Simulation-Based Power-Analysis for Factorial ANOVA Designs

Daniël Lakens[1] & Aaron Caldwell[2]

[1] Eindhoven University of Technology, The Netherlands

[2] Department of Health, Human Performance and Recreation, University of Arkansas, USA

Author Note

All code used to create this manuscript is provided in an OSF repository at https://osf.io/xxxxx/.

Correspondence concerning this article should be addressed to Daniël Lakens, ATLAS 9.402, 5600 MB, Eindhoven, The Netherlands. E-mail: D.Lakens@tue.nl

Abstract

Researchers need to design informative studies. When the goal of an experiment is to test a hypothesis based on a frequentist hypothesis test it is important to justify the sample size based on the statistical power of the study. However, researchers are faced with challenges when they try to calculate power for factorial ANOVA designs. First, current software solutions are limited as they do not allow power analyses for more complex designs involving multiple within factors. Second, they require partial eta-squared or Cohen's f as input, which are not the most intuitive way to specify predicted effects in an ANOVA, and do not generalize to different experimental designs. We have created R functions and an online Shiny app that performs simulations for ANOVA designs for up to three factors with an unlimited number of levels. Predicted effects are entered by specifying means, standard deviations, and correlations (for within factors). The simulation provides a-priori power analyses for all effects in the ANOVA, and all simple comparisons. No other software is currently available that allows researchers to so easily perform power analyses for a wide range of ANOVA designs. The app plots *p*-value distributions for all tests, and allows researchers to select a range of options to correct for multiple comparisons. This tutorial will teach researchers how to perform power analysis for ANOVA designs, and through simulations illustrate important factors that determine the statistical power of factorial ANOVA designs.

Keywords: power analysis, ANOVA, hypothesis test, sample size justification

Word count: 5000

Simulation-Based Power-Analysis for Factorial ANOVA Designs

Statistical power is the probability of observing a statistically signficant result, given a specified effect size, and assuming there is a true effect. When a researcher aims to analyze a study based on an analysis of variance (ANOVA), the sample size of the study should be justified based on the statistical power of the test. A study has low power to detect effects the researcher is interested in has a high Type 2 error rate, and leads to a high probability of saying there is no effect, when there actually is a true effect to be found. Whereas power analyses for simple comparisons are relatively easy to perform, an a-priori power analysis for factorial ANOVA designs is a challenge. Available software solutions do not provide easy options for more complex designs (e.g., a 2x2x2 design, where the first factor is manipulated between participants, and the last two factors are manipulated within participants). Popular software solutions such as G*power require participants to enter their predictions as Cohen's f or partial eta squared, which are not the most intuitive ways to specify a hypothesized pattern of results, and which do not generalize to different experimental designs.

Here, we demonstrate how to perform power analyses for factorial ANOVA designs based on simulations. We provide R code and a Shiny app that can be used to calculate the statistical power based on a predicted pattern of means, standard deviations, and correlations (for within factors). Simulating studies, and calculating their p-values and effect sizes, are a useful way to gain a better understanding of the factors that determine the statistical power of hypothesis tests. After providing an introduction to statistical power in general, and for the F-test specifically, we will demonstrate how the power of factorial ANOVA designs depends on the pattern of means across conditions, the number of factors and levels, and the sample size. We will also illustrate the importance of controlling Type 1 error rates in exploratory ANOVA's, and the importance to design studies that have high power for main effects and interactions in an ANOVA, but also for follow-up tests for simple effects.

Factors That Determine Power in ANOVA Designs

You perform a study in which participants interact with an artificial voice assistant who sounds either cheerful or sad. You measure how much 70 participants in each condition enjoy interacting with the voice assistant on a line marking scale (coded continously from -5 to 5) and observe a mean of 0 in the sad condition, and a means of 1 in the cheerful condition, with an estimated standard deviation of 2. After submitting your manuscript for publications, reviewers ask you to add a study with a neutral control condition to examine whether cheerful voices increase, or sad voices decrease enjoyment (or both). Depending on what the mean in the neutral condition is, which sample size would you need to have a high powered study for the expected pattern or means? A collaborator suggests to switch from a between design to a within design, to more efficiently collect the data. Which consequence will this switching to a within-subject design have on the sample size planning? Because the effect size in the first study could be considered "medium" based on the benchmarks by Cohen (1988), does it make sense to plan for a "medium" effect size in either the between of within ANOVA design? And if you justify the sample size based on the ANOVA, will the study also have sufficient statistical power for the simple effects (or vice versa)?

Performing simulation studies is an excellent way to develop intuitions about these questions. The power in ANOVA designs depends on the pattern of means, the number of groups, the standard deviation, the correlation between dependent measures, the alpha level, and the sample size. After these factors have been specified, a simulation study can be perform to run the planned statistical test on generated data many times, and summarize the results. Where analytic power solutions exists in software such as G*Power for ANOVA designs with up to one within-subject factor, simulation studies are more flexible (and can for example be used to see what happens in a 3x3x3 within-subject desing).

Statistical Power When Comparing Differences Among Group Means

Let's consider the initial study we described above, where two group means are compared. We can test the difference between two means with a t-test of a one-way single factor ANOVA, and the two tests are mathematically equivalent. Figure 1 and Figure 2 visualize the distribution of Cohen's d and partial eta-squared that should be observed when there is no effect (grey curves) and when the observed difference between means equals the true effect. In both figures the light grey areas under the curve mark the observed results that would lead to a Type 1 error (observing a statistically significant result if the null-hypothesis is true) and the dark grey areas under the curve marks the observed effect sizes that would lead to a Type 2 error. An observed effect is statistically significant if the observed effect size is larger than the critical value. Critical values are often expressed as t-values or F-values, but can be expressed as effect sizes (Cohen's d and partial eta-squared), and any observed effect size larger than the critical effect size will be statistically significant. The goal of power analysis is to choose a sample size so that the probability of observing a statistically significant effect for a specified effect size reaches a desired probability. In both Figures, 83.5% of the expected effect sizes, if the true effect size is d = 0.5 or $\eta_p^2 = 0.0588$ and 70 participants in each condition are collected, will be more extreme than the critical effect size (which is d = 0.334 or $\eta_p^2 = 0.028$).

Two relationships between the t-test and the F-test when comparing two means are worth pointing out. First, $F = t^2$, or the F-value equals the t-value, squared. Whereas we typically think of a t-test as the difference between means, the relationship between the t-test and F-test reveals that we can also use the group means to calculate the variance of the difference between the two means $(m_1 - m_2)^2$. The F-test is used to compute the ratio of the between group variance and the within group variance, and the between group variance can be used to compare more than two means. The second relationship between a t-test and a F-test for two groups is that the effect size for an ANOVA, Cohen's f, is half the

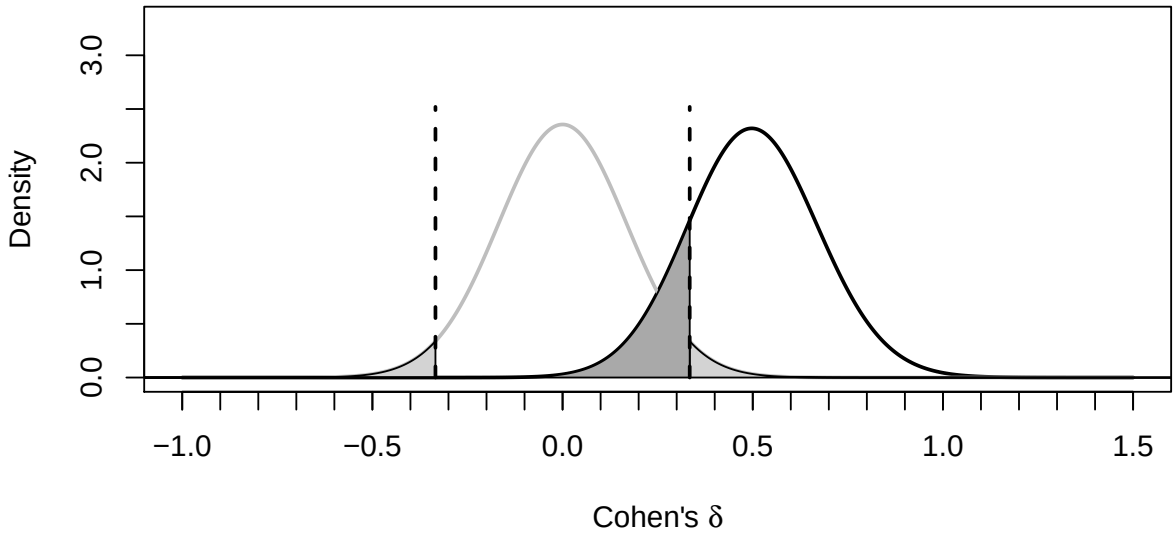

Figure 1. Distribution of Cohen's d under the null-hypothesis (grey curve) and alternative hypothesis assuming d = 0.5 (black curve). Note that

size of the effect size for standardized mean differences, Cohen's d, of $f = \frac{1}{2}d$. Cohen's d is calculated by dividing the difference between means by the standard deviation, or

$$d = \frac{m_1 - m_2}{\sigma}. \tag{1}$$

If we have two groups with means of 1 and 2, and the standard deviation is 2, Cohen's d is (2-1)/2, or 0.5. Cohen's f is the standard deviation of the population means divided by the population standard deviation (Cohen, 1988), or:

$$f = \frac{\sigma_m}{\sigma} \tag{2}$$

where for equal sample sizes,

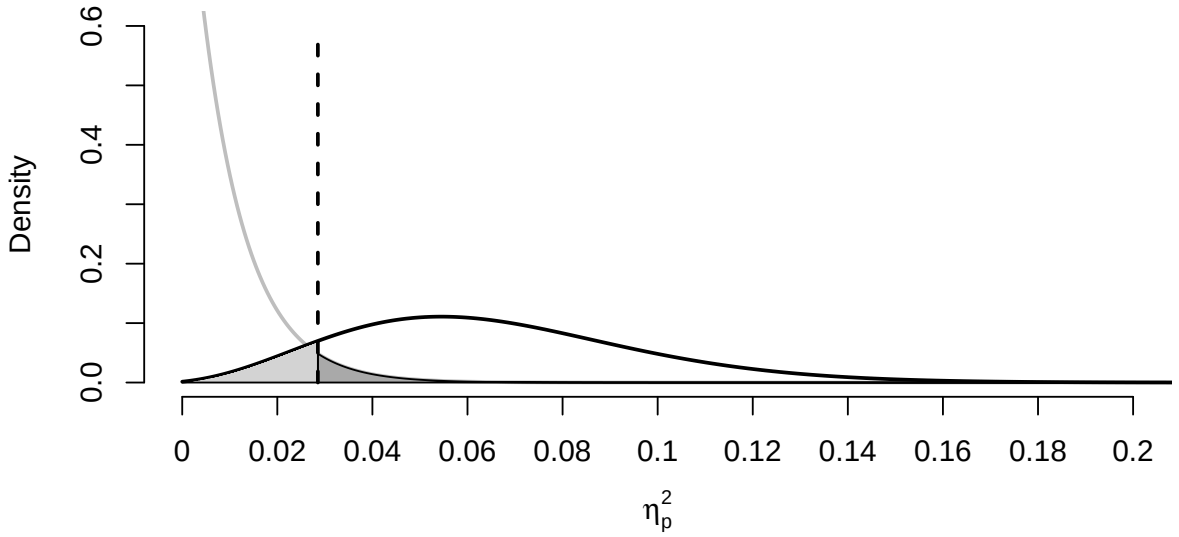

Figure 2. Distribution of eta-squared under the null-hypothesis (grey curve) and alternative hypothesis assuming partial eta-squared = 0.0588 (black curve).

$$\sigma_m = \sqrt{\frac{\sum_{i=1}^{k}(m_i - m)^2}{k}}. \tag{3}$$

Because Cohen's f is an essential part of power power analyses for factorial ANOVA designs, it is worth illustrating how it is calculated in an example. If we again take two means of 1 and 2, and a standard deviation of 2, the grand mean is 1.5. We subtract each condition mean from the grand mean, take the square, calculate the sum of squares, divide it by two, and take the square root. $\sigma_m = \sqrt{\frac{(1-1.5)^2+(2-1.5)^2}{2}} = \sqrt{\frac{0.25+0.25}{2}} = 0.5$, and $f = \frac{0.5}{2} = 0.25$. We see Cohen's f is half as large as Cohen's d. Power analyses for ANOVA are based on Cohen's f, but popular power anaysis software such as G*Power (Faul, Erdfelder, Lang, & Buchner, 2007) also allows researchers to specify the effect size as partial eta-squared (η_p^2). Partial eta-squared can be converted into Cohen's f:

$$f = \sqrt{\frac{\eta_p^2}{1 - \eta_p^2}} \tag{4}$$

and Cohen's f can be converted into partial eta-squared:

$$\eta_p^2 = \sqrt{\frac{f^2}{f^2 + 1}} \tag{5}$$

In the example above, $\eta_p^2 = 0.25^2/(0.25^2 + 1) = 0.0588$.

Because Cohen's f is calculated based on the sum of squares, the value is always positive, just as the F-value for an ANOVA is always positive.

References

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, N.J: L. Erlbaum Associates.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). GPower 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. doi:10.3758/BF03193146