

Power analysis for simple experiments With two conditions and equal numbers of observations

Marc Brysbaert
Royal Holloway College
University of London

1. INTRODUCTION

- inferential statistics = techniques that allow us to study samples and then make generalisations about the populations from which they were drawn
- Most frequent case: we have two conditions and we want to know whether the average performance in one condition is reliably different from the average performance in the other condition. To find out, we use a t-test or an analysis of variance to get a p-value, which informs us whether or not the observed difference between the means is likely to be due to sampling error.
- However, by limiting our analysis to this kind of t-test or ANOVA, we are only looking at half of the reality: we assume that if the sample statistic is significant, there is a difference between the population means (we can reject the null-hypothesis) and if the sample statistic is not significant, there is no difference between the population means (we fail to reject the null-hypothesis).
- Actually, the situation is slightly more complicated. Because of the sampling error, there is always a possibility that the sample statistic provides us with the wrong information about the population distributions. In case we find a significant difference between two sample means, there is a certain probability that this is simply due to chance fluctuations and that actually (i.e., at the population level) there is no difference. We all know that this probability is a function of the alpha-level we choose. If $\alpha = .05$, we have 5% chance that we will reach the wrong conclusion of a significant difference at the population level; if $\alpha = .01$, we have 1% chance that we will reach the wrong conclusion. This is called the Type I error.
- There is another error we can make. Suppose, we fail to find a significant difference between the two sample means. How likely is it in this case that there is indeed no difference between the population distributions? Or how large is the possibility that we will have to conclude that we fail to reject the null-hypothesis whereas in reality (i.e. at the population level) there is a difference? This is the question about the power of an experiment. A wrong conclusion about a null-effect is called a Type II error: we think there is no difference between the populations (because we fail to obtain a p-value that is small enough), whereas in reality there is one.

- In summary, inferential statistics involves educated guesses about population distributions. We would like our guesses to be correct as often as possible, although it is never possible to completely exclude erroneous conclusions due to sampling error. There are two types of errors, as indicated below

		difference at the population level	
		yes	no
significant difference between sample means	yes	correct decision	Type I error
	no	Type II error	correct decision

- The chances of a Type I error are known by looking at the alpha-level. But what about the chances of a Type II error? How likely are we to find no significant difference between our sample statistics, whereas in reality there is one? How can we minimise this possibility? These are questions about the **power** of an experiment, the probability of obtaining a significant effect in the sample data when there is an effect at the population level.
- In what follows, I will considerably simplify the situation by limiting the question of power analysis to (a) the situation in which there are only two conditions, and (b) the numbers of observations in the conditions are the same. This is the most common (and interesting) situation and it should encourage researchers to make sure that they have an equal number of observations in each condition (which is always the best option, because it makes the analyses robust against violations of the underlying assumptions, such as the form of the distribution – normal – and the equality of the variances in both conditions).

2. POWER ANALYSIS FOR REPEATED MEASURES

- In a design with two repeated measures, we get two observations from the same participants, for instance, naming times for very familiar words and naming times for less familiar words. Normally, we analyse them with a t test for related samples (also called paired t test, or t test for dependent means) or an analysis of variance with repeated measures (also called within-subjects design). Below, I give the data of an experiment in which 6 persons named 10 words that were preceded by a semantically related word (e.g., the word “butter” preceded by the prime “bread”) and 10 words preceded by a semantically unrelated word (e.g., the word “nurse” preceded by the prime “cat”). Each data point is the mean naming latency for the 10 words (measured in milliseconds).

	Unrelated	Related	Diff
person 1	540	520	20
person 2	579	565	14
person 3	523	501	22
person 4	550	555	-5
person 5	512	499	13
person 6	543	547	-4
	-----	-----	-----
\bar{X} =	541.2	531.2	10.0
sd_x =	23.2	28.4	11.7

- If we calculate the t test, we obtain the following result: $t(5) = 2.09$, $p = .09$. If we run an ANOVA, we get $F(1,5) = 4.35$, $MSe = 69.0$, $p = .09$. This is a situation in which we find some difference between the conditions, but we fail to reject the null-hypothesis at the .05 level. What can we conclude: That in reality there is no difference because we failed to find reliable evidence for it? Or that in reality there is a difference, but that our experiment was not strong (powerful) enough to reveal such a difference. In order to answer this question, we need some information of the **a priori** likelihood that we would be able to detect the difference with our experiment (which by the way consisted of only 6 subjects, as most of you probably have noticed already). This is the question of the power of an experiment, the power to detect a genuine difference at the population level.
- There are two important factors determining the power of an study (besides 2 other, in practice less important variables: the α -level of the test, and whether the test is one-tailed or two-tailed). These two variables are the effect size and the sample size.

2.1. Effect size based on the standardised mean difference

- The **effect size** refers to the magnitude of the effect you can expect in the population. Suppose, for instance, that we know that the effect of semantic priming usually is in the order of 100 ms (you can name the word “nurse” 100 ms faster after having seen the prime “doctor” than after having seen the prime “cat”). In such a situation, six participants might indeed have been enough to detect a reliable difference between the conditions. On the other hand, suppose we know that the effect of semantic priming in reality is only 10 ms, then chances are much lower that we will be able to detect such a small difference with a sample of only 6 persons (because a few participants may have negative values due to the variability of the underlying processes and sampling error). Statistics not only provide us with these general intuitions, but also with quite precise information about the likelihood of obtaining a significant effect as a function of the effect size and the sample size, as I will show below.
- Although the effect size in raw values (e.g., RTs in milliseconds) is an interesting statistic (and will be used in many literature reviews), statisticians prefer to express the effect size in a measure that does not depend on the specific dependent variable used. There are two ways of doing this. The first is to divide the raw effect size by its standard deviation, which results in a statistic called “Cohen’s d”. The second way is to express the effect size as the proportion of variance accounted for by the independent variable, which results in a statistic called η^2 (eta squared). We will discuss the first measure in this section, and leave the η^2 measure till the next section.
- Cohen’s d measure is a measure based on the standardised mean difference, that is on the mean difference between the conditions divided by the standard deviation of the difference scores:

$$d = \frac{\bar{X}_1 - \bar{X}_2}{sd} = \frac{\bar{D}}{sdb}$$

In our example:

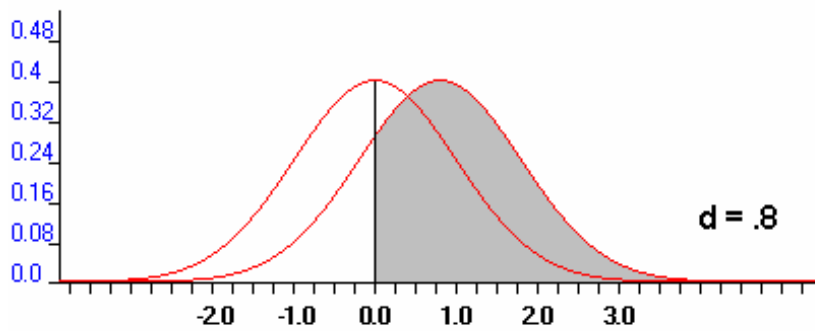
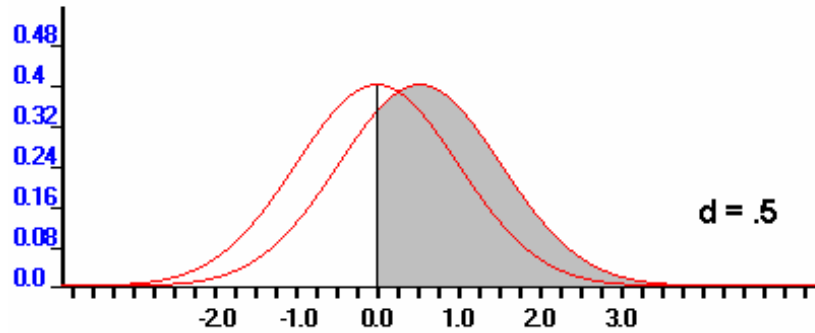
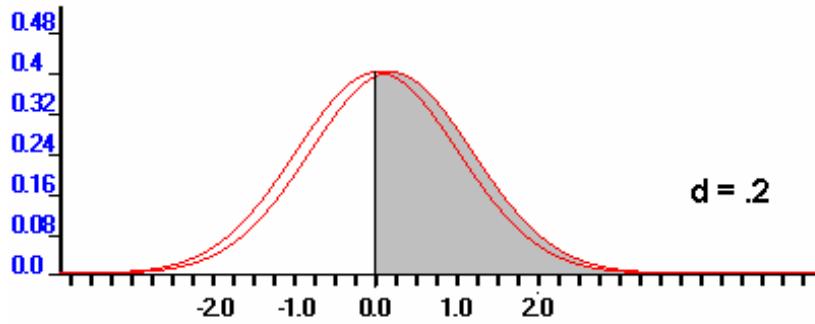
$$d = \frac{10.0}{11.7} = .85$$

- Of course, this is only a post-hoc calculation of the effect size of the experiment. We can use this as an *a priori* estimate of the effect size that we can expect for our next experiment. However, in general we would like to have an idea of the effect size before we start the experiment. There are three main ways in which we can obtain this information.

1. Postulate a rough estimate. In his seminal work on the power of statistical tests, Cohen (1969, 1988) postulated the following rules of thumb:

- a. If you expect a **small** effect, assume $d = .2$
- b. If you expect a **medium** effect, assume $d = .5$
- c. If you expect a **large** effect, assume $d = .8$

- To get an idea of what these values mean, you have to return to the standard normal distribution (on which the d-statistic is based). A normal distribution centred around a z-value of 0.0 means that half of the observations (i.e. the D-values in our example) will be above zero and half will be below zero. A normal distribution centred around a z-value of .2 means that 57.9% of the observations will be above 0 and 42.1% below zero. A normal distribution centred around a z-value of .5 means that 69.1% of the values will be above 0 and 30.9% below 0. Finally, a normal distribution centred around .8, means that 78.8% of the values will be positive and 21.2% will be negative. On the basis of this classification the effect size of .85 we obtained in our experiment is rather big.
- You may have the feeling that $d = .8$ is not that large after all. However, subsequent research showed that Cohen's intuitions were basically right. It is very difficult in psychology to get effect sizes above 1.0. To some extent this can be understood, because large effect sizes basically mean that you do not need statistics to "see" the difference. For instance, the difference in length between newborns and adults yields a d-value way above 2. However, no-one feels a need for statistics to test whether this effect is significant (do we?).



2. The second way to get an idea of the effect size you're testing, is to look for meta-analyses in the literature. In these review papers, the effect sizes of a lot of previous studies have been gathered and an estimate of the overall effect size is given. For instance, you may find that Lucas (2000) reviewed tens of studies on semantic priming and found an overall effect size of $d = .50$ (so, a medium size effect, according to Cohen). Other values that have been found are:

- the effect of computer-based instruction over traditional instruction in class-rooms:
 $d = .30$
- the effect of having supportive parents on resilience in adolescence
 $d = .34$
- the effectiveness of memory rehabilitation after traumatic brain injury
 $d = .47$
- efficacy of group therapy for depression
 $d = 1.03$

- effect of AIDS risk-reduction interventions for drug abusers
d = .31

- difference between females and males in amount of fear during thriller watching
d = .41

3. The third way to get an idea of the effect size of the phenomenon you're investigating is to determine the effect size yourself on the basis of studies that have been published before. In general, this can be a difficult enterprise if you have complicated designs and no (longer) access to the original data, but for the designs we're discussing, there exist a few quick and easy procedures that will give you a fair estimate. These are:

- if you have access to the mean difference score and the standard deviation of the difference scores, you just calculate d as we have done above

$$d = \frac{\bar{D}}{s_{db}} = \frac{10.0}{11.7} = .85$$

- if you have access to the F-value and the mean square of errors, this is the formula

$$d = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{2MSe}} = \frac{10}{\sqrt{2 \times 69}} = .85$$

- there are even quicker ways, which only require you to have access to the t- and F-statistics. They are the following

$$d = \frac{t}{\sqrt{n}} = \frac{2.09}{\sqrt{6}} = .85$$

$$d = \sqrt{\frac{F}{n}} = \sqrt{\frac{4.37}{6}} = .85$$

2.2. Effect size based on the proportion of variance explained

- Cohen's d statistic is straightforward to understand as long as only two conditions are involved. Then it simply refers to the positions of the two normal distributions relative to one another (as shown in the figure above). However, this statistic becomes more difficult to understand (and calculate) in more complex situations, when an independent variable has more than two levels or when the design involves interactions between two or more independent variables. For these reasons, statistical software packages often use another measure of the effect size: the proportion of variance explained by the effect.
- Remember that in the data from a study there are two possible sources of variation: systematic variation and random (or unsystematic) variation. The former is the variation due to the manipulation introduced by the experimenter; the latter is due to natural differences that exist between individuals.
- Suppose the data of our semantic priming example had looked as follows:

		Unrelated	Related	Diff
person 1		540	530	10
person 2		579	569	10
person 3		523	513	10
person 4		550	540	10
person 5		512	502	10
person 6		543	533	10
		-----	-----	-----
\bar{X}	=	541.2	531.2	10.0
sd_x	=	23.2	28.4	0.0

- In this case, it is clear that all variability in the difference scores is due to the semantic priming and that none of the variability is due to individual differences. We therefore can conclude that 100% of the variance in the difference scores is explained by the manipulation we introduced.
- Suppose now the data had looked as follows:

		Unrelated	Related	Diff
person 1		740	520	220
person 2		379	565	-186
person 3		723	501	222
person 4		712	499	213
person 5		350	555	-205
person 6		343	547	-204
		-----	-----	-----
\bar{X}	=	541.2	531.2	10.0
sd_x	=	23.2	28.4	228.3

- In this example, we see that the mean 10 ms difference we have between the unrelated and the related condition is swamped by the unsystematic variation between the individuals. Only a very small percentage (i.e., 0.2%) of the variation observed in the difference scores is due to the manipulation.
- The η^2 (eta squared) statistic is a statistic that indicates how much of the variation is due to the effect being tested. When $\eta^2 = 1.00$, then 100% of the variation is due to the effect; when $\eta^2 = 0$, then no variation is due to the effect. Because of this characteristic, the meaning of η^2 closely resembles that of a correlation (actually, that of a squared correlation: R^2). The more η^2 approaches 1.00 the better the values predicted on the basis of the effect agree with the observed values.
- There are several ways to calculate η^2 . The first is when you have access to the ANOVA table (by the way, SPSS gives you the option to have a column with the η^2 values included in your ANOVA table). Then the formula to use is:

$$\eta^2 = \frac{SS_{effect}}{SS_{total}} = \frac{SS_{effect}}{SS_{effect} + SS_{error}}$$

- For the example of the semantic priming study introduced at the beginning of the section on the power analysis for repeated measures, this is the SPSS output table:

Tests of Within-Subjects Effects

Measure: MEASURE_1

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Eta Squared	Noncent. Parameter	Observed Power ^a	
FACTOR1	Sphericity Assumed	300.000	1	300.000	4.348	.091	.465	4.348	.394
	Greenhouse-Geisser	300.000	1.000	300.000	4.348	.091	.465	4.348	.394
	Huynh-Feldt	300.000	1.000	300.000	4.348	.091	.465	4.348	.394
	Lower-bound	300.000	1.000	300.000	4.348	.091	.465	4.348	.394
Error(FACTOR1)	Sphericity Assumed	345.000	5	69.000					
	Greenhouse-Geisser	345.000	5.000	69.000					
	Huynh-Feldt	345.000	5.000	69.000					
	Lower-bound	345.000	5.000	69.000					

a. Computed using alpha = .05

- From this table, we can read that $SS_{effect} = 300$ and that $SS_{error} = 345$. So,

$$\eta^2 = \frac{300}{300 + 345} = .465$$

- A second way to calculate η^2 is to start from the t-values or the F-values reported in the ms. This is particularly interesting when we have no access to the ANOVA table. This is the formula you have to use when you only have a t-value:

$$\eta^2 = \frac{t^2}{t^2 + df}$$

- For the example with $t(5) = 2.09$, this gives

$$\eta^2 = \frac{(2.09)^2}{(2.09)^2 + 5} = \frac{4.3681}{4.3681 + 5} = .466$$

- When we have the F-value, this is the formula to use :

$$\eta^2 = \frac{F}{F + df_{error}}$$

- In this formula, you can easily see that for a repeated measures design with one independent variable and two levels, the F-value of the ANOVA is the square of the t-value of the t-test. For our example, with $F(1,5) = 4.35$, we get

$$\eta^2 = \frac{4.35}{4.35 + 5} = .465$$

- Finally, the third way to calculate η^2 is to convert it from a d-value (or vice versa). This can be done with the following equations

$$\eta^2 \approx \frac{d^2}{d^2 + 1} \quad d \approx \sqrt{\frac{\eta^2}{1 - \eta^2}}$$

- For the example:

$$\eta^2 \approx \frac{(.85)^2}{(.85)^2 + 1} = \frac{.7225}{1.7225} = .419 \quad d \approx \sqrt{\frac{.465}{1 - .465}} = .93$$

- The reason why the formula only works approximately, is because it assumes that $d = t / \sqrt{df}$ rather than $d = t / \sqrt{n}$ ($2.09 / \sqrt{5} = .93$). For this reason, the d-statistic will always yield a slightly smaller effect size than the η^2 statistic. The smaller the sample size, the bigger the difference between both statistics.

- Rules of thumb for small, median, and large effect sizes, are

- * small effect size : $\eta^2 = .01$ (this agrees with a correlation of .10)
- * medium effect size : $\eta^2 = .09$ (this agrees with a correlation of .30)
- * large effect size : $\eta^2 = .25$ (this agrees with a correlation of .50)

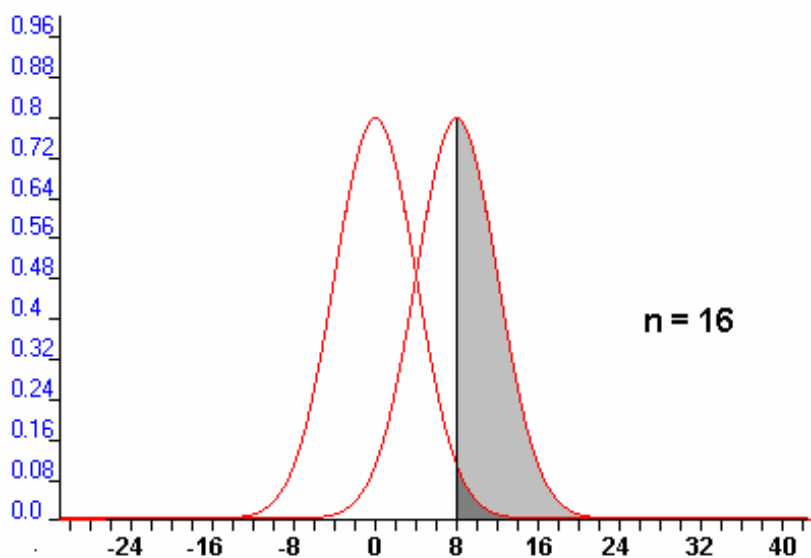
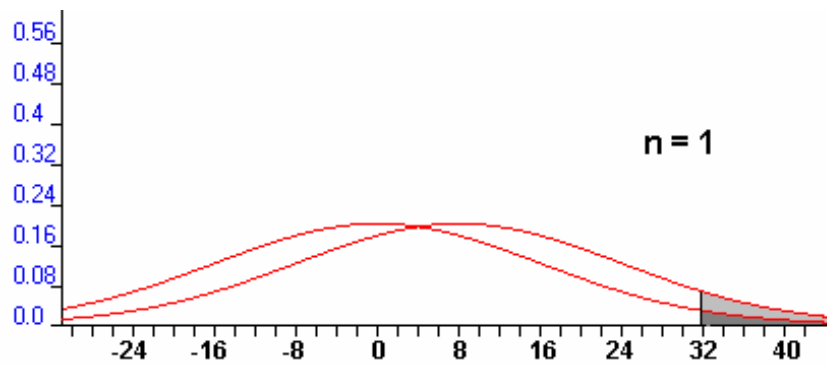
2.3. Sample size and power

- The importance of the **sample size** for the power of an experiment can be understood as follows. As you may have noticed, the effect size is based on the standard deviation of the differences, and not on the standard error of the mean deviation. Remember that the standard error of a mean is

$$SE_{\text{mean}} = \frac{sd}{\sqrt{n}}$$

meaning that there will be less variability in the means of samples when the samples are large than when they are small [remark that the formula $d = t / \sqrt{n}$ does nothing else than “uncorrect” the t-statistic for the sample size].

- On the next page I draw the hypothetical distributions of \bar{D} s we would obtain if we repeated the above semantic priming effect many times under the assumption that the population value of $D = 8$ ms and that the standard deviation of the differences is 16 ms (i.e., an effect size of $d = 8/16 = .5$). I also include the distribution under the null-hypothesis that the population value of $D = 0$ ms (no difference between both conditions) with the same standard deviation of 16 ms. The upper panel shows the distributions for sample sizes $n = 1$; the lower panel for $n = 16$. As you can see, the distributions become thinner as the sample size increases (because $SE = sd / \sqrt{n}$). Now, again theoretically speaking, we will only reject the null-hypothesis of $\bar{D} = 0$ when we obtain a \bar{D} -value larger than $0 + 1.96 * SE$ (or smaller than $0 - 1.96 * SE$). For $n = 1$, this will be for values larger than $0 + 1.96 * 16 / \sqrt{1} = 31.4$ ms; for $n = 16$, this will be values larger than $0 + 1.96 * 16 / \sqrt{16} = 7.8$ ms. The theoretical chances that we will find a \bar{D} larger than 31.4 ms for $n = 1$ are pretty low (i.e., 7%), whereas the chances that we will find a \bar{D} larger than 7.8 for $n = 16$ are much higher (50%). In reality, the chances will be even slightly lower because we have to work with empirical distributions rather than theoretical distributions (i.e., the critical value will be larger than 1.96, because the calculations are based on empirical t-distributions rather than the theoretical normal distribution).



- The power is nothing else than the probabilities we have just calculated. How likely are we a priori to find a significant effect, given the effect size and the sample size? The power of our theoretical experiment with $d = .5$ and $n = 1$ would be .07; the power of our theoretical experiment with $d = .5$ and $n = 16$ is .50. Similarly, we can calculate the power of our theoretical experiment about semantic priming, assuming that $d = .5$ (Lucas, 2000) and with $n = 6$. This value would be .23. That is, from the onset we could have known that the chances of obtaining a significant effect with a sample size of 6 were less than 25%! So, it would be quite unwise of us to conclude that we found evidence supporting the null-hypothesis of no semantic priming in our experiment. The tables on the next page give you the approximate power for some repeated-measures tests you may want to perform. The missing values can easily be extrapolated. They will help you to estimate how many participants you should include in your experiment (Cohen recommends to aim for a power of .80) and will help you to correctly interpret insignificant findings.

Table 1
Approximate Power for Studies Using the *t* Test for Dependent Means in Testing Hypotheses at the .05 Significance Level, Two-tailed tests

Sample Size	Effect size		
	Small (<i>d</i> = .2)	Medium (<i>d</i> = .5)	Large (<i>d</i> = .8)
10	.09	.32	.66
20	.14	.59	.93
30	.19	.77	.99
40	.24	.88	≈1.00
50	.29	.94	≈1.00
100	.55	≈1.00	≈1.00

Table 2
Approximate Number of Research Participants Needed to Achieve 80% Power for the *t* Test for Dependent Means in Testing Hypotheses, Two-tailed tests

Alpha Level	Effect size		
	Small (<i>d</i> = .2)	Medium (<i>d</i> = .5)	Large (<i>d</i> = .8)
.05	198	33	14
.01	295	50	22

Note: You can also calculate these values by using a simple internet applet (e.g., <http://www.cs.uiowa.edu/~rlenth/Power>). Choose one-sample t-test; set Sigma = 1, True difference score = .2, .5, or .8, and play with the values of *d*, *n*, α , and one- vs. two-tailed! Notice that these values are empirical values and, therefore, will be slightly lower than the theoretical ones calculated above (e.g., *d* = .5, *n* = 6, α = .05, two-tailed, power = .17). These values have been used to make Tables 1 and 2.

3. POWER ANALYSIS FOR INDEPENDENT MEASURES

- The logic behind the power analysis for independent measures is very much the same as the one for dependent measures, just that here we are working with two distributions (one for each condition) rather than with one distribution (the one of the difference scores). I will not discuss the details of the analysis (see the books listed below), but will simply give you the formula for the effect size and the table of the sample sizes. These equations only work for sample sizes that are the same (i.e., the number of observations in condition 1 is the same as the number of observations in condition 2). I will apply these equations to the same data set as before, but this time we will assume that different participants saw the words preceded by the semantically related prime and by the unrelated prime. So, in total 12 persons were tested. These are the data

	Unrelated	Related
p1	540	p7 520
p2	579	p8 565
p3	523	p9 501
p4	550	p10 555
p5	512	p11 499
p6	543	p12 547
	-----	-----
\bar{X}	= 541.2	531.2
sd _x	= 23.2	28.4

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\text{var}_1/n_1 + \text{var}_2/n_2}} = \frac{541.2 - 531.2}{\sqrt{538.9/6 + 806.6/6}} = .67$$

So $t(10) = .67$, $p = .52$, or if we did an analysis of variance $F(1,10) = .45$, $MSe = 672.8$, $p = .52$). Notice that because we are working with independent samples, the variability between the individuals is included in the calculation of the statistics, so that the t- and the F-values are much lower. As we will see in the next sections, because of this increased variation, the effect size and the power of the experiment will be lower as well.

3.1. Effect size based on the standardised mean difference

- To calculate Cohen's d statistic for a design with independent measures, we use the following formula on the basis of the t-test:

$$d = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\text{var}_1 + \text{var}_2}{2}}} = \frac{541.2 - 531.2}{\sqrt{\frac{538.9 + 806.6}{2}}} = .39$$

or still easier

$$d = \frac{2t}{\sqrt{n_1 + n_2}} = \frac{1.34}{\sqrt{12}} = .39$$

or (approximately) :

$$d = \frac{2t}{\sqrt{df}} = \frac{1.34}{\sqrt{10}} = .42$$

- The following are the formula on the basis of an ANOVA (see the table below):

$$d = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{MSe}} = \frac{10}{\sqrt{672.8}} = .39$$

or (approximately)

$$d = \frac{2\sqrt{F}}{\sqrt{df(\text{error})}} = \frac{2\sqrt{.446}}{\sqrt{10}} = .42$$

3.2. Effect size based on the proportion of variance explained

- To calculate η^2 for a design with independent measures, we will use the following equation on the basis of a t-test (notice that for this statistic, unlike the d statistic, the formula is the same for independent measures as for repeated measures):

$$\eta^2 = \frac{t^2}{t^2 + df} = \frac{(0.67)^2}{(0.67)^2 + 10} = \frac{.449}{10.449} = .043$$

- When we have access to the SPSS output table, we can calculate the η^2 measure as follows.

Tests of Between-Subjects Effects

Dependent Variable: VAR00002

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Eta Squared	Noncent. Parameter	Observed Power ^a
Corrected Model	300.000 ^b	1	300.000	.446	.519	.043	.446	.093
Intercept	3449696	1	3449696	5127.627	.000	.998	5127.627	1.000
VAR00001	300.000	1	300.000	.446	.519	.043	.446	.093
Error	6727.667	10	672.767					
Total	3456724	12						
Corrected Total	7027.667	11						

a. Computed using alpha = .05

b. R Squared = .043 (Adjusted R Squared = -.053)

$$\eta^2 = \frac{SS_{effect}}{SS_{total}} = \frac{SS_{effect}}{SS_{effect} + SS_{error}} = \frac{300}{300 + 6727.667} = .043$$

- Alternatively, we can also calculate the η^2 value on the basis of the F-statistic:

$$\eta^2 = \frac{F}{F + df_{error}} = \frac{.446}{.446 + 10} = .043$$

- Formulas to easily convert from d to η^2 and vice versa are as follows (assuming that $d = 2t / \sqrt{df}$, rather than $d = 2t / \sqrt{n}$; that is $d = 1.34 / \sqrt{10} = .42$):

$$\eta^2 = \frac{d^2}{d^2 + 4} = \frac{(.42)^2}{(.42)^2 + 4} = .043$$

$$d = \frac{2\sqrt{\eta^2}}{\sqrt{1 - \eta^2}} = \frac{2\sqrt{.043}}{\sqrt{1 - .043}} = .42$$

3.3. Sample size

- As for related samples, the power of a study with independent samples depends on both the effect size and the sample size. For an equal effect size, the distributions of the means will be thinner as the sample sizes increase. Remember that all our analyses have been done under the assumption that the sample sizes in the two conditions are equal!

3.4. Power

Below, you find the tables needed to determine the power of an experiment (based on a priori estimates of the effect size). As you can see, many more participants are needed in a design with independent measures than in a design with dependent measures to achieve the same power.

Table 3
Approximate Power for Studies Using the *t* Test for Independent Means in Testing Hypotheses at the .05 Significance Level, Two-tailed tests

Sample Size of each group	Effect size		
	Small ($d = .2$)	Medium ($d = .5$)	Large ($d = .8$)
10	.07	.18	.39
20	.09	.33	.69
30	.12	.47	.86
40	.14	.60	.94
50	.17	.70	.98
100	.29	.94	≈1.00

Table 4
Approximate Number of Research Participants Needed in Each Group to Achieve 80% Power for the *t* Test for Independent Means in Testing Hypotheses, Two-tailed tests

Alpha Level	Effect size		
	Small ($d = .2$)	Medium ($d = .5$)	Large ($d = .8$)
.05	393	64	26
.01	586	95	38

Note: You can also calculate these values by using a simple internet applet (e.g., <http://www.cs.uiowa.edu/~rlenth/Power>). Choose two-sample t-test; set Sigma1 and Sigma2 = 1, True difference score = .2, .5, or .8, and play with the values of d , n_1 , n_2 , α , and one- vs. two-tailed!

4. POWER ANALYSIS FOR PEARSON CORRELATION COEFFICIENTS

- In the preceding sections, we have seen how many participants need to take part in an experiment, for the experimenter to have a reasonable (80%) chance of finding the effect they are looking for, provided the effect exists at the population level.
- In this final section, we will do the same for correlation coefficients. How many participants do you need in order to find a significant ($\alpha = .05$, two-tailed) effect in your sample given that there is a correlation at the population level?
- As for the previous power analyses, the power of a correlational study depends on the effect size and the sample size.
- The **effect size** is very easy to understand in correlational research, because the correlation coefficient itself is a measure of the percentage of variance accounted for by the covariation between the variables. When we introduced the η^2 statistic, we said that it had the same meaning as R^2 , which is the square of the correlation coefficient.
- So, effect sizes are easy to define in correlational research:

small effect size : $r = .10$
medium effect size : $r = .30$
large effect size $r = .50$

- With respect to the **sample size**, the most important question that comes to mind is: How many people do I need to test, in order to have a reasonable chance of finding a significant correlation? For example, suppose that a Health agency wants to know whether living close to a carriageway increases the risk of health problems. For a sample of people the distance from their home to the nearest two-lane road will be measured, and the number of times they have visited their GP in the last month will be registered. Because nobody has raised the issue yet, the risk is assumed to be a small effect ($r = .10$ at the population level). Because the agency does not want to raise alarm unnecessarily, the α -level is set at .01 two-tailed. How many data pairs need to be collected, in order to have 95% chance of finding a significant correlation in the sample tested?
- Cohen (1969) was one of the first to address questions like these and his staggering answer was that for this particular example, it would require the agency to collect data from 1790 independent individuals. Table 5 below shows the numbers of participants needed to reach 80% power of finding a significant correlation coefficient in a sample given the different effect sizes at the population level.

Table 5
Approximate Number of Research Participants Needed to
Achieve 80% Power for a Pearson correlation coefficient,
Two-tailed tests

Alpha Level	Effect size		
	Small ($r = .10$)	Medium ($r = .30$)	Large ($r = .50$)
.05	783	84	28
.01	1163	124	41

Interesting references

- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.
- Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155-159.
- Aron, A. & Aron, E.N. (1999). *Statistics for Psychology* (2nd Edition), chapter 8.
- Macdonald, R. R. (1997). On statistical testing in psychology. *British Journal of Psychology*, *88*, 333-347.
- Campbell, J.I.D., & Thompson, V.A. (2002). More power to you: Simple power calculations for treatment effects with one degree of freedom. *Behavior Research Methods, Instruments, & Computers*, *34*, 332-337.