

Simulation-Based Power-Analysis for Factorial ANOVA Designs

Daniël Lakens¹ & Aaron R. Caldwell^{2,3}

¹ Human-Technology Interaction Group, Eindhoven University of Technology, The
Netherlands

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

³ Thermal and Mountain Medicine Division, U.S. Army Research Institute of Environmental
Medicine, USA

Author Note

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 often rely on analysis of variance (ANOVA) when they report results of experiments. To ensure a study is adequately powered to yield informative results when performing an ANOVA, researchers can perform an a-priori power analysis. However, power analysis for factorial ANOVA designs is often a challenge. Current software solutions do not enable power analyses for complex designs with several within-subject factors. Moreover, power analyses often need partial eta-squared or Cohen's f as input, but these effect sizes are not intuitive and do not generalize to different experimental designs. We have created the R package Superpower and an online Shiny app to enable researchers without extensive programming experience to perform simulation-based power analysis for ANOVA designs of up to three within- or between-subject factors, with an unlimited number of levels. Predicted effects are entered by specifying means, standard deviations, and correlations (for within-subject factors). The simulation provides the statistical power for all ANOVA main effects, interactions, and individual comparisons, and allow researchers to correct for multiple comparisons. The simulation plots p -value distributions for all tests, and power plots across a range of sample sizes. This tutorial will demonstrate how to perform power analysis for ANOVA designs, and highlights important factors that determine the statistical power of factorial ANOVA designs.

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

Word count: 4654 words.

Simulation-Based Power-Analysis for Factorial ANOVA Designs

When a researcher aims to test hypotheses with an analysis of variance (ANOVA), the sample size of the study should be justified based on the statistical power of the test. The statistical power of a test is the probability of rejecting the null-hypothesis, given a specified effect size, alpha level, and sample size. When power is low there is a high probability of concluding there is no effect when an underlying effect may exist in the population of interest. Several excellent resources exist that explain power analyses, including books (Aberson, 2019; Cohen, 1988), general reviews (Maxwell, Kelley, & Rausch, 2008), and practical primers (Brysbaert, 2019; Perugini, Gallucci, & Costantini, 2018). Whereas power analyses for individual comparisons are relatively easy to perform, power analyses for factorial ANOVA designs are a bigger challenge. There are many current software solutions that *can* calculate power for factorial ANOVAs (Campbell & Thompson, 2012; Faul, Erdfelder, Lang, & Buchner, 2007; Lang, 2017; J. Westfall, 2015a), but the options within these packages are often limited (e.g, Gpower has limited options for within subjects factors). Available software solutions do not provide easy options to specify 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). The predicted effects often need to be specified as a standardized effect size such as Cohen's f or partial eta squared (η_p^2), which are not the most intuitive way to specify a hypothesized pattern of results, and these effect sizes do not generalize to different experimental designs. Simulations based on a specified pattern of means and a covariance matrix (based on the expected standard deviation and correlation between within participant factors) provide a more flexible approach to power analyses. However, such simulations typically require extensive programming knowledge.

In this manuscript we introduce Superpower, an R package and Shiny app that can be used to perform power analyses for factorial ANOVA designs based on simulations. Superpower can be used to perform a-priori power analyses based on a predicted pattern of

means, standard deviations, and (for within-subject factors) correlations. By simulating data for factorial designs with specific parameters researchers can gain a better understanding of the factors that determine the statistical power of an ANOVA, and learn how to design well-powered experiments. After a short introduction to statistical power, focusing on the F -test, we will illustrate through simulations how the power of factorial ANOVA designs depend on the pattern of means across conditions, the number of factors and levels, the sample size, and whether you need to control the alpha level for multiple comparisons.

A basic example

Imagine you plan to perform a study in which participants interact with an artificial voice assistant who sounds either cheerful or sad. You measure how much 80 participants in each condition enjoy interacting with the voice assistant on a line marking scale (coded continuously from -5 to 5) and observe a mean of 0 in the sad condition, and a mean of 1 in the cheerful condition, with an estimated standard deviation of 2. After submitting your manuscript for publication, 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 enjoyment in the neutral condition is, what sample size would you need to collect for a high powered test of the expected pattern of means? A collaborator suggests to switch from a between-subject design to a within-subject design to collect data more efficiently. What impact will switching to a within-subject design have on the required sample size? The effect size in the first study is sometimes referred to as a “medium” effect size based on the benchmarks by Cohen (1988). Does it make sense to perform an a-priori power analysis for a “medium” effect size if we add a third between-subject condition, or switch to a within-subject ANOVA design? And if you justify the sample size based on the power for the main effect for the ANOVA, will the study also have sufficient statistical power for the independent comparisons between conditions (or vice versa)? Before we answer these

questions, let's consider some of the basic concepts of statistical power and how power calculations are typically performed.

Calculating Power for ANOVA Designs

Let's consider the initial design described above, where enjoyment is measured when 80 participants per condition interact with a cheerful or sad voice assistant. We can test the difference between two means with a t -test or a one-way ANOVA, and the two tests are mathematically equivalent. Figure 1 and Figure 2 visualize the distribution of the effect sizes Cohen's d (for the t -test) and η_p^2 (for the F -test) that should be observed when there is no effect (grey curves) and when the observed difference between means equals the true effect (black curves)¹. In both figures the light grey areas under the null-distribution mark the observed effect sizes 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 mark the observed effect sizes that would lead to a Type 2 error (observing a non-significant result when there is a true effect). To perform an a-priori power analysis, researchers need to specify an effect size for the alternative hypothesis (for calculations, see Box 1).

¹Note that we refer to sample level statistics by default, and explicitly mention whenever we refer to population parameters instead.

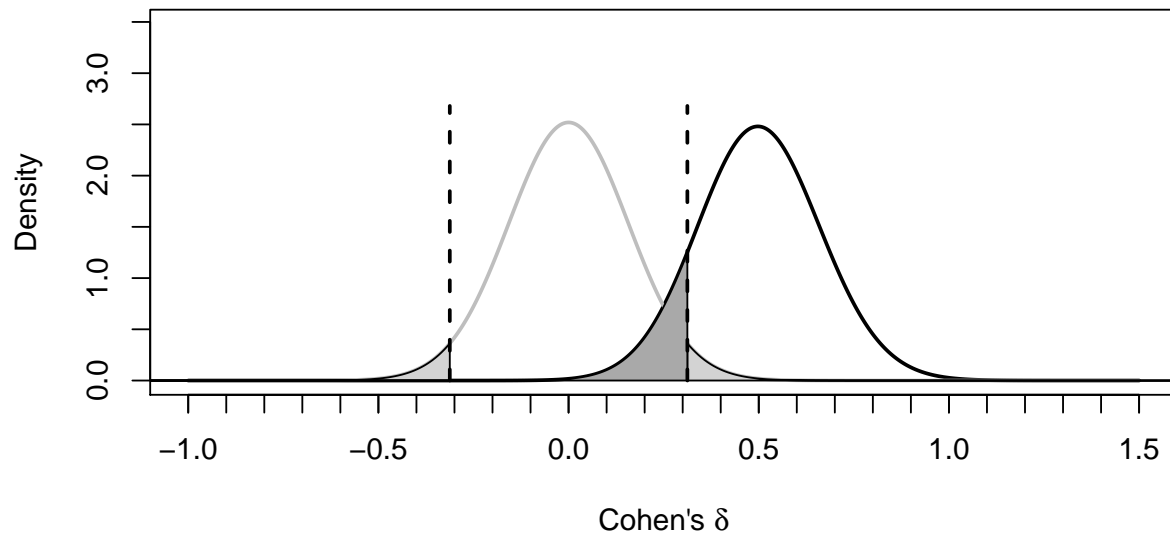


Figure 1. Distribution of Cohen's d under the null-hypothesis (grey curve) and alternative hypothesis assuming $d = 0.5$ (black curve) given $n = 80$.

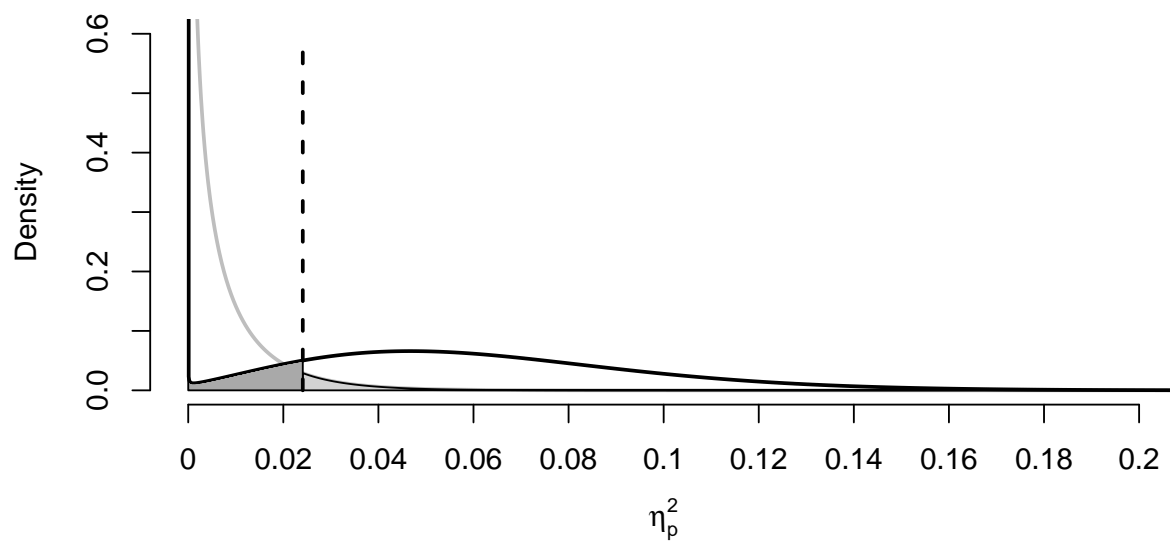


Figure 2. Distribution of eta-squared under the null-hypothesis (grey curve) and alternative hypothesis assuming partial eta-squared = 0.0588 (black curve) given $n = 80$.

A test result is statistically significant when the p -value is smaller than the alpha level, or when the test statistical (e.g., a F -value) is larger than a critical value. For a given sample size we can also calculate a critical *effect size*, and a result is statistically significant if the observed effect size is more extreme than the critical effect size. Given the sample size of 80 participants per group, observed effects are statistically significant when they are larger than $d = 0.31$ in a t -test, or $\eta_p^2 = 0.02$ for the F -test. The goal of an a-priori power analysis is to design a study with a desired probability of observing a significant effect. To calculate the sample size required to reach a desired statistical power one has to specify the alternative hypothesis, the sample size, and the alpha level. Based on λ (the noncentrality parameter, which together with the degrees of freedom specifies the shape of the expected effect size distribution under a specified alternative hypothesis, illustrated by the black curves in Figure 1 and 2) we can calculate the area under the curve that is more extreme than the critical effect size (i.e., Figure 2 to the right of the critical effect size). Under the alternative hypothesis that the true effect size is $d = 0.5$ or $\eta_p^2 = 0.0588$, data are collected from 80 participants in each condition, and an alpha of 0.05 is used, in the long run 8,816.02% of the observed data will yield statistically significant results.

Box 1. Formula for effect sizes for ANOVA designs

For two independent groups, the t -statistic can easily be translated to the F -statistic $F = t^2$. Cohen's d , a standardized effect size, is calculated by dividing the difference between means by the standard deviation, or

$$d = \frac{m_1 - m_2}{\sigma}. \quad (1)$$

The generalization of Cohen's d to more than two groups is Cohen's f , which is the standard deviation of the means divided by the standard deviation (Cohen, 1988), or:

$$f = \frac{\sigma_m}{\sigma} \quad (2)$$

where for equal sample sizes,

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

For two groups Cohen's f is half as large as Cohen's d , or $f = \frac{1}{2}d$. Partial eta-squared, which is often used as input in power analysis software, can be converted into Cohen's f :

$$f = \sqrt{\frac{\eta_p^2}{1 - \eta_p^2}} \quad (4)$$

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

$$\eta_p^2 = \frac{f^2}{f^2 + 1} \quad (5)$$

Power calculations rely on the noncentrality parameter (lambda, (λ) .) In a between-participants one-way ANOVA lambda is calculated as:

$$\lambda = f^2 \times N \quad (6)$$

where f is Cohen's f and N is the total sample size.

Simulating Statistical Power for Different Factorial Designs

Superpower can be used in R or in an online Shiny app. The code underlying the Superpower R package and the Shiny app generates data for each condition in the design and performs an ANOVA and *t*-tests for all comparisons between conditions. The `ANOVA_exact` function simulates one function that has exactly the desired statistical properties, and computes power directly from the test results on this perfect dataset. The `ANOVA_power` function simulates datasets repeatedly based on the specified parameters and calculates the percentage of statistically significant results. The simulation can be performed based on any design specified using the `ANOVA_design` function. Users specify the design based on the number of levels for each factor (e.g., 2) and whether the factor is manipulated within or between participants (by entering a “w” or a “b”). Superpower can handle up to three factors (separated by “’”). *A 2b design means a single factor with two groups manipulated between participants, whereas a 2b2w design is a 2 x 2 mixed ANOVA where the first factor is manipulated between, and the second within participants.* Users specify the sample size per group (*n*), the predicted pattern of means across all conditions, the expected standard deviation, and the correlation between variables (for within designs). To make it easier to interpret the output users can specify factor names and names for each factor level (e.g., ‘condition, cheerful, sad’). Detailed examples for a wide range of designs are available in an online manual at <http://arcaldwell49.github.io/SuperpowerBook>.

An example of the R code is:

```
design_result <- ANOVA_design(  
  design = "2b",  
  n = 80,  
  mu = c(1, 0),  
  sd = 2,
```

```
labelnames = c("condition", "cheerful", "sad"))
```

An example of the input in the Shiny app is:

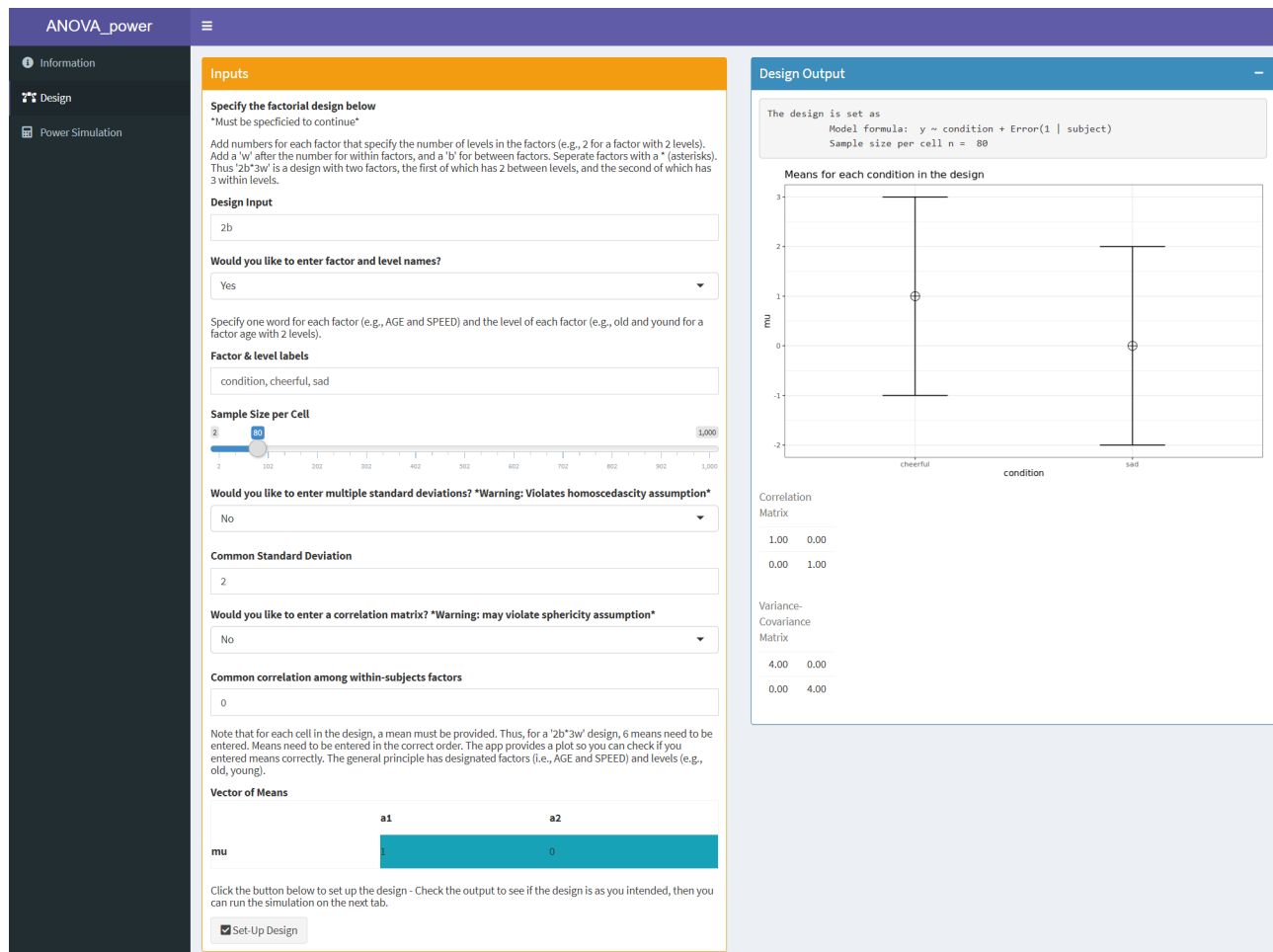


Figure 3. Screenshot of ANOVA powerShiny app

For a visual confirmation of the input a figure is created that displays the means and standard deviation (see the right side of Figure 3). After the design has been specified there are two ways to calculate the statistical power of an ANOVA through simulations. The first is to repeatedly simulate datasets and compute the percentage of statistically significant results. This is completed with the ANOVA_power function which performs a Monte Carlo simulation.

```
power_result_monte <- ANOVA_power(design_result, nsims = 100000)
```

The second is to simulate a dataset that has *exactly* the desired properties, perform an ANOVA, and use the ANOVA results to compute the statistical power. This is completed with the `ANOVA_exact` function which performs a “exact” simulation.

```
power_result_exact <- ANOVA_exact(design_result)
```

The first approach is a bit more flexible (e.g., it allows for sequential corrections for multiple comparisons such as the Holm procedure), but the second approach is much faster (and generally recommended). Because the true pattern of the data are unknown in an a-priori power analysis, there is often uncertainty about the values that need to be entered in a power analysis. It makes sense to examine power across a range of assumptions, from more optimistic scenarios, to more conservative estimates. In many cases researchers should consider collecting a sample size that guarantees sufficient power for the smallest effect size of interest, instead of the effect size they expect (for examples, see Lakens, Scheel, & Isager (2018)). This guarantees the study will yield an informative answer, even when there is uncertainty about the true effect size. If `ANOVA_power` is used the results from the simulation will vary each time the simulation is performed (unless a seed is specified, e.g., ‘set.seed = 2019’). A user should specify the number of simulations (the more simulations, the more accurate the results are,, but the longer the simulation takes), the alpha level for the tests, and any adjustments for multiple comparisons that are required. The output from `ANOVA_exact` and `ANOVA_power` is similar, and provides the statistical power for the ANOVA and all simple comparisons between conditions.

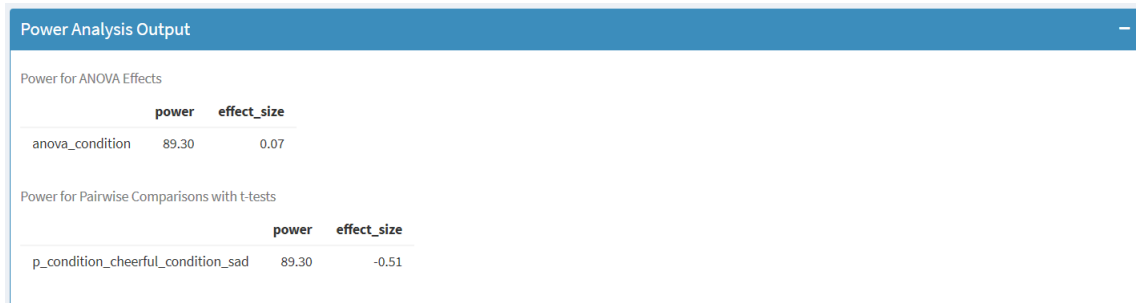
Power and Effect sizes for ANOVA tests

	power	effect_size
anova_condition	88.191	0.06425

Power and Effect sizes for pairwise comparisons (t-tests)

	power	effect_size
p_condition_cheerful_condition_sad	88.191	-0.5017

The same results are returned in the online Shiny app, but here users can also choose a “download PDF report” option to receive the results as a PDF file that can be saved to be included in a preregistration.



Power Analysis Output		
Power for ANOVA Effects		
	power	effect_size
anova_condition	89.30	0.07
Power for Pairwise Comparisons with t-tests		
	power	effect_size
p_condition_cheerful_condition_sad	89.30	-0.51

Figure 4. Screenshot of the results of the power analysis in the ANOVA_power Shiny app.

From these results we see that when 100,000 simulations are performed for our two group between subjects design with means of 1 and 0, a standard deviation of 2, and 80 participants in each group, with a seed set to 2019 (these settings will be used for all simulation results reported in this manuscript), the statistical power (based on the percentage of $p < \alpha$ results) is 88.19% and the average η_p^2 is 0.06. The simulation also provides the results for the individual comparisons based on t -tests. Since there are only two groups in this example, the statistical power for the individual comparison is identical to the ANOVA, but the expected effect size is given in Cohen’s d : -0.50.

Now that the basic idea behind power analyses in Superpower is illustrated, we can use it to explore how changes to the experimental design influence power, and answer some of the questions our hypothetical researcher is confronted with when designing a follow-up study. We will first examine what happens if we add a third, neutral, condition to the design. Let’s assume a researcher expects the mean enjoyment rating for the neutral voice condition

to fall either perfectly between the cheerful and sad conditions, or to be equal to the cheerful condition. The researcher wonders if simply collecting 80 additional participants in the neutral condition is sufficient to find a significant effect in the One-Way ANOVA. The R code to specify this design is:

```
design_result_1 <- ANOVA_design(  
  design = "3b",  
  n = 80,  
  mu = c(1, 0.5, 0),  
  sd = 2,  
  labelnames = c("condition", "cheerful", "neutral", "sad"))
```

The design now has 3 between-participant conditions, and we can explore what happens if we would collect 80 participants in each condition.

If we assume the mean falls exactly between between the cheerful and sad conditions, the simulations show the statistical power for a 3-groups one-way ANOVA F -test is reduced to 81.14 %, and the effect size (partial eta-squared) is 0.05. If we assume the mean is equal to the cheerful condition, the power increases to 91.03%. Compared to the two group design (where the power was 88.19%), three things have changed. First, the numerator degrees of freedom has increased because an additional group is added to the design, which makes the non-central F -distribution more similar to the central F -distribution, which reduces the statistical power. Second, the total sample size is 50% larger after adding 80 participants in the third condition, which increases the statistical power of the ANOVA. Third, the effect size, Cohen's f , has decreased from 0.25 to either 0.20 or 0.24, which reduces the statistical power. The most important take-home message is that changing an experimental design can have several opposing effects on the power of a study, depending of the pattern of means. The exact effect of these three changes on the statistical power is difficult to anticipate from one ANOVA design to the next. This highlights the importance of performing an a-priori

power analysis based on a specific pattern of means that you predict.

Although an initial goal when performing an ANOVA might be to test the *omnibus null hypothesis*, which answers the question whether there are *any* differences between group means, we often want to know which conditions differ from each other. Thus, an ANOVA is often followed up by individual comparisons (whether *planned* or *post-hoc*). Superpower automatically provides the statistical power for all individual comparisons that can be performed. By default, the power and effect size estimates are based on simple *t*-tests. It is also possible to combine variance estimates from all conditions and calculate the estimated marginal means (Lenth, 2019) when performing individual comparisons by setting `emm = TRUE` within the `ANOVA_power` or `ANOVA_exact` functions, or checking this option in the Shiny app. Doing so can have power benefits (Maxwell, Delaney, & Kelley, 2017), depending on whether the assumption of equal variances (also known as the homogeneity assumption) is met, which may not be warranted in psychological research (Delacre, Lakens, Mora, & Leys, 2018). The degree to which violations of the homogeneity assumption affect Type 1 error rates can be estimated with the `ANOVA_power` function (see Assumptions section below).

Power analysis for individual comparisons is relatively straightforward and can easily be done in all power analysis software, but providing power for all individual comparisons alongside the ANOVA result by default hopefully nudges researchers to take into account the power for follow-up tests. Depending on the pattern of means in the three conditions, we are interested in the statistical power for a mean difference of 0.5, or a mean difference of 1. Furthermore, when performing multiple individual comparisons, we need to choose the alpha level such that we prevent an inflated Type 1 error rate. By adjusting for multiple comparisons we ensure that we do not conclude there is an effect in *any* of the individual tests more often than the desired Type 1 error rate. Several techniques to control error rates exist, of which the best known is the Bonferroni adjustment. The Holm procedure is slightly more powerful than the Bonferroni adjustment, without requiring additional assumptions

(for other approaches, see Bretz, Hothorn, & Westfall, 2011). Power analyses using a manually calculated Bonferoni correction can be performed with the `ANOVA_exact` function by specifying the adjusted alpha level, but the sequential Holm approach can only be performed in the `ANOVA_power` simulation approach. Because the adjustment for multiple comparisons lowers the alpha level, it also lowers the statistical power. For the paired comparisons we see we have approximately 80% power for differences of 0.5 after controlling for multiple comparisons with the Holm procedure (compared to 88% power without correcting for multiple comparisons. As the number of possible paired comparisons increases, the alpha level is reduced, and power becomes lower, all else equal. These power analyses reveal the cost (in terms of the statistical power) of exploring across all possible paired comparisons while controlling error rates. The reduction in power due to the lower alpha level should be counteracted by increasing the sample size to maintain an adequate level of power for both the ANOVA as the individual comparisons. If a researcher is only interested in specific comparisons it is advisable to preregister and test only these comparisons instead of correcting the alpha level for all possible comparisons.

Power for Within-Subject Designs

What happens if we would perform the second study as a within-participants design? Instead of collecting three groups of participants, we only collect one group, and let this group evaluate the cheerful, neutral, and sad voice assistants. The `ANOVA_design` function below specifies this design. Note the design has changed from `3b` (a one factor between design with three levels) to `3w` (a one factor within design with three levels) and the correlation parameter `r = 0.5` is added, which specifies the expected correlation between dependent variables in the population.

```

design_result_within_1 <- ANOVA_design(
  design = "3w",
  n = 80,
  mu = c(1, 0.5, 0),
  sd = 2,
  r = 0.5,
  labelnames = c("condition", "cheerful", "neutral", "sad")
)

```

A rough but useful approximation of the sample size needed in a within-subject design (N_W), relative to the sample needed in between-design (N_B), is (from Maxwell & Delaney, 2004, p. 562, formula 47):

$$N_W = \frac{N_B(1 - \rho)}{a} \quad (7)$$

Here a is the number of within-participant levels, ρ is the correlation between measurements in the population. From this formula we see that switching from a between to a within design reduces the required sample size simply because each participant contributes data to each condition, even if the correlation between measurements is 0. In our example a within design would require three times as few participants as a between subjects design with three conditions. If the correlation between dependent variables is positive, the standard deviation of the difference scores is smaller in a within design than in a between design. Because the standardized effect size is the mean difference divided by the standard deviation of the difference scores, a positive correlation increases the standardized mean difference in a within-subject design, which increases the statistical power.

Box 1. Formula for effect sizes for within designs

The effect size in a within-design is referred to as Cohen's d_z (because it is the effect size of the difference score between x^* and y^* , z^*). The relation is:

$$\sigma_z = \sigma \sqrt{2(1 - \rho)} \quad (8)$$

Cohen's d_z is used in power analyses for dependent t^* -tests, but there is no equivalent Cohen's f_z for a within-participant ANOVA, and Cohen's f is identical for within and between designs. Instead, the value for lambda (λ) is adjusted based on the correlation. For a one-way within-participant design lambda is identical to Equation [eqrefeq:lambda](#), multiplied by u^* , a correction for within-subject designs, calculated as:

$$u = \frac{k}{1 - \rho} \quad (9)$$

where k is the number of levels of the within-participant factor, and ρ is the correlation between dependent variables. Equations 4 and 5 no longer hold when measurements are correlated. The default settings in GPower expects an f or η_p^2 that does not incorporate the correlation, while the correlation is incorporated in the output of software packages such as SPSS. One can enter the η_p^2 from SPSS output in GPower after checking the 'as in SPSS' checkbox in the options window, but forgetting this is a common mistake in power analyses for within designs in GPower. For a one-way within-subject design, Cohen's f can be converted into the Cohen's f SPSS uses through:

$$f_{SPSS}^2 = f^2 \times \frac{k}{k - 1} \times \frac{n}{n - 1} \times \frac{1}{1 - \rho} \quad (10)$$

and subsequently transformed to η_p^2 through Equation 5.

We can perform the simulation-based power analysis with the `ANOVA_power` or

ANOVA_exact functions.

```
power_result_within_1 <- ANOVA_power(design_result_within_1,
                                     nsims = 100000)

exact_result_within_1 <- ANOVA_exact(design_result_within_1)
```

Revisiting our between-participant design, power was 81.14% when the enjoyment scores were uncorrelated. If we want to examine the power for a within design we need to enter our best estimate for the true population value of the correlation between dependent measurements. Ideally this value is determined based on previous studies, and when there is substantial uncertainty about the true population value it often makes sense to explore a range of plausible correlations. Let's assume our best estimate of the correlation between enjoyment ratings in a within-subject design is $r = 0.5$. The power for a repeated-measures ANOVA based on these values, where ratings for the three conditions are collected from 80 participants, is 98.38%. Because of the positive correlation between dependent variables, the effect size η_p^2 is much larger for the within-subject design ($\eta_p^2 = 0.12$) than for the 3 group between participants design ($\eta_p^2 = 0.05$), and a researcher might decide to collect somewhat less participants, while staying having a sufficiently low Type 2 error rate. The Superpower package allows researchers to enter a correlation matrix that specifies the expected correlations between each individual pair of measurements, instead of assuming the correlations between all dependent variables are identical.

Power for Interactions

So far we have explored power analyses for one factor designs. Superpower can easily provide statistical power for designs with up to three factors of up to 999 levels (e.g., a 4b*2w2w would specify a mixed design with two within factors which 2 levels, and one

between factor with 4 levels). Let's assume the researcher plans to perform a follow-up experiment where in addition to making the voice sound cheerful or sad, a second factor is introduced by making the voice sound more robotic compared to the default human-like voice. Different patterns of results could be expected in this 2 by 2 design. Either the same effect is observed for robotic voices, or no effect is observed for robotic voices, or the opposite effect is observed for robotic voices (we enjoy a sad robotic voice more than a cheerful one, a "Marvin-the-Depressed-Robot Effect"). In the first case, we will only observe a main effect of voice, but in the other two scenarios there is an interaction effect between human-likeness of the voice and the emotional tone of the voice. We specify the design with the `ANOVA_design` function.

```
design_result_cross_80 <- ANOVA_design(  
  design = "2b*2b",  
  n = 80,  
  mu = c(1, 0, 0, 1),  
  sd = 2,  
  labelnames = c("condition", "cheerful", "sad",  
                 "voice", "human", "robot")  
)
```

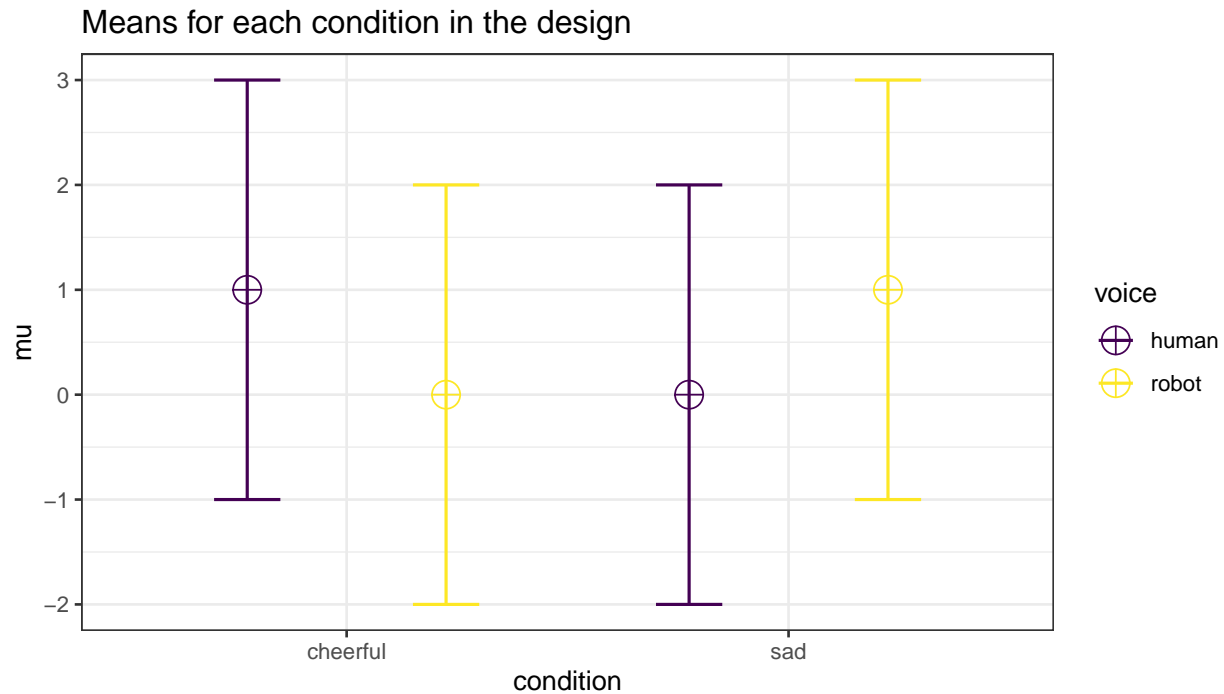


Figure 5. Visualization for the expected means and standard deviations for a crossover interaction. Error bars represent one standard deviation.

We can simulate a cross-over interaction for a 2x2 between-participant design with 80 participants in each group to examine the statistical power (see Figure 5 for the expected pattern of means). Again, we can do this with both simulation functions.

```
power_result_cross_80 <- ANOVA_power(design_result_cross_80,
                                     nsims = 100000)

exact_result_cross_80 <- ANOVA_exact(design_result_cross_80)
```

Mathematically the interaction effect is computed as the cell mean minus the sum of the grand mean, the marginal mean in each row minus the grand mean, and the marginal mean in each column minus grand mean (see Maxwell et al. (2017)). For example, for the cheerful human-like voice condition this is 1 (the value in the cell) - (0.5 [the grand mean] + 0.5 [the cell mean minus the marginal mean in row 1] + 0.5 [the cell mean minus the

marginal mean in column 2]). Thus, $1 - (0.5 + 0.5 + 0.5) = -0.5$. Completing this for all four cells gives the values -0.5, 0.5, 0.5, -0.5. Cohen's f is then $f = \sqrt{\frac{-0.5^2 + 0.5^2 + 0.5^2 + -0.5^2}{4}} = 0.25$. Simulations show we have 99.38% power when we collect 80 participants per condition. Power is high, because we collected 80 participants in each condition. Compared to the two-group comparison with 80 participants per group a cross-over (also called “disordinal”) interaction with two levels per factor has the same power as the initial two-group design if we halve the sample size per condition (i.e., $n = 40$ in each cell). This is accomplished simply by changing the `n` in the `ANOVA_design` function.

```
design_result_cross_40 <- ANOVA_design(
  design = "2b*2b",
  n = 40,
  mu = c(1, 0, 0, 1),
  sd = 2,
  labelnames = c("condition", "cheerful", "sad",
                 "voice", "human", "robot"),
  plot = FALSE
)
```

And, power can be estimated with the simulation functions.

```
power_result_cross_40 <- ANOVA_power(design_result_cross_40,
                                     nsims = 100000)

exact_result_cross_40 <- ANOVA_exact(design_result_cross_40)
```

Once we make that change, power with 40 participants per condition is 88.31%. Main effects in an ANOVA are based on the means for one factor averaged over the other factors (e.g., the main effect of a human-like versus robot-like voice, irrespective of whether it is

cheerful or sad). The interaction effect can be contrast coded as 1, -1, -1, 1, and thus tests the scores of 80 participants against 80 other participants, just as for the t -test or the main effects. The key insight here is that not the sample size per condition, but the total sample size over all other factors determines the power for the main effects and the interaction (cf. J. Westfall, 2015b).

We can also examine the statistical power for a pattern of results that indicated that there was no difference in interacting with a cheerful or sad conversational agent with a robot voice. In this case, we expect an “ordinal” interaction (the means for the human-like voice are never lower than the means for the robot-like voice, and thus there is no cross-over effect). The expected pattern of means is 1, 0, 0, 0, with only a single mean that differs from the others, and we create this new design.

```
design_result_ordinal <- ANOVA_design(  
  design = "2b*2b",  
  n = 160,  
  mu = c(1, 0, 0, 0),  
  sd = 2,  
  labelnames = c("condition", "cheerful", "sad",  
                 "voice", "human", "robot")  
)
```

As has been pointed out (Giner-Sorolla, 2018; Simonsohn, 2014) these designs require larger samples sizes to have the same power to detect the interaction, compared to the two-group comparison. The reason for this is that the effect size is only half as large, with Cohen’s $f = 0.125$ (compared to 0.25 in the cross-over interaction).

By steadily increasing the sample size in the simulation, we can examine the required sample size to get the same power as for the cross-over interaction with 40 participants per

cell (88.31% power). A total sample size of 635 is required, almost four times as large as the total sample size for the two-group comparison (160). To make it easier to find the sample size per group that is required to calculate the expected power the `plot_power` function in Superpower can be used to plot the power for a specific design across a range of sample sizes (see Figure 7; code included below).

```
plot_data <- plot_power(design_result_ordinal,
                        min_n = 10, max_n = 200,
                        plot = TRUE)
```

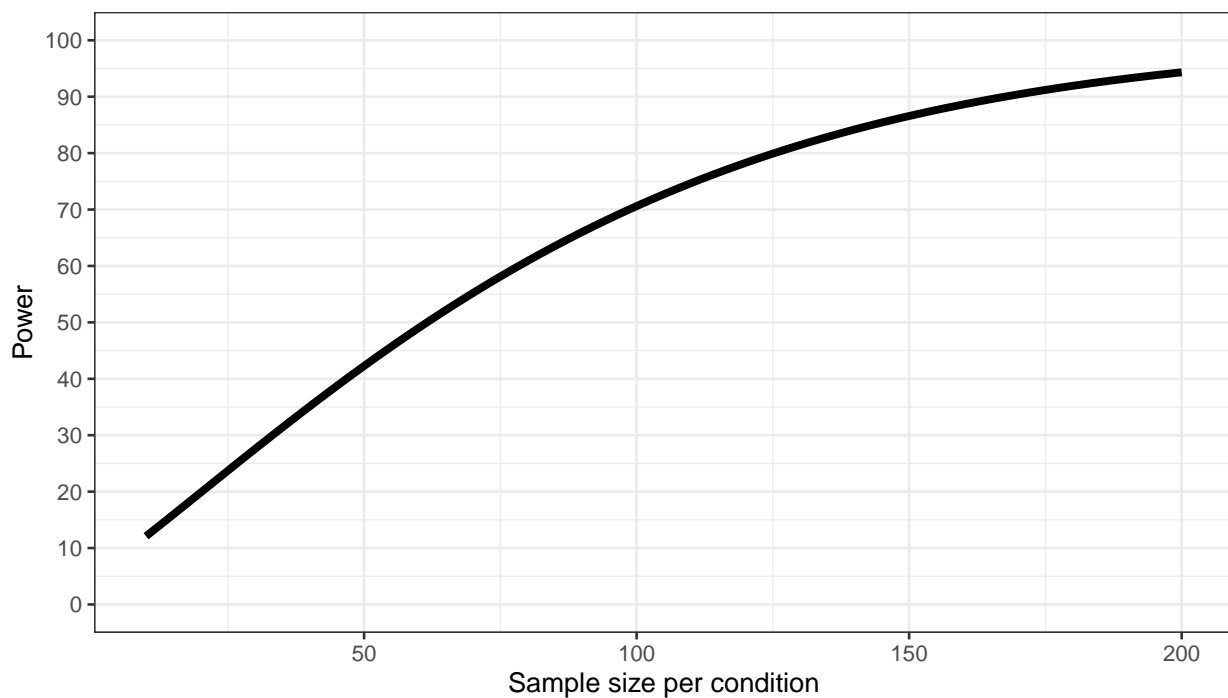


Figure 6. Power curves across a range of sample sizes per group from $n = 10$ to $n = 200$ for the expected main effects and ordinal interaction.

Where a two group comparison has a Cohen's f of 0.25 when the means are 1 and 0, a 2x2 ordinal interaction has a Cohen's f of 0.25 when the pattern of means is 2, 0, 0, 0, and a 2x2x2 interaction where only one cell differs from the other means has a Cohen's f of 0.25 when the pattern of means is 4, 0, 0, 0, 0, 0, 0, 0 across the eight cells. The take-home

message is that a “medium” effect size translates into a much more extreme pattern of means in an ordinal interaction than in a disordinal (crossover) interaction, or in a 2x2x2 interaction compared to a 2x2 interaction (see also Perugini et al. (2018)). It might therefore be more intuitive to perform a power analysis based on the expected pattern of means, and compute Cohen’s f based on this pattern, than to specify an effect size for the power analysis directly.

Type 1 Error Control in Exploratory ANOVA’s

In a 2x2x2 design, an ANOVA will give the test results for three main effects, three two-way interactions, and one three-way interaction. Because seven statistical tests are performed, the probability of making at least one Type 1 error in a single exploratory 2x2x2 ANOVA is $1 - (0.95)^7 = 30\%$. It is therefore important to control error rates when performing multiple comparisons (Cramer et al., 2016).

When designing a study where you will perform multiple comparisons, it is important to take into account the corrected alpha level when determining the required sample size, both for the ANOVA tests, as for the individual comparisons.

If we revisit the 2x2 between-condition ANOVA where a cross-over interaction was predicted and 40 participants per condition were collected, and apply a Holm correction for multiple comparisons.

```
design_result_holm_correction <-  
ANOVA_design(  
  design = "2b*2b",  
  n = 40,  
  mu = c(1, 0, 0, 1),  
  sd = 2,  
  labelnames = c("condition", "cheerful", "sad",
```



```

      "voice", "human", "robot")
)

```

And we perform a Monte Carlo simulation to obtain the power analysis. Please note that the `ANOVA_power` function must be used for situations where we want to simulate power when multiple comparisons corrections are applied.

```

power_result_holm_correction <- ANOVA_power(design_result_holm_correction,
      p_adjust = "holm",
      nsims = 100000)

```

We see power for the interaction is reduced from 88.31% without applying a correction, to 77.28% after correcting for multiple comparisons among the F -tests. There are three tests for a 2x2 design (two main effects, one interaction), but 6 individual comparisons between the four conditions, and thus the correction for multiple comparisons has a stronger effect on the individual comparisons. For example, the power for the t -test comparing two between-participant conditions falls from 60.17% without correcting for multiple comparisons, to 35.86% when adjusting the alpha level for the 6 possible individual comparisons between the four conditions. Please note, if a researcher would prefer to compare the estimated marginal means the R code could be simply modified (see below). The results of this code, saved under `power_result_holm_correction$emm_results`, would be approximately the same as the individual comparisons above.

```

power_result_holm_correction <- ANOVA_power(design_result_holm_correction,
      p_adjust = "holm",
      emm = TRUE,
      emm_model = "univariate",
      contrast_type = "pairwise",
      emm_comp = "condition+voice",

```

```
emm_p_adjust = "holm",  
nsims = 100000)
```

Violation of Assumptions

So far in manuscript we have shown how simulations can be useful for ANOVA power analysis purposes. However, the utility of Monte Carlo simulations extends beyond simple power analysis or general sample size planning purposes. In Monte Carlo simulations we *can* generate data that, when analyzed via ANOVA, violate the assumptions of underlying an ANOVA. In fact, it is common to see Monte Carlo simulation studies published in the statistics literature when comparing the performance of various statistical analysis techniques in certain scenarios that may affect the type 1 error rate.

As an example, we could use Superpower to estimate the impact of violating the assumption of sphericity on the type 1 error rate for a simple one-way repeated measures ANOVA. For most experiments in real life, this assumption could be violated since the correlations between levels within factors may vary. This is problematic because the univariate ANOVAs assume sphericity, which means, for all intents and purposes, that the correlations between factor-levels are equal and the standard deviations at each level are equal as well. This assumption is tenuous at best and is typically “adjusted” for by applying a sphericity correction (e.g, Greenhouse-Geisser or Huynh-Feldt adjusted output).

An inquiring researcher may wonder, how much will this violation inflate the type 1 error rate? We can simulate an example (below) wherein there are no differences between the repeated measures and varying correlations between the levels ($r = .05-.90$) of repeated measures. Again, we set this up in the ANOVA_design function.

```
design_result_s1 <- ANOVA_design(  
  "4w",  
  n = 29,  
  r = c(.05, .15, .25, .55, .65, .9),  
  sd = 1,  
  mu = c(0, 0, 0, 0)  
)
```

We then perform the Monte Carlo simulation using the `ANOVA_power` function.

```
power_result_s1 <- ANOVA_power(design_result_s1, nsims = 100000)
```

The simulated type 1 error rate for the univariate ANOVA is 7.33%. The multivariate ANOVA (MANOVA) does not assume sphericity and therefore should be robust to this pattern of correlations. We can then check the MANOVA results and see this analysis approach keeps the type 1 error rate at 4.86%. Alternatively, we could re-run the simulation with a Greenhouse-Geisser (`correction = "GG"`) or Huynh-Feldt (`correction = "HF"`) corrections for sphericity the typical univariate ANOVA to determine how well these procedures would preserve the type 1 error rate as well. After we determine which analysis approach best preserves the type 1 error rate we could then re-run the simulation with the pattern of means we want to detect and estimate the power for the given design.

P-value distributions

Statistical power is the long run probability of observing a p -value smaller than the alpha level. One intuitive way to illustrate this concept is to plot the distribution of p -values for all simulations. The `ANOVA_power` function automatically stores plots for p -value distributions for each test in the simulation. In Figure 7 we see that for the initial 2 group

between-participant design most p -values fall below the alpha level (in the figure, the vertical line as 0.05). We can produce by calling the stored results from the ANOVA_power function and selecting the plot1 object.

```
power_result$plot1
```

For the $1e+05$ simulations, 88.19% fall below the alpha level. As power is increased the p -value distribution becomes steeper, and when there is no effect the p -values are uniformly distributed.

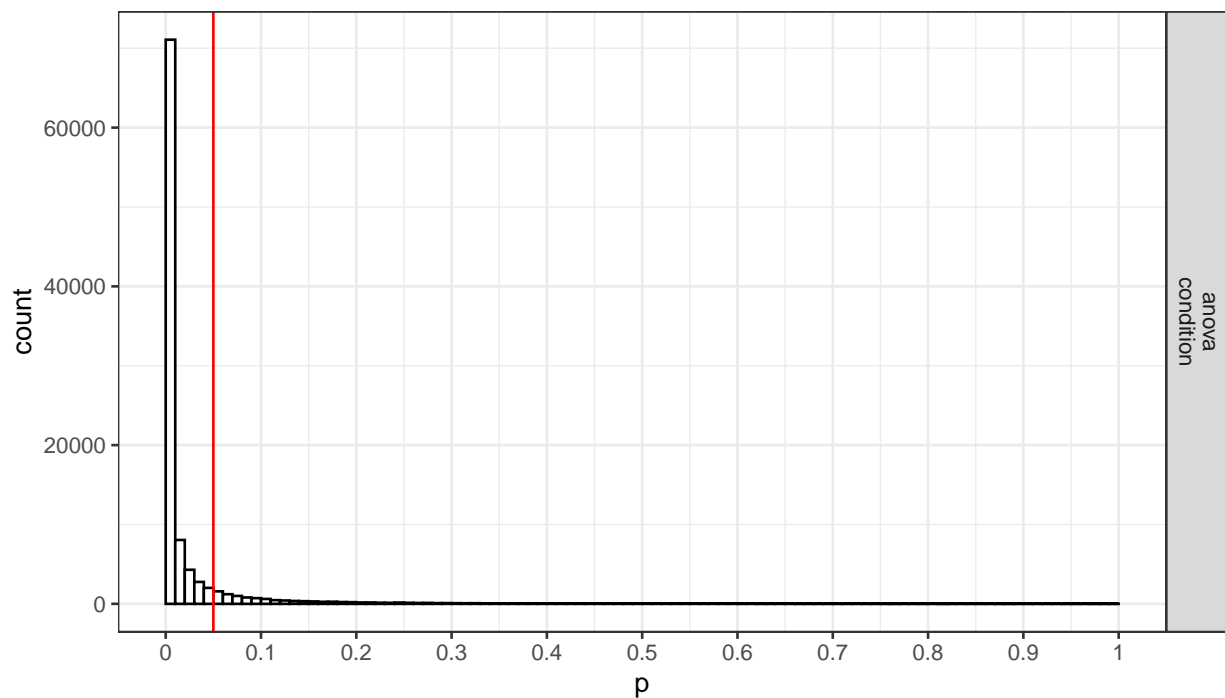


Figure 7. Distribution of p -values for the initial two-group between subjects design.

Conclusion



It is important to carefully justify the sample size when designing informative studies. Simulation based approaches can help to provide insights into the factors that affect the statistical power for factorial ANOVA designs. The effect of violating the assumptions of ANOVAs can also be evaluated using a simulation based approach. Exploring the power for

designs with specific patterns of means, standard deviations, and correlations between variables can be used to choose a design and sample size that provides the highest statistical power for future studies. The R package (<https://github.com/arcaldwell49/Superpower>), guide book (<https://arcaldwell49.github.io/SuperpowerBook>), and Shiny app (http://shiny.ieis.tue.nl/anova_power/) that accompany this paper enable researchers to perform simulations for factorial experiments of up to three factors and any number of levels, making it easy to perform simulation-based power analysis without extensive programming experience.

Author Contributions

D. Lakens and A. R. Caldwell collaboratively developed the Superpower R package. D. Lakens wrote the initial draft, and both authors revised the manuscript. A. R. Caldwell created the Shiny apps.

ORCID iD's

Daniël Lakens  <https://orcid.org/0000-0002-0247-239X> Aaron R. Caldwell 
<https://orcid.org/0000-0002-4541-6283>

Acknowledgements

Many improvements to Superpower are based on feedback from Lisa DeBruine and the `sim_design` function in her “faux” R package. The `ANOVA_exact` function was inspired by Chris Aberson’s `pwr2ppl` package. We are grateful to Jonathon Love for the development of a jamovi module based on Superpower which should be released soon.

Declaration of Conflicting Interests

The opinions or assertions contained herein are the private views of the author(s) and are not to be construed as official or reflecting the views of the Army or the Department of Defense. Any citations of commercial organizations and trade names in this report do not constitute an official Department of the Army endorsement of approval of the products or services of these organizations. No authors have any conflicts of interest to disclose.

Approved for public release; distribution is unlimited. The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Funding

This work was funded by VIDI Grant 452-17-013 from the Netherlands Organisation for Scientific Research.

Open Practices

The code to reproduce the analyses reported in this article has been made publicly available via the Open Science Framework and can be accessed at <https://osf.io/pn8mc/>.

References

- Aberson, C. L. (2019). *Applied Power Analysis for the Behavioral Sciences* (2nd ed.). New York: Routledge.
- Bretz, F., Hothorn, T., & Westfall, P. H. (2011). *Multiple comparisons using R*. Boca Raton, FL: CRC Press.

- Brysbaert, M. (2019). How many participants do we have to include in properly powered experiments? A tutorial of power analysis with some simple guidelines. *Journal of Cognition*.
- Campbell, J., & Thompson, V. A. (2012). MorePower 6.0 for anova with relational confidence intervals and bayesian analysis. *Behavior Research Methods*, 44, 1255–1265. Retrieved from <https://doi.org/10.3758/s13428-012-0186-0>
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, N.J: L. Erlbaum Associates.
- Cramer, A. O. J., van Ravenzwaaij, D., Matzke, D., Steingroever, H., Wetzels, R., Grasman, R. P. P. P., ... Wagenmakers, E.-J. (2016). Hidden multiplicity in exploratory multiway ANOVA: Prevalence and remedies. *Psychonomic Bulletin & Review*, 23(2), 640–647. doi:10.3758/s13423-015-0913-5
- Delacre, M., Lakens, D., Mora, Y., & Leys, C. (2018). Why Psychologists Should Always Report the W-test Instead of the F-Test ANOVA. *PsyArXiv*. doi:10.17605/OSF.IO/WNEZG
- 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
- Giner-Sorolla, R. (2018, January). Powering Your Interaction. *Approaching Significance*. <https://approachingblog.wordpress.com/2018/01/24/powering-your-interaction-2/>.
- Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence Testing for Psychological Research: A Tutorial. *Advances in Methods and Practices in Psychological Science*, 1(2), 259–269. doi:10.1177/2515245918770963

- Lang, A.-G. (2017). Is intermediately inspecting statistical data necessarily a bad research practice? *The Quantitative Methods for Psychology*.
- Lenth, R. (2019). Emmeans: Estimated marginal means, aka least-squares means. Retrieved from <https://CRAN.R-project.org/package=emmeans>
- Maxwell, S. E., Delaney, H. D., & Kelley, K. (2017). *Designing Experiments and Analyzing Data: A Model Comparison Perspective, Third Edition* (3rd ed.). New York, NY: Routledge.
- Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008). Sample Size Planning for Statistical Power and Accuracy in Parameter Estimation. *Annual Review of Psychology*, 59(1), 537–563. doi:10.1146/annurev.psych.59.103006.093735
- Perugini, M., Gallucci, M., & Costantini, G. (2018). A Practical Primer To Power Analysis for Simple Experimental Designs. *International Review of Social Psychology*, 31(1), 20. doi:10.5334/irsp.181
- Simonsohn, U. (2014, March). No-way Interactions. *Data Colada*. <http://datacolada.org/17>.
- Westfall, J. (2015a). PANGEA: Power analysis for general anova designs. *Unpublished Manuscript*. Available at <Http://Jakewestfall.Org/Publications/Pangea.Pdf>.
- Westfall, J. (2015b, May). Think about total N, not n per cell. *Cookie Scientist*. <http://jakewestfall.org/blog/index.php/2015/05/26/think-about-total-n-not-n-per-cell/>.
- Aberson, C. L. (2019). *Applied Power Analysis for the Behavioral Sciences* (2nd ed.). New York: Routledge.
- Bretz, F., Hothorn, T., & Westfall, P. H. (2011). *Multiple comparisons using R*. Boca Raton, FL: CRC Press.

Brysbaert, M. (2019). How many participants do we have to include in properly powered experiments? A tutorial of power analysis with some simple guidelines. *Journal of Cognition*.

Campbell, J., & Thompson, V. A. (2012). MorePower 6.0 for anova with relational confidence intervals and bayesian analysis. *Behavior Research Methods*, 44, 1255–1265. Retrieved from <https://doi.org/10.3758/s13428-012-0186-0>

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

Cramer, A. O. J., van Ravenzwaaij, D., Matzke, D., Steingroever, H., Wetzels, R., Grasman, R. P. P. P., ... Wagenmakers, E.-J. (2016). Hidden multiplicity in exploratory multiway ANOVA: Prevalence and remedies. *Psychonomic Bulletin & Review*, 23(2), 640–647. doi:10.3758/s13423-015-0913-5

Delacre, M., Lakens, D., Mora, Y., & Leys, C. (2018). Why Psychologists Should Always Report the W-test Instead of the F-Test ANOVA. *PsyArXiv*. doi:10.17605/OSF.IO/WNEZG

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

Giner-Sorolla, R. (2018, January). Powering Your Interaction. *Approaching Significance*. <https://approachingblog.wordpress.com/2018/01/24/powering-your-interaction-2/>.

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence Testing for Psychological Research: A Tutorial. *Advances in Methods and Practices in Psychological Science*, 1(2), 259–269. doi:10.1177/2515245918770963

- Lang, A.-G. (2017). Is intermediately inspecting statistical data necessarily a bad research practice? *The Quantitative Methods for Psychology*.
- Lenth, R. (2019). Emmeans: Estimated marginal means, aka least-squares means. Retrieved from <https://CRAN.R-project.org/package=emmeans>
- Maxwell, S. E., Delaney, H. D., & Kelley, K. (2017). *Designing Experiments and Analyzing Data: A Model Comparison Perspective, Third Edition* (3rd ed.). New York, NY: Routledge.
- Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008). Sample Size Planning for Statistical Power and Accuracy in Parameter Estimation. *Annual Review of Psychology*, 59(1), 537–563. doi:10.1146/annurev.psych.59.103006.093735
- Perugini, M., Gallucci, M., & Costantini, G. (2018). A Practical Primer To Power Analysis for Simple Experimental Designs. *International Review of Social Psychology*, 31(1), 20. doi:10.5334/irsp.181
- Simonsohn, U. (2014, March). No-way Interactions. *Data Colada*. <http://datacolada.org/17>.
- Westfall, J. (2015a). PANGAEA: Power analysis for general anova designs. *Unpublished Manuscript*. Available at <Http://Jakewestfall.Org/Publications/Pangea.Pdf>.
- Westfall, J. (2015b, May). Think about total N, not n per cell. *Cookie Scientist*. <http://jakewestfall.org/blog/index.php/2015/05/26/think-about-total-n-not-n-per-cell/>.