

## APPENDIX 18.1: POWER ANALYSIS FOR THE ONE-FACTOR BETWEEN-SUBJECTS ANOVA

We've noted at many points that significance tests are frequently underpowered, thus wasting time and resources. Obviously we should aim for high power when conducting ANOVAs, but we will see that there are two perspectives on this question that have to do with the two levels of analysis that we have considered. That is, we can think of power in terms of the omnibus test of

$$H_0: \mu_1 = \mu_2 = \dots = \mu_k$$

or in terms of specific contrasts, as follows:

$$H_0: \Psi_i = 0.$$

G\*Power makes it very easy to do a power analysis at both levels. Although we will deal with these two perspectives in turn, ensuring adequate *power* for specific contrasts is far more important than ensuring adequate power for the omnibus test.

### An Example

Let's think about a hypothetical graduate student in psychology who theorizes that motor skills gained playing video games transfer to other motor tasks. He would like to study gamers who have experience with action games (such as *Grand Theft Auto* and *Call of Duty*) and conceptual games (such as *Portal* and *Tetris*). The action games require fast and immediate reactions to unpredictable events, whereas the conceptual games require fast reactions following some deliberation. This student predicts that gamers who have more experience with action games will be able to respond more quickly to unpredictable stimuli than those who have more experience with conceptual games. He also predicts that gamers as a group will have faster reaction times than those who don't play video games.

The student's plan for this quasi-experiment is to recruit three groups of participants. One group will have extensive experience with action games and another will have extensive experience with conceptual games. The third will be a control group of individuals with minimal experience with either type of game.

To assess his prediction about motor responses, the student will have all participants perform a simple reaction-time task that requires them to press a button as quickly as possible when a dim light appears at an unpredictable location on a computer screen. Each participant will go through 256 trials, and the dependent

variable is the average time it takes to respond to each of the 256 stimuli. The student plans to run a one-way between-subjects ANOVA with three levels. He predicts that there will be a statistically significant omnibus effect. He also specifies the following alternative hypotheses:

$$H_1: \Psi_1 = \mu_{\text{control}} - \frac{\mu_{\text{action}} + \mu_{\text{conceptual}}}{2} > 0$$

$$H_1: \Psi_2 = \mu_{\text{conceptual}} - \mu_{\text{action}} > 0.$$

Our student really should not test both the omnibus effect and the two planned contrasts. However, for the purposes of this exercise, we will use the same example to illustrate these two levels of analysis.

### Power for the Omnibus Effect

All power analyses require some measure of effect size. For the omnibus ANOVA, we use the quantity

$$f = \frac{\sigma_m}{\sigma}. \quad (18.A1.1)$$

In equation 18.A1.1,  $\sigma$  is the within-groups standard deviation and  $\sigma_m$  is the standard deviation of  $k$  hypothetical population means. To compute  $f$  requires an actual hypothesis about what the population means are. We can denote this hypothesis this way:

$$H_1: \mu_1, \mu_2, \dots, \mu_k.$$

This means we have to specify values for the  $k$  population means, which is a theoretical question, not a statistical question.

Whatever population means we specify, the variability among them is defined as follows:

$$\sigma_m = \sqrt{\sum_{i=1}^k (\mu_i - \mu)^2 / k}, \quad (18.A1.2)$$

where

$$\mu = \sum_{i=1}^k \mu_i / k.$$

Equation 18.A1.1 shows that the effect size ( $f$ ) is a ratio of two standard deviations, so the absolute values of  $\sigma$  and  $\sigma_m$  are not important. Therefore, a simple way to compute  $f$  is to set  $\sigma = 1$  and let the differences between the  $k$  means ( $\mu_1, \mu_2, \dots, \mu_k$ ) represent standard deviation

units. In this way, we are specifying  $(k^2 - k/2)$  contrasts simultaneously.

Our hypothetical graduate student expects the difference between the two gamer groups to be on the order of  $\frac{1}{2}$  a standard deviation, which Cohen’s model would classify as a medium effect size. The student also assumes that the control participants will have reaction times that are about a  $\frac{1}{2}$  standard deviation greater than those in the conceptual games group. Therefore, to conduct his power analysis, he assumes the following:

$$\begin{aligned} \mu_{\text{control}} &= 1.0, \\ \mu_{\text{conceptual}} &= 0.5, \text{ and} \\ \mu_{\text{action}} &= 0.0. \end{aligned}$$

(Please note that any three numbers separated by  $\frac{1}{2}$  would serve equally well.) The mean of these three means is  $\mu = \sum_{i=1}^k \mu_i / k = .5$ . When we calculate  $\sigma_m$  from these numbers, we find

$$\sigma_m = \sqrt{\sum_{i=1}^k (\mu_i - \mu)^2 / k} = \sqrt{(0.25 + 0.0 + .025) / 3} = .41.$$

Therefore,  $f = \sigma_m / \sigma = 0.41 / 1 = 0.41$ .

Figure 18.A1.1 shows how to do a power analysis for the omnibus effect in G\*Power. The drop-down list in Test family is set to F tests and the Statistical test is set to ANOVA: Fixed effects, omnibus, one-way. In the Input parameters region, we can enter  $f$ ,  $\alpha$ , power, and number of groups. Our effect size  $f$  was calculated to be 0.41, and we will set  $\alpha = .05$ , power = .8, and number of groups to 3. When these quantities have been entered in the appropriate boxes in the Input parameters section, pushing the **Calculate** button produces the total sample size required to achieve the requested power. Figure 18.A1.1 shows that to achieve power = .8 when  $\alpha = .05$  requires 63 participants, or  $63/3 = 21$  per group.

### Power Analysis for Planned Contrasts

A power analysis for contrasts is conducted in much the same way as for the omnibus test. We will continue to set  $\sigma = 1$ , so that differences between our hypothesized means can be seen as effect sizes. To compute  $\sigma_m$  for a contrast, we first put our predicted means in vector form. For our student’s hypothesis, we would have

$$h = \{\mu_{\text{control}}, \mu_{\text{conceptual}}, \mu_{\text{action}}\} = \{1.0, 0.5, 0.0\}.$$

We also specify a contrast in a similar way:

$$w_2 = \{0, 1, -1\}.$$

With this information, we compute  $\sigma_m$  as follows:

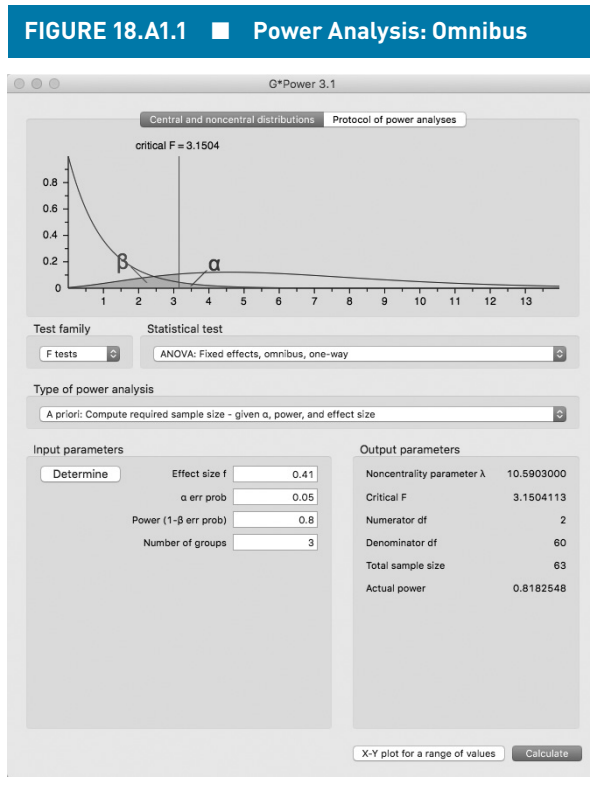
$$\sigma_m = \frac{|c_2|}{\sqrt{k \sum_{i=1}^k w_{2i}^2}}. \tag{18.A1.3}$$

The following steps provide a very general method to determine the sample size required to achieve any desired power for any possible contrast or trend. Whether we are considering a trend or a contrast, the basic requirements are weights ( $w$ ) and hypothesized means ( $h$ ). So, let’s work through the steps required to compute  $f = \sigma_m / \sigma$  for the contrast that compares the response times of the two gamer groups.

**Step 1.** Specify the contrast of interest as a set of weights. Because we are interested in the difference between the means of the two gamer groups, our weights are

$$w_2 = \{0, 1, -1\}.$$

**Step 2.** Specify the predicted means, where the separations between means are assumed to be in



units of standard deviation. *This is the hard part* because it requires specifying our expectations, or what we judge to be important; this is not a statistical issue. For this illustration, we have stated that the means will be

$$h = \{1.0, 0.5, 0.0\}.$$

**Step 3.** Compute the contrast, which is the numerator of equation 18.A1.3. We do this using equation 18.6:

$$\begin{aligned} c_2 &= \sum_{i=1}^k w_{2i} h_i = \sum \{0, 1, -1\} * \{1.0, 0.5, 0.0\} \\ &= 0.0 + 0.5 - 0.0 = 0.5. \end{aligned}$$

**Step 4.** Compute the denominator of equation 18.A1.3. We do this as follows:

$$\sqrt{k \sum_{i=1}^k w_{2i}^2} = \sqrt{3 * \{0^2 + 1^2 + 1^2\}} = \sqrt{3 * 2} = 2.449.$$

**Step 5.** Compute  $\sigma_m$  using equation 18.A1.3. Substituting in the quantities we've computed, we obtain

$$\sigma_m = \frac{|c_2|}{\sqrt{k \sum_{i=1}^k w_i^2}} = \frac{0.5}{\sqrt{6}} = 0.2041.$$

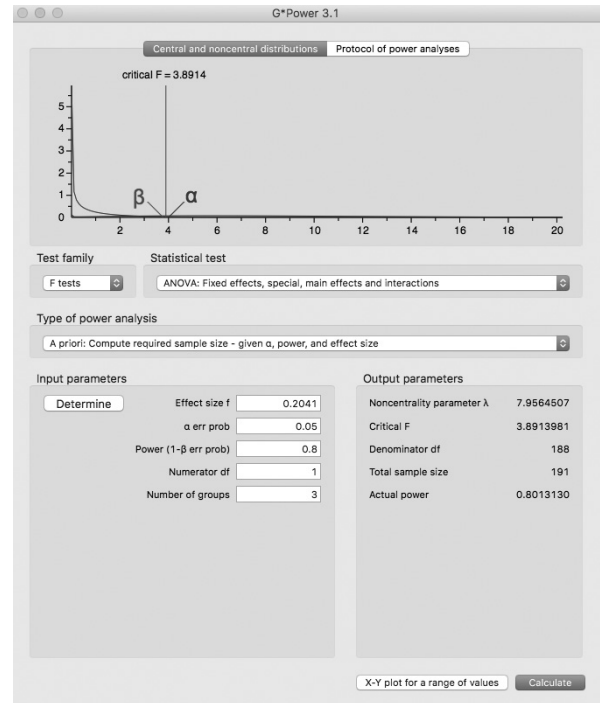
**Step 6.** Compute  $f$  using equation 18.A1.2. We assume  $\sigma = 1$ , so  $f = \sigma_m = 0.2041$ .

These steps can be implemented very easily in Excel.

The final step in this process is to use G\*Power to compute the number of participants required to achieve power = .8. Figure 18.A1.2 shows how this is done. The main difference between Figures 18.A1.1 and 18.A1.2 is that a different selection has been made in the Statistical test drop-down list. The selection made in Figure 18.A1.2 is ANOVA: Fixed effects, special, main effects and interaction. This produces one additional box in the Input parameters region. The new option is Numerator df. Remember that all contrasts are essentially  $t$ -tests and thus use one degree of freedom. Therefore, we enter a 1 in this new box. We just computed  $f = 0.2041$  for the effect size associated with our contrast, so we enter this in the first box, Effect size  $f$ . The quantities  $\alpha = .05$ , power = .8, and number of groups = 3 are the same as before.

When these quantities have been entered in the appropriate boxes in the Input parameters section, clicking the **Calculate** button produces the total sample size required to achieve the requested power. The

FIGURE 18.A1.2 ■ Power Analysis: Contrasts



Output parameters region in Figure 18.A1.2 shows that to achieve power = .8 when  $\alpha = .05$  requires 191 participants, or  $191/3 \approx 64$  per group.

It is important to note that more participants are required to achieve power = .8 for the contrast ( $n \approx 64$  per group) than for the omnibus test ( $n = 21$  per group). Because contrasts are more informative than the omnibus effect, you should always make sure that you have enough participants to achieve your desired power (typically at least .8) for all contrasts. This means computing  $f = \sigma_m$  for all contrasts and then determining the sample size required to detect the smallest of these.

### Power Planning From Published Results

In the preceding discussion, our graduate student generated a set of hypothetical standardized means based on his judgment about what would be interesting. We've seen before, however, that predictions or expectations about population means can also come from published results. Imagine someone had run the same quasi-experiment that our student wants to run and used

12 subjects in each of the three conditions. The results are shown in Table 18.A1.1.

Let's say that the authors of this study submitted these results to a one-factor between-subjects ANOVA and found the omnibus  $F$  to be statistically significant,  $F_{2,33} = 6.88, p = .003, \hat{\eta}^2 = .29$ . They then used Bonferroni-corrected  $t$ -tests to answer several questions using the contrast weights shown in columns  $w_1$  and  $w_2$ . They found, as expected, that the control group differed from the mean of the two gaming groups,  $c_1 = 80, t(33) = 2.93, p = .01$ . They then asked whether the mean of the action group differed from the mean of the conceptual group. This contrast was in the predicted direction,  $c_2 = 28$ , meaning that those who had experience with conceptual games responded more slowly than those who had experience with action games. However, this difference was not statistically significant,  $t(33) = 0.90, p = .74$ .

**TABLE 18.A1.1 ■ Data for Three Groups<sup>a</sup>**

Condition	$i$	$m$	$s$	$s^2$	$w_1$	$w_2$
Control	1	348	48.64	2366	1.00	0.00
Conceptual	2	282	84.25	7098	-0.50	1.00
Action	3	254	88.32	7800	-0.50	-1.00
					$c_1$	$c_2$
					80.00	28.00

<sup>a</sup> $n = 12, s_{pooled}^2 = 5754.67, s_{pooled} = 75.86$ .

Because we're not really interested in the omnibus effect, we will focus on using the data from Table 18.A1.1 to do a power analysis for contrast 2, which compared the means of the two gamer groups. To conduct a power analysis, we need values for  $\sigma$  and  $\sigma_m$  so that we can calculate  $f$ .

First, we can compute  $\sigma$  from the data as follows:

$$\sigma = \sqrt{\sum_{i=1}^k s_i^2 / k} \tag{18.A1.4}$$

This quantity is a version of  $s_{pooled}$ , which is the square root of the average sample variance. Table 18.A1.1 shows that  $s_{pooled} = 75.86$ ; so for our purposes,  $\sigma = 75.86$ . Next,  $\sigma_m$  can be computed exactly as before using equation 18.A1.4. Therefore,

$$\sigma_m = \frac{|c|}{\sqrt{k \sum_{i=1}^k w_i^2}} = \frac{28}{\sqrt{3(0^2 + 1^2 + (-1)^2)}} = 11.43.$$

With  $\sigma_m = 11.43$  and  $\sigma = 75.86$ , we determine that,

$$f = \sigma_m / \sigma = 11.43 / 75.86 = 0.15.$$

We use G\*Power exactly as before and find that we would need 351 participants altogether, or 117 in each group, to achieve power = .8 when  $\alpha = .05$ . Yikes!

### Conclusion

Our graduate student was right to conduct a prospective power analysis before undertaking his quasi-experiment. Now that he knows something about how big his samples need to be, what action should he take? It seems to me that he shouldn't bother with this quasi-experiment. In the absence of any published data, the student thought that a difference between the action and conceptual gaming groups of  $\frac{1}{2}$  a standard deviation ( $\delta = 0.5$ ) would be interesting. A quasi-experiment designed to detect this difference with power = .8 would require 64 participants per condition. When considering the published data on this effect, the observed effect size was even smaller,  $d = (282 - 254) / 75.86 = 0.35$ , and detecting an effect of this size would require 117 participants per condition.

A quasi-experiment with three conditions having 64 to 117 participants per condition would be enormous. Finding 64 participants for each of the two gamer conditions would be a huge and expensive undertaking in itself. Running each participant through the quasi-experiment would require at least half an hour, so this means about  $3 * 64 / 2 = 96$  hours of testing. Things get far worse when we consider the possibility of having to run 117 participants per condition.

The expected return from this quasi-experiment is an answer to the question of whether gaming experience produces a statistically significant transfer of motor skills from gaming to other tasks. It is not clear what value such a result would have outside a theoretical context, but the student might be able to make a compelling case for the general importance of the result. Without this kind of context, however, one must question the value of investing in the quasi-experiment.

Finally, the proposed study is, after all, a quasi-experiment. This means that one can't infer causality from observed differences between the groups. Those drawn to action games may find them enjoyable precisely because they have naturally fast reactions and these lead to success in the games. Those not drawn to gaming may not find them enjoyable precisely because they are not naturally quick and so such games are not rewarding. On the whole, then, pursuing this study seems ill advised.

We can be grateful for freely available programs like G\*Power that allow us to plan our experiments easily. The importance of power analysis cannot be overstated. Without it, we run on hunches, tradition, and little more. Underpowered experiments waste (i) experimenters'

time, (ii) taxpayer-supported grant money, (iii) participants' time, and (iv) animals' lives (in the case of animal research). These days there can be no excuse for running a significance test without having conducted a prospective power analysis.

## LEARNING CHECK 1

1. A researcher would like to know how motor activity in rats relates to how much cocaine has been injected into their bloodstreams. Her dependent variable is the number of minutes per hour rats spend on a running wheel following drug injection. There will be five experimental conditions. The rats in the five conditions will receive cocaine injections of either 0, 25, 50, 75, or 100  $\mu\text{g}$ /injection prior to testing. The researcher would

like to assess the omnibus effect as well as the linear and quadratic trends in the data. She expects the means to be proportional to the following:  $h = \{0.0, 0.3, 0.6, 0.9, 0.5\}$ . That is, she assumes  $\sigma = 1$ . Use G\*Power to determine the samples sizes required to detect (a) the omnibus effect, (b) the linear trend, and (c) the quadratic trend. Assume in all cases that power = .8 and  $\alpha = .05$ . Note that the trend coefficients can be obtained from Table 18.A1.1.

### Answers

1. (a) For the omnibus effect,  $f = .300$ . The test requires a total of 140 animals or  $140/5 = 28$  per condition. (b) For the linear trend,  $f = 0.226$ . The test requires a

total of 156 animals or  $156/5 \approx 32$  per condition. (c) For the quadratic trend,  $f = 0.167$ . The test requires a total of 284 animals or  $284/5 \approx 57$  per condition.