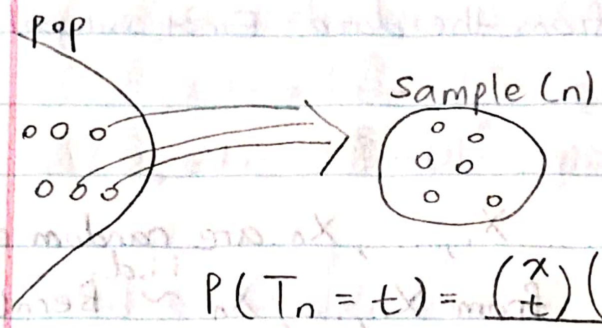
Lecture 2 8/31/2020



$$T_n = \underbrace{X_1 + \dots + X_n}_{\text{Hypergeometric dist. } (n, x, N)}$$

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

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

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

$$\lim P(X_2 = X_1 = 1) = \lim \frac{x-1}{N-1} = \theta \Rightarrow X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \text{Bern}(\theta)$$

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 $X_1 = 1, X_2 = 1, \dots, X_{100} = 1$

What population are you sampling from? What's N?

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

DGPs and ∞ population sampling is the same thing.

We no longer care about whether the population is "real", we just assume an i.i.d. DGP from now on...

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

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

$\overset{\shortparallel}{\bar{x}} \quad \overset{\shortparallel}{\hat{p}}$

e.g. $\bar{x} = [1\,0\,0\,1\,0] \Rightarrow \hat{\hat{\theta}} = 0.4 \Rightarrow \theta$ random (could be any thing)

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

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

One property is the estimator's expectation, the mean overall ~~[illegible]~~ samples of size n.

$E[\hat{\theta}] = E[\frac{1}{n}(X_1 + \ldots + X_n)] = \frac{1}{n}\sum E[X_i] = \frac{1}{\cancel{n}}\cancel{n}\, E[X_i] = \theta \Rightarrow$ "$\hat{\theta}$ is unbiased"

over all $X_1, \ldots, X_n$

in our i.i.d. Bern(θ) setting

Bias$[\hat{\theta}] := E[\hat{\theta}] - \theta$. If

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

" $\neq 0 \Rightarrow$ " biased"

How far is $\hat{\hat{\theta}}$ from θ?

We define a distance fn a.k.a. "loss function", ("error fn")

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

There are many loss fns e.g.

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

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

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

$$\ell(\hat{\hat{\theta}}, \theta) := \int_{\bar{x} \in \mathcal{X}} \ln\left(\frac{f(x;\theta)}{f(x;\hat{\theta})}\right) f(x;\theta)\, d\bar{x}$$ Kullblack - Leibler (KL) loss for continuous R.V.S.

How far away on average are we?

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

Risk of an estimator

If we used 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 i.i.d. Bern(θ)~~
~~$R(\hat{\theta}, \theta)$~~

If the estimator is unbiased, what is its MSE simplify?

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

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

MSE = variance

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

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

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

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

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

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

$SE[\hat{\theta}]$ i $= \sqrt{Var[\hat{\theta}]}$ "Std error of the estimation"