lec01Claros

Andrew Claros

August 26, 2020

- $n = 20 : \{0, 1, 1, 1, 0, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 0, 0, 1, 1, 0\}$ iphone vs ~iphone
- Meaning: $\{x_1 = 0, x_2 = 1, x_2 = 1, \ldots, x_{20} = 0\}$

if $n = 20$ elements $\in$ a (larger) super-set called the population, this is called the "population model sampling assumption"

Define population for this sample:

- all people on Earth?
- All people in America?
- etc.
- All QC students? this is a lot better

Is this sample representative of the population?

- (all on earth) there are individuals that would not even have a cell phone
- all college students (in NYC)? Better.
- QC students? Maybe

Problem: Given a sample, find a population model (i.e identify the representative population)

- fuzzy idea
- Data Science perspective

In classical statistics this is reversed

- First clearly define population of size N and then sample from the population (sample of size n.
- Population has size N. is population = "all Americans",

Sample defined:

- $x_1, x_2, \ldots, x_n$, in sample but not other data in population
- Can we learn about the population from sample? Yes, called "inference"
- Use sample to infer properties about population
- Usually properties of the random variable model creates the population
- Infer means to make an educated guess from specific occurrences to universal properties.

- This is a guess: meaning it can be wrong
- Synonym is induction
- Opposite is deduction: properties are universal to a subset

How is inference done?

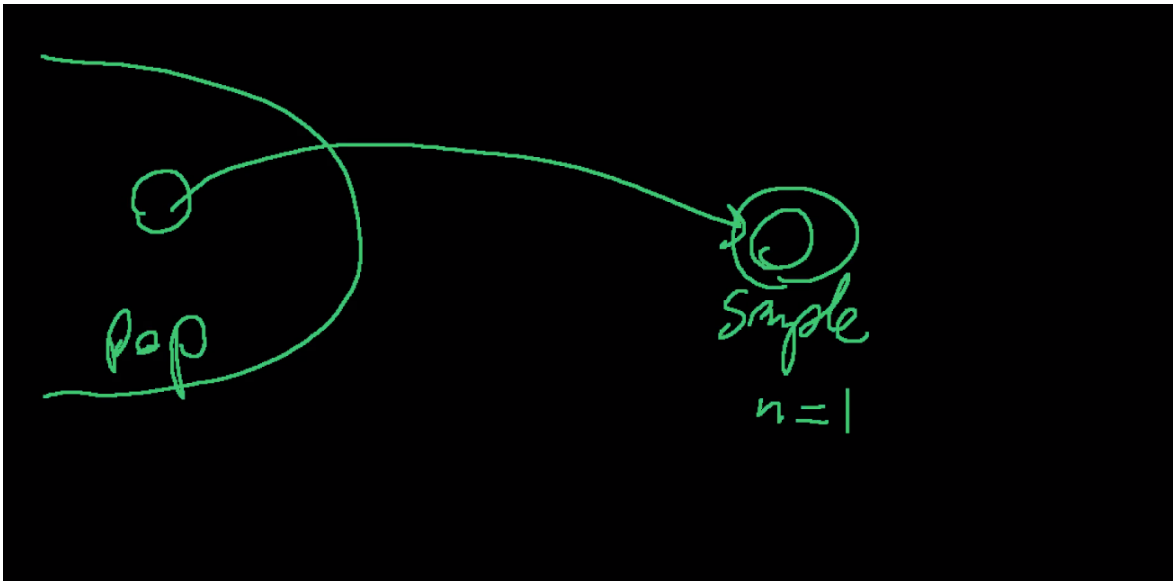
- Point estimation
- Confident set
- theory testing (testing a theory about a specific value of θ at a "certainty level" α)

Point estimation

- Generate statistics which are functions of the data/sample
- In general $\hat{\hat{\theta}} = w(x_{1,\ldots,x_n})$ w is a function of data. $\hat{\hat{\theta}}$ is usually scalar
- $\hat{\hat{\theta}} = \frac{1}{n}\sum x_i = 0.6$
- What can you infer with this statistic? Infer θ (the population) which is the true proportion of iPhones. Statistics to make inference.
- $\$\theta := \chi_{\overline{N\$-\chi}}$ -> number of people in population that have iphones (unknown)
- N # of elements in population (known)
- $\theta \in \Theta = [0, \frac{1}{N}, \ldots, \frac{N-1}{N}, 1]$, the parameter space Θ
- Convention, Greeks letters represent unknown quantities, Roman represent known quantities
- $\hat{\hat{\theta}}$ is a point estimate for Θ, (i.e a specific value, one number) which you believe is a good guess for the value of Θ.

Theory testing:

- Sample one element from the population



- How should this element be chosen for a "representative sample?"- randomly but superficially. uniformly meaning every element has probability of $\frac{1}{N}$ of being chosen. This is called a "simple random sample" (SRS)

What is the probability $P(X_1 = x_1 = 1)$?

- $P(X_1 = x_1 = 1) = \frac{X}{N} = \theta$
- X_1 = random variable modeling the survey
- x_1 = the realization (a value in the support of X_1)