# Lecture 1

Introduction

STAT 8010 Statistical Methods I August 20, 2020



Who is the instructor?

Class Policies Schedule

Tell us about yourself

Class Overview

Terminology and Basic Concepts

Whitney Huang Clemson University



Who is the instructor'

Class Policie Schedule

Tell us about yoursel

Class Overview

Terminology and Basic Concepts

# Who is the instructor?

#### Who am I?

- Second 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 a Ph.D. (Statistics) in 2017 at Purdue University.





Who is the instructor?

Class Policies Schedule

Tell us about yourself

lass Overview

### How to reach me?



Who is the instructor?

Class Policies Schedule

Tell us about yourself

lass Overview

Terminology and Basic Concepts

Email: wkhuang@clemson.edu

Office: O-221 Martin Hall

 Office Hours: TR 11:00am – 12:00pm via Zoom and by appointment



Who is the instructor

#### Class Policies / Schedule

Tell us about yoursel

Class Overview

Terminology and Basic Concepts

# Class Policies / Schedule

### Logistics

- We will meet TR 3:30pm 4:45pm via Zoom
- There will be two online exams and a (comprehensive) online final. The (tentative) dates for the two exams are:
  - Exam I: Sept. 24, Thursday
  - Exam II: Oct. 29, Thursday

The **Final Exam** will be given on Wednesday, Dec. 11, 11:30 am - 2:00 pm.

- There will be some homework assignments (~ 6):
  - To be uploaded to Canvas by 3:30 pm on the due dates
  - Worst grade will be dropped
- No classes on Nov. 3 (Fall Break) & 26 (Thanksgiving)



Who is the instructor?

lass Policies / chedule

Tell us about yourself

ss Overview

#### **Class Website**

# CANVAS and my teaching website (link:

https://whitneyhuang83.github.io/STAT8010/Fall2020/stat8010\_2020Fall.html)

- Course syllabus [Link] / Announcements
- Lecture slides/notes
- Homework assignments
- Exam and homework schedule
- Data sets for lectures and homework
- R codes



Who is the instructor?

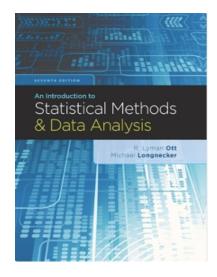
Class Policies / Schedule

Tell us about yourself

lass Overview

#### **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





Who is the instructor?

Class Policies Schedule

Tell us about yoursel

Class Overview

#### **Evaluation**

CLEMS N

Who is the instructor

Class Policies Schedule

Tell us about yoursel

Class Overview

Terminology and Bas Concepts

Homework: 20% Exam I 25% Exam II 25% Final Exam 30%

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

Letter Grade:

Grade Distribution:

# **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	Exam II
12	Inference on Proportions I
13	Inference on Proportions II
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

## Computing

We will use software to perform statistical analyses. The recommended software for this course are  ${\tt JASP}$  and

- R/Rstudio
   JASP
  - a free/open-source graphical program for statistical analysis
  - available at https://jasp-stats.org/
  - R Studio
    - 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

ss Overview



Who is the instructor?

Class Policies / Schedule

#### Tell us about yourself

Class Overview

Terminology and Basic Concepts

# Tell us about yourself

# Tell us about yourself

- CLEMS#N
  - Who is the instructor?
  - Class Policies Schedule
  - Tell us about yourself
    - Class Overview
  - Terminology and Basic Concepts

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



Who is the instructor'

Class Policies /

Tell us about yourse

#### Class Overview

Terminology and Basic Concepts

# Class Overview

# **Motivation: Why Study Statistics?**



Who is the instructor?

Class Policie: Schedule

Tell us about yourse

#### Class Overview

Terminology and Basic

To be able to effectively conduct (empirical) research

# **Motivation: Why Study Statistics?**



Who is the instructor?

Class Policie: Schedule

Tell us about yoursel

#### Class Overview

- To be able to effectively conduct (empirical) research
- To be an informed "consumer"

# **Motivation: Why Study Statistics?**



Who is the instructor?

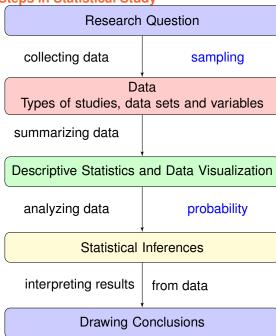
Class Policies Schedule

Tell us about yourself

#### Class Overview

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

# Typical Steps in Statistical Study





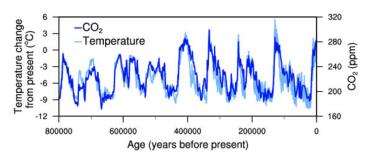
Who is the instructor'

Class Policies

Tell us about yourse

#### Class Overview

# **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 (Jouzel 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>?



Who is the instructor?

Class Policie Schedule

Tell us about yourself

#### ss Overview



Who is the instructor'

Class Policie Schedule

Tell us about yourse

ass Overview

Terminology and Basic Concepts

# **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



Who is the instructor?

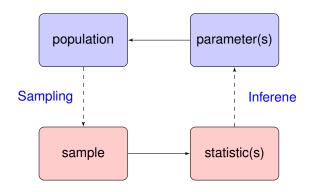
lass Policies / chedule

Tell us about yourself

Diass Overview

# 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





Who is the instructor?

Class Policies Schedule

Tell us about yourself

ass Overview

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



who is the instructor?

Class Policies Schedule

Tell us about yourself

Terminology and Basic

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



Who is the instructor?

Class Policies Schedule

Tell us about yourself

Terminology and Basic

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



Who is the instructor?

Class Policies Schedule

Tell us about yourself

Terminology and Basic

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



Who is the instructor?

Class Policies Schedule

Tell us about yourself

Terminology and Basi

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



Who is the instructor?

Class Policies : Schedule

Tell us about yourself

ass Overview

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



Who is the instructor's

Class Policies . Schedule

Tell us about yourself

ass Overview

## **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

- How many units are in the data set?
- 4 How many variables are in the data set?
- 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

ss Overview

## **Example**

Answer what type of variable each of the following are

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



vino is the instructor?

Bass Policies / Schedule

Tell us about yourself

ass Overview

# **Observational vs. Experimental Studies**



Who is the instructor?

Class Policies Schedule

Tell us about yourself

lass Overview

Terminology and Basic Concepts

Depending on how a study was 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.

## **Example**



Who is the instructor?

Class Policies Schedule

Tell us about yourself

Class Overview

Terminology and Basic Concepts

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.

## **Types of Data sets**

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



Who is the instructor's

Class Policies Schedule

Tell us about yoursel

Class Overview

Terminology and Basic

- 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

### **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.



Who is the instructor?

lass Policies / chedule

Tell us about yourself

ass Overview

### **Summary**

CLEMS N

Who is the instructor?

Class Policies Schedule

Tell us about yourself

Terminology and Basic Concepts

- 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