Math 369 Lec 1

Let's do a "survey". Who has an iphone? Let's begin with me.

I do not $X_1 = 0$ ← code for "No". Yes will be 1.

"raise your hand"

standard symbol for datum

I'm the first survey element. I'm numero uno.

$X_1 = 0, X_1 = 1, X_3 = 1, \ldots\ldots, X_{20} = 0$

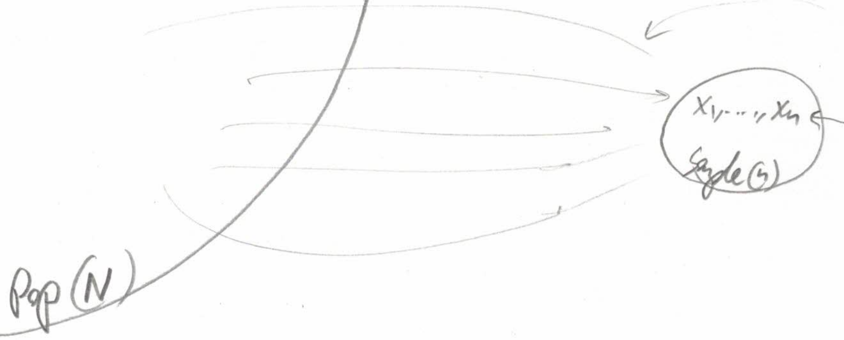
Do we believe this survey has a "sample" of $n = 20$ elements from a superset called the population? (i.e. a subset) This is the "population model sampling assumption". Let's assume it. What is the population?

- All people on Earth?
- All people in America?
- All college students?
- " " " " " in NYC?
- All public college students in NYC?

This is a typical situation. Given a sample, assume population model, then identify the population. This happens in data science all the time.

The more typical situation in classical statistics, is you start with a conception of a population e.g. Pop = All Americans. Then you take a sample of the population elements, size n and survey those. The pop. has N total elements. You should have some idea as to what N is. You define the population!

$N \approx 333$ million in this case.



sampling procedure / survey procedure

these are called the "data".

We see the data in the sample, but not the population. Can we use the sample to tell us something about the population?

Yes, the sample data is used to "infer" properties about the population. Numeric properties are called "population parameters".

"Infer" means to make an educated guess from the particular → the universal. A synonym is "induction".

The opposite is "deduction" which goes from universal → particular.

The process of "inference" is difficult and because it's a guess, you can __never__ be sure your inference is correct.

Assume: all swans are white $\xRightarrow{\text{deduce}}$ these 5 swans are white. | observe 5 white swans $\xRightarrow{\text{infer}}$ all swans are white.

this inference may be wrong.

How is inference done? We generate "statistics" by running functions on the data.

double-hat them is my own notation

$$\hat{\hat{\theta}} = w(x_1, \ldots, x_n) \quad \text{where } \hat{\hat{\theta}} \text{ is usually a scalar}$$

e.g.

in our iphone survey

$$\hat{\hat{\theta}} = \frac{1}{n}\sum_{i=1}^{n} x_i = .527$$

$\hat{\hat{\theta}} = \hat{p}$ sample proportion

$\hat{\hat{\theta}} = \bar{x}$ sample average

What can you infer from this statistic? The true population proportion parameter θ. "Statistical Inference": using statistics to make inferences.

What is θ?

$$\theta_1 = \frac{x}{N}$$

x ← # of people with iphones in pop (unknown)

N ← pop size (sometimes known)

$\theta \in \Theta$ All possible values

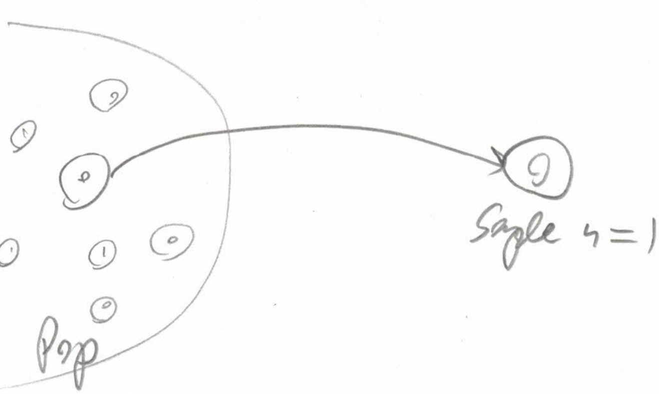
$\Theta = \{0, \frac{1}{N}, \frac{2}{N}, \ldots \frac{N-1}{N}, 1\}$ param space

Convention: Greek letters are used for unknowable quantities and Roman letters are used for knowable quantities.

$\hat{\theta}$ is a "point estimate" for θ. "Point" meaning one single value that you believe is a "good" guess for the value of θ.

"Point estimation" is one type of statistical inference. The two common other goals are "confidence set creation" (giving an interval of possible values of θ at a "certainty level" $1-\alpha$) and "theory testing" (testing a theory about the true value θ at a "certainty level" α).

Let's discuss sampling and surveying more. Let's take one sample + survey.



How is this element chosen? Randomly. Technically uniformly sampled i.e. every element of the pop. has prob. $\frac{1}{N}$ of being chosen.

Review of 241...

Representative sample: a sample that faithfully reflects the population. Uniformly random samples are representative.

What is the probability that $x_1 = 1$?

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

uses the naive definition of probability

$$\frac{\#\text{ elements satisfying event}}{\#\text{ of total elements}}$$

the r.v. modeling the first survey drawn (capital letter)

its realization $\Downarrow$ (lowercase letter)

a possible value of its realization

data: realizations of r.v.'s. Surveying: forcing a r.v. to realize.

Bernoulli r.v. with parameter θ

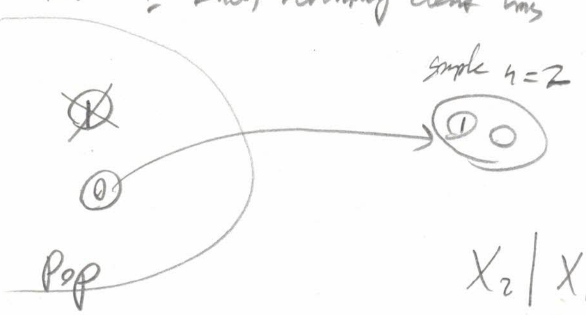
$$\Rightarrow X_1 \sim \text{Bern}\left(\frac{x}{N}\right) = \text{Bern}(\theta) = \underbrace{\theta^{x_1}(1-\theta)^{1-x_1}}_{\text{prob. mass function (PMF)}}$$

$\text{Supp}[X_1] = \{0, 1\}$ the support of the r.v. is the set of all possible realization values.

Note: the parameter of the survey r.v. is the same as the population parameter we would like to draw inference about.

Let's draw a second sample assuming $x_1 = 1$. And ask some questions

How? Each remaining element has $\frac{1}{N-1}$ prob. of being drawn.



$$P(X_2 = 1 \mid X_1 = 1) = \frac{x-1}{N-1} < \theta = P(X_1 = 1)$$

conditional prob.

$$X_2 \mid X_1 = 1 \sim \text{Bern}\left(\frac{x-1}{N-1}\right)$$

cond. r.v. model

$\Rightarrow X_1, X_2$ are dependent r.v.'s

If $x_1 = 0 \Rightarrow P(X_2 = 1 \mid X_1 = 0) = \frac{x}{N-1} > \theta$

either way dependent

What is $P(X_2 = 1) = \frac{x}{N} = \theta \Rightarrow X_2 \sim \text{Bern}(\theta)$

↑

unconditional prob. X_1 was realized but... you don't know what it is thus you pretend it doesn't exist

$\Rightarrow X_1 \overset{d}{=} X_2$ they're "identically distributed" since they have the same PMF.

Let's sample all n. Let $T_n = X_1 + \ldots + X_n$ i.e. the r.v. that tallies the total # of 1's.

$$P(T_n = t) = \frac{\binom{x}{t}\binom{N-x}{n-t}}{\binom{N}{n}} = \text{Hyper}(n, x, N)$$

sample size; total # of 1's in pop; total pop.

# of unique samples!

Hypergeometric r.v. model

How did it get this complicated? Because $\frac{x}{N} \neq \frac{x-1}{N-1}$!

Let's make a simplifying assumption. Let $x, N \to \infty$ with $\frac{x}{N} = \theta$.

$$\Rightarrow \frac{x}{N} \approx \frac{x-1}{N-1} = \frac{x-k_1}{N-k_2} = \theta \quad \text{for all } k_1, k_2.$$

with this limit, we assume $\Theta = [0,1]$ all possible real #'s

Now $P(X_2 = 1 \mid X_1 = 1) = \frac{x-1}{N-1} \overset{\text{assumed}}{=} \theta = P(X_2 = 1)$

$\Rightarrow X_1, X_2$ are independent $\Rightarrow X_1, X_2 \overset{iid}{\sim} \text{Bern}(\theta)$

For all n samples, $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \text{Bern}(\theta)$

$$\Rightarrow T_n = X_1 + \ldots + X_n \sim \text{Binomial}(n, \theta) = \binom{n}{t}\theta^t (1-\theta)^{n-t}$$

Let's consider a new sampling problem. At the iphone factory, they check every new iphone to make sure it works. Let's say they check the first one, $x_1 = 1 \leftarrow$ it works, the second $x_2 = 0 \leftarrow$ it doesn't work, ..., $x_{100} = 1$. What population is this sample from? All iphones? $N = ?$

Are you drawing one sample of n from $\binom{N}{n}$? Not really. What is θ? Is it a "population" parameter?

Would you agree $X_1, \ldots, X_n \overset{iid}{\sim} \text{Bern}(\theta)$?

Is it a "process" parameter? Process or infinite population...

We still have a r.v. model that describes the survey. At this point we no longer care whether the pop. is real or if its a process, we just need an iid r.v. model assumption called the "data generating process".

Let's return to our main goal: inference. Specifically: point estimation of a parameter θ.

$$\hat{\hat{\theta}} = \frac{1}{n}(x_1+\dots+x_n) \approx \theta.$$ How approximate is it?

Since $x_1, \dots, x_n$ were random realizations of $X_1, \dots, X_n$, $\hat{\hat{\theta}}$ could have been different. e.g. if $\vec{x} = [1\,0\,0\,1\,0] \Rightarrow \hat{\hat{\theta}} = 0.4$

but if $\vec{x} = [1\,1\,1\,0\,1] \Rightarrow \hat{\hat{\theta}} = 0.8$. Thus $\hat{\hat{\theta}}$ is a realization itself from a r.v. $\hat{\theta}_n = \frac{1}{n}(X_1 + \dots + X_n)$ called a "statistical estimator" or just "estimator". So... $\hat{\hat{\theta}}$ is a realization from $\hat{\theta}_n$. The distr. of $\hat{\theta}_n$ is called the "sampling distr." The properties of the estimator are very important because they ~~tell~~ us a lot about our estimate.

One property is the expectation,

$$E[\hat{\theta}_n] = E\left[\frac{1}{n} T_n\right] = \frac{1}{n} E[T_n] = \frac{1}{\cancel{n}} \cancel{n}\theta = \theta \quad \forall n$$

$X_1, \dots, X_n$ over all t

$$\Rightarrow E[\hat{\theta}_n] = \theta$$ this is special. It means $\hat{\theta}$ is "unbiased"

this expectation is taken over the Supp$[X_1], \dots,$ Supp$[X_n] := \mathcal{X}$ weighted by the joint mass function. Not over any distr for θ! θ is one #! $p(x_1, \dots, x_n)$. $\forall n$

In general, for any estimator and any pop. param.

$$\text{Bias}[\hat{\theta}_n] := E[\hat{\theta}_n] - \theta$$

If $\text{Bias}[\hat{\theta}_n] = 0$, $\hat{\theta}$ is "unbiased".
If $\text{Bias}(\hat{\theta}_n) \neq 0$, $\hat{\theta}$ is "biased".

Across every possible sample of any size n, the average estimate will be the pop. param θ.

This is certainly reasonable. How "far" is $\hat{\theta}$ away from θ?

Let's define "far" by a loss function $\ell(\hat{\theta},\theta)$ where

$\ell : \Theta \times \Theta \to [0,\infty)$ and $\ell(\hat{\theta},\theta) = 0$ only when $\hat{\theta} = \theta$.

Some examples...

* default $\ell(\hat{\theta},\theta) = (\hat{\theta}-\theta)^2$ "squared error loss" or "L_2 loss"

$\ell(\hat{\theta},\theta) = |\hat{\theta}-\theta|$ "absolute error loss" or "L_1 loss"

$\ell(\hat{\theta},\theta) = |\hat{\theta}-\theta|^p$ "L_p loss" for $p > 0$.

$\ell(\hat{\theta},\theta) = \int_{\mathcal{X}} \ln\left(\frac{f(x;\theta)}{f(x;\hat{\theta})}\right) f(x;\theta)\, d\vec{x}$ Kullback-Leibler loss for cont. r.v.'s $X_1, \ldots, X_n$

estimator $\Rightarrow$ loss has a distribution

$R(\hat{\theta},\theta) := E_{\mathcal{X}}[\ell(\hat{\theta},\theta)]$

Risk of an estimator: what is the average loss acc'd. to our loss function. If it's sqd error loss...

mean squared error

$R(\hat{\theta},\theta) = E_{\mathcal{X}}[(\hat{\theta}-\theta)^2] = MSE[\hat{\theta}] = E_{\mathcal{X}}[(\hat{\theta}-E_{\mathcal{X}}[\hat{\theta}])^2] = Var_{\mathcal{X}}[\hat{\theta}]$.

If $\hat{\theta}$ is unbiased $\Rightarrow E[\hat{\theta}] = \theta$

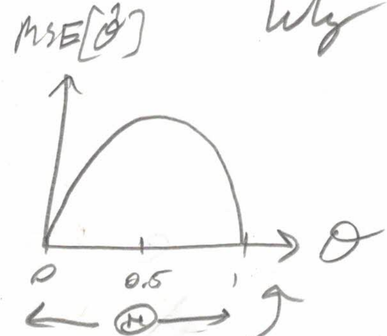
Risk = MSE = Variance for an unbiased estimator under L_2 loss.

under L_2 loss

For a biased estimator,

$$MSE = E_X[(\hat{\theta}-\theta)^2] = E_X[\hat{\theta}^2 - 2\theta\hat{\theta} + \theta^2]$$

Why?



$$= E_X[\hat{\theta}^2] - 2\theta E[\hat{\theta}] + \theta^2$$

Recall: $Var(\hat{\theta}) = E[\hat{\theta}^2] - E(\hat{\theta})^2$

$$= Var[\hat{\theta}] + E[\hat{\theta}]^2 - 2\theta E[\hat{\theta}] + \theta^2$$

$$= Var[\hat{\theta}] + (E[\hat{\theta}] - \theta)^2$$

$$= Var[\hat{\theta}] + Bias[\hat{\theta}]^2$$ "Bias-Variance decomposition of MSE."

$SE[\hat{\theta}] := \sqrt{Var[\hat{\theta}]}$ Standard error is the standard deviation of an estimator.

$E[\hat{\theta}_n] = \theta$ for all iid DGP's $X_1, \dots X_n$ with mean θ, variance σ^2

Back to our example...

$$SE(\hat{\theta}_n) = \sqrt{Var[\hat{\theta}_n]} = \sqrt{Var\left[\tfrac{1}{n} T_n\right]} = \sqrt{\tfrac{1}{n^2} Var[T_n]} = \sqrt{\tfrac{1}{n^2}\, n\, Var(X)} = \frac{\overbrace{SD(X)}^{\sigma}}{\sqrt{n}} = \sqrt{\frac{\theta(1-\theta)}{n}}$$

variance law for iid r.v.'s — for all iid DGP's — Bernoulli

Law of Large #'s for $\hat{\theta}_n = \bar{X}_n$

$\Rightarrow n \to \infty \Rightarrow SE[\hat{\theta}_n] \to 0 \Rightarrow \underset{\text{r.v.}}{\hat{\theta}_n} \to \underset{\#}{\theta}$ w.p. 1

this limit isn't precisely defined. But ignore this detail...

For all estimators which are sample averages of iid r.v.'s, we're kind of done with Goal #1: Pt. Est.

Now that we have an idea about how variable the estimator is we can move to goal #3: standard inference testing. Turns out confidence sets which is goal #2 is harder!