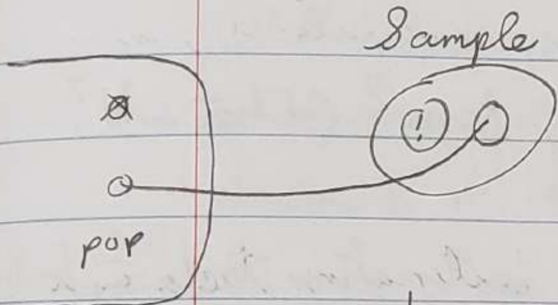
8/31/20

$x_1 \sim Bern(\theta) = Bern(\frac{x}{N})$

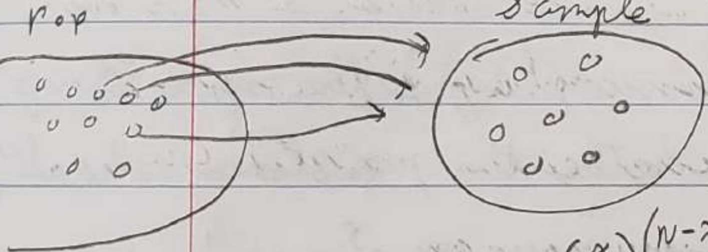
Let's draw a second sample from the population assuming $x_1 = 1$



$P(x_2 = 1 \mid x_1 = 1)$

$$= \frac{x-1}{N-1} < \frac{x}{N} = \theta$$

$$\Rightarrow x_2 \mid x_1 = 1 \sim Bern\left(\frac{x-1}{N-1}\right)$$



$T_n = x_1 + \ldots + x_n \sim Hyper(n, X, N)$

Hypergeometric distribution

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

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

Simplifying assumption

Let $x, N \to \infty$ but $\theta = \frac{x}{N}$

$\lim P(x_2 = 1 \mid x_1 = 1) = \lim \frac{x-1}{N-1} = \theta \Rightarrow x_1, \ldots, x_n \overset{iid}{\sim} Bern(\theta)$

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 theta, you're estimating theta in a "process", ie a "data generating process" (DGP), iid Bern(th

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 theta from the data. First subgoal: point estimation. Recall,

$$\hat{\hat{\theta}} = \frac{1}{N}(x_1 + \ldots x_n) \quad x_1, \ldots, x_n \text{ are random realizations from } X_1 \ldots X_m \overset{iid}{\sim} \text{Bern}(\theta).$$

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

eg $\bar{x} = [10010] \Rightarrow \hat{\hat{\theta}} = 0.4$ $\Rightarrow \hat{\hat{\theta}}$ random

eg $\bar{x} = [11101] \Rightarrow \hat{\hat{\theta}} = 0.8$

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

Think about θ, any est. be?

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

$$\underset{\substack{\uparrow \\ \text{over all} \\ x_1 \dots x_n}}{E}\left[\hat{\theta}\right] = E\left[\frac{1}{n}(x_1 + \dots + x_n)\right] = \frac{1}{n}\sum E[x_i] = \frac{1}{n} n E[x_1] \underset{\substack{\uparrow \\ \text{in our} \overset{iid}{\sim} \text{Bern}(\theta) \\ \text{setting}}}{=} \theta \Rightarrow \text{"}\hat{\theta} \text{ is unbiased"}$$

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

How far is thetahathat from theta?
We define a distance function AKA "loss function" ("error function")
$\ell(\hat{\hat{\theta}}, \theta)$, $\ell : \Theta \times \Theta \to [0, \infty)$. $\ell = 0$ only if $\hat{\hat{\theta}} = \theta$

There are many loss function e.g.

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

default

* $\ell(\hat{\theta}, \theta) := |\hat{\theta} - \theta|^2$ square error loss (L_2 loss)

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

$\ell(\hat{\theta}, \theta) := \int \ln\left(\frac{f(x;\theta)}{f(x;\hat{\theta})}\right) f(x;\theta)\, d\vec{x}$ Kullback-Leibler loss for continuous rv's

$\frac{1}{x} \in \mathcal{X}$

How far away on average are we?

$R(\hat{\theta}, \theta) := \underset{\substack{\uparrow \\ \text{over } x_1, \dots, x_n}}{E}\left[\ell(\theta, \hat{\theta})\right]$

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)

under squared error loss and DGP: iid Bern(theta)

$R(\hat{\theta}, \theta) =$

If the estimator is unbiased, does its MSE simplify?

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

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

MSE = Variance

For a biased estimator (ie the general case),

$$\begin{aligned} MSE[\hat{\theta}] &= E[(\hat{\theta}-\theta)^2] = E[\hat{\theta}^2 - 2\hat{\theta}\theta + \theta^2] \\ &= E[\hat{\theta}^2] - 2\theta E[\hat{\theta}] + \theta^2 \\ &= Var[\hat{\theta}] + E[\hat{\theta}]^2 - 2\theta E[\hat{\theta}] + \theta^2 \\ &= Var[\hat{\theta}] + E[\hat{\theta}] - \theta)^2 \\ &= Var[\hat{\theta}] + Bias[\theta]^2 \end{aligned}$$

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

Bias-variance decompos of MSE

$SE[\hat{\theta}] := \sqrt{var[\hat{\theta}]}$ "standard error of the estimation"