

## Tests of Differences between Means

### Standard Error of the Mean

In the previous chapter, descriptive statistics were discussed. The standard deviation is used to quantify the variation of a distribution. A related statistic is the standard error of the mean (SEM). The standard error of the mean represents your confidence that the mean of your sample truly reflects the mean of the population you are sampling from. The SEM is calculated as the standard deviation (SD) divided by the square root of the sample size (n).

$$SEM = \frac{SD}{\sqrt{n}}$$

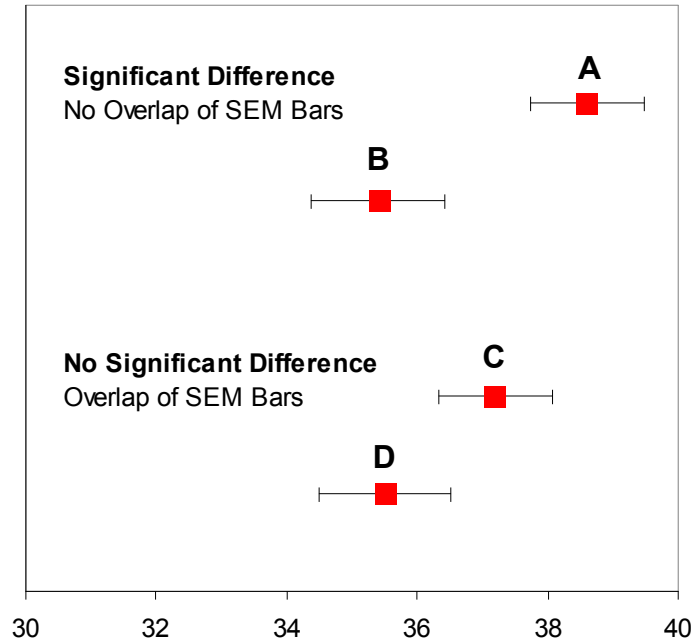
If you take a sample from your population, you could calculate a mean of that sample. If you sampled again you would get a different mean. After repeated sampling the distribution of those means would be close to a normal distribution and the mean of those means would also be your best estimate of the mean of the population. The standard deviation of this distribution of means is estimated by our calculated SEM. Since this distribution tends to be normally distributed you could state that the 95% confidence interval for the population mean is within 1.96 SEMs above and below the mean, or you are 68.26% confident the population mean is within 1 SEM above and below the sample mean. For example, for a sample with a mean of 76, standard deviation of 4 and sample size of 64, the calculated SEM would be 4 divided by the square root of 64 (=0.5). This SEM can then be applied to the sample mean to make inference about the population mean. You have 68.26% confidence that the population mean will be within  $76 \pm 0.5$  (75.5 to 76.5). The 95% confidence estimate would be  $76 \pm (1.96 \times 0.5)$  or 75.02 to 77.98.

The larger your sample is, the more confident you would be that your mean was a good estimate of the population mean. Although the SEM will tend to get smaller as the sample size increases, the standard deviation does not tend to change appreciably with increasing sample size.

### Visual Test of a Difference between Means

The SEM provides for a useful technique for a visual appraisal of likely differences between group means. Although this visual test does not replace the use of the formal test of inference, it is sometimes presented in publications. Graphically this is shown in figure 2-7.1, which depicts

four means A, B, C & D. When comparing A versus B, the mean of A minus its SEM is 37.73 and the mean of B plus its SEM is 36.41.



	Mean	SEM	SD	n	Mean - SEM	Mean + SEM
<b>A</b>	38.6	0.87	7.8	81	37.73	39.47
<b>B</b>	35.4	1.01	8.1	64	34.39	36.41
<b>C</b>	37.2	0.87	7.8	81	36.33	38.07
<b>D</b>	35.5	1.01	8.1	64	34.49	36.51

**Figure 2-7.1: Visual Test of Difference Between Means**

There is therefore no overlap between the means  $\pm$  their associated SEMs, inferring there is a significant difference between the means at about the 95% confidence level. In the C versus D comparison, there is overlap between the means  $\pm$  their associated SEMs, therefore there is no significant difference between the means.

**Student's t test**

You may want to compare the means of two different groups, such as men vs. women, athletes vs. non-athletes, young vs. elderly, or you may want to compare means measured on a single group under two different experimental conditions or at two different times, such as a pretest posttest design. The comparison of multiple group means ( $>2$ ) is done with the Analysis of

Variance (ANOVA) and will be discussed in the next chapter. The simplest test for the comparison of two means is the Student's t test, which can be applied to relatively small samples (<150). This test was developed by a statistician named William Gossett who worked for the Guinness Brewery in Dublin as a quality control manager. When he published his work he used the pseudonym of Student. The assumption of the t-test are:

1. The dependent variable is distributed normally in the population.
2. The dependant variable is continuous (however ordinal scales usually can be analyzed).
3. Samples of the population are randomly selected, as are subjects to treatments.
4. Homogeneity of variances, such that the variances in each group are the same in the population (they can be slightly different for the two groups).

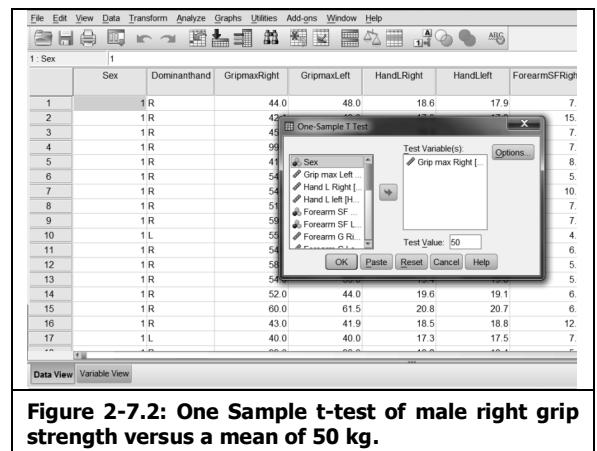
There are two different types of t tests:

**Independent t test:** Used to test for a difference in means between two variables that are not related to each other. An example would be, testing for a difference in mean arm girth between a group of men and a group of women. The sample sizes of the two groups can be the same or different. The *null hypothesis* being tested in the independent t test is that there is **no difference between the means of the two samples**.

**Paired t test:** Sometimes called the t-test for correlated data, this is used to test for a difference between means where the two variables are paired, typically for a within-subjects design. An example would be a pretest post design where the mean difference in arm girth is compared before and after a 6 week training program in the same group of men. Another example would be bilateral symmetry. Is there a mean difference in left side arm girth versus the right side arm girth in a sample of individuals? Because of the paired nature of the variables the sample sizes are always the same for the two. The *null hypothesis* being tested in the independent t test is that the **mean difference between two samples is equal to zero**.

### One Sample t-test

A one sample t-test is used when the investigator wants to test whether the mean of the sample is different to the mean of another group, when only the mean of the other group is known. In SPSS, a one sample t-test can be run by selecting ANALYZE - COMPARE MEANS – ONE SAMPLE T TEST. Figure 2-7.2 shows the dialog box for the test set up for



**Figure 2-7.2: One Sample t-test of male right grip strength versus a mean of 50 kg.**

testing whether the mean of the right grip strength in males is different from a mean of 50 kg. Figure 2-7.3 shows the result of One Sample t-test of male right grip strength versus a mean of 50 kg. The results show that the males (N = 19) had a mean right grip strength of 53.2 kg (sd = 9.0). The calculated t of 1.531 (p = 0.143) indicates that there was no significant difference (because  $p > 0.05$ ) between the mean of the sample and the hypothesized mean of 50 kg.

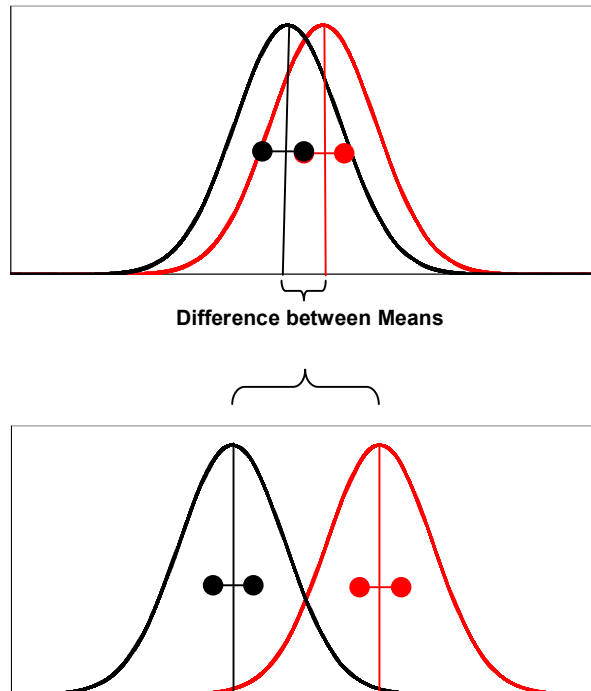
One-Sample Statistics				
	N	Mean	Std. Deviation	Std. Error Mean
Grip max Right	19	53.168	9.0209	2.0695

One-Sample Test						
	Test Value = 50					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
Grip max Right	1.531	18	.143	3.1684	-1.180	7.516

**Figure 2-7.3: Result of One Sample t-test of male right grip strength versus a mean of 50 kg.**

**Independent t-test**



**Figure 2-7.4: The t statistic is a ratio of the difference between the means to the amount of variance in the two groups.**

When using an independent t-test, the question you are asking is, "What is the probability that the difference between the means of the two groups could have occurred by chance?" The null hypothesis therefore is that the difference between the two population means equals zero. If the probability is small enough to satisfy our criterion, then we state that there is a significant difference between the two means.

A general characteristic of inferential tests is the calculation of a test statistic. In this case it is the t statistic, which is compared to a distribution of t in order to determine the probability of there being a difference in the two population means.

Previously the standard error of mean of a variable was defined as the standard deviation divided by the square root of sample size. In the calculation of the t statistic, the difference in the group means is divided by the standard error of the difference between means. When the size of the groups are the same, the standard error of the difference between means is the square root of the sum of the squared standard deviations for each group divided by their respective sample sizes. There is a unique t distribution for each sample size. Figure 2-7.4 illustrates the independent t-test.

The distribution is defined by the degrees of freedom (*df*), where  $df = (n_1 - 1) + (n_2 - 1)$ .

The general formula for the independent t-test is:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}} \quad s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

The formula for equal n has  $df = (2n_1 - 2)$  and is:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

The t statistic is compared to a critical value of t for the predetermined probability level of acceptance. The critical value of t is not only determined by the probability level of acceptance but also the sample size. Table 2-7.1 is a table of critical values of the t-statistics for different probabilities and sample sizes (degrees of freedom) The most common purpose of a t-test is to determine whether two sample means are significantly different at a preselected probability level. The calculated t value is compared with the critical value. If the calculated t equals or exceeds the critical t, the null hypothesis is rejected, and the conclusion is that the difference is significant. The assumption is that the observed difference was due to a real difference in the populations, not to mere sampling variations. If the calculated t is less than the critical t, the null hypothesis is

accepted or retained; the conclusion is that the difference is non-significant. The assumption is that the observed difference was due to chance variations caused by sampling.

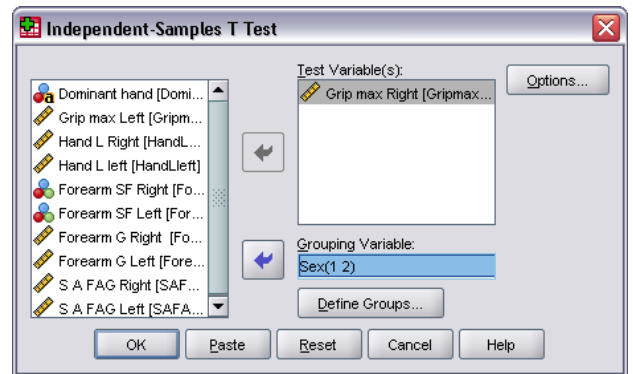
Degrees of Freedom	Probability		
	0.050	0.025	0.010
1	12.706	25.452	6.675
2	4.303	6.205	9.925
3	3.182	4.176	5.841
4	2.776	3.495	4.604
5	2.571	3.163	4.032
6	2.447	2.969	3.707
7	2.365	2.841	3.499
8	2.306	2.752	3.355
9	2.262	2.685	3.250
10	2.228	2.634	3.169
11	2.201	2.593	3.106
12	2.179	2.560	3.055
13	2.160	2.533	3.012
14	2.145	2.510	2.977
15	2.131	2.490	2.947
16	2.120	2.473	2.921
17	2.110	2.458	2.898
18	2.101	2.445	2.878
19	2.093	2.433	2.861
20	2.086	2.423	2.845
21	2.080	2.414	2.831
22	2.074	2.406	2.819
23	2.069	2.398	2.807
24	2.064	2.391	2.797
25	2.060	2.385	2.787
26	2.056	2.379	2.779
27	2.052	2.373	2.771
28	2.048	2.368	2.763
29	2.045	2.364	2.756
30	2.042	2.360	2.750
35	2.030	2.342	2.724
40	2.021	2.329	2.704
45	2.014	2.319	2.690
50	2.008	2.310	2.678
55	2.004	2.304	2.669
60	2.000	2.299	2.660
70	1.994	2.290	2.648
80	1.989	2.284	2.638
90	1.986	2.279	2.631
100	1.982	2.276	2.625
120	1.980	2.270	2.617
∞	1.9600	2.2414	2.5758

**Table 2-7.1: Critical values of the t statistic**

**Calculation of the Independent Samples t-Test using SPSS**

The Independent t test can be found under the ANALYZE menu COMPARE MEANS option in SPSS. Figure 2-7.5 shows the dialog box for the Independent t test in SPSS. In this example, a t test of the difference in means of the maximum grip strength of the right hand between men and women is being tested.

The variable Max Grip Strength Right is sent to the Test Variable(s) box and sex is sent to the Grouping Variable box. In this case, the variable sex has two codes: 1 for men and 2 for women. This is indicated with the 1 and 2 in parentheses. If a variable for grouping has more than two codes; ie. AgeGroup 1, 2 or 3, a t test can only compare the means of two groups, so you would need to indicate in the parentheses which two codes to use.



**Figure 2-7.5: Independent Samples t-test of male versus female right grip strength**

E.g. AgeGroup(1,3) would compare the means of the score for individuals with codes 1 and 3 for AgeGroup.

The dialog box setup shown in Figure 2-7.5 would produce the output shown in Figure 2-7.6.

Group Statistics					
	Sex of Subject	N	Mean	Std. Deviation	Std. Error Mean
Max Grip Strength Right	Male	21	53.248	8.7694	1.9136
	female	23	35.304	6.8536	1.4291

Independent Samples Test										
		Levene's Test for Equality of Variances		t-test for Equality of Means					95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Max Grip Strength Right	Equal variances assumed	1.042	.313	7.598	42	.000	17.943	2.3616	13.1773	22.7092
	Equal variances not assumed			7.513	37.831	.000	17.943	2.3883	13.1076	22.7789

**Figure 2-7.6: Independent Samples t-test of male versus female right grip strength**

The red box outlines the most important part of the output. There is a significant difference between the means of 17.9 kg ( $p < 0.05$ ) based upon a calculated t statistic of 7.598. Always use the test assuming Equal Variances, unless you have shown that the variances of the two groups

are significantly different using a statistical test of differences. In Figure 2-7.6 Levene's test for equality of variances evaluates this. Note in this case  $p = 0.313$ , therefore, there is no significant difference in variances.

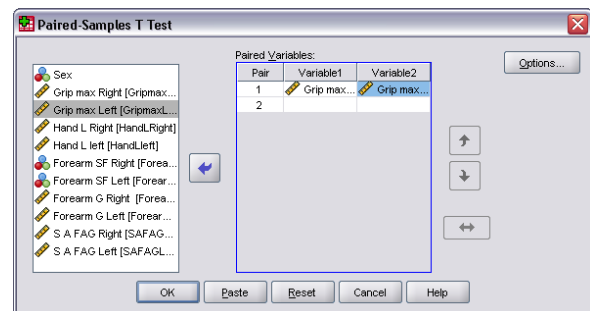
### Paired Samples t- Test

Also called the t-test for correlated data, the hypothesis tested is whether the mean difference between paired observations is significantly different than zero.

The formula for the t statistic calculation with  $df=(n-1)$  is:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2 + s_2^2 - 2rs_1s_2}{n}}}$$

Figure 2-7.7 shows the SPSS dialog box for the Paired t test. In this case, because the test is paired, you pick the two variables you wish to compare. In this case **grip<sub>r</sub>** (right grip strength) and **grip<sub>l</sub>** (left grip strength). This will test if the mean difference between sides is significantly different than zero. Or more simply stated is right grip strength different from left grip strength. In Figure 2-7.7 the SPSS output for this analysis is shown. The red boxes highlight the most important parts of the output. Notice there are two analyses here. Because we wanted to keep men and women in separate analyses there was a SPLIT FILE by SEX in place at the time of the analysis. The men have means of 53.2 kg and 52.7 kg for right and left grip strength, respectively. While the women have means of 35.3 kg and 34.2 kg, respectively. The results show a significant difference for women ( $p < 0.05$ ) but not for men.



**Figure 2-7.7: Dialog Box for Paired Samples t-test of right versus left grip strength.**



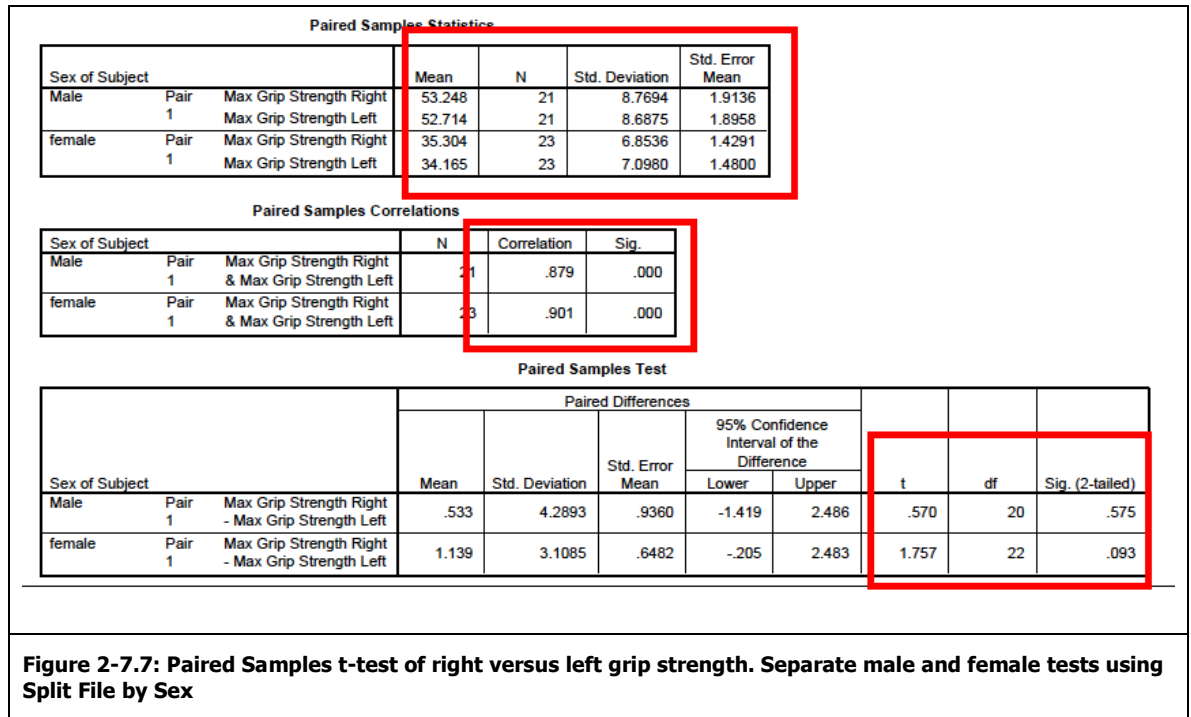


Figure 2-7.8 shows the results of a hypothetical weight loss experiment. Each of the 9 subjects lost weight, therefore intuitively you would expect there to be a significant difference between mean weight before and after. This is a paired experiment, therefore a paired t-test would be appropriate. If you did not recognize this fact and ran an independent test you would get the output shown as the WRONG ANALYSIS. This happens to be the EXCEL output for this test but SPSS would give the same answer. The computer will run the test because it does not know you have run the wrong test. Because there is so much overlap in the distributions of weight before and after, the independent t test shows no significant difference ( $p = 0.918$ ). Whereas when the CORRECT ANALYSIS is run, that being the paired t test, the correct significant difference between means is shown ( $p = 0.0002$ ).

# 2-7.10

## Tests

t

Paired Weight Loss Data		
Weight Before (kg)	Weight After (kg)	Weight Loss (kg)
89.0	87.5	1.5
67.0	65.8	1.2
112.0	111.0	1.0
109.0	108.5	0.5
56.0	55.5	0.5
123.5	122.0	1.5
108.0	106.5	1.5
73.0	72.5	0.5
83.0	81.0	2.0

t-Test: Two-Sample Assuming Equal Variances (MS EXCEL)		
	Before	After
<b>Mean</b>	<b>91.16666667</b>	<b>90.03333333</b>
Variance	537.875	531.11
Observations	9	9
Pooled Variance	534.4925	
Hypothesized Mean Difference	0	
df	16	
<b>t Stat</b>	<b>0.103990367</b>	
P(T<=t) one-tail	0.459234679	
t Critical one-tail	1.745884219	
P(T<=t) two-tail	0.918469359	
t Critical two-tail	2.119904821	

t-Test: Paired Two Sample for Means (MS EXCEL)		
	Before	After
<b>Mean</b>	<b>91.16666667</b>	<b>90.03333333</b>
Variance	537.875	531.11
Observations	9	9
Pearson Correlation	0.999741718	
Hypothesized Mean Difference	0	
df	8	
<b>t Stat</b>	<b>6.23354978</b>	
P(T<=t) one-tail	0.000125066	
t Critical one-tail	1.85954832	
P(T<=t) two-tail	0.000250133	
t Critical two-tail	2.306005626	

**Table 2-7.8: Independent and Paired t test Analyses of Paired Weight Loss Data**

## Power

The level of significance is equivalent to the alpha level (typically 0.05). That is, if we reject  $H_0$  it is very unlikely ( $p < 0.05$ ) that there was no difference between means when we regard them as being significantly different. If the null hypothesis is really true, then once in a while we may make a wrong conclusion and reject  $H_0$  when it is true. This is termed a Type I error. By setting alpha we control the probability of making a Type I error.

Type II error ( $B$ ) is the probability of failing to reject  $H_0$  when it is false. In other words, it is the probability of not obtaining a significant t-statistic when the null hypothesis is incorrect and there really is a difference between population means.

Power ( $1 - B$ ) is the probability of rejecting  $H_0$  when it is false and the alternate hypothesis is true.

The relationship of these errors to power is shown in table 2-7.7.

	$H_0$ true	$H_0$ false
Reject $H_0$	Type I error $P = \alpha$ (Level of significance)	Correct decision $P = 1 - \beta = \text{power}$
Fail to reject $H_0$	Correct decision $P = 1 - \alpha$ (Level of confidence)	Type II error $P = \beta$

**Table 2-7.7: Relationship of power to type of error**

A number of factors can affect the power of

the independent t-test.

These include:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

1.  $s^2$  – a decrease in  $s^2$  (in either or both samples) increases the calculated t – and therefore increases power.
2.  $\alpha$  – for an increase in alpha (e.g. 0.01 to 0.05) although there is an increase in the probability of a Type 1 error, the critical t is smaller, and therefore there is a greater likelihood of obtaining significance and therefore an increase in power.
3. N – a larger sample size results in a larger calculated t (the critical value of t is also smaller) and therefore an increase in power .
4. One tailed hypothesis – has smaller critical t and thus an increase in power. But if the direction of the difference in means is opposite to what was hypothesized there is no power.

5. Difference in means – the greater the difference of means in the population the greater the power
6.  $r$  (for paired t-tests) – the larger the  $r$  the greater the power.

For any given statistical test (t-test, ANOVA, correlation etc) you can find equations or tables for the calculation of power or sample size as available in such reference books as Zar's Biostatistical Analysis (New Jersey: Prentice-Hall Inc., 1984). However, in addition to these there are a number of statistical packages and interactive software available to compute sample size needed or power of the test. A very useful website can be found at <http://statpages.org/>, where links to multiple programs are provided, in order that you can carry out the power, or sample size calculation for your specific research design.