Lecture 4. Intro to Statistics

Basic Concepts: Population, Sample, Parameter, Statistics

- *Sample*= a single draw of data from all potential realizations of that data (random vector $\mathcal{X}$)
- *Population*= the distribution F_X of data vector $\mathcal{X}$.
- *Probability Theory*: given population what can you say about distribution of $g(\mathcal{X})$?
- *Statistics*: given a single realization of $\mathcal{X}$ what can you say about F_X?

Basic Concepts: Population, Sample, Parameter, Statistics

3 types of data:

- *Cross-section* = a set of i.i.d. random vectors $(\mathcal{X} = (X_1, ..., X_n), x = (x_1, ..., x_n))$. If we assume that $X_i \sim F$, then $F_X(x) = \prod_{i=1}^{n} F(x_i)$.
- *Time-series* $X_t, t = 1, ..., T$ allows dependence between consecutive observations . $\mathcal{X} = (X_1, ..., X_T)$ is one realization of a path of a dynamic process.
- *Panel data* $\mathcal{X} = \{X_{it}, i = 1, .., n, t = 1, ..., T\}$ a draw from n independent identically distributed dynamic processes.

Basic Concepts: Population, Sample, Parameter, Statistics

→ fixed, constant, not random, unknown & trying to learn

- *Parameter*= functional of the unknown distribution F_X (population concept!). cdf

p = prob of success

- *Inference*= to render a judgement about a parameter (or population F_X) based on a single draw (sample).
- 3 types of inference: estimation, confidence sets, testing

Ex.1) coin toss $x_i = \begin{cases} 1 & p \\ 0 & 1-p \end{cases}$

pop — $x_1, \dots, x_n$: iid p: parameter

sample $[x_1, \dots, x_n]$ based on one realizn infer p

Ex2) r.v $\leftarrow (x_i, y_i) \sim$ iid F_{xy}

↑ height ↑ income

one draw $\leftarrow \{(x_1, y_1), \dots (x_n, y_n)\}$

excel, sample

i think θ exists

$\theta = \text{cov}(x_i, y_i)$

so look at sample and say something about pop

Basic Concepts: Population, Sample, Parameter, Statistics

- *Statistic*= function of sample $Y = g(\mathcal{X}) = g(X_1, ..., X_n)$
- It is a random variable! so interested in its distⁿ & what we can tell about it
- The distribution of a statistic =*sampling distribution* how it appear from different samples

Sample Mean and Sample Variance

- sample mean ($\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$) — "average"
- sample variance ($s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$)

$\mathbb{E}[\bar{x}_n] = \mu$
mean of $\bar{x}_n$ is μ
var of $\bar{x}_n = \sigma^2/n$
$\mathbb{V}ar[\bar{x}_n]$

$\mathbb{V}ar(x_i) = \sigma^2 = \mathbb{E}(x_i - \mu)^2$

Lemma 1.

If $X_1, \ldots, X_n \sim$ i.i.d.F is a random sample of size n from a population distribution with mean $\mu = \mathbb{E}X_i$ and variance $\sigma^2 = \mathbb{V}ar(X_i)$, then $\mathbb{E}[\overline{X}_n] = \mu$ and $\mathbb{E}[s^2] = \sigma^2$.

$\mathbb{E}\bar{x}_n = \mu$ ~ pop; parameter; unknown
can be calc from from sample

$\mathbb{V}ar(\bar{x}_n) = \sigma^2/n$

$s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ (sample var) rv

& wts mean of s^2 i.e $\mathbb{E}s^2 = \sigma_n^2$

Empty slide

□ $y_i = x_i - \mu$ $\quad var(y_i) = \sigma^2$ $\quad \bar{y}_n = \bar{x}_n - \mu$ $\quad y_i - \bar{y}_n = x_i - \bar{x}_n$

so $s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{Y})^2 = \frac{1}{n-1}\sum_{i=1}^{n}[y_i^2 - 2\bar{Y}y_i + \bar{Y}^2]$

$= \frac{1}{n-1}\sum_{i=1}^{n} y_i^2 - 2\bar{Y}\underbrace{\sum_{i=1}^{n} y_i}_{n\bar{Y}} + n\bar{Y}^2]$

$s_n^2 = \frac{1}{n-1}[\sum y_i^2 - n(\bar{Y})^2]$

$EY_i = 0$

Taking Exp,

$E s_n^2 = \frac{1}{n-1}[\sum_{i=1}^{n} E y_i^2 - n\, E(\bar{Y})^2]$

$Var(Y_i) = \sigma^2$ $\quad Var(\bar{Y}) = \sigma^2/n$

$$= \frac{1}{n-1}\left[n \cdot \sigma^2 - n\, E(\bar{\xi})^2\right] = \sigma^2$$

$E(X-a)^2 \to \min_a$ $\quad a = EX = u$

$\frac{1}{n}\sum (x_i - a)^2 \to \min_a$

$a = \frac{1}{n}\sum x_i = \bar{x}$ (avg.)

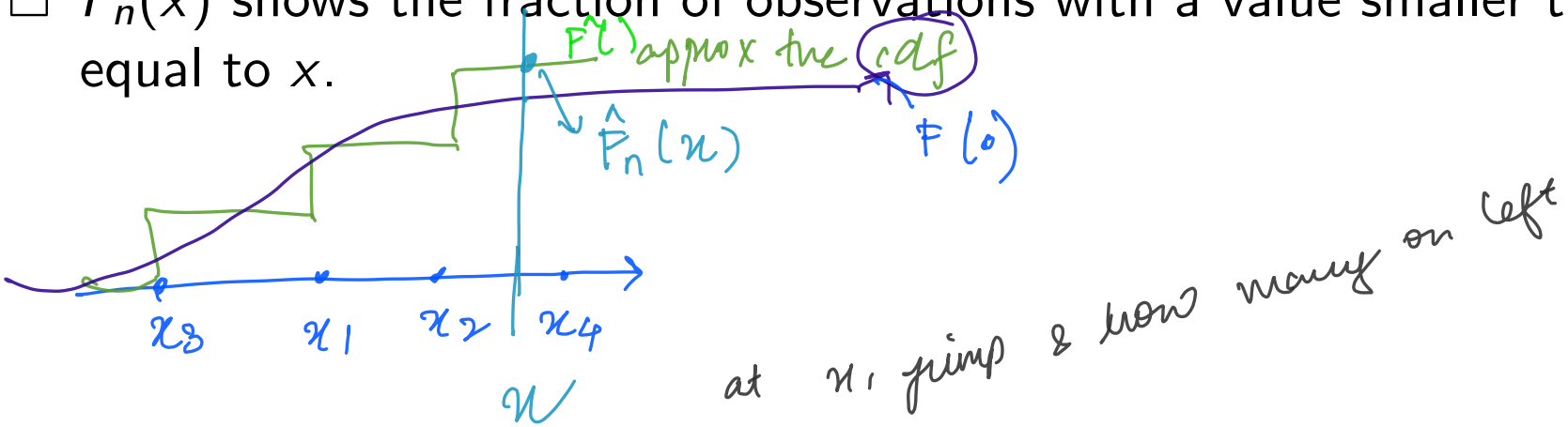
But u is pop, if given from outside

$E\frac{1}{n}\sum (x_i - u)^2 = \sigma^2$ // pop. variance

$\frac{1}{n}\sum (x_i - u)^2 \geq \frac{1}{n}\sum (x_i - \bar{x})^2$

Empirical distribution function

- $\widehat{F}_n$ is the cdf of the distribution that places mass $1/n$ at each data point X_i.
- $\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(X_i \leq x)$,
- $\widehat{F}_n(x)$ shows the fraction of observations with a value smaller than or equal to x.



Empirical distribution function

LLN : Law of large no's

Theorem 1.

If we have a random sample $X_1, \ldots, X_n$ of size n from a distribution with cdf F, then for any $x \in \mathbb{R}$, $\mathbb{E}[\widehat{F}_n(x)] = F(x)$ and $\mathbb{V}ar(\widehat{F}_n(x)) \to 0$ as $n \to \infty$. As a consequence, $\widehat{F}_n(x) \to_p F(x)$ as $n \to \infty$.

$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{X_i \le x\}$

$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} Y_i = \bar{Y}$

$Y_i = \mathbb{1}\{X_i \le x\} \sim \text{Ber}(p)$

$p = P\{X_i \le x\} = F_X(x)$

$\mathbb{E}\,\hat{F}_n(x) = \mathbb{E}\,\bar{Y} = \mathbb{E}\,Y_i = p = F_X(x)$

LLN: $\hat{F}_n(x) \xrightarrow{P} F_X(x)$ as $n \to \infty$

my fn at every point does converge to cdf

a fn converges to fn (both monotonic not have much

Empirical distribution function

space to depart from each other)

Theorem 2 (Glivenko-Cantelli).

If $X_1, ..., X_n$ is a random sample from a distribution with cdf F, then

$$\sup_{x \in \mathbb{R}} |\widehat{F}_n(x) - F(x)| \to_p 0.$$

Ways to find the distribution of a statistic

analytical, rarely used!

Method 1: exact distribution

- Question: If you know the distribution of $\mathcal{X}$, what is the distribution of $Y = g(\mathcal{X})$?
- Can you do this analytically?
- Very rare case.

6x1 $X_1, \ldots, X_n \sim iid\ N(\mu, \sigma^2)$

$Y = \bar{X} \sim N(\mu, \sigma^2/n)$ —

6x2 $X_1, \ldots, X_n \sim iid$ $x_i = \begin{cases} 1 & p \\ 0 & 1-p \end{cases}$ || Bernoulli

$Y = \sum_{i=1}^{n} X_i$ $P\{Y=k\} = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$ || Binomial

(Y can take values 1, 2 ... upto n)

Ways to find the distribution of a statistic :

(simulation)

Method 2: Monte- Carlo (Numerical way to implement Method 1)

- Question: If you know the distribution of $\mathcal{X}$, what is the distribution of $Y = g(\mathcal{X})$?
- For $b = 1, ..., B$, simulate $\mathcal{X}_b^* = (X_{1b}^*, ..., X_{nb}^*)$, where $X_{ib}^* \sim$ i.i.d F
- Calculate $Y_b^* = g(\mathcal{X}_b^*)$.
- We can approximate distribution of Y:
  - $F_Y(s) \approx \frac{1}{B}\sum_{b=1}^{B} \mathbb{I}\{Y_b^* \le s\}$;
  - $P\{Y \in A\} \approx \frac{1}{B}\sum_{b=1}^{B} \mathbb{I}\{Y_b^* \in A\}$;
  - α-quantile of Y: $q_Y(\alpha) \approx Y^*_{(\lfloor \alpha B \rfloor)}$;
  - $\mathbb{E}Y \approx \frac{1}{B}\sum_{b=1}^{B} Y_b^*$;
  - $\mathbb{V}ar(Y) \approx \frac{1}{B-1}\sum_{b=1}^{B}\left(Y_b^* - \frac{1}{B}\sum_{s=1}^{B} Y_s^*\right)^2$.

You control the quality of approximation!

How do I simulate from same distn as pop
v.good simulation ; # control B can do as many & get good-

Empty slide

Ways to find the distribution of a statistic

CLT: start w unknown distribution — stabilize & get normal

Method 3: Asymptotic approximation (most often used method)

- What if you do not know the distribution of $\mathcal{X}$? Is it helpless to figure out the distribution of $Y = g(\mathcal{X})$?
- You can approximate distribution of some statistics by the limit as $n \to \infty$
- Looking for statement

$$F_{Y,n}(t)\{Y = g(\mathcal{X}_n) \leq t\} \Rightarrow F_\infty(t).$$

- We rely on CLT, delta-methods, Slutsky theorem
- We do not control the quality of approximation (depends on n and speed of convergence)

$X_1, \ldots, X_n \sim$ iid

- to get a distb that can be normal.

Empty slide

Ways to find the distribution of a statistic

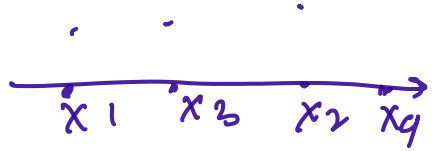
* simulate data from my empirical cdf which is actually close to the true/real cfd of pop.

Method 4: Bootstrap*

- What if you know the distribution of $\mathcal{X}$? Is it helpless to figure out the distribution of $Y = g(\mathcal{X})$?
- Idea: to approximate F_X by some close distribution.
- If Y depends *continuously* on F_X, and if $F_X \approx \widehat{F}$, then we approximate Y
- The last step can be done by Monte-Carlo.

empirical cdf: draw from my sample.

I have a sample



Ways to find the distribution of a statistic

Method 4: Bootstrap Glivenko-Cantelli' theorem: $\widehat{F}_n(x) \approx F(x)$.

- For $b = 1, ..., B$, simulate $\mathcal{X}_b^* = (X_{1b}^*, ..., X_{nb}^*)$, where $X_{ib}^* \sim$i.i.d$\widehat{F}_n$ draw X_{ib}^* independently with replacement from the set $\{x_i, i = 1, ..., n\}$.
- Calculate $Y_b^* = g(\mathcal{X}_b^*)$.
- We can approximate distribution of Y by that of Y_b^*:
  - $F_Y(s) \approx \frac{1}{B}\sum_{b=1}^{B} \mathbb{I}\{Y_b^* \le s\}$;
  - $P\{Y \in A\} \approx \frac{1}{B}\sum_{b=1}^{B} \mathbb{I}\{Y_b^* \in A\}$;
  - α-quantile of Y: $q_Y(\alpha) \approx Y^*_{(\lfloor \alpha B \rfloor)}$;
  - $\mathbb{E}Y \approx \frac{1}{B}\sum_{b=1}^{B} Y_b^*$;
  - $\mathbb{V}ar(Y) \approx \frac{1}{B-1}\sum_{b=1}^{B}\left(Y_b^* - \frac{1}{B}\sum_{s=1}^{B} Y_s^*\right)^2$.

You do not control the quality of approximation.

Empty slide

Ways to find the distribution of a statistic

Plug-in approach

- A random sample $X_1, ..., X_n$ from a population F
- Want to estimate $\theta = T(F)$ (functional of population)
- Plug-in idea: if $\widehat{F}_n(x) \approx F(x)$, then $T(\widehat{F}_n) \approx \theta$.
- This is informal idea.
- Example

$$\mu = \mathbb{E}X_i = \int x dF(x) \approx \int x d\widehat{F}_n = \overline{X}_n$$

if you know cdf know mean

if you know F, can get mean, var.

$\int x \, d\hat{F}_n(x)$ puts equal weight on all obs

$= \frac{1}{n} \sum x_i = \overline{X}$

$m = \text{med } x_i : F(m) = 1/2$

Try to find $\hat{m}$ s.t. $\hat{m} : \hat{F}_n(\hat{m}) = 1/2$

$X_{(\lceil n/2 \rceil)}$

Empty slide

Sample

$$Var = \int (x - \int x dF)^2 dF$$

- plug in Var $\quad \hat{Var} = \frac{1}{n} \sum (x_i - \bar{X})^2$

i