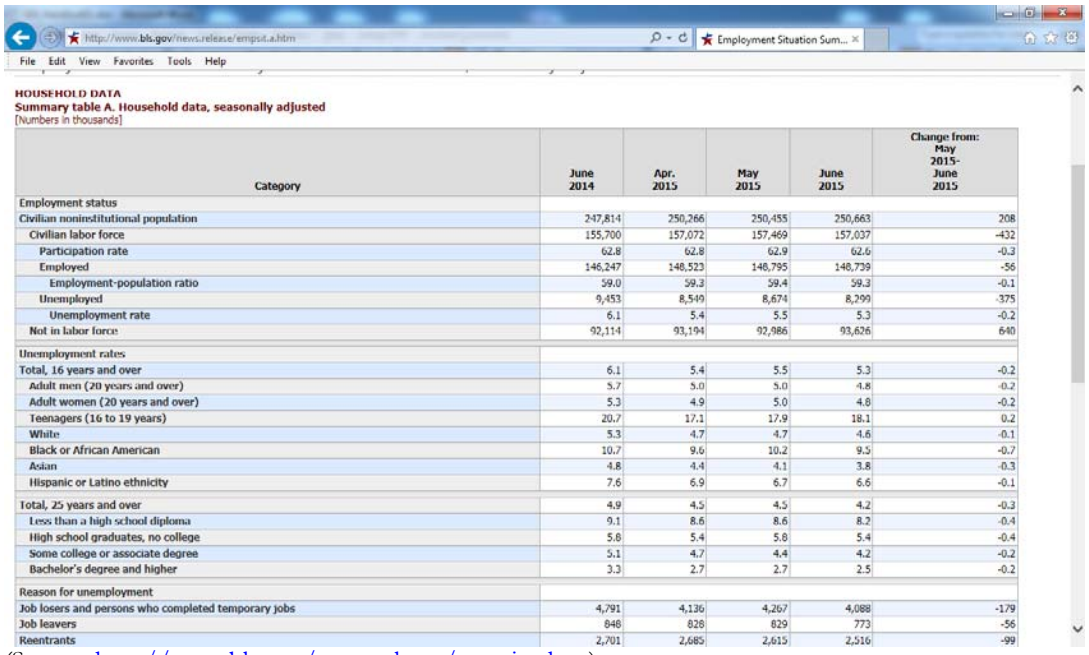
Handout #1

Title: Foundations of Econometrics — Fall/2015
Course: Econ 367 — Instructor: Dr. I-Ming Chiu

POPULATION vs. SAMPLE

From the Bureau of Labor web site (http://www.bls.gov), we can find the unemployment rate for each month in the U.S. For example, the unemployment rate is 5.3% in June 2015. How does the Bureau of Labor measure the unemployment rate?

Table 1 Data from the Bureau of Labor



HOUSEHOLD DATA
Summary table A. Household data, seasonally adjusted
[Numbers in thousands]

Category	June 2014	Apr. 2015	May 2015	June 2015	Change from: May 2015-June 2015
Employment status					
Civilian noninstitutional population	247,814	250,266	250,455	250,663	208
Civilian labor force	155,700	157,072	157,469	157,037	-432
Participation rate	62.8	62.8	62.9	62.6	-0.3
Employed	146,247	148,523	148,795	148,739	-56
Employment-population ratio	59.0	59.3	59.4	59.3	-0.1
Unemployed	9,453	8,549	8,674	8,299	-375
Unemployment rate	6.1	5.4	5.5	5.3	-0.2
Not in labor force	92,114	93,194	92,986	93,626	640
Unemployment rates					
Total, 16 years and over	6.1	5.4	5.5	5.3	-0.2
Adult men (20 years and over)	5.7	5.0	5.0	4.8	-0.2
Adult women (20 years and over)	5.3	4.9	5.0	4.8	-0.2
Teenagers (16 to 19 years)	20.7	17.1	17.9	18.1	0.2
White	5.3	4.7	4.7	4.6	-0.1
Black or African American	10.7	9.6	10.2	9.5	-0.7
Asian	4.8	4.4	4.1	3.8	-0.3
Hispanic or Latino ethnicity	7.6	6.9	6.7	6.6	-0.1
Total, 25 years and over	4.9	4.5	4.5	4.2	-0.3
Less than a high school diploma	9.1	8.6	8.6	8.2	-0.4
High school graduates, no college	5.8	5.4	5.8	5.4	-0.4
Some college or associate degree	5.1	4.7	4.4	4.2	-0.2
Bachelor's degree and higher	3.3	2.7	2.7	2.5	-0.2
Reason for unemployment					
Job losers and persons who completed temporary jobs	4,791	4,136	4,267	4,088	-179
Job leavers	848	828	829	773	-56
Reentrants	2,701	2,685	2,615	2,516	-99

(Source: http://www.bls.gov/news.release/empsit.a.htm)

We would like to choose a "representative" (i.e., avoid sampling bias) sample to make "inference" about the population. How do we achieve that? Answer: random sampling method should be applied.
Statistical inference: make a conclusion about the population using the collected sample data.

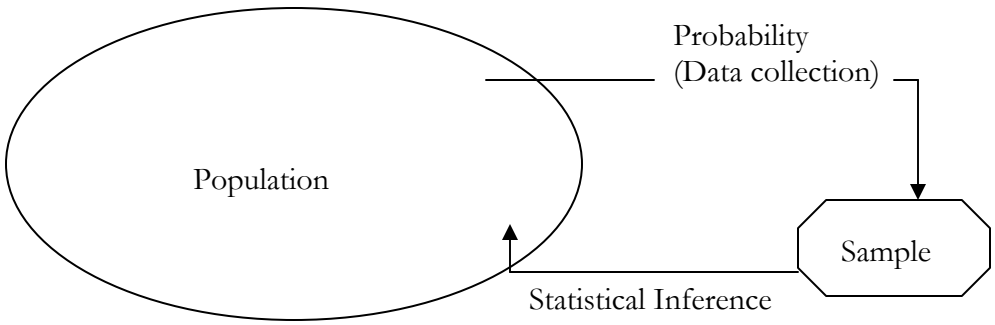


Table 2 Parameters and the corresponding Statistics

	Population parameter	Sample statistic
Mean	u	$\bar{x}$
Median	$\tilde{u}$	$\tilde{x}$
Variance	σ^2	s^2
Standard Deviation	σ	s
Proportion	p	$\hat{p}$
Correlation	ρ	r
Slope (regression)	β	b

*The square of the standard deviation is called variance (σ^2 vs. s^2)

THE STRUCTURE OF DATA

Statistics: the science of collecting, describing, and analyzing data.

I. Cases & Variables

Case: the subjects/objects that we obtain information about.
Variable: a variable is any character that is recorded for each case (unit).

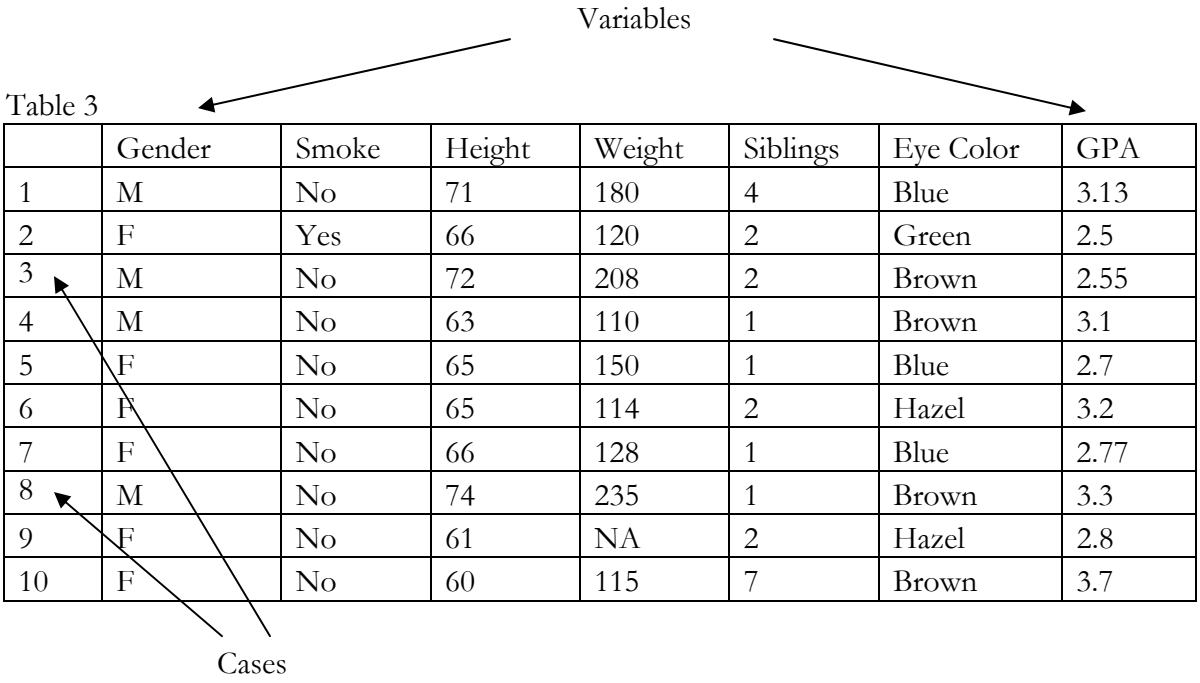


Table 3

	Gender	Smoke	Height	Weight	Siblings	Eye Color	GPA
1	M	No	71	180	4	Blue	3.13
2	F	Yes	66	120	2	Green	2.5
3	M	No	72	208	2	Brown	2.55
4	M	No	63	110	1	Brown	3.1
5	F	No	65	150	1	Blue	2.7
6	F	No	65	114	2	Hazel	3.2
7	F	No	66	128	1	Blue	2.77
8	M	No	74	235	1	Brown	3.3
9	F	No	61	NA	2	Hazel	2.8
10	F	No	60	115	7	Brown	3.7

*Table 3 represents a small data set that is retrieved from "StudentSuvey.csv" file. Noticed that cases are in rows and variables are in columns.

II. Data Classification

A. Types of Variables

a) Quantitative variables
e.g. Height, weight, stock prices, trading volumes, etc.

b) Categorical variables (nominal vs. ordinal)
e.g. Sex (male, female), working status (employed, unemployed), political affiliation (Dem, Rep, Ind) $\in$ nominal.
e.g. cup size at coffee shops (small, medium, large), grade (F, D, C, B, A) $\in$ ordinal.

When there are only two levels in a categorical variable, there is no need to differentiate whether it is nominal or ordinal.

B. Types of Variables

In the field econometrics the data can be categorizes as

a) Cross-section
e.g. Students' 1st exam scores of class FE 367 in fall 2015.

b) Time series
e.g. Daily Dow Jones Industrial Average Index.

c) Longitudinal/Panel
e.g. Average household income in 50 states between 1999 and 2014.

C. Types of Variables

a) Observational
e.g. Temperatures in NJ, crime rates in Camden, etc.

b) Experimental
e.g. Clinical Trials

How do we utilize data?
a) Study a single variable
b) Study the relationship between variables; i.e., response vs. explanatory variable.

Exercise: What do you want to know about a single variable such as GPA *in Table 3?*
Do those students who don't smoke have a higher GPA *than those who do?*
Are students' heights (Height) *and weights* (Weight) *related?*

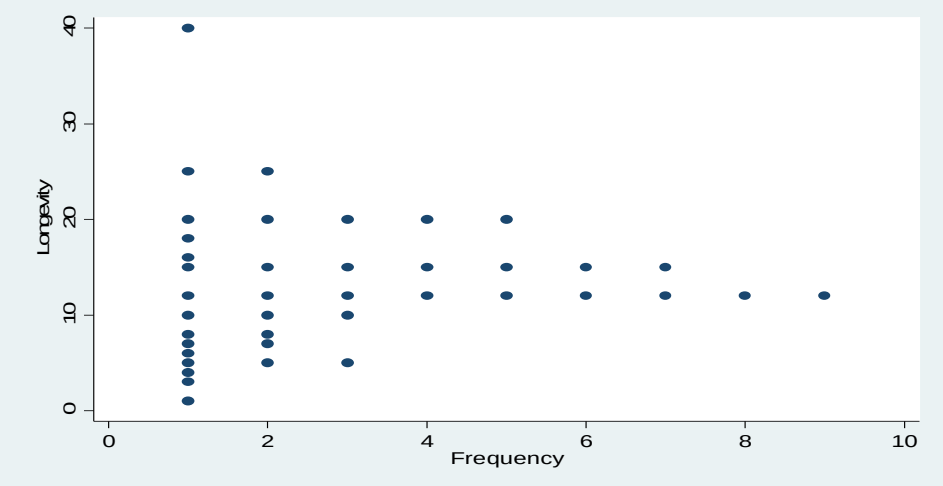
GRAPHICAL PRESENTATION OF DATA

A. Stem-and-Leaf

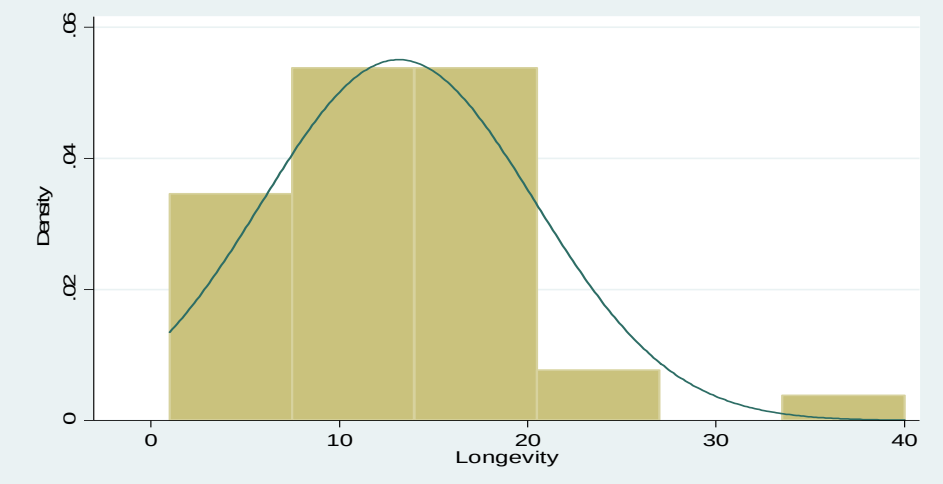
Stem-and-leaf plot for longevity (Longevity)

```
0*  | 134
0.  | 55567788
1*  | 000222222222
1.  | 555555568
2*  | 00000
2.  | 55
3*  |
3.  |
4*  | 0
```

B. Dotplot
Longevity Data



C. Histogram



MEASURES OF LOCATION AND VARIABILITY

I. Location

$$\text{Mean } (\bar{x}) = \frac{\sum_{i=1}^{n} x_i}{n}$$

Notice: the symbol of mean is μ for population and $\bar{x}$ for sample.

Median ($\tilde{x}$): the middle value
e.g. 1, 2, 5, 7, 9 $\Rightarrow$ the median is 5 ($\tilde{x} = 5$)
e.g. 1, 5, 7, 9 $\Rightarrow$ the median is $\frac{5+7}{2} = 6$ ($\tilde{x} = 6$)

Outliers: extreme values
Resistance: If a "statistic" is not affected by outliers, it's resistant.

Q: Is mean or median more resistant?
Q: For each of the following variables:
a) Find the mean
b) Find the median
c) Identify any outliers

8, 12, 3, 18, 15
41, 53, 38, 12, 115, 47, 50
15, 22, 12, 28, 58, 18, 25, 18
110, 112, 118, 119, 122, 125, 129, 135, 138, 140

II. Variability

$$\text{Variance } (s^2) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} \quad \nearrow S_{xx}$$

$$S_{xx} = \sum_{i=1}^{n} (x_i - x)^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$

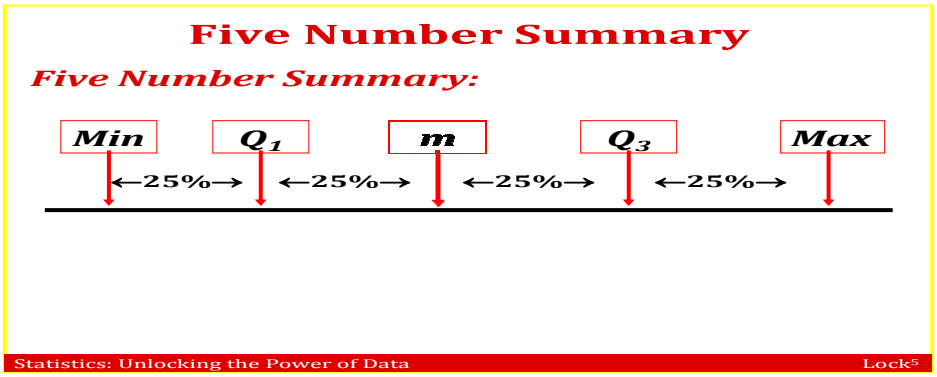
$$\text{Standard deviation } (s) = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$$

e.g. Using data 8, 12, 3, 18, 15 to find variance.

Notice: Notice: the symbol of standard deviation is σ for population and s for sample.

Other measures

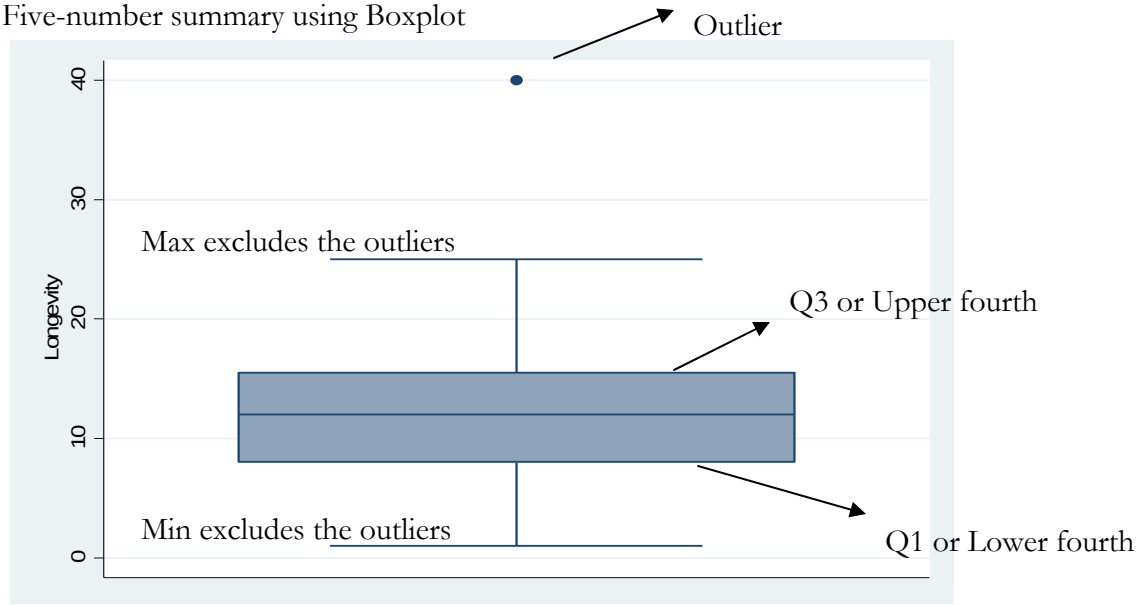
Percentile: the P^{th} percentile is the value of a quantitative variable which is greater than P percent of the data.



Min: minimum, Max: maximum
Q_1: 1^{st} quartile (25^{th} percentile or lower fourth)
$\tilde{x}$: median (50^{th} percentile)
Q_3: 3^{rd} quartile (75^{th} percentile or upper fourth)
Range = Max – Min
Interquartile range (IQR or fourth spread) = Q3 – Q1

Definition of outliers: if a sample value is smaller than Q_1 – 1.5*IQR or greater than Q_3 + 1.5*IQR.

Five-number summary using Boxplot



Q: use the "MammalLongevity.csv" data file to find these five number summary as well as range and IQR.
(We will do this in handout #2 once you know how to use the commands in Stata)