Lec 1.

Let's do a survey. Who has an iPhone?

Standard notation for a "datum"

$X_1 = 0$

first survey respondent

where
0 is no
1 is yes

$X_2 = 0, X_3 = 1, X_4 = 1, X_5 = 0$

$X_6 = 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 0, 0, 1, 1, 0$

$n = 20$ in our "sample."

12 (1's), 8 (0's)

Now:

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 of Earth
- All people of America
- All people in Q.C.
- All college 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

i.e everyone who has als over 60, clear population. Then sample n elements from that population.

Population has size N. You have some idea of what N is.

If pop = all Americans ⇒ N = 330 million.

Can we learn about the population from the sample?

Yes. This is called "inference". We use the sample to "infer" properties about the populati
Usually the properties are parameters of the rv 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 deducation 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:

This isn't the symbol used in textbooks

the iphone survey

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

$$\shortparallel \quad \shortparallel$$

$$\bar{x} \quad \hat{p}$$

statistic (usually scalar)

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 the~~t~~a?

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

the true # of people in the population that have i.e iphones (unknown)

Parameter (unknown property)

# of elements in the population (known)

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

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

$\hat{\hat{\theta}}$ is a "point estimate" for the unknown theata. "point" meaning one specific value which you believe is a good guess for the value of theta. "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.

$n=1$

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" (SR

i.e. imagine a hat with names in it.

So, what is the probability 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.)

specific value