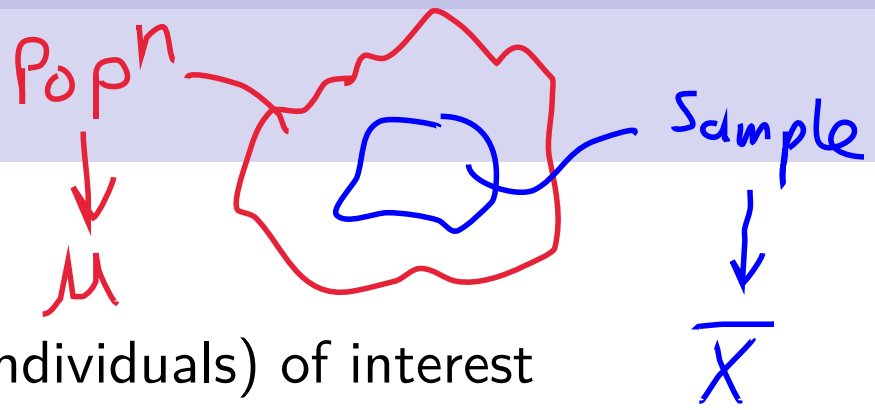
Lecture I: Statistics Review

Hammad Shaikh

January 4, 2018

Inferential Statistics Overview



- Population: set of all items (ex. individuals) of interest
  → Ex. All Canadian adults
- Parameter: number describing a characteristic about the population
  → μ = Avg. salary of candian adults
- Sample: subset of the population
  → Salaries of n people $x_1, x_2, \ldots, x_n$
- Statistic: number describing a characteristic about the sample
  → $\bar{x} = \frac{x_1 + \ldots + x_n}{n}$ = sample mean
- We want to make inferences about the population parameters given the sample

Types of Data

- Cross sectional: variable(s) in same time period measured for different units
  - Math and reading scores for students in grade 4
- Time series: variable(s) for same unit measured at different time periods
  - Yearly average GPA at UTM for the past 10 years
- Panel data: variable(s) measured for a range of units and time periods
  - High school graduation rates for all provinces for past 10 years

→ Cross sectional and panel most common in education economics.

Cross Sectional Data Example

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

Time Series Data Example

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 (University)

Panel Data Example

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? HS Grad. and Educ.
- What is the time period? Year
- What is the unit of observation? Province

Summary Statistics

- The first table in a research paper generally describes the data
  - Known as the “Summary Stats" table
  - Usually contains mean, variance, range, and number of observations

- 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: $Var(X) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$

↳ Avg. squared distance from mean

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?

21396 students, Gender data is complete

- Why are the N's different?

Missing data since people don't respond to all questions on survey.

- Average student owns 73 books?

Outliers in data and high variability.

Random Variables

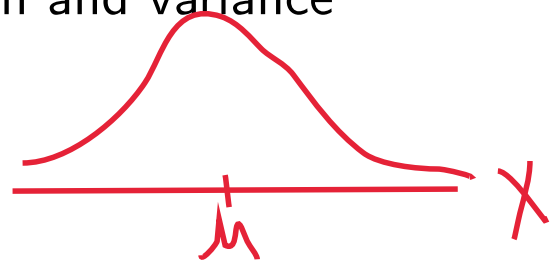
- Random process: A procedure, involving a population, that can conceptually be repeated, and produces outcomes
- A random variable assigns a number to each outcome of a random process
  - Can be discrete or continuous
    - Discrete RV takes on finite number of values
    - Continuous RV takes on infinite number of values
  - Example: Letter grade (A, B, C, and D)
    - ↳ discrete RV
  - Example: Average SAT score in a school
    - ↳ continuous RV

Distribution of Random Variables

- 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: $x_1, \ldots, x_n$
- Most common distribution is the "Normal distribution"
  - Characterized by two components: mean and variance

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

$E(X) = \mu, \; V(X) = \sigma^2$



Estimating Parameters

- Recall population parameters are typically unknown
  - Population in economics are generally very large
- Estimator: a rule that maps underlying RVs into another RV $\widehat{X_n}$ that is informative about the population parameter
  - Is an estimator associated with a probability distribution?
- 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

Estimator: $\hat{X}_n = f(X_1, X_2, \ldots, X_n)$ is a RV

Estimate: $\tilde{x} = f(\underbrace{x_1, x_2, \ldots, x_n}_{\text{sample}})$ is a number

Properties of Estimators

Ex. $\hat{p} = \frac{\#\text{ tails}}{n \text{ flips}}$ (assume fair coin) $p = \frac{1}{2}$ is true prob. of tails

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

$\hat{p}$ is unbiased

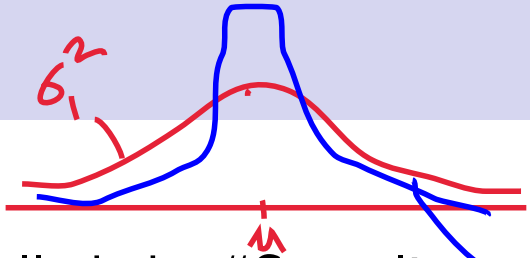
- Unbiasedness: on average the estimator is right
  - $E(\widehat{X_n}) = \mu$ for all n $\rightarrow$ $E(\hat{p}) = \frac{n/2}{n} = \frac{1}{2} = p$

- 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 $Pr(\widehat{X_n} \to \mu) = 1$
  - Example: if you flip a coin a very large number of times, the proportion of heads will be close to 0.5
    - Very likely since prob. of heads $= 0.5$, but not guaranteed

$\hat{p}$ is consistent

$V(\hat{p}) = \frac{p(1-p)}{n} \xrightarrow[n \to \infty]{} 0 \Rightarrow \hat{p} \xrightarrow{p} \frac{1}{2} = p$

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

$\sigma_2^2 < \sigma_1^2$

- We are often interesting in the sampling distribution of $\bar{X}$
- 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
- Law of large numbers: as $n \to \infty$ then $\widehat{X_n} \xrightarrow{p} \mu$ if
  - Sample is IID from the population

$\bar{X}$ is Consistent

$\rightarrow V(\bar{X}) = \frac{\sigma^2}{n} \to 0$, $\bar{X}$ collapses to μ

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

Estimator Example

- Want to estimate average salary of UTM graduate
  - Parameter of interest: μ = 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? - Yes

→ LLN says $\bar{X}$ consistent if sample iid

$E[\bar{X}] = \mu$ since $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$

hence $\bar{X}$ unbiased

Hypothesis Testing

good $\begin{cases} H_0: p = \frac{1}{2} \text{ (fair coin)} \\ H_1: p \neq \frac{1}{2} \text{ (unfair coin)} \end{cases}$ wrong $\begin{cases} H_0: \hat{p} = \frac{1}{2} \\ H_1: \hat{p} \neq \frac{1}{2} \end{cases}$

- A hypothesis is a statement about population parameters
  - Sample statistics don't belong in a hypothesis
- Null hypothesis: statement relating to the status quo (innocent until found guilty beyond reasonable doubt)
  - Example, H_0 : Teacher did not cheat, fair coin
- Alternative hypothesis: statement taking the opposite stance than the null hypothesis
  - Example, H_1 : Teacher cheated, biased coin

Conducting a Hypothesis Test

- Starting point is a estimator for the parameter(s) of interest
  - Realization from the estimator using a sample also required
- Assume H_0 is true
  - Identifies distribution for estimator $\widehat{X}$
- Compute the probability of obtaining a value for $\widehat{X}$ at least as extreme as that obtained from sample
  - This is known as the p-value
- Define significance level $\alpha \in \{0.01, 0.05, 0.1\}$
  - Fail to reject H_0 if p-value $> \alpha$
  - Reject H_0 if p-value $< \alpha$

Pvalue $< \alpha$ says result is very surprising if H_0 is true, hence H_1 more favourable.

Hypothesis Test Example

- UofT has around 6000 students that enrol in ECO100. Suppose the dean claims that ~~at least~~ more than 80% of them complete the course. The dean asks you to test this claim. You have a survey of 500 students who initially enrolled in ECO100, 420 students report completing the course

(a) What is the population parameter of interest?

p = Prop. of initially enrolled students who completed EC100 (out of 6000)

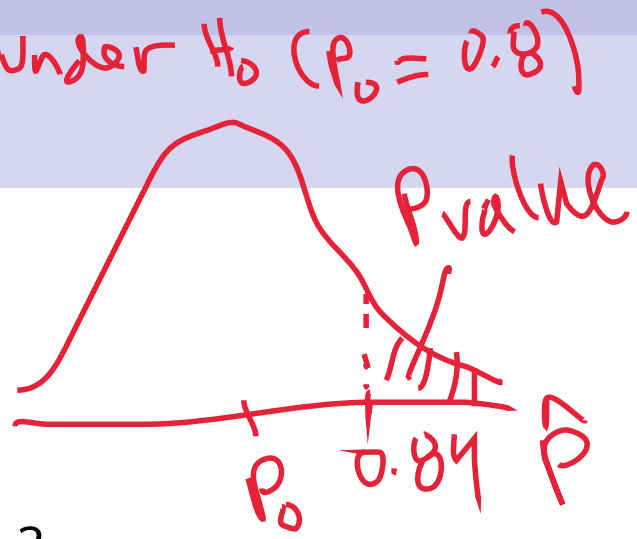
(b) How to estimate population parameter?

$\hat{p} = \frac{\#\text{ completed EC0100}}{500}$ (out of survey)

(c) How to set up hypothesis test (H_0 and H_1)?

$H_0: p \leq 0.80 \ (p = 0.8)$, $H_1: p > 0.80$

Hypothesis Test Example Cont.



(d) What is the distribution of estimator?

It can be shown: $\hat{p} \sim N\left(p_0, \frac{p_0(1-p_0)}{n}\right)$

Using CLT

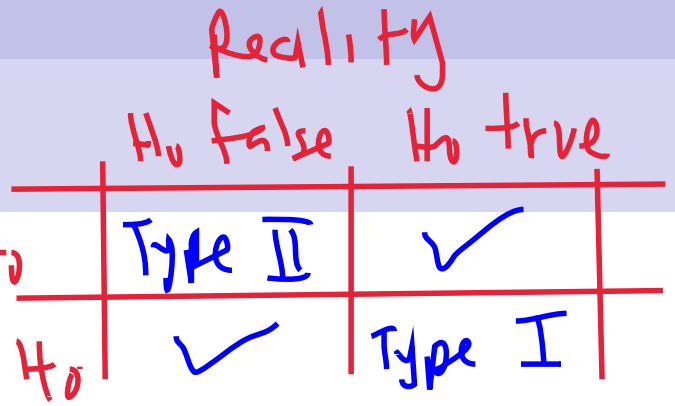
where $p_0 = 0.80$, $\bar{p} = 0.84$ estimate

(e) What is p-value and conclusion?

$\text{Pvalue} = Pr(\hat{p} > 0.84) = 0.009 < \alpha = 0.05$

Since Pvalue $< \alpha$, reject H_0 in favour of H_1.

Type I and Type II Errors



- We can never be 100% sure whether the conclusion obtained from the hypothesis test is correct
  - The conclusion may be incorrect (mistakes are possible)
- Type I Error: Rejecting a true null hypothesis ("false positive")
  - Hypothesis test says a honest teacher cheated
- Significance level $\alpha = \Pr(\text{Type I Error})$
- Type II Error: Failing to reject a false null ("false negative")
  - Hypothesis test says a cheating teacher did not cheat

→ There is a tradeoff between type I and type II error. ↓ type I will ↑ type II.

Summary of Statistics Review

- Economic policies are always associated with some degree of uncertainty regarding its effectiveness
  - Need to use probability theory to model this uncertainty
- Setup probability framework
  - Population, random variable, and distribution
- Estimation
  - Define estimator for parameter of interest (hopefully unbiased and consistent)
- Hypothesis testing
  - Try rule out the null hypothesis (ex. policy not effective) beyond a reasonable doubt

Econometrics Overview

- What is Econometrics?
  - Statistics applied to economics with emphasis on causal inference
  - Why do we need Econometrics?
    - Economics theory suggests important relationships, but usually doesn't suggests quantitative magnitudes of causal effects
- What is the quantitative effect of reducing class size on student achievement?
- How does another year of education change earnings?
- What are the long term effects to subsidizing preschool programs?