Lecture 2

Text: Chapter I, II

STAT 8010 Statistical Methods I
August 23, 2019

Whitney Huang
Clemson University

Agenda

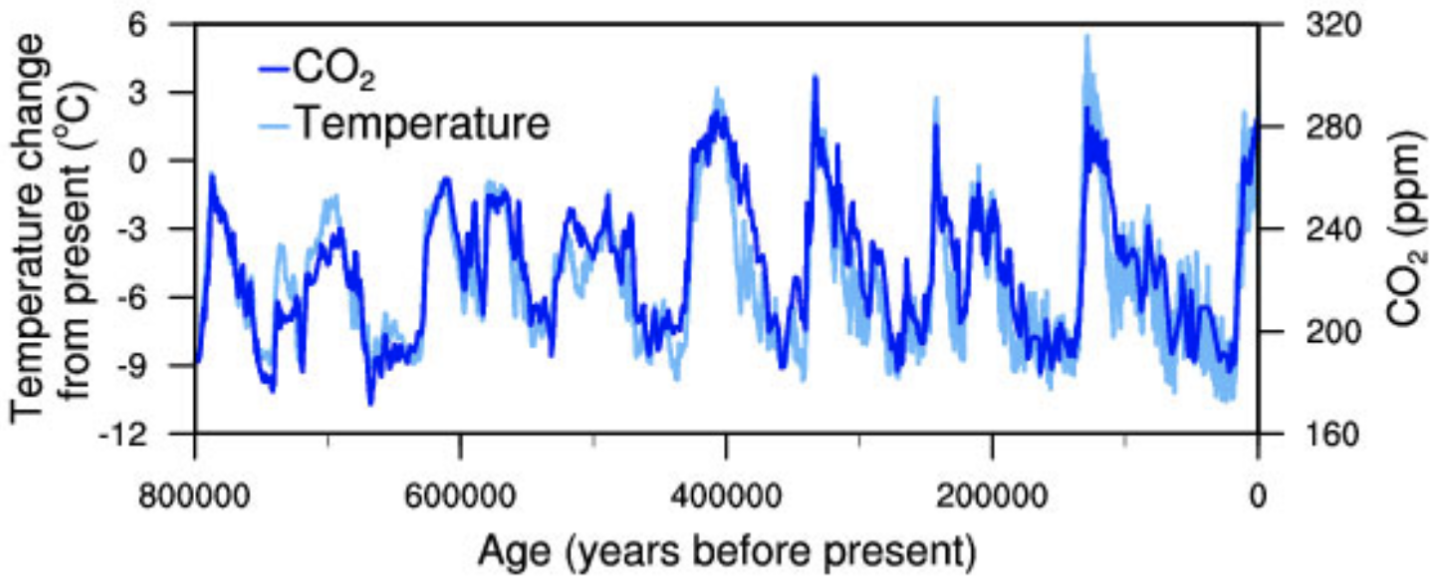
Announcements

- Syllabus and lecture notes are in CANVAS and my personal website (link: https://whitneyhuang83.github.io/stat8010_2019Fall.html)
- **Academic Continuity Statement** is added in the updated syllabus (link: https://whitneyhuang83.github.io/STAT8010_Syllabus_2019_Fall.pdf)
- Please talk to me if you would like to share your data set to be used for this class

Motivation: Why Study Statistics?

- To be able to effectively conduct (empirical) research
- To be an informed “consumer”
- To further develop critical and analytic thinking skills

Temperature and Carbon Dioxide CO_2

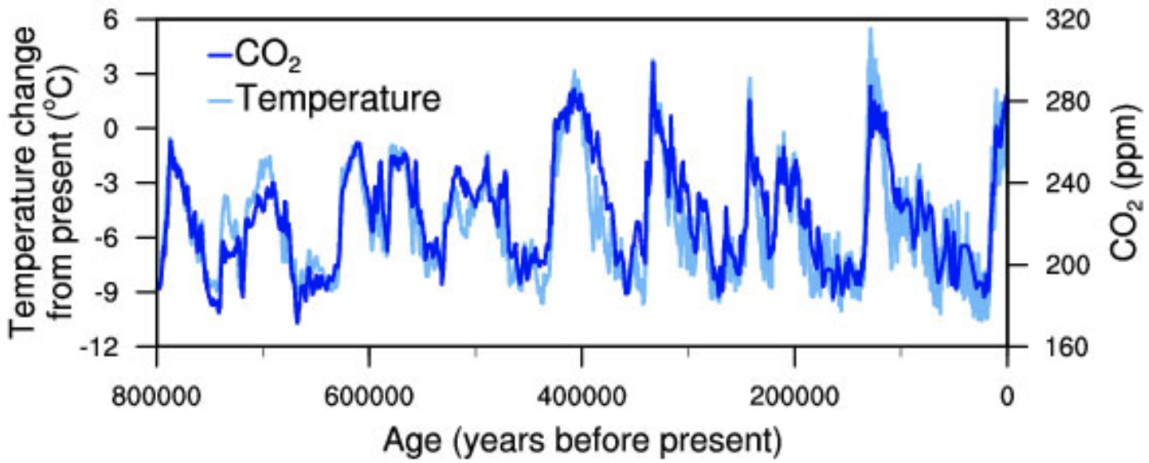


Temperature change (light blue) and carbon dioxide change (dark blue) measured from the EPICA Dome C ice core in Antarctica (Jouzel et al. 2007; Lüthi et al. 2008).

Research questions:

- Does temperature correlate with CO_2? If so, how to "predict" temperature using CO_2?
- Can we make some statement about the causation between temperature and CO_2?

Temperature CO_2 data revisited



Temperature change (light blue) and carbon dioxide change (dark blue) measured from the EPICA Dome C ice core in Antarctica (Jouzel et al. 2007; Lüthi et al. 2008).

- Stating the problem, identifying the variable(s) of interest, and gathering data
- Summarizing the data
- Analyzing the data
- Reporting and interpreting the results

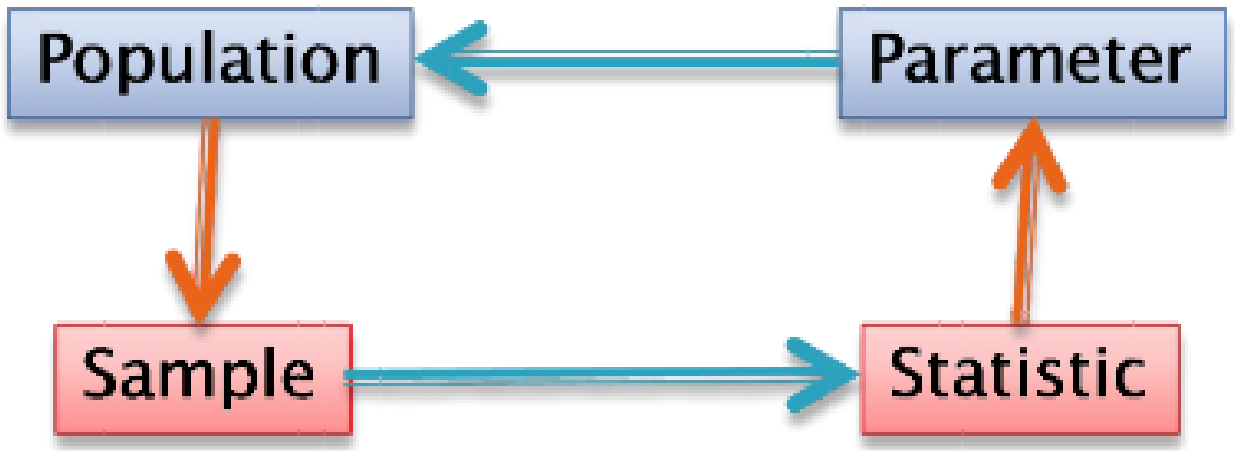
Terminology

- A **unit** is a single entity (person or object) whose characteristics are of interest
- A **population of units** is the complete collection of units about which information is sought
- A **population** is a set of all measurements corresponding to each unit in the entire collection of units about which information is sought
- A **sample** is a subset of measurements selected from the population of interest

Statistical Science concerned with using sample information to make inference about populations

Population (Parameters) vs. Sample (Statistics)

- We use parameters to describe the population
- We use statistics to describe the sample with respect to the population



Types of variables

A **variable** is a characteristic of an individual or object that may vary for different observations

There are two main types of variables, qualitative (aka categorical) and quantitative (aka numerical)

- Qualitative variable: has labels or names used to identify an attribute of an element. Qualitative data use either the **nominal** or **ordinal** scale of measurement

Types of variables

A **variable** is a characteristic of an individual or object that may vary for different observations

There are two main types of variables, qualitative (aka categorical) and quantitative (aka numerical)

- Qualitative variable: has labels or names used to identify an attribute of an element. Qualitative data use either the **nominal** or **ordinal** scale of measurement
  - Nominal: order does not matter e.g. Gender

Types of variables

A **variable** is a characteristic of an individual or object that may vary for different observations

There are two main types of variables, qualitative (aka categorical) and quantitative (aka numerical)

- Qualitative variable: has labels or names used to identify an attribute of an element. Qualitative data use either the **nominal** or **ordinal** scale of measurement
  - Nominal: order does not matter e.g. Gender

Types of variables

A **variable** is a characteristic of an individual or object that may vary for different observations

There are two main types of variables, qualitative (aka categorical) and quantitative (aka numerical)

- Qualitative variable: has labels or names used to identify an attribute of an element. Qualitative data use either the **nominal** or **ordinal** scale of measurement
  - Nominal: order does not matter e.g. Gender
  - Ordinal: order does matter e.g. Education levels

Types of variables

A **variable** is a characteristic of an individual or object that may vary for different observations

There are two main types of variables, qualitative (aka categorical) and quantitative (aka numerical)

- Qualitative variable: has labels or names used to identify an attribute of an element. Qualitative data use either the **nominal** or **ordinal** scale of measurement
  - Nominal: order does not matter e.g. Gender
  - Ordinal: order does matter e.g. Education levels

Types of variables

A **variable** is a characteristic of an individual or object that may vary for different observations

There are two main types of variables, qualitative (aka categorical) and quantitative (aka numerical)

- Qualitative variable: has labels or names used to identify an attribute of an element. Qualitative data use either the **nominal** or **ordinal** scale of measurement
  - Nominal: order does not matter e.g. Gender
  - Ordinal: order does matter e.g. Education levels
- Quantitative variable: has numeric values that indicate how much or how many of something. Quantitative data uses either the **interval** or **ratio** scale

Types of variables

A **variable** is a characteristic of an individual or object that may vary for different observations

There are two main types of variables, qualitative (aka categorical) and quantitative (aka numerical)

- Qualitative variable: has labels or names used to identify an attribute of an element. Qualitative data use either the **nominal** or **ordinal** scale of measurement
  - Nominal: order does not matter e.g. Gender
  - Ordinal: order does matter e.g. Education levels
- Quantitative variable: has numeric values that indicate how much or how many of something. Quantitative data uses either the **interval** or **ratio** scale
  - Interval: difference of quantities that are meaningful but ratios of quantities that cannot be compared e.g. temperature with the Celsius scale

Types of variables

A **variable** is a characteristic of an individual or object that may vary for different observations

There are two main types of variables, qualitative (aka categorical) and quantitative (aka numerical)

- Qualitative variable: has labels or names used to identify an attribute of an element. Qualitative data use either the **nominal** or **ordinal** scale of measurement
  - Nominal: order does not matter e.g. Gender
  - Ordinal: order does matter e.g. Education levels
- Quantitative variable: has numeric values that indicate how much or how many of something. Quantitative data uses either the **interval** or **ratio** scale
  - Interval: difference of quantities that are meaningful but ratios of quantities that cannot be compared e.g. temperature with the Celsius scale

Types of variables

A **variable** is a characteristic of an individual or object that may vary for different observations

There are two main types of variables, qualitative (aka categorical) and quantitative (aka numerical)

- Qualitative variable: has labels or names used to identify an attribute of an element. Qualitative data use either the **nominal** or **ordinal** scale of measurement
  - Nominal: order does not matter e.g. Gender
  - Ordinal: order does matter e.g. Education levels
- Quantitative variable: has numeric values that indicate how much or how many of something. Quantitative data uses either the **interval** or **ratio** scale
  - Interval: difference of quantities that are meaningful but ratios of quantities that cannot be compared e.g. temperature with the Celsius scale
  - Ratio: ratios of quantities that are meaningful e.g. Height

Example

Grade	Major	GPA	Credit hours
Sophomore	Psychology	3.14	30
Senior	Spanish	2.89	105
Senior	Religion	3.01	99
Freshman	Philosophy	2.45	12

1. How many units are in the data set?
2. How many variables are in the data set?
3. What type of variable is each variable in the data set (be sure to answer both qualitative or quantitative as well as nominal, ordinal, interval, or ratio).

Example

For this example, answer what type of variable each of the following are

1. Smoking status
2. Income
3. Level of satisfaction
4. Clothing size (s, m, l, xl)
5. Time taken to run a mile

Observational vs. Experimental Studies

- Observational study: a study in which the investigator observes characteristics of members of an existing population(s) in order to draw conclusions

- Experimental Study: a study in which the investigator observes how a response variable behaves when the researcher manipulates one or more factors in order to determine the effect of those factors on the response.

Example

State whether the study is observational or experimental

- A researcher wants to know if smoking during pregnancy leads to children with lower IQ scores. She looks at 200 pregnant women and records smoking status along with the subsequent IQ score (measured a few years after birth)

- A scientist tries his weight loss drug on a group of monkeys with identical diets. 40 monkeys are randomly assigned to either get the drug or not get the drug (20 in each group). The weight gained or lost was recorded for each monkey.

Cross-sectional vs. Time series data

We have two types of data set based on how the data were collecting

- Cross-sectional: data collected at the same or approximately the same point in time
- Time series: data collected over several time periods

Example

For this problem, state whether the variables included are cross-sectional or time series

1. United States current temperatures
2. Temperatures in Clemson from 1950-2015
3. Total salary of the LA Lakers throughout the 2010s
4. Salaries of all NBA teams in 2019.

Statistical Sampling

In Statistics, sampling is procedure to select a subset from a statistical population that is representative of the population. There are several types of sampling as follows:

- Simple random sampling (SRS): a sample selected such that each element in the population has the same probability of being selected

Statistical Sampling

In Statistics, sampling is procedure to select a subset from a statistical population that is representative of the population. There are several types of sampling as follows:

- Simple random sampling (SRS): a sample selected such that each element in the population has the same probability of being selected
- Stratified random sample: elements in the population are first divided into groups and a simple random sample is then taken from each group

Sampling cont'd

- Probability sampling: elements in the population are selected with a known probability of being included in a sample

Sampling cont'd

- Probability sampling: elements in the population are selected with a known probability of being included in a sample
- Cluster sampling: the elements in the population are first divided into separate groups called clusters and then a simple random sample of the clusters is taken that all elements in a selected cluster are part of a sample

Sampling cont'd

- Probability sampling: elements in the population are selected with a known probability of being included in a sample
- Cluster sampling: the elements in the population are first divided into separate groups called clusters and then a simple random sample of the clusters is taken that all elements in a selected cluster are part of a sample
- Systematic sampling: randomly select one of the first k elements from the population and then every k_{th} element thereafter is picked

Sampling cont'd

- Probability sampling: elements in the population are selected with a known probability of being included in a sample
- Cluster sampling: the elements in the population are first divided into separate groups called clusters and then a simple random sample of the clusters is taken that all elements in a selected cluster are part of a sample
- Systematic sampling: randomly select one of the first k elements from the population and then every k_{th} element thereafter is picked
- Convenience sampling: elements selected from the population on the basis of convenience

Sampling cont'd

- Probability sampling: elements in the population are selected with a known probability of being included in a sample
- Cluster sampling: the elements in the population are first divided into separate groups called clusters and then a simple random sample of the clusters is taken that all elements in a selected cluster are part of a sample
- Systematic sampling: randomly select one of the first k elements from the population and then every k_{th} element thereafter is picked
- Convenience sampling: elements selected from the population on the basis of convenience
- Judgment sampling: elements are selected from the population based on the judgment of the person doing the study.

What type of sampling was used?

1. A researcher randomly chooses houses in a town. Once a particular house is chosen everyone living in the house is surveyed

2. A school principal decides to performs an exit interview with every 14^{th} name from a list of graduating seniors

3. A biologist knows that 40% of bats are male and that 60% are female so she randomly selects 20 males and randomly selects 30 females to be in her sample

4. A graduate student wants to do a study on why people like bluegrass music and uses the people she meets at the next show she attends as her sample

5. To get an idea of the average weight of his cattle, a rancher randomly chooses to weigh 25 from his list of the animals