8/20/2
Enoch K

Let's do a survey, who has an iphone, I'll begin with me

standard notation for a "datum"

$x_1 = 0$ ("no") $x_2 = 0$, $x_3 = 1$, $x_4 = 1$, $x_5 = 0$, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1

$x_{16} = 0$, 0, 1, 1, $x_{20} = 0$

first survey response

$n = 20$ in our sample 12 1's
8 0's

Do we believe this survey is a "sample of $n = 20$ elements from a superset called the "population"? If we do, this is called the "population model sampling assumption"

If so, what is that population

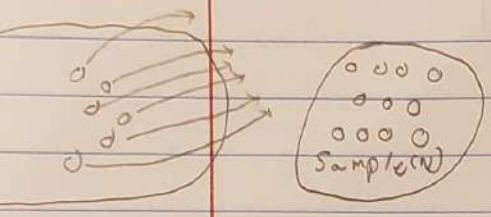
- all people on earth?
- all people in america
- all college students
- all college students in NYC
- all public college students in NYC
- all Bc students

Is this sample representative of the population?
This is typical, given a sample, assume a population model, then identify the representative population. This happens in data science all the time.

Housing prices in Census

In classical stat., this goes the opposite direction, you begin by defining the population clearly and then sample n element from that population

Population has size N. you have some idea of what N is
ex pop = all american $\Rightarrow N = 330$ million



We see the data $x_1, x_2, \ldots, x_n$
in the sample but not other data
in the population

Can we learn about the population from the sample
yes, This is called "inference". We use the sample to
"infer" properties about the population, usually the properties
are parameter of the r.v. model which creates the population

"Infer" means to make an educated guess from specific
things to universal properties. A synonym is "induction".
The opposite is deduction which is universal → particular
You can *never* be sure your inference is correct

How is inference done with data? you generate "statistics"
which are functions of the data

$\hat{\theta} = w(x_1, \ldots x_n)$ ← data
Stat. (usually scalar)

eg $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i = 0.6$ ← our iphone survey
$\quad = \bar{x} \quad = \hat{p}$

What can you infer with this statistic? Usually, you
infer theta the population parameter which is the
"true proportion of iphones". "Statistical inference" - using
stat. to make inferences

What is theta?

$\theta := \frac{x}{N}$ → # of people in the population that have iphones (unknown)
→ # elements in the population (known)

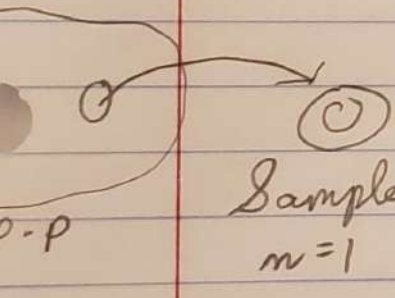
parameter (unknown property)

$\theta \in (H) = \{0, \frac{1}{N}, \frac{2}{N}, \ldots, \frac{N-1}{N}, 1\}$, the parameter space

Convention is that greek letter represent unknown quantities and roman letters represent known quantities.

$\hat{\theta}$ theta is a "point estimate" for the unknown theta. "Point" meaning one specific value which you believe is a good guess for the value of theta. (1) "point estimation" is one of the goals of stat. inference. The other two are (2) Confidence set creation and (3) theory testing (testing a theory about a specific value of theta at a "certainty level" alpha

Let's sample one element from the population, and do one survey



Sample
$n = 1$

How should this element be chosen? If I want a "representative" sample? Randomly but specifically uniformly meaning every element has probability of $1/N$ of being chosen. That's called a "simple random sample" (SRS)

What is prob. that $x_1 = 1$?

$$P(X_1 = x_1 = 1) = \frac{x}{n} = \theta$$

θ: Specific value

X_1: the r.v modeling the survey

x_1: the realization (a value in the support of x_1)