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 Samples Design** (also known as the between-subjects design) the observations in the one group are independent of those in the other group. If the research is experimental, we not only randomly select our subjects but we also randomly assign them to the treatment groups – thus, the name "completely randomized design." 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, 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 outcome 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 couples.

 a. If the matching variable is in fact well correlated with the outcome variable, the matched pairs design should provide a more powerful test (greater probability of rejecting the null hypothesis) than will the independent samples 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.

Matched Pairs *t*

 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 $D = Y_1 - Y_2$. 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 $\mu_D < 0$, $\mu_1 < \mu_2$].

Remember that null hypotheses always include an equality sign, for example, $\mu_1 = \mu_2$, or $\mu_1 \geq \mu_2$, or $\mu_1 \leq \mu_2$. Alternative hypotheses never include an equality sign, for example, $\mu_1 \neq \mu_2$, or $\mu_1 > \mu_2$, or $\mu_1 < \mu_2$.

This analysis is aka the related samples t, the correlated samples t, the within-subjects t, and the repeated measures t.

This test is appropriate whenever data have been collected in a way that results in the two samples of scores being related to each other. Suppose that Farmer Sol is evaluating the yield of two varieties of cucumbers that produce superior fruits, the SuperCuke and the Cukazilla. He has 12 small plots of land on which he is going to plant cucumbers. Each plot into split into two halves, and one half is planted with SuperCukes and the other half with Cukazilla, with assignment to halves, within each plot, being made randomly. Sol records how many bushels of cucumbers he harvests from each split plot. Here are the data.

Plot	SuperCuke	Cukazilla	Difference
A	1.4	1.6	.2
B	1.8	1.8	.0
C	2.3	2.2	-.1
D	2.7	2.7	.0
E	3.3	3.4	.1
F	3.4	3.7	.3
G	3.8	3.5	-.3
H	4.1	4.4	.3
I	4.7	5.0	.3
J	4.9	5.1	.2
K	5.8	6.1	.3
L	6.0	6.5	.5
Mean	3.683	3.833	0.15
SD	1.4899	1.6216	.2195

Since the data appear to be approximately normally distributed, Sol decides to employ a related samples t test to determine whether or not there is a significant difference between the two types of cucumbers, with respect to yield. The correlated samples t is computed by taking the difference between the two means and dividing by the standard error of the mean for the difference scores.

$$t = \frac{M_1 - M_2}{s_{Diff} / \sqrt{n}} = \frac{0.15}{.2195 / \sqrt{12}} = \frac{0.15}{.063364} = 2.367$$

A table in the back of Sol's stats text tells him that on 11 degrees of freedom one has to obtain a *t* whose absolute value is at least 2.201 to claim significance at the traditional .05 level. Since Sol's obtained *t* exceeds that critical value, he concludes that the yield of Cukazilla cucumbers is significantly greater than that of Supercuke cucumbers.

Sol also should construct a 95% confidence interval for the difference between the means. First he needs to find the two-tailed critical value of t for alpha = 1-.95 = .05 and *df* = 11. That is the same critical value used with the test of the null hypothesis, 2.210. The confidence interval is now the difference in means plus/minus the critical value of *t* times the standard error: $0.15 \pm 2.210(.063364)$ = $0.15 \pm .140$ = [.01, .19]. Notice that the confidence interval excludes zero, which will be the case when the difference between means was statistically significant.

Independent Samples *t*

The **independent sampling 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 (in footnote on page 354 of Howell, 7th ed.) 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 ρ (rho) is the correlation between scores in population 1 and scores in population 2. In matched pairs designs this ρ 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 ρ, 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:

$\sigma_{M_1-M_2} = \sqrt{\dfrac{\sigma_1^2}{N_1} + \dfrac{\sigma_2^2}{N_2}}$ where the variances are population variances. We could then test the null

hypothesis that $\mu_1 = \mu_2$ with: $Z = \dfrac{M_1 - M_2}{\sigma_{M_1-M_2}}$ 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 ρ with r). Assuming that $n_1 = n_2$, we use the same formulas shown above, except that we substitute s for σ and s^2 for σ^2 and r for ρ 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: $s_{M_1-M_2} = \sqrt{\dfrac{s_1^2 + s_2^2}{n}}$ and $t = \dfrac{M_1 - M_2}{s_{M_1-M_2}}$

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

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

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**, $s_{M_1-M_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$. If you do not have a computer available to conduct this test, you can evaluate this t with df equal to the smaller of $(n_1 - 1)$ and $(n_2 - 1)$. This is a very conservative procedure, meaning that alpha is probably even less than the a priori criterion you used. If you do have a computer available, use SPSS or another statistical program to conduct the analysis. It will adjust the degrees of freedom downwards to correct for any problem with heterogeneity of variances.

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, work by **Donald W. Zimmerman** (Some properties of preliminary tests of equality of variances in the two-sample location problem, *Journal of General Psychology*, 1996, *123*, 217-231) 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 as many 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.

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.

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 μ_1 - μ_2 by:

$$CI = (M_1 - M_2) \pm t_{critical} * s_{M_1 - M_2}$$

Computation of *t* by Hand.

See my slideshow .

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 *d*. The parameter being estimated here is $\delta = \dfrac{\mu_1 - \mu_2}{\sigma}$. Our estimator is $\hat{d} = \dfrac{M_1 - M_2}{s_{pooled}}$, also known as Hedges' *g*, 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 $s_{pooled} = \sqrt{.5(s_1^2 + s_2^2)}$. If you have unequal sample sizes, $s_{pooled} = \sqrt{\Sigma(p_j s_j^2)}$, where for each group s_j^2 is the within-group variance and $p_j = \dfrac{n_j}{N}$, the proportion of the total number of scores (in both groups, *N*) which are in that group (n_j). You can also estimate d as $\hat{d} = \dfrac{t\sqrt{n_1 + n_2}}{\sqrt{n_1 n_2}}$, where *t* is the pooled variances independent samples *t* comparing the two group means.

Constructing a confidence interval for *d* is unreasonably difficult to do by hand, but easy to do with SPSS or SAS. 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*.

I shall illustrate using the Howell data (participants were students in Vermont), comparing boys' GPA with girls' GPA. Here is the computer output.

Group Statistics

	gender	N	Mean	Std. Deviation	Std. Error Mean
gpa	1 male	55	2.2358	.80767	.10891
	2 female	33	2.8236	.83305	.14501

Independent Samples Test

	t-test for Equality of Means			
	t	df	Sig. (2-tailed)	95% Confidence Interval of the Difference

				Lower	Upper
gpa Equal variances assumed	-3.267	86	.002	-.94553	-.23010
Equal variances not assumed	-3.241	65.860	.002	-.94992	-.22572

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

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

*NoncT.sav [DataSet2] - IBM SPSS Statistics Data Editor

File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help

1 : d   .71936972413356   Visible: 8 of 8 Variables

	tval	df	conf	n1	n2	d	lowd	highd	
1	3.2670	86	.950	55	33	.7194	.2727	1.1621	
2									

Data View   Variable View

IBM SPSS Statistics Processor is ready

You should memorize the following benchmarks for $\hat{d}$, but keep in mind that they are not appropriate in all contexts:

Size of effect	d
small	.2
medium	.5
large	.8

Please see my document Cohen's Conventions for Small, Medium, and Large Effects .

Glass' Delta. $\Delta = \dfrac{M_1 - M_2}{s_{control}}$. That is, in computing the standardize difference between group means we use the control group standard deviation in the denominator.

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 :

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

$$r_{pb} = \sqrt{\frac{3.267^2}{3.267^2 + 86}} = .332.$$ 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, $r_{pb} = \dfrac{.588(.487)}{.861} = .33$.

Correlations		gpa
gender	Pearson Correlation	.332
	Sig. (2-tailed)	.002
	N	88

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, $\eta^2 = .332^2 = .11$. To put a confidence interval on eta-squared, use SPSS or SAS. 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 . This interesting but obscure statistic estimates the probability that a randomly selected score from the one population will be greater than a randomly sampled score from the other population.

Correlated Samples Designs. Treat the data as if they were from independent samples – that is, use this formula: $\hat{d} = \dfrac{M_1 - M_2}{s_{pooled}}$. See "D-EqN-- Estimate Cohen's d, equal n data (including correlated samples)" at my SPSS Programs page for the syntax to make SPSS compute estimated d for you.

If you base your effect size estimate on the correlated samples analysis, you will overestimate the size of the effect. You cannot use the SPSS scripts to construct a confidence interval for d when the data are from correlated samples. With correlated samples the distributions here are very complex, not following the noncentral t. You can construct an approximate confidence interval, $\hat{d} \pm Z_{CC} * SE$, where SE is $\sqrt{\dfrac{\hat{d}^2}{2(n-1)} + \dfrac{2(1-r_{12})}{n}}$, but such a confidence interval is not very accurate unless you have rather large sample sizes. I do have SAS scripts to conduct confidence intervals for d from correlated samples, but we shall not use them in this class.

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
- SPSS

You could compute $\hat{d}_{Diff} = \dfrac{M_1 - M_2}{s_{Diff}}$, 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. Cohen defined dz as the mean difference score divided by the standard deviation of the difference scores. The "z" was added to note that the unit of analysis is no longer X (the one mean) or Y (the other mean) but their difference, Z. While I and others (Dunlap, Cortina, Vaslow, & Burke, 1996) think it is most often not a good idea to use dz as an effect size estimator, there may be cases where it is defensible. In such cases the confidence interval for dz could be obtained in exactly the same manner that one would obtain the confidence interval for d for a one sample test.

Why not use the standard deviation of the difference scores as the standardizer? Think about the ρ in the variance sum law: $\sigma_{Y1-Y2} = \sqrt{\sigma_{Y1}^2 + \sigma_{Y2}^2 - 2 * \rho * \sigma_{Y1} * \sigma_{Y2}}$. Now suppose that you found a difference on a test of clerical accuracy between individuals tested after taking a placebo and individuals tested after taking a sedating antihistamine. The difference in means is 3. Within each group the standard deviation is 10. You have independent samples. The d is 3/10 = .3. Now suppose that you had correlated samples with the same difference in means and the same within-condition standard deviations. The correlation between samples will result in the standard deviation of the difference scores being less than the standard deviation within each group (again, remember the ρ in the variance sum law). Suppose it is reduced to a value of 7. If you were to use the standard deviation of the difference scores as the standardizer when estimating d, you would obtain a value of 3/7 = .43 – but the difference between the means is still just 3 points. Why should d = .3 with independent samples but .43 with correlated samples when the difference between means is the same in both cases and the within-conditions standard deviations the same as well?

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.

Assumptions of Parametric Tests

Parametric tests are those for which the hypothesis tested specifies that some parameter has some particular value – for example, μ = 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.

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 <u>not</u> 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.

Notice that I have set all of the statistical symbols in italic font. This is standard practice, and you should do the same.

- Link: Complete Example of Independent Samples t Test

Reference

- Dunlap, W. P., Cortina, J. M., Vaslow, J. B., & Burke, M. J. (1996). Meta-analysis of experiments with matched groups or repeated measures designs. *Psychological Methods*, *1*, 170-177.

Fair Use of This Document