Tutorial 2: Statistics Review

Hammad Shaikh

September 22, 2019

(parameter) $\mu \leftarrow$ Pop^n

sample

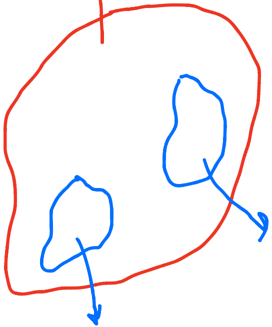
$\downarrow$

$x_1, x_2, \ldots, x_n$

$\downarrow$

(estimator) $\bar{X}$

Pop$^n \rightarrow \mu$



$S_2 \rightarrow \bar{x}_2 = \$5000$

$S_1 \rightarrow \bar{x}_1 = \$1000$

Inferential Statistics Overview

- Population: set of all items (ex. individuals) of interest

  ↳ UTM econ graduates

- Parameter: number describing a characteristic about the population

  ↳ μ = Avg. salary of UTM econ grad

- Sample: subset of the population

  ↳ $n = 100$ UTM econ grads

- Statistic: number describing a characteristic about the sample

  ↳ Data: $x_1, x_2, \ldots, x_{100} \rightarrow \bar{x} = \frac{\sum x_i}{n}$

Cross Sectional Data Example

Focus of ECO375

(multiple units in one time period)

Table: Grade 4 Achievement Outcomes

Student	Math	Reading	Science	Grade
Hammad	80	70	60	4
Alex	65	75	85	4
⋮	⋮	⋮	⋮	⋮
Bob	60	70	80	4

variables

multiple units

- Variables are math, reading, and science test scores
- Time period in this context is grade 4
- Unit of observation is students

Time Series Data Example

Studied more in ECO475

(Follow one unit over time)

Table: Annual Average GPA for UTM

School	Average GPA	Year
UTM	3.45	2000
⋮	⋮	⋮
UTM	3.61	2018

- What is the variable?

  ↳ Avg. GPA
- What is the time period?

  ↳ Year
- What is the unit of observation?

  ↳ School (UTM)

Panel Data Example

Studied more in ECO475

(multiple units over time)

Table: Educational Attainment in Canada

Province	HS Graduation Rate	Years of Education	Year
Ontario	70	13	2000
⋮	⋮	⋮	⋮
Ontario	86.5	16	2018
⋮	⋮	⋮	⋮
Alberta	55	10	2000
⋮	⋮	⋮	⋮
Alberta	70	14	2018

- What are the variables? Grad rate & Yrs of educ.
- What is the time period? Year
- What is the unit of observation? Provinces

Summary Statistics

- The first table in a research paper generally describes the data
  - Known as the “Summary Stats" table

- Common statistics used to describe variables:
  - Central tendency: mean and median
    - mean: $\bar{X} = \frac{x_1 + \ldots + x_n}{n}$
  - Variability: variance, standard deviation, and range
    - variance: $Var(X) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$

↳ range = max(X) − min(X)

Example of Summary Statistics Table

Summary Stats. of real survey data from U.S.

Table: Summary Statistics of Kindergarten Students

Variable	Mean	Std. Dev.	Min.	Max.	N
Male Student	0.512	0.5	0	1	21396
Age (months)	65.48	4.29	54	79	18066
No. Books	72.79	59.52	0	200	17912
Non-english	0.14	0.35	0	1	20007

- How big is the data?

  Around 21000 students

- Why are the N's different?

  Missing values due to non-survey response

- Average student owns 73 books?

  ↳ outliers and large variance

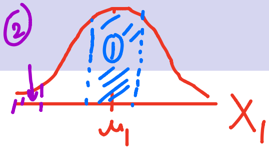
Random Variables

- Random process: A procedure, involving a population, that can conceptually be repeated, and produces outcomes
- A random variable assigns a number to each outcome of a random process
  - Discrete RV takes on finite number of values

    ↳ letter grade or GPA in course
  - Continuous RV takes on infinite number of values

    ↳ Course avg. or Salary

Distribution of Random Variables

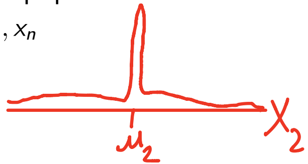


$X \sim N(\mu, \sigma^2)$

- Random variables (RVs) are associated with probability distribution function (pdf)
  - The pdf characterizes the likelihood that the RV takes on values in a particular set

- RVs are usually denoted by capital letters (X) and their realizations are lower case (x)

- Samples are drawn from the population distribution
  - Sample of size n: $x_1, \ldots, x_n$



Estimating Parameters

Estimator: $f(X_1, X_2, \ldots, X_n) \sim F$

unit 1

unit 2

- Recall population parameters are typically unknown
  - Population in economics are generally very large

↳ μ = avg. salary of UTM grad

- Estimator: a function of the observed RVs $\widehat{X_n}$ that is informative about the population parameter
  - Is an estimator associated with a probability distribution?

↳ Yes, since diff. sample → diff. estimates

- Estimate: a realization of $\widehat{X_n}$ obtained by evaluating the estimator at a particular data set
  - Different samples will likely lead to different estimates

↳ Dbn represent uncertainty in estimates

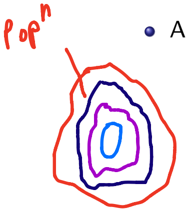
Properties of Estimators

- Suppose the population mean is μ and $\widehat{X_n}$ is its estimator

- Unbiasedness: on average the estimator is right
  - $E(\widehat{X_n}) = \mu$ for all n

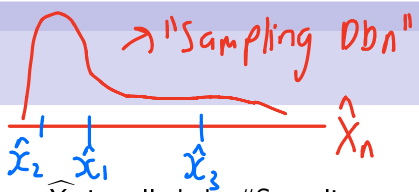
    ↳ Mean of many sample estimates approx the pop^n param.

- Consistency: the truth is eventually discovered
  - As $n \to \infty$ then $\widehat{X_n} \xrightarrow{p} \mu$ (convergence in probability)
    - A bit more formally, as $n \to \infty$, then $Pr(\widehat{X_n} \to \mu) = 1$



$n_1 < n_2 < n_3$

Sampling Distributions

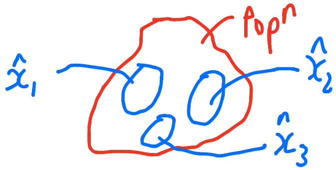


- The distribution of a estimator $\widehat{X_n}$ is called the "Sampling distribution"
  - Sampling distribution models uncertainty in the estimates produced from varying samples

- We are often interesting in the sampling distribution of $\bar{X}$

  ↳ $\bar{X}$ used to infer μ

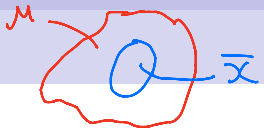
- Central limit theorem says that $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$ under:
  - The sample is independently and identically drawn (IID) from the population
  - Sample size is sufficiently large



Estimator Example

$\bar{X} \xrightarrow{P} \mu$

$\Uparrow$



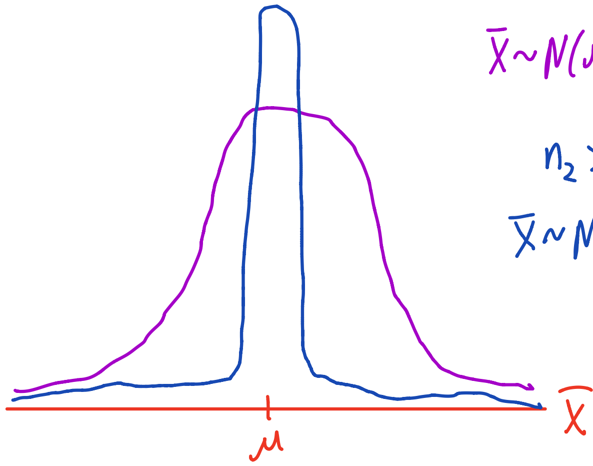
CLT $\Rightarrow \bar{X} \sim N(\mu, \frac{\sigma^2}{n})$, $n \to \infty$, $V(\bar{X}) \to 0$

- Want to estimate average salary of UTM graduate
  - Parameter of interest: $\mu =$ average salary of all UTM graduates (suppose there are N total graduates)
  - Estimate μ using $\bar{X} =$ average salary for n graduates (note n is usually much smaller than N)

- If CLT holds, is $\bar{X}$ a consistent and unbiased estimator of μ?

i) unbiased: $E(\bar{X}) = \mu$, $\bar{X} = \dfrac{X_1 + X_2 + \dots + X_n}{n}$

$\hookrightarrow E(\bar{X}) = \dfrac{E(X_1) + E(X_2) + \dots + E(X_n)}{n} = \dfrac{n \cdot \mu}{n} = \mu$



$\bar{X} \sim N(\mu, \frac{\sigma^2}{n_1})$

$n_2 \gg n_1$

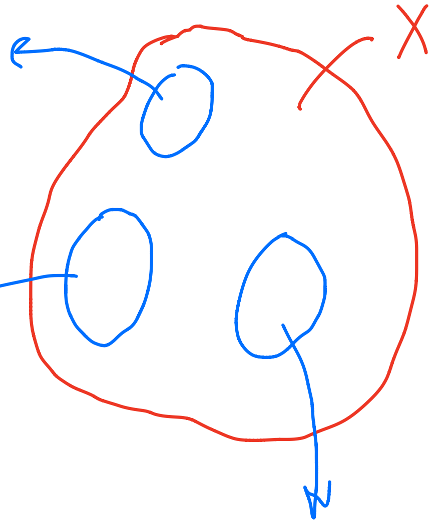
$\bar{X} \sim N(\mu, \frac{\sigma^2}{n_2})$

Stata

S_2

$\downarrow$

$0.21 = \bar{x}_2$



$X \sim Beta(\frac{3}{2}, 5)$

$\downarrow$

$E(X) = \mu$

$\| $

$\frac{1.5}{1.5+5} = 0.23$

$S_1 = x_1, \ldots, x_{100}$

$\downarrow$

$\bar{x}_1 = 0.224$

$S_3 \to \bar{x}_3 = 0.25$

Repeatdly draw samples and obtain:

$\bar{x}_1, \bar{x}_2, \bar{x}_3, \ldots, \bar{x}_{100}$

Hist

$\nearrow$ pbn of $\bar{X}$

(sampling dbn)