Two-Mean Inference®

Two-Group Research

1. We wish to know whether two groups (samples) of scores (on some continuous OV, outcome variable) are different enough from one another to indicate that the two populations from which they were randomly drawn are also different from one another.

2. The two groups of scores are from research units (subjects) that differ with respect to some dichotomous GV, grouping variable (treatment).

3. We shall compute an exact significance level, $p$, which represents the likelihood that our two samples would differ from each other on the OV as much (or more) as they do, if in fact the two populations from which they were randomly sampled are identical, that is, if the dichotomous GV has no effect on the OV mean.

Research Designs

1. In the Independent Sampling Design (also know as the between-subjects design) we have no good reason to believe there should be correlation between scores in the one sample and those in the other. With experimental research this is also known as the completely randomized design -- we not only randomly select our subjects but we also randomly assign them to groups - the assignment of any one subject to group A is independent of the assignment of any other subject to group A or B. Of course, if our dichotomous GV is something not experimentally manipulated, such as subject’s sex/gender, we do not randomly assign subjects to groups, but subjects may still be in groups in such as way that we expect no correlation between the two groups’ scores.

2. In the Matched Pairs Design (also called a randomized blocks design, a related samples design, or a correlated samples design) we randomly select pairs of subjects, with the subjects matched on some extraneous variable (the matching variable) thought to be well correlated with the dependent variable. Within each pair, one subject is randomly assigned to group A, the other to group B. Again, our dichotomous GV may not be experimentally manipulated, but our subjects may be matched up nevertheless—for example, GV = sex, subjects = married heterosexual couples.

   a. If the matching variable is in fact well correlated with the dependent variable, the matched pairs design should provide a more powerful test (greater probability of rejecting the null hypothesis) than will the completely randomized design. If not, it may yield a less powerful test.

   b. One special case of the matched pairs design is the Within Subjects Design (also known as the repeated measures design). Here each subject generates two scores: one after treatment A, one after treatment B. Treatments are counterbalanced so that half the subjects get treatment A first, the other half receiving treatment B first, hopefully removing order effects.

The Correlated Samples $t$ Test

The matched pairs design has the less complex analysis. We have two scores from each pair, one under condition 1, another under condition 2. For example, we investigate reaction time scores as affected by alcohol. The null hypothesis is that $\mu_{sober} = \mu_{drunk}$, that alcohol doesn’t affect reaction time. We create a new variable, $D$. For each pair we compute $D = X_1 - X_2$. The null hypothesis becomes $\mu_D = 0$ and we test it exactly as we previously tested one mean hypotheses,
including using one-tailed tests if appropriate [if the alternative hypothesis is \( \mu_D > 0 \), that is \( \mu_1 > \mu_2 \), or if it is \( \mu_D < 0 \), \( \mu_1 < \mu_2 \)].

**Calculation of the Related Samples \( t \)**

**The Independent Samples \( t \) Tests**

The **independent samples design** is more complex. The sampling distribution is the distribution of differences between means, which has a mean equal to \( \mu_1 - \mu_2 \). By the variance sum law, the standard deviation of the sampling distribution, the **Standard Error Of Differences Between Means**, is the square root of the variance of the sampling distribution:

\[
\sigma_{M_1-M_2} = \sqrt{\sigma_{M_1}^2 + \sigma_{M_2}^2 - 2 \rho \sigma_{M_1} \sigma_{M_2}}
\]

This formula for the standard error actually applies to both matched pairs and independent sampling designs. The \( \rho \) (rho) is the correlation between scores in population 1 and scores in population 2. In matched pairs designs this \( \rho \) should be positive and fairly large, assuming that the variable used to match scores is itself positively correlated with the dependent variable. That is, pairs whose Group 1 score is high should also have their Group 2 score high, while pairs whose Group 1 score is low should have their Group 2 score low, relative to other within-group scores. The larger the \( \rho \), the smaller the standard error, and thus the more powerful the analysis (the more likely we are to reject a false null hypothesis). Fortunately there is an easier way to compute the standard error with matched pairs, the difference score approach we used earlier.

In the independent sampling design we assume that \( \rho = 0 \), so the standard error becomes:

\[
\sigma_{M_1-M_2} = \sqrt{\frac{\sigma_{M_1}^2}{N_1} + \frac{\sigma_{M_2}^2}{N_2}}
\]

where the variances are population variances. We could then test the null hypothesis that \( \mu_1 = \mu_2 \) with:

\[
Z = \frac{M_1 - M_2}{\sigma_{M_1-M_2}}
\]

assuming that we know the population variances. Since we are not likely to know the population variances (or, with matched samples, the population rho) when making inferences about population means, we must estimate them with sample variances (and estimate \( \rho \) with \( r \)). Assuming that \( n_1 = n_2 \), we use the same formulas shown above, except that we substitute \( s \) for \( \sigma \) and \( s^2 \) for \( \sigma^2 \) and \( r \) for \( \rho \) and the test statistic is not the normally distributed \( Z \) but is rather Student’s \( t \), assuming that the dependent variable is **normally distributed** in both populations. The \( t \) is evaluated on \( n_1 + n_2 - 2 \) degrees of freedom if you can assume that the two populations have identical variances, the **homogeneity of variance** assumption. Such a test is called the **pooled variances \( t \)-test**. If \( n_1 = n_2 = n \), then:

\[
s_{M_1-M_2} = \sqrt{\frac{s_1^2 + s_2^2}{n}}\text{ and } t = \frac{M_1 - M_2}{s_{M_1-M_2}}
\]

If \( n_1 \neq n_2 \), the pooled variances standard error requires a more elaborate formula. Given the homogeneity of variance assumption, we can better estimate the variance of the two populations by using \((n_1 + n_2)\) scores than by using the \( n_1 \) and the \( n_2 \) scores separately. This involves pooling the sums of squares when computing the standard error:

\[
s_{M_1-M_2} = \sqrt{\frac{SS_1 + SS_2}{n_1 + n_2 - 2} \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}
\]

Remember, \( SS = s^2 (n - 1) \).

The \( t \) that you obtain is then evaluated using \( df = n_1 + n_2 - 2 \).
If you cannot assume homogeneity of variance, then using the pooled variances estimate is not reasonable. Instead, compute \( t \) using the separate variances error term,

\[
s_{M_1-M_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.
\]

T', the separate variances \( t \), is not, however, distributed the same as the pooled variances \( t \). Behrens and Fisher have tabled critical values of \( t' \) and Cochran and Cox invented a formula with which one can approximate the critical value of \( t' \) (see the formula on p. 186 of the 3rd edition of Howell if you are interested in this approach). So, when should one use the separate variances \( t \)? One "rule of thumb" which I have employed is: "If the ratio of the larger variance to the smaller variance exceeds 4 or 5, one should not pool variances, especially if sample sizes are also greatly unequal." However, Monte Carlo work by Donald W. Zimmerman (1996) has indicated that two stage testing (comparing the variances to determine whether to conduct a pooled test or a separate variances test) is not a good procedure, especially when the sample sizes differ greatly (3 or 4 times more subjects in one group than in the other, in which case the pooled test performs poorly even when the ratio of variances is as small as 1.5). Zimmerman was, by the way, a faculty member in our department here at ECU many years ago. Zimmerman's advice is that the separate variances \( t \) should be applied unconditionally whenever sample sizes are unequal. Given the results of his Monte Carlo study, I think this is good advice, and I suggest that you adopt the practice of using the separate variances test whenever you have unequal sample sizes. I still believe that the pooled test may be appropriate (and more powerful) when your sample sizes are nearly equal and the variances not greatly heterogeneous, but carefully defining "nearly equal sample sizes" and "not greatly heterogeneous variances" is not something I care to tackle.

I prefer the Welch-Satterthwaite solution over the Behrens and Fisher or Cochran and Cox procedures. With this solution one adjusts the \( df \) downwards to correct for the amount of heterogeneity of variance indicated by the samples (the greater the heterogeneity, the more the \( df \) lost) and then uses a standard \( t \)-table or \( t \) probability density function. The formula for \( df' \) appears in your textbook (as does a slight modification suggested by Welch in 1947). Note that \( df' \) is never smaller than the smaller of \( (n_1 - 1) \) and \( (n_2 - 1) \), so if \( t' \) is significant with the smaller of those, it is also significant with the completely adjusted \( df' \). The \( df' \) is never larger than \( (n_1 + n_2 - 2) \), so if \( t' \) is not significant at \( (n_1 + n_2 - 2) \) \( df \), it is not significant at the completely adjusted \( df' \) either.

Heterogeneity of variance may also be removed by transforming the data prior to analysis. For example, square root and logarithmic transformations may reduce the heterogeneity of variance. These transformations may also reduce positive skewness in the data, helping to meet the normality assumption.

The \( t \)-test is often considered robust, that is, little affected by violations of its assumptions, within limits. With equal sample sizes, violation of the homogeneity of variance assumption has little effect on the validity of the \( p \) obtained from a pooled \( t \). If sample sizes are unequal and there is considerable heterogeneity of variance, the pooled \( t \) should not be trusted, but the separate variances \( t' \) with adjusted \( df \) should still be OK. If both populations have similar shapes, or if both are symmetric even if their shapes differ, violating the normality assumption should not be too serious -- but see Bradley (1982).

Earlier I noted that use of a matched-pairs design may provide a more powerful analysis than use of the independent sampling design, since the standard error in the matched pairs \( t \) test is reduced by the term \( \sqrt{-2 \rho \sigma_{M_1} \sigma_{M_2}} \). Please note that the matched pairs test also involves a reduction in \( df \) from \( (n_1 + n_2 - 2) \) to \( (n - 1) \), that is, for the same total number of scores, the matched \( t \) has half the \( df \) of the independent \( t \). If total \( N < 30 \), loosing half the \( df \) produces a considerable increase in the critical value of \( t \), which reduces power. One generally hopes that the reduction in
power caused by the loss of \( df \) is more than offset by the increase in power resulting from having a smaller standard error (which increases the computed value of \( t \)).

One may compute a confidence interval for \( \mu_1 - \mu_2 \) by:

\[
CI = (M_1 - M_2) \pm t_{critical} \cdot s_{M_1-M_2}
\]

It is good practice always to report a confidence interval for the difference between means.

**Effect Size Estimates**

You should always provide an estimate of the size of the effect that you are reporting. There are several effect size estimates available for two group designs. I recommend that you use the standardized difference between group means. You should also present a confidence interval for the effect size.

**Cohen’s \( \delta \).** \( \delta = \frac{\mu_1 - \mu_2}{\sigma} \). This is the parameter. Most use “\( d \)” for the statistic.

You should memorize the following benchmarks for \( \delta \), but keep in mind that they are not appropriate in all contexts:

<table>
<thead>
<tr>
<th>Size of effect</th>
<th>( \delta )</th>
</tr>
</thead>
<tbody>
<tr>
<td>small</td>
<td>.2</td>
</tr>
<tr>
<td>medium</td>
<td>.5</td>
</tr>
<tr>
<td>large</td>
<td>.8</td>
</tr>
</tbody>
</table>

Please see my document [Cohen’s Conventions for Small, Medium, and Large Effects](#).

**Estimated Cohen’s \( \delta \), \( d \).** The parameter being estimated here is \( \delta = \frac{\mu_1 - \mu_2}{\sigma} \). Our estimator is \( d = \frac{M_1 - M_2}{s_{pooled}} \), where the pooled standard deviation is the square root of the within groups mean square (from a one-way ANOVA comparing the two groups). If you have equal sample sizes, the pooled standard deviation is \( s_{pooled} = \sqrt{\frac{s_1^2 + s_2^2}{2}} \). If you have unequal sample sizes, \( s_{pooled} = \sqrt{\frac{\Sigma (p_j s_j^2)}{\Sigma p_j}} \), where for each group \( s_j^2 \) is the within-group variance and \( p_j = \frac{n_j}{N} \), the proportion of the total number of scores (in both groups, \( N \)) which are in that group (\( n_j \)). You can also compute \( d \) as \( d = \frac{t \sqrt{n_1 + n_2}}{\sqrt{n_1 n_2}} \), where \( t \) is the pooled variances independent samples \( t \) comparing the two group means.

You can use the program [Conf Interval-d2.sas](#) to obtain the confidence interval for the standardized difference between means. It will require that you give the sample sizes and the values of \( t \) and \( df \). Use the pooled variances values of \( t \) and \( df \). Why the pooled variances \( t \) and \( df \)? See [Confidence Intervals, Pooled and Separate Variances T](#). Also see [Standardized Difference Between Means, Independent Samples](#). Also see [Getting the CI with SPSS or R](#).
I shall illustrate using the Howell data (participants were students in Vermont), comparing boys’ GPA with girls’ GPA. Please look at the computer output. For the girls, $M = 2.82$, $SD = .83$, $n = 33$, and for the boys, $M = 2.24$, $SD = .81$, $n = 55$.

$$s_{pooled} = \sqrt{\frac{33}{88} (.83)^2 + \frac{55}{88} (.81)^2} = .818.$$  

$$\hat{d} = \frac{2.82 - 2.24}{.818} = .71.$$ Also, $$\hat{d} = \frac{3.267 \sqrt{33} + 55}{\sqrt{33(55)}} = .72$$ (there is a little rounding error in the earlier computations).

**Glass’ Delta.** $\Delta = \frac{M_1 - M_2}{s_{control}}$. That is, in computing the standardized difference between group means, we use the control group standard deviation in the denominator. You want the standardizer to be a good estimate of the population to which you wish to generalize your results. Your treatment is an experimental manipulation that does not currently exist outside of the lab. Accordingly, that treatment does not account for any of the variance in the population of interest. The standard deviation of the control group would likely be the better estimate of the standard deviation in the population of interest, so use the control group standard deviation in the denominator.

You are most likely to consider using Glass’ delta after noticing that your treatment has affected the variance in the dependent variable. In this case, the observed heterogeneity of variance should be treated as an effect of interest, not just a nuisance. Suppose that educational achievement is the dependent variable and the treatment increased the mean and reduced the variance. This might result from the treatment increasing the achievement of the students who otherwise would have performed poorly but having little effect on the students who otherwise would have performed well. Suppose that the treatment did not much affect the mean but increased the variance. This might result from the treatment increasing performance in one subset of the students but decreasing performance in the other subset. In this case one should attempt to identify the subsets so that future application of the treatment will directed only to those it is likely to help.

**Point-Biserial $r$.** This is simply the Pearson $r$ between the grouping variable (coded numerically) and the criterion variable. It can be computed from the pooled variances independent $t$:

$$r_{pb} = \frac{t}{\sqrt{t^2 + df}}.$$  

For the comparison between girls’ and boys’ GPS, $r_{pb} = \frac{3.267}{\sqrt{3.267^2 + 86}} = .332$. This is the standardized slope for the regression line for predicting the criterion variable from the grouping variable. The unstandardized slope is the difference between the group means. We standardize this difference by multiplying it by the standard deviation of the grouping variable and dividing by the standard deviation of the criterion variable. For our comparison, $r_{pb} = \frac{.588(.487)}{.861} = .33$.

**Eta-squared.** For a two-group design, this is simply the squared point-biserial correlation coefficient. It can be interpreted as the proportion of the variance in the criterion variable which is explained by the grouping variable. For our data, $\eta^2 = .11$. For a confidence interval, use my program **Conf-Interval-R2-Regr.sas**. It will ask you for $F$ (enter the square of the pooled $t$), $df\_num$ (enter 1), and $df\_den$ (enter the $df$ for the pooled $t$). For our data, a 95% confidence interval runs from .017 to .240.

**Omega-squared.** Eta-squared is a biased estimator, tending to overestimate the population parameter. Less biased is the omega-squared statistic, which we shall study when we cover one-way independent samples ANOVA.
Common Language Effect Size Statistic. See [http://core.ecu.edu/psyc/wuenschk/docs30/CL.pdf](http://core.ecu.edu/psyc/wuenschk/docs30/CL.pdf). CL is the probability that a randomly selected score from the one population will be greater than a randomly sampled score from the other distribution. Compute \( Z = \frac{|M_1 - M_2|}{\sqrt{S_1^2 + S_2^2}} \) and then find the probability of obtaining a \( Z \) less than the computed value. For the data here, \( Z = \frac{|2.82 - 2.24|}{\sqrt{.83^2 + .81^2}} = 0.50 \), which yields a lower-tailed \( p \) of .69. That is, if one boy and one girl were randomly selected, the probability that the girl would have the higher GPA is .69. If you prefer odds, the odds of the girl having the higher GPA = .69/(1-.69) = 2.23 to 1.

Point Biserial \( r \) versus Estimated \( d \). Each of these has its advocates. Regardless of which you employ, you should be aware that the ratio of the two sample sizes can have a drastic effect on the value of the point-biserial \( r \) (and the square of that statistic, which is \( \eta^2 \)), but does not affect the value of estimated \( d \). See [Effect of \( n_1/n_2 \) on Estimated \( d \) and \( r_{pb} \)](http://core.ecu.edu/psyc/wuenschk/docs30/CL.pdf).

Correlated Samples Designs. You could compute \( \hat{d}_{Diff} = \frac{M_1 - M_2}{s_{Diff}} \), where \( s_{Diff} \) is the standard deviation of the difference scores, but this would artificially inflate the size of the effect, because the correlation between conditions will probably make \( s_{Diff} \) smaller than the within-conditions standard deviation. You should instead treat the data as if they were from independent samples. If you base your effect size estimate on the correlated samples analysis, you will overestimate the size of the effect. You cannot use my [Conf_Interval-d2.sas](http://core.ecu.edu/psyc/wuenschk/docs30/CL.pdf) program to construct a confidence interval for \( d \) when the data are from correlated samples. See my document [Confidence Interval for Standardized Difference Between Means, Related Samples](http://core.ecu.edu/psyc/wuenschk/docs30/CL.pdf) for details on how to construct an approximate confidence interval for the standardized difference between related means and Algina’s SAS programs for constructing confidence intervals for the standardized difference between related means.

You can use the apps linked below to compute \( d \) (but not the confidence interval) from means and standard deviations, as long as the samples sizes are equal, as they always will be with correlated samples.

- [Excel](http://core.ecu.edu/psyc/wuenschk/docs30/CL.pdf)
- [SAS](http://core.ecu.edu/psyc/wuenschk/docs30/CL.pdf)
- [SPSS](http://core.ecu.edu/psyc/wuenschk/docs30/CL.pdf)

Importance of Providing Effect Size Estimates. IMHO, effect size estimates, especially when presented as confidence intervals, are a lot more informative than are \( p \) values. Consider the research of Kramer, Guillory, and Hancock (2014). They manipulated the number of positive and negative entries in Facebook users’ new feeds and estimated the size of the effect of that manipulation on the subsequent number of positive and negative posts by those users. Do note that the only consent obtained from these users was in the fine print that nobody ever reads (basically, you consent to let Facebook do whatever it wants to with your posts). The sample size was 689,003. Such an enormous sample size almost certainly lead to “statistically significant” effects, regardless of whether or not the actual effect size in trivial in magnitude. Accordingly, getting a small \( p \) value is pretty much meaningless. If you look at page 8789 you will see that the estimates of Cohen’s \( d \) ranged from .008 to .02. Cohen suggested the values of \( d \) below .2 represent trivial effects, and these are closer to zero than to .2. If you were to put confidence intervals on these effect size
estimates, it would be apparent that the conclusion to be drawn from this research is that the effects of the manipulations were so very small that they might as well be zero.

**Tests of Equivalence.** Sometimes we want to test the hypothesis that the size of an effect is not different from zero by more than a trivial amount. For example, we might wish to test the hypothesis that the effect of a generic drug is equivalent to the effect of a brand name drug. Please read my document *Tests of Equivalence and Confidence Intervals for Effect Sizes.*

**Testing Variances**

One may be interested in determining the effect of some treatment upon variances instead of or in addition to its effect on means. Suppose we have two different drugs, each thought to be a good treatment for lowering blood cholesterol. Suppose that the mean amount of cholesterol lowering for drug A was 40 with a variance of 100 and for drug B the mean was 42 with a variance of 400. The difference in means is trivial compared to the difference in variances. It appears that the effect of drug A does not vary much from subject to subject, but drug B appears to produce very great lowering of blood cholesterol for some subjects, but none (or even elevated cholesterol) for others. At this point the researcher should start to look for the mystery variable which interacts with drug B to determine whether B’s effect is positive or negative.

To test the null hypothesis that the treatments do not differ in effect upon variance, that is, \( \sigma^2_A = \sigma^2_B \), one may use an *F-test*. Simply divide the larger variance by the smaller, obtaining an *F* of 400/100 = 4.0. Suppose we had 11 subjects in Group A and 9 in Group B. The numerator (variance for B) degrees of freedom is \( n_b - 1 = 8 \), the denominator (variance for A) *df* is \( n_a - 1 = 10 \). From your statistical program [ in SAS, \( p = 2 \times (1 - \text{PROBF}(4, 8, 10)) \) ] you obtain the two-tailed probability for \( F(8, 10) = 4 \), which is \( p = .044 \).

We can do one-tailed tests of *directional hypotheses* about the relationship between two variances. With directional hypotheses we must put in the numerator the variance which we predicted (in the alternative hypothesis) would be larger, even if it isn’t larger. Suppose we did predict that \( \sigma^2_A > \sigma^2_B \). \( F(8, 10) = 4.00 \), we don’t double *p*, \( p = .022 \). What if we had predicted that \( \sigma^2_B > \sigma^2_A \)? \( F(10, 8) = 100 / 400 = 0.25 \). Since the *p* for \( F(x, y) \) equals 1 minus the *p* for 1 / \( F(y, x) \), our *p* equals 1 - .022 = .98, and the null hypothesis looks very good. If you wish, you can use SAS to verify that \( p = 1 - \text{PROBF}(.25, 10, 8) \); returns a value of .98.

Although *F* is often used as I have shown you here, it has a robustness problem in this application. It is not robust to violations of its normality assumption. There are, however, procedures that are appropriate even if the populations are not normal. Levene suggested that for each score you find either the square or the absolute value of its deviation from the mean of the group in which it is and then run a standard *t*-test comparing the transformed deviations in the one group with those in the other group. Brown and Forsythe recommended using absolute deviation from the median or a trimmed mean. Their Monte Carlo research indicated that the trimmed mean was the best choice when the populations were heavy in their tails and the median was the best choice when the populations were skewed. Levene’s tests can be generalized to situations involving more than two populations – just apply an ANOVA to the transformed data. Please consult the document *Levene Test for Equality of Variances*. Another alternative, *Obrien’s test*, is illustrated in the 4th edition of Howell’s *Statistical Methods for Psychology*. As he notes in the 5th edition, it has not been included in mainstream statistical computing packages.

To test the null hypothesis of homogeneity of variance in two related (not independent) samples, use E. J. G. Pitman’s test (A note on normal correlation, *Biometrika*, 1939, 31, 9-12) *t*:
\[ t = \frac{(F - 1)\sqrt{n - 2}}{2\sqrt{F(1 - r^2)}} \]

where \( F \) is the ratio of the larger to the smaller sample variance, \( n \) is the number of pairs of scores, \( r \) is the correlation between the scores in the one sample and the scores in the other sample, and \( n - 2 \) is the \( df \).

**Assumptions of Parametric Tests**

Parametric tests are those for which the hypothesis tested specifies that some parameter has some particular value – for example, \( \mu = 100 \), or, for two-group research, \( (\mu_1 - \mu_2) = 0 \). Such parametric tests (including the \( t \) tests we have been studying) typically have certain assumptions which must be met if the obtained \( p \) value is to be correct. For the \( t \) tests, we assume that the population distribution(s) of scores is(are) normally distributed. For the pooled \( t \) test, we also assume that the two populations of scores have identical variances. Some parametric procedures are considered to be somewhat “robust” to violation of their assumptions – that is, the assumptions can be somewhat violated without having a great effect on the validity of the obtained \( p \) value.

**Testing Variances Prior to Testing Means.** Some researchers have adopted the bad habit of using a test of variances to help decide whether to use a pooled \( t \) test or a separate variances \( t \) test. This is now recognized to be poor practice. Box (1953) was an early critic of testing variances prior to conducting a test of means. He wrote “to make the preliminary test on variances is rather like putting to sea in a rowing boat to find out whether conditions are sufficiently calm for an ocean liner to leave port.” More recently, several statisticians have advised against the preliminary testing of variances, including Delacre, Lakens, & Leys (2017), Moser & Stevens (1992), Rash, Kublinger & Moder (2011), and Zimmermann (2004). Heterogeneity of variance is often accompanied by non-normal distributions, and some tests of variances are often not robust to their normality assumption, but even the preliminary use of robust tests of homogeneity of variance is no longer advised.

With large sample sizes the pooled \( t \) test is relatively robust to violations of its homogeneity of variance assumption, but those large sample sizes also make statistical tests of those assumptions so powerful that they will detect as “significant” even trivial differences in variances, differences of no consequence. When sample sizes are small, the pooled \( t \) test is less robust, and tests of variances have little power, so such tests may not detect as significant differences in variance large enough to be of concern. If you are going to use an independent samples \( t \) test, the best procedure may be use the separate variances \( t \) regardless of whether or not the sample variances differ much.

**Testing For Normality Prior to Testing Means.** There are procedures that can test the null hypothesis that the data were sampled from a normal distribution. In the past, some have recommended that these procedures be used to determine whether or not a standard \( t \) test should be employed with no transformation of the data. Such a two-step procedure is now considered unwise (Delacre, Lakens, & Leys, 2017; Moser & Stevens, 1992; Rash, Kublinger & Moder, 2011; Schucany, & Ng, 2006). Tests of normality are available in SAS Proc Univariate. The problem here is the same as that with testing for homogeneity of variance.

These tests will have very much power when sample size is large, and thus will detect as significant even very small differences in variance, differences that are of no concern given the pooled \( t \) test’s great robustness when sample sizes are large.
Writing an APA Style Summary for Two-Group Research Results

Using our example data, a succinct summary statement should read something like this: Among Vermont school-children, girls’ GPA \((M = 2.82, SD = .83, N = 33)\) was significantly higher than boys’ GPA \((M = 2.24, SD = .81, N = 55)\), \(t(65.9) = 3.24, p = .002, d = .72\). A 95% confidence interval for the difference between girls’ and boys’ mean GPA runs from .23 to .95 in raw score units and from .27 to 1.16 in standardized units.

Please note the following important components of the summary statement:

- The subjects are identified.
- The variables are identified: Method of motivation and time to cross the finish line.
- The group means, standard deviations, and sample sizes are given.
- Rejection of the null hypothesis is indicated (the difference is significant).
- The direction of the significant effect is indicated.
- The test statistic \((t)\) is identified, and its degrees of freedom, computed value, and \(p\)-value are reported.
- An effect size estimate is reported.
- Confidence intervals are reported for the difference between means and for \(d\).

The style for reporting the results of correlated \(t\) test would be the same.

If the result were not significant, we would not emphasize the direction of the difference between the group means, unless we were testing a directional hypothesis. For example, among school-children in Vermont, the IQ of girls \((M = 101.8, SD = 12.7, N = 33)\) did not differ significantly from that of boys \((M = 99.3, SD = 13.2, N = 55)\), \(t(69.7) = 0.879, p = .38, d = .19\). A 95% confidence interval for the difference between girls’ and boys’ mean IQ runs from -3.16 to 8.14 in raw score units and from -.24 to .62 in standardized units.

As an example of a nonsignificant test of a directional hypothesis: As predicted, the GPA of students who had no social problems \((M = 2.47, SD = 0.89, N = 78)\) was greater than that of students who did have social problems \((M = 2.39, SD = .61, N = 10)\), but this difference fell short of statistical significance, one-tailed \(t(14.6) = 0.377, p = .36, d = .09\). A 95% confidence interval for the difference between mean GPA of students with no social problems and that of students with social problems runs from -.38 to .54 in raw score units and from -.56 to .75 in standardized units.

References and Suggested Readings


Wuensch, K. L. (2009). The standardized difference between means: Much variance in notation. Also, differences between g and rpb as effect size estimates. Available [here](#).


**Links to Related Resources**

- [Complete Example of Independent Samples t Test](#)
- [Visualization Relating Cohen's d to Other Measures of Effect Size](#)
- [Other Two-Mean Effect Size Estimators](#) – Follow the link to Chapter 4
- [Summary of Effect Size Estimates](#) -- Lee Becker at the Univ. of Colorado, Colorado Springs
- [Two Groups and One Continuous Variable](#)
- [The Moments of Student's t Distribution](#)
- [Return to My Statistics Lessons Page](#)

Copyright 2018, Karl L. Wuensch - All rights reserved.

[Fair Use of This Document](#)