Power Analysis & Sample Size Estimation

Determine Required Sample Size at the Beginning of your Research

A sample size calculation is a standard part of a study design. Studies with low sample numbers result in inaccurate and inconsistent estimates. Studies with too many samples result in unjustifiable sampling cost. In addition, depending on the type of samples, there may be ethical reasons to optimize sample collection. For example, unnecessarily exposing too many people to a drug to test its effectiveness, or exposing too few people that will eventually end up with unreliable results, are examples of unethical sampling.

What Information Do We Need?

The first step in determining the appropriate sample size is to plan the study. We must determine what is it that we are trying to estimate and what we need this estimate for. We need to define the number and type of variables we are using, how we are going to categorize our subjects or groups, what data types we are dealing with for independent and dependent variables, and what statistical test we are going to use. Once we determined the statistical test to be used, we need to determine how precise we want our estimate to be. For example, we might be okay with having a 95% confidence in a weight loss drug, but we might need 99% confidence when we test a new high-pressure valve that if fails, would have catastrophic outcome.

The reliability of our test is determined through the level of statistical significance, and statistical power required. In order to estimate sample size, both statistical significance and statistical power need to be determined during experiment design.

Statistical Significance

In statistical analysis, we often define a null hypothesis, in which we state that any relationship between groups is caused by pure chance. We define a p-value which is a probability that an observed difference could have occurred at random. Therefore, the higher the p-value, the higher the probability that an observed difference occurred by pure chance (we accept the null hypothesis, the observed difference occurred at random). Lower p-values represent the lower probability that the observed difference occurred by chance (we reject the null hypothesis and there is a meaningful difference between the groups).

We need to decide on the significance level (alpha) we'd like to achieve at the beginning of our study. Note that alpha and p-value are used interchangeably. Alpha is used for a pre-chosen probability, whereas p-value is the probability we calculate after a study. The confidence of 95% (alpha=0.05) is widely accepted in literature. If higher precision is needed, alpha can be chosen to even lower numbers (for example 0.01 to reach 99% confidence).

A significance level (alpha) is determined to reduce the probability of making a Type I error. Type I error is when we reject the null hypothesis in favour of a false alternative hypothesis.

Statistical Power

In designing an experiment, we also need to make sure that we do not make a type II error. Type II error is defined as failing to reject a false null hypothesis. This is when we define statistical power. In other words, power is the probability of making a correct decision (to reject the null hypothesis) when the null hypothesis is false. In most studies, the power is usually set to 0.80 or higher. For most research studies, powers lower than 0.80 are considered too low.

The power of a statistical test depends on sample size and effect size. In order to determine the effect size, we need to determine the following:

- How much variation is in the population we are studying?
- How much of a change we'll expect as a result of treatment, not just how much variability we'd expect, but how large would that change be.

For example: We conduct a study to see if a particular diet was effective in weight loss. If the samples are from a group of obese individuals, we might expect more weight loss (so we'll expect to see a change pretty easily, so smaller sample sizes would suffice), and if we take samples from a group with normal BMI, we might expect a small change (so in order to detect a change we'll need a bigger sample size).

There are three main methods to determine the effect size:

- The effect size can be estimated based on pilot data
- The effect size can be estimated from previous research (meta analysis)
- The effect size can be estimated through an educated guess. An educated guess is generally based on our understanding of research

If pilot data or previous research data are available, effect size can be calculated as follows:

Effect Size = the estimated difference in means / pooled estimated standard deviations

It is also possible to use standardized effect size metric is Cohen's effect size, where "small", "medium" and "large" effects are defined as standardized effect sizes.

Determining Appropriate Sample Size

Once we have the information required, we can easily use a variety of web applications and software programs to calculate the minimum sample size required. Two methods are mentioned here:

- G*Power software (free to download)
- Statsmodels library in Python

An example of sample size required for conducting a mean comparison in two independent groups is provided below using both of these methods. We assumed 80% power at 0.05 significance with medium effect size of 0.5.

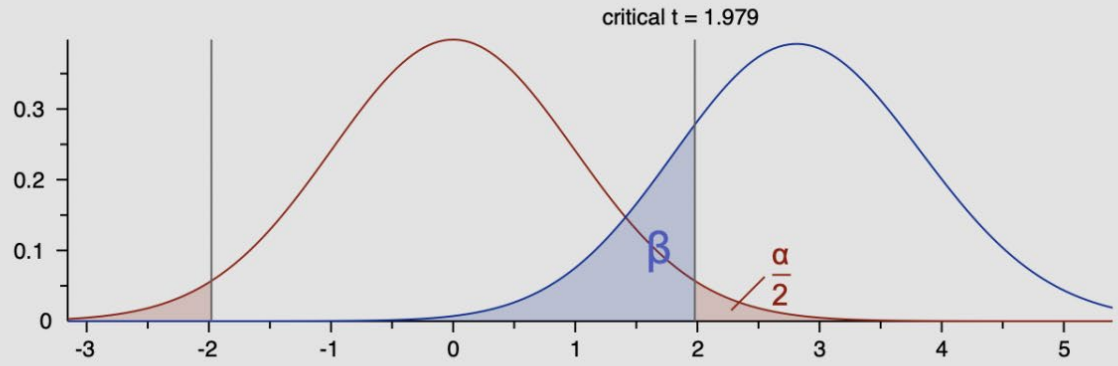
```
(base) Golpiras-MacBook-Pro:~ golpira$ ipython
Python 3.7.4 (default, Aug 13 2019, 15:17:50)
Type 'copyright', 'credits' or 'license' for more information
IPython 7.8.0 -- An enhanced Interactive Python. Type '?' for help.

In [1]: # estimate sample size via power analysis
   ...: from statsmodels.stats.power import TTestIndPower
   ...: # parameters for power analysis
   ...: effect = 0.5
   ...: alpha = 0.05
   ...: power = 0.8
   ...: # perform power analysis
   ...: analysis = TTestIndPower()
   ...: result = analysis.solve_power(effect, power=power, nobs1=None, ratio=1.0, alpha=alpha)
   ...: print('Sample Size For Each Group: %.0f' % result)
/Users/golpira/opt/anaconda3/lib/python3.7/site-packages/statsmodels/tools/_testing.py:19: FutureWarning: panda
s.util.testing is deprecated. Use the functions in the public API at pandas.testing instead.
  import pandas.util.testing as tm
Sample Size For Each Group: 64

In [2]:
```

G*Power 3.1

Central and noncentral distributions | Protocol of power analyses



Test family: t tests

Statistical test: Means: Difference between two independent means (two groups)

Type of power analysis: A priori: Compute required sample size - given α, power, and effect size

Input parameters

Tail(s)	Two
Effect size d	0.5
α err prob	0.05
Power (1-β err prob)	0.8
Allocation ratio N2/N1	1

Determine

Output parameters

Noncentrality parameter δ	2.8284271
Critical t	1.9789706
Df	126
Sample size group 1	64
Sample size group 2	64
Total sample size	128
Actual power	0.8014596

X-Y plot for a range of values | Calculate