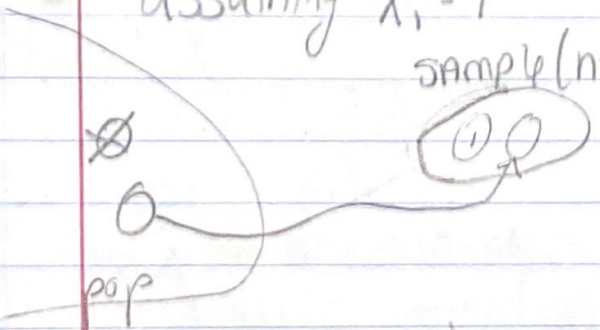
Victoria Lombardi

MATH 369 8/31/20

Lecture 2

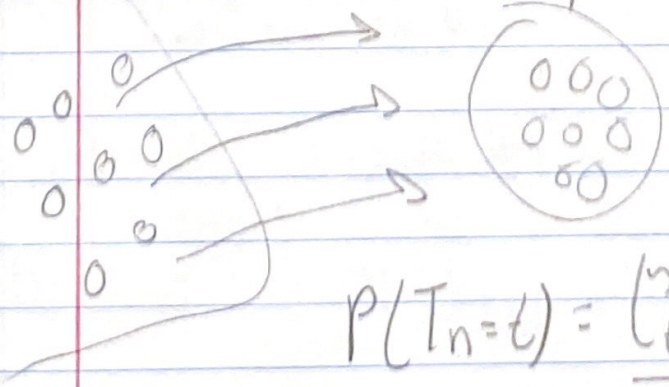
$X_1 \sim \text{Bern}(\theta) = \text{Bern}\left(\frac{\chi}{N}\right)$

Let's draw a second sample from the population assuming $X_1 = 1$



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

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



$T_n = X_1 + \ldots + X_n \sim \text{Hyper}(n, \chi, N)$

Hypergeometric distribution

$$P(T_n = t) = \frac{\binom{\chi}{t}\binom{N-\chi}{N-t}}{\binom{N}{n}}$$

Dealing with the hypergeometric is complicated (but doable). What can we assume to make this go away?

Let $Y, N \to \infty$ but $\theta = \frac{Y}{N}$ — Simplying assumption

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

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

Pretend you work at the iPhone factory, they sample new iphones to ensure they work to ensure the manufacturing is working properly. You check the first one $X_1 = 1$, $X_2 = 1, \ldots X_{100} = 1$.

What population are you sampling from? What is N?

When you estimate θ, you're estimating θ in a "process", i.e. a "data generating process" (DGP) iid Bern(θ).

DGPs and infinite population sampling is the same thing. We no longer care about whether the population is "real", we just assume an iid DGP from now on...

Returning to our main goal: inference
i.e. knowing something about θ from the data.
First subgoal: point estimation. Recall,

$\hat{\theta} = \frac{1}{n}(X_1 + \ldots + X_n)$. $X_1, \ldots, X_n$ are random
realizations from
$X_1, \ldots, X_n$ iid Bern(θ).

$\vec{x} \Rightarrow \hat{p}$ $\Rightarrow \hat{\theta}$

e.g. $\vec{x}[1\,0\,0\,1\,0] \Rightarrow \hat{\theta} = 0.4$

$\Rightarrow \hat{\theta}$ random

but e.g. $\vec{x}[1\,1\,1\,0\,1] \Rightarrow \hat{\theta} = 0.8$

$\hat{\theta}$ is a realization from the random variable
$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} X_i$ called a "statistical estimator"
$= w(X_1, \ldots, X_n)$
or just "estimator". The statistic (estimate,
statistical estimat) is a realization
from the estimator. The distribution
of the estimator $\hat{\theta}$ is called the "sampling
distribution". This sampling distribution and
its properties are very important because
it tells us a lot about our estimates

One property is the estimator's expectation, the
mean over all samples of size n.

$E[\hat{\theta}] = E\left[\frac{1}{n}(X_1 + \ldots + X_n)\right] = \frac{1}{n}\sum E[X_i]$

over all $X_1, \ldots, X_n$ $= \frac{1}{n} n\, E[X_i] = \theta \Rightarrow$ in our iid Bern(θ) setting $\hat{\theta}$ is "unbiased"

expectation

$\text{Bias}[\hat{\theta}] := E[\hat{\theta}] - \theta$. If $\text{Bias}[\hat{\theta}] = 0 \Rightarrow$
$\Rightarrow \hat{\theta}$ is unbiased.
$\text{Bias}[\hat{\theta}] \neq 0 \Rightarrow \hat{\theta}$ is biased.

How far is $\hat{\hat{\theta}}$ from θ?
We define a distance function AKA "loss function", ("error function") (parameter space)

$$l(\hat{\hat{\theta}}, \theta). \quad l: (H) \times (H) \to [0, \infty)$$

$l = 0$ only if $\hat{\hat{\theta}} = \theta$

There are many loss function e.g.

① $l(\hat{\hat{\theta}}, \theta) := |\hat{\hat{\theta}} - \theta|^1$ absolute error loss (L_1 loss)

default * ② $l(\hat{\hat{\theta}}, \theta) := |\hat{\hat{\theta}} - \theta|^2$ squared error loss (L_2 loss)

③ $l(\hat{\hat{\theta}}, \theta) := |\hat{\hat{\theta}} - \theta|^p, \; p > 0$ L_p loss

④ $l(\hat{\hat{\theta}}, \theta) := \int_{x \in X} \ln\left(\frac{f(x;\theta)}{f(x;\hat{\hat{\theta}})}\right) f(x;\theta)\, dx$ *

Kullback-Leibler (KL) loss for continuous rv's

How far away on average are we?

ave. $x_1, \ldots, x_n \to E[l(\theta, \hat{\hat{\theta}})]$

$R(\hat{\theta}, \theta) := E[l(\theta, \hat{\theta})]$
Risk of an estimator

If we use squared error loss,

$R(\hat{\theta}, \theta) = MSE[\hat{\theta}] = E[(\hat{\theta} - \theta)^2]$

"mean squared error" (MSE)

If the estimator is unbiased, does its MSE simplify?

if $\hat{\theta}$ is unbiased, $E[\hat{\theta}] = \theta$

$MSE[\hat{\theta}] = E[(\hat{\theta} - \theta)^2] \overset{\downarrow}{=} E[(\hat{\theta} - E[\hat{\theta}])^2]$

$= Var[\hat{\theta}]$

MSE = Variance

For a biased estimator (i.e. the general case)

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

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

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 of the estimation"