Effect Size, Power, and Type I Error Correction

Paul Stey

January 14, 2019

Table of Contents

1. Introduction
2. Statistical Power
   - Overview
   - Effect Size
   - A Philosophical Note
   - Power and Sample Size
3. Type I Error Correction
   - Applications of Type I Error Correction
   - Family-Wise Error
   - False Discovery Rate
4. Post Hoc Tests with ANOVA
   - Why Post Hoc Tests?
5. Conclusion
   - Summary

Two Books

- "*Designing Experiments and Analyzing Data*", Maxwell & Delaney
- "*The Art of R Programming*", Matloff

Statistical Power (informally)

In general, the power of a statistical test refers to that test's ability to detect an effect when one exists.

Statistical Power (formally)

Power is a probability. In particular, it is the probability that you will reject H_0 when H_1 is true.

$$\text{power} = \Pr\left(\text{Reject } H_0 \mid H_1 \text{ is True}\right)$$

Effect Size

1. Quantifies the magnitude of some phenomenon being studied
2. Many different measure of effect size exist for quantifying:
    - Mean differences (e.g., Cohen's d)
    - Correlation (e.g., regression coefficients)
    - Categorical relationships (e.g., odds ratio)

Cohen's d

Cohen's d is a very commonly used measure of effect size for mean differences. It is computed using

$$d = \frac{\overline{x}_1 - \overline{x}_2}{s},$$

where s is the pooled standard deviation

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}},$$

and s_j^2 is the variance for group j.

Guidelines for Cohen's d

Effect Size	d
Very Small	0.01
Small	0.2
Medium	0.5
Large	0.8
Very Large	1.2
Huge	2.0

Effect Size in ANOVA

Most common method is η^2, which is simply

$$\eta^2 = \frac{SS_{\text{Between}}}{SS_{\text{Total}}}.$$

Effect Size in ANOVA (*cont.*)

Guidelines for η^2

Effect Size	η^2
Small	0.02
Medium	0.13
Large	0.26

Effect Size Examples in R

<EXAMPLES_IN_R>

Statistical Power (formally)

Power is influenced by 3 things:

1. Effect size
2. Sample size
3. Significance threshold (α)

Statistical Power (formally)

Demo...

Source: https://sites.berry.edu/vbissonnette/

Detour: Meehl's Paradox

George Box's famously wrote that "*All models are wrong, but some are useful*".

The notion that all models are—in a sense—wrong is generally accepted, as they are only approximations of some real phenomenon.

Detour: Meehl's Paradox (*cont.*)

Meehl begins with the notion that all models are wrong. He goes on to observe that since this is the case, as we get increasingly precise estimates by gathering more data, we should become increasingly good at demonstrating the extent to which our models are wrong. So, as we get more data, our tests ought to become increasingly stringent.

But exactly the opposite is true. As we collect more data, we conduct an increasingly lenient test of the null hypothesis. As n grows, everything is significant.

Falsification in Science

Karl Popper's philosophical views:

1. We cannot prove a hypothesis to be true
2. Hypotheses can only be disproved
3. We should focus on falsifying hypotheses
4. Those hypotheses that resist falsification remain credible

The Myth of Falsification-ism

In reality, although science has been greatly influenced by Popper, we do a bad job of adhering to his ideas.

If we did, we would be trying to falsify H_1 and not H_0. Our current system essentially sets up H_0 as a straw man.

Power and Sample Size

Relation between power and sample size:

1. As sample size increases, so does power
   - This has implications for experiment planning
   - Given desired power, and expected effect size, we can compute necessary sample size
2. Can compute post-hoc power
   - May be useful if you fail to reject H_0
   - Compute sample size needed in follow-up study

Power Analysis

Happily, for many basic statistical tests, there are simple, closed-form methods for determining the necessary sample size for a given power and effect size.

The pwr Package in R

Features of the `pwr` package in R:

1. Given a desired power and effect size, can compute sample size
2. Given sample size and effect size, can compute power
3. Works with many tests and models, including:
    - Binomial test
    - *t*-tests (one, two-sample, and paired)
    - χ^2 test
    - ANOVA
    - Correlation
    - Regression

Power and Sample Size Examples in R

<EXAMPLES_IN_R>

Simulation to Calculate Power

Once your models get more complex, you will need to use simulations to compute power.

Simulation to Calculate Power

<EXAMPLES_IN_R>

When to use Type I Error Correction

Type I error correction applies in any instances in which many statistical tests are being conducted.

Any time you find you are doing many statistical tests, consider a Type I error correction procedure

When to use Type I Error Correction

Previously discussed Type I error correction in context of ANOVA
This is mostly because multiple comparisons come up a lot with ANOVA, especially one they get complex

For a one-way ANOVA, when x has k levels, the number of comparisons is $\frac{k(k-1)}{2}$

Family-Wise Error

We discussed several Type I error correction procedures

1. Bonferroni
2. Tukey's HSD
3. Šidák

These all correct the family-wise error rate.

False Discovery Rate (FDR)

FDR is the proportion of “discoveries” (i.e., significant results) that are actually false positives.

Benjamini-Hochberg is probably the most commonly used FDR correction procedure

What is a *Post Hoc* Test? *cont.*

The key motivation for *post hoc* tests following ANOVA is to follow-up on a significant “omnibus” test.

Null Hypothesis in ANOVA

Recall that in simple ANOVA the null hypothesis is

$$H_0 : \mu_1 = \mu_2 = ... = \mu_k,$$

where k is the number of groups and μ_j is the population mean for group j. We are testing this hypothesis using $\overline{X}_1 = \overline{X}_2 = ... = \overline{X}_k$

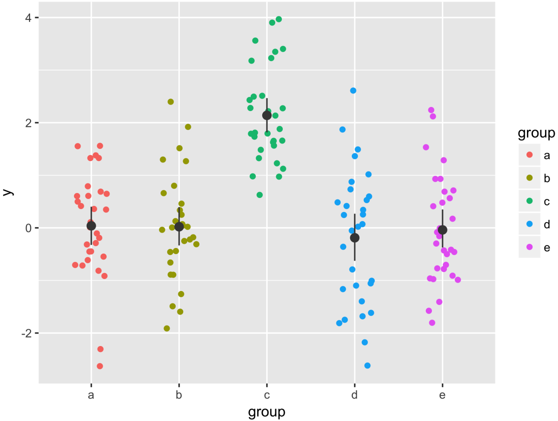
Null Hypothesis in ANOVA *cont.*

Suppose we run our ANOVA and reject H_0. That is, we reject the notion that

$$H_0 : \mu_1 = \mu_2 = ... = \mu_k.$$

But this only tells not all means are equal.

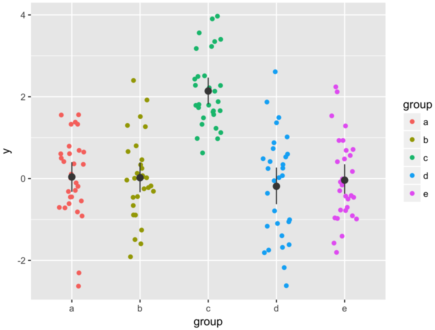
Null Hypothesis in ANOVA *cont.*



Consider the data above. This shows a significant effect from a simple ANOVA.

Beyond Omnibus Null Hypothesis in ANOVA

Suppose we are interested in a more nuanced test now. For example, suppose we want to compare every group against every other group.



ANOVA Post-Hoc Tests

<EXAMPLES_IN_R>

Summary

- Maxwell & Delaney, “*Designing Experiments and Analyzing Data*”