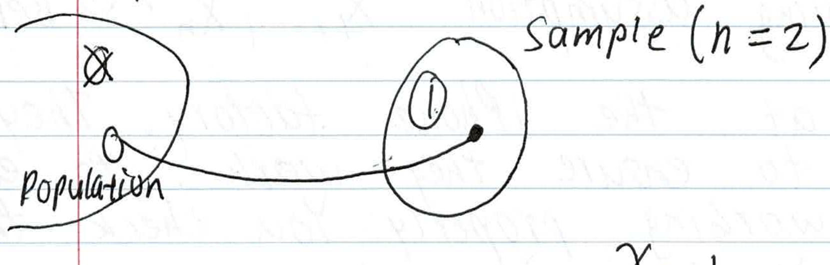
Lecture 2

08/31/20

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

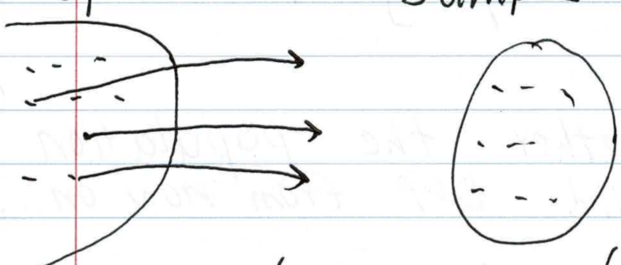
Let's draw a second sample from the population assuming $X_1 = 1$



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

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

Pop Sample (n) $T_1 = X_1 + \dots + X_n \sim Hyper(n, \mathcal{X}, N)$



Hypergeometric Distribution

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

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

Let $\mathcal{X}, N \longrightarrow \infty$ but $\theta = \frac{\mathcal{X}}{N}$

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

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

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 theta, you're estimating theta in a "process", i.e. a "data generating process" (DGP), iid Bernoulli (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 theta from the data

First subgoal: Point estimation. Recall,

$$\underset{\hat{p}}{\overset{\bar{X}}{\hat{\theta}}} = \frac{1}{n}(X_1 + \cdots + X_n)$$

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

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

$\Rightarrow \hat{\theta}$ random

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

$\hat{\hat{\theta}}$ is a realization from the r.v. $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} X_i$

called a "statistical estimator" or just "estimator." The statistic (statistical estimate) 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

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

overall $X_1, \dots, X_n$

$$= \frac{1}{\cancel{n}}\cancel{n}\, E[X_i] = \theta$$

in our iid Bern(θ) setting

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

$\Rightarrow \hat{\theta}$ is unbiased
$Bias[\hat{\theta}] \neq 0 \Rightarrow \hat{\theta}$ is biased

---

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

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

$l = 0$ 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)

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

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

$$\ell[\hat{\theta}, \theta) = \int_{\vec{X} \in X} \ln\left(\frac{f(x;\theta)}{f(x;\hat{\theta})}\right) f(x;\theta)\, d\vec{x}$$

kullblack-Leibler (kL) loss for continuous r.v.'s

How far away on average are we?

$R(\hat{\theta}, \theta) = E[\ell(\theta, \hat{\theta})]$

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 it's MSE simplify? $\Rightarrow$ Yes, MSE = variance

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

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

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 estimator"