Lab 7: Power and Sample Size

PHS2000B, Spring 2023

Susan Rattigan and Sudipta Saha

March 5, 2023

Learning Objectives

1. **Why Power and Sample Size?**
2. **Type I Error, Type II Error, Power**
3. **Calculating Sample Size**
4. **Estimating Outcome Variance & Effect Size**
5. **Calculations Based on Precision (CI Width)**
6. **Other Important Considerations**

Power and Sample Size

During the **design phase** of a study, there are two key questions we may wish to answer:

1. What is the required **sample size *N*** to identify some minimal detectable effect size η with a given power?
2. How much statistical **power** would I expect to have if I wanted to identify a minimal detectable effect size η with a fixed sample size N?[1]

Answering these questions can help 1) **define the number of individuals to be recruited** into a study and 2) **assess how much statistical power (or precision)** an analysis of *existing data* would be expected to have.

[1] Calculations are less relevant when the sample size is fixed but can be useful to investigate the extent to which we may be worried about false positives. Some have argued we shouldn't be performing power calculations for secondary data analyses: Hernán MA. Causal analyses of existing databases: no power calculations required. J Clin Epidemiol. 2021;S0895-4356(21)00273-0.

Type I Error, Type II Error, Power

Type I Error, Type II Error, Power

- Type I Error: $\alpha = Pr(\text{Reject } H_0 | H_0 \text{ is true})$
- Type II Error: $\beta = Pr(\text{Fail to reject } H_0 | H_1 \text{ is true})$
- Power: $1 - \beta = Pr(\text{Reject } H_0 | H_1 \text{ is true})$

	H_0 is True	H_1 is True
Reject H_0	Type I Error (α)	Correct! ($1 - \beta$)
Fail to Reject H_0	Correct! ($1 - \alpha$)	Type II Error (β)

Type I Error, Type II Error, Power

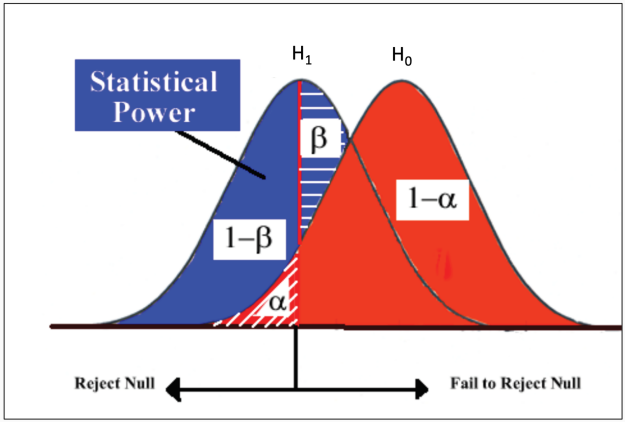


Figure 1: There is a trade-off between Type I and Type II Errors.

Calculating Sample Size

The most popular methods for **calculating sample size** rely on pre-specifying the desired statistical power. The sample size is affected by the following quantities:

1. The desired **statistical test** for the parameter of interest, including the underlying probability model
2. The acceptable **probability of Type I Error**
3. The acceptable level of **statistical power**
4. The **variance of the outcome** of interest in the population
5. The **true effect size**

Calculating Sample Size: Example

Let's estimate the required sample size for a **randomized trial** with **continuous outcome** and **equally sized groups.**

- To compute the minimum sample size N for each group to detect an effect size η using a t-test with a predetermined value of α and β and with population outcome variance V, which is assumed to be equal in the treatment and control groups, we can use the following formula:

Sample Size: T-test with 2 Equally Sized Groups

$$N = (Z_{1-\alpha/2} + Z_{1-\beta})^2 \times \frac{V}{\eta^2}$$

Where $V = 2\sigma^2$

Calculating Sample Size: Example

We are interested in designing a randomized trial for a new **hospital-based intervention to reduce postpartum depression symptoms.** We will compare the effect of intervention vs. standard of care on postpartum depression score. If we assume:

- $\alpha = 0.05 \implies Z_{1-\alpha/2} = 1.96$
- $1 - \beta = 0.80 \implies Z_{1-\beta} = 0.84$
- $\sigma = 20 \implies V = 2\sigma^2 = 2(20^2) = 800$
- $\eta = 10$

Then:

$$\begin{aligned} N &= (1.96 + 0.84)^2 \times \frac{800}{10^2} \\ &= 62.72 \end{aligned}$$

Under these assumptions, we would need to include **63 people** in each study arm in order to have 80% power to detect a 10 point difference in postpartum depression score.

Estimating The Variance of The Outcome

Several strategies exist for estimating the **variance of the outcome:**

1. Search the **literature**
2. Use **approximate methods** for estimating the variance
3. Conduct a **pilot study**
4. Make a **well-informed guess**
5. Consider **multiple plausible scenarios**

Lenth[2] recommends comparing the variance of the outcome that was observed at the end of your study with the value of the variance you used in calculating the required sample size to help with planning future studies.

[2] Lenth. Some Practical Guidelines for Effective Sample Size Determination. The American Statistician. 2001;55(3):187 - 193.

Estimating The True Effect Size

To estimate the **true effect size,** consider:

1. Setting the true effect size equal to a value considered to be **substantively important** (prone to investigator biases!).
2. Using **existing literature or preliminary data** (e.g., from a pilot study) to estimate a **plausible effect size** (not as easy as one might imagine, in part because the literature is full of low powered studies!).

Dangers of Statistical Significance Testing

Sample size and power calculations have been heavily influenced by a **statistical significance testing framework.** Unfortunately, this framework has led to many problems.

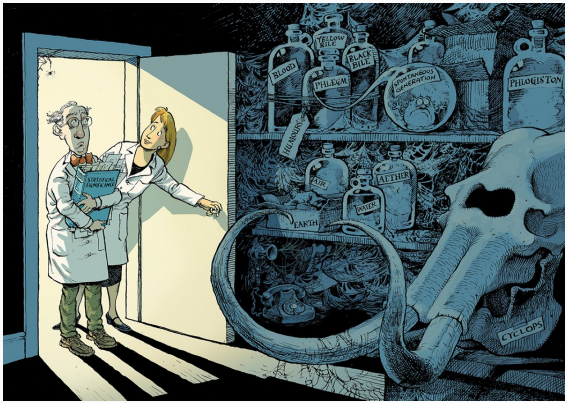


Figure 2: Scientists rise up against statistical significance. Nature. 2019;567(7748):305-307.

Dangers of Statistical Significance Testing

Dangers of statistical significance testing include:[3] [4]

- **"Dichotomania."** Rather than interpreting study results quantitatively, researchers dichotomize results as *"statistically significant"* or *"not statistically significant"* based on a p-value cutoff.
- **Overstated claims** that significant results "prove" the alternative hypothesis, or that non-significant results with $p>0.05$ "prove" the null hypothesis ("no difference").
- **Claims of "conflicting" study results** when, in fact, no conflict exists.

[3] Greenland S, Senn SJ, Rothman KJ, et al. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. Eur J Epidemiol. 2016;31(4):337-350. doi:10.1007/s10654-016-0149-3.

[4] Amrhein V, Greenland S, McShane B. Scientists rise up against statistical significance. Nature. 2019;567(7748):305-307.

Are These Study Results Conflicting?

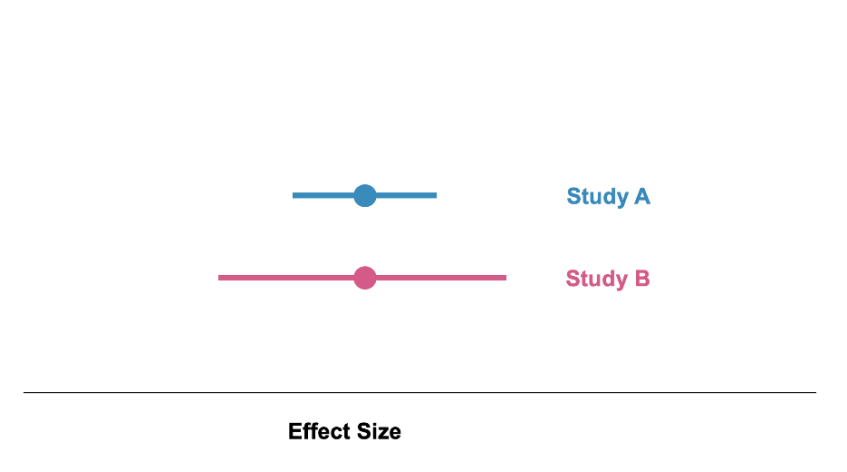


Figure 3: Scientists rise up against statistical significance. Nature. 2019;567(7748):305-307.

Are These Study Results Conflicting?

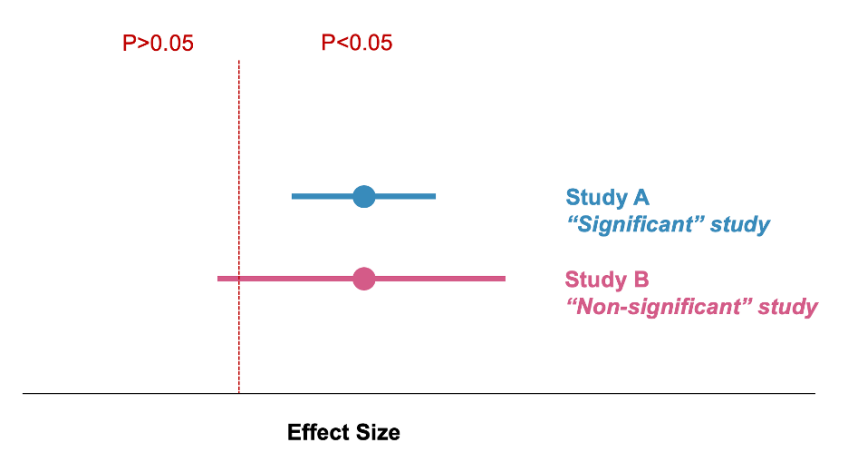


Figure 4: Scientists rise up against statistical significance. Nature. 2019;567(7748):305-307.

Other Important Considerations for Power and Sample Size

- In general the **sample size to detect interaction** is approximately **4x** the sample size required to detect a main effect of the same magnitude.
- **Power to detect indirect effect** is often (but not always!) greater than the power to detect a total effect, **while power to detect a direct effect** is lower than to detect a total effect.
- **Correlations in cluster randomized trials** need to be accounted for in study planning (e.g., by estimating the effective sample size, which depends on within-cluster correlation and average cluster size).
- **Control for covariates** affects power and sample size. In a randomized trial, controlling for covariates which predict the outcome **increases power.** In an observational study, controlling for covariates can result in either an increase or a decrease in power, but **often power decreases** when many covariates are controlled for.

Supplemental Slides

An Alternative: Calculations Based on Precision

An alternative approach which does not directly rely on a statistical significance testing framework is to perform calculations based on the **desired precision.**[5]

- We may design our study with the aim of obtaining a **desired confidence interval width** for the targeted effect.
- Example: Say we are interested in estimating disparities in risk of postpartum depression by citizenship status. We would like to be able to estimate a risk difference comparing non-citizens vs. citizens with a 95% confidence interval width of 0.1 (that is, a confidence interval that is equal to the point estimate +/- 0.05).

[5] Rothman KJ, Greenland S. Planning Study Size Based on Precision Rather Than Power. Epidemiology. 2018;29(5):599-603.

Example: Calculating Sample Size Based on Precision

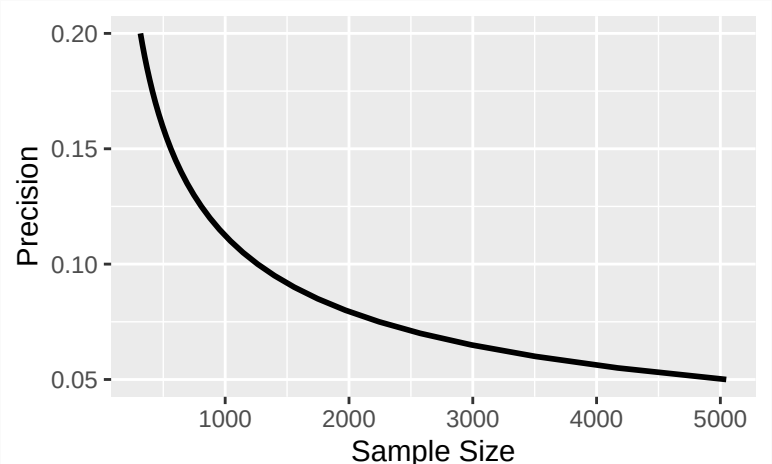
These calculations could be performed in the `precisely` package in `R`:

```
library(precisely)
n_risk_difference(precision = .1,
                  exposed = .3,
                  unexposed = .15,
                  group_ratio = 2,
                  ci = .95)
```

```
## # A tibble: 1 x 9
##   n_exposed n_unexposed n_total risk_dif~1 preci~2 exposed unexp~3 group~4    c
##       <dbl>       <dbl>   <dbl>      <dbl>   <dbl>   <dbl>   <dbl>   <dbl> <dbl
## 1      421.        841.   1262.       0.15     0.1     0.3    0.15       2  0.9
## # ... with abbreviated variable names 1: risk_difference, 2: precision,
## #   3: unexposed, 4: group_ratio
```

Example: Calculating Sample Size Based on Precision

```
map_precisely(n_risk_difference, precision = seq(from = .05, to = .20, by = .005)
exposed = .3, unexposed = .15, group_ratio = 2, ci = .95) %>% plot_sample_size()
```



Additional Resources

There are many **resources** for conducting sample size and power calculations. Here are a few R-based resources that may be helpful:

- Sample size and power calculations for **simple statistical tests** (e.g., chi-squared, t-test): pwr package.
- Simulation-based sample size and power calculations for **survival quantities** such as log-rank test or restricted mean survival: npsurvSS.
- Simulation-based power calculations for **generalized linear mixed effect models**: SIMR package.

In addition, the Open Science Foundation provides an online platform which promotes replication by providing tools to organize and share study protocols, analytic code, results, etc.