ECON 340
Economic Research Methods

Div Bhagia

Lecture 2
Empirical Distribution & Measures of Central Tendency

Describing Data

A dataset is a collection of variables. Each variable contains multiple observations of the same measurement.

Types of variables:

- *Categorical*: gender, race, education (*binary*: two categories)
- *Continuous*: income, age, GPA

How do we summarize the information contained in a variable?

The Empirical Distribution

How often do different values occur?

For categorical variables:

$$f_k = \frac{n_k}{n} = \frac{\text{observations in category } k}{\text{total observations}}$$

f_k captures the relative frequency of outcome k.

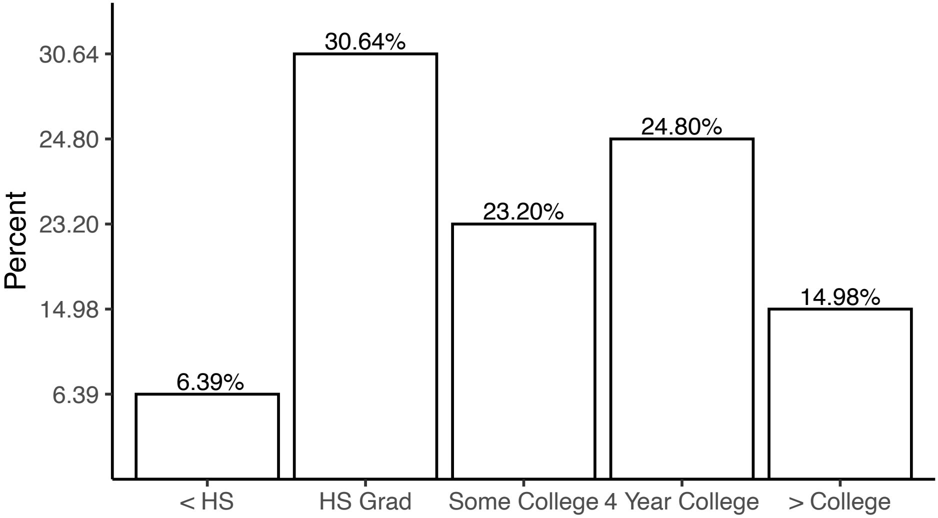
Frequency Distribution Table

Education	Count	Percent
$<$ HS	1540	6.39
HS Grad	7388	30.64
Some College	5595	23.20
4 Year College	5979	24.80
$>$ College	3611	14.98
Total	24113	100

Frequency Distribution Table

Education	Count	Percent	Cumulative
$<$ HS	1540	6.39	6.39
HS Grad	7388	30.64	37.03
Some College	5595	23.20	60.23
4 Year College	5979	24.80	85.02
$>$ College	3611	14.98	100.00
Total	24113	100	

Histogram: Education



The Empirical Distribution

What about continuous variables?

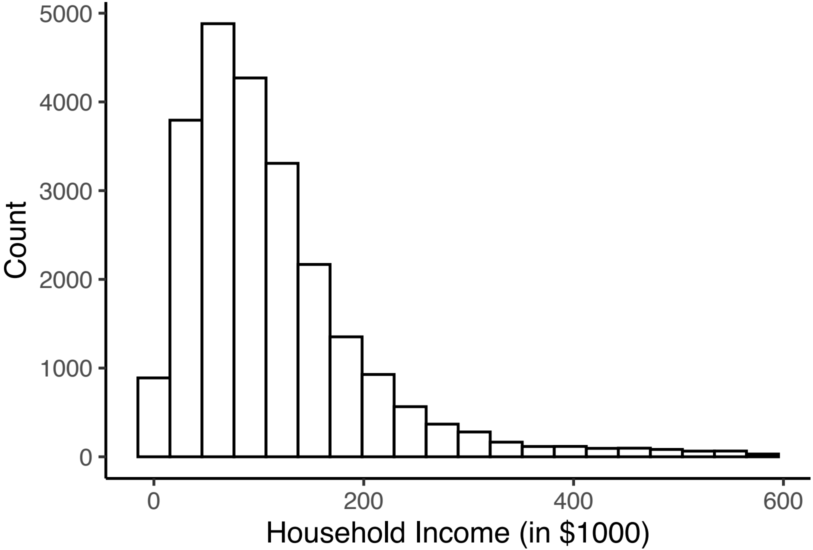
The Empirical Distribution

What about continuous variables?

How often do different values occur in a particular interval?

$$f_k = \frac{\text{observations in } \textit{interval } k}{\text{total observations}}$$

Histogram: Household Income



Source: American Community Survey (ACS) 2019

Measures of Central Tendency

Mean: is the average value

Median: is the middle value

Mode: is the number that is repeated more often than any other

Example: 5, 5, 10, 10, 10, 10, 20

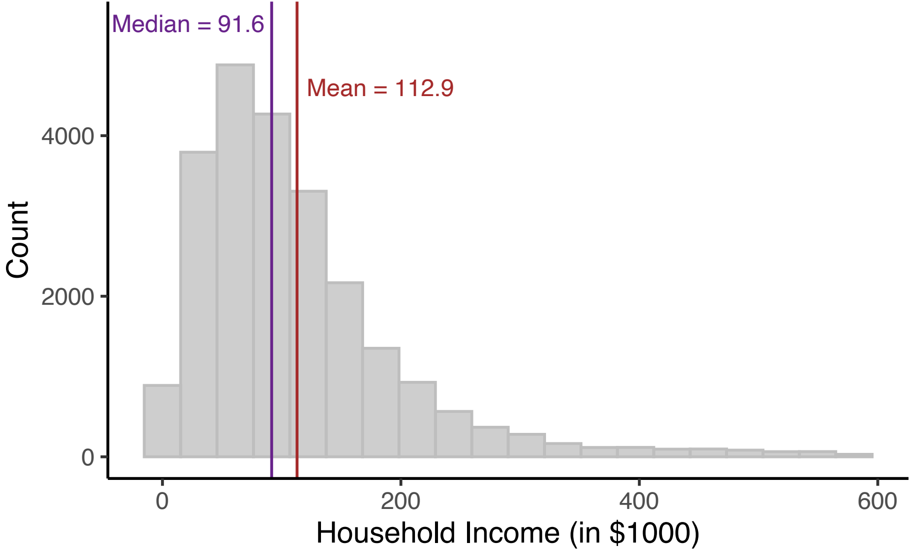
Mean

To calculate the mean:

$$\bar{X} = \frac{\text{sum of all observations}}{\text{number of observations}} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Use $\bar{X}$ to denote the sample mean and μ to denote the population mean.

Mean vs Median



Source: American Community Survey (ACS) 2019

Mean vs Median

- Mean household income: \$112,900
- Median household income: \$91,600

Why are mean earnings higher than the median?

Percentiles

The P^{th} **percentile** is a value such that $P\%$ of observations are at or below that number.

25th percentile a.k.a 1st quartile
75th percentile a.k.a 3rd quartile

What is the 50th percentile called?

More about Mean

- $\sum_{i=1}^{n} X_i = n\bar{X}$

More about Mean

- $\sum_{i=1}^{n} X_i = n\bar{X}$

- Deviations from the mean are always zero

$$\sum_{i=1}^{n}(X_i - \bar{X}) = \sum_{i=1}^{n} X_i - n\bar{X} = n\bar{X} - n\bar{X} = 0$$

More about Mean

- $\sum_{i=1}^{n} X_i = n\bar{X}$

- Deviations from the mean are always zero

$$\sum_{i=1}^{n}(X_i - \bar{X}) = \sum_{i=1}^{n} X_i - n\bar{X} = n\bar{X} - n\bar{X} = 0$$

- We can always write

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i = \sum_{i=1}^{n} \frac{X_i}{n}$$

An easier way to calculate mean

- If data is grouped, we can use the frequency distribution table to calculate the mean:

$$\bar{X} = \frac{\sum_{k=1}^{K} n_k X_k}{n} = \sum_{k=1}^{K} f_k X_k$$

- Previous example: 5, 5, 10, 10, 10, 10, 20

X_k	n_k	f_k	$X_k f_k$
5	2		
10	4		
20	1		
Total	7		

Weighted Mean

The weighted mean of a set of data is

$$\bar{X} = \frac{\sum_{i=1}^{n} w_i X_i}{\sum_{i=1}^{n} w_i}$$

where w_i is the weight of the i^{th} observation.

Why might we want to use a weighted mean?

2016 Election Predictions

The New York Times

TheUpshot

POLITICAL CALCULUS

A 2016 Review: Why Key State Polls Were Wrong About Trump

By Nate Cohn

May 31, 2017

193

Education weighting seems to explain a lot

Education was a huge driver of presidential vote preference in the 2016 election, but many pollsters did not adjust their samples — a process known as weighting — to make sure they had the right number of well-educated or less educated respondents.

It's no small matter, since well-educated voters are much likelier to take surveys than less educated ones. About 45 percent of respondents in a typical national poll of adults will have a bachelor's degree or higher, even though the census says that only 28 percent of adults (those 18 and over) have a degree. Similarly, a bit more than 50 percent of respondents who say they're likely to vote have a degree, compared with 40 percent of voters in newly released 2016 census voting data.

Things to do next

- Review this week's material; handouts, notes, and reading (NYT article) on Canvas
- You may be asked to summarize what you got out of the reading in the next class