Power(ful) Guidelines for Experimental Economists *†

Kathryn N Vasilaky ‡ J Michelle Brock §

August 1, 2020

Abstract

Statistical power is an important detail to consider in the design phase of any experiment. This paper serves as a reference for experimental economists on power calculations. We synthesize many of the questions and issues frequently brought up regarding power calculations and the literature that surrounds that. We provide practical coded examples and tools available for calculating power, and suggest when and how to report power calculations in published studies.

JEL: C9
Keywords: Power, Experiments, Design, Significance.

1 Introduction

> In spite of years of teaching and using statistics, we had not developed an intuitive sense of the reliability of statistical results observed in small samples. (Kahneman, 2011)

The purpose of this paper is to provide a concise and consolidated resource for experimental economists regarding power calculations. The significance of a test is central to our understanding of hypothesis testing, while its cousin, statistical power, remains peripheral. Power calculations are important for experiment and survey design. Nonetheless, researchers are either not performing power analyses, or are simply not reporting them (Czibor et al., 2019; Zhang and Ortmann, 2013). Thus, it is important to reiterate and provide additional resources for *ex ante* analysis of statistical power, and what to report *ex post*.[1]

*Thanks to contributions from the ESA discussion forum. We gratefully acknowledge the JESA editors and Eduardo Zambrano for very helpful comments.
†forthcoming in Journal of the Experimental Science Association
‡Cal Poly, Department of Economics and Columbia University, Intl Res Inst Climate & Society(e-mail: kvasilak@calpoly.edu)
§European Bank for Reconstruction and Development and CEPR (e-mail: BrockM@ebrd.com)
[1]For additional references on power in the behavioral sciences see Cohen (1988); Gerber and Green (2012b); Murphy et al. (2014)

In Section 2 we provide the formal definition of power, and provide intuition for how to operationalize it. Section 3 describes the contexts in which reporting power calculations is relevant and in Section 4 we discuss what calculations can be considered after an experiment has already been conducted. Section 5 considers some experimental designs for which it may be difficult to obtain sufficient power in smaller samples. In Section 6 we provide several options for computing power calculations, including simulated power calculations with sample code. Section 7 concludes.

2 What is statistical power?

Power is the probability that an experiment will lead to the rejection of the null hypothesis if it is indeed false, given a pre-specified target significance threshold (Gerber and Green, 2012a). In intervention research, this is referred to as sensitivity, or the ability to detect a difference between the treatment and control conditions for some outcome of interest. Choosing the statistical power of a test in the design phase of an experiment helps researchers determine how much data to collect, given their research question(s). Power is not linked to causal inference, nor is it a tool for analysing data. It is an experimental design tool, but is rarely reported in experimental economics papers.

Failure to consider power during design of an experiment can lead to incorrect, or even harmful, conclusions. Consider the context of a field experiment, where a researcher wants to estimate the impact of a cash incentive on a preventative health behavior, such as reducing alcohol consumption. She will conduct an experiment with a randomly assigned treatment group that receives a cash incentive and a control group that does not. Suppose there is in fact a positive relationship between receiving the particular incentive and alcohol consumption - people who receive the money actually increase alcohol consumption. If the test to find the relationship is under-powered, it indicates that the probability of observing the true relationship is low. This is problematic for two reasons. First, failing to reject a false null hypothesis (i.e. conclude that the incentive is neither definitively effective nor ineffective) could result in people being harmed if the intervention is adopted into policy under the assumption that it is harmless.[2] Second, a found effect in an under-powered study is likely to be far from the true effect. Over-estimating the harm of an intervention may then prevent adoption of an otherwise useful policy.[3]

The statistical power of an experiment depends on the null and alternative hypotheses, the targeted significance level and the (anticipated) features of the data (i.e. mean, variance and distribution). Before conducting the experiment, the researcher specifies a null and alternative hypothesis for each outcome variable of interest. We follow the convention of deriving power following the potential outcomes framework and Rubin causal model that is frequently

[2]Concluding that failure to find a harmful effect indicates a harmless intervention is conditional on interpreting an insignificant coefficient with a wide confidence interval as a precise zero.

[3]Note that the confidence interval of an estimated effect size would not immediately inform a reader of whether the researcher was powered to detect that effect size ex-ante. Power informs us of whether the effect size may be overstated, whereas the confidence interval around the ex-post effect size does not.

used to discuss randomized experiments (e.g. Athey and Imbens (2017); Chow et al. (2008); List et al. (2011)). Let the observed outcome of the treatment be denoted $\hat{\mu} \sim \mathcal{N}(\mu, \sigma^2/n)$.[4] Suppose there are two potential outcomes from the treatment, μ_0 and μ_1, where μ_0 refers to the outcome in the absence of the treatment (which is captured empirically via the control group), and μ_1 refers to the outcome in the presence of the treatment (which is captured empirically via the treatment group).

Let $\theta = \mu_1 - \mu_0$ denote the true treatment effect for the main variable of interest. Let the researcher's null hypothesis for θ be that $\theta = \theta_0$, where $\theta_0 \in \mathbb{R}$. For example, if the researcher wants to test the hypothesis of no difference between treatment and control groups, then θ_0 would be zero. The researcher chooses a one-sided alternative that there is some positive treatment effect greater than θ_0.

Null Hypothesis
H_0:$\theta = \theta_0$

Alternate Hypothesis
H_1:$\theta > \theta_0$

where $\theta_0 \geq 0$. Results from hypothesis tests are in terms of H_0; one either rejects H_0 or fails to reject H_0. When deciding whether to reject H_0 it is possible to make two kinds of errors.

A Type I error, or a false positive, occurs if one rejects the null hypothesis, when it is in fact true. The probability of a Type I error is denoted as α. It occurs if $\hat{\theta}$ falls "too far" from θ_0 for the researcher to believe that θ_0 is the true value of θ. What constitutes "too far" is decided by the researcher, and generally it is set so that $\alpha \leq 0.05$.[5] This is illustrated below. Let c denote the (non-standardized) cut-off such that the researcher will reject H_0 if $\hat{\theta} > c$.

$$\begin{aligned} \alpha &= Prob\left(\text{reject } H_0 | H_0\right) \\ &= Prob\left(\hat{\theta} \geq c | H_0\right) \end{aligned}$$

Since $\hat{\theta}$ is normally distributed, the researcher will use a z-test. We standardize c so that we can use the standard normal statistical tables to identify the critical value for any given α, as follows:

$$\alpha = Prob\left(\frac{\hat{\theta} - \theta_0}{\sqrt{V(\hat{\theta})}} \geq \frac{c - \theta_0)}{\sqrt{V(\hat{\theta})}}\right) \implies$$

[4] In Section 6 we touch on non-normally distributed outcomes.

[5] Traditionally, $\alpha \in \{0.10, 0.05, 0.01\}$. Recently, Benjamin et al. (2018) advocate for redefining statistical significance in economics according to an $\alpha = 0.005$ for claims of new discoveries.

$$1-\alpha = Prob\left(\frac{\hat{\theta}-\theta_0}{\sqrt{V(\hat{\theta})}} \leq \frac{c-\theta_0}{\sqrt{V(\hat{\theta})}}\right)$$

$$= \Phi\left(\frac{c-\theta_0}{\sqrt{V(\hat{\theta})}}\right) \Longrightarrow$$

$$\Phi^{-1}(1-\alpha) = \frac{c-\theta_0}{\sqrt{V(\hat{\theta})}} \Longrightarrow$$

$$B_{1-\alpha} = \frac{c-\theta_0}{\sqrt{V(\hat{\theta})}}$$

where $B_{1-\alpha}$ is the critical value associated with the $1-\alpha$ portion of the standard normal distribution that is centered around θ_0 (recall that $B_{1-\alpha} = -B_\alpha$). For example, for a normally distributed outcome variable, if the researcher chooses a one-sided test with $\alpha = 0.05$, then $B_{1-\alpha} = 1.645$. This means that the researcher will reject H_0 if the normalized $\hat{\theta}$ is greater than or equal to 1.645. For a two-sided test with $\alpha = 0.05$, $B_{1-\alpha} = 1.96$. This means that the researcher will reject H_0 if the normalized $\hat{\theta}$ is at least 1.96 units from 0.

Statistical power is related to the second type of error, the Type II error. A Type II error, or a false negative, occurs if one does not reject the null hypothesis when it is in fact false. A researcher would be committing a Type II error if the true treatment effect were something other than θ_0, but she fails to reject the hypothesis that it is θ_0. The probability of a Type II error is denoted as β. Power is 1-Pr(Type II error).

The approach to analyzing power depends on whether the researcher chooses a simple alternative, such as $\theta = \theta_1$, or a composite alternative, such as $\theta > \theta_0$. For the simple alternative, the power of the test is defined relative to a specific H_1 - the researcher must assert that θ will exclusively take either θ_0 or another value, say θ_1. This requires one power calculation, which we derive below, for a test statistic with a normal distribution. We return to composite hypotheses subsequently.

$$\beta = Prob\left(\text{fail to reject } H_0 | H_1\right)$$

$$= Prob\left(\hat{\theta} \leq c | H_1\right)$$

$$= Prob\left(\frac{\hat{\theta}-\theta_1}{\sqrt{V(\hat{\theta})}} \leq \frac{c-\theta_1}{\sqrt{V(\hat{\theta})}}\right)$$

$$= \Phi\left(\frac{c-\theta_1}{\sqrt{V(\hat{\theta})}}\right)$$

$$\implies \Phi^{-1}(\beta) = \frac{c - \theta_1}{\sqrt{V(\hat{\theta})}}$$

$$\implies B_\beta = \frac{c - \theta_1}{\sqrt{V(\hat{\theta})}}$$

where B_β is the critical value associated with the β portion of the standard normal distribution conditional on H_1 being true (e.g. for a distribution centered around θ_1). For example, for an outcome variable with a standard normal distribution, if the researcher chooses $\beta = 0.20$, then 1 - $\beta = 0.8$ and $B_\beta = 0.84$.

Note that c is the same in both the α and β formulas. On one side of c the researcher feels she cannot reject H_0 (and can reject H_1). On the other side of c, she feels she must reject H_0 (which implies she believes that H_1 is more likely to be true, given the data). B_β and $B_{1-\alpha}$ are different in so far as the normalized c is a different number of standardized units away from each of θ_0 and θ_1.

Solving both α and β equations for expressions of c obtains:

$$\frac{c}{\sqrt{V(\hat{\theta})}} = -B_\alpha + \frac{\theta_0}{\sqrt{V(\hat{\theta})}}$$

$$\frac{c}{\sqrt{V(\hat{\theta})}} = B_\beta + \frac{\theta_1}{\sqrt{V(\hat{\theta})}}$$

Returning to our example, since the researcher has two independent groups for treatment and control, $V(\hat{\theta}) = \sigma_1^2/n_1 + \sigma_0^2/n_0$, where n_i refers to the sample size in each of group i. Also, let $\sigma_1^2 = \sigma_0^2$. Then, setting the two critical values that satisfy Type I and Type II errors equal, we can solve for the sample size.[6] The sample size, $N = n_0 + n_1$, that the researcher needs to achieve the desired level of statistical power for her experiment is:

$$N = \frac{(B_\alpha + B_\beta)^2}{(\theta_1 - \theta_0)^2} \cdot \frac{\sigma^2}{\gamma(1-\gamma)} \tag{1}$$

where γ equals $(\frac{n_1}{n_1+n_0})$, the portion of the total sample, N, that is in the treatment group.

Replacing the parameters with values and solving for N prior to data collection, is what we refer to as "power calculations." Note that B_α determines B_β for a given θ_0, θ_1, σ and N. This makes it clear the trade-off between power and significance, and how this trade-off may change as we vary θ_1 (for composite null hypotheses). However, for the same α one can obtain additional power by increasing N, by choosing a wider distance between θ_0 and θ_1, or by using a one-sided, as opposed to a two-sided, alternative in hypothesis. Across

[6] For unequal variances see List et al. (2011) Equation 7.

disciplines, it is generally accepted to aim for a power of 0.8 (Lenth, 2001).[7]

Finally, note that rather than specifying θ_1, the researcher can use Equation 1 to determine the minimum detectable effect, $\theta_1 - \theta_0$, that she would be able to observe given a fixed sample size and power.

Composite tests and power functions In practice, single power calculations are frequently performed to correspond with composite tests. This is not necessarily misleading if the researcher has reliable estimates of θ_0, θ_1 and σ. It is, however, limiting. As we demonstrate above, a single power calculation requires a specific alternative value, θ_1. A composite hypothesis admits multiple values for θ_1. Thus, to properly assess power for a composite alternative, the researcher must estimate a power function. Using the arguments specified for the function, the researcher can determine the minimum θ_1 that will yield the minimum acceptable statistical power.

A power function is a mapping of effect sizes to power, given a fixed significance, sample size and variance. Calculating a power function (for a composite hypothesis) requires the same steps as a power calculation for the simple hypothesis, but rather than calculating a single β (for $\theta = \theta_1$), the researcher will calculate a β for each possible alternative value of θ (e.g. $\forall\theta_1 \in \mathbb{R}$ or $\forall\theta_1 \in \{-0.50, -0.25, ..., 1.25, 1.5, ...\}$, etc.).

Here we demonstrate how a power function is constructed for a test statistic that has a standard normal distribution. For this example, let $\hat{\theta} \sim \mathcal{N}(\theta, 2\sigma^2/0.5N)$, $\sigma^2 = \sigma_0^2 = \sigma_1^2 = 4$, $N = n_0 + n_1$ and $n_0 = n_1$. The researcher pre-specifies the hypothesis as $H_0{:}\theta = 0$ and $H_1{:}\theta > 0$, has a fixed sample size of $N = 64$, and chooses $\alpha = 0.05$. She will reject the null if $\hat{\theta}$ is greater than the critical value $c_\alpha = 0.823$. We include the "α" subscript to differentiate the critical values used for the one- and two-sided power functions.[8] To generate the values of the power function, we insert the parameter values, and then successive values of θ_1, into the power formula:

$$
\begin{aligned}
(1-\beta)_\alpha &= 1 - \Phi\left(\frac{c_\alpha - \theta_1}{\sqrt{2\sigma^2/0.5N}}\right) \\
&= 1 - \Phi\left(\frac{0.823 - \theta_1}{\sqrt{2 * 4/0.5 * 64}}\right)
\end{aligned}
$$

Let power evaluated for any given θ_1 be denoted $\mathit{\Pi}(\theta_1)$. The list of calculations below demonstrates how power evolves as the effect size increases. We use the known standardized value for $\alpha = 0.05$ to find the corresponding un-standardized cut-off value under the null hypothesis. We then standardize the cut-off relative to the distribution for the alternative hypothesis. One can think of this as shifting a curve centered on θ_1 further from that stationary cut-off value, as in Figure 1.

[7] Replications generally require higher power, e.g. Camerer et al. (2016).

[8] We obtain c_α by "unstandarizing" the critical value from the standard normal distribution for $\alpha = 0.05$. $c_\alpha = 0 + 1.645\sqrt{2 * 4/32} = 0.8225$.

$$
\begin{aligned}
\Pi(-0.50) &= 1 - \Phi\left(\frac{0.823 + 0.50}{0.5}\right) = 0.004 \\
&\vdots \\
\Pi(-0.25) &= 1 - \Phi\left(\frac{0.823 + 0.25}{0.5}\right) = 0.016 \\
&\vdots \\
\Pi(0) &= 1 - \Phi\left(\frac{0.823 - 0.0}{0.5}\right) = 0.050 \implies \Pi(0) = \alpha \\
&\vdots \\
\Pi(0.25) &= 1 - \Phi\left(\frac{0.823 - 0.25}{0.5}\right) = 0.126 \\
&\vdots \\
\Pi(1.25) &= 1 - \Phi\left(\frac{0.823 - 1.25}{0.5}\right) = 0.804
\end{aligned}
$$

The resulting values can be plotted on a graph, as in Figure 2. The solid line pertains to the one-sided power function. The negative values are included to demonstrate the long left tail for the one-sided power function with $H_1{:}\theta > 0$.[9]

The dotted line in Figure 2 is the graph of a two-sided power function. A two-sided power function is appropriate for a two-sided alternative hypothesis, e.g. $\theta \neq 0$. The formula for two-sided power is the union of the probabilities for rejecting the null for the left and right tails. To generate the values of the power function for a two-sided test, $\alpha/2$ determines the critical value, rather than α. The two-sided test is an "or" test - the hypothesis is either below zero (the hypothesized value) or above zero. It cannot be both. Power for the two-sided test is thus about the absolute value of the effect size. It is symmetrical about the vertical axis and is lower than the one-sided function for each value of θ_1 (because the $\alpha/2$ critical value is used).

$$
\begin{aligned}
(1-\beta)_{\alpha/2} &= Prob(\text{reject } H_0 | \theta \neq 0) \\
&= Prob(\hat{\theta} \geq c_{\alpha/2} | \theta \geq 0 \cup \theta < 0) \\
&= 1 - \Phi\left(\frac{c_{\alpha/2} - \theta_1}{\sqrt{2\sigma^2/0.5N}}\right) + \Phi\left(\frac{-c_{\alpha/2} - \theta_1}{\sqrt{2\sigma^2/0.5N}}\right)
\end{aligned}
$$

For values of θ_1 greater than θ_0 (zero in this example), the the second term in this equation is very small - this is the long left tail, with very little probability mass, from the one-sided test. The reverse is true for values of θ_1 that are less than θ_0; in this case the first term is the smaller of the two. The calculations below show how the power of the two-sided test converges to 1 as θ_1 increases. As with the one-sided test, we find the corresponding cut-off value of 0.98 given $\alpha/2$, and the values of N and σ in this example.

[9] For $H_1{:}\theta < 0$ there is a long right tail. $\Pi(\theta_1)$ converges to 1 as θ_1 decreases. It converges to zero as θ_1 moves further from θ_0 to the right.

$$
\begin{aligned}
\Pi(-0.5) &= 1 - \Phi\left(\frac{0.98 + 0.50}{0.5}\right) + \Phi\left(\frac{-0.98 + 0.50}{0.5}\right) = 0.170 \\
&\vdots \\
\Pi(0) &= 1 - \Phi\left(\frac{0.98 - 0.0}{0.5}\right) + \Phi\left(\frac{-0.98 - 0.0}{0.5}\right) = 0.050 \implies \Pi(0) = \alpha \\
&\vdots \\
\Pi(0.5) &= 1 - \Phi\left(\frac{0.98 - 0.50}{0.5}\right) + \Phi\left(\frac{-0.98 - 0.50}{0.5}\right) = 0.170 \\
&\vdots \\
\Pi(1.4) &= 1 - \Phi\left(\frac{0.98 - 1.4}{0.5}\right) + \Phi\left(\frac{-0.98 - 1.4}{0.5}\right) = 0.800
\end{aligned}
$$

A power function, therefore, returns the power associated with a range of alternative θ's under H_1. Each power function pertains to a given N (and σ^2), and one can follow a similar procedure to asses the trade-off between θ_1 and N for a given power. Power functions are easy to generate using statistical software.[10] See Section 6 for more detail. We have presented B_α and B_β as the critical values in the standard normal distribution. The assessment of statistical power follows similarly for test statistics that have other distributions.

For a general overview of performing power calculations, see the Jameel Poverty Action Lab's (J-PAL) note on power calculations in the course "Evaluating Social Programs" (JPAL, 2014). Zhong (2009) and Ledolter (2013) also provide a number of numerical examples and derivations.

3 When to report power?

Taking power into account in a study design is important for economists because doing so increases efficiency of experimental design. The idea is to avoid samples that are either unnecessarily large (and thus unnecessarily expensive) or too small to detect an effect. It also disciplines the researcher to focus on economically meaningful effects because one must specify an anticipated effect, $\theta_1 - \theta_0$ (along with α, β, and σ), in order to determine N. But once an experiment is completed, we might ask if it is necessary to report the power calculations used to arrive at its sample size. We posit that reporting power calculations is useful under two scenarios in particular: a) when a study was too under-powered *ex ante* to detect the statistically significant effect that it does find and b) for replicating studies and publishing well-designed studies with null effects.

It is important to emphasize that reporting power calculations does not help in the interpretation of the experimental results. Once an experiment has been completed we should rely on statistical inference to determine the impact of our result. This includes not just the point-estimate of the effect size and its p value, but also a discussion of the estimate's confidence interval. Power and threshold levels of statistical significance are determined prior to data collection, while the confidence intervals are a function of the data. Confidence

[10] For example, the following command will generate a similar power function to Figure 2 in Stata version 15: `power twomeans 0 (-2.5(.25)2.5), sd(2) knownsds n(64) graph.`

intervals are especially useful when results are not significant because they indicate a range of plausible values for the true θ. Power and confidence intervals are linked through sample size, and a low powered study will be reflected in a wide, and thus inconclusive, confidence interval. That interval could include treatment effects that are and *are not* economically meaningful, even if the point estimate is statistically significant.[11]

When researchers report their results from a study, they are also providing validated sample statistics for other researchers to use (or not to use). We refer again to Equation 1. What values should be used for μ_0, μ_1 and σ? Researchers can look to well-powered past studies for estimates. Overstated effect sizes from low powered studies should not be used to power future studies (Button et al., 2013; Gelman and Carlin, 2014; Ioannidis, 2005; Szucs and Ioannidis, 2017). If past studies do not report power, this distinction cannot be made. As seen in Figure 3, the lower the power of a test for a given σ, the closer that H_0 and H_1 will be. A more extreme point estimate is, therefore, needed in order to reject H_0 in favor of H_1 in low powered study. Using a t-distribution, as opposed to a standard normal distribution does not adjust for this. Since the tails of t-distribution are wider than those of the normal distribution, t-scores are larger than z-scores for the same level of significance. Therefore, only large deviations, far in the tail of the H_1 distribution, will classify as statistically significant in underpowered studies. Powering a study using an overstated effect size as a target for μ_1 would lead to yet another underpowered study.

Reporting power calculations is also important for replication exercises and qualifying null effects as recommended by the Journal of the Economic Science Association (Nikiforakis and Slonim, 2015). Publishing studies that failed to detect significance in a well-powered study, where others may have found a significant effect, is an important part of the scientific process. For example, Zethraeus et al. (2009) study the relationship between hormones and economic behavior in the lab. The authors are explicit about the power of their study, which is sufficiently high at over 90%. Participants are randomized into different hormone treatment groups and then play a series of games. The authors find no significant effect, a contradiction to existing correlative results (e.g. Apicella et al. (2008); Burnham (2007)). Such a result deserves consideration for publication, as it adds to a body of scientific evidence. Anything less than this contributes to publication bias.[12]

One way to accommodate publication of *ex ante* power calculations, which has taken hold in recent years, is pre-registration. Pre-registration essentially forces a researcher to publicize her intended hypothesis tests and the needed sample size for those tests.[13] The researcher would list any statistical software and associated commands that she used to perform the

[11] Goodman and Berlin (1994) provide a useful rule-of-thumb (Predicted 95% CI = observed difference +- 0.7 * (true difference 80% power)) for predicted confidence intervals, which depend on the observed effect size, β, and α.

[12] A few pooled replication papers have received considerable attention in economics and psychology (Camerer et al., 2016; Nosek, 2015), but it remains to be seen if individual, well-powered studies that find no effect will occupy space in top journals.

[13] For example, EGAP allows for pre-registration of experimental studies: https://egap.org/registration-license-options/.

calculations. If the researcher uses simulations to explore power, the code should be provided. This prevents the researcher from data mining, or running dozens of hypothesis tests, while only reporting the one or two significant results (Anderson, 2008), because the study is only powered to detect a certain number of effects. But, as Coffman and Niederle (2015) detail, the main downside to pre-analysis plans is that they tie researchers to particular analyses and inhibit exploratory work, which can be particularly taxing for young researchers or researchers without sufficient budgets to carry out pilot studies. As a result, Coffman and Niederle (2015) advocate for establishing a norm that journals publish well-powered replications of studies rather than tying researchers to pre-analysis plans. Banerjee et al. (2020) suggest that researchers can offer pre-analysis plans, but that the results be evaluated as a "distinct object" from the plan.

4 What can we compute ex-post?

Researchers may find themselves in a situation where the experiment has been completed and either a) they did not use power to determine sample size, or b) the parameter values they chose for the *ex ante* power calculations were inaccurate and actual effect sizes were much smaller (or larger) than anticipated, and/or c) their study faced considerable attrition (at random) such that the researcher was unable to maintain the sample size that her original power calculations dictated (as often occurs with field studies). The researcher then would like to know what can be computed ex post?

First, we can begin with what should *not* be done *ex post*. One temptation may be to retrospectively calculate an "observed" or "post hoc" power given the observed p value, treatment effect, variance and sample size from the completed experiment. This calculation is problematic because power is not an observable concept (Goodman and Berlin, 1994; Hoenig and Heisey, 2001; Lenth, 2001). Target significance, based on α, is an *ex ante* concept that is useful insofar as it helps us calculate power. But the observed significance has nothing to do with power. Moreover, observed power and the observed p value are inversely related, while the *ex ante* trade-off between α and $1-\beta$ is positive (Hoenig and Heisey, 2001). The difference is subtle, and often goes unrecognized. We provide a graphical example, which we believe best exhibits the fallacy of observed power.

Take a scenario where the researcher pre-specifies an $\alpha = 0.05$, and $\beta = 0.2$, and is testing H_0:$\hat{\theta} = 0$ against H_1:$\hat{\theta} \neq 0$, where $\hat{\theta}$ follows a standard normal distribution. $B_{1-\alpha/2} = 1.96$. Her observed test statistic is 1.9, with distribution under H_1'. She fails to reject the null (top panel of Figure 4). Now she decides to compute observed power. Observed power is the probability that the observed statistic falls to the left of 1.96 under H_1', the curve centered around 1.9 (Prob($\hat{\theta} \geq 1.96|\theta = 1.9$)). The bottom panel of Figure 4 has "observed power" shaded. We can see that this probability will always be 0.5 or less (since the probability of landing to the left of 1.9 is 0.5 and to the right of 1.9 is 0.5). We can also see that if α had been smaller than 0.05, then observed power would be even smaller.

Hoenig and Heisey (2001, pg 2) plot observed power against the pre-specified α, which shows

that any insignificant estimate from a study will, mechanically, exhibit an *ex post* power that is less than 50%. Conversely, any significant estimate from a study will, mechanically, exhibit ex post power that is greater than 50%. For this reason, any *ex post* power calculation where the effect is significant will result in high *ex post* power, and, conversely, insignificant effects will result in low *ex post* power.

Abandoning observed power does not conflict with performing the *ex post* calculation recommended by Nikiforakis and Slonim (2015), particularly, for the publication of studies with null effects. One can calculate the minimum detectable effect size given the sample size and estimates of σ from the data (as well as predesignated values for α and β), which, crucially, do not depend on whether the study found a significant result or not. The latter information cannot help with inference on the study's observed results, but it can provide a clue as to whether economically meaningful effect sizes might have been overlooked in the original study design, particularly, if the researcher was overly optimistic with respect to the magnitude of the intervention's effect.

Other *ex post* methodologies to replace observed power calculations have been proposed, but are beyond the main scope of this paper. Gelman and Carlin (2014) propose design analysis with accompanying code, which focuses on how to interpret results from studies with small sample sizes. In particular, they focus on the probability that a found effect size is the wrong sign or is far in magnitude from the true effect size, and whether the minimum detectable effect is scientifically meaningful.

In sum, once an experiment has been completed, and a statistically significant effect was not found, researchers have the option of computing the minimum detectable effect given their sample size and variance. They should also consider the confidence intervals around their effects to determine how well measured their effect sizes are. Both low power and poor measurement can result in wide confidence intervals. But one should not compute the $1-\beta$ associated with an observed effect.

5 Considerations for small samples

Power poses a particular challenge to researchers who are limited to using small samples, due to, for example, funding constraints or working with difficult to obtain samples (more common for lab-in-the-field studies). We discuss four design features that are often employed in lab experiments and explain how they reduce the power of a study. Each of these features demands a larger sample size to detect an effect with a given effect size, variance, α, and β.

Two- versus one-sided alternative hypotheses Powering a study for a one-sided alternative, as opposed to a two-sided alternative, is often appropriate if the researcher expects an effect in a particular direction. A one-sided test will also afford the researcher more power than a two-sided test, and so can be particularly useful for studies that face sample restrictions. However, for some studies, even when the researcher expects an effect in a specific direction, it is still prudent to minimize the risk of not detecting effects in either direction.

This applies when the underlying theory or past literature are limited, such that the basis for the expected effect direction is weak.[14] In such studies, a conservative approach would be to power the study for a two-sided alternative.

Between- versus within-subject experiments Most experiments use either a between- or a within-subjects experimental design (some may involve both). Between-subject designs require a larger sample size than within study designs to reach the same level of statistical power. Intuitively, for a within study, each participant serves as her own control; whereas in a between design a large fraction of the total sample must be designated as the control group. The formula to calculate required sample size for a within-subject study is different than for a between-subjects study due to the structure of the standard errors. In this section we compare these formulas to explain more formally where the difference in power comes from.

The example discussed so far, with the corresponding formula for calculating N, Equation 1, is known as a between-subjects design. In a simple between-subject experiment, one random sample receives the treatment and another serves as the control. The difference in mean outcomes in each of the two samples is interpreted as the treatment effect.

For a within-subject design, the experimenter takes multiple observations from the same participants but under different conditions (e.g. before the treatment and after the treatment).[15] The variance for the mean difference in outcomes in a within-subjects study is:

$$V(\hat{\theta}_W) = \frac{\sigma_1^2}{N_W} + \frac{\sigma_0^2}{N_W} - \frac{2\sigma_1\sigma_0\rho}{N_W}$$

where ρ is the correlation of subjects' round 1 and round 2 outcomes and N_W indicates the sample size for a within-subjects design. The analogous equation to Equation 1 for determining sample-size for a two-round within-subjects study then is:

$$N_W = \frac{(B_\alpha + B_\beta)^2}{(\theta_1 - \theta_0)^2} \cdot 2\sigma^2(1-\rho) \tag{2}$$

This equation assumes common variance for both rounds, $\sigma^2 = \sigma_1^2 = \sigma_2^2$.[16]

[14] It also applies if the true effect (in the unanticipated direction) is potentially harmful: failure to detect it can lead to damaging policy implications or directions for future research that can unintentionally cause harm. We do not belabor this point here since economics laboratory experiments rarely test potentially harmful interventions.

[15] In very simple within-subject study, every participant receives each treatment. The treatment is not randomly assigned. A within-subject experiment generates panel data, and the total sum of squares is the sum of the within and between sum of squares. One can include individual specific effects in the analysis. See Baltagi (2013), Searle et al. (2009) and Wooldridge (2010) for further reference. To determine the precise effect that added covariates, fixed effects, or random effects have on statistical power, one can run simulations, which is discussed in Section 6.

[16] In Stata `powerpairedmeans` can be used to specify a within experiment power calculation. Stata allows for unequal variance and one can perform calculations with various estimates of ρ. For example, `power pairedmeans 110 105, corr(0.25) sd(15)` will assume a 0.25 correlation between individuals' outcomes under different treatments.

Comparing Equation 1 and Equation 2, we obtain:

$$N_W = 2(1-\rho)\gamma(1-\gamma)N_B$$

If the between-subjects study allocates half of the subjects to each of the treatment and control groups, this simplifies to $N_W = 1/2(1-\rho)N_B$ (Maxwell et al., 2018, p. 650). Thus, with $\rho > 0$, the needed N_W will be less than half of N_B for a given α, β, μ_0, μ_1 and σ. Even with $\rho = 0$, the between-subjects design will need twice as many participants to achieve the same power as a corresponding within design. Bellemare et al. (2014) show that "between study designs require 4 to 8 times more subjects than a within study design to reach an acceptable level of statistical power." [17,18]

In this section we discussed why within-subject designs provide more power than between designs; however, they may not always be appropriate for the particular experiment at hand. Charness et al. (2012) provide an overview of the pros and cons of each design and the circumstances under which each may be appropriate.

Multiple treatments A between subjects study with multiple treatments puts additional constraints on power. This is because the study attempts to detect a separate effect for each treatment on the outcome variable, and a sufficiently large group is needed to detect the effect of each treatment.

One way to maximize power with multiple treatment arms is by using an unbalanced design. This implies that there is a different number of participants in each treatment arm. Many studies with several treatments distribute the same number of subjects into each treatment group, because this optimizes power when variances are equal across groups. But if we can more precisely anticipate the expected variance of the outcome for each treatment arm, then the number of participants assigned to each of the treatment arms and control group can be different.

To elaborate, we could at least expect the observations in the treatment arm(s) to exhibit more variation than observations in the control group. In particular, we can assign few participants to treatment arms in which we might expect a lower variance in the outcome variable. Thus, designs with equal sample sizes across all treatment and control groups require a larger sample size than is optimal, because the highest variance across all treatment cells is (implicitly) assumed. List et al. (2011) provide a derivation of power calculations

[17] ρ in this formula should not be confused with the intra-cluster correlation coefficient (ICC), which is often also denoted as ρ. The ICC measures the Pearson correlation between two individuals in the same cluster. Accounting for $ICC > 0$ in Equation 1 gives $N_{BC} = \left[\sigma_2(1 + \text{ICC}(m-1))(B_\alpha + B_\beta)^2\right] / \left[\gamma(1-\gamma)(\theta_1 - \theta_0)^2\right]$, where m is the cluster size (List et al., 2011; Matthews, 2006, p. 204-207). $ICC > 0$ thus increases required sample size for a given power. See Sainani (2010) for examples. See Frèchette (2012) for discussion of when a within-subjects study has $ICC > 0$.

[18] Some experiments involve both within- and between-subject treatments - e.g. split plot designs, where a field is divided into several plots where each receives different irrigation schemes (between), and then the plots are sampled repeatedly using different types of fertilizers (within). The appropriate power calculations can be made for each outcome. See Maxwell et al. (2018) for detail in analyzing such designs.

for a treatment and control group with unequal variances (their equations 6 and 7).[19] Another promising area of research with regards to between designs with multiple treatments is adaptive designs, in which more study participants are allocated to promising treatment arms over time. See Finucane McKenzie et al. (2018) and Xiong et al. (2019).

Multiple hypothesis testing Additional consideration must be taken if researchers want to examine power for multiple hypotheses. Multiple hypothesis tests (MHT) can arise for several reasons: multiple treatments - meaning, for a given outcome, applying k treatments will require conducting k hypothesis tests to test the effect of each treatment; multiple subgroups - meaning, for a given treatment and outcome, testing the treatment's effect for each subgroup (e.g. the effect for men versus women); and multiple outcomes - meaning, testing the effect of any one treatment on multiple outcomes (List et al., 2019). As the number of hypothesis tests increases, so does the probability that one of them will be significant. After M independent tests, the probability of making at least one type I error in M tests is $1-(1-\alpha)^M$.[20] Thus, after 10 tests, and 5% significance, the probability of falsely rejecting one or more null hypotheses is already 40%.

One can adjust power calculations for multiple hypotheses via simulation. For example, Porter (2018) provides R code to adjust for multiple hypothesis tests in simulated power calculations. Rather than using p values generated from single t-tests, rejection rates are based on p values adjusted by any of the many available procedures (Bonferroni, Holm, or Benjamini-Hochberg) that account for the number of tests being conducted, and by a chosen definition for rejection - e.g. "the probability of detecting effects of at least a particular size on at least one outcome, or the probability of detecting an effect of a particular size or larger for each particular hypothesis test" (Porter, 2018). List et al. (2019) provide a new procedure to correct for multiple hypothesis testing that outperforms the Bonferroni correction in terms of power. In Section 6 we also provide simple starter code to generate one's own simulated power calculations. Note that MHT can put more demands on one's power calculations, but this will depend on the definition for rejection that is chosen.

Even if one takes multiple hypotheses into account *ex ante* in the power analysis, *ex post* corrections are still needed in order to account for the probability of one or more false rejections.

6 Computing power

Most statistical programming languages offer packages that will compute sample size, given the choice of α, β, μ_0, μ_1, and σ, as well as additional parameters such as the ratio of the

[19] We provide an example with multiple treatment arms and different variances in Section 6. Also, Stata allows for different ratios of the treatment to control sample size using the `nratio` option with the `power` command.

[20] P(Making an error) = P(reject $H_0|H_0$) = α; P(Not making an error) = P(not reject $H_0|H_0$); P(Not making an error in M tests) = $(1-\alpha)^M$; P(Making at least 1 error in m tests) = 1 - $(1-\alpha)^M$.

sample size between treatment to control groups. Stata includes the commands `power`,[21] `sampsi` and `sampclus`. Note, importantly, that Stata's default output is framed in terms of composite tests $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$, but the command allows the user to specify a specific null and a single alternative, on one side of θ_0. The the two-sided aspect is then reflected only in the critical value. Stata's `power` command can, however, produce power function outputs in table and graph form, for both one-sided and two sided tests. To generate the power function output, the user inputs a list of possible values for one or more parameters, and specifies the option `table` or `graph`. The two-sided option is also easily changed to one-sided, using the option `onesided`. Open source languages such as R and Python also have their respective `power` libraries.[22]

Another tool for power analysis is the Optimal Design (OD) software (Raudenbush et al., 2011).[23] OD exclusively considers two-sided alternative hypotheses applied to designs with a single treatment and a control and an even split of subjects in each group. OD does not accommodate designs where each subject is subject to two or more treatments. The tool can produce graphs that depict the trade-offs between any two chosen parameters. For example, the researcher may input an anticipated effect size and standard deviations to generate a graph of power versus sample size. To learn the coordinates of any given point on a graph, the user must click on the desired point. Stata will also produce these graphs, but in OD they are the default output and very quick to generate. However, unlike Stata, OD will not output a table.

OD can calculate power for a variety of experimental designs, with a focus on multi-level and longitudinal designs. There are two design options that JESA readers will likely find most useful. They are under the "person randomized trials" sub-menu. A person randomized trial is an experiment where the unit of randomization is the individual subject.[24] The pertinent options in the "person randomized trials" category are labeled "single level trials" and "repeated measures." A single level trial is the same as a single period between subjects experiment: choosing this option in OD tells the software that the study will have one (post-treatment) outcome measure per subject and that each subject is either in the treatment group or the control group. Note, however, that power calculations for this design option assume a continuous outcome variable.

The repeated measures option pertains to a between subject design with multiple observations per person. This kind of design is useful for tracking treatment effects over multiple periods of decision making. It thus accommodates panel data. A multiple round public

[21] A one sample, two-sided sample size calculation with $\alpha = 0.05$, $1 - \beta = 0.8$, $\mu_0 - \mu_1 = 2$ - 2.5 and $\sigma = 0.8$ is coded as `power onemean 2 2.5, sd(0.8)`; A two sample, two-sided mean test where $\mu_0 - \mu_1 = 12$ - 15, $\sigma_0 = 5$, $\sigma_1 = 7$ and the treatment groups is twice the control group is coded as `power twomeans 12 15, sd1(5) sd2(7) nratio(2)`.

[22] For a comprehensive list of power packages see Bellemare et al. (2016).

[23] The software is free and available here. More information can be found at this site. A guide by J-PAL that includes tutorials is available here.

[24] In contrast, OD also accommodates multi-site and block randomized trials, which are more relevant to field experiments.

goods game (with no feedback) is an example of a repeated measures design in the lab. As with the single level trials, each subject is assumed to be in either the treatment group or the control group. In OD, power calculations for this design assume a single observation per subject prior to treatment being applied (in both treatment and control groups). To use the repeated measures feature, one would specify the frequency of observations (F), the duration of the study (D), and the total number of observations per subject (M=FD + 1, where the 1 refers to a pre-treatment observation). As before, the calculations for this option pertain only to continuous outcome variables.

Normal vs non-normal distributions Section 2 contextualized the derivation of sample size and power with respect to a normally distributed test statistic. But the principles apply more generally, as well. In the way of non-normal distributions there are several considerations. For normally distributed data, if σ^2 is unknown, it is typically estimated using sample variances. As a result, the test statistic will have a t-distribution, and critical values associated with B_α and B_β would be taken from a t-distribution rather than a standard normal distribution. Stata takes this approach as its default for `twomeans` and `onemean`.[25] Second, other types of data, such as studies with multiple observations per subject, will require test statistics with different distributions. Most statistical packages support a variety of test statistics (and their corresponding distributions). Both Stata and OD assign an appropriate probability distribution automatically, given user inputted information on the study design and type of variable being tested (mean, proportion, etc.). The programs' user manuals describe the possible test statistics for the various designs (and tests) that they support.

In large samples, normal approximations may be adequate, particularly for t-statistics and chi-squared statistics. To illustrate, for power calculations that use a t-distribution, "large" may apply to a sample as small as 25 observations. This can be seen in Figure 5. The figure demonstrates how quickly the critical values that achieve 80% power using a t-distribution converge to the critical value from a standard normal distribution, as the sample size increases. The standard normal critical value is marked with a dashed line. This figure shows that the t-distribution and standard normal distribution return very similar power as the sample size passes 25. At $N = 20$, the same critical value to achieve 80% power on a t-distribution will yield 80.54% power using a standard normal. At $N = 25$ it yields 80.43%. Finally, for very small samples where a specific distribution is not assumed, non-parametric tests are more appropriate. Rahardja et al. (2009) and, more recently, Happ et al. (2019), provide closed form calculations of power calculations for the Mann-Whitney U test.

Even with the many programs available, researchers may still face situations where closed form solutions for sample size and power may not exist. In such cases simulation based power calculations can be a useful tool to overcome the weaknesses of programmed commands, and is common among statisticians (van der Sluis et al., 2008). Essentially, the researcher generates k samples of size N following the distribution specified under H_0 and k samples of

[25] In this case, Stata uses an estimate of 1 for the unknown standard deviation. The user can override the default by adding the option `knownsds` to the command line. This default changes to a z-distribution if the study design will lead to clustered error terms, which is common in field experiments.

size N following the distribution specified under H_1 and compares the two samples k times using her preferred statistical test. β is the proportion of k tests that are not rejected, and power is $1 - \beta$.

Arnold et al. (2011) provide sample code for simulated power calculations in R and Stata, where their examples include calculations for cluster randomized trials and studies with two treatments arms with different outcome variances. Two user written Stata packages also exist for simulating power calculations including Bellemare et al. (2016)'s *powerBBK* package, Luedicke (2013)'s *powersim* package. Bellemare et al. (2016)'s package is remarkably versatile and can account for experimental design, order effects, budget constraints, differences in variances across treatment and control, multiple treatment arms, and panel data.[26] We also provide a simple benchmark example in the Appendix A using Monte Carlo simulations to calculate power in Python that can be easily adjusted for other distributions, sample sizes, effect sizes, variances, and number of simulations. The code defines the parameters for two distributions (N_0, N_1, α, μ_0, μ_1, σ_0, σ_1), reflecting the distribution of each random sample that would be drawn from the treatment and control groups, and the number of simulations.[27] For each random draw from the treatment and control group, the program calculates the mean difference between the groups and its related p value on a standard normal distribution. It then reports power, or the percentage of times where the null is rejected across all simulations.

6.1 Choosing inputs for a power calculation

There are no strict rules for how to determine the values of μ_0, μ_1 and σ^2 for power calculations. Effect sizes and standard errors from studies that examine similar populations and treatments are the most common source. But studies do not uniformly report μ_0, μ_1 and σ^2, or they may not contain values for these parameters in specific sub-samples of interest (e.g. women versus men). Pilot studies or pre-intervention surveys can be useful, but must be considered carefully, as these are often small sample exercises. For exploratory studies, researchers may not know what absolute effect to expect, such that discovering an effect of any size may be sufficient to meet research goals. Clearly, choosing specific values for μ_0, μ_1 and σ^2 is not a straightforward or trivial exercise.

Moreover, a power calculation inherently assumes that the parameters are equal to the values to which we set them, where in fact, point estimates lie somewhere within a confidence interval. The data from which that point estimate is derived could support a much smaller

[26] A sample command for a study of $t = 2$ rounds, a budget ranging from 40 to 800 in 40 dollar increments per round, a within design, an effect size of 0.1, where the baseline is 6.3, individual heterogeneity variance of 0.045, and variance of the error term of 0.02 is: `budget(40(40)800) t(2) design(both) beta(6.3 0.1) muvar(0.045) epsvar(0.02) command(regress) panel rep(100)`.

[27] The comparison of means can be replaced by a regression framework where we regress the outcome variable on an indicator for treatment and control and store the resulting p value on the indicator variable. The latter lends itself to more complicated designs - multiple treatments, multiple subgroups, multiple outcomes.

or larger value for the true μ_0, or μ_1, which could change the power calculation considerably. We might think that an estimate of σ derived from the same data would account for this issue; however, confidence intervals decrease at a slower rate - with one over the square root of the sample size for the z-statistic- than the standard deviation alone (Gelman and Hill, 2006). Therefore, using a specific μ_0 or μ_1 to calculate power is never conclusive, since the true values are inside a confidence interval which shrinks slowly with N. These uncertainties argue that, except in very particular cases, power can be calculated over a range of parameter values. For exploratory studies, in particular, this will help the researcher avoid overly conservative sample sizes. Authors can present plots of the power function (power graphed against effect size) for a given sample size. This should be accompanied by a discussion of how the researcher used the information to decide on a sample size.

Finally, when a specific expected effect is hard to determine, or when the researcher has limited control over the sample size, it is useful to calculate the minimum detectable effect (MDE), given assumptions about sample size, and power. For example, the researcher can present the MDE under 90%, 80%, and 70% power, and discuss the conditions under which these MDEs are attainable. See, for example, Drichoutis et al. (2015).

7 Conclusion

Using and reporting power in published articles is a practice economists conducting experiments should adopt. In lieu of power calculations, experimental economists have tended to apply rules of thumb (e.g. $N > 30$) for determining sufficient sample sizes. Rules of thumb are not without statistical underpinning, but power calculations bring to focus the importance of economically meaningful effect sizes and also shed light on how and why a particular subject pool is attained.

We discuss many of the topics frequently brought up in experimental design and analysis that are also related to power, including the fallacy of observed power, overstated effect sizes, publication bias, the importance of reporting power for null effects, and replication.

References

Anderson, M. L. (2008, December). Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects. *Journal of the American Statistical Association 103*(484), 1481–1495.

Apicella, C., A. Dreber, B. Campbell, P. Gray, M. Hoffman, and A. Little (2008, November). Testosterone and financial risk preferences. *Evolution and Human Behavior 29*(6), 384–390.

Arnold, B. F., D. R. Hogan, J. M. C. Jr, and A. E. Hubbard (2011). Simulation meth-

ods to estimate design power: an overview for applied research. *BMC Medical Research Methodology 11*(94).

Athey, S. and G. W. Imbens (2017). *The Econometrics of Randomized Experiments*, Volume 1 of *Handbook of Economic Field Experiments*.

Baltagi, B. (2013). *Econometric Analysis of Panel Data*. West Sussex, UK: Wiley.

Banerjee, A., E. Duflo, A. Finkelstein, L. F. Katz, B. A. Olken, and A. Sautmann (2020). In praise of moderation: Suggestions for the scope and use of pre-analysis plans for rcts in economics. *National Bureau of Economic Research* (No. w26993).

Bellemare, C., L. Bissonnette, and S. Kröger (2014). Statistical Power of Within and Between-Subjects Designs in Economic Experiments. *IZA Discussion Paper No. 8583 1*(2).

Bellemare, C., L. Bissonnette, and S. Kröger (2016). Simulating power of economic experiments: the powerBBK package. *Journal of the Economic Science Association 2*(2), 157–168.

Benjamin, D. J., J. O. Berger, M. Johannesson, B. A. Nosek, E. Wagenmakers, R. Berk, K. A. Bollen, B. Brembs, L. Brown, C. Camerer, D. Cesarini, C. D. Chambers, M. Clyde, T. D. Cook, P. D. Boeck, Z. Dienes, A. Dreber, K. Easwaran, C. Efferson, E. Fehr, F. Fidler, A. P. Field, M. Forster, E. I. George, R. Gonzalez, S. Goodman, E. Green, D. P. Green, A. Greenwald, J. D. Hadfield, L. V. Hedges, L. Held, T. H. Ho, H. Hoijtink, D. J. Hruschka, K. Imai, G. Imbens, J. P. A. Ioannidis, M. Jeon, J. H. Jones, M. Kirchler, D. Laibson, J. List, R. Little, A. Lupia, E. Machery, S. E. Maxwell, M. Mccarthy, D. Moore, S. L. Morgan, M. Munafó, S. Nakagawa, B. Nyhan, T. H. Parker, L. Pericchi, M. Perugini, J. Rouder, J. Rousseau, V. Savalei, F. D. Schönbrodt, T. Sellke, B. Sinclair, D. Tingley, T. V. Zandt, S. Vazire, D. J. Watts, C. Winship, R. L. Wolpert, Y. Xie, C. Young, J. Zinman, and V. E. Johnson (2018). Redefine statistical significance. *Nature Human Behavior 2*, 6–10.

Burnham, T. C. (2007). High-testosterone men reject low ultimatum game offers. *Proceedings of the Royal Society B: Biological Sciences 274*(1623), 2327–2330.

Button, K., J. Ioannidis, C. Mokrysz, B. Nosek, J. Flint, E. Robinson, and M. Munafo (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience 14*, 365.

Camerer, C., A. Dreber, E. Forsell, T.-H. Ho, J. Huber, M. Johannesson, M. Kirchler, J. Almenberg, A. Altmejd, T. Chan, E. Heikensten, F. Holzmeister, T. Imai, S. Isaksson, G. Nave, T. Pfeiffer, M. Razen, and H. Wu (2016). Evaluating replicability of laboratory experiments in economics. *Science 351*(6280), 1433–1436.

Charness, G., U. Gneezy, and M. A. Kuhn (2012). Experimental methods: Between-subject and within-subject design. *Journal of Economic Behavior Organization 81*, 1–8.

Chow, S.-C., J. Shao, and H. Wang (2008). *Sample size calculations in clinical research*. Boca Raton: Chapman Hall/CRC.

Coffman, L. C. and M. Niederle (2015). Pre-Analysis Plans Have Limited Upside, Especially Where Replications Are Feasible. *Journal of Economic Perspectives 29*(3), 81–98.

Cohen, J. (1988). *Statistical Power for the Behavioral Sciences*. New York: Academic Press.

Czibor, E., D. Jimenez-Gomez, and J. A. List (2019). The Dozen Things Experimental Economists Should Do (More of). *Southern Economic Journal 86*(2), 371–432.

Drichoutis, A. C., J. L. Lusk, and R. M. J. Nayga (2015). The veil of experimental currency units in second price auctions. *Journal of the Economic Science Association 1*(2), 182–196.

Finucane McKenzie, M., I. Martinez, and S. Cody (2018). What works for whom? a bayesian approach to channeling big data streams for public program evaluation. *American Journal of Evaluation 39*(1), 109–122.

Frèchette, G. R. (2012). Session-effects in the laboratory. *Experimental Economics 15*, 485–498.

Gelman, A. and J. Carlin (2014). Beyond Power Calculations : Assessing Type S (Sign) and Type M (Magnitude) Errors. *Perspectives on Pscychological Science 9*(6), 641–651.

Gelman, A. and J. Hill (2006). *Sample size and power calculations*, pp. 437â456. Analytical Methods for Social Research. Cambridge University Press.

Gerber, A. and M. Green, Donald (2012a). *Field Experiments: Design, Analysis and Interpretation*. Newbury Park, CA: W W Norton and Company.

Gerber, A. S. and D. P. Green (2012b). *Field Experiments Design, Analysis, and Interpretation*. New York, NY: W.W. Norton.

Goodman, S. and J. Berlin (1994). The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. *Ann Intern Med 121*, 200–206.

Happ, M., A. Bathke, and E. Brunner (2019). Optimal sample size planning for the Wilcoxon-Mann-Whitney test. *Statistical Medicine 38*(3), 363–375.

Hoenig, J. M. and D. M. Heisey (2001). The Abuse of Power : The Pervasive Fallacy of Power Calculations for Data Analysis. *The Americian Statistician 55*(1), 19–24.

Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLoS Medicine 2*(8), e124.

JPAL (2014). How to do Power Calculations in Optimal Design Software. Technical report, Abdul Latif Jameel Poverty Action Lab.

Kahneman, D. (2011). *Thinking, Fast and Slow*. New York: Penguin Books.

Ledolter, J. (2013). Economic Field Experiments: Comments on Design Efficiency, Sample Size and Statistical Power. *Journal of Economics and Management 9*(2), 271–290.

Lenth, R. V. (2001). Some Practical Guidelines for Effective Sample Size Determination. *The American Statistician 55*(3), 187–193.

List, J., S. Sadoff, and M. Wagner (2011, March). So you want to run an experiment, now what? Some simple rules of thumb for optimal experimental design. *Experimental Economics 14*(4), 439–457.

List, J., A. M. Shaikh, and Y. Xu (2019). Multiple hypothesis testing in experimental economics. *Experimental Economics*, 1–21.

Luedicke, J. (2013). Simulation-based power analysis for linear and generalized linear models. In *Stata Conference*, pp. 1–25.

Matthews, J. N. S. (2006). *Introduction to Randomized Controlled Clinical Trials, 2nd edition*. New York, NY: Chapman Hall.

Maxwell, S., H. Delaney, and K. Kelley (2018). *Designing Experiments and Analyzing Data: A Model Comparison Perspective, 3rd ed.* Newbury Park, CA: Routledge.

Murphy, B., B. Myors, and A. Wolach (2014). *Statistical Power Analysis*. New York, NY: Routledge.

Nikiforakis, N. and R. Slonim (2015). Editorsâ preface: statistics, replications and null results. *Journal of the Economic Science Association 1*(2), 127–131.

Nosek (2015). Estimating the reproducibility of psychological science. *Science 349*.

Porter, K. (2018). Statistical power in evaluations that investigate effects on multiple outcomes: A guide for researchers. *Journal of Research on Educational Effectiveness 1*(2), 267–295.

Rahardja, D., Y. D. Zhao, and Y. Qu (2009). Sample size determinations for the wilcoxonâmannâwhitney test: A comprehensive review. *Statistics in Biopharmaceutical Research 1*(3), 317–322.

Raudenbush, S., J. Spybrook, R. Congdon, A. Martinez, H. Bloom, and C. Hill (2011). Optimal Design Software for Multi-level and Longitudinal Research (Version 3.01).

Sainani, K. (2010). The importance of accounting for correlated observations. *American Academy of Physical Medicine and Rehabilitation 2*(9), 858–861.

Searle, S. R., G. Casella, and C. E. McCulloch (2009). *Variance components*, Volume 391. John Wiley & Sons.

Szucs, D. and J. Ioannidis (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLoS Biol 15*(3).

van der Sluis, S., C. V. Dolan, M. C. Neale, and D. Posthuma (2008, March). Power calculations using exact data simulation: a useful tool for genetic study designs. *Behavior genetics 38*(2), 202–11.

Wooldridge, J. (2010). *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press.

Xiong, R., S. Athey, M. Bayati, and G. Imbens (2019). Optimal experimental design for staggered rollouts. *Available at SSRN: https://ssrn.com/abstract=3483934 or http://dx.doi.org/10.2139/ssrn.3483934* .

Zethraeus, N., L. Kocoska-Maras, T. Ellingsen, B. von Schoultz, A. L. Hirschberg, and M. Johannesson (2009, April). A randomized trial of the effect of estrogen and testosterone on economic behavior. *Proceedings of the National Academy of Sciences 106*(16), 6535–8.

Zhang, L. and A. Ortmann (2013). Exploring the meaning of significance in experimental economics.

Zhong, B. (2009). How to Calculate Sample Size in Randomized Controlled Trial? *Journal of Thoracic Disease 1*(1), 51–54.

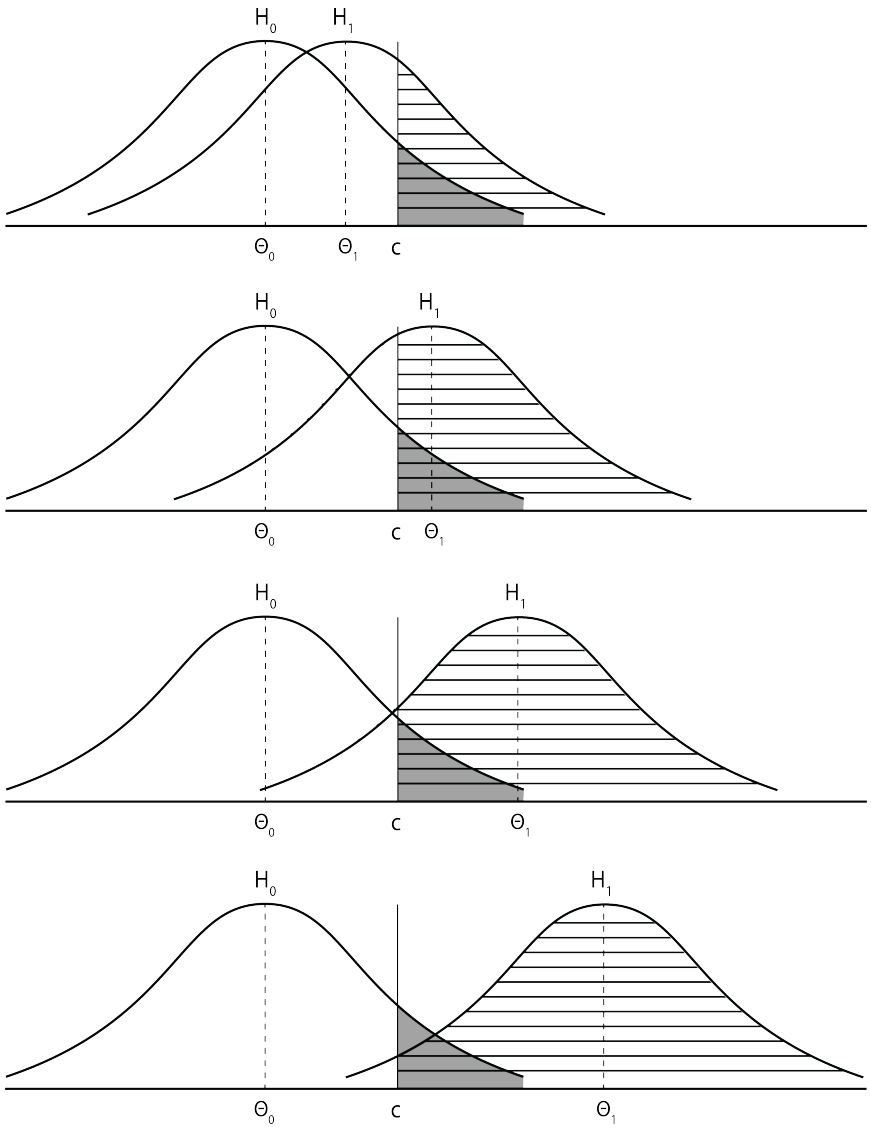


Figure 1: Power (one-sided) for successive values of θ_1. Power is the area marked by horizontal lines, c is the critical value for the chosen α and θ_0 and θ_1 are the values for the null and alternative hypotheses, respectively. The chart shows how power increases as θ_1 moves further from θ_0.

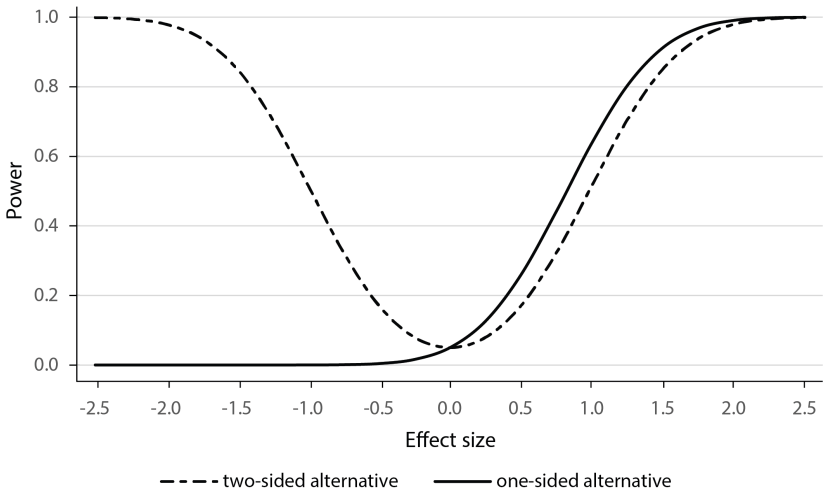


Figure 2: Power function for one- and two-sided alternatives. Power function is evaluated for alternative hypothesis effect sizes, θ_1.

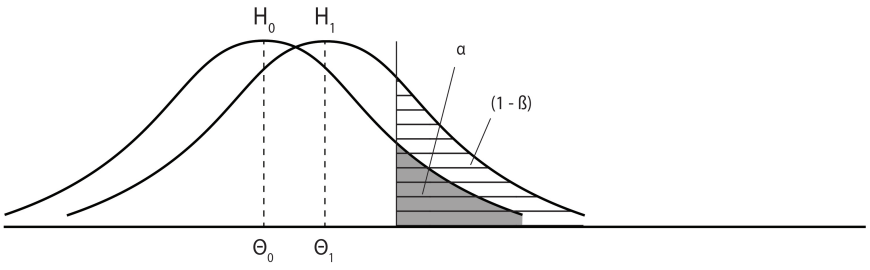
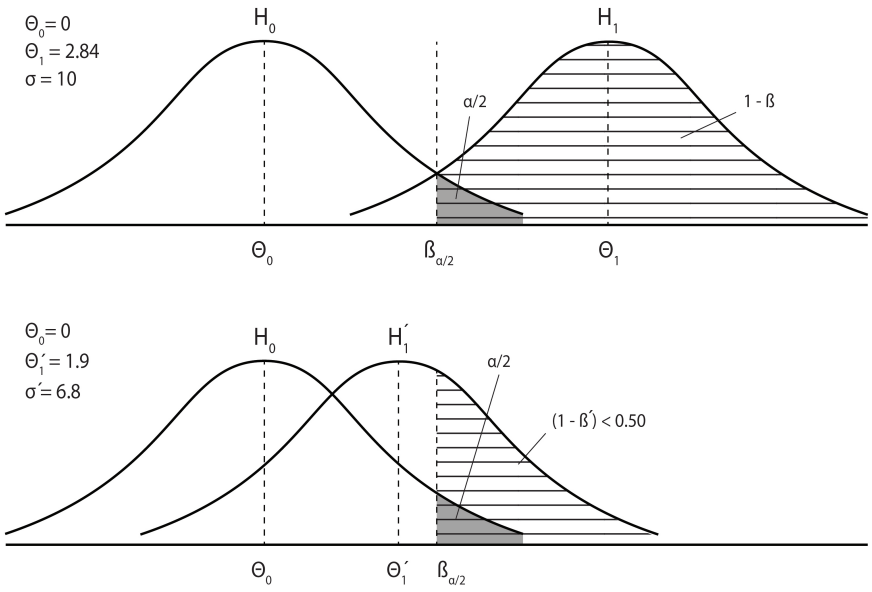


Figure 3: Overstated Effect Size in Underpowered Studies



$\alpha = 0.5$, $B_{\alpha/2} = 1.96$, $\beta = 0.2$, N = 100

Figure 4: Fallacy of Ex post Power

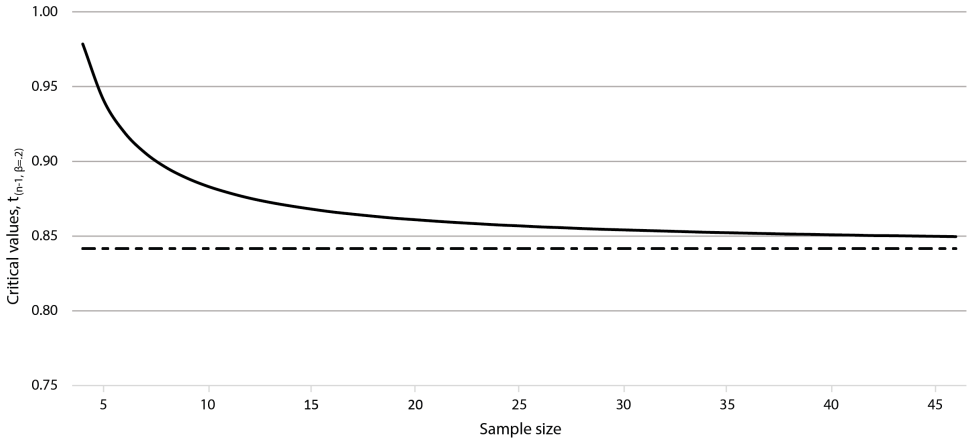


Figure 5: Convergence of t-distribution to normal distribution

A Appendix

A.1 Example: Simulation of Power Calculations

```
import numpy as np
import scipy.stats

# Set parameters
sample_0 = 30
sample_1 = 30
mean_0 = 0.0
effect_size = 0.8
sigma_0 = 1
sigma_1 = 1
simulations = 10000

# Empty list to store p values
p_values = []

# Draw samples from a normal distribution
for i in range(simulations):
    # Sample from control group
    control = np.random.normal(loc = mean_0, scale = sigma_0, size = sample_0)
     # Sample from treatment group
    treatment = np.random.normal(loc = mean_0 + effect_size, scale = sigma_1, size = sample_1)
    # ttest across control and treatment
    result = scipy.stats.ttest_ind(control, treatment)
    # Store p value from test
    p_values.append(result[1])

# Number of simulations where the null was rejected
p_values = np.array(p_values)
reject = np.sum(p_values < 0.05)

# Calculate percentage of times reject null
percent_reject = reject / float(simulations)

print("Power: ", percent_reject)
```

To alter the code to account for other distributions and statistical tests (e.g. non-parametric tests) we would simply replace the np.random.normal() function with another sampling distribution (e.g. np.random.chisquare()) and the scipy.stats.ttestind() function with another statistical test (e.g. such as scipy.stats.mannwhitneyu()).