

# Lecture 1

## Introduction

STAT 8010 Statistical Methods I  
January 9, 2020

Whitney Huang  
Clemson University



Who is the instructor?  
Class Policies / Schedule  
Tell us about yourself  
Class Overview  
Terminology and Basic Concepts

1.1

Notes

---

---

---

---

---

---

---

## Who is the instructor?



Who is the instructor?  
Class Policies / Schedule  
Tell us about yourself  
Class Overview  
Terminology and Basic Concepts

1.2

Notes

---

---

---

---

---

---

---

## Who am I?

- **First year** Assistant Professor of Applied Statistics and Data Science

- Born in Laramie, Wyoming, grew up in Taiwan



- With a B.S. in Mechanical Engineering, switched to Statistics in graduate school

- Got my Ph.D. (Statistics) in 2017 at Purdue

University



Who is the instructor?  
Class Policies / Schedule  
Tell us about yourself  
Class Overview  
Terminology and Basic Concepts

1.3

Notes

---

---

---

---


---

---

---

How to reach me?

- Email: [wkhuang@clermson.edu](mailto:wkhuang@clermson.edu)
- Office: O-221 Martin Hall
- Office Hours: TR 11:00am – 12:00pm and by appointment



[Who is the instructor?](#)  
[Class Policies / Schedule](#)  
[Tell us about yourself](#)  
[Class Overview](#)  
[Terminology and Basic Concepts](#)

1.4

Notes

---

---

---

---

---

---

---

Class Policies / Schedule



[Who is the instructor?](#)  
[Class Policies / Schedule](#)  
[Tell us about yourself](#)  
[Class Overview](#)  
[Terminology and Basic Concepts](#)

1.5

Notes

---

---

---

---

---

---

---

Logistics

- We will meet TR 9:30am – 10:45am at M-104 Martin
- There will be two in-class exams and a (comprehensive) final. The (tentative) dates for the two exams are:
  - Exam I: Feb. 13, Thursday
  - Exam II: Mar. 26, Thursday
- The Final Exam will be given on Wednesday, Apr. 29, 8:00 am - 10:30 am.
- There will be some homework assignments (~ 7):
  - Will be due Tues by 9:30am
  - Worst grade will be dropped
- No classes on Mar. 17 & 19 (Spring Break)



[Who is the instructor?](#)  
[Class Policies / Schedule](#)  
[Tell us about yourself](#)  
[Class Overview](#)  
[Terminology and Basic Concepts](#)

1.6

Notes

---

---

---

---

---

---

---

Class Website

CANVAS and my teaching website (link:  
[https://whitneyhuang83.github.io/stat8010\\_2020Sp.html](https://whitneyhuang83.github.io/stat8010_2020Sp.html))

- Course syllabus [\[Link\]](#) / Announcements
- Lecture slides/notes
- Homework assignments
- Exam and homework schedule
- Data sets for lectures and homework



Notes

---

---

---

---

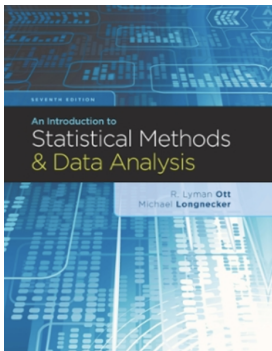
---

---

---

Recommended Textbook

An Introduction to Statistical Methods and Data Analysis,  
6<sup>th</sup> Edition. Lyman Ott and Micheal T. Longnecker,  
Duxbury, 2010; ISBN-13: 978-1305269477



Notes

---

---

---

---

---

---

---

Evaluation

- Grade Distribution:

Homework:	20%
Exam I	25%
Exam II	25%
Final Exam	30%
- Letter Grade:

>= 90.00	A
88.00 ~ 89.99	A-
85.00 ~ 87.99	B+
80.00 ~ 84.99	B
78.00 ~ 79.99	B-
75.00 ~ 77.99	C+
70.00 ~ 74.99	C
68.00 ~ 69.99	C-
<= 67.99	F



Notes

---

---

---

---

---

---

---

Tentative Topics and Dates

Week	Topic
1	Introduction
2	Data Summary and Display
3	Intro to Probability & Probability Distributions I
4	Probability Distributions II
5	Normal Distribution & Central Limit Theorem
6	Exam I
7	Statistical Inference for a Single Sample
8	Statistical Inference for Two Samples
9	One Way ANOVA & Multiple Comparisons
10	Randomized Complete Block Designs
11	No Classes–Spring Break
12	Exam II
13	Inference on Proportions
14	Contingency Table Analysis
15	Correlation and Simple Linear Regression I
16	Simple Linear Regression II



Who is the instructor?  
Class Policies / Schedule  
Tell us about yourself  
Class Overview  
Terminology and Basic Concepts

Notes

---

---

---

---

---

---

---

---

Computing

We will use software to perform statistical analyses. The recommended software for this course are JASP and R/Rstudio

- JASP
  - a free/open-source graphical program for statistical analysis
  - available at <https://jasp-stats.org/>
- R / RStudio
  - a free/open-source programming language for statistical analysis
  - available at <https://www.r-project.org/> (R); <https://rstudio.com/> (Rstudio)

You are welcome to use a different package (e.g. SAS, JMP, SPSS, Minitab) if you prefer



Who is the instructor?  
Class Policies / Schedule  
Tell us about yourself  
Class Overview  
Terminology and Basic Concepts

Notes

---

---

---

---

---

---

---

---

Tell us about yourself



Who is the instructor?  
Class Policies / Schedule  
Tell us about yourself  
Class Overview  
Terminology and Basic Concepts

Notes

---

---

---

---

---

---

---

---

Tell us about yourself

- Your name
- Degree program
- Your background in Statistics/Computing

CLEMSON  
UNIVERSITY

Who is the instructor?

Class Policies / Schedule

Tell us about yourself

Class Overview

Terminology and Basic Concepts

1.13

Notes

---

---

---

---

---

---

---

Class Overview

CLEMSON  
UNIVERSITY

Who is the instructor?

Class Policies / Schedule

Tell us about yourself

Class Overview

Terminology and Basic Concepts

1.14

Notes

---

---

---

---

---

---

---

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

CLEMSON  
UNIVERSITY

Who is the instructor?

Class Policies / Schedule

Tell us about yourself

Class Overview

Terminology and Basic Concepts

1.15

Notes

---

---

---

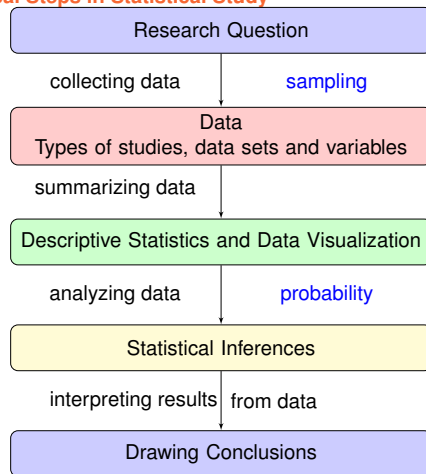
---

---

---

---

## Typical Steps in Statistical Study



## Notes

---

---

---

---

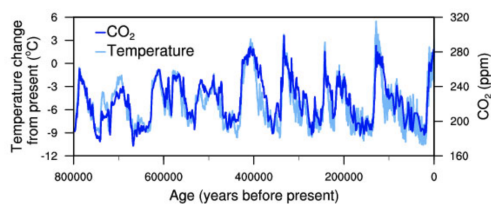
---

---

---

---

## Temperature and Carbon Dioxide CO<sub>2</sub>



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

### Research questions:

- Does temperature correlate with CO<sub>2</sub>? If so, how to "predict" temperature using CO<sub>2</sub>?
- Can we make some statement about the causation between temperature and CO<sub>2</sub>?

## Notes

---

---

---

---

---

---

---

---

## Terminology and Basic Concepts

## Notes

---

---

---

---

---

---

---

---

## Terminology

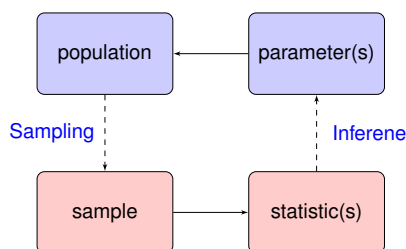
- A **unit** is a single entity (person or object) whose characteristics are of interest
- A **population of units** is the complete collection of units about which information is sought
- A **population** is a set of all measurements corresponding to each unit in the entire collection of units about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

## Notes

## Population (parameters) vs. Sample (statistics)

- We use **parameter(s)** to describe the population of interest
- We use **statistic(s)** to describe the sample with respect to the population of interest



## Notes

## Understanding Data: Types of variables

A **variable** is a characteristic of a unit 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 a unit. 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

## Notes

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

- ➊ How many units are in the data set?
- ➋ How many variables are in the data set?
- ➌ 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).



Who is the instructor?  
Class Policies / Schedule  
Tell us about yourself  
Class Overview  
Terminology and Basic Concepts

1.22

Notes

---

---

---

---

---

---

---

Example

Answer what type of variable each of the following are

- ➊ Smoking status
- ➋ Income
- ➌ Level of satisfaction
- ➍ Clothing size (s, m, l, xl)
- ➎ Time taken to run a mile



Who is the instructor?  
Class Policies / Schedule  
Tell us about yourself  
Class Overview  
Terminology and Basic Concepts

1.23

Notes

---

---

---

---

---

---

---

Observational vs. Experimental Studies

Depending on how the study were conducted, we have the following types of studies:

- **Observational study:** a study in which the investigator **observes** a variable of interest of an existing sample in order to draw conclusions
- **Experimental Study:** a study in which the investigator examines 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.



Who is the instructor?  
Class Policies / Schedule  
Tell us about yourself  
Class Overview  
Terminology and Basic Concepts

1.24

Notes

---

---

---

---

---

---

---



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.

CLEMSON  
UNIVERSITY

Who is the instructor?

Class Policies / Schedule

Tell us about yourself

Class Overview

Terminology and Basic Concepts

1.25

Notes

---

---

---

---

---

---

---

Types of Data sets

Depending on how the data were collected, we have the following types of data sets:

- **Cross-sectional dat**: data collected at the same or approximately the same point in time
- **Time series data**: data collected over several time periods
- **Spatio-temporal data**: data collected at different "locations" over several time periods

CLEMSON  
UNIVERSITY

Who is the instructor?

Class Policies / Schedule

Tell us about yourself

Class Overview

Terminology and Basic Concepts

1.26

Notes

---

---

---

---

---

---

---

Example

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

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

CLEMSON  
UNIVERSITY

Who is the instructor?

Class Policies / Schedule

Tell us about yourself

Class Overview

Terminology and Basic Concepts

1.27

Notes

---

---

---

---

---

---

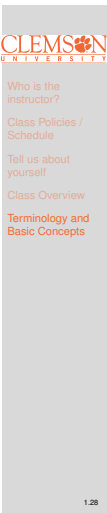
---

Summary

In this lecture, we learned

- Typical steps in statistical study
- Terminology
  - Population vs. Sample
  - Types of variables, studies, datasets

In next lecture we will learn how to summarize data both graphically and numerically



Notes

---

---

---

---

---

---

---

Notes

---

---

---

---

---

---

---

Notes

---

---

---

---

---

---

---