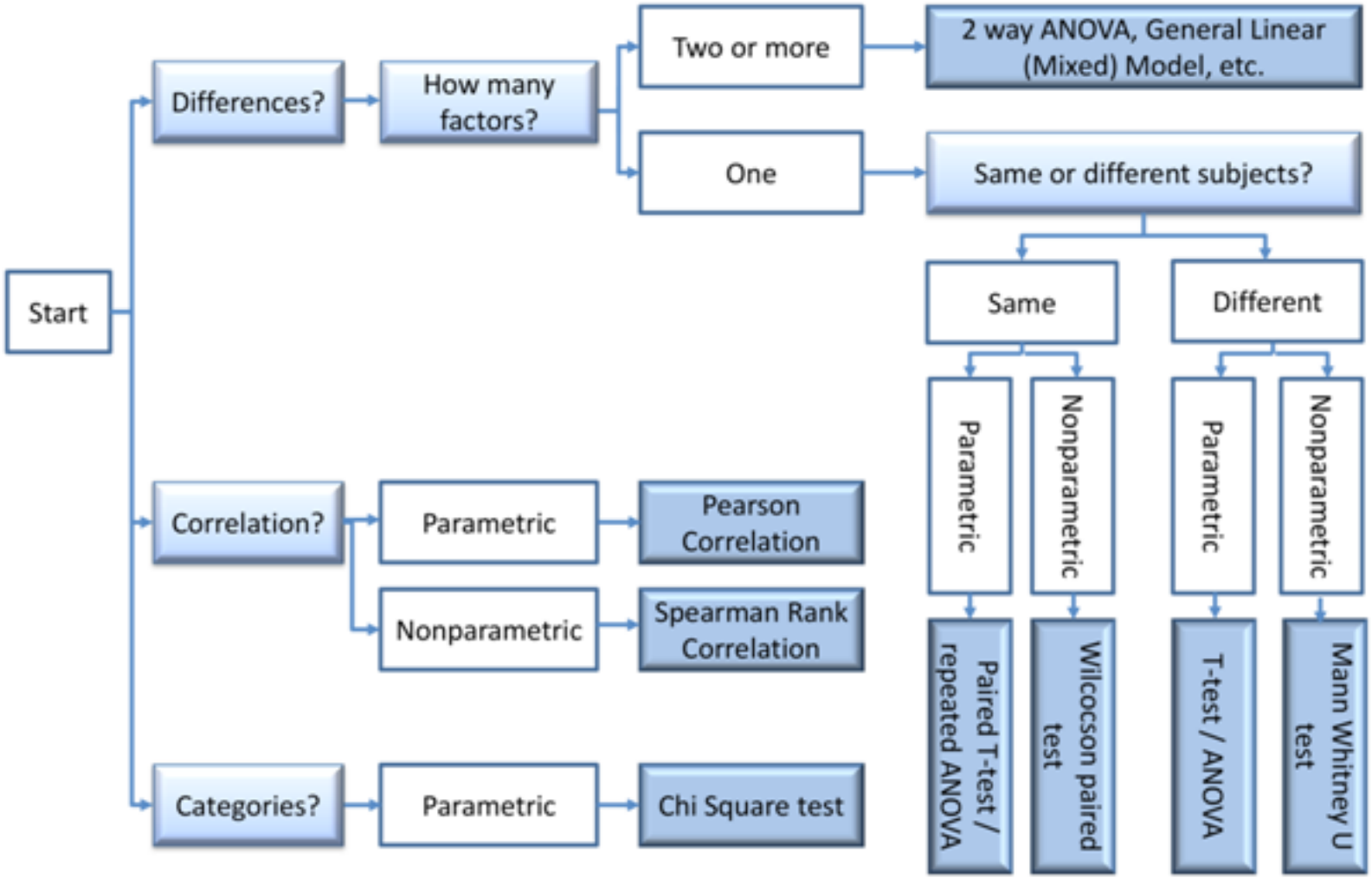
Data Science in the Wild

Lecture 9: Sampling

Eran Toch

Types of Tests



Sampling questions

- A sample is “a smaller (but hopefully representative) collection of **units** from a **population** used to determine truths about that population” (Field, 2005)
- What can we ask about sampling?
  - What is the population of interest?
  - What is the sampling procedure?
  - What is the sample size?

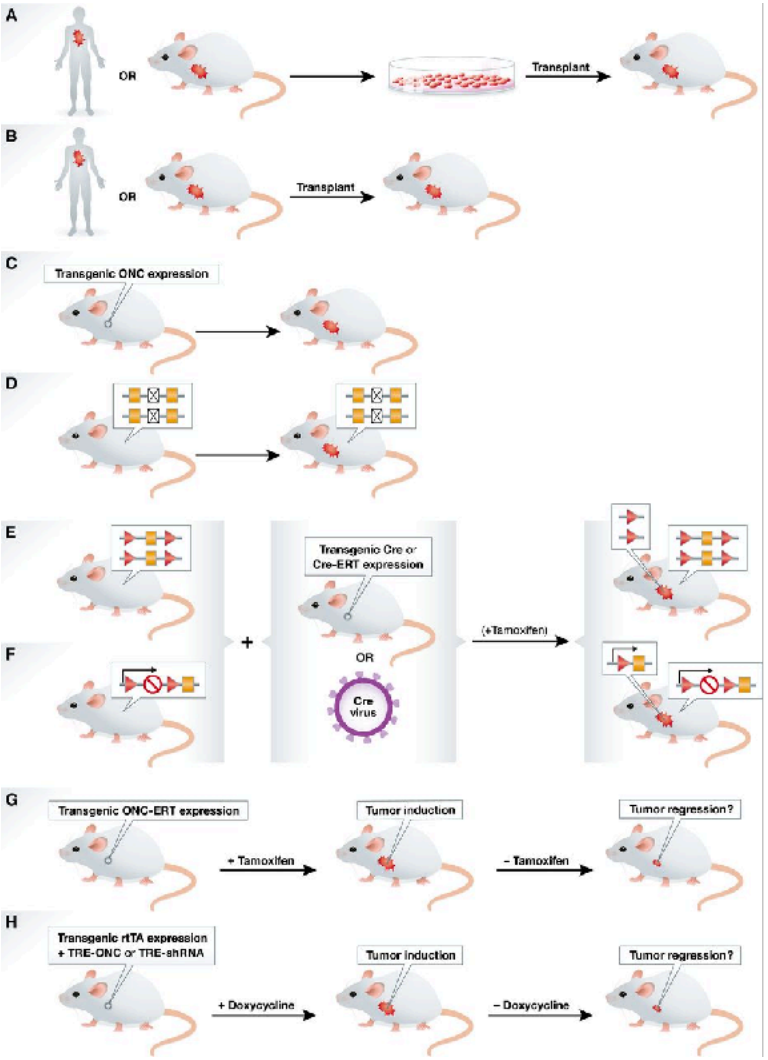
Sampling Process

1. Defining the population of concern
2. Specifying a sampling frame, a set of accessible items
3. Specifying a sampling method for selecting items or events from the frame
4. Determining the sample size
5. Implementing the sampling plan
6. Sampling and data collecting
7. Reviewing the sampling process

Sampling Procedure

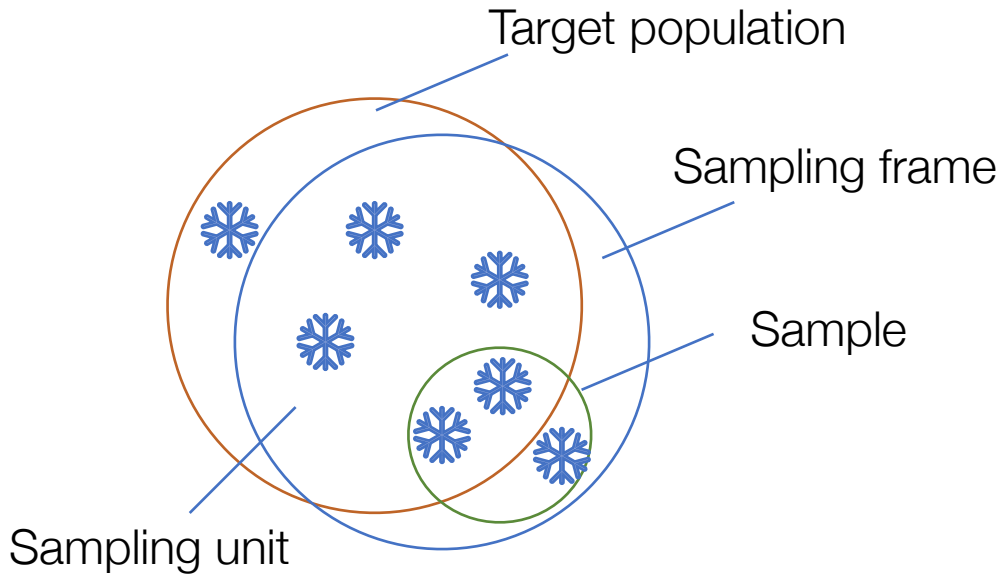
Defining the population of interest

- A population is all the units with the characteristic one wishes to understand
- People: Age, gender, education, computer experience, users of certain web sites, OS
- Other units of interest:
  - Wheat plants
  - Manufactured items
  - Mice (sometimes acting as models)
  - Mobile OS applications
  - Atoms
  - Schools



Sampling Frame

- We may not have access to the entire population
- So we call the accessible sampling units as the sampling frame
- Example:
  - Our target population is the entire US population
  - But not all will have phone numbers
  - The US population that can be communicated by phone numbers is the sampling frame



Ideal Sampling Frame Characteristics

- All units have a unique identifier
- All units can be found and accessed (e.g., contacted)
- The frame has additional meta-data about the units that allows advanced sampling frames
- Every element of the population is present in the frame
- Every element of the population is present only once in the frame
- No elements from outside the population of interest are present in the frame

Sampling method

- How do we reach our target population?
  - Is there a directory of targeted users?
  - An e-mail distribution list?
  - A postal mailing list?
  - A web site they all visit?
  - A social networking group?
  - Face-to-face meetings?
  - Membership in a certain organization
  - Job licensing or certification?

How to sample?

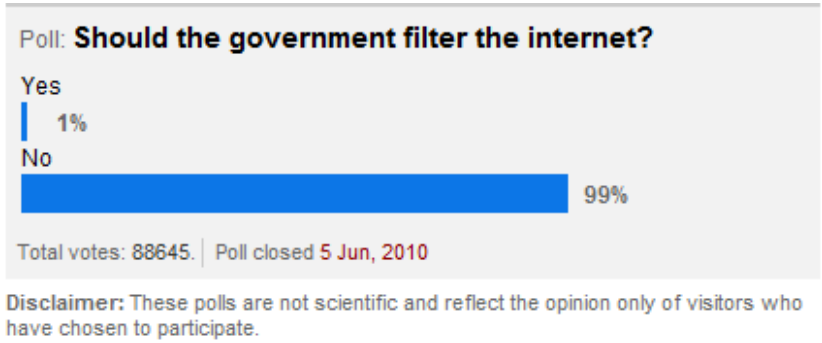
- Two major types of sampling methods:
  - **Probabilistic sampling**
    - Where there is a known probability of a unit being chosen
  - **Non-Probabilistic sampling**
    - The likelihood of being chosen is unknown

Non-probabilistic sampling

- Non-probabilistic sampling is used when:
  - You do not use a strict random sample
  - You do not know the likelihood of an individual being selected
  - You are not interested in a population estimate
  - There may not be a clearly defined population of interest

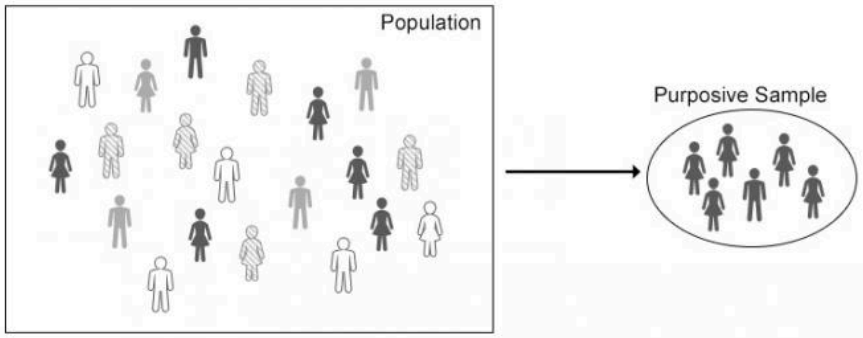
Non-Probabilistic Sampling

- **Convenience sample**: made up of people who are easy to reach
- **Quota sampling**: the sample has the same proportions of individuals as the entire population with respect to known characteristics, traits or focused phenomenon
- **Purposive sample**: Units are selected based on characteristics of a population and the objective of the study
- **Self-selected surveys**: Units decide for themselves whether to participate



Purposive Samples

- **Heterogeneous**: A maximum variation/heterogeneous purposive sample is one which is selected to provide a diverse range of cases
- **Typical case sampling**: a sample that relates to what are considered "typical" or "average" members of the effected population
- **Extreme/Deviant Case Sampling**: when a researcher wants to study the outliers that diverge from the norm as regards a particular phenomenon, issue, or trend
- **Critical case sampling**: one case is chosen for study because the researcher expects that studying it will reveal insights that can be applied to other like cases
- **Expert Sampling**: when research requires one to capture knowledge rooted in a particular form of expertise



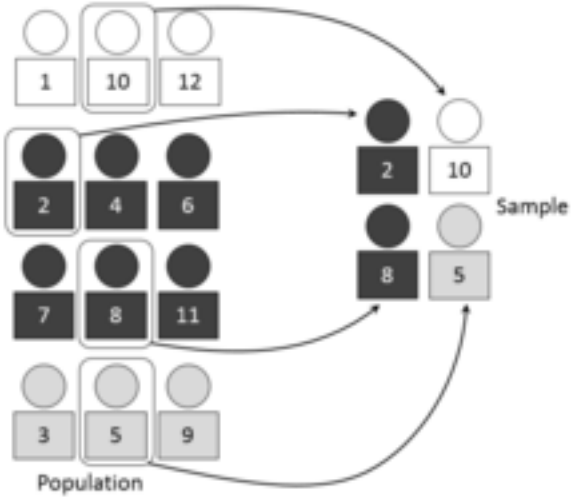
Purposive Sampling

Probabilistic sampling

- A probability sampling scheme is one in which every unit in the population has a chance (greater than zero) of being selected in the sample, and this probability can be accurately determined
  - **Census**
    - Where every single unit in the targeted population is chosen to take part in the sample
  - **Simple random sample**
    - All subsets of the frame are given an equal probability
    - Estimates are easy to calculate

Stratified sample

- A stratified sample is when you have an appropriate number of responses from each subset of your user population
- Every unit in a stratum has same chance of being selected
- Example: a random sample of college students would not have an equal number of freshman, sophomores, juniors, and seniors.
- A stratified random sample would have an equal number from each class year.
- But It doesn't need to be equal. It would still be stratified if you took 40% seniors, 40% juniors, 10% sophomores, and 10% junior. The researcher decides what is the appropriate breakdown.

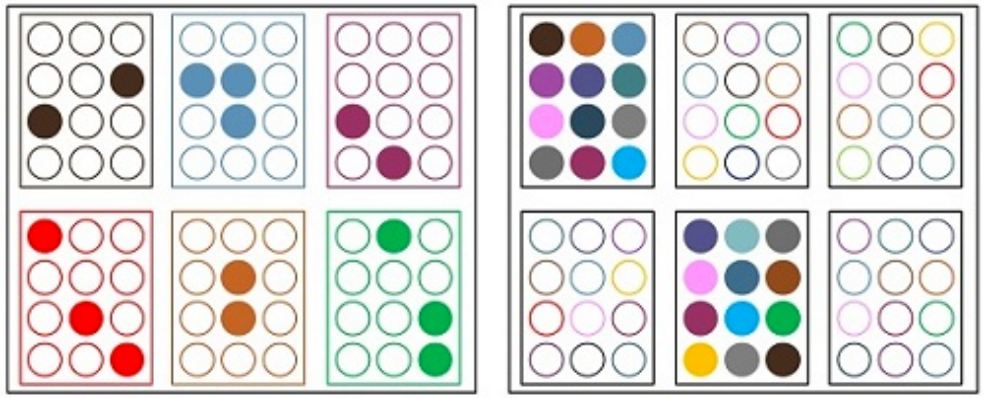


Cluster sample (or two-step sampling)

- In cluster sampling, we wish to sample some cluster of units as well as the units
- For example, we wish to randomly select some census tracks and then sample people in them
- Process:
  - At the first stage a sample of clusters is chosen
  - All units in the cluster are studied

Cluster sampling

- When to use?
  - Population divided into clusters of homogeneous units, usually based on geographical contiguity
  - Sampling units are groups rather than individuals.



Stratified Sampling Vs Cluster Sampling

Establishing informal validity

- If non-probabilistic surveys are used, both demographic information and response size both become important in establishing informal validity
- Demographic data can be used to ensure:
  - Respondents represent a diverse population.
  - Respondents are somewhat representative of already-established population.

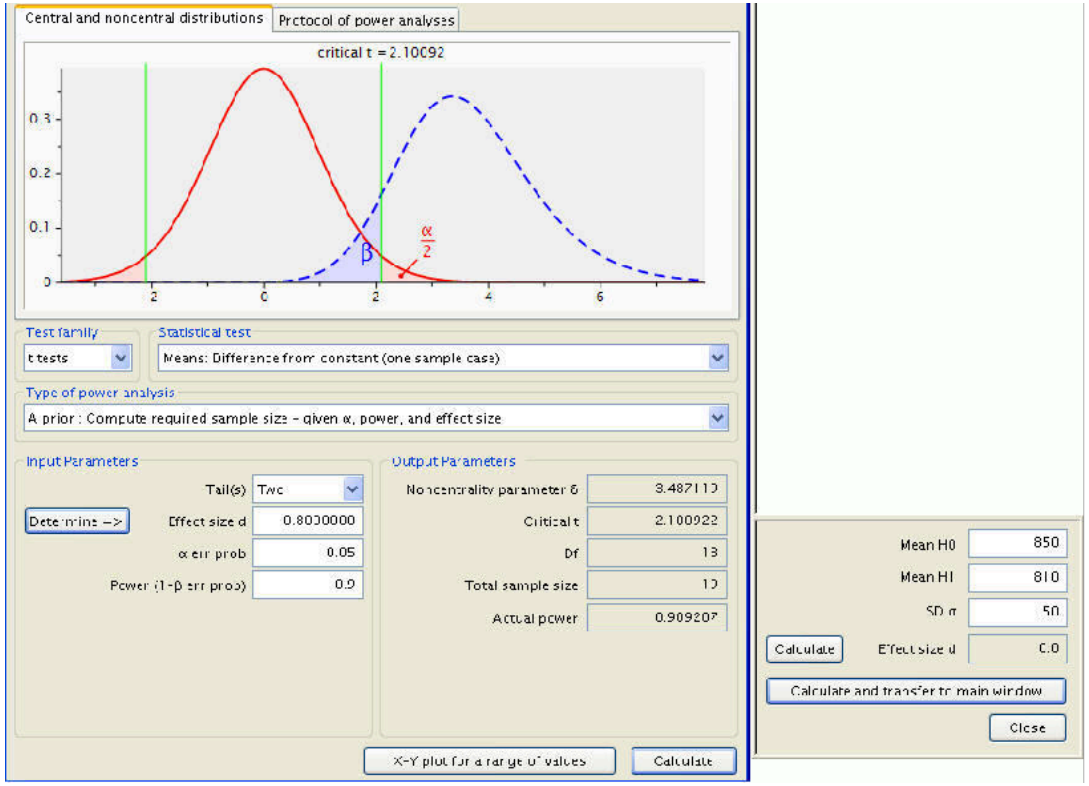
Sources of error and bias

- Sampling error (not enough responses)
- Coverage error (not all members of the population of interest have an equal likelihood of being sampled)
- Measurement error (questions are poorly worded)
- Non-response error (major differences in the people who were sampled and the people who actually responded)

Sampling Size

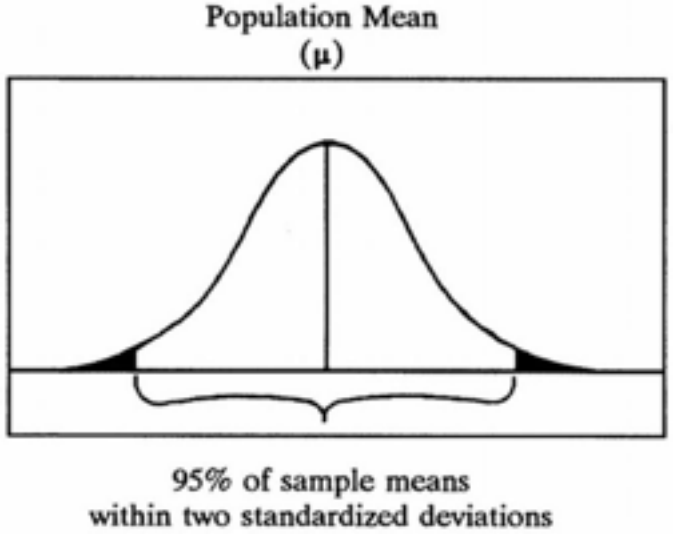
Sample size

- What sample size is considered to be sufficient for a random sample?
- It depends on what we are looking for:
  - Estimating values
  - Establishing hypotheses



Estimating values

- The sample size depends on the confidence level and margin of error you consider acceptable
- For instance, to get a 95% confidence level and +-5% margin of error, you need 384 responses.

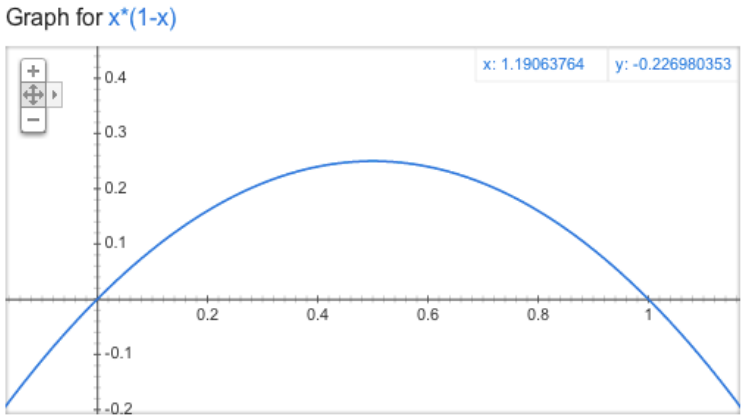


if a 95% confidence level is selected, 95 out of 100 samples will have the true population value within the range of precision specified earlier

Power analysis: Calculating the Sample Size

Formula:

$$n_0 = \frac{Z^2 pq}{e^2}$$



Where:

n_0 = required sample size
Z = confidence level at 95% (standard value of 1.96 in a normal distribution)
p = degree of variability, q=1-p
e = margin of error at 5% (standard value of 0.05)

Example

We wish to evaluate a program in which users were encouraged to adopt a new practice.

Assume there is a large population but that we do not know the variability in the proportion that will adopt the practice; therefore, assume p=.5 (maximum variability). Furthermore, suppose we desire a 95% confidence level and ±5% precision.

$$n_0 = \frac{Z^2 pq}{e^2} = \frac{(1.96)^2(.5)(.5)}{(.05)^2} = 385$$

Power analysis

- The power of a binary hypothesis test is the probability that the test rejects the null hypothesis (H_0) when a specific alternative hypothesis (H_1) is true
- The statistical power ranges from 0 to 1, and as statistical power increases, the probability of making a type II error (wrongly failing to reject the null) decreases
- Power analysis can be used to calculate the minimum sample size required so that one can be reasonably likely to detect an effect of a given size

Calculating Power

- To calculate the sample size of a given statistical test, the following are needed:
- significance level (let's say 0.05)
- effect size
- power (let's say π=0.8 or 0.9 in the next example)

Calculation with t-test

- The effect of the treatment can be analyzed using a one-sided t-test, the statistics is given by: $$T_n = \frac{\bar{D}_n - 0}{\hat{\sigma}_D / \sqrt{n}},$$
- Given a critical value $\alpha = 0.05$. The null hypothesis will be rejected if

$$T_n > 1.64.$$

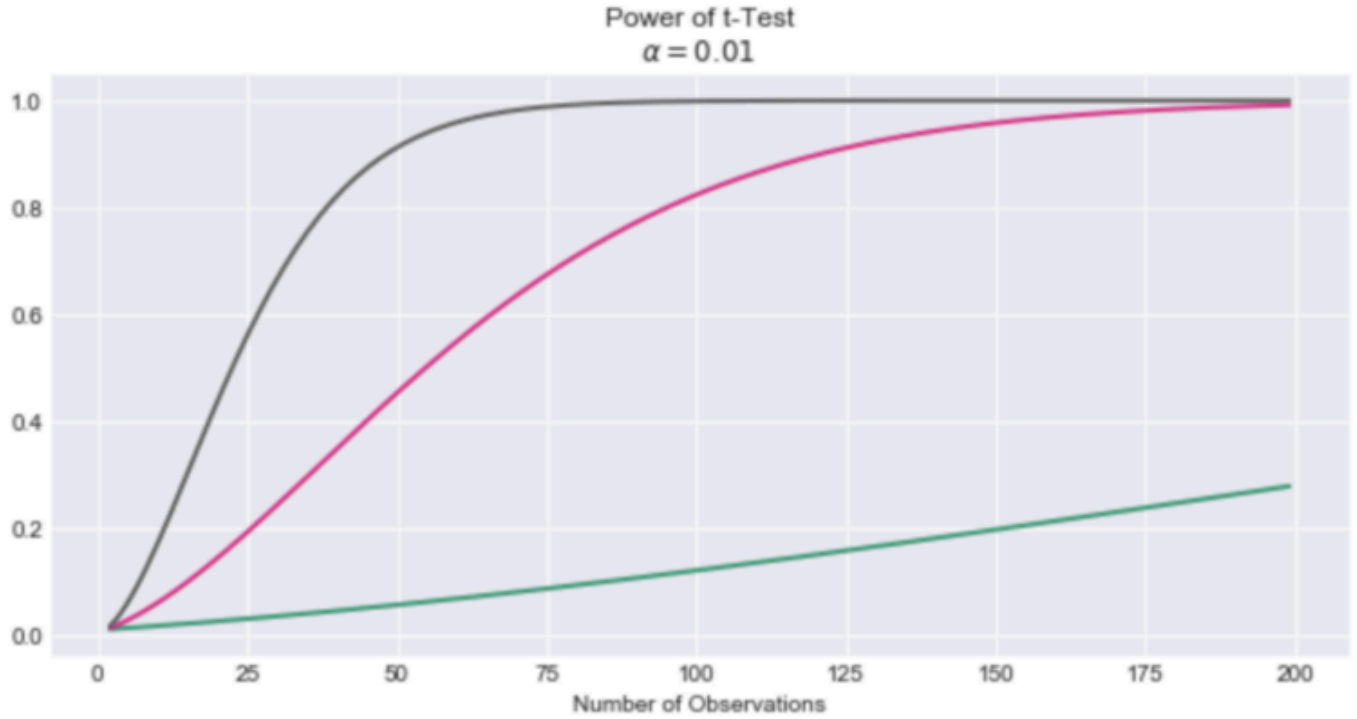
$$B(1) \approx 1 - \Phi\left(1.64 - \frac{\sqrt{n}}{\hat{\sigma}_D}\right) > 0.90,$$

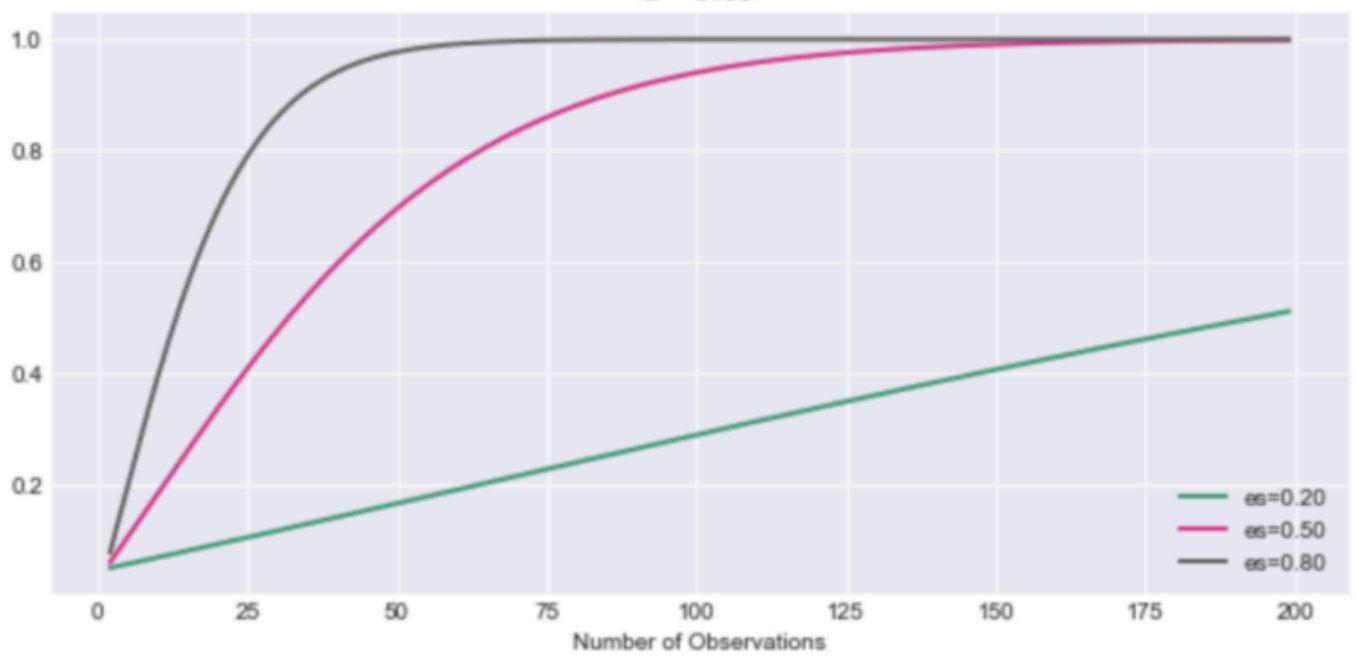
$$\frac{\sqrt{n}}{\hat{\sigma}_D} > 1.64 - z_{0.10} = 1.64 + 1.28 \approx 2.92 \qquad \text{or} \qquad n > 8.56\hat{\sigma}_D^2,$$

Python code

```
# parameters for the analysis

effect_size = 0.8
alpha = 0.05 # significance level
power = 0.8

power_analysis = TTestIndPower()
sample_size = power_analysis.solve_power(effect_size = effect_size,
                                         power = power,
                                         alpha = alpha)

print('Required sample size: {0:.2f}'.format(sample_size))
```

https://towardsdatascience.com/introduction-to-power-analysis-in-python-e7b748dfa26


Power of t-Test
$\alpha = 0.01$
1.0
0.8
0.6
0.4
0.2
0.0
0
25
50
75
100
125
150
175
200
Number of Observations


$\alpha = 0.05$
1.0
0.8
0.6
0.4
0.2
es=0.20
es=0.50
es=0.80
0
25
50
75
100
125
150
175
200
Number of Observations

Summary

- Every scientific activity has some questions of sampling
- Different types of sampling
- Sample size and power analysis