Notes for Applicants on Power Analysis in Experimental and Non-Experimental Designs

Updated: April 2024[1]

Thank you for your interest in Arnold Ventures. As part of the research proposal, we ask applicants to assess power for both experimental and non-experimental studies and to clearly define the assumptions made for power calculations within their proposals. We ask for power analyses for all submitted research proposals to ensure projects will provide a fair and valid test of the program or policy in question.

We recognize that the standard approach to power analysis in experiments (RCTs) is not transferable to non-experimental research designs used for causal analysis. Below we provide power analysis guidance for both experimental and non-experimental research projects.

When completing required power calculations, applicants are neither required to follow this guidance, nor do we expect applicants to solely use the outlined approaches to power calculations. Rather, this document is meant to serve as a reference and resource for applicants. If you have additional questions about your proposal, please contact the lead reviewer.

Guidance for Proposals

Within your full proposal submission, we ask that researchers include the following details from power analyses:

1. Target sample size
2. Minimal detectable effect size (MDE)
3. Anticipated effect size in real-world terms (e.g. dollars) and rationale for why this is the anticipated effect
4. Assumptions made in power calculations: fixed significance level, power level, variance of outcome variable, treatment allocation, intra-cluster correlation coefficient (if appropriate)
5. Discussion whether the study will use clustering, pooled, interaction effects, serial correlation, or hierarchical structures; the impact of this approach to power calculations; and an explanation for the methodological approach

These details can be included in the body of the proposal or in an Appendix.

Power Analysis Guidance: Experimental Designs

In a simple randomized experiment in which the outcome variable is normally distributed, has known variance σ^2, and has mean 0 under the null hypothesis, the following formula shows how to determine the sample size (N) needed to detect a treatment effect of size *MDE* with power $1 - \kappa$ (typically chosen to be 80%) using a two-sided test at significance level α, (typically chosen to be 5%):[2]

$$N = (z_{1-\kappa} + z_{1-\frac{\alpha}{2}})^2 \; \frac{1}{P(1-P)} \; x \; \frac{\sigma^2}{MDE^2}$$

Here, P is the pre-determined share of sample that is randomly allocated to treatment, and $z_{(1-\kappa)}$ and $z_{\alpha/2}$ are the $(1 - \kappa)$ and $\left(1 - \frac{\alpha}{2}\right)$ quantiles of the standard normal distribution (e.g., $z_{0.8}$=0.84 and $z_{0.975}$=1.96).[3] This equation can also be expressed as the *MDE* (minimal detectable effect) for a given sample size of *N*. Researchers can determine a reasonable effect size range based on a thorough review of related studies ideally with a similar context and research question. We recommend using a conservative MDE—one on the smaller end of plausible or previously estimated treatment effect magnitudes—to ensure the proposed study has sufficient power to detect plausible treatment effect sizes. Another reasonable approach would be to consider a range of values of MDE, if there is not an obvious focal value of interest.

A range of estimates is often not available for novel research questions, so researchers should also consider whether the size of the MDE is reasonable considering the role of other treatments/determinants of the outcome variable of interest. In practice, this can be calculated using any statistical software. Appendix A provides links to comprehensive guides to power analysis in Stata and R, including resources for when using more complex research designs, such as when data is in a panel format and power calculations must account for serial correlation.

Power Analysis Guidance: Non-Experimental Designs

In this guide, we outline simulation-based methods that can be used to assess power in non-experimental research designs. Sample R code and links to more in-depth resources are provided in an appendix to this guide.

Across the standard quasi-experimental designs used for causal analysis (Regression Discontinuity, Difference-in-Differences, Instrumental Variables), there exist numerous components of an empirical model (e.g. large sets of covariates, fixed effects, random effects, etc.) that make it very difficult, if not impossible, to evaluate power using an extension of typical approach employed in experimental studies. To assess power for a specific non-experimental research design, we recommend a simulation-based approach, and we provide guidance below based on several useful resources.

Below are key questions and examples of how we would set up a simulated power analysis for a non-experimental research design. Below, we describe three scenarios: when the research team *does not yet have access to the data* (e.g. a research team may be seeking funding to access or collect data), when researchers *do have access to the data*, and when researchers *have access simulated data*.

Questions to consider when conducting a power analysis include:

- What is the intervention? Does it vary across time and/or individuals or groups?
- What specific variables will be included in the analysis?
- What is the estimation equation?
- Are summary statistics such as the mean and standard deviation available for key variables?
- Is there an idea of the approximate sample size available?
- What is the smallest effect expected from the intervention? Median effect? Largest effect?
- What are acceptable levels of Type I error (the probability of *falsely rejecting a true null hypothesis of no impact, known as* α) and Type II error (the probability of *failing to detect an effect when there is one*, $1 - \kappa$) for your proposed study?

A. Scenario 1: Researchers do not have the data yet.

In this scenario, the key statistics needed to conduct an informative simulation-based power analysis are the mean and standard deviation of all variables used in the primary estimation equation along with the approximate sample size expected to be obtained. For projects that plan to analyze existing data, we recommend trying to obtain key summary statistics directly from the data provider to most accurately assess power via a simulation. In situations where this is not possible, summary statistics could be available from prior research evaluating a similar population (e.g. from summary statistics tables within related papers). In some situations, assumptions about other distributional properties, such as a normal or Poisson data generating process, may also be important determinants of power; however, most measures of program effect are distributed approximately asymptotically normal, so distributional properties may be of second-order importance in these cases.

Some general considerations and examples for researchers trying to assess power without access to data include:

a) **Sample size:** Ideally, the data provider can provide this information. If not, an informed approximation may be available using publicly available data or data from prior studies that use the dataset the researcher plans to use.

   i) *For example, if a project plans to use administrative public education records in a given county, publicly available U.S. census data could be used to determine the number of school-aged children along with the share of children attending public school in the area and time period of interest.*

b) **Summary statistics for relevant variables:** We recommend requesting this information from the data provider for projects that plan to analyze data that already exists. If a prior study uses the same (or a very similar) population, researchers can likely obtain summary statistics from descriptive tables. Publicly available data may also help inform researchers.

   i) *For example, if a project involves earnings outcomes, the Bureau of Labor Statistics data that tracks average earnings by local area may be helpful. Further, information about the mean and standard deviation of demographic covariates could be available through the suite of Census datasets.*

c) **Intervention effect size:** Determining the expected effect size of an intervention is challenging; however, previous evaluations of similar interventions can provide reasonable approximations. For example, a study of the effect of eliminating money bail on a Failure to Appear (FTA) could first draw on the existing literature on that topic to determine the range of estimated effects from the literature. Based on that range, researchers could use the median estimated effect as a benchmark effect size for the proposed study and could suggest that smaller effects serve as an MDE, if the researchers believe the smaller effects would be substantively important. Where possible, we recommend putting more weight on estimates from studies with a strong causal identification strategy. If no similar interventions exist, then we recommend researchers consider studies that use different interventions but may be expected to exhibit similar treatment effects on a given outcome. Note that it is important for researchers to explicitly outline assumptions made when estimating effect sizes.

   i) *For example, if a research team proposed evaluating the impact of a midnight basketball program on arrests among juveniles and could not find causal estimates of this type of program, then estimated treatment effects on arrests for policies that expanded school athletic opportunities or those that expanded organized activities outside of school hours could be used to approximate a reasonable expected effect size.*

B. Scenario 2: Researchers have access to some or all of the data.

This scenario requires fewer assumptions about the data's properties, which allows more accurate estimates of statistical power. In this case, the actual data may be used to simulate the relevant data generating process under the null hypothesis of zero treatment effect. Researchers need only to determine the effect size as discussed in part c of the first scenario above. When researchers have access to the data, the general goal of a simulated power analysis should be to determine the smallest effect size the empirical strategy can detect at a 0.05 significance level in 80% of the simulations performed.[4] We recommend that researchers exclude post-treatment data to avoid a change in outcomes due to treatment from contaminating an ex-post power analysis.

C. Scenario 3: Researchers have access to real or simulated data.

With either the analysis data or a simulated version of analysis data, simulations to evaluate power can be implemented in a variety of statistical programs. In an appendix to this guide, sample R code is provided for a simple linear regression model with fixed effects. For alternative methodologies, the linear regression model reference in Appendix B to this guide can be replaced with the estimation model planned for the proposed study. This code demonstrates how a researcher can input effect sizes in the simulation as well as output the power for a range of sample sizes. Below are additional resources for varying types of quasi-experimental designs:

- **Difference-in-differences designs (DDs)**. An excellent guide and clear example for a simulation-based power analysis is provided for DDs in Black, Hollingsworth, Nunes, and Simon (2020).[5] Stata code and documentation for the simulated power analysis are provided by Alex Hollinsworth here. We recommend research teams adopt a similar approach for proposed quasi-experimental approaches because they can replace the DD commands in this code with the estimation model of interest, e.g., two-way fixed effects, instrumental variables, and so on.
- **Regression discontinuity designs (RDDs).** Evaluating the power of a proposed RDD brings additional considerations.
  - ○ *The design effect*. One issue is that treatment assignment is necessarily correlated with the running variable. This introduces a design effect into an RDD, as discussed by Peter Z. Schochet's 2008 Technical Methods Report on educated-related RDDs: "School sample sizes typically need to be about three to four times larger under RD than [RCT] designs to achieve impact estimates with the same levels of precision."[6]
  - ○ *Bandwidth issues*. John Deke and Lisa Dragoset's Mathematica Policy Research Working Paper (2012) point out that when researchers use optimal bandwidth selection methods, which can be important to the credibility of an RDD-based study, the result is akin to introducing clustering. The end result of their analysis is that "accounting for these necessary components of RDD impact analysis further increases the RDD impact variance to be 9 to 17 times higher than an RCT impact variance in a study with the same sample size."[7]
  - ○ Cattaneo, Titiunik and Bare (2019) provide a Stata program, *rdpower*, that proposals involving a RDD should consider using to evaluate power using either pre-treatment data or simulated data.[8] Intuitively, RDDs leverage information local to the discontinuity, so larger sample sizes are often required by RDD for inference with the same level of statistical power in randomized experiments. Their command, *rdsampsi*, uses power calculations and can also be used in situations where the data is not available to provide the minimal sample sizes needed to achieve a desired level of power when the final data will be analyzed using *rdrobust* in Stata. We recommend this be used in combination with *rdpower* on a simulated dataset to provide a comprehensive evaluation of power and the appropriate sample size.
  - ○ David McKenzie provides an overview of power analysis for RD designs in three World Bank blog posts; here are Part 1, Part 2, and Part 3.

Appendix A: Resources

Below are resources to provide additional background on power calculations in experimental and non-experimental studies.

Resources for Experimental Studies

- Resource for RCTs: The Poverty Action Lab provides a thorough discussion of power calculations for Randomized Controlled Trials (RCTs).

- Six Rules of Thumb outlines best practices for determining sample size and statistical power.
- Comprehensive guide to power analysis in Stata
- Comprehensive guide to power analysis in R
- Stata package, pcpanel, for when data is in a panel format and power calculations must account for serial correlation

Resources for Non-Experimental Studies

- Illustrative power analysis for staggered difference-in-differences designs
- Six Questions about doing Power Calculations provides brief responses to frequently asked questions when conducted power analysis for non-experimental designs
- Several in-depth guides and examples for simulation-based power analysis are available online for both Stata and R:
  - *Stata: Power by Simulation*
  - *Power analysis by data simulation in R (Julian Quandt)*
  - *Power simulations in R (Cameron Raymond)*
  - *Using R for simulation (Charles DiMaggio and Steve Mooney)*
  - *Simulation and power analysis in R (Matthew Crump)*
- There are also some power-related resources in the Python *statsmodels* package; see Power and Sample Size Calculations.

Appendix B: R Code Example

Loads Packages and installs if needed

```
if (!require(“pacman”)) install.packages(“pacman”)
pacman::p_load(faux, pwr, simstudy, lfe, ggplot2)
```

The simulation approach to power analysis involves these steps:

1. Use R to sample numbers into each condition of any design.
2. You can set the properties (e.g., n, mean, sd, kind of distribution etc.) of each sample in each condition, and mimic any type of expected pattern
3. Analyze the simulated data to obtain a p-value (use any analysis appropriate to the design)
4. Repeat many times, save the p-values
5. Compute power by determining the proportion of simulated p-values that are less than your alpha criterion.

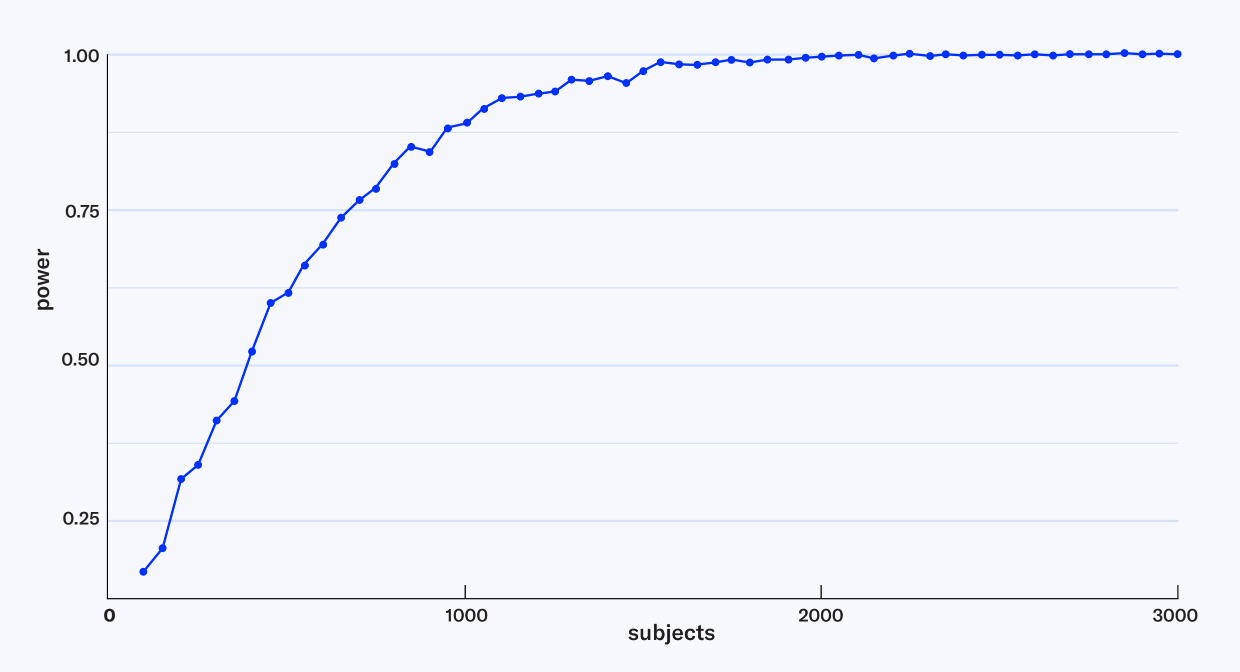
For all simulations, increasing the number of simulations will improve the accuracy of your results. Ideally use at least 1000 simulations.

Scenario 1) No data available but know distribution (mean,sd) of variables

```
#Make sure to set seed in order to reproduce results
set.seed(123)
```

```
#Creating a function to conduct power calculation for variations in sample size x
sim_power_reg <- function(x){

  #Create Normalized dependent variable. One could also use a non-normalized dependent variable
  def <- defData(varname = “dv”, dist = “normal”, formula = 0,
          variance = 1)

  #Build in effect size here in the formula input.
  #This example is for effect size of 0.2 std dev but one could also specify a specific size based
  # on raw numbers by simply adding amount to non-normalized dv
  def <- defData(def, varname = “treat”, dist = “binary”,
          formula = “dv*0.2”, link = “logit”)

  #Random Covariate with some correlation to treatment
   def <- defData(def, varname = “cov”, dist = “poisson”,
          formula = “dv*0.05 - 0.1*treat”, link = “log”)

  #Generate a Fixed Effect
   def <- defData(def, varname = “fe”, dist = “categorical”,
          formula = “0.5;0.1;0.1;0.1;0.05;0.05;0.025;0.025;0.025;0.025”, link = “log”)

  #Creates dataset of size x
  dd <- genData(x, def)

  #OLS model with fixed effects
  fit1 <- felm(dv ~ treat + cov | fe | 0 | 0 , dd )

  #Grabs the p-value from coefficient of the treatment variable
  reg_results <- summary(fit1)$coefficients[1,4]

  #return the pvalue
  return(reg_results)
}

#Vector of sample sizes
subjects <- seq(100,3000,50)

#Run simulation for each sample size 1000 times for 0.05 significance level.
```

```
power <- sapply(subjects,
            FUN = function(x) {
              sims <- replicate(1000,sim_power_reg(x))
              sim_power <- length(sims[sims<.05])/length(sims)
              return(sim_power)})

#Combine into dataframe
plot_df <- data.frame(subjects,power)

#Plot the power curve. Standard convention for minimum sample size needed for a study is #based on sample size in
this figure or dataset which generates a power of 0.8, but each #researcher should determine the power they would like
to implement in their study.

ggplot(plot_df, aes(x=subjects,
               y=power))+
  geom_point()+
  geom_line()
ggsave("PowerFigureFakeData20percent.pdf", height=6, width=9, dpi=600, device = cairo_pdf)
```

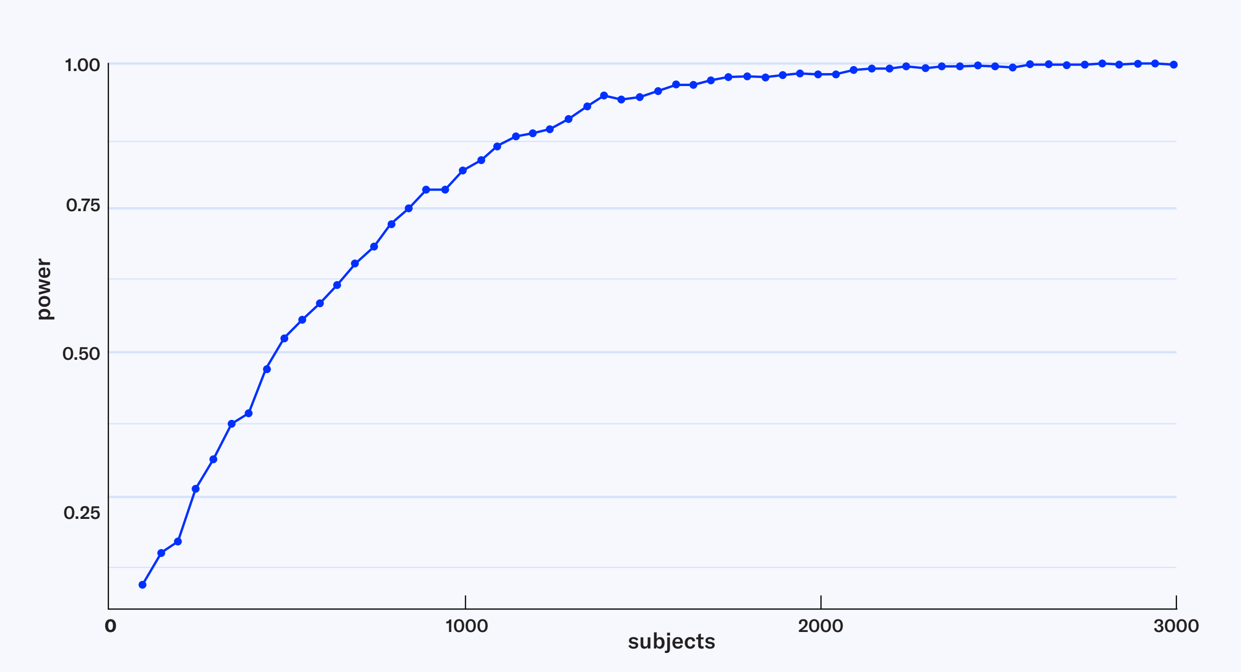


Scenario 2) Have some or all data available and know distribution (mean,sd) of any missing variables

```
#Example using R built-in dataset.
#What would happen to the number of home runs if the MLB decided to
#randomly move outfield fences closer in half of baseball parks?
```

```
#Expect a 20% increase in HRs in those ballparks

#baseball is built in dataset for R and provides player level statistics across a number of #seasons

baseball <- baseball %>% select(id,year,team,lg,g:so) %>% na.omit()

#Make sure to set seed in order to reproduce results
set.seed(123)

#Create random treatment variable. Could define this based on actual treatment assignment
#mechanism which may not be random

def <- defData(varname = “treat”, dist = “binary”,
        formula = -1, link = “logit”)

#Function that combines randomly generated treatment with actual data and provides power
#calculation for different sample sizes x

sim_power_reg <- function(x){
 policy <- genData(x, def) %>% select(treat)

#Randomly sample dataset of size x
 ball <- baseball[sample(nrow(baseball), x), ]
 ball <- cbind(ball,policy) %>% select(id,hr,treat,ab,team,year)

#Build in a 20% reduction in HRs for ballparks treated with closer fences
 ball <- ball %>% mutate(hr = ifelse(treat==1,(hr*0.8),hr))

#Empirical Model
 fit1 <- felm(hr ~ treat + ab | year | 0 |  year   , ball)
 reg_results <- summary(fit1)$coefficients[1,4]

 #return the pvalue
 return(reg_results)
}
# vector of sample sizes
subjects <- seq(100,3000,50)
```

```
# run simulation for each sample size 1000 times for 0.05 significance level.
power <- sapply(subjects,
            FUN = function(x) {
              sims <- replicate(1000,sim_power_reg(x))
              sim_power <- length(sims[sims<.05])/length(sims)
              return(sim_power)})


# combine into dataframe
plot_df <- data.frame(subjects,power)


# plot the power curve. Standard convention for minimum sample sized needed is based on #sample size in this figure or dataset which generates a power of 0.8, but each researcher
# should determine the power they would like to implement in this study.


ggplot(plot_df, aes(x=subjects,
               y=power))+
  geom_point()+
  geom_line()
ggsave("PowerFigureActualData20percent.pdf", height=6, width=9, dpi=600, device = cairo_pdf)
```



Endnotes

1 Prepared by Arnold Ventures' Evidence and Evaluation team in collaboration with Stephen B. Billings, Jonah B. Gelbach, and Kevin T. Schnepel.

2 This formula is based on the use of a *Z*-statistic, which is feasible when the variance, σ^2, is known. In two-sided testing cases, the formula is an approximation due to its omission of a small additional term; see the discussion surrounding equation (2.13) on page 31 of Xiaofeng Steven Liu, *Statistical Power Analysis for the Social and Behavioral Sciences*. Routledge, 2014. When a one-sided test is appropriate, the formula in the text holds exactly, provided that $z_{1-\alpha/2}$ is replaced with $z_{1-\alpha}$.

3 In practice, the standard deviation σ usually must be estimated, and a *T*-statistic rather than a *Z*-statistic is used to test the null hypothesis. This means that the exact formula would use critical values of the Student's *t* distribution, rather than the standard normal distribution. However, the appropriate Student's *t* critical values are a function of the sample size due to the degrees-of-freedom parameterization of the Student's *t* distribution. Using these critical values in place of $z_{1-\kappa}$ and $z_{1-\alpha/2}$ (would make the right-hand side of the formula in the text a function of sample size, making implicit-function solution methods necessary. A further complication arises in the commonly encountered case when the share of the sample allocated to the treatment group is random (as would happen if assignment were determined randomly at the unit level, e.g., with a coin toss); in that case, *P* is random rather than fixed, so that even with Student's *t* critical values, the formula would be correct only conditional on the realized allocation of observations across the treatment and control groups. As long as the sample size is reasonably large, and as long as the probability of treatment assignment is not too close to 0 or 1, the formula in the text—using standard normal critical values, with (i) the observed share of treated observations as *P* and (ii) a consistent estimate, $\hat{\sigma}$, used in place of σ—will yield a suitable approximation to the required sample size, thanks to the continuous mapping theorem.

4 The 80% figure is a conventional value rather than one written in stone. Researchers could use a different value with a reasonable explanation.

5 Bernard Black, Alex Hollingsworth, Letícia Nunes, and Kosali Simon, "Simulated Power Analyses for Observational Studies: An Application to the Affordable Care Act Medicaid Expansion." *Journal of Public Economics* 213 (2022): 104713.

6 Peter Z. Schochet, "Technical Methods Report: Statistical Power for Regression Discontinuity Designs in Education Evaluations," NCEE 2008-4026 (August 2008).

7 John Deke and Lisa Dragoset, "Statistical Power for Regression Discontinuity Designs in Education: Empirical Estimates of Design Effects Relative to Randomized Controlled Trials," Mathematica Policy Research Working Paper (June 2012).

8 Matias D. Cattaneo, Rocío Titiunik,and Gonzalo Vazquez-Bare, "Power calculations for regression-discontinuity designs." The Stata Journal, 19(1), 210-245. (2019).