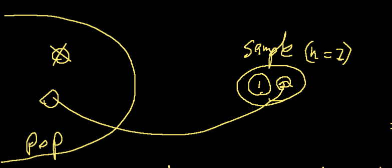
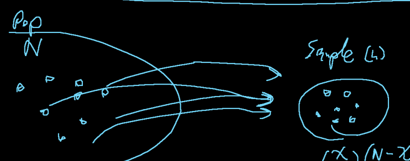
$X_1 \sim \text{Bern}(\theta) = \text{Bern}\left(\frac{\mathcal{X}}{N}\right)$

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



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

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



$T_n = X_1 + \dots + X_n \sim \text{Hyper}(n, \mathcal{X}, N)$ Hypergeometric distribution

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

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

Let $\mathcal{X}, N \to \infty$ but $\theta = \frac{\mathcal{X}}{N}$ $\Rightarrow$ simplifying assumption

$\lim P(X_2 = 1 \mid X_1 = 1) = \lim \frac{\mathcal{X}-1}{N-1} = \theta$ $X_1, \dots, X_n \overset{iid}{\sim} \text{Bern}(\theta)$.

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, ..., 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(theta).

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,

$\hat{\theta} = \frac{1}{n}(x_1 + \dots + x_n)$. $x_1, \dots, x_n$ are random realizations from $X_1, \dots, X_n \overset{iid}{\sim} \text{Bern}(\theta)$.

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

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

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

$\hat{\hat{\theta}}$ is a realization from the rv $\underset{\wedge}{\hat{\theta}} := \frac{1}{n}\sum_{i=1}^{n} X_i = w(X_1, \dots, X_n)$ called a

"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". 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.

$E[\hat{\theta}] = E\left[\frac{1}{n}(X_1 + \dots + X_n)\right] = \frac{1}{n}\sum E[X_i] = \frac{1}{n} n E[X_i] \overset{\text{in our iid Bern}(\theta)\text{ setting}}{=} \theta \Rightarrow$ "$\hat{\theta}$ is unbiased"

over all $X_1, \dots, X_n$

$\text{Bias}[\hat{\theta}] := E[\hat{\theta}] - \theta$. 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 functions e.g.

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

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

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

$\ell[\hat{\theta}, \theta] := \int_{x \in \mathcal{X}} \ln\left(\frac{f(x;\theta)}{f(x;\hat{\theta})}\right) f(x;\theta)\,dx$ Kullblack-Leibler (KL) loss for continuous rv's.

How far away on average are we?

$R(\hat{\theta}, \theta) := \underset{\text{over } X_1, \dots, X_n}{E}[\ell(\theta, \hat{\theta})]$

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? MSE=Variance

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

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

For a biased estimator (ie 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 $Var[\hat{\theta}] = E[\hat{\theta}^2] - E[\hat{\theta}]^2$

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

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

$= Var[\hat{\theta}] + \text{Bias}[\hat{\theta}]^2$ Bias-variance decomposition of MSE

$SE[\hat{\theta}] := \sqrt{Var[\hat{\theta}]}$ "standard error of the estimator".