Introduction to Statistical Methods SOC-GA 2332

Lecture 1: Introduction

Siwei Cheng

NEW YORK UNIVERSITY

Lecture Outline

- Syllabus and course overview
- Data and variables
- Descriptive statistics
- From description to inference

Syllabus and Course Overview

Open the course syllabus...

Important Announcement

When you see a cute animal picture on the slide, that's my cue to stop for questions.

Data

- Data is a collection of information.
- Data consists of facts and statistics.
- Data can be generated by surveys, experiments, and direct observation.

What does data look like?

Edit Browse Filter Variables Properties Snapshots

var11[16]

	sid	state	crime	murder	pctmetro	pctwhite	pcths	poverty	single
1	1	ak	761	9	41.8	75.2	86.6	9.1	14.3
2	2	al	780	11.6	67.4	73.5	66.9	17.4	11.5
3	3	ar	593	10.2	44.7	82.9	66.3	20	10.7
4	4	az	715	8.6	84.7	88.6	78.7	15.4	12.1
5	5	ca	1078	13.1	96.7	79.3	76.2	18.2	12.5
6	6	co	567	5.8	81.8	92.5	84.4	9.9	12.1
7	7	ct	456	6.3	95.7	89	79.2	8.5	10.1
8	8	de	686	5	82.7	79.4	77.5	10.2	11.4
9	9	fl	1206	8.9	93	83.5	74.4	17.8	10.6
10	10	ga	723	11.4	67.7	70.8	70.9	13.5	13
11	11	hi	261	3.8	74.7	40.9	80.1	8	9.1
12	12	ia	326	2.3	43.8	96.6	80.1	10.3	9
13	13	id	282	2.9	30	96.7	79.7	13.1	9.5
14	14	il	960	11.4	84	81	76.2	13.6	11.5
15	15	in	489	7.5	71.6	90.6	75.6	12.2	10.8
16	16	ks	496	6.4	54.6	90.9	81.3	13.1	9.9

What does data look like?

	caseid	year	sex	familyid1968	sampletype	age_reported	racehd	reltohd	region	longiwt	edu_hd	edu_wf	train_outside_hd	train_outside_wf	hrlywage_hd	hrlywage_wf
1	1001	1968	1	1	1	52	1	1	2	23.000	2	2	0	NA	1.43	1.95
2	1001	1969	1	1	1	54	1	1	2	24.900	2	NA	0	NA	0.00	2.03
3	1002	1968	2	1	1	46	1	2	2	23.000	2	2	0	NA	1.43	1.95
4	1002	1969	2	1	1	48	1	2	2	24.900	2	NA	0	NA	0.00	2.03
5	1002	1970	2	1	1	48	1	1	2	24.900	2	NA	5	NA	2.37	0.00
6	1002	1971	2	1	1	49	1	1	2	24.900	2	NA	5	NA	1.80	0.00
7	1002	1972	2	1	1	50	1	1	2	24.900	2	0	5	NA	2.30	0.00
8	1002	1973	2	1	1	51	1	1	2	24.900	2	0	5	NA	2.59	0.00
9	1002	1974	2	1	1	52	1	1	2	27.400	2	0	5	NA	2.57	0.00
10	1002	1975	2	1	1	53	1	1	2	27.400	2	0	0	0	3.08	0.00
11	1002	1976	2	1	1	55	1	1	2	27.400	2	0	0	0	3.05	0.00
12	1002	1977	2	1	1	55	1	1	2	27.400	2	0	0	0	4.13	0.00
13	1003	1972	1	1	1	25	1	1	2	24.900	4	4	5	NA	2.11	3.42
14	1003	1973	1	1	1	25	1	1	2	24.900	4	4	5	NA	2.40	5.00
15	1003	1974	1	1	1	27	1	1	2	30.400	4	4	5	NA	3.75	4.99
16	1003	1975	1	1	1	28	1	1	2	30.400	4	4	0	5	4.01	5.10
17	1003	1976	1	1	1	30	1	1	2	30.400	4	4	0	5	4.40	5.61
18	1003	1977	1	1	1	30	1	1	2	30.400	4	4	0	5	4.79	4.08
19	1003	1978	1	1	1	31	1	1	2	30.400	4	4	0	5	4.53	5.68
20	1003	1979	1	1	1	32	1	1	2	32.500	4	4	0	5	8.20	5.92
21	1004	1974	2	1	1	25	1	2	2	30.400	4	4	5	NA	3.75	4.99
22	1004	1975	2	1	1	26	1	2	2	30.400	4	4	0	5	4.01	5.10

Data

- This course focuses mainly on how to describe and analyze data.
- But methods for analyzing data will also inform us of how the data should be *collected*.
- Statistics consists of a body of methods that help social scientists describe characteristics of a sample, make inferences about the population, or test hypotheses.

Variable

- A variable is a characteristic of a statistical unit being observed.
- A variable can assume different values. The value of a variable usually varies in a population or a sample (otherwise this will be a constant).
- The values can be categorical or numeric.

Types of Variable

$$\begin{cases} \textit{Categorical} \begin{cases} \textit{Nominal} \text{ (e.g. marital status)} \\ \textit{Ordinal} \text{ (e.g. educational level)} \end{cases} \\ \\ \textit{Numeric} \begin{cases} \textit{Discrete} \text{ (e.g. number of child(ren))} \\ \textit{Continuous} \text{ (e.g. income, neighborhood poverty rate)} \end{cases} \end{cases}$$

Types of Variable in Real-world Research - 1

- In empirical studies, the type of variable depends on how this variable is measured in your data.
- e.g. Neighborhood socioeconomic status can be measured as a categorical variable (poor and non-poor) or a numeric variable (median household income).

Types of Variable in Real-world Research - 2

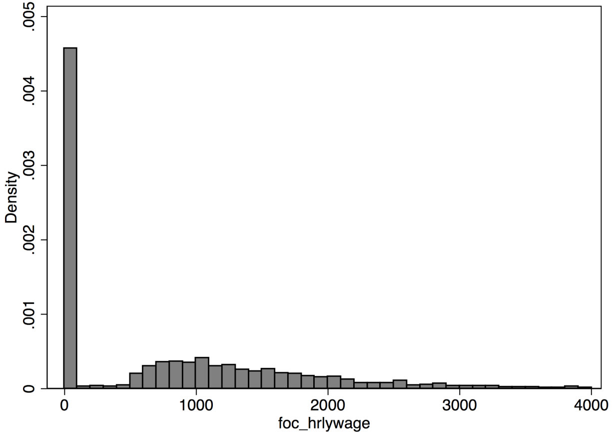
- In statistical analysis, sometimes categorical variables can be coded as numeric values. But we should always be careful about how these variables are coded.
- e.g. Adding up K-6 items to construct an index of depression.

The following questions ask about how you have been feeling during the **past 30 days**. For each question, please circle the number that best describes how often you had this feeling.

Q1. During the past 30 days, about how often did you feel …	**All of the time**	**Most of the time**	**Some of the time**	**A little of the time**	**None of the time**
a. …nervous?	**1**	**2**	**3**	**4**	**5**
b. …hopeless?	**1**	**2**	**3**	**4**	**5**
c. …restless or fidgety?	**1**	**2**	**3**	**4**	**5**
d. …so depressed that nothing could cheer you up?	**1**	**2**	**3**	**4**	**5**
e. …that everything was an effort?	**1**	**2**	**3**	**4**	**5**
f. …worthless?	**1**	**2**	**3**	**4**	**5**

Types of Variable in Real-world Research - 3

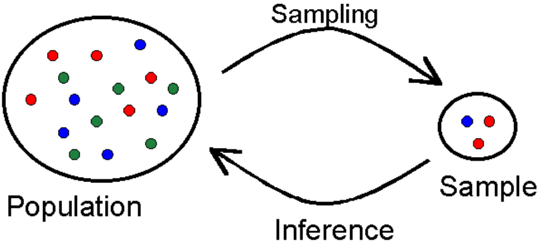
- There can be more complicated cases in real world research:
- e.g. Hourly wage. Is this a numeric or categorical variable? What does a zero wage mean?



Sample and Population

Sample and Population

- Population: total set of subjects of interest in a study.
- Sample: subset of the population on which we collect data.
- Goal of quantitative data analysis: learn about the population using the sample data.



Descriptive Statistics and Inferential Statistics

- Descriptive statistics summarize the information in the sample.
- Descriptive statistics help us: (a) reduce the data to simpler and more understandable forms; and (b) describe the sample distribution in the entire sample or in different subgroups of the sample.

- Inferential statistics provide predictions about a population, based on data from a sample of that population.
- *But, does a population have to actually exist?*

Descriptive Statistics and Inferential Statistics

- A parameter is a numerical summary of the population.
- A statistic is a numerical summary of the sampled data.
- We can think of a statistic is a random variable. Why?

Descriptive Statistics and Inferential Statistics

- A parameter is a numerical summary of the population.
- A statistic is a numerical summary of the sampled data.
- We can think of a statistic is a random variable. Why?
- Any function of a random variable is itself a random variable.
- A statistic could take on different values, depending on the different samples we could collect. (Once we collect a single sample, we can calculate a specific value of the statistic.)

Descriptive Statistics and Inferential Statistics

- A parameter is a numerical summary of the population.
- A statistic is a numerical summary of the sampled data.
- We can think of a statistic is a random variable. Why?
- Any function of a random variable is itself a random variable.
- A statistic could take on different values, depending on the different samples we could collect. (Once we collect a single sample, we can calculate a specific value of the statistic.)
- We make statistical inferences about the population based on properties of statistics from the sample.


@EricScottBJWT

Descriptive Statistics

- As a first step of quantitative data analysis, we describe our data with tables and graphs.
- We will begin with univariate statistics, that is, statistics on a single variable (numerical descriptions of center, variability/dispersion, position).
- Then, we will discuss bivariate statistics, that is, statistics describing relationships between two variables.

Frequency Distribution

- A listing of possible values for a variable, together with the number of observations or relative frequency at each value.
- Example: frequency distribution of the number of awards earned by students at a high school in a year:

```
. tab num_awards
```

num_awards	Freq.	Percent	Cum.
0	124	62.00	62.00
1	49	24.50	86.50
2	13	6.50	93.00
3	9	4.50	97.50
4	2	1.00	98.50
5	2	1.00	99.50
6	1	0.50	100.00
Total	200	100.00	

Frequency Distribution

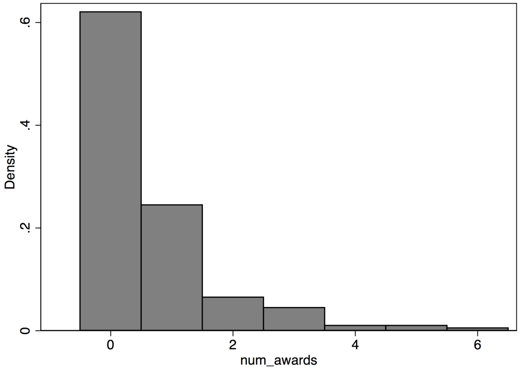
- Frequency distribution for marital status in Current Population Survey (1979-2914)

```
. tab marital if marital>=0
```

Marital status	Freq.	Percent	Cum.
Married, Civilian Spouse Present	6,502,406	55.88	55.88
Married, Armed Forces Spouse Present	45,270	0.39	56.27
Married, Spouse Absent (exc. Separated)	190,119	1.63	57.90
Widowed	1,039,332	8.93	66.83
Divorced	779,700	6.70	73.53
Separated	170,732	1.47	75.00
Never Married	2,909,042	25.00	100.00
Total	11,636,601	100.00	

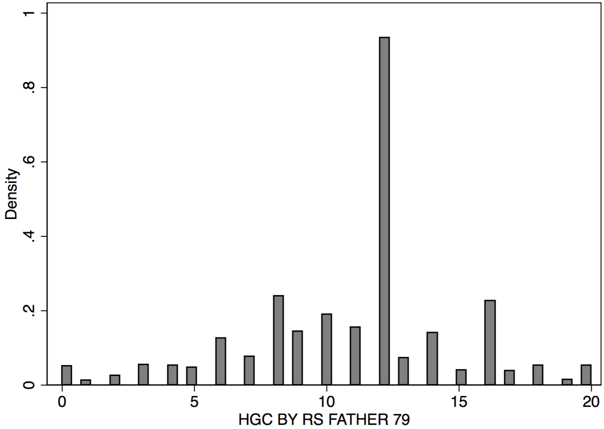
Bar Graph

- showing relative frequency in each category:



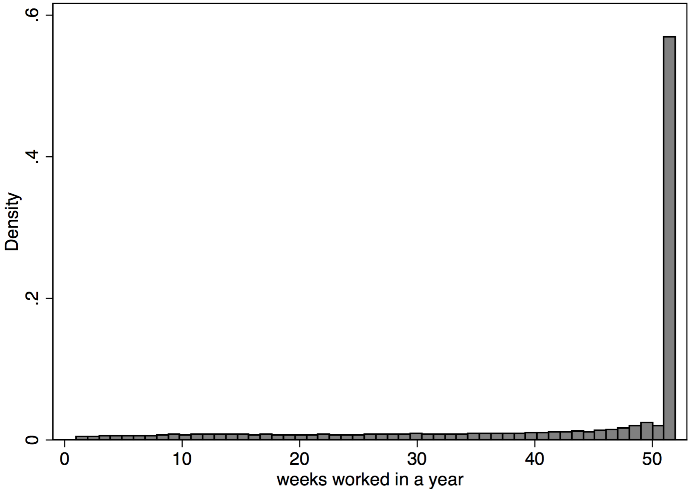
Histogram for **numeric** variables -Example 1

- Highest grade completed of the respondent's father (NLSY79 data)



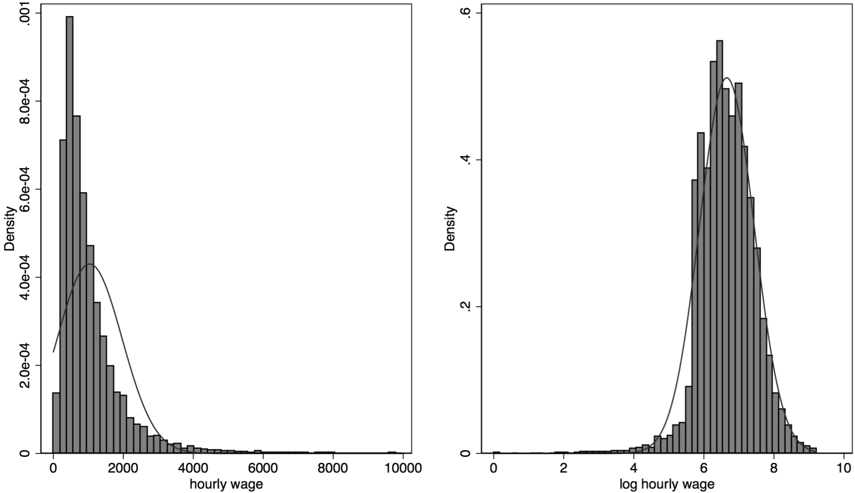
Histogram for **numeric** variables -Example 2

- Number of weeks worked in a year (NLSY79 data)



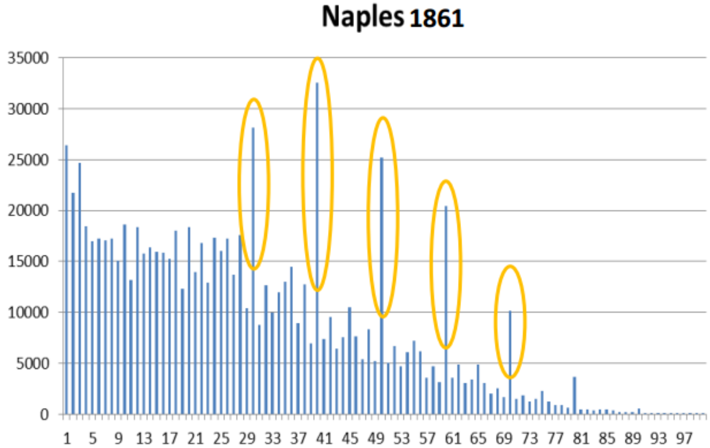
Histogram for **numeric** variables -Example 3

- Histogram of *hourly wage* and *log hourly wage*. (NLSY79 data)



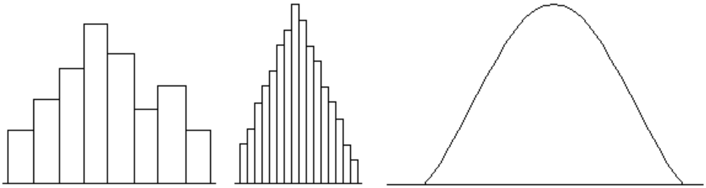
Histogram for numeric variables -Example 4

- Age heaping/digit preference



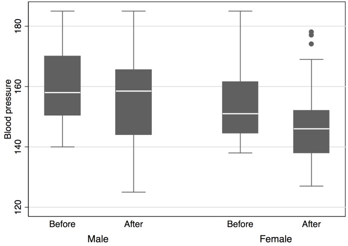
Sample Size and Histogram

- Histogram gradually approaches a smooth curve as sample size gets larger.



Boxplot

- Showing the median (line in the middle), 75th (top of box) and 25th (bottom of box) percentile. The box shows 50% of the observations.
- "outliers" are individual observations that are over 1.5*IQR(inter quartile range) from the upper/lower quartile.
- Example, blood pressure by time and gender



Describing the data with sample statistics

- Central tendency
- Positions of distributions
- Variability

Measures of Central Tendency

- The most important descriptive statistics for the center of a sample is the mean.
- Let y denote a quantitative variable, with n observations $y_1, y_2, y_3, ...y_n$.
- Sample mean (a statistic):

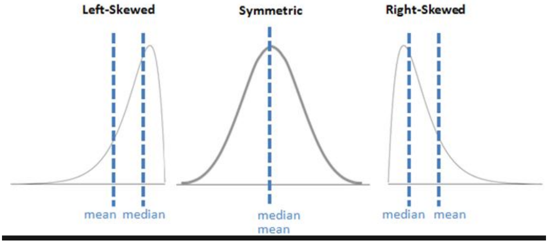
$$\bar{y} = \frac{y_1 + y_2 + y_3 + ... + y_n}{n} = \frac{\sum_{i=1}^{n} y_i}{n}$$

- We use sample mean as an estimator of population mean μ (a parameter):

$$\hat{\mu} = \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

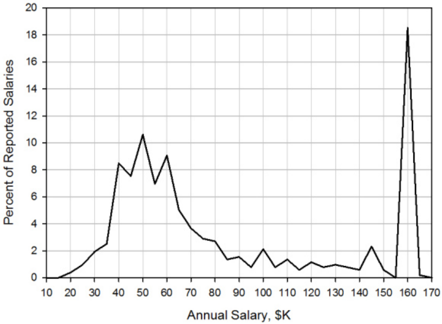
Measures of Central Tendency

- Median: middle measurement of ordered sample.
- For symmetric distributions, such as standard normal distribution, the median and mean are identical.
- For skewed distributions, mean and median can be very different:
- Mean is sensitive to outliers, so median is preferred for highly skewed distributions.



Measures of Central Tendency

- Mode measures the most common response
- Some data has two distinct mounds (bimodal distribution):
- Example: Distribution of Reported Full-Time Salaries among Lawyers, 2010 (Source: NALP)



Describing positions of distributions:

- pth percentile: p percent of observaionts below it, (100-p) percent above it.
- p=50: median;
- p=25: lower quartile (LQ)
- p=75: upper quartile (UQ)
- Interquartile range:

$$IQR = UQ - LQ$$

Describing Variability

- Sample variance (a statistic):

$$s^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$$

- Population variance (a parameter):

$$\sigma^2 = \frac{\sum_{i=1}^{N}(y_i - \bar{y})^2}{N}$$

- Sample standard deviation (a statistic):

$$s = \sqrt{s^2}$$

- Population standard deviation (a parameter):

$$\sigma = \sqrt{\sigma^2}$$

Z-score: Describing Deviation from the Mean

- We describe the deviation from the mean of an observation y_i using its z-score:

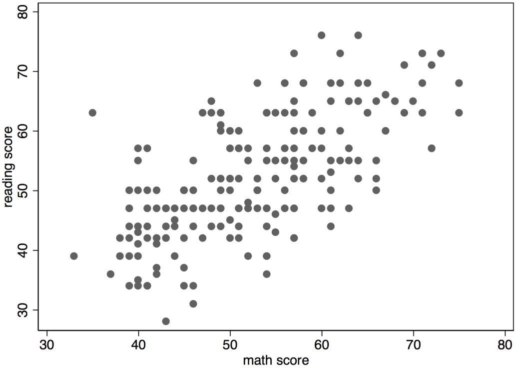
$$z = \frac{y_i - \bar{y}}{s}$$

- Intuitively, the z-score is the deviation of an individual observaiton divided by the "average" deviation (the standard deviation).

- Question: when would z-score be useful?

Describing Relationship between Two Variables

- Scatterplot can be used to visualize the distribution of two variables.



Describing Relationship between Two Variables

Describing Relationship between Two Variables

- For categorical variables, we can use contingency tables:

```
. tab race female, col
```

Key
frequency *column percentage*

race	female: male	female: female	Total
hispanic	13 14.29	11 10.09	24 12.00
asian	3 3.30	8 7.34	11 5.50
african-amer	7 7.69	13 11.93	20 10.00
white	66 72.53	77 70.64	143 71.50
5	2 2.20	0 0.00	2 1.00
Total	91 100.00	109 100.00	200 100.00

Describing Relationship between Two Variables

- For categorical variables, we can use contingency tables:

```
> table(dataorig$edu4cate_lab, dataorig$sex_lab)

               Female  Male
  BA+           38899 37358
  HS grad       46424 39217
  Less then HS  23517 21357
  Some college  35809 30985
> 
```

Describing Relationship between Two Variables

- We use covariance and correlation to describe the relationship between two variables.
- Sample covariance of X and Y is given by:

$$Cov(x, y) = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

- Covariance tells us whether two variables are related to each other. Covariance can be positive, negative, or zero.
- Note that the covariance of a variable with itself is the variance.

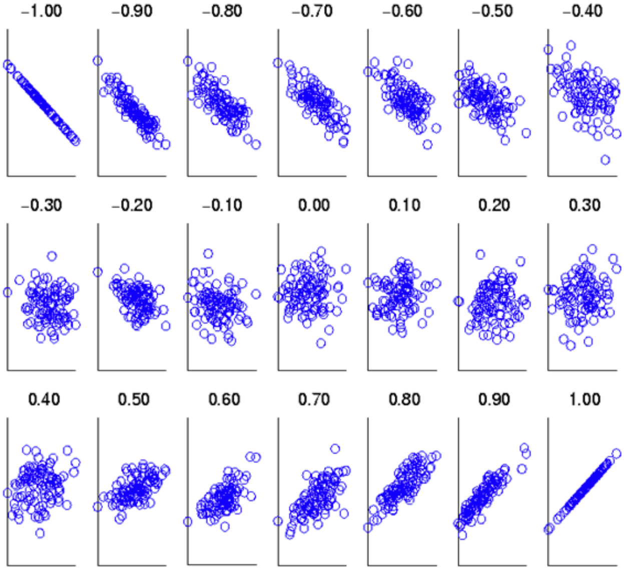
Describing Relationship between Two Variables

- The correlation is a *rescaled* version of the covariance, where the covariance is divided by the standard deviation of each variable.

$$\rho = Corr(x, y) = \frac{Cov(x, y)}{s_x s_y}$$

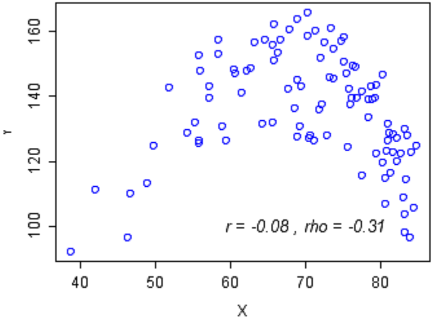
- Correlation ranges between -1 and 1. This allows for a better sense of how strong the relationship between two variable is.

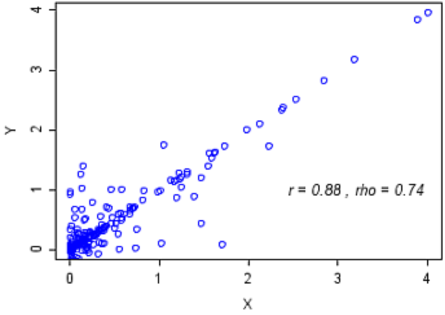
Correlation - An Illustration



Correlation - A Drawback

Correlation only captures linear association...





A Quick Review

- We use statistical methods to describe or analyze quantitative data.
- **Descriptive statistics** describes distributions and generate statistics in the sample.
- **Inferential statistics** will enable us to learn about parameters in the population based on sample statistics.

Key Concepts So Far...

- Sample mean (a statistic):

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

- Population mean (a parameter):

$$\mu = \frac{\sum_{i=1}^{N} y_i}{N}$$

- Sample variance (a statistic):

$$s^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1}$$

- Population variance (a parameter):

$$\sigma^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{N}$$

Some More Thoughts on Samples

- In the above discussions, we are assuming that the sample is a **probabilistic sample** and the sample units drawn via **random selection** from a population.
- But that's rarely the case in real social science research!
- **Examples**:
  - response bias/non-response bias;
  - attrition bias in panel data;
  - convenience sample;
  - unobserved outcomes (e.g. potential wage levels among current inmates).

Moving from descriptive to inferential statistics...

Sampling Distribution

- A sampling distribution describes the distribution of a statistic, such as a sample mean or variance, of that statistic was measured over a range of different samples within a population.
- (Recall that a statistic is a random variable.)
- For example, consider the sampling distribution of *the sample mean of earnings*...
- We can collect **many repeated samples** from the population, calculate the mean of earnings in each sample, and then describe the probability distribution of the sample mean of earnings over the range of these different samples.

Sampling Distribution

- Let's do an experiment!

Sampling Distribution

- Let's do an experiment!
- Please randomly write down an integer from 1 to 10.

Sampling Distribution

- Let's do an experiment!
- Please randomly write down an integer from 1 to 10.
- Suppose I am interested in the sampling distribution of the population mean of this number μ, with the population defined as everyone in this class.

Sampling Distribution

- Let's do an experiment!
- Please randomly write down an integer from 1 to 10.
- Suppose I am interested in the sampling distribution of the population mean of this number μ, with the population defined as everyone in this class.
- Now, I am going to collect our sample $\#1$...

Sampling Distribution

- Let's do an experiment!
- Please randomly write down an integer from 1 to 10.
- Suppose I am interested in the sampling distribution of the population mean of this number μ, with the population defined as everyone in this class.
- Now, I am going to collect our sample $\#1$...
- And sample $\#2$...

Sampling Distribution

- Let's do an experiment!
- Please randomly write down an integer from 1 to 10.
- Suppose I am interested in the sampling distribution of the population mean of this number μ, with the population defined as everyone in this class.
- Now, I am going to collect our sample $\#1$...
- And sample $\#2$...
- And sample $\#3$...

Sampling Distribution

- Let's do an experiment!

- Please randomly write down an integer from 1 to 10.

- Suppose I am interested in the sampling distribution of the population mean of this number μ, with the population defined as everyone in this class.

- Now, I am going to collect our sample #1...
- And sample #2...
- And sample #3...
- I can calculate the mean of every sample, call it $\bar{X}_k$, where k is the index for all the samples.

Sampling Distribution

- Let's do an experiment!
- Please randomly write down an integer from 1 to 10.
- Suppose I am interested in the sampling distribution of the population mean of this number μ, with the population defined as everyone in this class.
- Now, I am going to collect our sample #1...
- And sample #2...
- And sample #3...
- I can calculate the mean of every sample, call it $\bar{X}_k$, where k is the index for all the samples.
- Imagine that I have taken an infinite number of such samples.
- The sampling distribution of the populaiton mean is simply the distribution of the means of all my imaginary samples.

Sampling Distribution

- In other words, we can think of the mean of a given sample, $\bar{y}$, as a variable with a value that varies from sample to sample around the population mean μ.
- So the sampling distribution of $\bar{y}$ has mean μ.
- But how about the standard deviation of the sampling distribution? We use a new concept called the standard error.

Sampling Distribution and Standard Error

- The standard deviation of a sampling distribution is called the standard error.
- For example, the sampling distribution of $\bar{y}$ in a sample of n observations has standard error:

$$\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}}$$

- But since we typically don't know the population standard deviation σ, we estimate it with the sample standard deviation s:

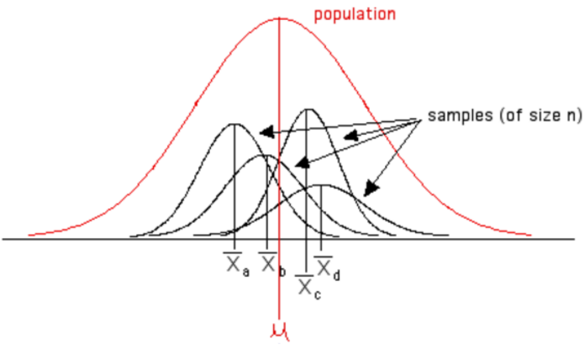
$$\hat{\sigma}_{\bar{y}} = \frac{s}{\sqrt{n}}$$

- Over large samples, sampling distribution is approximately normal (even though sample or population distribution may not be).

Comparing Three Types of Distributions

- Population distribution: described by parameters (usually unknown) such as mean (μ) and standard deviation (σ).
- Sample distribution: described by sample statistics such as sample mean ($\bar{y}$) and sample standard deviation (s).
- Sampling distribution: probability distribution of a sample statistic, such as sample mean. The sampling distribution of a sample mean equals population mean (μ), and the standard deviation of the sampling distribution (call STANDARD ERROR) is ($\hat{\sigma}_{\bar{y}} = \frac{s}{\sqrt{n}}$).

Comparing Three Types of Distributions



Sampling Distribution

- Sampling distribution is important, because it determines the **probability** that a statistic falls within certain distance of population parameter.
- That will lead us to the next concept (to be continued in next lecture): Confidence Interval.