Power and Sample Size

Clayton W. Schupp

Galvanize

Overview

- Revisit Frequentist Hypothesis Testing
- Sample size, power, and effect size
- How to calculate power
- Determining sample size (n) needed for a given power
- Relation to A/B Testing

Frequentist Hypothesis Testing

	H_0 is true	H_0 is false
Accept H_0	Correct Decision ($1-\alpha$)	Type II Error (β)
Reject H_0	Type I Error (α)	Correction Decision ($1-\beta$)

Design of Experiment

When designing an experiment, you usually need to decide on what sample size you should collect

- This is especially true when it is particularly important to minimize the number of individuals selected
  - Perhaps you're proposing something that could be painful such as new surgical procedure
  - Or your experiment is very time-consuming or expensive
- Performing a power analysis can allow you to determine the sample size needed to detect a particular effect

If instead, you're sample size is predetermined, a power analysis can allow you to determine the minimum effect size detectable

Power/Sample Size Analysis

The following are the components needed to perform a power/sample size analysis

- Level of Significance, Alpha (α)
- Beta (β) and its complement power (sometimes noted π)
- Effect size
- Standard deviation
- Sample size

In general, you define 4 of these components (or a range of them) and solve for the remaining component

Type I and Type II Errors and Power

Type I Error $=$ P(Reject $H_0 | H_0$ is true)
Type II Error $=$ P(Fail to reject $H_0 | H_0$ is false)
Power $=$ P(Reject $H_0 | H_0$ is false) $\longleftarrow$ Correct Decision

	H_0 is true	H_0 is false
Accept H_0	Correct Decision ($1-\alpha$)	Type II Error (β)
Reject H_0	Type I Error (α)	Correction Decision ($1-\beta$)

Effect Size

Effect size is the difference between the null hypothesis and the alternative hypothesis that you, as the researcher, hope to detect

For applied and clinical biological research, there may be a definite effect size that you want to detect. For example, from yesterday's sprints:

- The increase of 1 point in average productivity if we replace the monitors at Company X
- The 1% lift when comparing two types of web page

Often you don't know how big a difference you're hoping to detect, a power analysis will allow you to determine the range of detectable effects for a set power level when you vary sample size

Standard Deviation and Sample Size

- Standard Deviation: measure of spread of distribution
  - More variable the measurement is, the less powerful the test all other components held constant
  - Need an estimate of the population standard deviation, which can come from a pilot experiment or the published literature
- Sample Size
- Often the goal of a power analysis is to identify the sample size needed when holding the other components fixed
- The main component the researcher actually has direct control over

Calculating Power

Example: One-sample Test of Mean

$$H_0 : \mu = \mu_0 \qquad H_1 : \mu = \mu_1 (> \mu_0)$$

Recall: Power = P(reject $H_0 | H_0$ is false) = P(reject $H_0 | H_1$ is true)

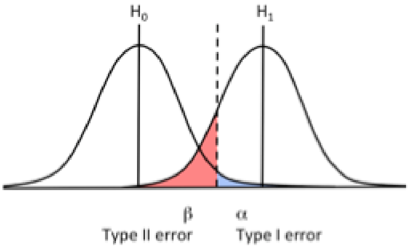
- We want to find the critical value, under H_0, beyond which we would reject H_0

$$X^* = \mu_0 + Z_\alpha \frac{s}{\sqrt{n}}$$

- Now find the corresponding probability of this value under H_1

$$Power = P(X > X^* | H_1) = P\left(Z > \frac{X^* - \mu_1}{s/\sqrt{n}} \right)$$

Calculating Power



	H_0 is true	H_0 is false
Accept H_0	Correct Decision (1-α)	Type II Error (β)
Reject H_0	Type I Error (α)	Correction Decision (1-β)

What happens to power as we adjust each of the other parameters?

α , effect size , standard deviation , sample size