Econ 1620: Introduction to Econometrics

Week 1, Lecture 1

Introduction (SW ch.1)

Some descriptive statistics (ASW ch. 1)

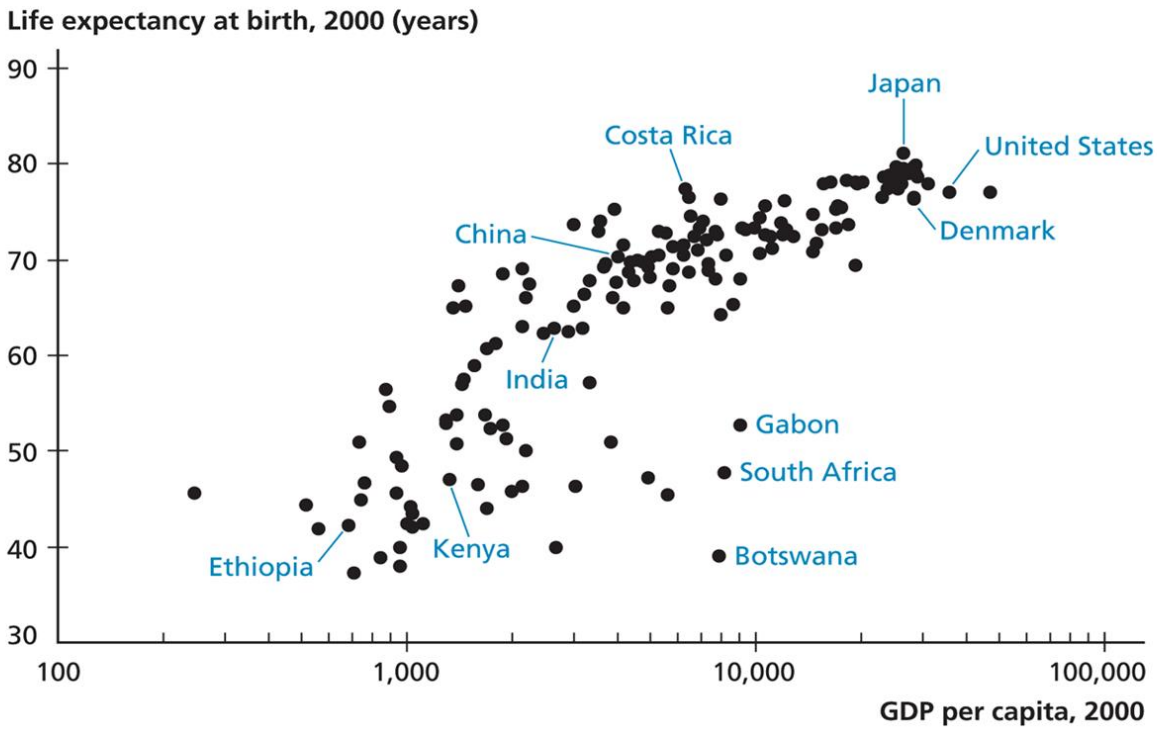
Dimitra Politi -- Brown University -- Spring 2015

Why this class is important

- You are bombarded with information every day, but you have to learn how to learn from it, as well as understand its limitations.
- Processing raw data and making sense of it is an essential asset in workplace and daily life.
- You will always deal with uncertainty in your life, and you need to understand what its implications are.
- Learning more about probability and statistics is the first step you have to take in that direction.

A simple graph: is it really that simple?

FIGURE 6.2
Life Expectancy Versus GDP per Capita



Source: Heston et al. (2002), World Bank (2003b).

How would you interpret this graph?

So what exactly are Statistics and Econometrics?

- Statistics is a tool used to collect, process, summarize, analyze, and interpret data.
- The term *statistics* can also refer to numerical facts such as averages, medians, percents, and index numbers.

- Econometrics is the application of statistical methods to the study of economic data and problems.

- Statistics and econometrics help us make sense of information, and thus help us in decision-making.
- Many applications!
  - Economics and related fields (finance, accounting, marketing…)
  - Biology and medicine
  - Psychology
  - Political science
  - Sports
  - Engineering

Econometrics: why bother?

A) Test Economic Theory, evaluate competing hypotheses

- Example: Is there discrimination in the labor market?
- Economic theory: If there is no discrimination, then two similar candidates should have similar chances of getting a callback interview.
- Economic theory: If there is no discrimination, similar workers should get similar wages.
- Get testable implications of a theory and bring them to the data.

More uses of econometrics

B) Inform policymakers, make policy recommendations

- There might be a lot of good ideas on the table, but resources are limited.
- How do you pick best value for (taxpayers') money?
- Example #1: Project evaluation: Does reducing class size improve students' performance?
- Example #2: What is the best way to improve school attendance in developing countries?

Project evaluation

- Deriving meaningful relationships from data can be hard.
- Causal effects are tricky to measure.
- Ideal: Construct the counterfactual.
- Contrast with randomized controlled experiments, where researcher has direct control over conditions.
- Uncovering social interactions is much harder than medical trials! Treatment and control groups are often hard to define!

More uses of Econometrics

C) Fit mathematical models to data

- Quantify economic relationships
- Example #1: What is the price elasticity of cigarettes?
- Example #2: How much does government spending crowd-out private investment?
- Economic theory provides us with relationships between variables, but provides no numerical values.
- Very complicated to estimate in practice, lots of variables moving at the same time.

More uses of Econometrics

D) Economic forecasting

- Example #1: What will be the rate of inflation next year?
- Example #2: By how much will GDP grow in the next semester?
- Example #3: What will happen to house prices in the next three months?
- Example #4: How will a firm's sales change after it invests in new equipment?
- Economists often use past data, economic theory, and statistical tools to make predictions.

The beauty of uncertainty

- Uncertainty is a major aspect of data analysis.
- How do you know that your predictions wouldn't change if you looked at different data from the same population?
- How do you know that the relationships in your sample are true for other populations?
- The statistical framework we develop should allow us to give numerical values to relationships, but also to measure how precise these answers are.

Data: Where do they come from?

- Experimental data
  - Specifically designed for research purposes, similar idea to lab experiments.
  - Can measure causal effects.
  - Very rare in social sciences: expensive, hard, and often unethical to collect.

- Observational data: the most common case
  - Collected from observing real-world behaviour.
  - Means of collection: surveys, administrative records, historical archives, national censuses…
  - Real-world data pose big challenges in the estimation of causal effects. Econometrics develops methods to deal with these challenges.

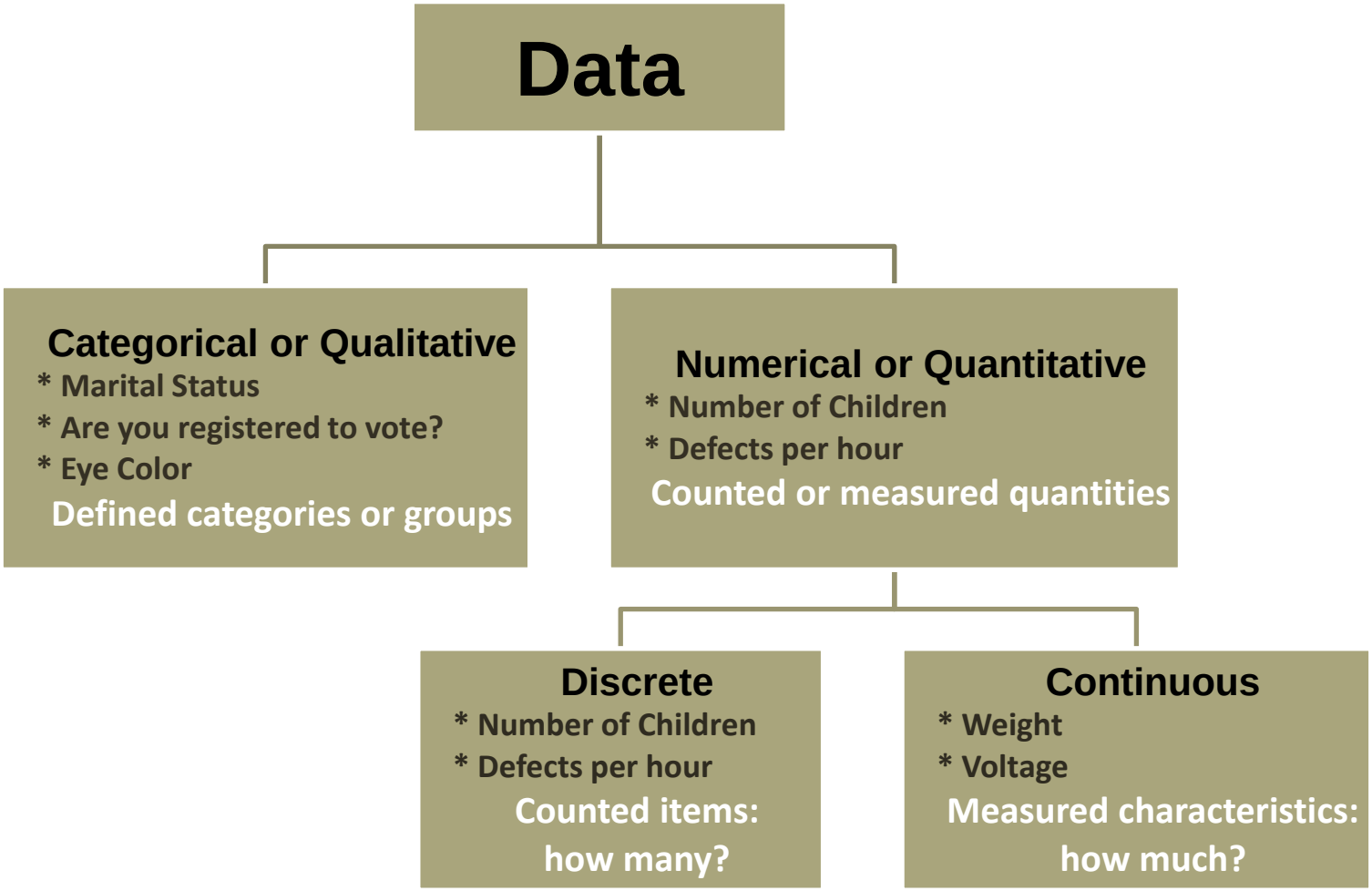
What types of data are there?

- Cross-sectional data
  - Multiple entities (individuals, firms, countries…) at a single moment in time.
- Time-series data
  - A single entity observed in multiple time periods.
- Panel (or longitudinal) data
  - Multiple entities, each one observed in multiple time periods.

Cross-section, time series or panel?

- A representative survey of 400 French households in May 2010.
- Canadian hospital patients medical records.
- Average height in UK soldiers from 1850 to 2000.
- Data on cigarette consumption for US states from 1985 to 1995.
- Australian inflation and unemployment numbers over the last decade.
- University graduation rates for OECD countries in 2000.

More data distinctions



Data measurement levels

- The level of measurement determines how much information is contained in our data and what kind of analysis is best.
- **Nominal data**: categories without ordering
  - Example: In a dataset of students, each student's declared concentration (Economics, Biology, etc) classifies our observations into categories.
  - The label can be numeric (e.g. 1 for Economics, 2 for Biology) or non-numeric, but there is no natural order: 1 is not bigger or smaller than 2.

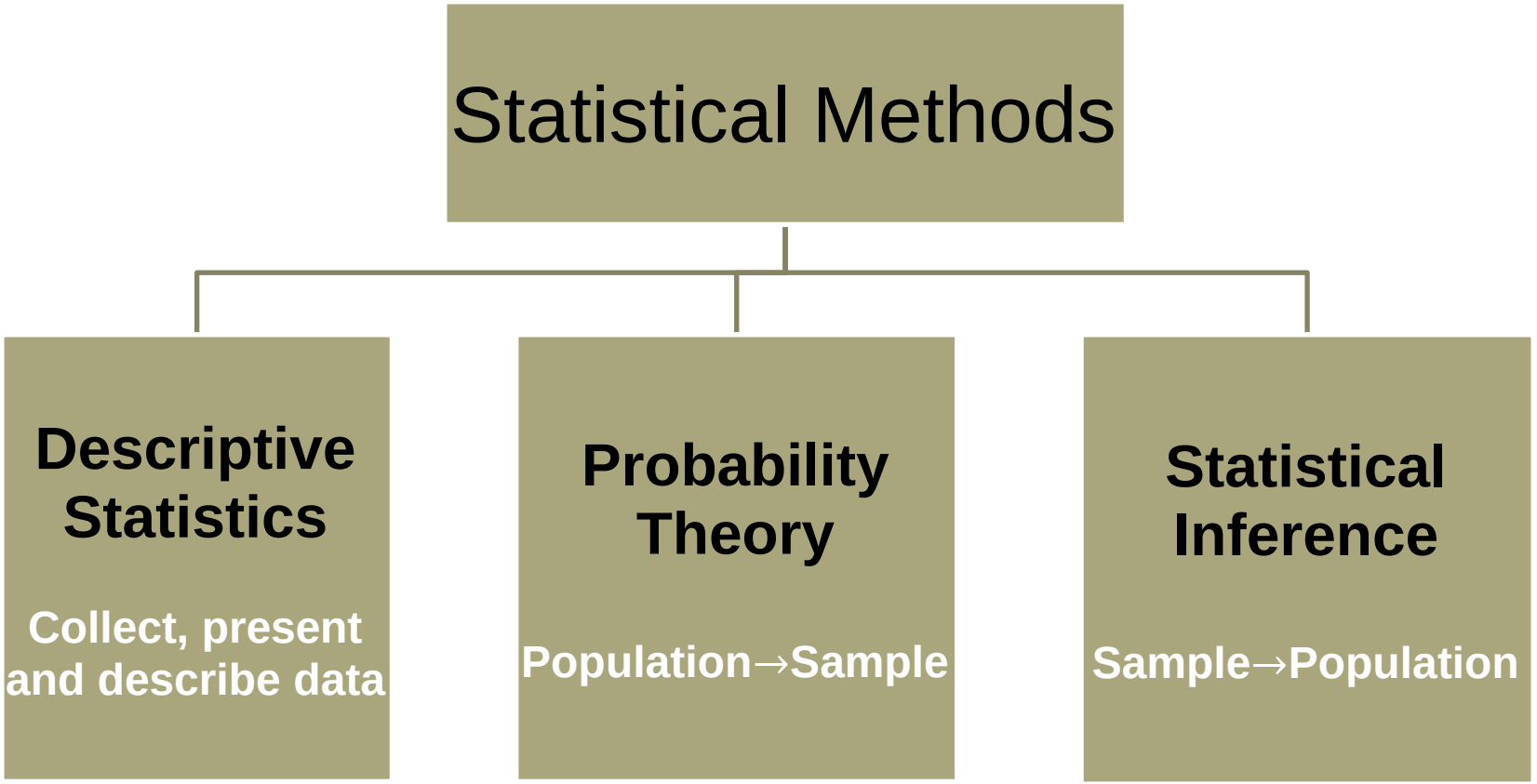
- **Ordinal data:** ordered categories
  - Example: In a dataset of students, each student's year (Freshman, Sophomore, etc) classifies our observations into categories.
  - The label can be numeric (e.g. 1 for Freshman, 2 for Sophomore) or non-numeric, and order has meaning here.

Measurement levels (continued)

- **Interval data**: the interval between observations is expressed in terms of a fixed unit of measure
  - Example: In a dataset of temperatures by location, we might have records of 80 degrees Fahrenheit in Orlando, Florida, and only 20 degrees Fahrenheit in St. Paul, Minnesota. The difference in temperature is 60 degrees Fahrenheit but we cannot say that it is four times as warm in Orlando as it is in St. Paul!
  - Interval data are always numeric.

- **Ratio data**: the ratio between two values is meaningful, zero is part of the scale.
  - Example: In employee data on hours worked, if Melissa's record shows 72 hours, while Kevin's record shows 36 hours, we can say that Melissa has worked twice as many hours as Kevin.
  - Ratio data are always numeric.

What do we do with the data?



Descriptive Statistics

- Collect information, process it and analyze it, in order to present/describe it in a useful way.

  - Example: Claire, an Econ 1620 student, would like to have a better understanding of the difficulty of the course. She examines 100 scores from last year's midterm exam.
  - Sample of 100 test scores:

80	80	70	60	30	30	90	60	60	80
50	50	60	70	60	90	40	70	100	70
70	40	100	70	90	60	40	70	50	80
80	30	50	80	60	90	90	70	70	50
70	80	20	70	70	90	80	50	70	90
20	70	80	70	70	70	80	90	60	60
100	80	60	60	70	60	40	80	90	80
70	70	60	100	70	50	80	90	90	50
50	50	80	70	80	80	90	90	90	60
90	70	80	100	60	40	90	50	90	80

Descriptive statistics example

- How can Claire present the data in an informative way?
  - She can, for example, calculate **frequencies**, or **relative frequencies** for each score.
  - She can plot a **histogram** showing the **distribution** of scores.
  - She can calculate various **sample statistics**: the **average** score, the **median** score, the **maximum/minimum**, the **standard deviation**…

Statistic	Value in the sample
Class average	69.4
Median	70
Maximum	100
Minimum	20
Standard deviation	18.35

  - All of the above are examples of descriptive statistics.

The distribution of test scores

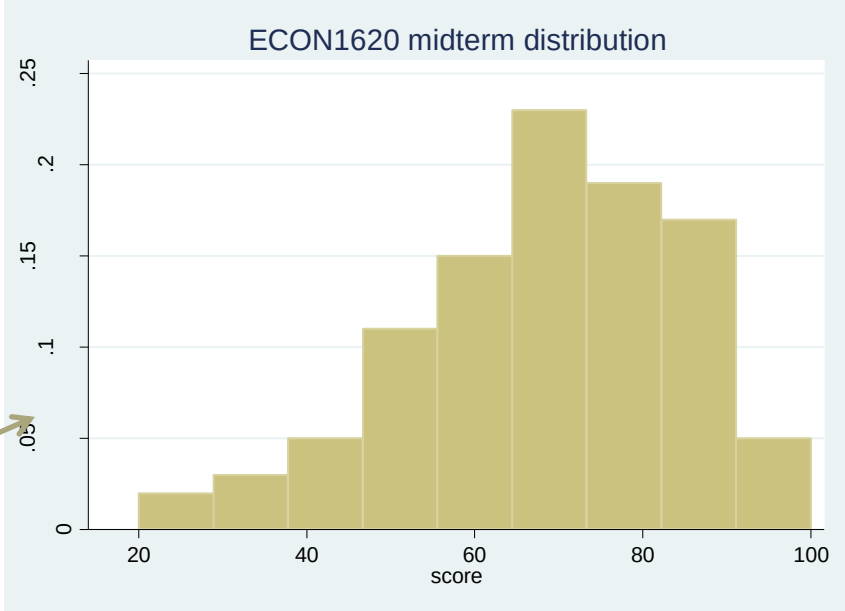
Score	Absolute Frequency	Percent Frequency
20-40	10	10%
50	11	11%
60	15	15%
70	23	23%
80	19	19%
90-100	22	22%
	100	100%

$$\frac{Absolute\ frequency}{Sample\ size} \cdot 100$$

observations = sample size

Relative frequencies

Histogram of scores



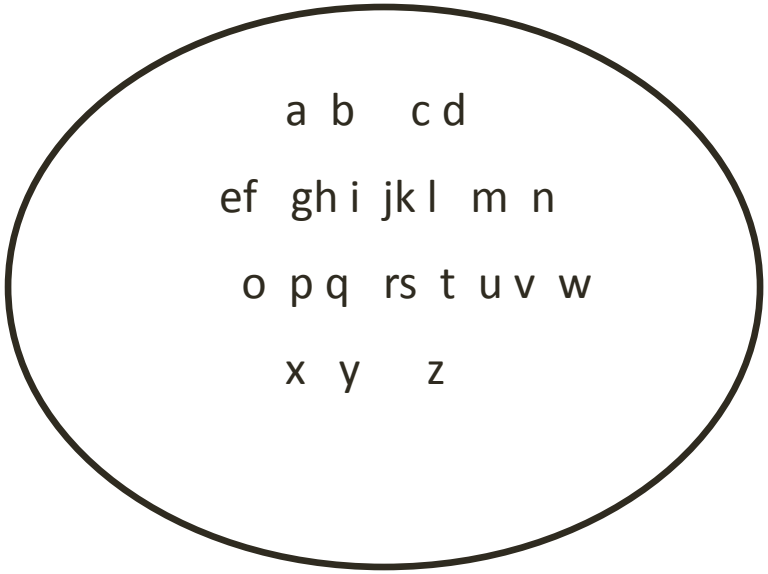
Sample vs. Population

- In her quest for information, Claire might want to think about the fact that she only has access to 100 midterm scores. This was not the entire class.

- Claire has a **sample** from the **population** of test scores, and her calculations are based on the scores in her sample, not the whole population.

Sample vs. Population

Population



Values calculated using population data are called parameters. Any characteristic of the population is a parameter.

Sample

Values computed from sample data are called statistics. Any function of the data in a sample is a statistic.

Examples of populations

- Names of all registered voters in the United States
- Incomes of all families living in Daytona Beach
- Annual returns of all stocks traded on the New York Stock Exchange
- Grade point averages of all the students in your university

Probability Theory

- Probability theory is the cornerstone of statistics. Much more on that next week…

- Reasoning is from population to sample: we know (or assume) some characteristics about a population, and we determine how likely we are to obtain a given sample from that population.

- Examples:
  - What's the probability of getting a head on the toss of a single fair coin?
  - What is the probability of winning lotto?

Statistical Inference

Inference is the process of drawing conclusions or making decisions about a population based on sample results.

- Estimation and confidence intervals
  - Estimate the mean weight of a population using the sample mean weight
  - Get a range of likely values for the mean weight in the population
- Hypothesis testing
  - Test the claim that the population mean weight is 140 pounds

Random sampling

- When we conduct statistical inference, the quality of the sample we have is very important. A bad or unsuitable sample might lead us to the wrong conclusions about the population.

- Simple random sampling is a procedure in which
  - each member of the population is chosen strictly by chance,
  - each member of the population is equally likely to be chosen,
  - every possible sample of n objects is equally likely to be chosen

- The resulting sample is called a random sample (we will apply a more technical definition later in this course).

Summary

- We defined statistics and Econometrics, and caught a glimpse of their coolness!
- Reviewed types of data and measurement levels
- Introduced key definitions:
  - Population vs. Sample
  - Parameter vs. Statistic
  - Descriptive vs. Inferential statistics
- Described random sampling