# Power Analysis with examples in R

## Session I

Xiaomeng Li, Katherine Zavez & Tim Moore

October 21, 2020

Statistical Consulting Services

Introduction and Overview

Sample size calculation in R

# Why should I care about sample size and power?

# Why should I care about sample size and power?

As a researcher, you want answers to interesting questions, by way of detecting important effects. So you conduct an experiment to test if an effect is significant. For example, you might be interested in if:

- Drivers who text and drive perform worse on a driving test than drivers who do not text and drive.
- The proportion of registered Democrats who support gun control laws is lower than the proportion of registered Republicans who support gun control laws.

To detect important effects reliably, your hypothesis test needs to have enough power. . . you need a large enough sample size!

# What does a power analysis tell me?

Power is the probability of detecting an effect in your sample, given that the effect is really there in the population. An a priori power analysis can help you:

- Determine the minimum number of subjects (samples) you need for your experiment to detect a minimum effect size.
- Prevent from "overshooting", as recruiting subjects are expensive!
- Improve the reliability of the results from your experiment.

There are serious consequences to having underpowered studies. We'll talk about these later on.

# Important Terms

To speak about power, we need to learn a few terms:

- Null and Alternative Hypotheses
- Type I and Type II error rates
- Power
- Significance Level
- Effect Size

Let's quickly run through this list!

# Null Hypothesis and Alternative Hypothesis

The null hypothesis ($H_0$) is often the “no effect” hypothesis.

The alternative hypothesis ($H_a$) is the hypothesis of interest to us. If the effect is real, we would like to reject $H_0$ in favor of $H_a$.

# Type I and Type II error rates

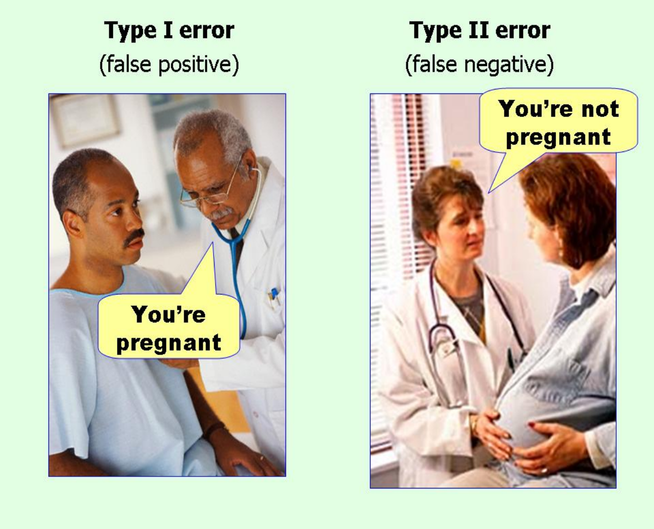
Keep in mind that reality could be different from the outcome of a hypothesis test!

| | $H_0$ True | $H_0$ False |
|---|---|---|
| **Reject $H_0$** | Type I Error | Correct Rejection |
| **Fail to Reject $H_0$** | Correct Decision | Type II Error |

Type I error = false positive and Type II error = false negative.

Type I error rate = P(making a Type I error) = $\alpha$
Type II error rate = P(making a Type II error) = $\beta$

# Type I and Type II error rates



# Power

Power is defined as the probability of correctly rejecting a false null hypothesis.

power = P(Reject $H_0$ | $H_0$ is false)

= 1 - P(Fail to reject $H_0$ | $H_0$ is false)

= 1 - $\beta$

Thus power equals one minus the Type II error rate!

# Significance Level ($\alpha$)

For the purposes of power analysis, we are going to define the significance level simply as the maximum Type I error rate that we will tolerate.

We will use the same symbol $\alpha$ to denote the significance level.

# Effect Size

Intuitively, the concept of effect size is:

- Magnitude of the effect of interest
- Measure of how far the alternative hypothesis (effect exists) is from the null hypothesis (effect doesn't exist)

But the devil is in the details!

- The precise definition of effect size is specific to the study design, analysis endpoint and statistical model/test.
- Even if all of the above are specified, altenative effect size measures exist!

# Important Terms

Here's a recap of the terms we've just discussed:

- Null and Alternative Hypotheses
- Type I and Type II error rates
- Power
- Significance Level
- Effect Size

Now we're going to understand the relationship between power and these terms.

# Factors that affect power

Power, effect size, sample size and the level of significance are all iterrelated:

- If you fix three of them, the fourth is completely determined (Cohen, 1988).
- This is important because by increasing one of them, you can increase(or decrease) another, holding the others fixed.
- We're going to focus on fixing power, effect size and the level of significance, and computing the resulting sample size, i.e, sample size determination.

Let's look at some guidelines on setting the effect size, the significance level and the power, and see how they affect the required sample size.

# How to choose effect size

It's important for you to decide what effect size is clinically relevant.
NOTE: the effect size you specify in the power analysis is the minimum effect size that can be detected reliably.

If you don't know relevant effect size, you could use one of these approaches:

- Pilot Study: can be useful to get a rough estimate
- Literature Review: effect size from similar research
- Cohen's recommendations: Cohen (1988) gave guidelines for large, medium and small effect sizes for various problem types. A last resort only!

The larger the effect size, the lower the required sample size.

# How to choose power

Once you've settled on a minimum effect size, you should now ask yourself "What's the minimum level of sureity I need that my experiment will detect an existing effect?"

Why .80?

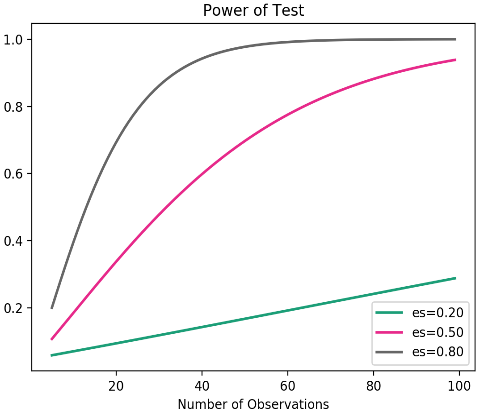
- It's the "default" standard
- Suitable for studies that are important, but not life-and-death decisions

Why .90?

- Used for studies with life-and-death implications (Think clinical trials for life-saving drugs)

The higher the required power, the larger the required sample size.

# Power, effect size and sample size of T-test



# How to choose the level of significance ($\alpha$)

Now the question to ask yourself is "What the maximum rate of type I errors that is acceptable to me?". This determines your $\alpha$. The criteria for selecting an appropriate $\alpha$ usually boil down to:

- What is the norm in your field of study?
- How costly are type I errors?
  - For very costly situations such as medical research, set $\alpha$=.01 or lower.
  - For more relaxed situations such as marketing, you might be able to afford a higher $\alpha$ such as .10 or even higher.
  - Setting $\alpha$=.05 is the "defacto" standard, but this really shouldn't be done without careful thought.

The lower the $\alpha$, the greater the required sample size.

# Other factors that affect the required sample size

Besides the effect size, power and the level of significance, there are other design-related factors that can affect how many samples you need:

- One- versus two-tailed hypothesis tests
- Experimental design (some designs are more efficient)
- Type of response variable (continuous, binary, count)
- Statistical procedure (parametric needs less samples than non-parametric)
- Number of primary hypotheses to test (more hypotheses need more samples)
- Missing data (dropouts, inclusion/exclusion criteria)

# Consequences of low power

Low powered studies make interpretation of both positive and negative results difficult to interpret, and can easily lead to wasted resources:

- False negatives: Missing real effects
- Low positive predictive value: Even your significant results are suspect!
- Inflated effect sizes: Your estimates of signifiant effect sizes is an overestimate!

# Sounds complicated? We're here to help!

We can help you:

- Select an efficient design
- Compute the minimum number of samples you'll need without losing power
- Help you write a grant proposal
- Analyze the data once you collect it
- Interpret and present the results

If you have a statistical problem (power-related or otherwise!), check out our services at: statsconsulting.uconn.edu

SCS
Statistical Consulting Services

# Install pwr package

```r
install.packages("pwr")

# load into current session
library("pwr")
```

# Working Example

We suspect we have a loaded coin that lands heads 75% of the time instead of the expected 50%. We wish to test if this coin is biased.

Hypotheses

- Our null hypothesis ($H_0$) is that the coin is fair and lands heads 50% of the time (prop. of heads $= 0.5$).
- Our alternative hypothesis ($H_a$) is that the coin is loaded to land heads more than 50% of the time (prop. of heads $> 0.5$).

Question

- How many times should we flip the coin to have a high power, 0.80, of correctly rejecting the null of 50% heads if our coin is indeed loaded to land heads 75% of the time?