Lecture 1

Text: Chapter 1,2,3

STAT 8010 Statistical Methods I
August 21, 2019

Whitney Huang
Clemson University

Agenda

1. Logistics
2. Introduction
3. Definitions
4. Sampling
5. Types of Data
6. Summarizing data
7. Population vs. Sample
8. Numerical Summaries
9. Graphical Summary: Boxplots

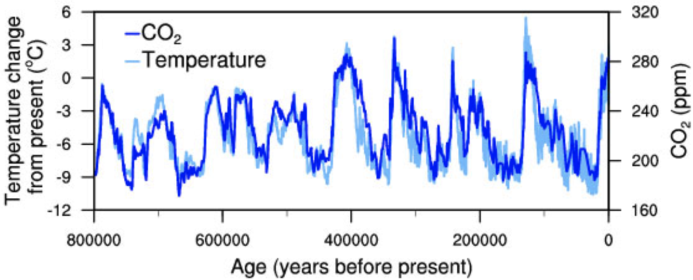
STAT 8010, Section 003

- We will meet MWF 1:25pm – 2:15pm at M-104 Martin
- We will have two Exams (late Sept. and Oct.), one Final Exam (Dec. 13, 3:00pm–5:30pm), and some Homework assignments
- No classes on Oct. 14 (Fall break) and Nov. 27, 29 (Thanksgiving)
- Office hours: TBD

Motivation: Why Study Statistics?

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

An Example



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).

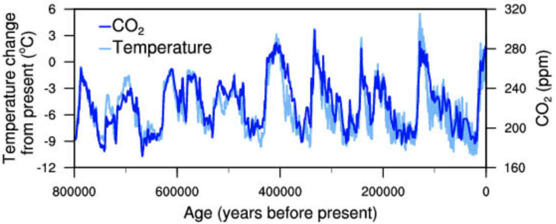
Research questions:

- 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$_2$?

STAT 8010 Overview

1. **Data collection**: sampling methods, data types
2. **Descriptive statistics**: plot the data, numerical summary
3. **Tools and Concepts**: random variable, probability distributions
4. **Inferential statistics**: one sample/two samples test, ANOVA, RCBD, Correlation and (linear) regression

Temperature CO_2 data revisited



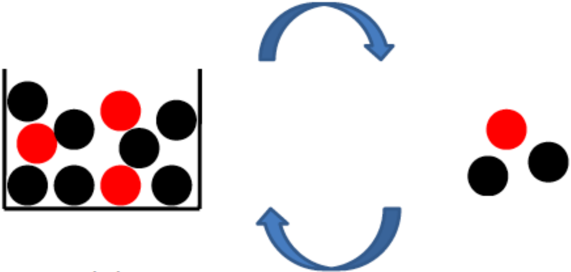
Temperature change (light blue) and carbon dioxide change (dark blue) measured from the EPICA Dome C ice core in Antarctica (Jouzel et al. 2007; Lüthi et al. 2008).

- Stating the problem
- Gathering data
- Summarizing the data
- Analyzing the data
- Reporting and interpreting the results

Probability vs. Statistics

Probability:

What is the probability to get 1 red and 2 black balls?



Statistics:

What percentage of balls in the box are red?

Figure: Taken from JHU Statistical Computing by Hongkai Ji

Definitions

- Statistics is the science of collecting, analyzing, presenting, and interpreting data

Definitions

- Statistics is the science of collecting, analyzing, presenting, and interpreting data
- Data set: all the data collected in a particular study

Definitions

- Statistics is the science of collecting, analyzing, presenting, and interpreting data
- Data set: all the data collected in a particular study
- Elements are the individual entities of a data set

Definitions

- Statistics is the science of collecting, analyzing, presenting, and interpreting data
- Data set: all the data collected in a particular study
- Elements are the individual entities of a data set
- A variable is a characteristic of interest for the elements

Definitions

- Statistics is the science of collecting, analyzing, presenting, and interpreting data
- Data set: all the data collected in a particular study
- Elements are the individual entities of a data set
- A variable is a characteristic of interest for the elements
- An observation is the set of measurements obtained for a particular element

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.

Types of variables

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

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

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

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

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

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

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

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 elements 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 cont'd

Solution.

1. 4 elements in total
2. 4 variables in this data set. They are Grade, Major, Credit hours, and GPA
3. Grade: qualitative (ordinal); Major: qualitative (nominal); GPA: quantitative (interval); Credit hours: quantitative (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

Example cont'd

Solution.

1. qualitative (nominal)
2. quantitative (ratio) or qualitative (ordinal)
3. qualitative (ordinal)
4. qualitative (ordinal)
5. quantitative (ratio)

Example

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

1. Current GPAs of Purdue Statistics Graduate Students
2. GPA of Sanvesh during his time at Purdue
3. Value of Gordan Gecko's portfolio over the previous 3 years
4. Value of all portfolio's at Charles Schwaab in January 2008
5. Total salary of the LA Lakers throughout the 1990s
6. Salaries of all NBA teams in 1994.

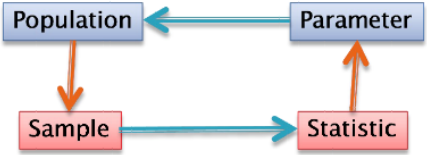
Example 63 cont'd

Solution.

1. cross-sectional
2. time series
3. time series
4. cross-sectional
5. time series
6. cross-sectional

Population vs. Sample

- The term “population" is used in Statistics to represent all possible outcomes that are of interest in a particular study
- The term “sample" refers to a portion of the population that is representative of the population
- We use parameters to describe the population
- We use statistics to describe the sample with respect to the population



Statistics provides a way to make inferences of the population by using sample data

Numerical Summaries of data

- Mean: the average/expected value of a set of numbers

Numerical Summaries of data

- Mean: the average/expected value of a set of numbers
  - Population mean: μ_x

Numerical Summaries of data

- Mean: the average/expected value of a set of numbers
  - Population mean: μ_x
  - Sample mean: $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$

Numerical Summaries of data

- Mean: the average/expected value of a set of numbers
  - Population mean: μ_x
  - Sample mean: $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
- Variance: measures how far a set of numbers is spread out

Numerical Summaries of data

- Mean: the average/expected value of a set of numbers
  - Population mean: μ_x
  - Sample mean: $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
- Variance: measures how far a set of numbers is spread out
  - Population variance: $\sigma_x^2 = \frac{\sum_{i=1}^{N}(x_i - \mu_x)^2}{N}$

Numerical Summaries of data

- Mean: the average/expected value of a set of numbers
  - Population mean: μ_x
  - Sample mean: $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
- Variance: measures how far a set of numbers is spread out
  - Population variance: $\sigma_x^2 = \frac{\sum_{i=1}^{N}(x_i - \mu_x)^2}{N}$
  - Sample variance: $s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$

Numerical Summaries of data

- Mean: the average/expected value of a set of numbers
  - Population mean: μ_x
  - Sample mean: $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
- Variance: measures how far a set of numbers is spread out
  - Population variance: $\sigma_x^2 = \frac{\sum_{i=1}^{N}(x_i - \mu_x)^2}{N}$
  - Sample variance: $s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$
- Mode: the value that appears most often in a set of numbers

Numerical Summaries of data

- Mean: the average/expected value of a set of numbers
  - Population mean: μ_x
  - Sample mean: $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
- Variance: measures how far a set of numbers is spread out
  - Population variance: $\sigma_x^2 = \frac{\sum_{i=1}^{N} (x_i - \mu_x)^2}{N}$
  - Sample variance: $s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$
- Mode: the value that appears most often in a set of numbers
- Range: the largest value $-$ the smallest value in a set of numbers

Example 59

Suppose we have the data set 1, 2, 3, 4, and 5. Find the mean of the data. Also compute variance in 2 ways (one assuming that this is a sample, the other assuming that this represents the entirety of the population)

Solution.

- Mean: $\bar{x} = \frac{1+2+3+4+5}{5} = 3$
- Sample variance: $s^2 = \frac{\sum_{i=1}^{5}(x_i-3)^2}{5-1} = \frac{10}{4} = 2.5$
- Population variance: $\sigma^2 = \frac{\sum_{i=1}^{5}(x_i-3)^2}{5} = \frac{10}{5} = 2$

Numerical Summaries: Percentiles

Numerical Summaries: Percentiles

- Percentile: The p_{th} percentile is a value of the data set such that at least p% of the data set is less than or equal to this value

Numerical Summaries: Percentiles

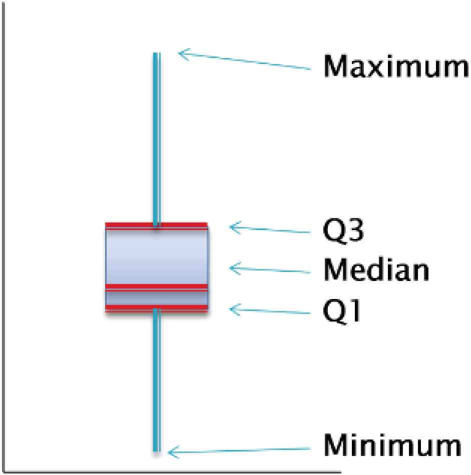
- Percentile: The p_{th} percentile is a value of the data set such that at least p% of the data set is less than or equal to this value
- Calculation of Percentiles using the indexing method:
  1. Sort the set of numbers in an increasing order
  2. For p_{th} percentile, compute $i = \frac{np}{100}$ where n is the sample size
  3. If i is an integer then p_{th} percentile is the average of i_{th} value and $(i+1)_{th}$ value, otherwise take the $(i+1)_{th}$ value

Numerical Summaries: Percentiles

- Percentile: The p_{th} percentile is a value of the data set such that at least $p\%$ of the data set is less than or equal to this value
- Calculation of Percentiles using the indexing method:
  1. Sort the set of numbers in an increasing order
  2. For p_{th} percentile, compute $i = \frac{np}{100}$ where n is the sample size
  3. If i is an integer then p_{th} percentile is the average of i_{th} value and $(i+1)_{th}$ value, otherwise take the $(i+1)_{th}$ value
- Quartiles:
  1. $Q1$: first quartile
  2. M or $Q2$: median or second quartile
  3. $Q3$: third quartile
  4. Interquartile range or IQR: $Q3 - Q1$

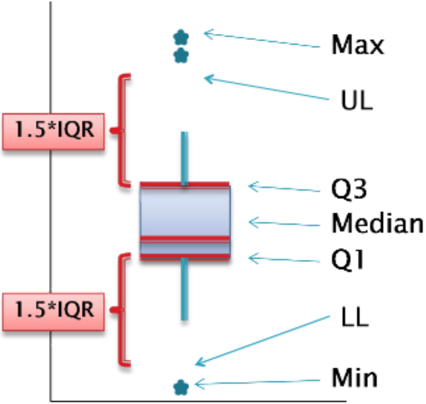
Regular Boxplots

A boxplots is a visual representation of the 5 number summary: *Min*, *Q*1, *Median*, *Q*3, *Max*



Modified Boxplots

- The modified boxplot will highlight if there are outliers
- Outliers: an outlier is a number that is far from other numbers
- LL (Lower Limit): $LL = Q1 - 1.5IQR$
  UL (Upper Limit): $UL = Q3 + 1.5IQR$
- A number is considered as an outlier if it is $\leq LL$ or $\geq UL$



Example 60

Hank Aaron hit an astounding 755 home runs in his career. His career spanned from 1954 through 1976. In those 23 seasons he hit 13, 27, 26, 44, 30, 39, 40, 34, 45, 44, 24, 32, 44, 39, 29, 44, 38, 47, 34, 40, 20, 12, 10 home runs

1. What is the mode of the data set?
2. What is the range of the data set?
3. Create both a regular and a modified boxplot for the number of home runs that Hank Aaron hit in a season
4. Find the 61st percentile

Example 60 cont'd

Solution.

First, we sort the number of home run in an increasing order: 10, 12, 13, 20, 24 ,26, 27, 29, 30 ,32, 34, 34, 38, 39, 39, 40, 40, 44, 44, 44, 44, 45, 47

1. 44
2. range=max-min=$47 - 10 = 37$
3. The index value for each quartile is $5.75, 11.5, 17.25$ respectively, so we have $Q1 = 26$, $Q2 = 34$, $Q3 = 44$
   The min and max are 10 and 47
   For Modified boxplot we need LL and UL.
   $IQR = Q3 - Q1 = 18 \Rightarrow LL = 26 - 1.5(18) = -1, \ UL = 44 + 1.5(18) = 71$
4. $i = \frac{np}{100} = \frac{(23)(61)}{100} = 14.03 \Rightarrow$ the 61st percentile is 39

Frame Title

Frame Title

Frame Title

Frame Title

Frame Title

Frame Title

Frame Title