

UNDERSTANDING THE INDEPENDENT-SAMPLES t TEST

The independent-samples t test evaluates the difference between the means of two independent or unrelated groups. That is, we evaluate whether the means for two independent groups are significantly different from each other. The independent-samples t test is commonly referred to as a between-groups design, and can also be used to analyze a control and experimental group. With an independent-samples t test, each case must have scores on two variables, the grouping (independent) variable and the test (dependent) variable. The grouping variable divides cases into two mutually exclusive groups or categories, such as boys or girls for the grouping variable gender, while the test variable describes each case on some quantitative dimension such as test performance. The t test evaluates whether the mean value of the test variable (e.g., test performance) for one group (e.g., boys) differs significantly from the mean value of the test variable for the second group (e.g., girls).

HYPOTHESES FOR THE INDEPENDENT-SAMPLES t TEST

Null Hypothesis: $H_0: \mu_1 = \mu_2$ where μ_1 stands for the mean for the first group and μ_2 stands for the mean for the second group.

-or- $H_0: \mu_1 - \mu_2 = 0$

Alternative (Non-Directional) Hypothesis: $H_a: \mu_1 \neq \mu_2$ -or- $H_a: \mu_1 - \mu_2 \neq 0$

Alternative (Directional) Hypothesis: $H_a: \mu_1 < \mu_2$ -or- $H_a: \mu_1 > \mu_2$
(depending on direction)

NOTE: the subscripts (1 and 2) can be substituted with the group identifiers

For example: $H_0: \mu_{Boys} = \mu_{Girls}$ $H_a: \mu_{Boys} \neq \mu_{Girls}$

ASSUMPTIONS UNDERLYING THE INDEPENDENT-SAMPLES t TEST

1. The data (scores) are independent of each other (that is, scores of one participant are not systematically related to scores of the other participants).

This is commonly referred to as the *assumption of independence*.

2. The test (dependent) variable is normally distributed within each of the two populations (as defined by the grouping variable).

This is commonly referred to as the *assumption of normality*.

3. The variances of the test (dependent) variable in the two populations are equal.

This is commonly referred to as the *assumption of homogeneity of variance*.

Null Hypothesis: $H_0: \sigma_1^2 = \sigma_2^2$ (if retained = assumption met)
(if rejected = assumption not met)

Alternative Hypothesis: $H_a: \sigma_1^2 \neq \sigma_2^2$

TESTING THE ASSUMPTION OF INDEPENDENCE

One of the first steps in using the independent-samples t test is to test the assumption of independence. Independence is a methodological concern; it is dealt with (or should be dealt with) when a study is set up. Although the independence assumption can ruin a study if it is violated, there is no way to use the study's sample data to test the validity of this prerequisite condition. It is assessed through an examination of the design of the study. That is, we confirm that the two groups are independent of each other?

The assumption of independence is commonly known as the unforgiving assumption (r.e., robustness), which simply means that if the two groups are not independent of each other, one cannot use the independent-samples t test.

TESTING THE ASSUMPTION OF NORMALITY

Another of the first steps in using the independent-samples t test is to test the assumption of normality, where the Null Hypothesis is that there is no significant departure from normality, as such; retaining the null hypothesis indicates that the assumption of normality has been met for the given sample.

The Alternative Hypothesis is that there is a significant departure from normality, as such; rejecting the null hypothesis in favor of the alternative indicates that the assumption of normality has not been met for the given sample.

To test the assumption of normality, we can use the Shapiro-Wilks test. From this test, the Sig. (p) value is compared to the *a priori* alpha level (level of significance for the statistic) – and a determination is made as to reject ($p \leq \alpha$) or retain ($p > \alpha$) the null hypothesis.

Tests of Normality

Stress Condition	Kolmogorov-Smirnov ^a			Shapiro-Wilk			
	Statistic	df	Sig.	Statistic	df	Sig.	
Percentage of time talking	Low Stress	.229	15	.033	.917	15	.170
	High Stress	.209	15	.076	.888	15	.062

a. Lilliefors Significance Correction

For the above example, where $\alpha = .001$, given that $p = .170$ for the Low Stress Group and $p = .062$ for the High Stress Group – we would conclude that each of the levels of the Independent Variable (Stress Condition) are normally distributed. Therefore, the assumption of normality has been met for this sample. The *a priori* alpha level is typically based on sample size – where .05 and .01 are commonly used. Tabachnick and Fidell (2007) report that conventional but conservative (.01 and .001) alpha levels are commonly used to evaluate the assumption of normality.

NOTE: Most statisticians will agree that the Shapiro-Wilks Test should not be the sole determination of normality. It is common to use this test in conjunction with other measures such as an examination of skewness, kurtosis, histograms, and normal Q-Q plots.

In examining skewness and kurtosis, we divide the skewness (kurtosis) statistic by its standard error. We want to know if this standard score value significantly departs from normality. Concern arises when the skewness (kurtosis) statistic divided by its standard error is greater than $z \pm 3.29$ ($p < .001$, two-tailed test) (Tabachnick & Fidell, 2007).

We have several options for handling non-normal data, such as deletion and data transformation (based on the type and degree of violation as well as the randomness of the missing data points). Any adjustment to the data should be justified (i.e., referenced) based on solid resources (e.g., prior research or statistical references). As a first step, data should be thoroughly screened to ensure that any issues are not a factor of missing data or data entry errors. Such errors should be resolved prior to any data analyses using acceptable procedures (see for example Howell, 2007 or Tabachnick & Fidell, 2007).

TESTING THE ASSUMPTION OF HOMOGENEITY OF VARIANCE

Another of the first steps in using the independent-samples t test statistical analysis is to test the assumption of homogeneity of variance, where the null hypothesis assumes no difference between the two group's variances ($H_0: \sigma_1^2 = \sigma_2^2$). The Levene's F Test for Equality of Variances is the most commonly used statistic to test the assumption of homogeneity of variance. The Levene's test uses the level of significance set *a priori* for the t test analysis (e.g., $\alpha = .05$) to test the assumption of homogeneity of variance.

		Levene's Test for Equality of Variances	
		F	Sig.
VISUAL Visualization Test	Equal variances assumed	17.570	.000
	Equal variances not assumed		

For Example: For the VISUAL variable (shown above), the F value for Levene's test is 17.570 with a Sig. (p) value of .000 ($p < .001$). Because the Sig. value is less than our alpha of .05 ($p < .05$), we reject the null hypothesis (no difference) for the assumption of homogeneity of variance and conclude that there is a significant difference between the two group's variances. That is, the assumption of homogeneity of variance is **not** met.

If the assumption of homogeneity of variance is not met, we must use the data results associated with the "Equal variances not assumed," which takes into account the Cochran & Cox (1957) adjustment for the standard error of the estimate and the Satterthwaite (1946) adjustment for the degrees of freedom. In other words, we will use the bottom line of the t test for equality of means results table and ignore the top line of information.

Had the Sig. (p) value been greater than our *a priori* alpha level, we would have retained the null hypothesis and concluded that there is *not* a significant difference between the two group's variances. If the assumption of homogeneity of variance is met, we must use the data results associated with the "Equal variances assumed," and interpret the data accordingly. That is, we would use the top line of information for the t test.

Independent Samples Test

		t-test for Equality of Means						
		t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
							Lower	Upper
VISUAL Visualization Test	Equal variances assumed	2.532	188	.008	1.22289	.346194	-1.903070	-.542710
	Equal variances not assumed	-3.442	414.057	.001	-1.22289	.355259	-1.921226	-.524553

For this example (testing the difference between males and females on the Visualization test), since the t value (-3.442, which indicates that the second group was higher than the first group) resulted in a Sig. (p) value that was less than our alpha of .05 ($p < .05$, which puts the obtained t in the tail) – we reject the null hypothesis in support of the alternative hypothesis, and conclude that males and females differed significantly on their Visualization test performance. By examining the group means for this sample of subjects (not shown here), we see that males (with a mean of 6.382) performed significantly higher on the Visualization test than did females (with a mean of 5.159).

VIOLATION OF THE ASSUMPTIONS OF THE t TEST FOR INDEPENDENT GROUPS

The independent-samples t test is what we refer to as a *robust* test. That is, the t test is relatively insensitive (having little effect) to violations of normality and homogeneity of variance, depending on the sample size and the type and magnitude of the violation.

If $n_1 = n_2$ and the size of each sample is equal to or greater than 30, the t test for independent groups may be used without appreciable error despite moderate violations of the normality and/or the homogeneity of variance assumptions (Pagano, 2004, p. 339).

Sample sizes can be considered equal if the larger group is not more than 1½ times larger than the smaller group (Morgan, Leech, Gloeckner, & Barrett, 2004).

If the variance in one group is more than 4 or 5 times larger than the variance in the other group – they are considered very different – and the homogeneity of variance assumption is violated (i.e., not met). A variance ratio (F_{max}) analysis can be obtained by dividing the lowest variance of a group into the highest group variance. Concern arises if the resulting ratio is 4-5 times, which indicates that the largest variance is 4 to 5 times the smallest variance (Tabachnick & Fidell 2007).

If there are extreme violations of these assumptions – with respect to normality and homogeneity of variance – an alternate (non-parametric) test such as the *Mann-Whitney U* test should be used instead of the independent-samples t test.

DEGREES OF FREEDOM

Because we are working with two independent groups, we will lose (restrict) one df to the mean for each group. Therefore, df for an independent-samples t test will be $(n_1 - 1) + (n_2 - 1)$, where n_1 and n_2 are the sample sizes for each of the independent groups, respectively. Or we can use, $N - 2$, where N is the total sample size for the study.

EFFECT SIZE STATISTICS FOR THE INDEPENDENT-SAMPLES *t* TEST

Cohen's *d* (which can range in value from negative infinity to positive infinity) evaluates the degree (measured in standard deviation units) that the mean scores on the two test variables differ. If the calculated *d* equals 0, this indicates that there are no differences in the means. However, as *d* deviates from 0, the effect size becomes larger.

$$d = t \sqrt{\frac{n_1 + n_2}{n_1 n_2}} \quad \text{where} \quad t \text{ is the obtained } t \text{ value and } n_1 \text{ is the total sample size for group 1 and } n_2 \text{ is the total sample size for group 2.}$$

So what does this Cohen's *d* mean? Statistically, it means that the difference between the two sample means is (e.g., .31) standard deviation units (usually reported in absolute value terms) from zero. A mean difference of zero is the hypothesized difference between the two population means. Effect sizes provide a measure of the magnitude of the difference expressed in standard deviation units in the original measurement. It is a measure of the practical importance of a significant finding.

SAMPLE APA RESULTS

Using an alpha level of .05, an independent-samples *t* test was conducted to evaluate whether males and females differed significantly on a visualization test. The test was significant, $t(414.06) = 3.44$, $p < .01$, $d = .31$. The 95% confidence interval for the visualization test mean ranged from -1.92 to -0.52. An examination of the group means indicate that males ($M = 6.38$, $SD = 4.29$) performed significantly higher on the visualization test than did females ($M = 5.16$, $SD = 3.45$).

Note: there are several ways to interpret the results, the key is to indicate that there was a significant difference between the two groups at the *a priori* alpha level – and include, at a minimum, reference to the group means and effect size.

In looking at a sample statistical stand from an independent-samples *t* test, we see

$$t(414.06) = 3.44, p < .01, d = .31$$

<i>t</i>	Indicates that we are using a <i>t</i> -Test
(414.06)	Indicates the degrees of freedom associated with this <i>t</i> -Test
3.44	Indicates the obtained <i>t</i> statistic value (t_{obt})
$p < .01$	Indicates the probability of obtaining the given <i>t</i> value by chance alone
$d = .31$	Indicates the effect size for the significant effect (measured in standard deviation units)

INTERPRETING THE INDEPENDENT-SAMPLES *t* TEST

The first table (**TESTS OF NORMALITY**), which is produced through the **EXPLORE** command in SPSS (not a part of the *t*-test default options), shows the Shapiro-Wilks test of normality. We use this statistic to test whether the levels of the independent variable are statistically normal. If the *p* (Sig.) values are less than or equal to (\leq) our *a priori* alpha level, the level(s) are considered to be non-normal and will require attention (e.g., transformation). If the *p* values are greater than ($>$) our *a priori* alpha level, we consider the level(s) to be statistically normal (i.e., normally distributed). In our example, using an *a priori* alpha level of .001, we find that neither level (low stress or high stress) are significant, and as such, consider both levels of the independent variable to be normally distributed.

THE FOLLOWING TABLES ARE PART OF THE *t* TEST OPTIONS PRODUCED FROM SPSS:

The second table, (**GROUP STATISTICS**) shows descriptive statistics for the two groups (low-stress and high-stress) separately. Note that the means for the two groups look somewhat different. This might be due to chance, so we will want to test this with the *t* test in the next table.

The third table, (**INDEPENDENT SAMPLES TEST**) provides two statistical tests. In the left two columns of numbers, is the **Levene's Test for Equality of Variances** for the assumption that the variances of the two groups are equal (i.e., assumption of homogeneity of variance). **NOTE, this is *not* the *t* test; it only assesses an assumption!** If this *F* test is not significant (as in the case of this example), the assumption is not violated (that is, the assumption is met), and one uses the **Equal variances assumed** line for the *t* test and related statistics. However, if Levene's *F* is statistically significant (Sig., $p \leq .05$), then variances are significantly different and the assumption of equal variances is violated (not met). In that case, the **Equal variances not assumed** line would be used – for which SPSS adjusts the ***t***, ***df***, and ***Sig.*** as appropriate.

Also in the third table... we obtain the needed information to test the equality of the means. Recall that there are three methods in which we can make this determination.

METHOD ONE (most commonly used): comparing the Sig. (probability) value ($p = .022$ for our example) to the *a priori* alpha level ($\alpha = .05$ for our example). If $p \leq \alpha$ – we reject the null hypothesis of no difference. If $p > \alpha$ – we retain the null hypothesis of no difference. For our example, $p \leq \alpha$, therefore we reject the null hypothesis and conclude that the low-stress group ($M = 45.20$) talked significantly more than did the high-stress group ($M = 22.07$).

METHOD TWO: comparing the obtained *t* statistic value ($t_{obt} = 2.430$ for our example) to the *t* critical value (t_{cv}). Knowing that we are using a two-tailed (non-directional) *t* test, with an alpha level of .05 ($\alpha = .05$), with $df = 28$, and looking at the Student's *t* Distribution Table – we find the critical value for this example to be 2.048. If $|t_{obt}| \geq |t_{cv}|$ – we reject the null hypothesis of no difference. If $|t_{obt}| < |t_{cv}|$ – we retain the null hypothesis of no difference. For our example, $t_{obt} = 2.430$ and $t_{cv} = 2.048$, therefore, $t_{obt} \geq t_{cv}$ – so we reject the null hypothesis and conclude that there is a statistically significant difference between the two groups. More specifically, looking at the group means, we conclude that the low-stress group ($M = 45.20$) talked significantly more than did the high-stress group ($M = 22.07$).

METHOD THREE: examining the confidence intervals and determining whether the upper (42.637 for our example) and lower (3.630 for our example) boundaries contain zero (the hypothesized mean difference). If the confidence intervals do not contain zero – we reject the null hypothesis of no difference. If the confidence intervals do contain zero – we retain the null hypothesis of no difference. For our example, the confidence intervals (+3.630, +42.637) do not contain zero, therefore we reject the null hypothesis and conclude that the low-stress group ($M = 45.20$) talked significantly more than did the high-stress group ($M = 22.07$).

Note that if the upper and lower bounds of the confidence intervals have the same sign (+ and + or – and –), we know that the difference is statistically significant because this means that the null finding of zero difference lies *outside* of the confidence interval.

CALCULATING AN EFFECT SIZE

Since we concluded that there was a significant difference between the average amount of time spent talking between the two groups – we will need to calculate an effect size to determine the magnitude of this significant effect. Had we not found a significant difference – no effect size would have to be calculated (as the two groups would have only differed due to random fluctuation or chance).

To calculate the effect size for this example, we will use the following formula:

$$d = t \sqrt{\frac{n_1 + n_2}{n_1 n_2}} \quad \text{where } t \text{ is the obtained } t \text{ value and } n_1 \text{ is the total sample size for group 1 and } n_2 \text{ is the total sample size for group 2.}$$

$$\text{Where,} \quad t = 2.43 \quad n_1 = 15 \quad n_2 = 15$$

Substituting the values into the formula – we find:

$$d = 2.43 \sqrt{\frac{15 + 15}{(15)(15)}} = 2.43 \sqrt{\frac{30}{225}} = 2.43 \sqrt{.133333} = (2.43)(.365148) = .887310 = \mathbf{.89}$$

SAMPLE APA RESULTS

Using an alpha level of .05, an independent-samples t test was conducted to evaluate whether the average percentage of time spent talking differed significantly as a function of whether students were in a low stress or high stress condition. The test was significant, $t(28) = 2.43$, $p < .05$, $d = .89$. The 95% confidence interval for the average percentage of time spent talking ranged from 3.63 to 42.64. An examination of the group means indicate that students in the low stress condition ($M = 45.20$, $SD = 24.97$) talked (on average) significantly more than students in the high stress condition ($M = 22.07$, $SD = 27.14$).

Tests of Normality

Stress Condition	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Percentage of time talking Low Stress	.229	15	.033	.917	15	.170
High Stress	.214	15	.063	.810	15	.005

a. Lilliefors Significance Correction

Independent-Samples T-Test Example

Group Statistics

STRESS	N	Mean	Std. Deviation	Std. Error Mean
TALK 1 Low Stress	15	45.20	24.969	6.447
2 High Stress	15	22.07	27.136	7.006

Independent Samples Test

	Levene's Test for Equality of Variances		t-test for Equality of Means						
	F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
								Lower	Upper
TALK Equal variances assumed	.023	.881	2.430	28	.022	23.13	9.521	3.630	42.637
Equal variances not assumed			2.430	27.808	.022	23.13	9.521	3.624	42.643

REFERENCES

- Cochran, W. G., & Cox, G. M. (1957). *Experimental Designs*. New York: John Wiley & Sons.
- Green, S. B., & Salkind, N. J. (2003). *Using SPSS for Windows and Macintosh: Analyzing and Understanding Data* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.
- Hinkle, D. E., Wiersma, W., & Jurs, S. G. (2003). *Applied Statistics for the Behavioral Sciences* (5th ed.). New York: Houghton Mifflin Company.
- Howell, D. C. (2007). *Statistical Methods for Psychology* (6th ed.). Belmont, CA: Thomson Wadsworth.
- Huck, S. W. (2004). *Reading Statistics and Research* (4th ed.). New York: Pearson Education Inc.
- Morgan, G. A., Leech, N. L., Gloeckner, G. W., & Barrett, K. C. (2004). *SPSS for Introductory Statistics: Use and Interpretation* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.
- Pagano, R. R. (2004). *Understanding Statistics in the Behavioral Sciences* (7th ed.). Belmont, CA: Thomson/Wadsworth.
- Satterthwaite, F. W. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin*, 2, 110-114.
- Tabachnick, B. G., & Fidell, L. S. (2007). *Using Multivariate Statistics* (5th ed.). Needham Heights, MA: Allyn & Bacon.