Data Analytics

Yong Zheng

Illinois Institute of Technology
Chicago, IL, 60616, USA

School of Applied Technology
ILLINOIS INSTITUTE OF TECHNOLOGY

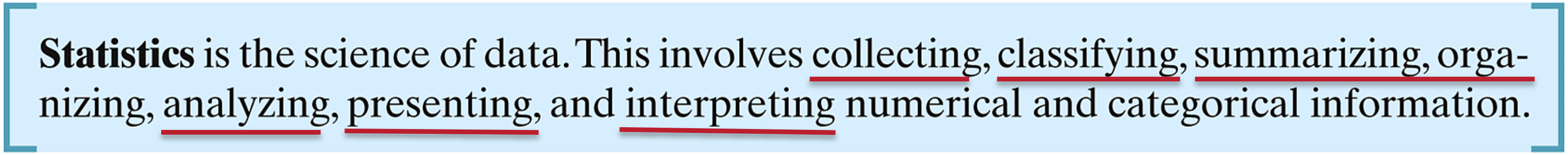
Schedule

- Statistical Applications
- Case Study 1: Student Grade and Behaviors
- Get to Know Data: Data Types
- Data: Population and Samples
- Descriptive Statistics
  - For Categorical Data
  - For Numerical Data

Schedule

- Statistical Applications
- Case Study 1: Student Grade and Behaviors
- Get to Know Data: Data Types
- Data: Population and Samples
- Descriptive Statistics
  - For Categorical Data
  - For Numerical Data

Statistics

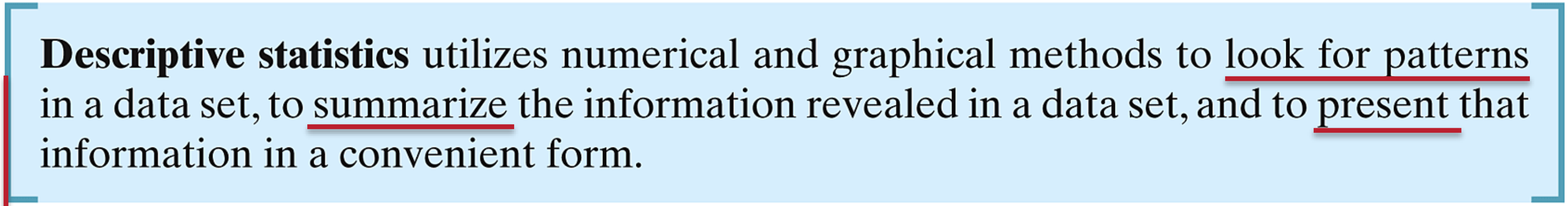


Statistics is the science of data. This involves collecting, classifying, summarizing, organizing, analyzing, presenting, and interpreting numerical and categorical information.

- Data and Data Visualization
- Probability Distributions
- Descriptive Statistics and Statistical Inference
- Predictive Analytics and Models
- Predictive Models for Data Mining
- Statistical Fundamentals for Machine Learning
- Statistical Interpretations

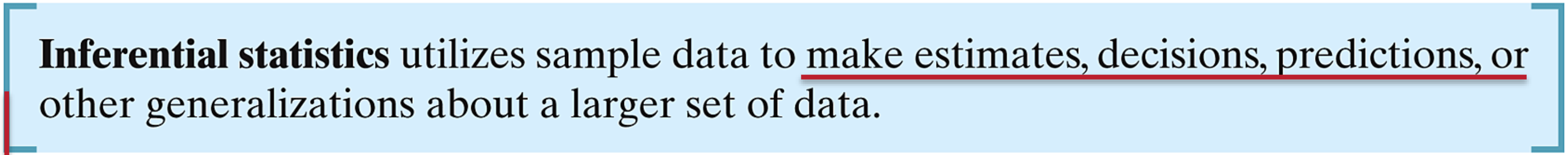
Statistical Applications

Types of Statistical Applications



Descriptive statistics utilizes numerical and graphical methods to look for patterns in a data set, to summarize the information revealed in a data set, and to present that information in a convenient form.

It is used to understand and/or visualize our data at the beginning.



Inferential statistics utilizes sample data to make estimates, decisions, predictions, or other generalizations about a larger set of data.

It is used to

- ❑ Estimate a variable → Estimating population by samples
- ❑ Validate hypothesis or assumptions → Hypothesis Testing
- ❑ Make predictions → Predictive Models or Analysis

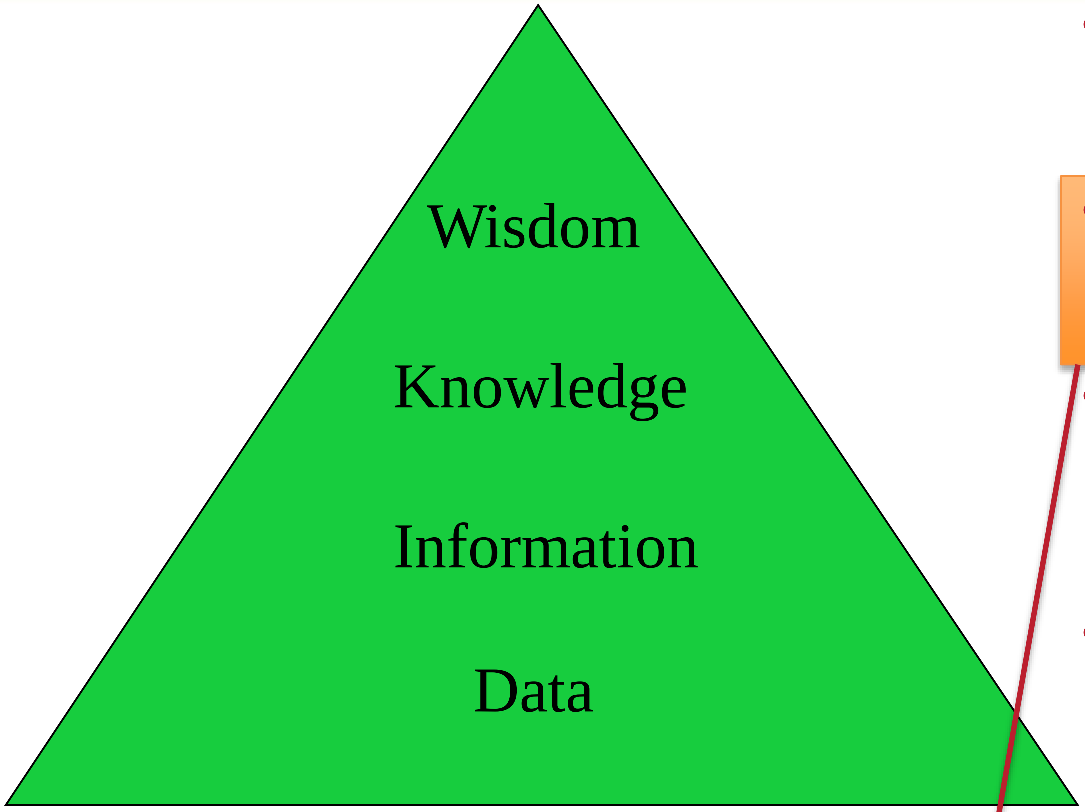
Schedule

- Statistical Applications
- Case Study 1: Student Grade and Behaviors
- Get to Know Data: Data Types
- Data: Population and Samples
- Descriptive Statistics
  - For Categorical Data
  - For Numerical Data

Case Study 1: Student Grade and Behaviors

- **Data Files**
  - Three csv files for small, regular and large data size
  - We usually use the regular size for example
  - You can use small and large size for self-practice
- **Descriptions**
  - Each row represent a student and his or her grades
  - Student Info: ID, Nationality, Gender, Age, Program
  - Student Behaviors: # of hours on activities / week
  - Student Grade: Exam score and Letter grade

Getting to Know Data



The Information Hierarchy

- Data
  - The raw material of information
- Information
  - Data organized and presented by someone
- Knowledge
  - Information read, heard or seen and understood and integrated
- Wisdom
  - Distilled knowledge and understanding which can lead to decisions

Collect your data and prepare your data in tables (such as csv)

Schedule

- Statistical Applications
- Case Study 1: Student Grade and Behaviors
- Get to Know Data: Data Types
- Data: Population and Samples
- Descriptive Statistics
  - For Categorical Data
  - For Numerical Data

Get to Know Data

Once you got a data set, you should do:

- Step 1. Know the data size and format
  - Format: csv/txt/sql/pdf/videos/music, and so on
  - Data size: how many rows and columns
- Step 2. Understand the columns/variables
  - What do they mean
  - What are the data types and values in the variables
  - Descriptive Statistics
- Step 3. Any possible concerns or questions?

Get to Know Data

Once you got a data set, you should do:

- Step 1. Know the data size and format
  - Format: csv/txt/sql/pdf/videos/music, and so on
    - Let’s look at the data in regular size. It is in csv file
  - Data size: how many rows and columns
    - 600 rows, note that the first row is header
    - 12 columns

Get to Know Data

Once you got a data set, you should do:

- Step 2. Understand the columns/variables
  - What do they mean
    - Student info, behaviors, and grades
  - What are the data types and values in the variables
  - Descriptive Statistics

Getting to Know Data

- **Types of the Data**
  - **Qualitative (Categorical/Nominal)**
    - Nominal
    - Binary
    - Ordinal
  - **Quantitative (Numerical)**
    - Discrete
    - Continuous

Qualitative/Categorical/Nominal Data

- **Nominal data** are categories, states, or “names of things”
  - *color* = {*auburn*, *black*, *blond*, *brown*, *grey*, *red*, *white*}
- **Special type: Binary variable**
  - Nominal attribute with only 2 states (0 and 1)
  - Symmetric binary: both outcomes equally important, e.g., gender
  - Asymmetric binary: outcomes not equally important.
    - e.g., medical test (positive vs. negative)
- **Special Type: Ordinal variable**
  - Values have a meaningful order (ranking)
  - *Size* = {*small*, *medium*, *large*}, university rankings

Quantitative/Numeric Data

In general, they are numbers.

- **Discrete Data**
  - **They are counted:** Has only a finite or countably infinite set of values
  - Examples: zip codes, counts, etc
  - Note: binary attributes are a special case of discrete attributes

- **Continuous Data**
  - **They are measured:** Has real numbers as attribute values
  - Examples: temperature, height, or weight.
  - Continuous attributes are typically represented as floating-point variables.

Quantitative/Numeric Data

However, not all of the numbers are numerical data!!!!

ID	Nationality	Gender
1	China	1
2	France	0
3	France	0
4	India	1
5	India	1
6	India	1

We code nationality

by using numbers

ID	Nationality	Gender
1	0	1
2	1	0
3	1	0
4	2	1
5	2	1
6	2	1

Quantitative/Numeric Data

However, not all of the numbers are numerical data!!!!

ID	Nationality	Gender
1	China	1
2	France	0
3	France	0
4	India	1
5	India	1
6	India	1

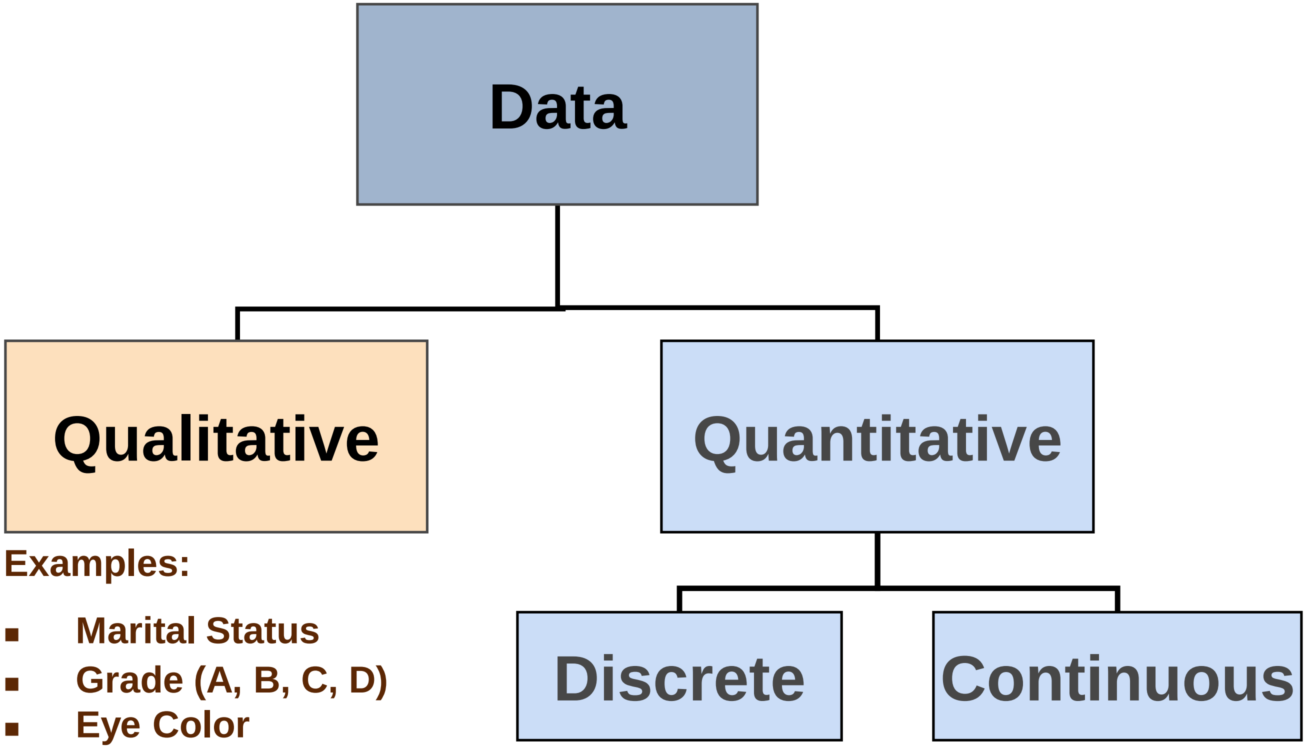
We code nationality

by using numbers

ID	Nationality	Gender
1	0	1
2	1	0
3	1	0
4	2	1
5	2	1
6	2	1

How to determine a variable is categorical or numerical if the values in the variable are numbers? ➔ it depends on that fact that whether you can explain the difference in the numbers

Types of the Data



Examples:

- Marital Status
- Grade (A, B, C, D)
- Eye Color

Two special data types

- Binary
- Ordinal

Examples:

- Number of Children
- Defects per hour

(Counted items)

Examples:

- Weight
- Voltage

(Measured characteristics)

Practice: What are the data types?

- Letter grades
- Number of students in the class
- Student ID (A Number@IIT)
- Zip code
- HIV test result
- Length

Practice: What are the data types?

- Letter grades → Categorical: Ordinal
- Number of students → Numerical: Discrete
- Student ID → Categorical: Nominal
- Zip code → Numerical: Discrete
- HIV test result → Categorical: Asymmetric Binary
- Length → Numerical: Continuous

Case Study 1

- ID
- Nationality
- Gender
- Age
- Degree
- Hours on readings/assignments/Games/Internet
- Exam score
- Grade
- Letter Grade

Get to Know Data

Once you got a data set, you should do:

- Step 2. Understand the columns/variables
  - What do they mean
    - Student info, behaviors, and grades
  - What are the data types and values in the variables
  - Descriptive Statistics

Schedule

- Statistical Applications
- Case Study 1: Student Grade and Behaviors
- Get to Know Data: Data Types
- Data: Population and Samples
- Descriptive Statistics
  - For Categorical Data
  - For Numerical Data

Where can we obtain the data

Data Is Collected From Either A Population or A Sample

POPULATION

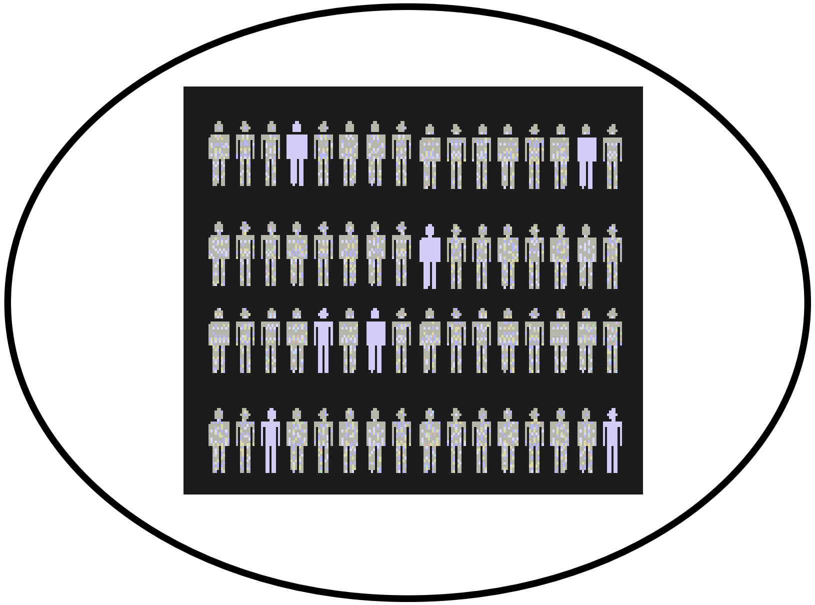
A **population** consists of all the items or individuals about which you want to draw a conclusion. The population is the “large group”

SAMPLE

A **sample** is the portion of a population selected for analysis. The sample is the “small group”

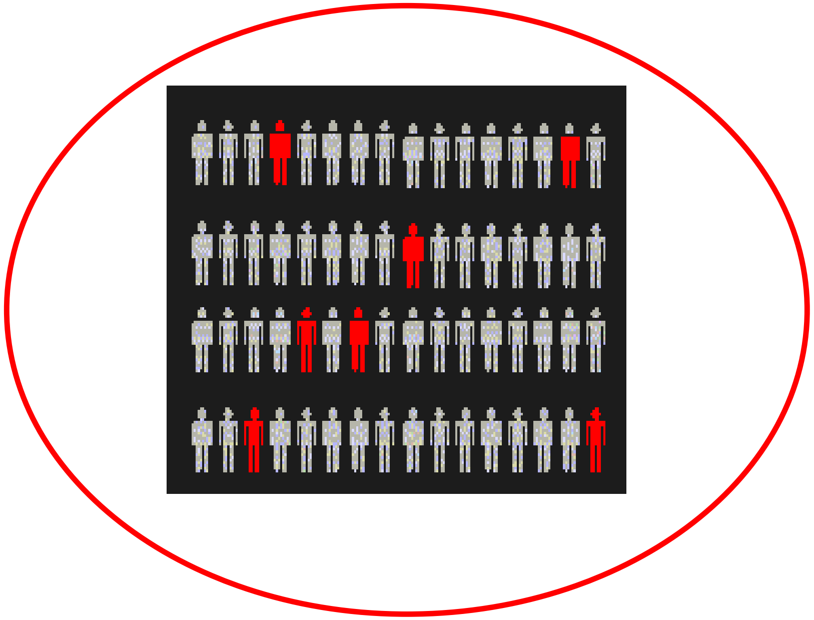
Population vs. Sample

Population (All)



All the items or individuals about which you want to draw conclusion(s)

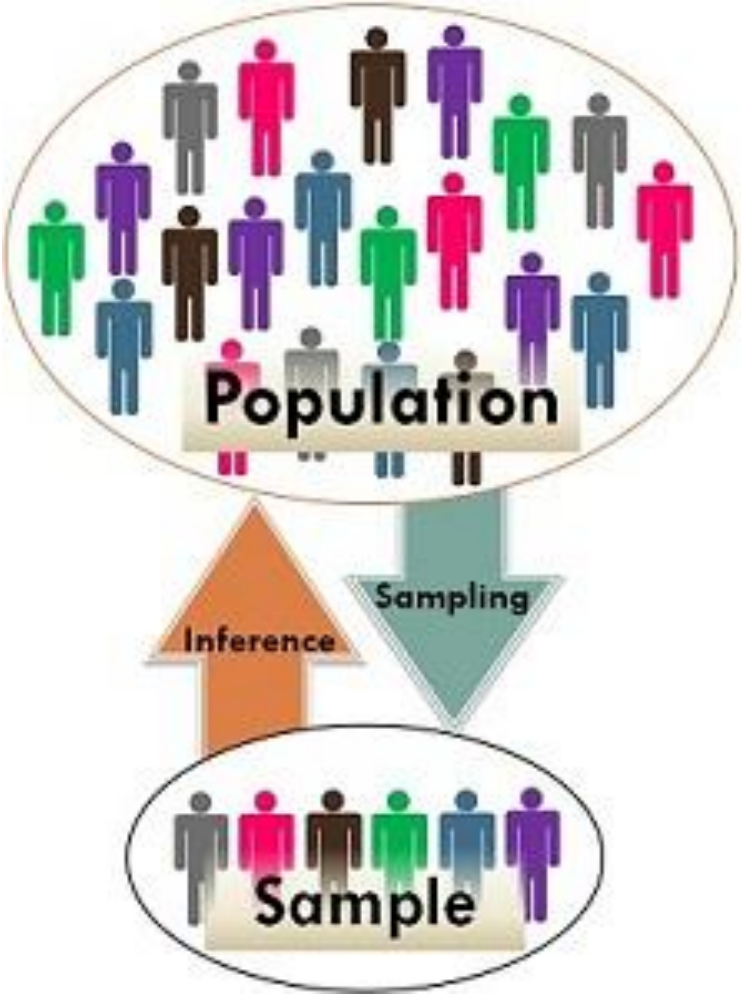
Sample (Red ones)



A portion of the population of items or individuals

Population vs. Sample

We are not able to calculate population statistics directly. Instead, we get a sample from the population, and use sample statistics to estimate population statistics.



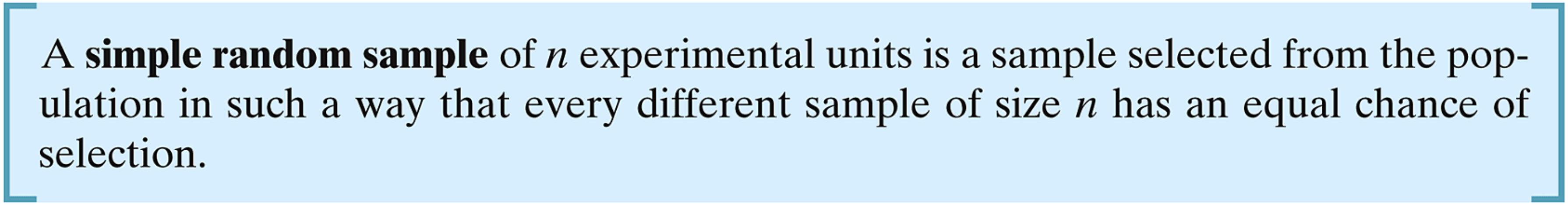
Example

According to a report, the average age of viewers of the major network's TV news programming is 50 years old. Suppose a cable network manager hypothesizes that the average age of cable TV news viewer is less than 50. To test the hypothesis, she selects 500 cable TV news viewers and determine the age of each.

Question: what is population, what is the sample? And what inference she'd like to make?

Random Sampling

- Only a representative sample is able to well-estimate the popular statistics.
- How to extract a representative sample? Random sampling is usually used as a reliable way.



A **simple random sample** of *n* experimental units is a sample selected from the population in such a way that every different sample of size *n* has an equal chance of selection.

Note: without explicit information, the data you have or get is usually sample data, NOT the population data

Schedule

- Statistical Applications
- Case Study 1: Student Grade and Behaviors
- Get to Know Data: Data Types
- Data: Population and Samples
- Descriptive Statistics
  - For Categorical Data
  - For Numerical Data

Describe Qualitative Data

- Qualitative data are categorical or discrete values

ID	Nationality	Gender	Age	Degree
1	China	1	21	BS
2	France	0	26	PHD
3	France	0	20	PHD
4	India	1	18	MS
5	India	1	18	MS
6	India	1	18	BS
7	India	1	18	BS
8	India	1	20	BS
9	France	0	19	BS
10	France	0	20	BS
11	India	0	19	BS

Describe Qualitative Data

- Qualitative data are categorical or discrete values

ID	Nationality	Gender	Age	Degree
1	China	1	21	BS
2	France	0	26	PHD
3	France	0	20	PHD
4	India	1	18	MS
5	India	1	18	MS
6	India	1	18	BS
7	India	1	18	BS

Describe Qualitative Data

- Describe qualitative data Numerically
  - By class frequency
  - By class relative frequency
- Describe qualitative data by visualizations
  - By bar graph
  - By pie chart

Describe Qualitative Data

- Describe qualitative data Numerically

1). By class frequency (cf)

cf = # of observations associated with a class

ID	Nationality	Gender	Age	Degree
1	China	1	21	BS
2	France	0	26	PHD
3	France	0	20	PHD
4	India	1	18	MS
5	India	1	18	MS
6	India	1	18	BS
7	India	1	18	BS

cf (BS) = 3

cf (MS) = 2

cf (PHD) = 2

Describe Qualitative Data

- Describe qualitative data Numerically

2). By class relative frequency (crf)

crf = cf/n, n = total number of observations

ID	Nationality	Gender	Age	Degree
1	China	1	21	BS
2	France	0	26	PHD
3	France	0	20	PHD
4	India	1	18	MS
5	India	1	18	MS
6	India	1	18	BS
7	India	1	18	BS

cf (BS) = 3
crf (BS) = 3/7

cf (MS) = 2
crf (MS) = 2/7

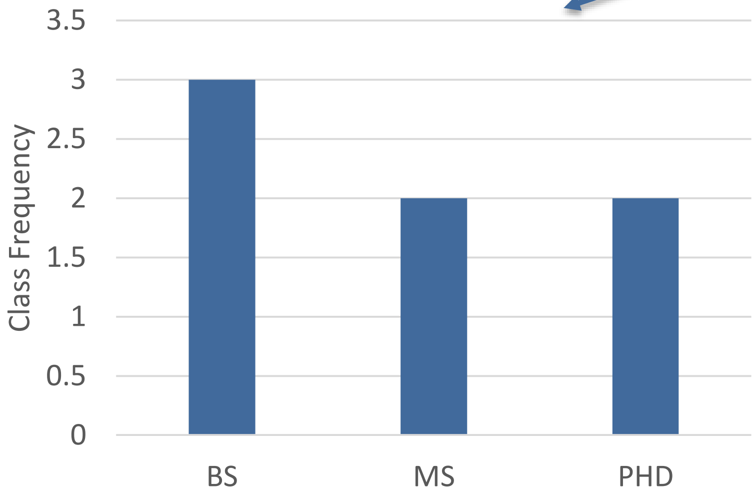
cf (PHD) = 2
crf (PHD) = 2/7

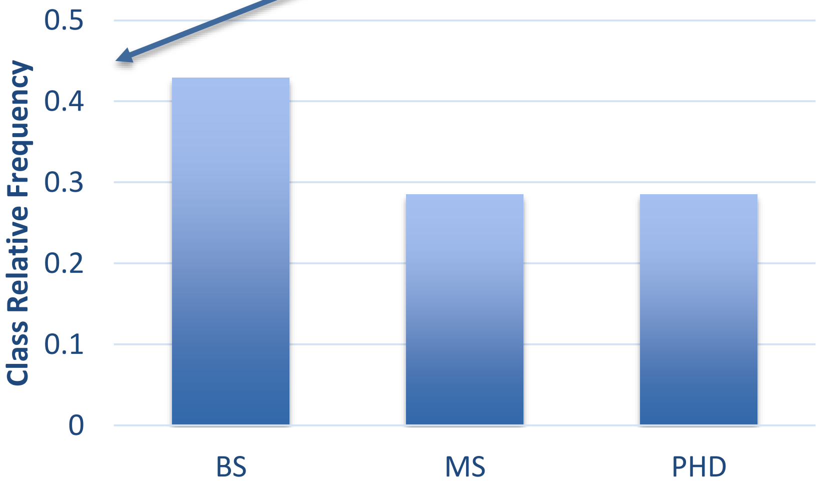
Describe Qualitative Data

- Describe qualitative data by visualizations

1). By bar graph

Y axis could be class frequency or class relative frequency



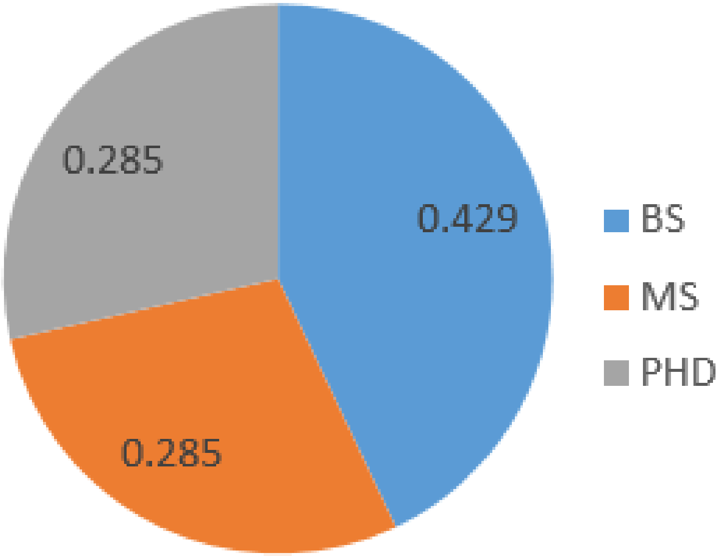


Describe Qualitative Data

- Describe qualitative data by visualizations

2). By pie chart

It is usually used to
depict class relative frequency



Summary: Describe Qualitative Data

- Describe qualitative data Numerically
  - By class frequency
  - By class relative frequency
- Describe qualitative data by visualizations
  - By bar graph
  - By pie chart

After-class practice: Describe categorical variables in our case study 1 data

Describe Quantitative Data

- Quantitative Data are numerical

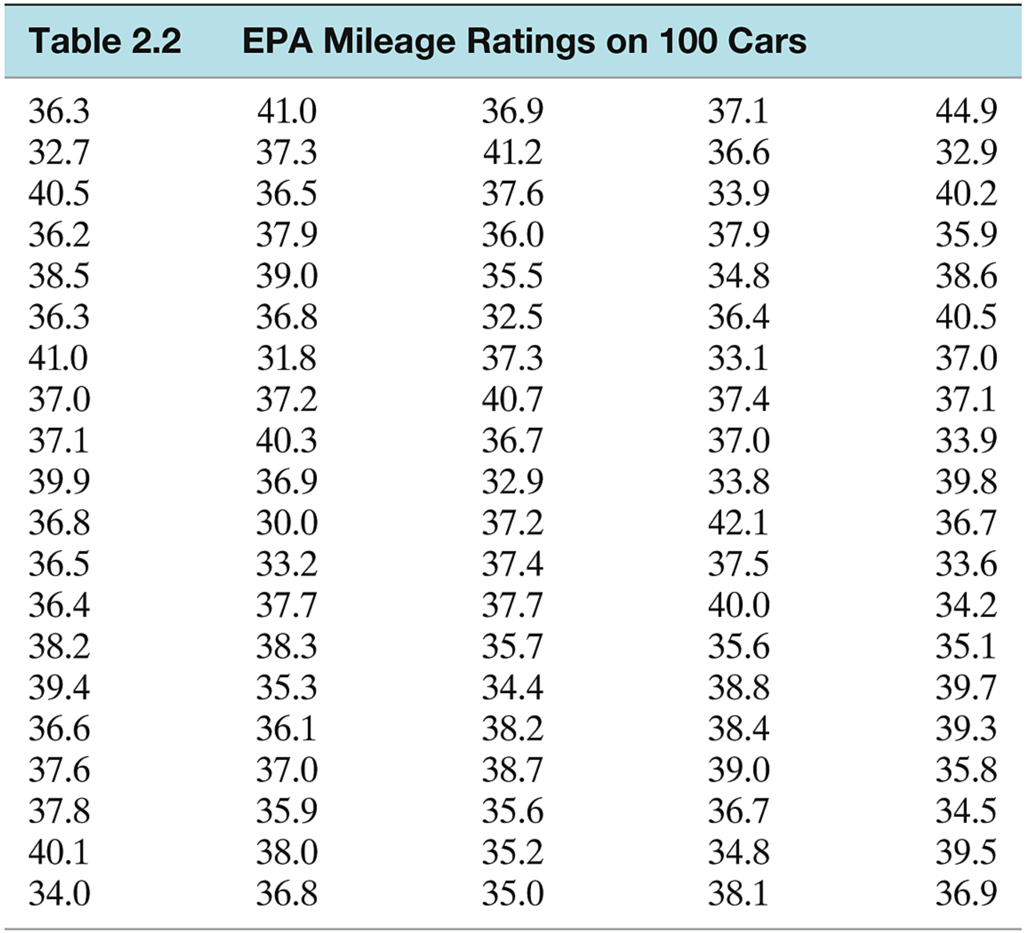


Table 2.2 EPA Mileage Ratings on 100 Cars

36.3	41.0	36.9	37.1	44.9
32.7	37.3	41.2	36.6	32.9
40.5	36.5	37.6	33.9	40.2
36.2	37.9	36.0	37.9	35.9
38.5	39.0	35.5	34.8	38.6
36.3	36.8	32.5	36.4	40.5
41.0	31.8	37.3	33.1	37.0
37.0	37.2	40.7	37.4	37.1
37.1	40.3	36.7	37.0	33.9
39.9	36.9	32.9	33.8	39.8
36.8	30.0	37.2	42.1	36.7
36.5	33.2	37.4	37.5	33.6
36.4	37.7	37.7	40.0	34.2
38.2	38.3	35.7	35.6	35.1
39.4	35.3	34.4	38.8	39.7
36.6	36.1	38.2	38.4	39.3
37.6	37.0	38.7	39.0	35.8
37.8	35.9	35.6	36.7	34.5
40.1	38.0	35.2	34.8	39.5
34.0	36.8	35.0	38.1	36.9

Describe Quantitative Data

- Describe quantitative data Numerically
  - By range, min, max, mean, median, mode
  - By variance, standard deviation
  - By q1, q2, q3
- Describe quantitative data by visualizations
  - By stem-and-leaf
  - By histogram
  - By box plot
  - By probability distribution

Describe Quantitative Data

- Describe quantitative data Numerically
  - By range, min, max, mean, median, mode

Min = minimal value

Max = maximal value

Range = the difference between largest & smallest values

Mean = average value

Median = the value in the middle

Mode = the value that occurs most often

Describe Quantitative Data

- Describe quantitative data Numerically
  - By range, min, max, mean, median, mode

  Example-1: 3, 4, 3, 3, 1, 0, 0

  Example-2: 3, 4, 3, 2, 0, 0

  Example-3: 1, 2, 3, 4, 5

Describe Quantitative Data

- Describe quantitative data Numerically
  - By range, min, max, mean, median, mode

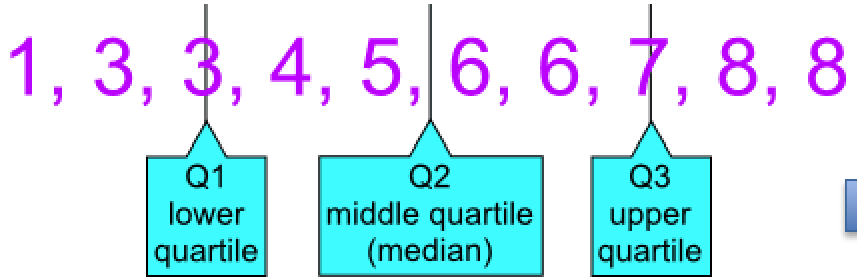
❑ If the # of observations is even, median is the average value of the most two centered numbers.

❑ If there are multiple values that occur the most often, all of them are modes.

❑ If each number appears for the same times, the mode is “none”.

Describe Quantitative Data

- Describe quantitative data Numerically
  - By q1, q2, q3



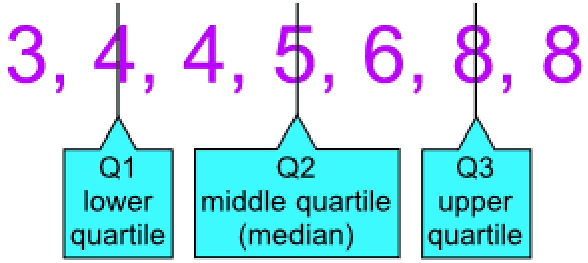
1, 3, 3, 4, 5, 6, 6, 7, 8, 8

Q1 lower quartile

Q2 middle quartile (median)

Q3 upper quartile

Size: Even Numbers



3, 4, 4, 5, 6, 8, 8

Q1 lower quartile

Q2 middle quartile (median)

Q3 upper quartile

Size: Odd Numbers

Describe Quantitative Data

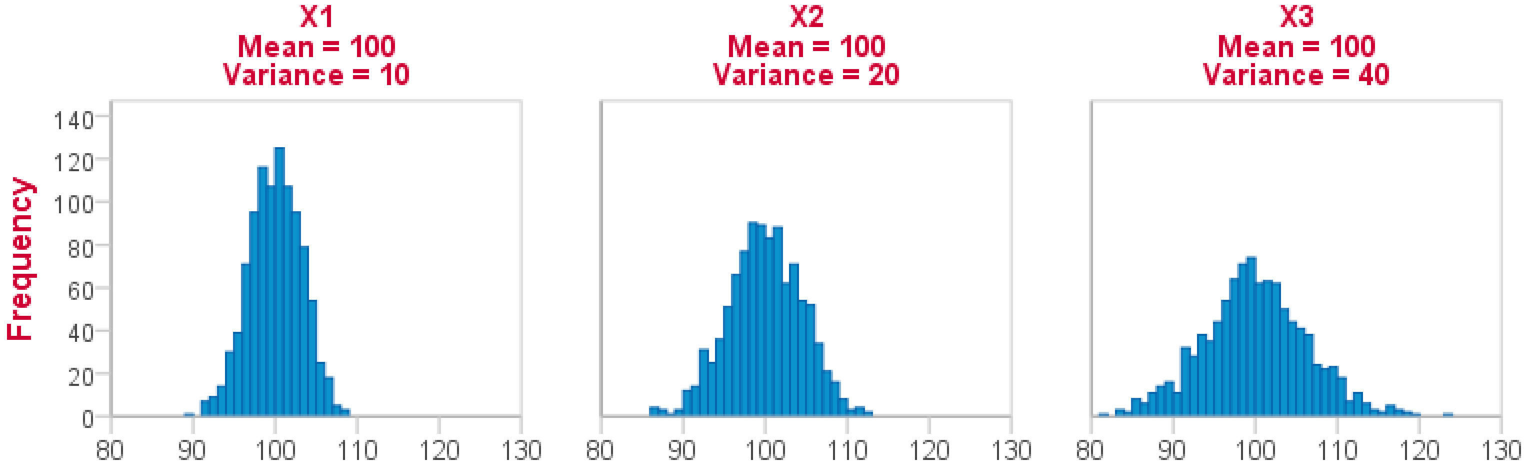
- Describe quantitative data Numerically
  - By q1, q2, q3
- Important notes
  - There are no unified ways to calculate q1 and q3
  - It seems that there are 19 methods
  - Excel and R may give you different results
  - For manually calculations, use my methods
  - By using R and Excel, just report the results
  - For more information about it, you can visit this webpage to collect more information: https://lagunita.stanford.edu/courses/DB/RD/SelfPaced/wiki/HRP258/example-r-classwork-solutions-using-r/calculating-inner-quartile-range-r/

Describe Quantitative Data

- Describe quantitative data Numerically
  – By variance and standard deviation

❑Variance and standard deviation are used to describe the variation of the data

❑Standard deviation = square root of the variance



Describe Quantitative Data

- Describe quantitative data Numerically
  - By variance and standard deviation

❑ We need to distinguish population statistics and sample statistics, when we use mean, variance, standard deviation to describe quantitative data.

❑ Recall that the sample statistics is just used to estimate the population statistics.

Describe Quantitative Data

- Population Statistics

μ = Population mean

σ^2 = Population variance

σ = Population standard deviation

$$\sigma^2 = E[(x_i - \mu)^2] = \frac{\sum_{i=1}^{n}(x_i - \mu)2}{n}$$

It will underestimate the value if the sample variance is divided by n.

Note: usually popular mean & var are unknown

- Sample Statistics

$\bar{x}$ = Sample mean

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Sample Variance:

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}$$

sample standard deviation

$$s = \sqrt{s^2}$$

Describe Quantitative Data

- Describe quantitative data Numerically
  - By variance and standard deviation

❑ Standard deviation (STD) is used more often than variance to describe the data variation.

❑ How to interpret STD?

STD is the deviations of measurement values from the mean value (sample or population mean)

Describe Quantitative Data

- Describe quantitative data Numerically
  - By variance and standard deviation

Example: 1, 2, 3, 4, 5

mean = ?

STD = ?

variance = ?

Does ¾ of measurements lie in the two STDs of mean?

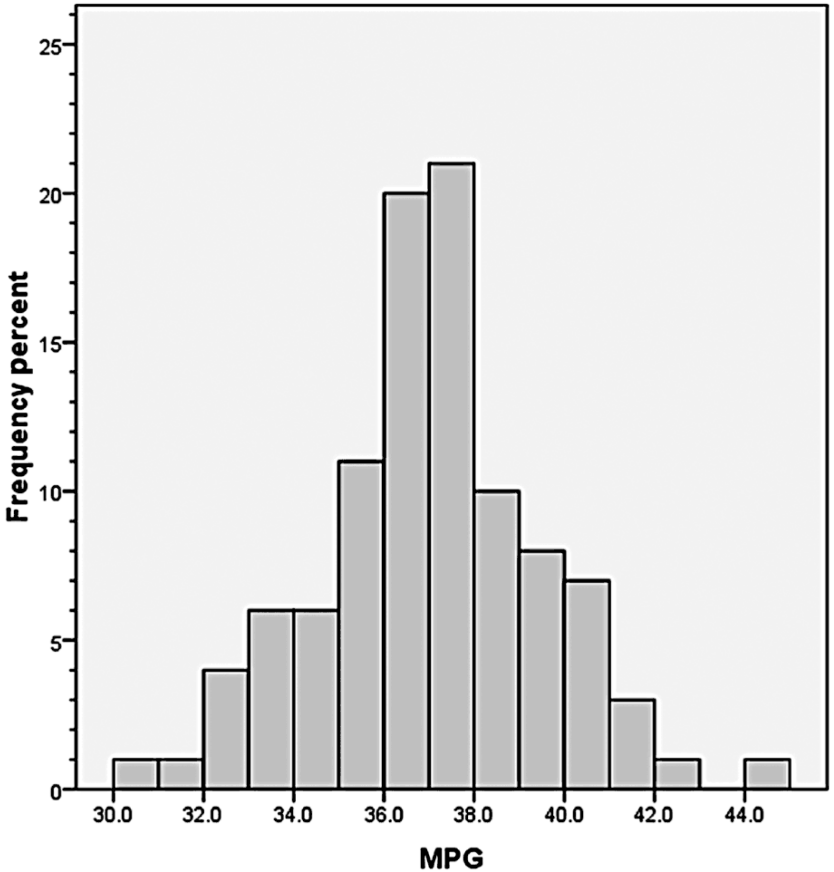
Note: By default the statistics in these questions are asking for sample statistics.

Describe Quantitative Data

- Describe quantitative data by visualizations
  - By stem-and-leaf [Optional]
  - By histogram
  - By box plot
  - By probability distribution

Describe Quantitative Data

- Describe quantitative data by visualizations
  - By histogram



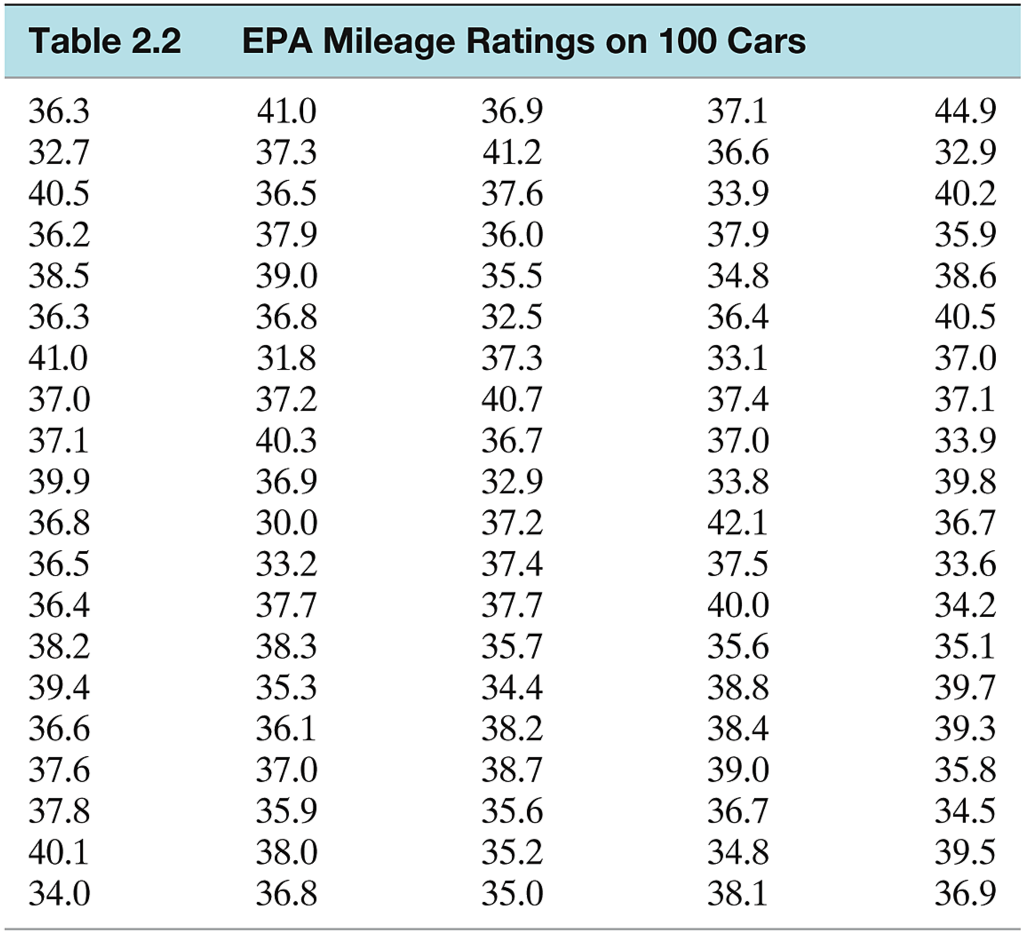


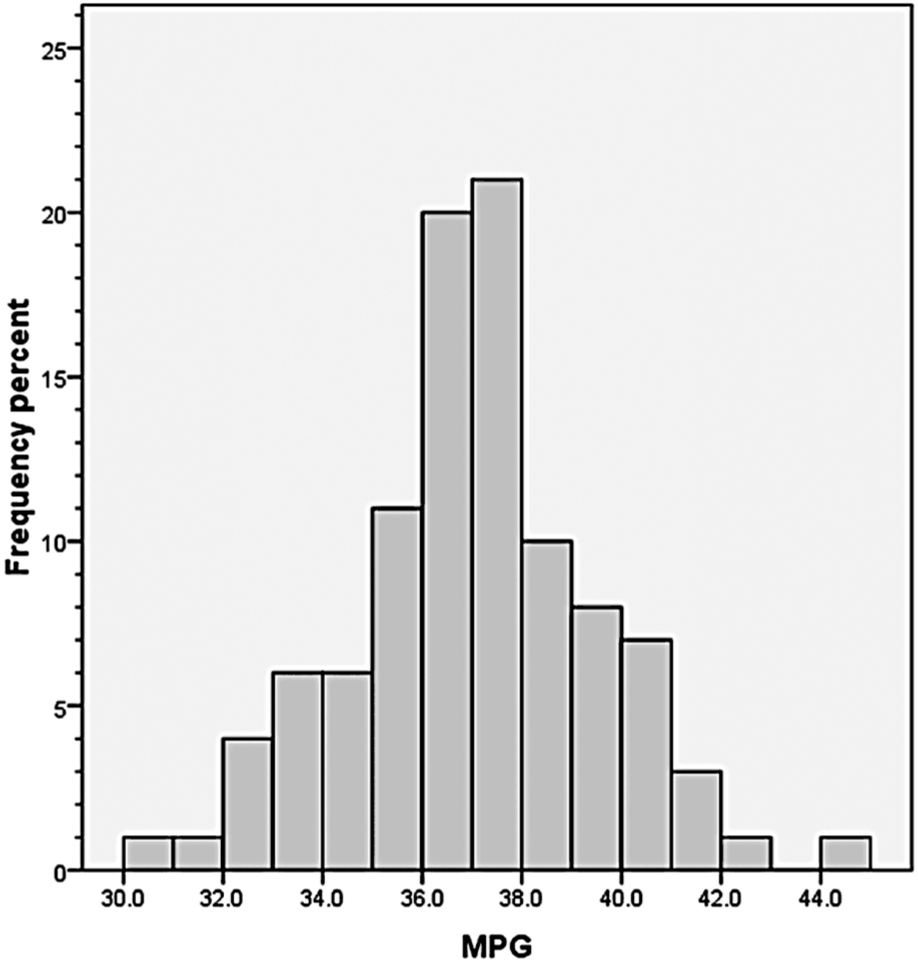
Table 2.2 EPA Mileage Ratings on 100 Cars

36.3	41.0	36.9	37.1	44.9
32.7	37.3	41.2	36.6	32.9
40.5	36.5	37.6	33.9	40.2
36.2	37.9	36.0	37.9	35.9
38.5	39.0	35.5	34.8	38.6
36.3	36.8	32.5	36.4	40.5
41.0	31.8	37.3	33.1	37.0
37.0	37.2	40.7	37.4	37.1
37.1	40.3	36.7	37.0	33.9
39.9	36.9	32.9	33.8	39.8
36.8	30.0	37.2	42.1	36.7
36.5	33.2	37.4	37.5	33.6
36.4	37.7	37.7	40.0	34.2
38.2	38.3	35.7	35.6	35.1
39.4	35.3	34.4	38.8	39.7
36.6	36.1	38.2	38.4	39.3
37.6	37.0	38.7	39.0	35.8
37.8	35.9	35.6	36.7	34.5
40.1	38.0	35.2	34.8	39.5
34.0	36.8	35.0	38.1	36.9

Describe Quantitative Data

- Describe quantitative data by visualizations
  – By histogram

It is similar to the bar graph used to describe categorical data.
Here, we present class frequency for a range of values, e.g., [30, 32]

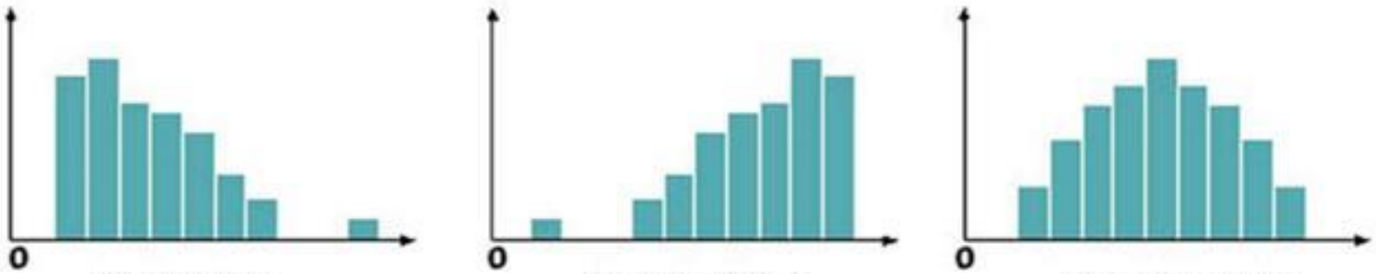


Describe Quantitative Data

- Describe quantitative data by visualizations
  - By histogram

  How to interpret histogram? (skewness and outlier)

Analyzing Shape:



Positive Skew

Data is skewed to the right. The long tail of the data is on the right side of the peak.

Negative Skew

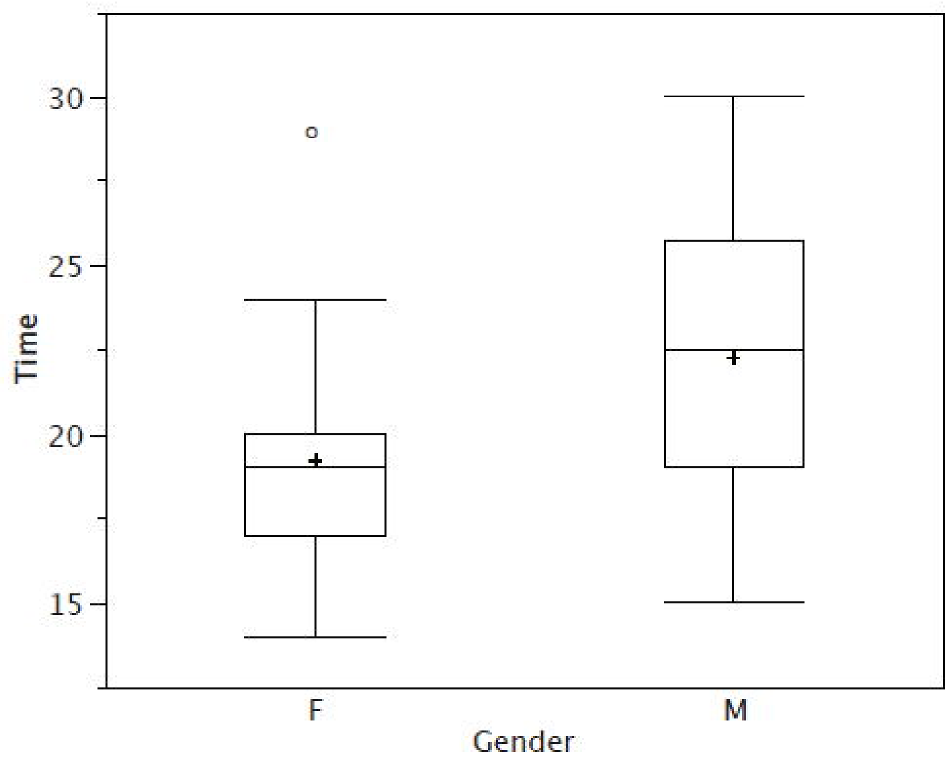
Data is skewed to the left. The long tail of the data is on the left side of the peak.

Normal Distribution

Data is not skewed to the right or left. The data is evenly distributed on both sides of the peak.

Describe Quantitative Data

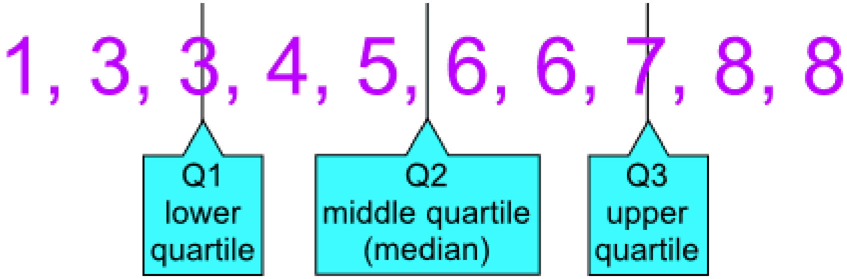
- Describe quantitative data by visualizations
  - By box plot



Describe Quantitative Data

- Describe quantitative data by visualizations
  - By box plot: Interpretations

1). Quartile

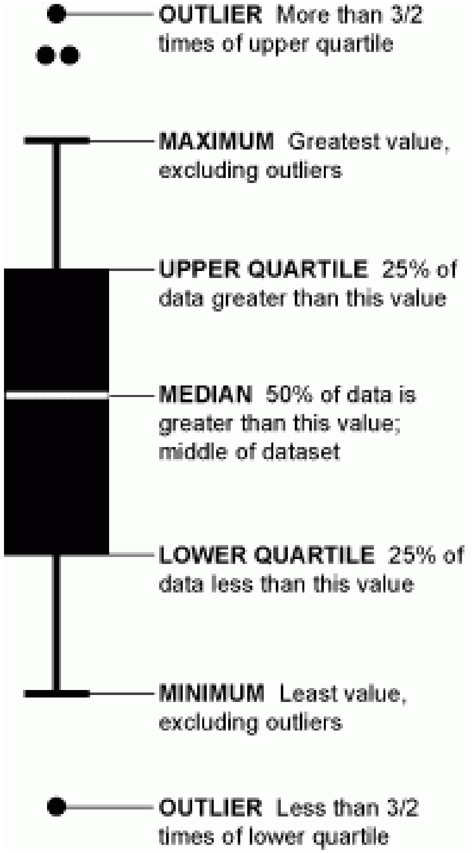


1, 3, 3, 4, 5, 6, 6, 7, 8, 8

Q1 lower quartile

Q2 middle quartile (median)

Q3 upper quartile

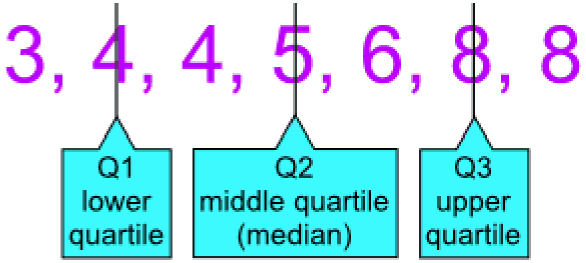


OUTLIER More than 3/2 times of upper quartile
MAXIMUM Greatest value, excluding outliers
UPPER QUARTILE 25% of data greater than this value
MEDIAN 50% of data is greater than this value; middle of dataset
LOWER QUARTILE 25% of data less than this value
MINIMUM Least value, excluding outliers
OUTLIER Less than 3/2 times of lower quartile

Describe Quantitative Data

- Describe quantitative data by visualizations
  - By box plot: Interpretations

1). Quartile

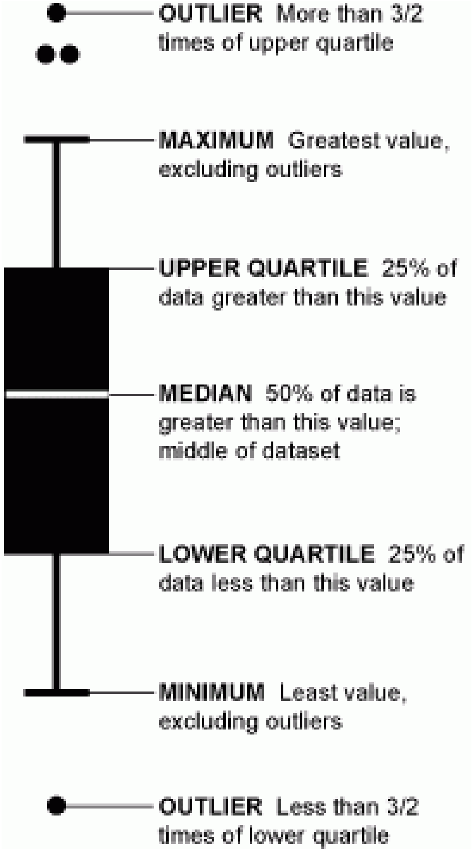


3, 4, 4, 5, 6, 8, 8

Q1 lower quartile

Q2 middle quartile (median)

Q3 upper quartile



OUTLIER More than 3/2 times of upper quartile

MAXIMUM Greatest value, excluding outliers

UPPER QUARTILE 25% of data greater than this value

MEDIAN 50% of data is greater than this value; middle of dataset

LOWER QUARTILE 25% of data less than this value

MINIMUM Least value, excluding outliers

OUTLIER Less than 3/2 times of lower quartile

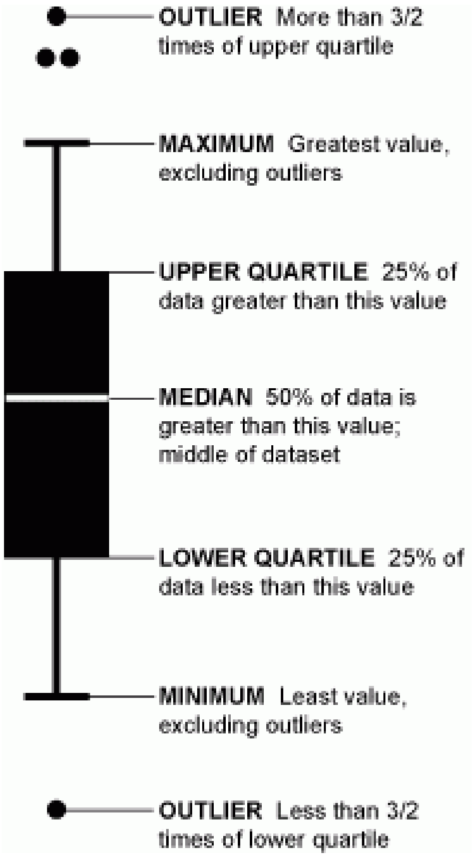
Describe Quantitative Data

- Describe quantitative data by visualizations
  - By box plot: Interpretations

2). Median

Median = 2nd quartile = q2

Note: we usually use either mean or median to represent a set of quantitative data



OUTLIER More than 3/2 times of upper quartile
MAXIMUM Greatest value, excluding outliers
UPPER QUARTILE 25% of data greater than this value
MEDIAN 50% of data is greater than this value; middle of dataset
LOWER QUARTILE 25% of data less than this value
MINIMUM Least value, excluding outliers
OUTLIER Less than 3/2 times of lower quartile

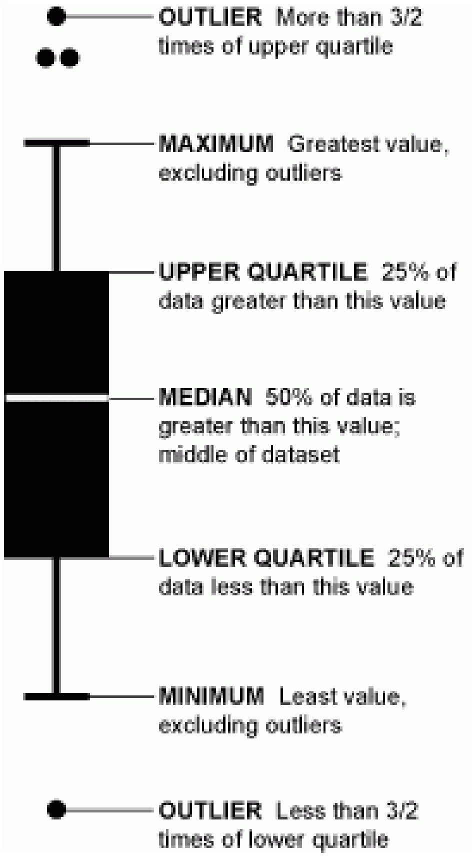
Describe Quantitative Data

- Describe quantitative data by visualizations
  - By box plot: Interpretations

3). Min, Max, Outlier

Here, the min and max values are the ones without considering outliers.

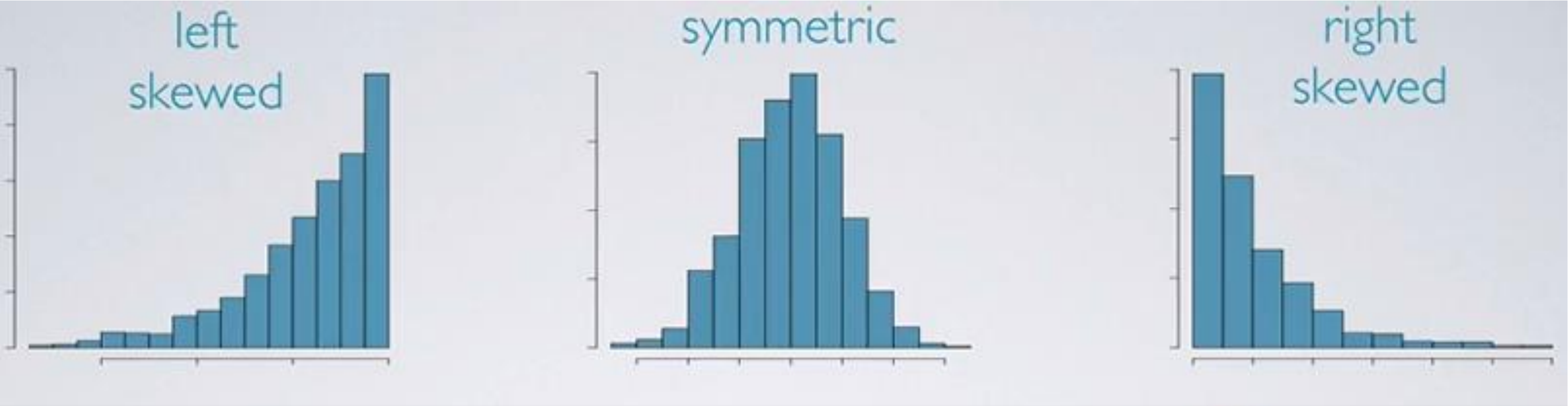
So, range ≠ Max-Min <u>from the box plot</u>!!!!!!!


OUTLIER More than 3/2 times of upper quartile
MAXIMUM Greatest value, excluding outliers
UPPER QUARTILE 25% of data greater than this value
MEDIAN 50% of data is greater than this value; middle of dataset
LOWER QUARTILE 25% of data less than this value
MINIMUM Least value, excluding outliers
OUTLIER Less than 3/2 times of lower quartile

Describe Quantitative Data

- Describe quantitative data by visualizations
  - By box plot: Interpretations

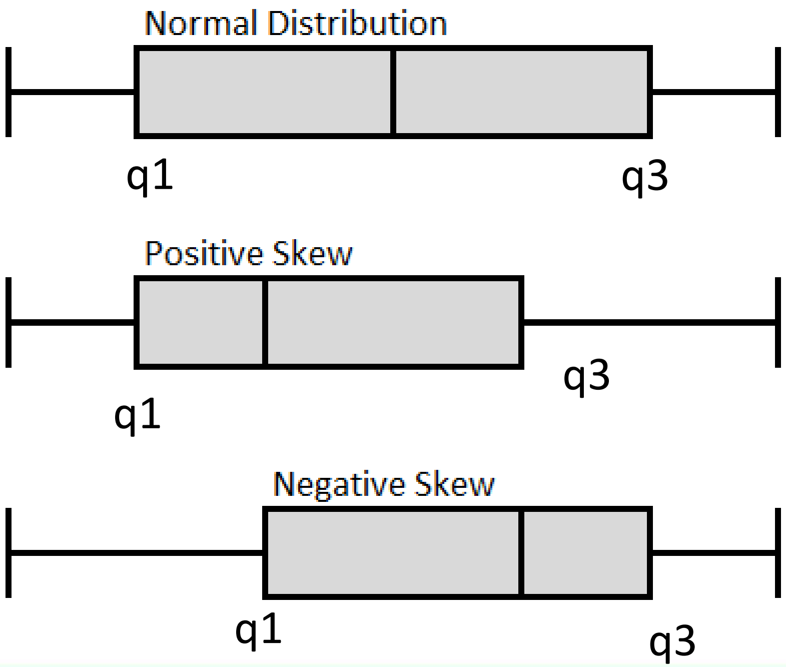
4). Skewness



Describe Quantitative Data

- Describe quantitative data by visualizations

How to make a decision about skewness from the box plot? We focus on the median and box only



Median is exactly in the middle

Median is closer to the q1

Median is closer to q3

Describe Quantitative Data

- Describe quantitative data by visualizations
  - By probability distribution

  We will introduce this topic in the next class.

After-class practice: Describe categorical variables in our case study 1 data