

Power, Effective Size, Sample Size, Type I and II Errors

How to determine sample size requirements, how to
calculate an effect size, how to assess statistical power

Nagiza F. Samatova, samatova@csc.ncsu.edu
Professor, Department of Computer Science
North Carolina State University

Learning Objectives

- Find out how big a **sample size** is needed for a **statistical test** to be conducted
- Pre-define the minimum size of the effect to be detected (i.e., **effect size**)
- Specify the required probability of detecting the effective size (**power**)
- Specify the significance level (**alpha**) at which the test will be conducted
- Correctly use the vocabulary:
 - **Effect size**
 - **Power**
 - **Significance level, alpha**
 - **Sample size**

Goal: Maximize the power of statistical tests, while maintaining an acceptable α with small sample size

- **Under more direct control:**
 - **sample size**
 - **significance level**
- **Indirect control**
 - **power**
 - **effect size**
- **Relax significance level**
 - **make it easier to reject H_0**
 - **power increases**
- **Increase sample size**
 - **power increases**

Maximize chances of finding a real effect and minimize the chances of finding an effect that isn't there, while keeping the study costs reasonable.

Sample size affects the ability to detect the real difference between treatments A and B

- **Web testing of a new feature**
 - How long should it run?
 - How many impressions per treatment are needed?
- **Will a hypothesis test reveal a difference between treatments A and B?**
- **The outcome of a hypothesis test (the p -value) depends on:**
 - what the real difference is between two treatments A and B
 - on the luck of the draw: who gets selected for the groups in the experiment
- **The real difference between treatments vs. sample size**
 - the bigger the actual difference between the two treatments, the greater the probability that the experiment will reveal it
 - the smaller the difference between treatments A and B, the more data will be needed to detect it

Power

- **Power**

- the probability of detecting a specified *effect size* with specified sample characteristics (size and variability)

- **Example**

- The probability of distinguishing between a 0.33 hitter and a 0.2 hitter in $n = 25$ at-bats is 0.75
- The effect size is a difference of $0.13 = 0.33 - 0.2$
- Detection: a hypothesis test will reject the null hypothesis of no difference and concludes that there is a real effect
 - The experiment of sample size $n = 25$ for two hitters, with an effect size of 0.13, has (hypothetical) power of 0.75 or 75%

Moving Parts for Calculating Power or Sample Size

- **Effect size you want to detect**
- **Sample size**
- **Significance level (alpha) at which:**
 - the hypothesis test is conducted or
 - the power is calculated
- **Power**

**Specify any three of the moving parts,
and the fourth can be calculated.**

Motivation

- **A/B Tests**

- Collecting and processing the data involves some cost
- Knowing approximately how much data to collect to avoid the situations where the result ends up being **inconclusive**

Sample Size

- **Applications of Power Calculations**

- To estimate how big a sample needs to be

- **Motivating Example**

- Experiment: A click-through rates (clicks as a percentage of exposure) for a new add should be tested against an existing ad
 - Policy: a new ad must do better must do better than an existing ad by some percentage (e.g., 10%)
 - otherwise, the existing ad will remain in place
 - This goal, **the “effect size” the drives the sample size**

- **How many clicks to accumulate in the study?**

- If you are interested in results that show a huge difference (e.g., 50%), then a relatively small sample might be enough
 - But if even a minor difference would be of interest, then a much larger sample is needed

Summary: Key Ideas and Concepts

- Find out how big a **sample size** is needed for a **statistical test** to be conducted
- Pre-define the minimum size of the effect to be detected (i.e., **effect size**)
- Specify the required probability of detecting the effective size (**power**)
- Specify the significance level (**alpha**) at which the test will be conducted
- The vocabulary:
 - **Effect size**
 - **Power**
 - **Significance level**

Type I and II Errors, Power, Significance Level (α)

Actual	Decision	
	Reject H0	Fail to Reject H0
	H0 True	Type I Error
H0 False	correct	Type II Error

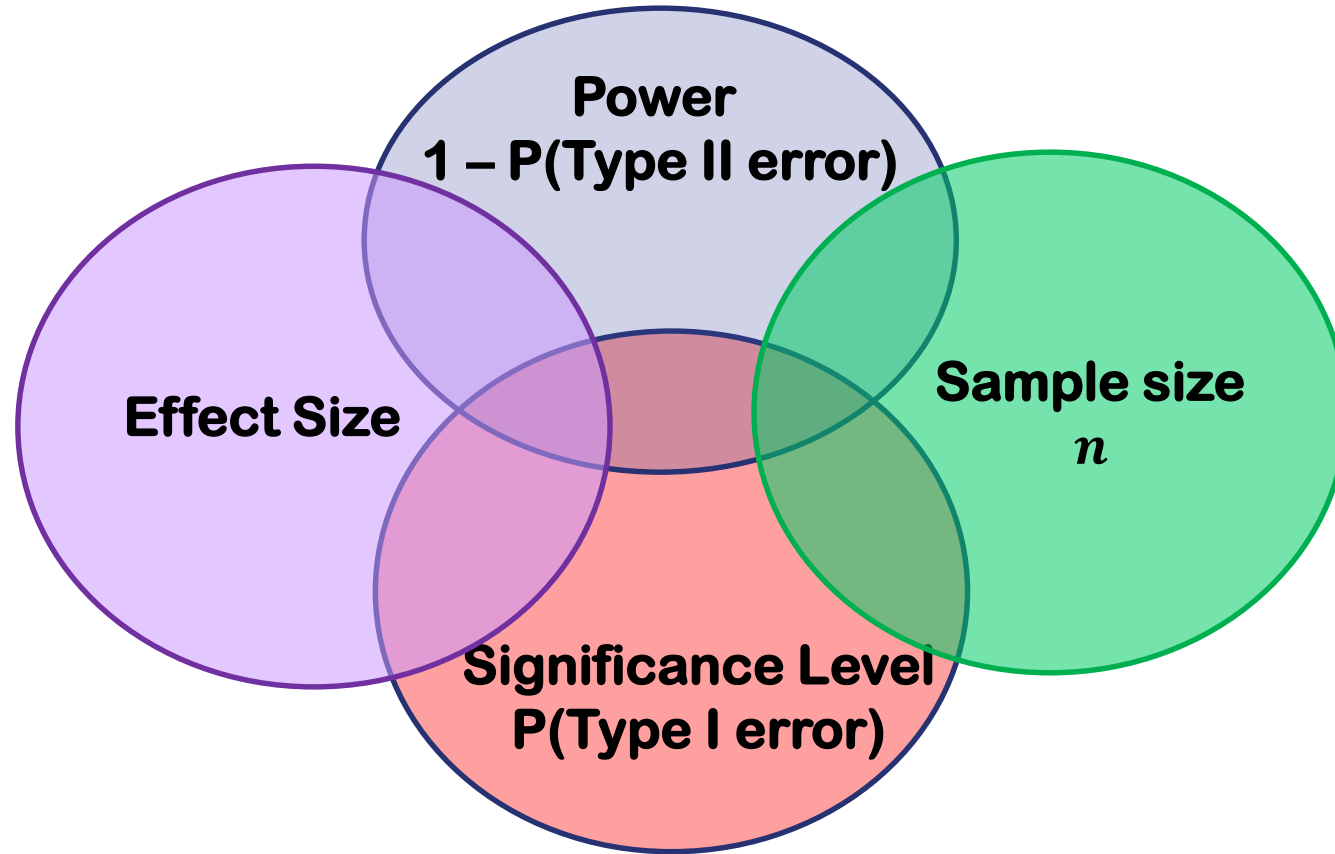
Power = 1 – Probability (Type II Error)

α = Probability (Type I Error)

- **Type I Error**
 - Probability of False Positives
 - Fail to find an effect that IS there
- **Type II Error**
 - The probability of False Negatives
 - Find an effect that is NOT there
- **Power** is defined as one minus the probability of making a Type II error
 - The probability of finding an effect that *IS* there
- **Significance level** (α) is the probability of making a Type I error
 - The probability of finding an effect that is *NOT* there

Type I & II Errors, Power, Significance Level, Effect Size

The moving parts



*given any three,
the fourth can be determined*

Power Calculations in R with **pwr** Package

Function	Samples	Power Calculations for	Compare>Data Type
<code>pwr.2p.test()</code>	2	Two proportions (equal n)	Proportions
<code>pwr.2p2n.test()</code>	2	Two proportions (un-equal n)	Proportions
<code>pwr.p.test()</code>	1	Proportion (one sample)	Proportions
<code>pwr.t.test()</code>	1, 2, paired	t -tests	Continuous, means
<code>pwr.t2n.test()</code>	2	t -test (un-equal n)	Continuous, means
<code>pwr.chisq.test()</code>	2	Relationship between 2	Categorical variables
<code>pwr.anova.test()</code>	Multiple	Balanced one-way ANOVA	Continuous, means
<code>pwr.r.test()</code>	2	Correlation between two continuous variables	Continuous, correlation coefficients
<code>pwr.f2.test()</code>	2	General linear model	Impact of a set of predictors on an outcome or impact of one set of predictors above and beyond a second set of predictors (or covariates)

Power Analysis with `pwr.t.test()`: t-tests of means

(one sample, two samples and paired samples)

```
pwr.t.test(n = NULL, d = NULL, sig.level = 0.05, power = NULL,  
           type = c("two.sample", "one.sample", "paired"),  
           alternative = c("two.sided", "less", "greater"))
```

- | | |
|----------------------|----------------------------------------------------------------------------------------------------------------------------|
| • n | Number of observations (per sample) |
| • d | Effect size |
| • sig.level | Significance level (Type I error probability) |
| • power | Power of test (1 minus Type II error probability) |
| • type | Type of t test : one- two- or paired-samples |
| • alternative | a character string specifying the alternative hypothesis, must be one of "two.sided" (default), "greater" or "less" |

Example: t-test of means: two.sample

Case Study: Cell phone usage and driving reaction time

Null Hypothesis: $H_0: \mu_1 - \mu_0 = 0$ ← two sample

The mean response time (μ_1) for drivers using a cell phone is the same as the mean response time (μ_0) for drivers that are cell phone free.

Alternative Hypothesis: $H_1: \mu_1 \neq \mu_0$ ← two sided

Empirical Data (past experience):

- 1-second difference in reaction time is considered an important difference
- reaction time has a standard deviation of 1.25 seconds

$$\text{Effect size (d)} = \frac{\mu_1 - \mu_0}{\sigma} = \frac{1}{1.25} = 0.8 \text{ or larger}$$



standardized mean difference

Example: **t-test of means**: Equal Participants: **R code**

Case Study: **Cell phone usage and driving reaction time**

```
1 # File: power_analysis.R
2
3 #install.packages("pwr")
4 library(pwr)
5
6 # Cell phone usage and driving reaction time
7 # d = 0.8 or larger: standardized mean difference
8 pwr.t.test(d=0.8, sig.level=0.05, power=0.9,
9            type="two.sample",
10            alternative="two.sided")
```

Output

Two-sample t test power calculation

→ n = 33.82555
d = 0.8
sig.level = 0.05
power = 0.9
alternative = two.sided

→ NOTE: n is number in *each* group

sample size = 34 participants in each group in order to detect an **effect size** of 0.8 with 90% certainty (**power**) and no more than a 5% chance (**alpha**) of erroneously concluding that a difference exists when, in fact, it does not

Example: **t-test of means**: Un-equal Participants: **R code**

Case Study: **Cell phone usage and driving reaction time**

- Constraints:
 - Assume that you want to detect 0.5 standard deviation difference in population means
 - You limit the chances of falsely declaring population means to be different to 1 out of 100
 - You can only afford to include 40 participants in the study, with equal number in each group
- Question:
 - What is the probability to detect a difference between the population means that is that large, given the above constraints?

```
pwr.t.test(n = 20, d=0.5, sig.level=0.01,  
           type="two.sample",  
           alternative="two.sided")
```

you have less than 14% chance of declaring a difference of 0.625 seconds or less ($d = 0.5 = 0.625 / 1.25$).

Conversely, there is 86% chance that you will miss the effect that you are looking for.

Output

```
Two-sample t test power calculation  
  
      n = 20  
      d = 0.5  
sig.level = 0.01  
power = 0.1439551  
alternative = two.sided
```

NOTE: n is number in *each* group

BONUS: Calculating Power via Statistical Simulation

- 1. Start with some hypothetical data that represent your best guess about the data that will result (e.g., e.g., based on prior data)**
 - a hat with 20 ones and 80 zeros to represent a 0.2 hitter or
 - a hat with some observations of “time spent on website”
- 2. Create a second sample by adding the desired effect size to the first sample**
 - a second hat with 33 ones and 67 zeros or
 - a second hat with 25 seconds added to each initial “time spent on website”
- 3. Draw a bootstrap sample of size n from each box**
- 4. Conduct a hypothesis test (permutation- or formula-based) on the two bootstrap samples and record whether the difference between them is statistically significant**
- 5. Repeat Steps 3 and 4 many times:**
 - To determine how often the difference was significant
 - This is the estimated **power**