Experimental Design: Model Fits, Power, and Complex Designs

7

Contents

What You Will Learn in This Chapter

An ANOVA is one option to cope with the multiple testing problem. A much simpler way to cope with multiple testing is to avoid it by clever experimental design. Even if you need to measure many variables, there is no need to subject all of them to a statistical test. As we will show in this chapter, by collapsing many data into one meaningful variable or simply by omitting data, you may increase your statistical power. Simple and simplified designs are also easier to interpret, which can be a problem in many complex designs. In this chapter, we also show how to compute the power of an experiment, which is for example important to determine the sample size of your experiment.

7.1 Model Fits

When comparing two means, a *t*-test has a high power and is straightforward to interpret. Experiments with more group comparisons suffer from the multiple testing problem. The more comparisons we compute, i.e., the more groups or levels there are, the lower is the power. Experiments with more groups are also more complex to analyse because interactions can occur, which are not present in simple *t*-tests (Chap. 6). Hence,

M. H. Herzog et al., *Understanding Statistics and Experimental Design*,
Learning Materials in Biosciences, https://doi.org/10.1007/978-3-030-03499-3_7

simple experimental designs are usually preferred. However, complexity is sometimes unavoidable. A classic example is a learning experiment, where performance needs to be measured at many time points. For example, participants train on a simple visual task for 10 blocks containing 80 trials each. For each block, we determine the percentage of correct responses (Fig. 7.1a) and look for increases in performance across blocks. How can we quantify the learning success and compute statistics? The null hypothesis is that no learning occurs, i.e., performance is identical for all 10 blocks. Intuitively, one might think of using a repeated measures ANOVA with 10 levels, one for each block. However, this is not a good idea because, first, ANOVAs are about nominal variables, i.e., the order of the blocks plays no role. Second, were performance increasing for the first five blocks and then decreasing, the ANOVA would indicate a significant result when performance for block 5 is significantly different from block 1. However, a result like this is not about learning but a strange concatenation of learning and unlearning. Third, we would lose quite some power. What to do? Here, we show it is by no means necessary to subject all data to statistical analysis.

As shown in Fig. 7.1b for the learning experiment, one approach is to discard all blocks except for the first and the last one (the intermediate blocks are relevant to the experiment because they promote learning, but they are not relevant to the statistical analysis). The null hypothesis is that performance in these two blocks does not differ. We can use a repeated measures t-test to test this hypothesis. However, learning data are often noisy and thus we are losing power with this procedure. To obtain less noisy data, we may average the first and last two blocks and subject the two averages to a repeated measures t-test (Fig. 7.1c).

In both cases, we are discarding a large amount of data and thus do not take full advantage of our data. We might do better by fitting a model to the data. We may know, for example, from previous experiments that learning is reflected by a linear increase in performance, which we can model by the equation $mx + b$, where m is the slope of the learning curve, b is the y-intercept, and x is the block number. We can use a computer program to compute the optimal parameters for m and b, for each observer individually. Since we are only interested in the slope, we can discard b. Our null hypothesis is: $m = 0$. Hence, for each observer we obtain one m-value. If 12 observers joined the experiment, we compute a one-sample t-test with these 12 values of m and see whether they are significantly different from 0.

There is great flexibility in this approach. For example, if learning is not linear but follows an exponential rather than a linear function, then we can fit an exponential function, which also contains a "slope" parameter. When we are interested in cyclic processes, such as changes in temperature across a day or the numbers of insects across a year, we can fit a sine function. In general, we can fit any type of function to our data and extract one or a few parameters. We thus take full advantage of the data and do not lose power. It is the choice of the experimenter how to proceed. However, the experimenter must make the choice before the experiment is conducted. One cannot decide after having looked at the data and then try many possibilities until finding a significant result (see Sect. 11.3.5).

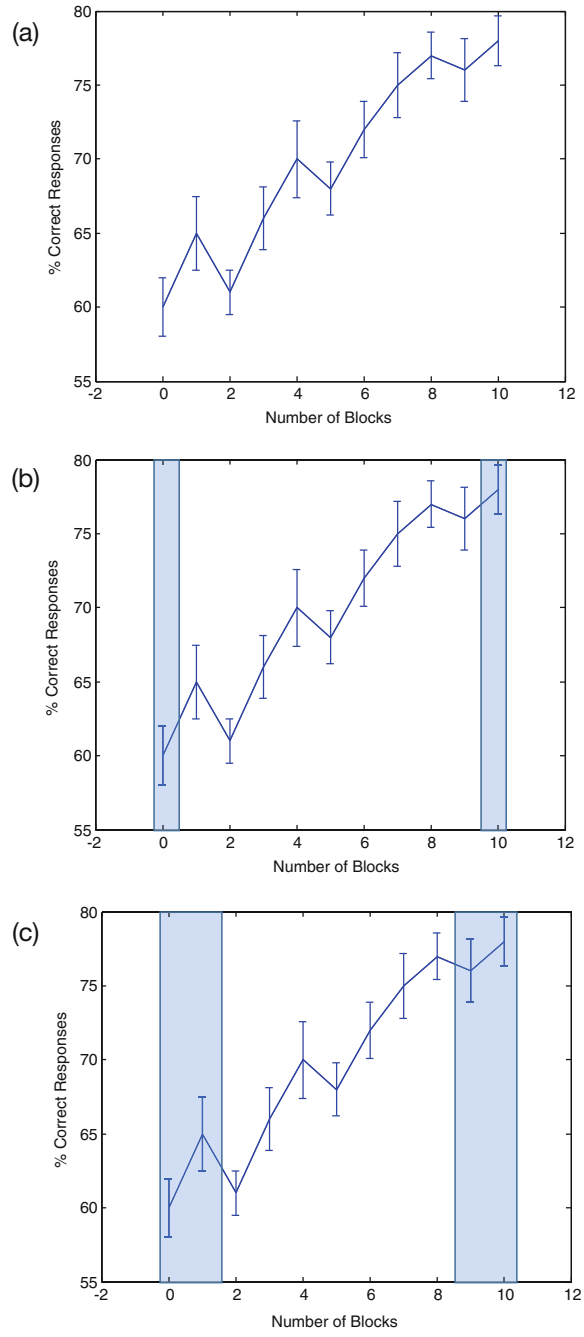

Fig. 7.1 Analyzing learning data. (**a**) Performance improves with number of blocks. (**b**) A statistical analysis might just compare the first and last blocks. (**c**) Alternatively, the analysis might average the first two and last two blocks and then compare the averages

The above example shows how to simplify statistics by reducing the number of variables. As shown, there is no need to subject all your data in its original form to statistical analysis. There are no general rules on how to simplify your analysis because each experiment is different. However, it is always a good idea to think about what is the main question your experiment is aimed to answer. Then, you decide what variables address the question best and how you can compute statistics. The simpler the design and the fewer variables, the better.

7.2 Power and Sample Size

7.2.1 Optimizing the Design

It often takes a lot of effort and resources to run an experiment. Thus, it is usually worthwhile to estimate whether the experiment is likely to succeed and to identify sample sizes that provide a high probability of success (if there is actually an effect to detect). Success generally means producing a large t-value and obtaining a significant result. We can do this in a couple of ways.

First, try to increase the population effect size $\delta = \frac{\mu_1 - \mu_2}{\sigma}$. You can do this by considering situations where you anticipate the difference between population means to be large. For example, you may try to find the optimal stimuli for a visual experiment or the most discriminative tests for a clinical study.

In addition, try to reduce σ. It is the ratio of the population mean differences and the standard deviation that determines δ, and thus t and p. You may try to make your measuring devices less noisy, for example, by calibrating every day. You may try to homogenize the sample, for example, by testing patients at the same time every day, making their coffee consumption comparable, using the same experimenter etc. You may think about excluding certain patients, for example, by imposing age limits to not confuse deficits of a disease with age effects. However, such stratifications limit the generality of your research (see Chap. 3, Implications 4). There are many ways to reduce σ and it is always a good idea to think about it.

Second, increase the sample size, n. Even if δ happens to be small, a large enough sample will produce a large t-value. With a large enough sample size it will be possible to discriminate even small differences between means (signal-and-noise) from a situation where there is actually no difference between means (noise-alone). Note, for this approach to be meaningful, you have to be confident that a small effect size matters. There is no point in running a large sample study to detect a trivially small effect (see Chap. 3, Implications 1 and 2).

7.2.2 Computing Power

Even when $\delta \neq 0$, experiments do not always produce significant results because of undersampling (Chap. 3). Here, we show how likely it is that for a given $\delta \neq 0$ and a given sample size n a significant result occurs. Vice versa, we show how large n needs to be to produce a significant result with a certain probability.

We estimate an experiment's success probability by computing power. Power is the Hit rate. It is the probability of selecting a random sample that allows you to correctly reject the null hypothesis. It supposes that the null hypothesis is false, meaning that there is a non-zero effect. As noted in Chap. 3, computing power requires a specific population standardized effect size. Where this specific population effect size comes from is situation-specific. Sometimes it can be estimated from other studies that have previously investigated the same (or a similar) phenomenon. Sometimes it can be derived from computational models that predict performance for a novel situation. Instead of predicting an effect size, it is sometimes worthwhile to identify a value that is deemed to be interesting or of practical importance.

Once a population effect size is specified, we turn to computer programs to actually calculate power (there is no simple formula). Figure 7.2 shows the output of a free program called G*Power. Here, we selected t-test from the *Test family* and a *Statistical test* of a difference between two independent samples for means. For *Type of power analysis* we selected "Post hoc." Under *Input parameters* we selected for a two-tailed test, entered an estimated population effect size $d = 0.55$, chose our Type I error rate to be $\alpha = 0.05$, and entered planned sample sizes of $n_1 = n_2 = 40$. The program provides graphs at the top and *Output parameters* on the bottom right. The graphs sketch the sampling distributions (see Fig. 3.7) that should be produced by the null (red curve) and the specific alternative hypothesis (blue curve). The shaded blue area is labeled β to indicate the Type II error rate. This is the probability for a non-significant result if $\delta = 0.55$. Power is the complement of the Type II error rate. As indicated, for the provided input parameters, the computed power is 0.68. This means that there is a probability of 0.68 that under the specified conditions you will obtain a significant result.

Suppose that we were unsatisfied with the 0.68 probability and wanted to identify sample sizes that would have a 90% chance of rejecting the null hypothesis. From the *Type of power analysis* menu, we select "A priori" and in the revised *Input parameters* we change the Power value from 0.68 to 0.9. Figure 7.3 shows the program output for the new situation. In the *Output parameters* panel, we see that the sample sizes needed to have a power of 0.9 for a two-tailed, two-sample t-test, when the population effect size is $\delta = 0.55$ are $n_1 = n_2 = 71$.

In general, for a given population effect size, one can identify the smallest sample sizes so that an experiment has the specified power. Calculating such sample sizes is an important part of experimental design. It usually makes little sense to run an experiment without knowing that it has a reasonable probability of success, i.e., reasonable power. Unfortunately, many scientists run experiments without doing a power analysis because

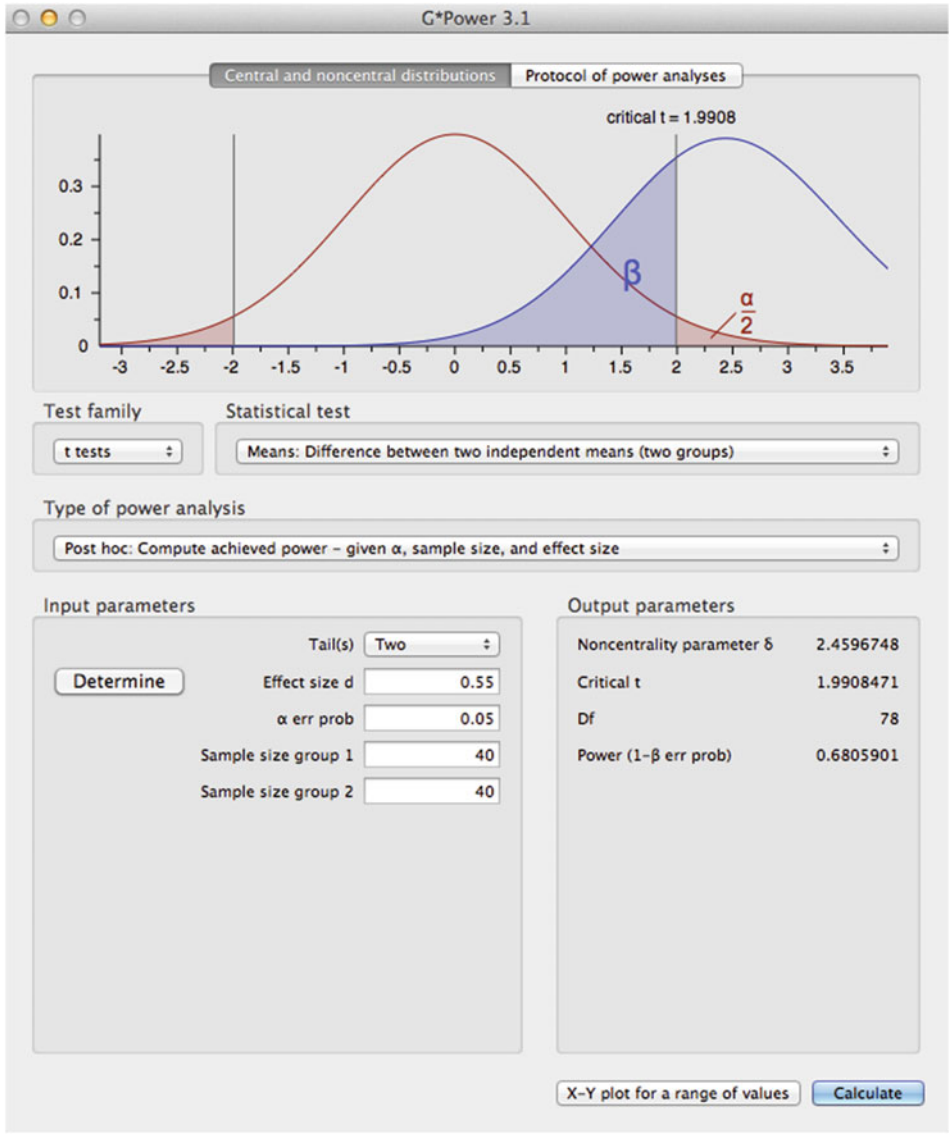

Fig. 7.2 Output from the G*Power program to compute power for a t-test with specified sample sizes. In this case, the effect size (0.55) and the sample sizes ($n_1 = n_2 = 40$) are known and we are searching for power, i.e., how likely it is that we obtain a significant result with this effect and sample size for an independent t-test and $\alpha = 0.05$. The output parameters are the noncentrality parameter δ, which is not the same as the population effect size and we ignore it here, the critical t-value, the degrees of freedom Df and, most importantly, the power

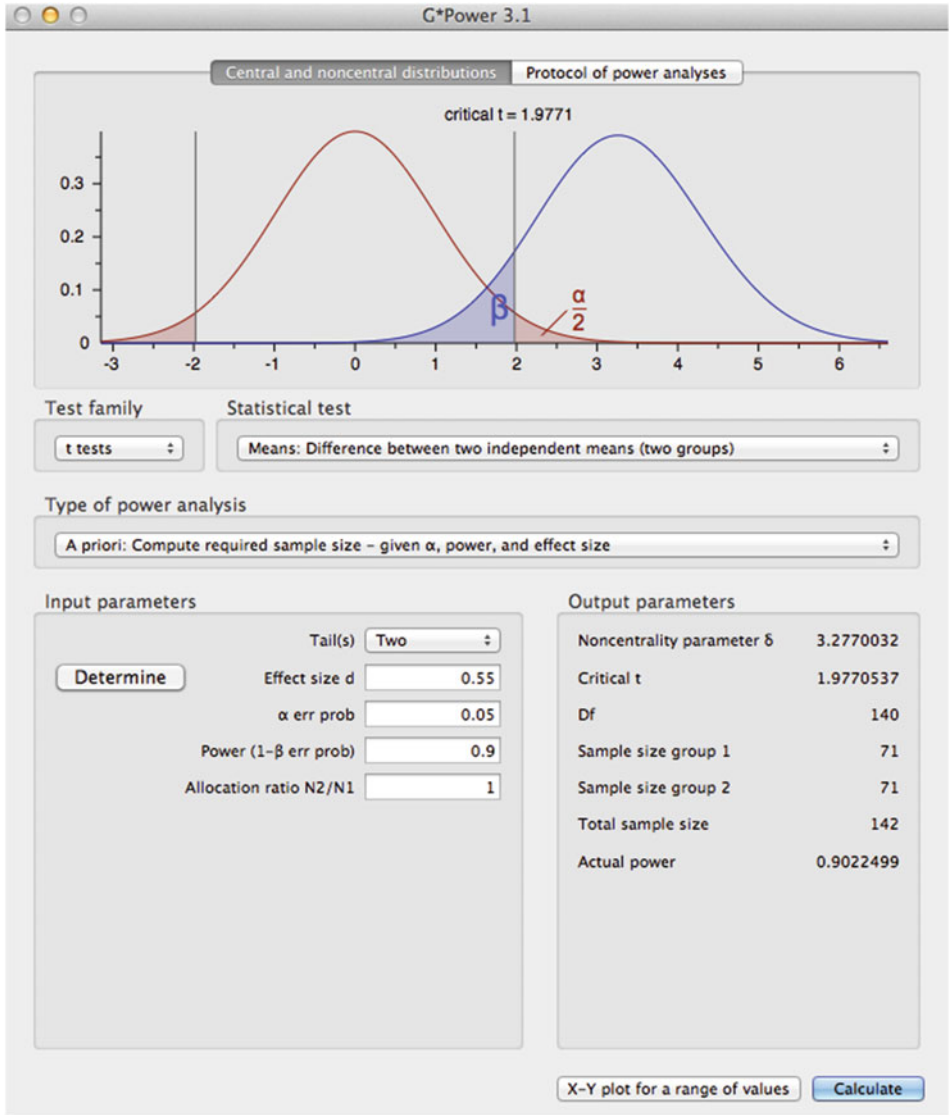

Fig. 7.3 Output from the G*Power program to compute the necessary sample sizes so that an experiment has a *t*-test with at least 90% power. In this case, the effect size (0.55) is known or desired and we are searching for the sample size to obtain a significant result with a probability of 0.9

they do not have a particular effect size in mind. Such investigations might turn out to be valuable, but whether they work or not is largely a matter of luck. If you cannot perform a meaningful power analysis (with a justified effect size), then the best you can do is "hope" that your experiment produces a significant outcome. If the experiment fails to produce a significant outcome, you can hardly be disappointed because you never really had any (quantitative) reason to expect that your sample was large enough to show an effect. Many times scientists are doing exploratory work even when they think they are doing confirmatory work. Confirmatory work almost always is built on knowledge about an effect size that can be used to design an experiment with high power.

7.3 Power Challenges for Complex Designs

Ideally, a power analysis is done before gathering any data; this is called a priori power. However, it is also possible to estimate power in a post hoc fashion by using the sample sizes and the estimated effect size from the data. For simple cases (e.g., a two-sample t-test) the post-hoc power analysis does not tell you anything beyond the test for significance. If you use G*Power to calculate power for different combinations of t and sample sizes you will discover that, if your t-test gives you $p > 0.05$ then your power calculation will be less than 0.5. Likewise, if your t-test gives you $p < 0.05$, then your power calculation will be greater than 0.5. If your t-test gives you $p = 0.05$, then your power calculation will give approximately 0.5. Here, we show that post hoc power calculations can be more useful for complicated statistical analyses that involve multiple tests on a set of data.

We saw above how to use G*Power to compute power for simple experimental designs. This program, and similar alternatives, tends to focus on just one statistical test at a time. In practice, scientists often use a combination of statistical tests to argue for a theoretical interpretation. Estimating power for a combination of statistical tests often requires generating simulated data sets that correspond to the experiment's design and sample sizes. This simulated data is then analyzed in the same way that the experimental data will be analyzed. By repeating this process thousands of times, one can simply count how often the full set of statistical outcomes matches the outcomes needed to support a theoretical claim. This simulation approach allows a researcher to consider "success probability", which generalizes the concept of power.

We will see that complex experimental designs with multiple tests can struggle to have high power. Even if individual tests have reasonable power, it can be the case that the full set of tests has low power.

To demonstrate this generalization, it may be helpful to consider a concrete example. The example is purposely complicated because the complications highlight important characteristics of power analyses. A prominent study published in 2017 reported empirical evidence that performance on a memory task was related to breathing through the nose. The motivation for the study was that nasal breathing can entrain the hippocampus of

the brain, which is related to memory processing. In contrast, oral breathing does not entrain the hippocampus and so should not influence memory performance. Subjects were asked to breath either orally (through the mouth) or nasally (through the nose) while viewing pictures during a memory encoding phase, and then during a retrieval test subjects identified pictures they had seen before. During both the encoding and retrieval phases the pictures were presented at random times so that sometimes the picture was presented while the subject was inhaling and sometimes the picture was presented while the subject was exhaling. The main conclusion was that identification accuracy was better for pictures that were presented to nasal breathers during inspiration (breathing in). This was true for encoding pictures and for retrieving pictures. In contrast, oral breathers showed no significant effect of inward versus outward breathing.

The study and its analysis is rather complicated, so it is useful to characterize all the hypothesis tests. For convenience, we also list the relevant statistics from the study. All tests compared memory performance of subjects.

1. Nasal breathers ($n_1 = 11$) showed a significant ($F(1, 10) = 6.18$, $p = 0.03$) main effect of breathing phase (inhale or exhale) on memory performance.
2. Nasal breathers showed enhanced memory for pictures that had been retrieved while inhaling compared to pictures that had been retrieved while exhaling ($t(10) = 2.85$, $p = 0.017$).
3. Oral breathers ($n_2 = 11$) did not show enhanced memory for pictures that had been retrieved while inhaling compared to pictures that had been retrieved while exhaling ($t(10) = -1.07$, $p = 0.31$).
4. There was no significant difference between nasal and oral breathers overall ($F(1, 20) = 1.15$, $p = 0.29$).
5. There was a significant interaction of breathing phase (inhale and exhale) with breath route (nasal and oral) when pictures were labeled by how they were encoded (inhale or exhale) ($F(1, 20) = 4.51$, $p = 0.046$).
6. There was also a significant interaction of breathing phase (inhale or exhale) with breath route (nasal and oral) when pictures were labeled by how they were retrieved (inhale or exhale) ($F(1, 20) = 7.06$, $p = 0.015$).

If you are confused, then take comfort in knowing that you are not alone. This study and its analysis is very complicated, which makes it difficult for a reader to connect the reported statistics to the theoretical conclusions. Moreover, some of the comparisons seem inappropriate. For example, the authors of the study used tests 2 and 3 to demonstrate a difference of significance for the nasal and oral breathers (comparing retrieval during inhaling versus exhaling). We noted in Chap. 3 (Implication 3b) that a difference of significance is not the same as a significant difference. Likewise, the authors of the study took the null result in test 4 as indicating "no difference" in performance of nasal and oral breathers overall. We saw in Chap. 3 (Implication 3a) that absence of proof is not proof of absence.

Table 7.1 Estimated success probabilities for the findings of a study relating memory performance to breathing orally or nasally

Test	Probability of success
Nasal: main effect of breath phase	0.690
Nasal retrieval: effect of breath phase	0.655
Oral retrieval: null effect of breath phase	0.809
Nasal vs. oral breathers: null main effect	0.820
During encoding: interaction for breath phase and route	0.604
During retrieval: interaction for breath phase and route	0.708
All tests	0.216

For the moment let us set aside our concerns about the appropriateness of the tests. Success for this study required four significant outcomes and two non-significant outcomes. If any of these outcomes were unsuccessful, it would call into doubt some of the conclusions made by the authors. As it turns out, the data supported every one of these necessary outcomes. We will show that with so many outcomes that must be satisfied by a single data set, such full success should be rare even if the effects are real and close to the values estimated by the experimental data. To estimate the probability of such a level of success, a statistical software program, R, was used to generate 100,000 simulated experiments with the reported sample sizes, means, standard deviations, and correlations (for within-subject aspects of the experiment). Table 7.1 shows how often each test produced the desired outcome. The success probability for any given hypothesis test varies between 0.60 and 0.82. For each significant test, the success probability of that specific test corresponds to power. For tests 3 and 4 listed above, a successful outcome was a non-significant result, and the table lists the probability of *not* rejecting the null hypothesis.

However, the probability of *every* test being successful for a given simulation is much lower than the probability for an individual test being successful because the data needs to have just the right properties to deliver a significant result for certain tests and to deliver a non-significant result for other tests. Based on the simulations, the joint probability that all of the tests would be successful in a single experiment is only 0.216. This low probability suggests that, simply due to random sampling, a direct replication of the study with similar sample sizes would have a rather small probability of producing the same pattern of outcomes.

A researcher replicating this study would want to pick sample sizes that give a high probability of success. Larger samples increase the power of a test, so that a study with just one test is more likely to find an effect if it exists. However, when the theoretical claims are based on both significant and non-significant tests, there are limits to the maximum probability of success because with large sample sizes small effects generate significant results (even for the studies where the authors hope for a null finding). The limit for this study can be investigated with additional simulated experiments that vary the

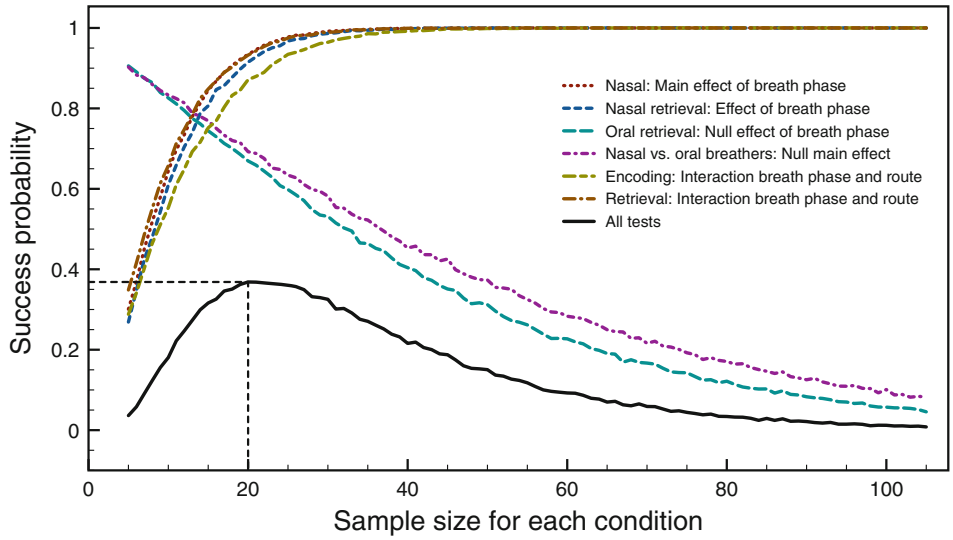

Fig. 7.4 Each colored line shows the estimated probability of success as a function of sample size for a test from a study investigating the effects of breathing on memory performance. The solid black curve shows the estimated success probability for all of the tests. The dashed black lines mark the sample size with the highest possible success probability for all of the tests combined. Each value is based on 10,000 simulated experiments

sample size for each condition. The colored lines in Fig. 7.4 plot the estimated probability of success for each of the six tests as a function of sample size (assuming the same sample size for each condition). For the four tests where success corresponds to producing a significant result, the probability of success increases with sample size and converges on the maximum value of 1 at around a sample size of 40. For the two tests where success corresponds to producing a non-significant result, the probability of success decreases with sample size (because some random samples show significant differences). The dashed black lines in Fig. 7.4 show that considering all six tests together (the black line), the maximum possible success probability is 0.37 with around $n_1 = n_2 = 20$ subjects in each condition.

This success probability analysis suggests that a better investigation of breathing and memory performance needs a different experimental design. Simpler designs are generally better because the more requirements you impose on a set of data (e.g., to produce many significant or non-significant outcomes) the lower the probability that any particular dataset will produce the required set of outcomes. Given the low estimated probability of success for this study, one might wonder how the original authors were so fortunate as to pick random samples that happened to reject/not reject results in exactly the pattern they needed to support their theoretical claims. We address this issue in Chap. 10 by considering how statistics should be interpreted across replications.

Take Home Messages

1. Keep your design simple: consider compressing raw data into intermediate variables, which then are subjected to statistical analysis.
2. Compute a power analysis before you do your experiment to check whether there is a real chance that it may show an existing effect.
3. Keep your design simple: if a theory presupposes both significant and null results your power may be strongly reduced.