LECTURE 1

POPULATIONS, SAMPLES, AND STATISTICAL INFERENCE

(Appendix C, sections 1.A and 2.A)

DEFINITION: The **population** consists of all units (everything/everyone) we want to measure.

E.g. You want to learn what is the *average height* of *all UM undergraduate students* (28,283 students according to Wikipedia).

Population: all UM undergraduate students

Populations have parameters.

DEFINITION: Population parameter – a measure of a characteristic (attribute) of a population

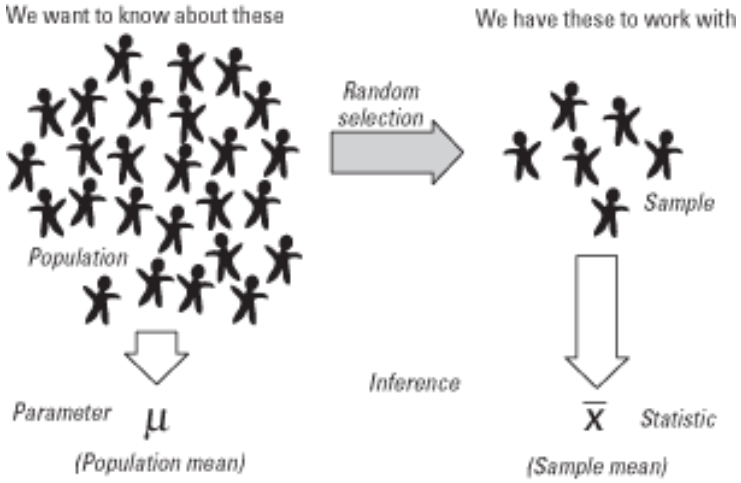
Example of a population parameter

Population mean (average) $\boldsymbol{\mu}$: intuitively – the average of a certain characteristic of a population (sum divided by the population size).

E.g. population average height of all UM undergraduate students (μ)

Population parameters often cannot be calculated due to the large size of the population. Often we want to know things about populations but do not have data for every person or unit in the population. If we wanted to know something about all UM students, it would not be practical (or too costly) to contact every students. Instead, we might select a **sample** of the population.

Figure 1: Illustration of the relationship between samples and populations.



DEFINITION: A **sample** is a smaller group of objects from our population, which is *randomly selected*. A random sample is one in which every unit from the population has an equal chance of being selected. For examples, we can talk about "500 women randomly selected amongst all women in the Ann Arbor area between the ages of 18 and 54", "100 male students randomly drawn from the Economics Department", "30 students randomly drawn from the ECON 251 class", etc.

E.g.: random sample of 500 students

Question: non-random sample – basketball/volleyball team; only men; only women

Samples have statistics.

DEFINITION: Sample statistic – a measure of a characteristic (attribute) of a sample

The difference between a statistic and a parameter is that statistics describe a **sample**, whereas a parameter describes an entire **population.**

Notice that different symbols are used to denote statistics and parameters.

Example of a sample statistic

Sample mean $\bar{\mathbf{X}}$ (also called the sample average), $\bar{\mathbf{X}}$: the sum of a certain characteristic of a sample divided by the sample size N.

A: sample mean height $\bar{\mathbf{X}} = (X_1 + X_2 + ... + X_{500}) / 500$ = **5ft 6 in / 170cm**

KEY: Based on a ***statistic*** computed from a *sample randomly* drawn from the ***population*** we learn about the ***population parameter***. *This is called* ***statistical inference.***

Two more terms:

The sample mean is an example of a so-called **estimator.**

DEFINITION: an estimator (of a population parameter) is a rule which tell us how we are going to calculate a best guess (**estimate)** given a random sample.

DEFINITION: An **estimate** is just a numerical result.

Estimator: rule $\bar{\mathbf{X}} = (X_1 + X_2 + ... + X_N) / N$

We can have another rule: $(X_1+X_2)/2$, $(X_1+X_N)/2$, 6 feet/183 cm, [Min(Xi)+Max(Xi)]/2, etc.

Estimate: 5ft 6 in / 170cm (number)

Another example to test your understanding

You want to learn about the average weight of *all individuals in the US*. You randomly select *5,000 persons* and determine that their average weight is 165 lb/75 kg.

Population?

A: all individuals in the US

Population parameter?

A: the average (of variable weight)

Sample?

A: all 5000 persons, randomly selected from the population (so, N=5000)

Sample statistic?

A: the sample mean (of variable weight)

$\bar{\mathbf{X}} = (X_1 + X_2 + ... + X_N) / N$

Estimate?

A: the numeric result we obtained, 165 lb/75 kg

Statistical inference?

Based on this sample mean, we can conclude that the populations mean weight μ is likely to be close to 165 lb/75 kg.

THE SUMMATION OPERATOR

APPENDIX A, SECTION A.1

a) Definition: $\sum_{i=1}^{N} X_i = X_1+X_2+...+X_N$

b) Properties

b1) if C=*contsant* ⇒ $\sum_{i=1}^{N} C = NC$

b2) if C=*contsant* $\Rightarrow \sum_{i=1}^{N} CX_i = \mathrm{C}\sum_{i=1}^{N} X_i$

b3) $\sum_{i=1}^{N}(X_i + Y_i) = \sum_{i=1}^{N} X_i + \sum_{i=1}^{N} Y_i$

b4) $\sum_{i=1}^{n} \frac{X_i}{Y_i} \neq \frac{\sum_{i=1}^{n} X_i}{\sum_{i=1}^{n} Y_i}$

Illustration

Example 1

Given data on a variable X, the sum of the deviations from the sample average (or sample mean) is always zero: $\sum_{i=1}^{n}(X_i - \bar{X}) = 0$

Proof:

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

$$\sum_{i=1}^{N} X_i - \sum_{i=1}^{N} \bar{X} = \ (regrouping\ the\ summation\ terms)$$

$$\sum_{i=1}^{N} X_i - N\bar{X} = \ (using\ the\ fact\ \bar{X}\ is\ a\ constant)$$

$$\sum_{i=1}^{N} X_i - N\frac{\sum_{i=1}^{N} X_i}{N} = \ (using\ the\ definition\ of\ \bar{X})$$

$$\sum_{i=1}^{N} X_i - \sum_{i=1}^{N} X_i = (cancelling\ out\ N\ with\ \frac{1}{N})$$

= 0.

Example 2 (sections)

Given data on two variables X and Y, and the sample means $\bar{X}$ and $\bar{Y}$ the following holds:

$$\sum_{i=1}^{N}(X_i - \bar{X})\,(Y_i - \bar{Y}) = \sum_{i=1}^{N} X_i Y_i \; - \; N\,\bar{X}\bar{Y}$$

Proof:

$$\sum_{i=1}^{N}(X_i - \bar{X})\,(Y_i - \bar{Y}) =$$

$$\sum_{i=1}^{N}(X_i Y_i - X_i\bar{Y} - Y_i\bar{X} + \bar{X}\bar{Y}\,) = (multiplying\ the\ terms)$$

$$\sum_{i=1}^{N} X_i Y_i - \sum_{i=1}^{N} X_i\bar{Y} - \sum_{i=1}^{N} Y_i\bar{X} + \sum_{i=1}^{N} \bar{X}\bar{Y} = (regrouping\ the\ summation\ terms)$$

$$\sum_{i=1}^{N} X_i Y_i - \bar{Y}\sum_{i=1}^{N} X_i - \bar{X}\sum_{i=1}^{N} Y_i + \; N\bar{X}\bar{Y} = (using\ the\ fact\ that\ \bar{X}\ \ and\ \bar{Y}\ are\ constants)$$

$$\sum_{i=1}^{N} X_i Y_i - \bar{Y}(N\bar{X}) - \bar{X}(N\bar{Y}) + \; N\bar{X}\bar{Y} =$$

$$= \; (using\ the\ definitions\ of\ \bar{X}\ \ and\ \bar{Y}, and\ expressing\ \sum_{i=1}^{N} X_i \ \ and\ \sum_{i=1}^{N} Y_i)$$

$$\sum_{i=1}^{N} X_i Y_i \; - \; N\,\bar{X}\bar{Y}\ (\text{as} - \bar{X}(N\bar{Y})\ \text{and}\ \ N\bar{X}\bar{Y}\ \text{cancel})$$

This completes the proof.

Note: from here it follows that when X=Y, then:

$$\sum_{i=1}^{N}(X_i - \bar{X})^2 = \; \sum_{i=1}^{N} X_i^2 - \; N\bar{X}^2$$

We are going to see these in econometrics part the course when we start talking about regression analysis.