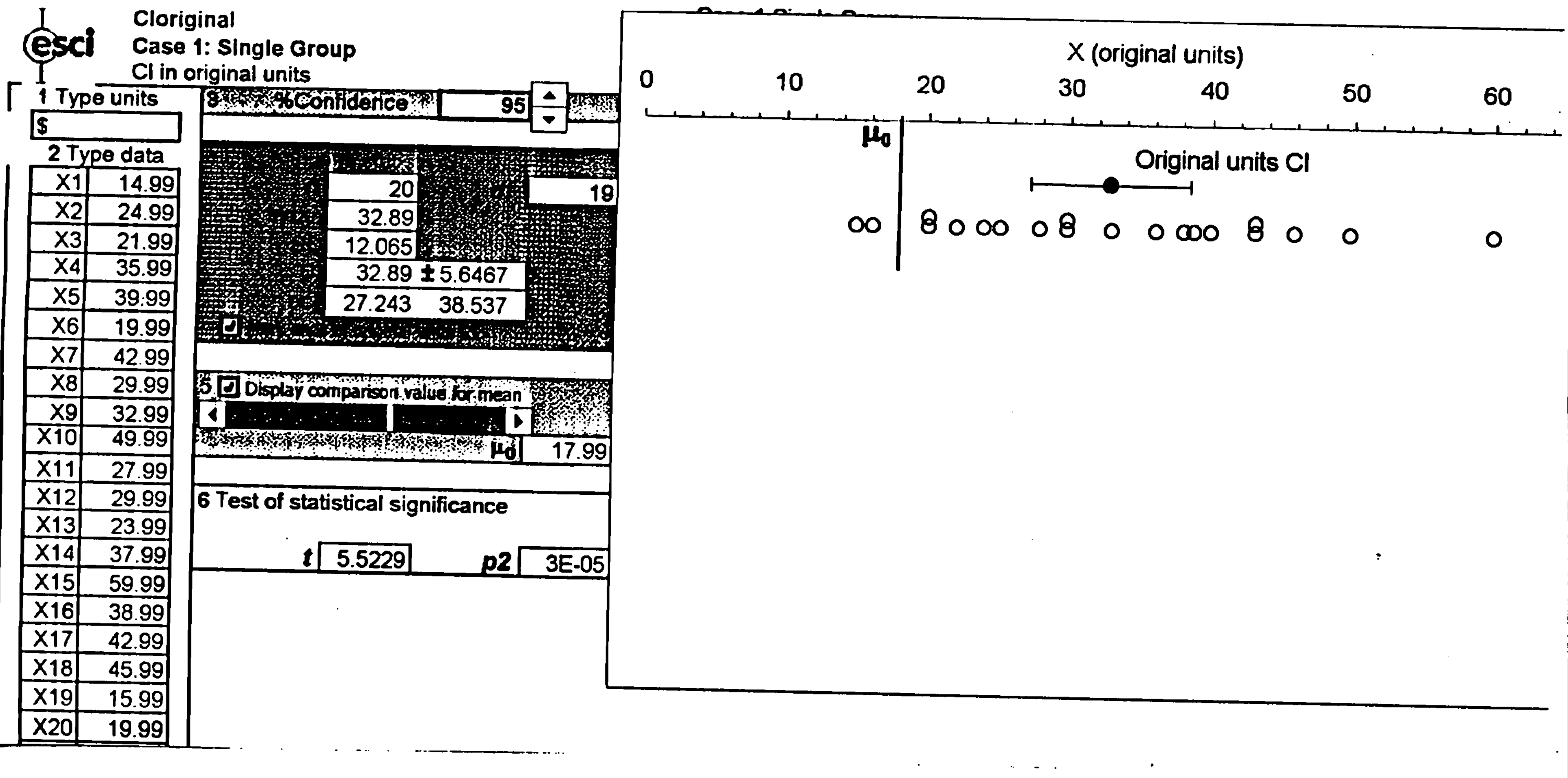


Figure 2



esci
Cloriginal
Case 1: Single Group
CI in original units
1 Type units
$
2 Type data
X1 14.99
X2 24.99
X3 21.99
X4 35.99
X5 39.99
X6 19.99
X7 42.99
X8 29.99
X9 32.99
X10 49.99
X11 27.99
X12 29.99
X13 23.99
X14 37.99
X15 59.99
X16 38.99
X17 42.99
X18 45.99
X19 15.99
X20 19.99
%Confidence 95
20
19
32.89
12.065
32.89 ± 5.6467
27.243 38.537
5 Display comparison value for mean
μ_0 17.99
6 Test of statistical significance
t 5.5229
p2 3E-05
X (original units)
0
10
20
30
40
50
60
μ_0
Original units CI

Figure 3

esci

Cloriginal

Case 1: Single Group

CI in original units

1 Type units	
$	
2 Type data	
X1	14.99
X2	24.99
X3	21.99
X4	35.99
X5	39.99
X6	19.99
X7	42.99
X8	29.99
X9	32.99
X10	49.99
X11	27.99
X12	29.99
X13	23.99
X14	37.99
X15	59.99
X16	38.99
X17	42.99
X18	45.99
X19	15.99
X20	19.99

3 %Confidence 95

20	19
32.89	
12.065	
32.89 ± 5.6467	
27.243	38.537

5 Display comparison value for mean

μ_0 29.99

6 Test of statistical significance

t 1.0749 $p2$ 0.2959

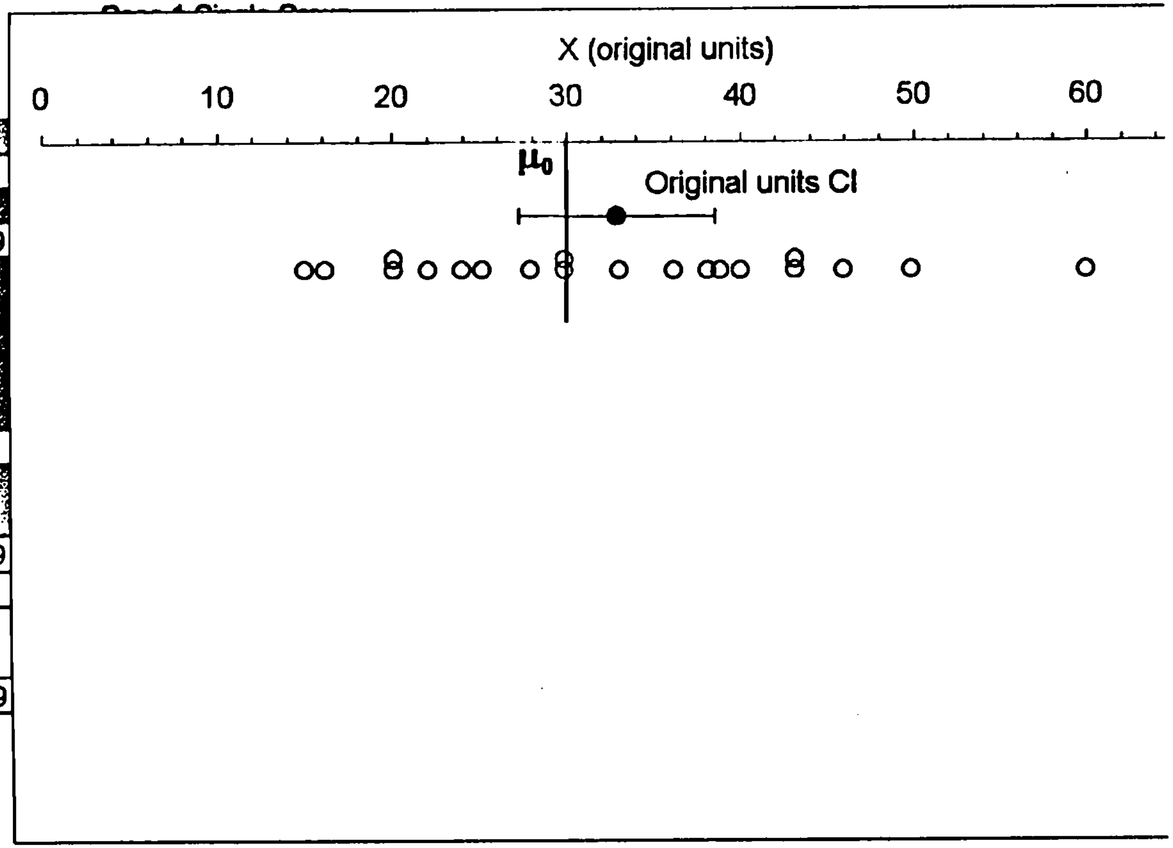
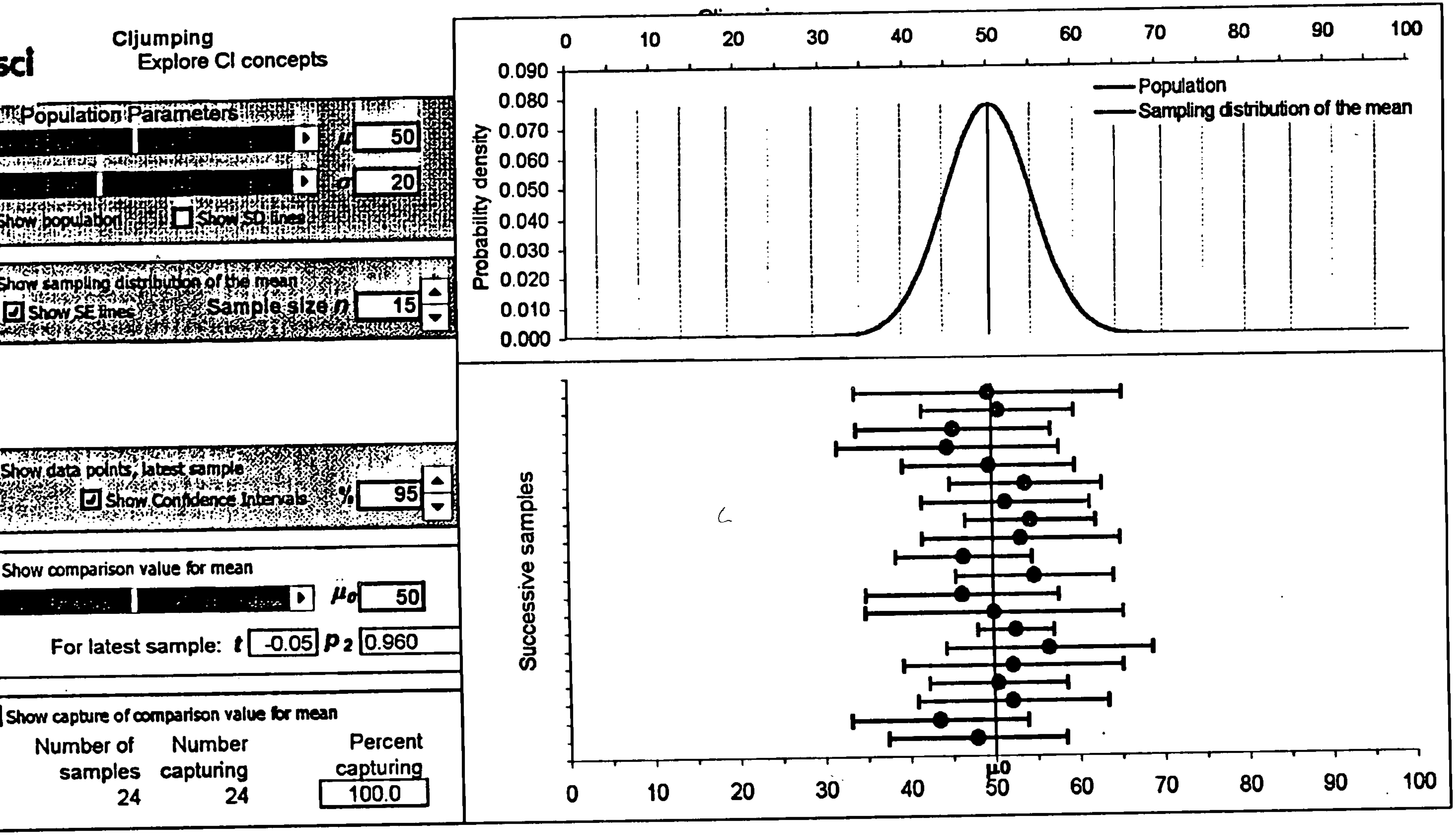
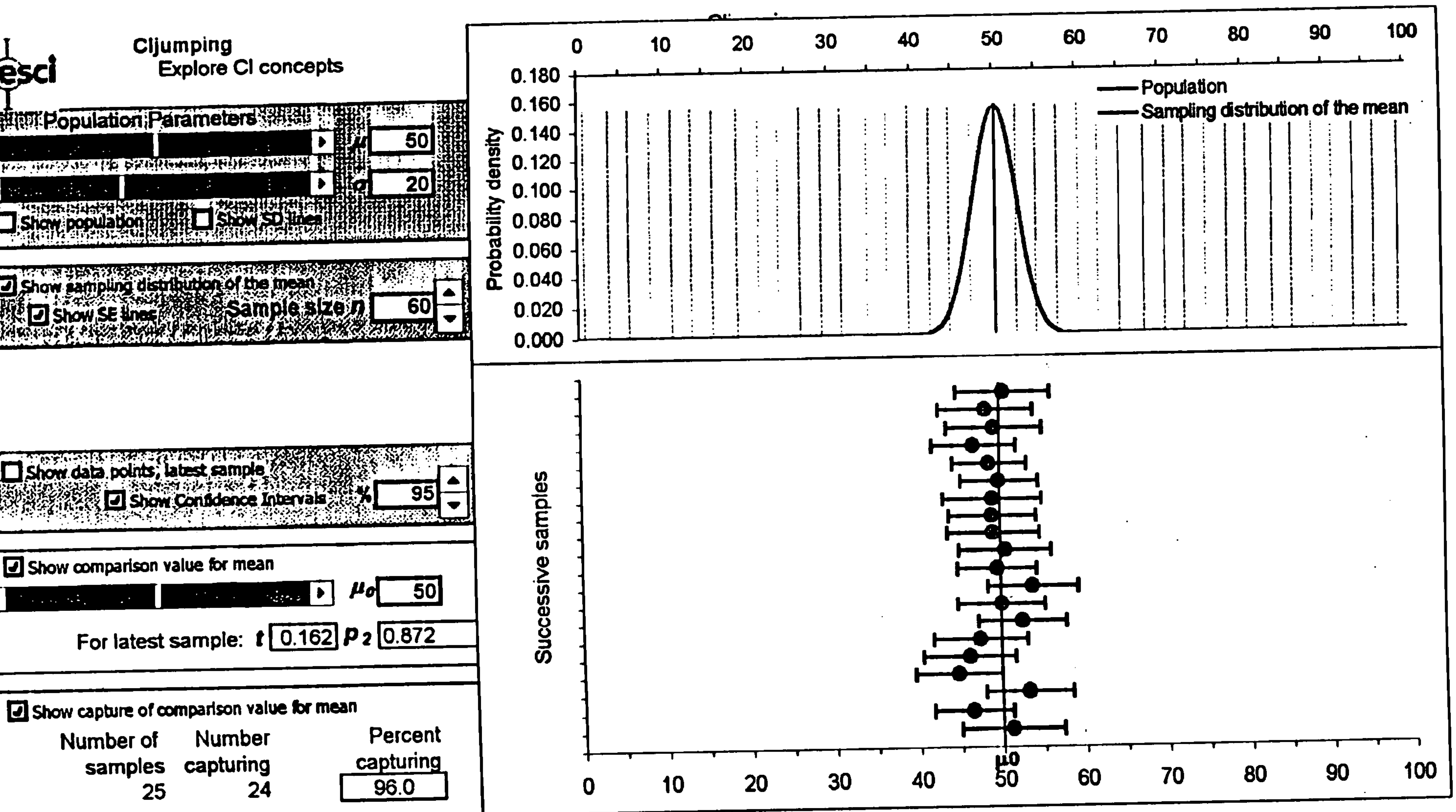


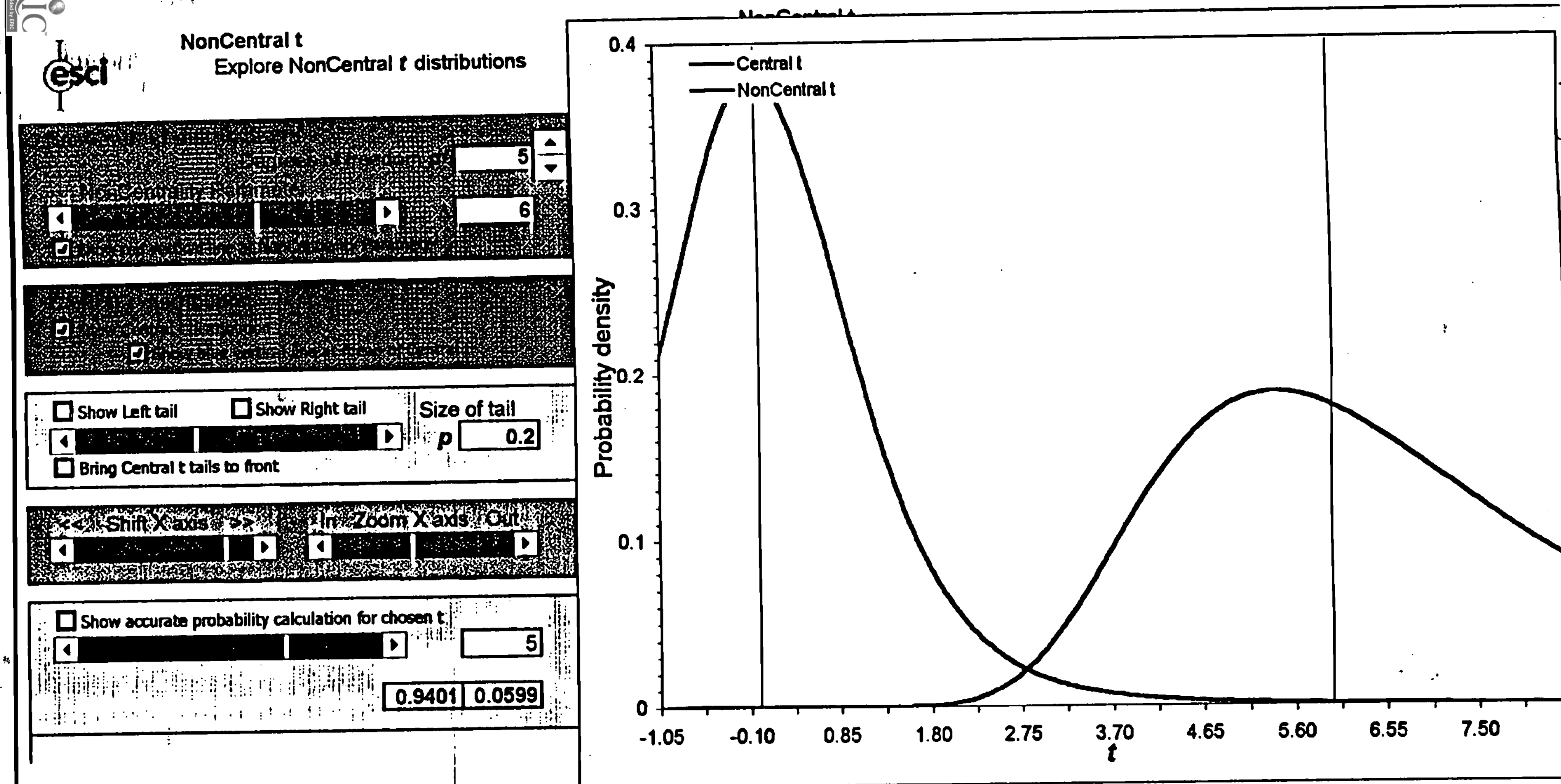
Figure 4



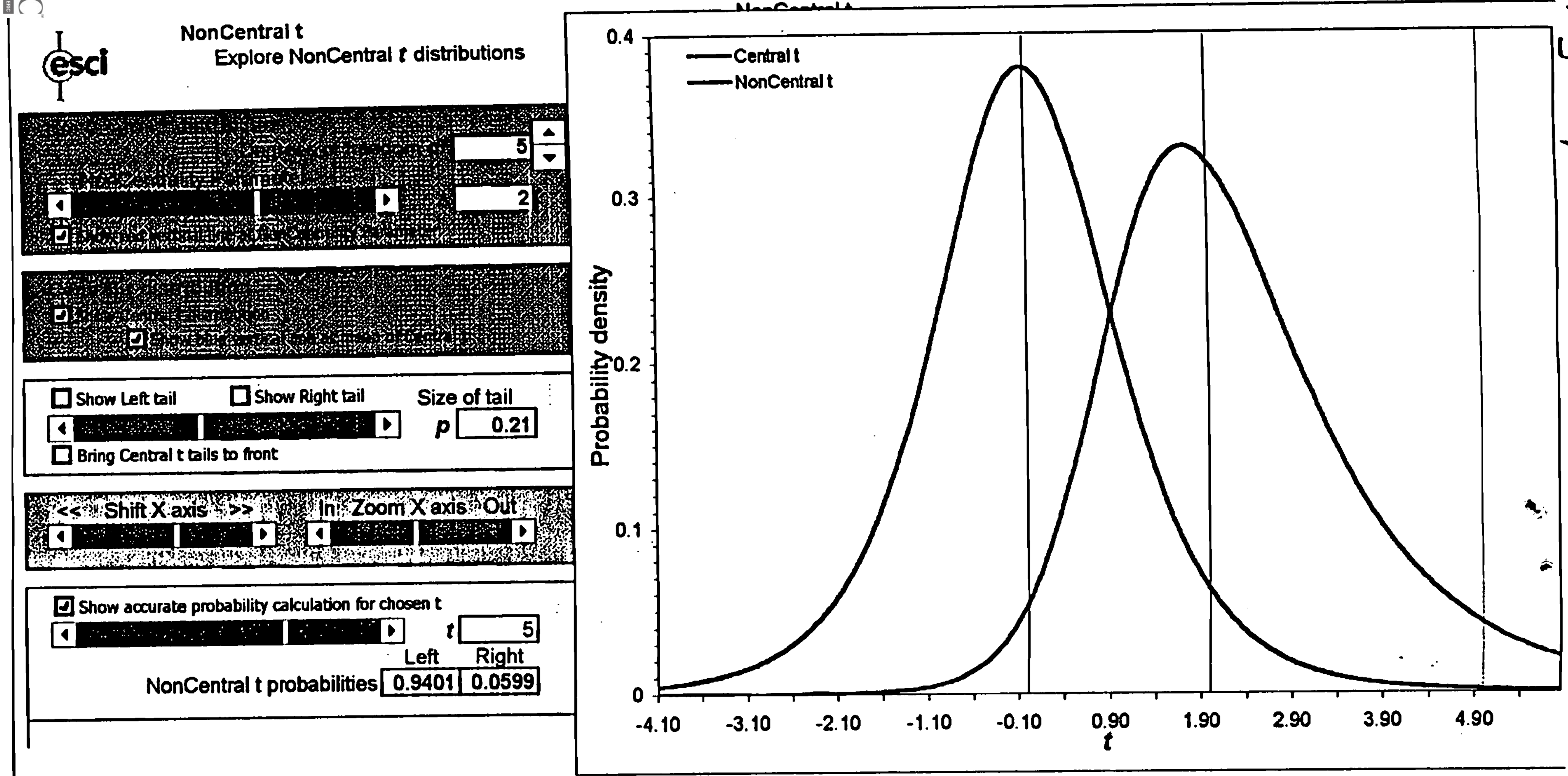
sci
CIjumping
Explore CI concepts
Population Parameters
μ 50
σ 20
Show population
Show SD lines
Show sampling distribution of the mean
Show SE lines
Sample size n 15
Show data points, latest sample
Show Confidence Intervals % 95
Show comparison value for mean
μ0 50
For latest sample: t -0.05 p2 0.960
Show capture of comparison value for mean
Number of samples 24
Number capturing 24
Percent capturing 100.0
Population
Sampling distribution of the mean
Probability density
0.090
0.080
0.070
0.060
0.050
0.040
0.030
0.020
0.010
0.000
0 10 20 30 40 50 60 70 80 90 100
Successive samples
μ0

Figure 5


esci
CIjumping
Explore CI concepts
Population Parameters
μ 50
σ 20
Show population
Show SD lines
Show sampling distribution of the mean
Show SE lines
Sample size n 60
Show data points, latest sample
Show Confidence Intervals
% 95
Show comparison value for mean
μ0 50
For latest sample: t 0.162 p2 0.872
Show capture of comparison value for mean
Number of samples 25
Number capturing 24
Percent capturing 96.0
Probability density
0.180
0.160
0.140
0.120
0.100
0.080
0.060
0.040
0.020
0.000
0 10 20 30 40 50 60 70 80 90 100
Population
Sampling distribution of the mean
Successive samples
μ0



NonCentral t
Explore NonCentral t distributions
5
6
Show Left tail
Show Right tail
Size of tail
p
0.2
Bring Central t tails to front
<< Shift X axis >>
In Zoom X axis Out
Show accurate probability calculation for chosen t
5
0.9401
0.0599
Central t
NonCentral t
Probability density
0.4
0.3
0.2
0.1
0
-1.05
-0.10
0.85
1.80
2.75
3.70
4.65
5.60
6.55
7.50
t


NonCentral t
Explore NonCentral t distributions
5
2
Show Left tail
Show Right tail
Size of tail
p 0.21
Bring Central t tails to front
<< Shift X axis >>
In Zoom X axis Out
Show accurate probability calculation for chosen t
t 5
Left
Right
NonCentral t probabilities 0.9401 0.0599
Central t
NonCentral t
Probability density
0.4
0.3
0.2
0.1
0
-4.10
-3.10
-2.10
-1.10
-0.10
0.90
1.90
2.90
3.90
4.90
t

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: An Introduction to Confidence Intervals for Both Statistical Estimates & Effect Sizes

Author(s): Mary Margaret Capraro

Corporate Source: TAMU

Publication Date: 2/14/02

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 1	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2A	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2B
Level 1	Level 2A	Level 2B
[X]	[ ]	[ ]
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, → please

Signature: Mary Margaret Capraro	Printed Name/Position/Title: Mary Margaret Capraro
Organization/Address: Texas A+M Univ.	Telephone: 979-845-8227 FAX: 979-845-9663
	E-Mail Address: mmcapraro@coe.tamu.edu Date: 2/15/02

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

THE CATHOLIC UNIVERSITY OF AMERICA
ERIC CLEARINGHOUSE ON ASSESSMENT AND EVALUATION
210 O'BOYLE HALL
WASHINGTON, DC 20064
Attn: Acquisitions

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
1100 West Street, 2nd Floor
Laurel, Maryland 20707-3598

Telephone: 301-497-4080
Toll Free: 800-799-3742
FAX: 301-953-0263
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com