ECN 102: Analysis of Economics Data

Chapter 1: Introduction to Data

Remy Beauregard

Chapter 1: Introduction to Data

Population vs. Sample Data

We will now begin our discussion of data. Data can be a powerful tool to answer important questions in Economics and beyond, but we must be sure we are understanding and using our data correctly. Today we will go over different types of data we will encounter in this class.

Population vs. Sample Data

We will now begin our discussion of data. Data can be a powerful tool to answer important questions in Economics and beyond, but we must be sure we are understanding and using our data correctly. Today we will go over different types of data we will encounter in this class.

- If we can collect data for every possible individual, we call this **population** data. Descriptives based on these are called *population parameters*.

Population vs. Sample Data

We will now begin our discussion of data. Data can be a powerful tool to answer important questions in Economics and beyond, but we must be sure we are understanding and using our data correctly. Today we will go over different types of data we will encounter in this class.

- If we can collect data for every possible individual, we call this **population** data. Descriptives based on these are called *population parameters*.
- If we can collect data for only a subset of the population, we call that **sample** data. Descriptives based on these are called *sample statistics*.

Population vs. Sample Data

We will now begin our discussion of data. Data can be a powerful tool to answer important questions in Economics and beyond, but we must be sure we are understanding and using our data correctly. Today we will go over different types of data we will encounter in this class.

- If we can collect data for every possible individual, we call this **population** data. Descriptives based on these are called *population parameters*.
- If we can collect data for only a subset of the population, we call that **sample** data. Descriptives based on these are called *sample statistics*.

To be able to make assertions about the broader population, our typical goal, we want our sample data to be *representative*. We will talk more about what this means later.

Types of Data

There are many types of data we may encounter as well:

Types of Data

There are many types of data we may encounter as well:

- **Categorical** data encodes which of many categories something belongs to, e.g. are you employed?

Types of Data

There are many types of data we may encounter as well:

- **Categorical** data encodes which of many categories something belongs to, e.g. are you employed?
- **Numerical** or **quantitative** data encodes numerical information about something, e.g. income level

Types of Data

There are many types of data we may encounter as well:

- **Categorical** data encodes which of many categories something belongs to, e.g. are you employed?
- **Numerical** or **quantitative** data encodes numerical information about something, e.g. income level

Furthermore, *continuous* numerical data can take any real number as a value, e.g. GDP, while *discrete* numerical data can take only separate, countable values (often whole numbers), e.g. number of classes taken in a quarter

Types of Data

There are many types of data we may encounter as well:

- **Categorical** data encodes which of many categories something belongs to, e.g. are you employed?
- **Numerical** or **quantitative** data encodes numerical information about something, e.g. income level

Furthermore, *continuous* numerical data can take any real number as a value, e.g. GDP, while *discrete* numerical data can take only separate, countable values (often whole numbers), e.g. number of classes taken in a quarter

Finally, we may need or want to *transform* our data in some way for analysis, e.g. calculating a growth rate

Observational vs. Experimental Data

The method we use to obtain our data also matters:

Observational vs. Experimental Data

The method we use to obtain our data also matters:

- **Observational** data comes from studying an uncontrolled environment, e.g. UCD enrollment. While this type of data is typically easier to obtain, it often makes it challenging to determine cause and effect in a relationship between variables.

Observational vs. Experimental Data

The method we use to obtain our data also matters:

- **Observational** data comes from studying an uncontrolled environment, e.g. UCD enrollment. While this type of data is typically easier to obtain, it often makes it challenging to determine cause and effect in a relationship between variables.
- **Experimental** data comes from a randomized experiment specifically designed by a researcher, e.g. strategic economic behavior in a lab setting. This type of data is harder to obtain but may yield much easier interpretation of cause and effect.

Data Structure: Cross-Section, Time Series, Panel

We can further classify our data by thinking about how many *units of observation* and *points in time* it covers:

Data Structure: Cross-Section, Time Series, Panel

We can further classify our data by thinking about how many *units of observation* and *points in time* it covers:

- **Cross-sectional** data covers many units at a single point in time

Data Structure: Cross-Section, Time Series, Panel

We can further classify our data by thinking about how many *units of observation* and *points in time* it covers:

- **Cross-sectional** data covers many units at a single point in time
- **Time-series** data covers one unit over multiple points in time

Data Structure: Cross-Section, Time Series, Panel

We can further classify our data by thinking about how many *units of observation* and *points in time* it covers:

- **Cross-sectional** data covers many units at a single point in time
- **Time-series** data covers one unit over multiple points in time
- **Panel** or **longitudinal** data covers multiple (repeated) units over multiple points in time

Data Structure: Cross-Section, Time Series, Panel

We can further classify our data by thinking about how many *units of observation* and *points in time* it covers:

- **Cross-sectional** data covers many units at a single point in time
- **Time-series** data covers one unit over multiple points in time
- **Panel** or **longitudinal** data covers multiple (repeated) units over multiple points in time
- **Repeated cross-sectional** data covers multiple sets of *non-repeated* units over multiple points in time

Variable Subscript Notation

We denote a variable $x_{\dots}$ with its necessary subscripts depending on the kind of data it contains:

- The i subscript corresponds to units of observation
- The t subscript corresponds to points in time

What kind of variable would x_i be? x_t? x_{it}?

Variable Subscript Notation

We denote a variable $x_{\dots}$ with its necessary subscripts depending on the kind of data it contains:

- The i subscript corresponds to units of observation
- The t subscript corresponds to points in time

What kind of variable would x_i be? x_t? x_{it}?

x_i would be a cross-sectional variable, x_t would be a time-series variable, and x_{it} would be a panel variable.

Correlation vs. Causation

The main goal for a field like Economics is ultimately *causality*, or the robust determination that x *causes* y. Experimental and quasi-experimental methods speak to this concern, but are often unfeasible or impractical to implement. For this class, we will restrict our discussion to *correlational* variable relationships, where x and y are covarying but we cannot claim causality.

We *cannot* say: x going up *causes* y to go up

We *can* say: x going up *is associated with* y going up as well

Chapter 2: Summary Statistics

Summation Notation

Summation notation: how we add across observations

$$x_1 + x_2 + x_3 + x_4 = \sum_{i=1}^{4} x_i$$

Summation Notation

Summation notation: how we add across observations

$$x_1 + x_2 + x_3 + x_4 = \sum_{i=1}^{4} x_i$$

"Add up the first four observations of a variable x"

Summation Notation: General Form

Generally for a variable x with sample size n we write:

$$\sum_{i=1}^{n} x_i$$

Summation Notation: General Form

Generally for a variable x with sample size n we write:

$$\sum_{i=1}^{n} x_i$$

"Sum up all observations in our sample from 1 to n"

Summation Notation: Operations

We can also add/subtract and multiply/divide elements within a summation:

$$
\begin{aligned}
X &= \sum_{i=1}^{3} 2x_i + 7 \\
&= \underbrace{2x_1 + 7}_{i=1} + \underbrace{2x_2 + 7}_{i=2} + \underbrace{2x_3 + 7}_{i=3} \\
&= 2(x_1 + x_2 + x_3) + 7(3) \\
&= 2\sum_{i=1}^{3}(x_i) + 21
\end{aligned}
$$

Summation Notation: Properties

For a constant c with sample size n:

$$\sum_{i=1}^{n} c = nc \text{ and } \sum_{i=1}^{n} c(x_i) = c\sum_{i=1}^{n} x_i$$

Summation Notation: Practice

What do the following sums add to?

$$\sum_{i=5}^{6} i$$

$$\sum_{i=1}^{3}(2i+1) = 2\sum_{i=1}^{3}(i) + 3(1)$$

$$\sum_{i=1}^{3} 2^i$$

Summation Notation: Solutions

What do the following sums add to?

$$\sum_{i=5}^{6} i = 5 + 6 = 11$$

$$\sum_{i=1}^{3} (2i + 1) = 3 + 5 + 7 = 15$$

$$\sum_{i=1}^{3} 2^i = 2 + 4 + 8 = 14$$

End of Lecture Material

Knowledge Check 1

What types of data are the following examples and why?

a) Quarterly data on average US mortgage rates 2005-2015

b) Data on the mode of transit used to commute around Davis from a sample of 200 students

c) Data on the number of students attending each elementary school in Yolo county this year

d) Results from a double-blind medical trial of a new drug on immediate patient outcomes