Tutorials in Quantitative Methods for Psychology
2007, Vol 3(2), p. 60-62.

DOI: 10.20982/tqmp.03.2.p060

Computing the power of a t test

Denis Cousineau
Université de Montréal

We show how to compute the power of a 2-group *t* test using SPSS or *Mathematica*. To do so, it is necessary to estimate the hypothetical effect size, if an effect is to be found.

Power is defined as the probability of correctly detecting an effect. It is often noted $1-\beta$, where the converse, β is the probability of a type-II error (not rejecting H_0 when there is an effect). When planning a new experiment, it is generally recommended to have a power of at least .80. Suppose that your design has very little power and suppose further that you found a significant effect. Since you were unlikely to detect it (low power), there are many chances that this effect is truly a type-I error (whose probability, often 5%, is noted α). Given the fact that you found an effect (and that the presence of an effect is as likely as its absence), the posterior probability that your finding is a type-I error is given by $\frac{\alpha}{\alpha+1-\beta}$. Hence, with a power of .10 (very low power), it represents a 33% chance of a type-I error rather than a true effect.

The power of a 2-group *t* test depends, as with any statistical test, on three factors:

the effect size, the level of significance and the sample size (Cohen, 1992). The larger the expected effect size is, the more powerful the test is likely to be. Likewise, setting the criterion level α higher (e.g. .10 instead of .05) will increase power. Figure 1 shows the distribution of the possible results of a *t* test if there is truly no effect (blue) and if there is a medium effect (red). Increasing the criterion level causes the green line to be moved to the left (smaller critical value), increasing the probability of a type-I error but increasing power. Finally, increasing the sample size increases the power by increasing the *t* statistic, as we will see next.

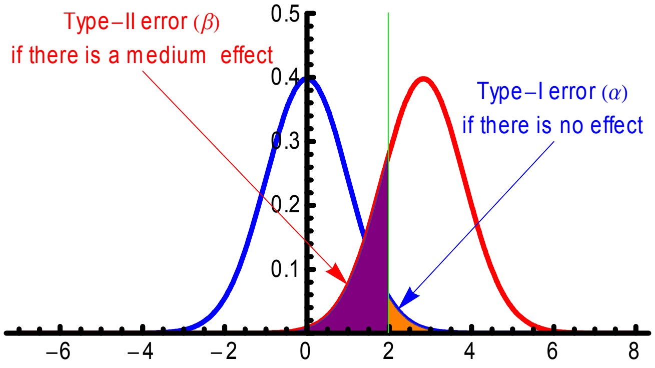

Figure 1. Distributions of the possible results of a *t* test if there is no effect (blue) or a medium-size effect (ES = ½, red). The green line is the critical value for two groups of 64 participants (equal to 1.979). The orange area represents the proportion of type-I error if there is no effect; the purple area represents the proportion of type-II error if there is a

The power of a 2-group *t* test is the probability of rejecting the null hypothesis given the fact that there is an effect (i.e. the effect size is different from zero). In planning an experiment, we need to assume what would be the effect size if there is one. The raw effect size is the difference between the two populations' mean, $|\mu_1-\mu_2|$, but in general, the effect size is given relative to the population standard deviation (a "standardized" effect size; Cohen, 1992, 1969, Rosnow and Rosenthal, 2003). Hence,

$$ES=\frac{|\mu_1-\mu_2|}{\sigma}.$$

The population standard deviation is estimated by the sample standard deviation across groups (the "pooled"

The author would like to thank an anonymous reviewer. Request for reprint should be addressed to Denis Cousineau, Département de psychologie, Université de Montréal, C. P. 6128, succ. Centre-ville, Montréal (Québec) H3C 3J7, CANADA, or using e-mail at Denis.Cousineau@Umontreal.CA. This research was supported by the *Conseil pour la Recherche en Sciences Naturelles et en Génie* du Canada.

standard deviation).

For example, suppose that you want to compare the time to find the exit from a maze for men and women. From informal pilot studies, you know that the average time (irrespective of the sex of the participants) is 17 seconds. More importantly, you found a standard deviation of 2 seconds in your pilot, again irrespective of sex. This is the pooled standard deviation (i.e. pooling together the groups). You believe that if a difference exists, it is probably in the order of 1 second. Relative to the sample's standard deviation, it represents an effect size of ½ (1 s / 2 s). This is considered a "medium" effect size (Cohen, 1992). Conversely, the ES times the pooled standard deviation yields back the expected raw effect size. Here, ½ × 2 s indeed yields back 1 s.

The probability β of a type-II error for a *t* test is given by

$$\beta = Pr(t_{statistic} < t_{critical} | ES \neq 0)$$

where

$$t_{statistic} = \frac{|\overline{X}_1 - \overline{X}_2|}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

in which $\overline{X}_1$ and $\overline{X}_2$ are the means of the two samples, n_1 and n_2 are the group sizes and S_p is the pooled standard deviation across the two groups. The easiest way to compute S_p is to take the standard deviation in the sample, irrespective of group. On the other hand, if each group's variance are known (say, $S^2_{x_1}$ and $S^2_{x_2}$), then S^2_p is the average of those, weighted by the groups' degrees of freedom. Hence:

$$S_p = \sqrt{\frac{(n_1 - 1)S^2_{x_1} + (n_2 - 1)S^2_{x_2}}{n_1 + n_2 - 2}}$$

which is the usual formula found in any textbook (e.g. Howell, 2004). The critical value $t_{critical}$ is read in a table with $n_1 + n_2 - 2$ degrees of freedom.

If a "non-central" *t* distribution with non-centrality parameter ES existed, we could directly compute power (Hélie, this issue). However, such distribution does not exist in current statistical packages.

To simplify the equation, let define the observed raw effect size $\Delta X = |\overline{X}_1 - \overline{X}_2|$ and note that

$$\frac{1}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \sqrt{\frac{1}{\frac{1}{n_1} + \frac{1}{n_2}}} = \sqrt{\frac{\tilde{n}}{2}}$$

in which $\tilde{n}$ is the harmonic mean of n_1 and n_2. Hence

$$t_{statistic} = \frac{\Delta X}{S_p} \times \sqrt{\frac{\tilde{n}}{2}}.$$

If there is an effect (the observed raw effect size $\Delta X \neq 0$), it will be magnified by being multiplied by a factor $\sqrt{\tilde{n}}$. Hence, with larger sample sizes, it is more probable that $t_{statistic}$ will exceed the critical value therefore increasing power.

With these notations, we have

$$\begin{aligned}\beta &= Pr\left(\frac{\Delta X}{S_p} \times \sqrt{\frac{\tilde{n}}{2}} < t_{critical} \mid ES \neq 0\right) \\ &= Pr\left(\frac{\Delta X}{S_p} \times \sqrt{\frac{\tilde{n}}{2}} - ES\sqrt{\frac{\tilde{n}}{2}} < t_{critical} - ES\sqrt{\frac{\tilde{n}}{2}}\right)\end{aligned}$$

obtained by subtracting the same quantity from both sides. Hence,

$$\beta = Pr\left(\frac{\Delta X - ES\ S_p}{S_p} \times \sqrt{\frac{\tilde{n}}{2}} < t_{critical} - ES\sqrt{\frac{\tilde{n}}{2}}\right)$$

The term $\Delta X - ES\ S_p$ is the difference between the expected raw effect and the observed raw effect size, which should be zero under our assumption. Hence, the left part of the inequality is a regular *t* statistic with mean zero and degrees of freedom $n_1 + n_2 - 2$ whose distribution is available on many statistical packages:

$$\beta = Pr\left(t < t_{critical} - ES\sqrt{\frac{\tilde{n}}{2}}\right)$$

and power is 1 minus the above.

For the previous example in which n_1 and n_2 were 64, the critical value is $t_{critical}$ = 1.979, $\tilde{n}$ = 64 (since the two groups are equal) and the expected effect size is ½. The power can be computed with SPSS using the following syntax (make sure that there is at least one line of data in your data editor):

```
COMPUTE power = 1 - CDF.T( 1.979 - (1/2) *
                 SQRT(64/2), 64 + 64 - 2 ).
EXECUTE.
```

In *Mathematica* 6.0, it is obtained with the similar commands (*Mathematica* is case-sensitive):

```
1 - CDF[ StudentTDistribution[64 + 64 - 2],
                 1.979 – (1/2) Sqrt[64/2] ]
```

The two software use the *cumulative distribution function* (CDF) which returns the probability that a $t_{statistic}$ is smaller than a given value.

In both cases, the power is found to be 0.80 (0.8014, to be exact). By going from 64 to 85 participants, the critical value jumps to 1.974 and the power would go from .80 to .90. A further increase of 5% to reach a power of .95 would require 105 participants per group. As seen, the increase is not linear.

References

Cohen, J. (1969). Statistical power analysis for the behavioral sciences. San Diego: Academic Press.

Cohen, J. (1992). *A power primer*. Psychological Bulletin, 112, 155-159.

Hays, W. L. (1973). Statistics for the social sciences. New York: Holt, Rinehart and Winston, inc.

Hélie, S. (in press). Understanding statistical power using noncentral probability distributions: Chi-squared, G-squared and ANOVA. Tutorials in Quantitative Methods for Psychology, (in press), 1-21.

Rosnow, R. L., & Rosenthal, R. (2003). *Effect sizes for experimenting psychologists*. Canadian Journal of Experimental Psychology, 57, 221-237.

Manuscript received June 9th, 2007