European Bank
for Reconstruction and Development

Power(ful) Guidelines for Experimental Economists

Kathryn N. Vasilaky and J. Michelle Brock

Abstract
Statistical power is an important detail to consider in the design phase of any experiment. This note 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.

Keywords: statistical power, experiments, design, significance.

JEL Classification: C9.

Contact details: J. Michelle Brock, Principal Economist, One Exchange Square, London, EC2A 2JN, UK
Phone: +44 20 7338 7193; *email*: brockm@ebrd.com.

Kathryn N. Vasilaky is an Assistant Professor at California Polytechnic University and Research Affiliate at the Columbia University International Research Institute for Climate and Society. J. Michelle Brock is a Principal Economist at the EBRD and a Research Affiliate at the Centre for Economic Policy Research (CEPR).

Thanks to contributions from the ESA discussion forum.

Working Paper No. 239 **Prepared in December 2019**

Power(ful) Guidelines for Experimental Economists *

Kathryn N Vasilaky † J Michelle Brock ‡

December 19, 2019

Abstract

Statistical power is an important detail to consider in the design phase of any experiment. This note serves as a reference on power calculations for experimental economists. 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.

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 note 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 (Zhang and Ortmann, 2013; Czibor et al., 2019). Thus, it is important to reiterate and provide additional resources for *ex ante* analysis of statistical power, and

*Thanks to contributions from the ESA discussion forum.

†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)

what, if anything, to report *ex post*.[1]

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 (??, Ger). 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, and is rarely reported in experimental economics papers.

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 strong positive relationship between receiving the particular incentive and alcohol consumption. If the test to find the relationship is under-powered, it indicates that the probability of observing the relationship is low. This is problematic because 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.

Before conducting the experiment, the researcher specifies a null and alternative hypothesis for each outcome variable of interest.[2] We follow the convention of deriving power following the potential outcomes framework and Rubin causal model that is frequently used to discuss randomized experiments (e.g. Athey and Imbens (2016); Chow et al. (2008); List et al. (2011)). Let the observed outcome of the treatment be denoted $\hat{\mu} \sim N(\mu, \sigma^2/n)$.[3] 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

[1] For additional references on power in the behavioral sciences see Cohen (1988); Murphy et al. (2014); Gerber and Green (2012)

[2] Just as each hypothesis test has its own significance, each test also has its own power.

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

treatment effect (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 $\theta_0 \in \Re$. For example, if the researcher wants to test the hypothesis of no difference between treatment and control, 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 negative, occurs if one rejects the null hypothesis, when it is in fact true. The probability of a Type I error is denoted as α.[4] 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 $\alpha \leq 0.05$. This is illustrated below. Let c denote the (non-standardized) cut-off such that the researcher will reject H_0 if $\hat{\theta} > c$.

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

We standardize c so that we can use the standard normal statistical tables to identify the critical value for any given α, as follows:

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

$$
\begin{aligned}
\alpha &= Prob([\frac{\hat{\theta}-\theta_0}{\frac{\sigma}{\sqrt{n}}}] \geq [\frac{c-\theta_0)}{\frac{\sigma}{\sqrt{n}}}]) \implies \\
1-\alpha &= Prob([\frac{\hat{\theta}-\theta_0}{\frac{\sigma}{\sqrt{n}}}] \leq [\frac{c-\theta_0}{\frac{\sigma}{\sqrt{n}}}]) \\
&= \Phi(\frac{c-\theta_0}{\frac{\sigma}{\sqrt{n}}}) \implies \\
\Phi^{-1}(1-\alpha) &= \frac{c-\theta_0}{\frac{\sigma}{\sqrt{n}}} \implies \\
B_{1-\alpha} &= \frac{c-\theta_0}{\frac{\sigma}{\sqrt{n}}}
\end{aligned}
$$

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. Note that ($B_{1-\alpha} = -B_\alpha$). For example, for a normally distributed outcome variable, if the researcher chooses a one-sided test with α=0.05, then $B_{1-\alpha}$=1.645. This means that the researcher will reject H_0 if the normalized $\hat{\theta}$ exceeds 1.645.

Statistical power is related to the second type of error, the Type II error. A Type II error, or a false positive, occurs if one does not reject the null hypothesis when it is in fact false. Our 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 analysing 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.

Single power calculations, for a specific H_1, are frequently done for composite alternatives. However, to truly calculate power for a composite alternative, the researcher must estimate a power function. Calculating a power function requires the same steps as a power calculation for the simple hypothesis, but rather than calculating a single β (for θ =θ_1), the researcher will calculate a β for each possible alternative value of θ (e.g. all integers greater than zero). A power function, therefore, returns the power associated with a range of alternative θ's under H_1. The researcher can then determine the minimum effect that will yield the lowest acceptable statistical power, for a given n.

$$
\begin{aligned}
\beta &= Prob(\text{fail to reject} H_0 | H_1) \\
&= Prob(\hat{\theta} \leq c | H_1) \\
&= Prob([\frac{\hat{\theta} - \theta_1}{\frac{\sigma}{\sqrt{n}}}] \leq [\frac{c - \theta_1)}{\frac{\sigma}{\sqrt{n}}}]) \\
&= \Phi(\frac{c - \theta_1)}{\frac{\sigma}{\sqrt{n}}}) \implies \\
\Phi^{-1}(\beta) &= \frac{c - \theta_1}{\frac{\sigma}{\sqrt{n}}} \implies \\
B_\beta &= \frac{c - \theta_1}{\frac{\sigma}{\sqrt{n}}}
\end{aligned}
$$

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 β=0.20, then 1 - β = 0.8, and B_β=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 (in which implies she thinks that 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:

$$
\begin{aligned}
\frac{c}{\frac{\sigma}{\sqrt{n}}} &= -B_\alpha + \frac{\theta_0}{\frac{\sigma}{\sqrt{n}}} \\
\frac{c}{\frac{\sigma}{\sqrt{n}}} &= B_\beta + \frac{\theta_1}{\frac{\sigma}{\sqrt{n}}}
\end{aligned}
$$

Setting the two critical values that satisfy Type I and Type II errors, we can solve for the sample size, n:

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

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 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, or by choosing a wider distance between θ_0 and θ_1. Across disciplines, it is generally accepted to aim for a power of 0.8 (Lenth, 2001).[5]

We have presented B_α and B_β as the critical values in the standard normal distribution, and this would indeed hold if σ were known. But if σ is unknown, it is typically estimated using sample variances. As a result, the critical values associated with B_α and B_β will be taken from a t-distribution rather than a standard normal distribution.

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

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 an effect size must be chosen (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 underpowered *ex ante* to then 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

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

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 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.[6]

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 either run pilot studies to estimate μ_0 and σ (while choosing an anticipated μ_1) or they can look to past studies for these sample statistics. One caveat to the latter is that past studies' effect sizes be representative of a intervention's true effect in a population. In particular, overstated effect sizes in low powered studies should not be used to power future studies (Gelman and Carlin, 2014; Button et al., 2013; Ioannidis, 2005; Szucs and Ioannidis, 2017). As seen in Figure 1, the lower the power of a test, 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 failure 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 the publication bias.[7]

One practice that has taken hold in recent years that asks researchers to report power calculations for their study before it is conducted is pre-registration. Pre-registration essentially forces a researcher to publicize her intended hypothesis tests and the needed sample size for those tests.[8] The

[6] 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.

[7] 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.

[8] For example, AEA RCT Registry.

researcher would list any statistical software and associated commands that they used to perform the 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.

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, 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) and were unable to maintain the sample size that their original power calculations dictated (as often occurs with field studies). The researcher then wants 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 (Lenth, 2001; Hoenig and Heisey, 2001; Goodman and Berlin, 1994). 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 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 small graphical example, which we believe best exhibits the fallacy of observed power.

Take a scenario where the researcher pre-specifies an α = 0.05, and β = 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 H_1'. She fails to reject the null (top panel of Figure 2). 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 2 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.

Note that 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, α, and β from the data, which, crucially, does 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. Bayarri et al. (2016) focus on the use of Bayes factors in evaluating the pre- and post- rejection ratios of statistical tests, and is also a suggested response to improving replications of studies (Benjamin et al., 2018). Given the recent focus on this methodology, we provide examples from Benjamin et al. (2019) in the **Online Supplementary Materials**. 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, a researcher has 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 limited to using small samples, because of, for example, funding constraints or working with hard to reach 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 study. Each of these features demand a larger sample size to detect an effect with a given size, variance, α, and β.

Between-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. A between subjects design involves comparing the mean outcomes between separate samples drawn from the same population. A within study design involves comparing the mean outcome for each individual across different treatments. Intuitively, for a within study, each participant serves as their own control, whereas in a between design a large fraction of the total sample has to be designated as the control group. Bellemare et al. (2014), for example, shows that "between study designs require 4 to 8 times more subjects than a within study design to reach an acceptable level of statistical power."

The relationship between the required sample sizes for a between, N_B, versus within, N_W, design with two treatments is: $N_W = N_B \frac{(1-\rho)}{2}$, where ρ is the correlation between the outcomes under each treatment in a within-subjects design (Maxwell et al., 2004)[p. 561]. When ρ is 0, N_W is half of N_B for a given α, β, μ_0, μ_1 and σ. Because every participant can provide two data points in the within design, the between design will need twice as many participants. And when ρ is positive, the correlation between a subject's outcomes across treatments further reduces the sample size needed in a within subject design.

While within-subject designs provide more power than between designs, 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 would like 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 optimises 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 for a treatment and control group with unequal variances (their equations 6 and 7).[9] Another promising area of research with regards to between designs with multiple treatments are 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 readers want to examine power for multiple hypotheses. Power calculations should account for the number of hypothesis tests that will be conducted on an outcome variable. As the number of hypothesis tests increases the probability that one of them will be significant rapidly increases. After M independent tests, the probability of making at least one type I error in M tests is $1-(1-\alpha)^M$.[10] Thus, after 50 tests, and 5% significance, the probability of falsely rejecting the null is already 92%. There are two paths to account for multiple testing: by adjusting the power calculations before the experiment, or after the experiment by adjusting the Type I error rate.[11] Our setup here focuses is on a single outcome hypothesis test. For more on power under multiple hypotheses, see Lin et al. (2010), who present methods for calculating sample size to detect a specified proportion of effects.

6 Computing power

Section 2 demonstrated the derivation of sample size or power for the most basic types of hypotheses. For an overview of performing power calculations more generally 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.

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 sample size between treatment to control groups. Stata includes the commands *power*,[12] *sampsi* and *sampclus*. Note, importantly, that Stata's default output is framed in terms of composite tests $H_0 : \theta = \theta_0$

[9] We provide an example with multiple treatment arms and different variances in Section 6.

[10] 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 - $(1-\alpha)^m$.

[11] A Bonferroni correction accounts for multiple testing after the experiment is conducted at the hypothesis testing stage. More recently, List et al. (2019) provide a new correction for multiple hypothesis testing that outperforms the Bonferroni correction in terms of power.

[12] A one sample, two-sided sample size calculation with α = 0.05, $1-\beta$ = 0.8, $\mu_0 - \mu_1$ = 2 - 2.5 and σ = 0.8 is **power onemean 2 2.5, sd(0.8)**; A two sample, two-sided mean test where $\mu_0 - \mu_1$ = 12 - 15, σ_0 = 5, σ_1 = 7 and the treatment groups is twice the control group is **power twomeans 12 15, sd1(5) sd2(7) nratio(2)**

versus $H_1 : \theta \neq \theta_0$, but the command does require the user to specify a specific null and a specific alternative. The two-sided aspect is reflected only in the critical value. To calculate power for a range of alternatives, a power function is necessary, which Stata's *power* command can produce both in table and graph form for both one-sided and two sided tests. Open source languages such as R and Python also have their respective *power* libraries.[13]

Another useful tool for power analysis is the J-PAL software Optimal Design (OD).[14] 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. 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 of values.

OD has two design options that would be most useful to JESA readers, both are under the option "person randomized trials." These are single level trials (between subject design) and repeated measures (between subject design with multiple observations per person, or within design with only one treatment). To use the single level trial feature is straight forward. 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). So with a within subject design with one control observation and 1 treatment, F = 1, D = 1, M = (1 + 1). OD does not accommodate designs where each subject is subject to two or more treatments.

In the way of non-normal distributions there are several considerations. Certain non-normal distributions may have closed form solutions for power calculations such as skewed distributions (Cundill and Alexander, 2015) or chi-squared distributions (Guenther, 1977). However, 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 size n following the distri-

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

[14]The software is free and available here https://www.povertyactionlab.org/research-resources/software-and-tools. The user guide has straightforward tutorials.

bution specified under H_1 and compares the two samples k times using their 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.[15] 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. 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. Pilot studies or pre-intervention surveys can also 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, so discovering an effect of any size may be sufficient to meet research goals. In that case, power should still be consulted, and calculated over a range of effect sizes, to 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.

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 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).

[15] 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)

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 (List et al., 2011). Rules of thumb are not without statistical underpinning (Berenson et al., 1988)[pg 227], 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.

Sharing details of power calculations will help the profession to develop accepted standards for how inputs (i.e. standard error estimates) should be decided in the absence of empirically motivated and context specific priors. We reiterate that the gains to calculating power outweigh the (small) effort of using them over the rules-of-thumb that currently pervade the experimental literature.

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

Athey, S. and G. W. Imbens (2016). The Econometrics of Randomized Experiments. *Arxiv*.

Bayarri, M. J., D. J. Benjamin, J. O. Berger, and T. M. Sellke (2016). Rejection odds and rejection ratios : A proposal for statistical practice in testing hypotheses . *Journal of Mathematical Psychology 72*, 90–103.

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., C. Camerer, and N. Vesterlund (2019). Panel discussion on research methods in experimental economics. Experimental Science Association, Los Angeles, CA.

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.

Berenson, M. L., D. M. Levine, and D. Rindskopf (1988). *Applied statistics: a first course*. Englewood Cliffs: Prentice Hall.

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.

Cundill, B. and N. D. E. Alexander (2015). Sample size calculations for skewed distributions. *BMC medical research methodology 15*(28).

Czibor, E., D. Jimenez-gomez, and J. A. List (2019). The Dozen Things Experimental Economists Should Do (More of) . *Working Paper*, 1–74.

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.

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.

Gerber, A. S. and D. P. Green (2012). *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.

Guenther, W. (1977). Power and Sample Size for Approximate Chi-Square Tests. *The American Statistician 31*(2), 83–85.

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).

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.

Lin, W.-J., H.-M. Hsueh, and J. J. Chen (2010). Power and sample size estimation in microarray studies. *Bioinformatics 11*(48).

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

List, J. a., 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.

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

Maxwell, S., H. Delaney, and K. Kelley (2004). *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). Editors' 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*.

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

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.

Xiong, R., S. Athey, M. Bayati, and G. Imbens (2019). Optimal experimental design for staggered rollouts.

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.

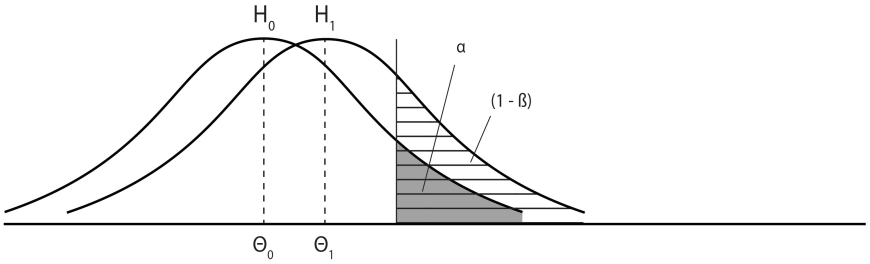
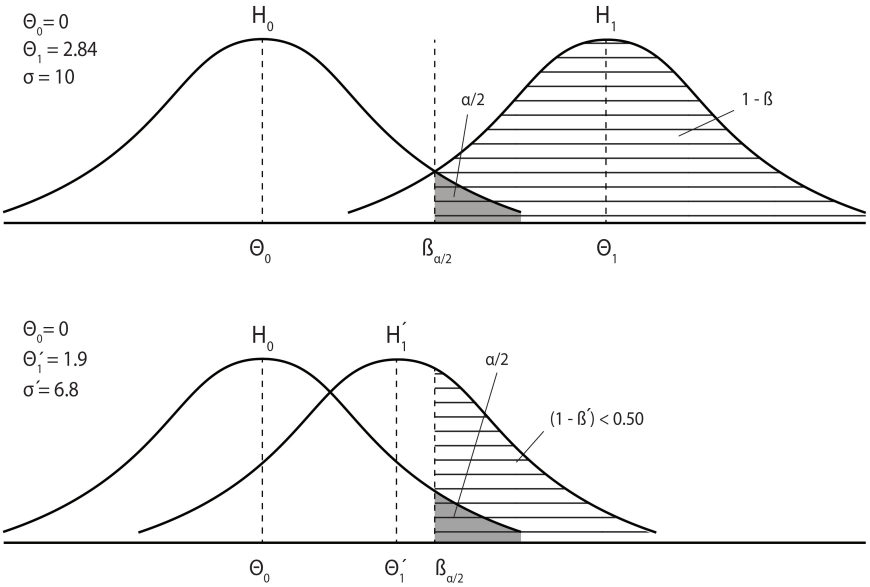


Figure 1: Overstated Effect Size in Underpowered Studies



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

Figure 2: Fallacy of Ex post Power

A Appendix

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()).