

## One Mean Inference<sup>©</sup>

---

### Z Population Standard Deviation Known

To use  $Z$  as the test statistic when testing the null hypothesis that  $\mu =$  some value, one must know the  $\sigma$  and be able to assume that the distribution of sample means is normal. The distribution of  $Z$  will be normal if:

- the population distribution of the variable being tested ( $Y$ ) is normally distributed or
- the **central limit theorem** [CLT] applies.

The CLT states that the distribution of sample means will become approximately normal with large sample sizes, even if  $Y$  is not normally distributed. If  $Y$  is nearly normal, the distribution of sample means will approach normal very quickly with increasing  $N$ , but if  $Y$  is very nonnormal,  $N$  may need be relatively large (30 or more) for the sampling distribution to be close to normal.

How large  $N$  must be to produce approximately normal sampling distributions can be investigated with computer **Monte Carlo** simulations. One instructs the computer to sample randomly a large number of samples of given  $N$  from a population of a specified shape, mean, and variance. The shape of the resulting sampling distribution is noted and then the whole process is repeated with a larger  $N$ , etc. etc. until the sampling distribution looks approximately normal.

The **standard error of the mean**, the standard deviation of the distribution of sample means, is:  $\sigma_M = \frac{\sigma}{\sqrt{N}}$ .

For example, suppose we wish to test the  $H_0$  that for IQ,  $\mu = 100$  for ECU students. Assuming a population  $\sigma$  of 15 (that found in the general population) and a normal sampling distribution, we compute:  $Z = \frac{M - \mu_0}{\sigma_M}$  Suppose that our sample of 25

students has  $M = 107$ .  $Z = \frac{107 - 100}{15/\sqrt{25}} = \frac{7}{3} = 2.33$ .

Now,  $P(Z \geq +2.33) = .0099$ ; doubling for a **two-tailed test**,  $p = .0198$ . Thus, we could reject the  $H_0$  with an  $\alpha$ -criterion of .05, but not at .01. Were we to have done a **one-tailed test** with  $H_0$  being  $\mu \leq 100$ ,  $H_1$  being  $\mu > 100$ ,  $p = .0099$  and we could reject the  $H_0$  even at .01.

With a **one-tailed test**, if the direction of the “effect”, (sample mean is  $>$  or  $<$   $\mu_0$ ) is as specified in the alternative hypothesis, one always uses the “**smaller portion**” column of the normal curve table to obtain the  $p$ . If the direction is opposite that specified in the alternative hypothesis, one uses the “**larger portion**” column. For a **two-tailed test**, always use the **smaller portion** column and **double** the value that appears there.

**Confidence intervals** may be constructed by taking the point estimate of  $\mu$  and going out the appropriate number of standard errors. The general formula is:

$$CI = M - CV * \sigma_M \leftrightarrow M + CV * \sigma_M$$

where  $CV$  = the critical value for the appropriate sampling distribution. For our sample problem  $CI_{.95} = 107 - 1.96(15/5) \leftrightarrow 107 + 1.96(15/5) \Rightarrow 101.12 \leftrightarrow 112.88$ . Once we have this confidence interval we can decide whether or not to reject a hypothesis about  $\mu$  simply by determining whether the hypothesized value of  $\mu$  falls within the confidence interval or not. The hypothesized value of 100 does not fall within 101.12 to 112.88, so we could reject the hypothesis that  $\mu = 100$  with at least 95% confidence – that is, with alpha not greater than  $1 - .95 = .05$ .

The wider the confidence interval is, the more error there is in our estimation of  $\mu$ . To reduce error and obtain a more narrow confidence interval, one can take steps to reduce the standard error of the mean. For example, one can increase the sample size. For our sample problem, imagine that we have 100 cases rather than just 25. The standard error of the mean is then reduced from  $15/5 = 3$  to  $15/10 = 1.5$ . The confidence interval is then narrowed from  $[101.12, 112.88]$  to  $107 \pm 1.96(1.5) = [104.06, 109.94]$ .

The meaning of the confidence coefficient (typically 95%) is difficult to grasp for some students, so I shall review it here. If you were to spend an eternity constructing an uncountably large number of 95% confidence intervals, 95% of those intervals would contain the true value of the estimated parameter (the  $\mu$ ) and 5% of them would not.

### Student's $t$ . Population Standard Deviation Not Known

One big problem with what we have done so far is knowledge of the population  $\sigma$ . If we really knew the  $\sigma$ , we would likely also know  $\mu$ , and thus not need to make inferences about  $\mu$ . The assumption we made above, that  $\sigma_{IQ}$  at ECU = 15, is probably not reasonable. Assuming that ECU tends to admit brighter persons and not persons with low IQ, the  $\sigma_{IQ}$  at ECU should be lower than that in the general population. We shall usually need to estimate the population  $\sigma$  from the same sample data we use to test the mean. Unfortunately, **sample variance,  $SS / (N - 1)$ , has a positively skewed sampling distribution**. Although unbiased [the mean of the distribution of sample variances equals the population variance], more often than not sample  $s^2$  will be smaller than population  $\sigma^2$  and sample  $s$  smaller than population  $\sigma$ .

Thus, the quantity  $t = \frac{M - \mu_0}{s / \sqrt{N}}$  will tend to be larger than  $Z = \frac{M - \mu_0}{\sigma / \sqrt{N}}$ . The result of

all this is that the sampling distribution of the test statistic will not be normally distributed, but will rather be distributed as **Student's  $t$** , a distribution developed by **Gosset** (his employer, Guinness Brewers, did not allow him to publish under his real name). For more information on Gosset, point your browser to:

<http://www-gap.dcs.st-and.ac.uk/~history/Mathematicians/Gosset.html>.

At Guinness, [Gosset](#) applied his skills in chemistry and mathematics towards developing new methods for evaluating data address questions like how to identify the

best yielding strains of barley (a primary ingredient in beers). Most of the statisticians of his day were interested in dealing with large samples of data, but Student typically had small data sets, and thus he developed statistical methods appropriate for working with small data set. It was [R. A. Fisher](#) who recognized the importance of Gosset's work and who named the statistic "t."

Student's  $t$ -distribution is **plumper in its tails** (representing a greater number of extreme scores in its tails) than is the normal curve. That is, Student's  $t$  has kurtosis  $> 0$ , it is leptokurtic, with the kurtosis becoming greater as the sample size becomes smaller. Because the distribution of sample variances is more skewed with small sample sizes than when  $N$  is large, the  $t$  distribution becomes very **nearly normal when  $N$  is large**. Pull out a table of critical values. Find the column that gives the critical value of Student's  $t$  for marking off the extreme 5% (two-tailed). Note that when  $df = 1$ , the critical value is 12.706 – wow, that  $t$  distribution has some seriously fat tails. Now observe what happens as the  $df$  increase. With  $df = 2$ , critical  $t$  drops to 4.303, with  $df = 5$  to 2.571, with  $df = 10$  to 2.228, with  $df = 20$  to 2.086, with  $df = 40$  to 2.021, with  $df = 100$  to 1.984, and with  $df$  uncountably large to 1.96. As  $df$  increase, Student's  $t$  becomes more and more nearly normal.

One of the parameters going into the probability density function of  $t$  is  **$df$ , degrees of freedom**. We start out with  $df = N$  and then we **lose one  $df$  for each parameter we estimate** when computing the standard error. We compute the sample standard error

as  $s_M = \frac{s}{\sqrt{N}}$ . When computing the sample  $s$  we estimate the population mean when

using  $(Y \text{ minus sample mean})$  rather than  $(Y \text{ minus } \mu)$  to compute the sum-of-squares. That one estimation cost us one  $df$ , so  **$df = N - 1$** . The fewer the  $df$ , the plumper the  $t$  is in its tails, and accordingly the greater the absolute critical value of  $t$ . With infinite  $df$ ,  $t$  has the same critical value as  $Z$ .

Here is an abbreviated table of critical values of  $t$  marking off the upper 2.5% of the area under the curve. Notice how the critical value is very large when  $df$  are small, but approaches 1.96 (the critical value for  $z$ ) as  $df$  increase.

$df$	1	2	3	10	30	100	$\infty$
Critical Value	12.706	4.303	3.182	2.228	2.042	1.984	1.960

When  $df$  are small, a larger absolute value of computed  $t$  is required to reject the null hypothesis. Accordingly, low  $df$  translates into low power. When  $df$  are low, sample size will be low too, and that also reduces power.

I shall illustrate the use of Student's  $t$  for testing a hypothesis about the mean score that my students in undergraduate statistics get on the math section of the Scholastic Aptitude Test. I shall use self-report data provided by students who took my undergraduate statistics class between 2000 and 2004. During that five year period the national mean score on the math SAT was 516. For North Carolina students it was 503. For the 114 students on whom I have data, the mean is 534.78 and the standard deviation is 93.385. I shall test the null hypothesis that the mean of the population from

which my students' scores were randomly drawn is 516. I shall employ the usual .05 criterion of statistical significance.

$s_M = \frac{93.385}{\sqrt{114}} = 8.746$ , and  $t = \frac{534.78 - 516}{8.746} = 2.147$ . To get  $P(t > 2.147)$  go to the  $t$  table.  $Df$  is 113, but there is not a row for  $df = 113$  in the table, so use the row for  $df = 100$ . Then go across looking for  $t = 2.147$ .  $T = 2.147$  is not there, but 1.984 and 2.364, between which 2.147 falls, are. Since these two  $t$ 's have two-tailed  $p$ 's of .05 and .02, our  $p$  is between .02 and .05. Were we doing a one-tailed test with correct prediction of the direction of effect, we would now say  $.01 < p < .025$ . In both cases,  $p \leq .05$ , and we reject the  $H_0$ . If we were doing a one-tailed test but the predicted direction were wrong,  $p$  would be 1 minus the value for the one-tailed  $p$  with direction correct, that is,  $.975 < p < .99$ . We can use SPSS or SAS to get the exact  $p$ , which is, for these data, .034.

A confidence interval should also be constructed.  $CI = M - CV * s_M \leftrightarrow M + CV * s_M$ . For  $CC = 95\%$ ,  $\alpha = 1 - .95 = .05$ ,  $= .025$  in upper tail. From the  $t$  table for  $df = 100$ ,  $CV = 1.984$ .  $CI_{.95} = 534.78 - 1.984(8.746) \leftrightarrow 534.78 + 1.984(8.746) \Rightarrow 517.43 \leftrightarrow 552.13$ .

### Effect Size

When you test a hypothesis about a population mean, you should report an estimate of  $(\mu - \mu_0)$ , where  $\mu$  is the true population mean and  $\mu_0$  is the hypothesized population mean, and you should put a confidence interval about that estimate. For our SAT data the estimated  $(\mu - \mu_0) = 534.78 - 516 = 18.78$ . Note that this is the numerator our the  $t$  ratio. To get a confidence interval for this difference, just take the confidence interval for the mean and subtract the hypothesized mean from both the lower and the upper limits. For our SAT data, the 95% confidence interval for  $(\mu - \mu_0)$  is 1.43 to 36.13.

When you are dealing with data where the unit of measurement is easily understood by most persons (such as inches, pounds, dollars, etc.), reporting an effect size in that unit of measurement is fine. Psychologists, however, typically deal with data where the unit of measurement is not so easily understood (such as score on a personality test). Accordingly, it useful to measure effect size in standard deviation units. The

standardized effect size parameter for the one-sample design is  $\delta = \frac{\mu - \mu_0}{\sigma}$ .

Confusingly,  $\delta$  is also used as the symbol for the noncentrality parameter, which we shall soon discuss.

We can estimate  $\delta$  with the statistic:  $d = \frac{M - \mu_0}{s} = \frac{18.78}{93.385} = .20$ . This statistic is also

known as **Hedges'  $g$**  (see McGrath & Meyer, 2006). You can also compute  $d$  simply by dividing the obtained  $t$  by the square root of the sample size. Our best point estimate of the amount by which my students' mean math SAT exceeds that of the national norm is 1/5 of a standard deviation. Jacob Cohen suggested that a  $d$  of 0.2 is small but not trivial, a  $d$  of 0.5 is medium sized, and a  $d$  of 0.8 is large. These are very general guidelines, and are not likely appropriate in all contexts.

Constructing a confidence interval for  $d$  is unreasonably difficult to do by hand, but easy to do with [SPSS](#) or [SAS](#). It involves an iterative procedure and use of the noncentral  $t$  distributions. Central  $t$  distributions have only one parameter,  $df$ . Noncentral  $t$  distributions have a second parameter, the noncentrality parameter. The noncentrality parameter is a function of the amount by which the null hypothesis is incorrect.

**Non-Normal Data.** Using the  $t$  table involves the assumption that  $Y$  is normally distributed. If that is not a reasonable assumption, and if the data cannot be transformed (via SQRT, LOG, etc.) to make them normal, one can use different statistical procedures called “nonparametric” or “distribution-free” tests that do not assume that  $Y$  is normally distributed (they do, however, have other assumptions which must be met). For more information on the use of data transformation to reduce skewness, see my documents [Using SAS to Screen Data](#) and [Using SPSS to Screen Data](#).

**Summary Statements.** When you test hypotheses for this class, I want you to state your conclusion as it would be stated in an APA (American Psychological Association) style journal. Your summary statement should include each of the following:

- who or what the research units were (sometimes called “subjects” or “participants”)
- what the null hypothesis was (implied, not explicitly stated)
- descriptive statistics such as sample sizes, means and standard deviations
- whether or not you rejected the null hypothesis (implied, not explicitly stated)
- if you did reject the null hypothesis, what was the observed direction of the difference between the obtained results and those expected under the null hypothesis
- what test statistic (such as  $t$ ) was employed
- the degrees of freedom
- if not obtainable from the degrees of freedom, the sample size
- the computed value of the test statistic
- the  $p$  value (use SPSS or SAS to get an exact  $p$  value)
- an effect size estimate
- and a confidence interval for the effect size parameter.

For the  $t$ -test that we did earlier, here is a APA-style summary statement: **The mean math SAT of my undergraduate statistics students ( $M = 535$ ,  $SD = 93.4$ ) was significantly greater than the national norm (516),  $t(113) = 2.147$ ,  $p = .034$ ,  $d = .20$ . A 95% confidence interval for the mean runs from 517 to 552. A 95% confidence interval for  $d$  runs from .015 to .386.**

Suppose that the sample mean we obtained was not 534.78 but 532. Our summary statement would read: **The mean math SAT of my undergraduate statistics students ( $M = 532$ ,  $SD = 93.4$ ) was not significantly different from the national norm (516),  $t(113) = 1.83$ ,  $p = .07$ ,  $d = .17$ . A 95% confidence interval for the mean runs from 515 to 549. A 95% confidence interval for  $d$  runs from -.014 to .356.** Note that I did not indicate a direction of difference with this “nonsignificant” result -- to do so would imply that I was testing directional rather than nondirectional hypotheses.

Suppose that I was testing directional hypotheses, with the alternative hypothesis being that the mean is greater than 516, and the obtained sample mean being 532. Now my summary statement would read: **Employing a one-tailed test, the mean math SAT of my undergraduate statistics students ( $M = 532$ ,  $SD = 93.4$ ) was significantly greater than the national norm (516),  $t(113) = 1.83$ ,  $p = .035$ ,  $d = .17$ . A 90% confidence interval for the mean runs from 517 to 547. A 90% confidence interval for  $d$  runs from .016 to .326.** Notice that I shifted to a 90% confidence interval, because with the one-tailed test I put all of alpha in one tail rather than splitting it into two tails – but confidence intervals are, IMHO, naturally bidirectional, so I put 5% in both tails for the confidence interval. If I did not make this change in the confidence coefficient, the confidence interval would include the null value, which would be in disagreement with the prior decision to reject the null hypothesis.

Suppose that the mean was only 530, still testing the directional hypotheses. Now my summary statement would read: **Employing a one-tailed test, the mean math SAT of my undergraduate statistics students ( $M = 530$ ,  $SD = 93.4$ ) was not significantly greater than the national norm (516),  $t(113) = 1.60$ ,  $p = .057$ ,  $g = .15$ . A 90% confidence interval for the mean runs from 515 to 545. A 90% confidence interval for  $d$  runs from -.005 to .305.** Even though the result is not “significant,” I use the phrase “not significantly greater than” rather than “not significantly different from” because the test was directional.

#### *Reference*

McGrath, R. E. & Meyer, G. J. (2006). When effect sizes disagree: The case of  $r$  and  $d$ . *Psychological Methods*, 11, 386-401. doi: 10.1037/1082-989X.11.4.386

#### **Table from McGrath & Meyer**

Table 1  
*Comparison of Symbols Used for Standardized Mean Difference Statistics*

Study	Parameter	Sample statistics		
		Pooled within sample sums of squares divided by:		Corrected for bias
		$N$	$N - 2$	
Cohen (1988)	$d$		$d_s$	
Hedges & Olkin (1985)	$\delta$	$\hat{\delta}^a$	$g$	$d$
Hunter & Schmidt (2004)	$\delta$	$d^a$	$d$	$d^*$
Lipsey & Wilson (2001)			$ES_{SM}$	
Rosenthal (1991)	$d$	$d$	$g$	$g^a$
Recommended	$\delta$	$d$	$g$	$\hat{\delta}$

<sup>a</sup> Discussed only as the maximum likelihood estimate of the population value.

Copyright 2017, Karl L. Wuensch - All rights reserved.