Let's do a survey. Who has an iPhone? I'll begin with me.

$x_1 = 0$ standard notation for a "datum"

first survey respondent

"No"

$x_2 = 0, x_3 = 1, x_4 = 1, x_5 = 0, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, \underbrace{0}_{x_{16}}, 0, 1, 1, \underbrace{0}_{x_{20}}$

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

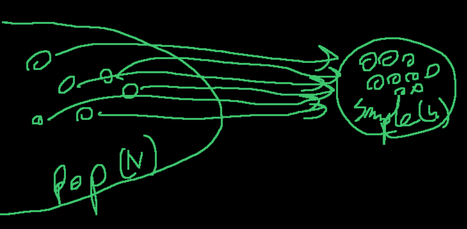
Do we believe this survey is a "sample" of n=20 elements from a superset called the "population"? If we do, this is called the "population model sampling assumption".

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 QC 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. In classical statistics, this goes the opposite direction. You begin by defining the population clearly and then sample n elements from that population.

Population has size N. You have some idea of what N is.
If pop = all Americans => N = 330million.



We see the data x_1, x_2, ..., 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 parameters of the random variable model which creates the population. "Infer" means to make an educations 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:

$$\hat{\hat{\theta}} = w(\overbrace{x_1, \ldots, x_n}^{\text{data}}) \quad \text{e.g.} \quad \hat{\hat{\theta}} = \frac{1}{n}\sum_{i=1}^{n} x_i = 0.6$$

statistic (usually scalar)

our iphone survey

$\hat{\hat{\theta}} = \bar{x} = \hat{p}$

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

What is theta?

$$\theta := \frac{x}{N}$$

x → # of people in the population that have iphones (unknown)

N → # elements in the population (known)

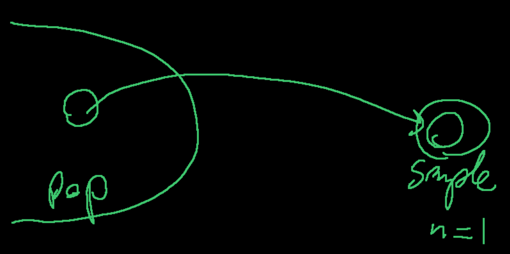
parameter (unknown property)

$\theta \in \Theta = \left\{0, \frac{1}{N}, \frac{2}{N}, \ldots, \frac{N-1}{N}, 1\right\}$, the parameter space.

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

theta-hathat 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 statistical 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.



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 the prob. that $X_1 = 1$?

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

the r.v. modeling the survey

the realization (a value in the support of X_1)

specific value