

# Lecture 2

## Exploratory Data Analysis I

Readings: IntroStat Chapters 2-3; OpenIntro Chapter 2

*STAT 8010 Statistical Methods I*

May 17, 2023

Whitney Huang  
Clemson University

Sampling Techniques

Summarizing  
Categorical Data

Summarizing  
Numerical Data

- 1 Sampling Techniques
- 2 Summarizing Categorical Data
- 3 Summarizing Numerical Data

- Stating the problem, identifying the variable(s) of interest, and gathering data
  - Types of variables and datasets
  - Observational vs. Experimental Studies
  - Methods of sampling
- Summarizing the data
- Analyzing the data
- Reporting and interpreting the results

- Stating the problem, identifying the variable(s) of interest, and gathering data
  - Types of variables and datasets
  - Observational vs. Experimental Studies
  - Sampling Techniques
- Summarizing the data
- Analyzing the data
- Reporting and interpreting the results

## Collecting Data: Statistical Sampling

Statistical sampling is the procedure to select a subset from a statistical **population** that is representative of the population. There are several types of sampling:

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

Simple random sample



## Collecting Data: Statistical Sampling

Statistical sampling is the procedure to select a subset from a statistical **population** that is representative of the population. There are several types of sampling:

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

Simple random sample



- **Stratified sample**: elements in the population are first divided into groups and a simple random sample is then taken from each group

Stratified sample



## Sampling cont'd

- **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

Cluster sample



## Sampling cont'd

- **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

Cluster sample



- **Systematic sampling**: randomly select one of the first  $k$  elements from the population and then every  $k_{th}$  element thereafter is picked

Systematic sample





## Sampling cont'd

- **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

Cluster sample



- **Systematic sampling**: randomly select one of the first  $k$  elements from the population and then every  $k_{th}$  element thereafter is picked

Systematic sample



- **Convenience sampling**: elements selected from the population on the basis of convenience

## 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<sup>th</sup> 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

# Summarizing Categorical Variables

## Example: Sport Injuries

The paper *“Profile of sport/leisure injuries treated at emergency rooms of urban hospitals.”* by Pelletier et al. 1991 examined the nature and number of sport/leisure injuries treated in hospital emergency rooms in a large metropolitan city. They classified non-contact sports injuries by sport, resulting in the following data set:

Sport
Soccer
Basketball
Others
Basketball
Touch Football
Others
Touch Football
Volleyball
Baseball/softball
⋮

**Question:** How to summarize this data set?

- A **frequency table** for **categorical data** is a table that displays the possible categories along with the associated **frequencies** or **relative frequencies**
- The **frequency** for a particular category is the number of times the category appears in the data set
- The **relative frequency** for a particular category is the fraction or proportion of the time that the category appears in the data set.

# Frequencies and Relative Frequencies

```
> table(sport)
sport
Baseball/softball      Basketball      Bicycling      Jogging/running
           11           19           11           11
      Others      Soccer      Touch Football      Volleyball
           47           24           38           17

> table(sport) / dim(sport)[1]
sport
Baseball/softball      Basketball      Bicycling      Jogging/running
    0.06179775    0.10674157    0.06179775    0.06179775
      Others      Soccer      Touch Football      Volleyball
    0.26404494    0.13483146    0.21348315    0.09550562
```

# Frequencies and Relative Frequencies

```
> table(sport)
sport
Baseball/softball      Basketball      Bicycling      Jogging/running
           11              19              11              11
      Others      Soccer      Touch Football      Volleyball
           47              24              38              17

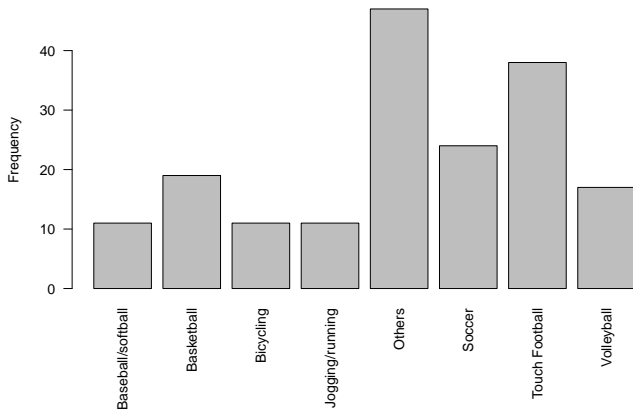
> table(sport) / dim(sport)[1]
sport
Baseball/softball      Basketball      Bicycling      Jogging/running
    0.06179775      0.10674157      0.06179775      0.06179775
      Others      Soccer      Touch Football      Volleyball
    0.26404494      0.13483146      0.21348315      0.09550562
```

How could we visualize these information?

⇒ Making a **bar chart** and/or a **pie chart**

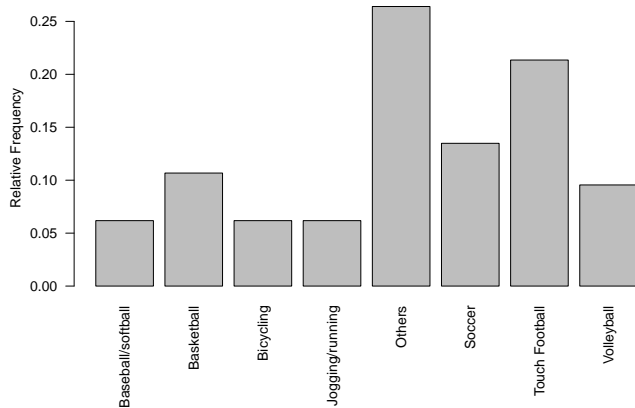
## Bar Charts

A **bar chart** draws a bar with a height proportional to the count in the table:

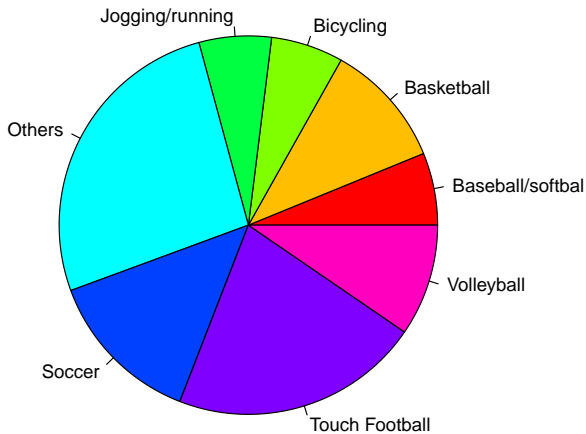




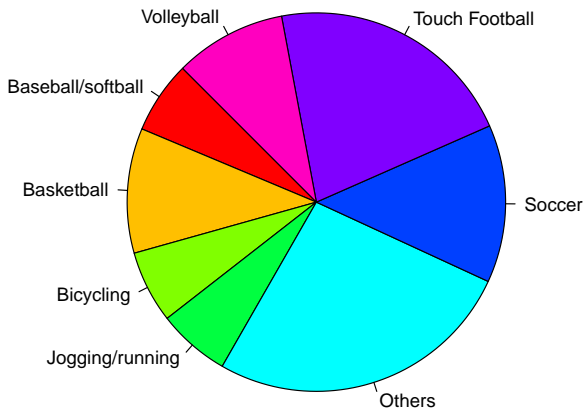
## Bar Charts cont'd



# Pie Charts

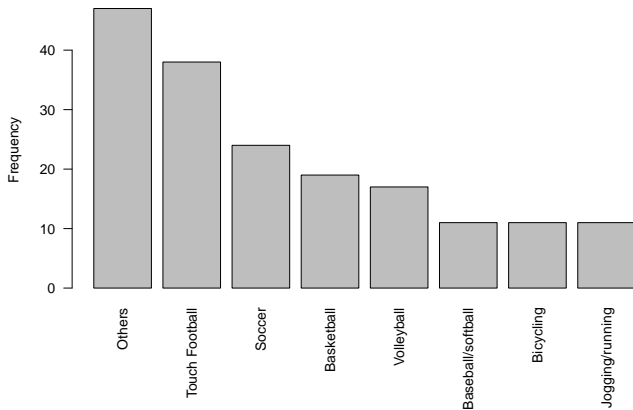


## Pie Charts cont'd



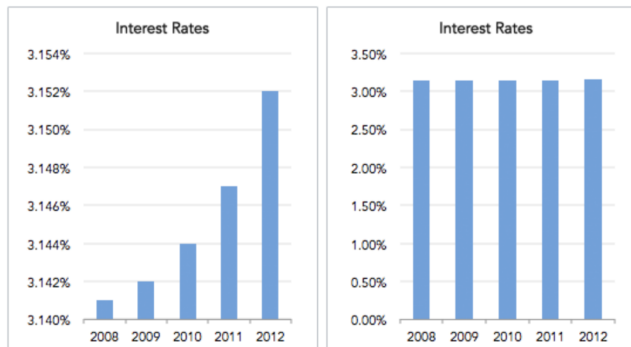
Discussion: Which one you prefer to visualize categorical variables. Why?

# A Good Bar Chart



# A (Potential) Misleading Bar Chart

**Same Data, Different Y-Axis**



## Example: O'Hare Airport Flight Data



	carrier	origin
--	---------	--------

1	UA	EWR
2	AA	LGA
3	AA	LGA
4	AA	LGA
5	UA	LGA
6	UA	EWR

In this example, we have two categorical variables, `carrier` and `origin`, respectively. How to summarize/visualize this dataset?

Sampling Techniques

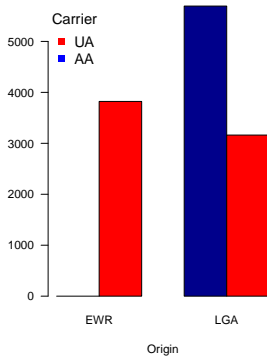
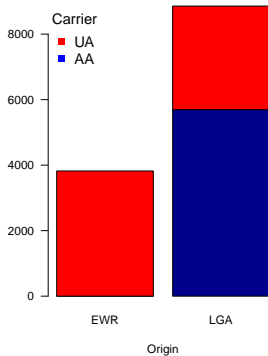
Summarizing  
Categorical Data

Summarizing  
Numerical Data

## ORD Flight Data Cont'd

	EWR	LGA
AA	0	5694
UA	3822	3162

	EWR	LGA
AA	0.00	0.45
UA	0.30	0.25





# Summarizing Numerical Variables

## Example: Murder arrests (per 100,000) in US States in 1973

**Data:** 13.2, 10.0, 8.1, 8.8, 9.0, 7.9, 3.3, 5.9,  
15.4, 17.4, 5.3, 2.6, 10.4, 7.2, 2.2, 6.0,  
9.7, 15.4, 2.1, 11.3, 4.4, 12.1, 2.7, 16.1,  
9.0, 6.0, 4.3, 12.2, 2.1, 7.4, 11.4, 11.1,  
13.0, 0.8, 7.3, 6.6, 4.9, 6.3, 3.4, 14.4, 3.8,  
13.2, 12.7, 3.2, 2.2, 8.5, 4.0, 5.7, 2.6, 6.8.

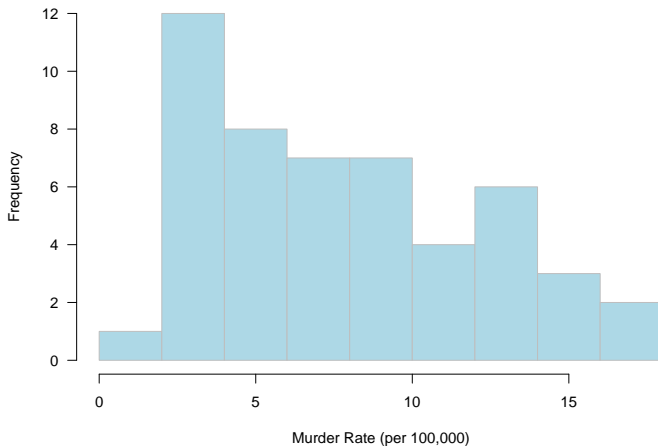
**Question:** How to graphically summarize this data set?

# Stem-and-Leaf Plot

The decimal point is at the |

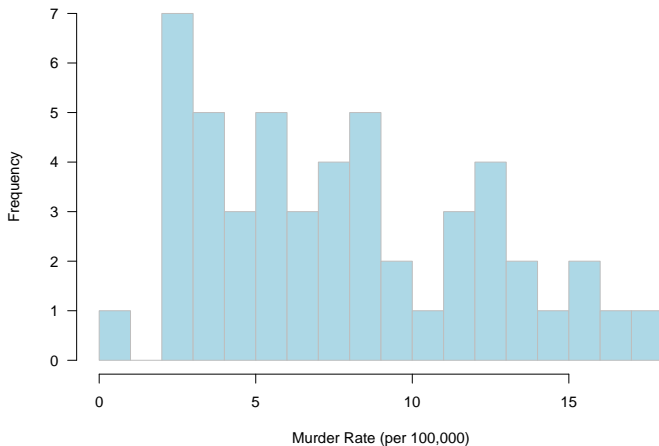
```
0 | 8
1 |
2 | 1122667
3 | 2348
4 | 0349
5 | 379
6 | 00368
7 | 2349
8 | 158
9 | 007
10 | 04
11 | 134
12 | 127
13 | 022
14 | 4
15 | 44
16 | 1
17 | 4
```

**Histogram of US Murder Rate in 1973**



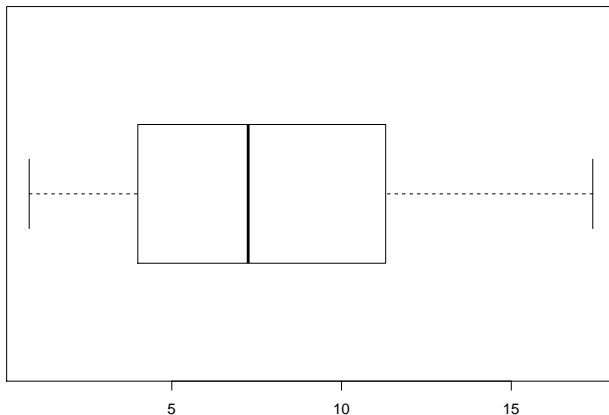
# Histogram

Histogram of US Murder Rate in 1973

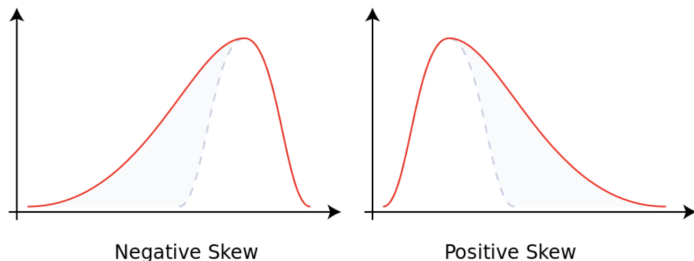


# Box-and-Whisker Plot

**Murder Rate (per 100,000)**



# Shape of Distributions



**Source:** [Skewness - Wikipedia](#)

In the rest of the class, we will talk about how to summarize a numerical variable in terms of its **center** and **spread**

- A **measure of center** attempts to report a “typical” value for the variable
- When a measure of center is calculated with **sample data** it is a **statistic**
- When a measure of center is calculated with popular (e.g., census data) it is a **parameter**
- **Measures:** Mean, Median, Mode



- The **population mean**, denoted by  $\mu_X$ , is the sum of all the population values ( $\{X_i, \dots, X_N\}$ ) divided by the size of the population ( $N$ ). That is,

$$\mu_X = \frac{\sum_{i=1}^N X_i}{N}$$

- The **sample mean**, denoted by  $\bar{X}$  is the sum of all the sample values ( $\{X_1, \dots, X_n\}$ ) divided by the sample size ( $n$ ). That is,

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$$

The **median** is the value separating the higher half from the lower half of a data sample

**How to compute the median:** Order the  $n$  observations in a data set from smallest to largest, then

$$\text{Median} = \begin{cases} \text{the single middle value,} & n \text{ odd} \\ \text{the average of the middle two values,} & n \text{ even} \end{cases}$$

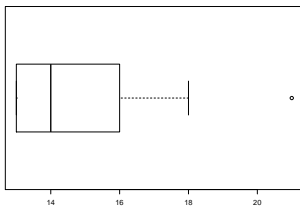
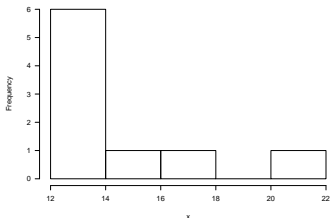
The **mode** is the value of the observation that appears most frequently

**How to compute the mode(s):** Order the observations in a data set from smallest to largest, then find the number that is repeated more often than any other

## Example

Suppose we have the following list of values: 13, 18, 13, 14, 13, 16, 14, 21, 13

- Plot this “data set” and describe the shape of the distribution



## Example cont'd

Suppose we have the following list of values: 13, 18, 13, 14, 13, 16, 14, 21, 13

- Find the sample mean

$$\bar{X} = \sum_{i=1}^9 \frac{13 + 18 + 13 + 14 + 13 + 16 + 14 + 21 + 13}{9} = 15$$

- Find the sample median

1 Order the data first: 13, 13, 13, 13, 14, 14, 16, 18, 21

## Example cont'd

Suppose we have the following list of values: 13, 18, 13, 14, 13, 16, 14, 21, 13

- Find the sample mean

$$\bar{X} = \sum_{i=1}^9 \frac{13 + 18 + 13 + 14 + 13 + 16 + 14 + 21 + 13}{9} = 15$$

- Find the sample median

1 Order the data first: 13, 13, 13, 13, 14, 14, 16, 18, 21

Suppose we have the following list of values: 13, 18, 13, 14, 13, 16, 14, 21, 13

- Find the sample mean

$$\bar{X} = \sum_{i=1}^9 \frac{13 + 18 + 13 + 14 + 13 + 16 + 14 + 21 + 13}{9} = 15$$

- Find the sample median

- 1 Order the data first: 13, 13, 13, 13, 14, 14, 16, 18, 21
- 2 Compute the sample size  $n$  and identify (or compute) the median value

Suppose we have the following list of values: 13, 18, 13, 14, 13, 16, 14, 21, 13

- Find the sample mean

$$\bar{X} = \sum_{i=1}^9 \frac{13 + 18 + 13 + 14 + 13 + 16 + 14 + 21 + 13}{9} = 15$$

- Find the sample median

- 1 Order the data first: 13, 13, 13, 13, 14, 14, 16, 18, 21
- 2 Compute the sample size  $n$  and identify (or compute) the median value



Suppose we have the following list of values: 13, 18, 13, 14, 13, 16, 14, 21, 13

- Find the sample mean

$$\bar{X} = \sum_{i=1}^9 \frac{13 + 18 + 13 + 14 + 13 + 16 + 14 + 21 + 13}{9} = 15$$

- Find the sample median

- 1 Order the data first: 13, 13, 13, 13, 14, 14, 16, 18, 21
- 2 Compute the sample size  $n$  and identify (or compute) the median value
- 3  $n = 9 \Rightarrow$  the median is the 5th number, which is 14

- Find the mode
  - ① Order the data first: 13, 13, 13, 13, 14, 14, 16, 18, 21

- Find the mode
  - ① Order the data first: 13, 13, 13, 13, 14, 14, 16, 18, 21

- Find the mode

- 1 Order the data first: 13, 13, 13, 13, 14, 14, 16, 18, 21

- 2 We have 3 13 and 2 14  $\Rightarrow$  13 is the mode

## Example: Resistant (Robust) Statistics

Suppose we have the following list of values: 13, 18, 13, 14, 13, 16, 14, 210, 13

- Find the sample mean

$$\bar{X} = \sum_{i=1}^9 \frac{13 + 18 + 13 + 14 + 13 + 16 + 14 + 210 + 13}{9} = 36$$

- Find the sample median

1 Order the data first: 13, 13, 13, 13, 14, 14, 16, 18, 210

## Example: Resistant (Robust) Statistics

Suppose we have the following list of values: 13, 18, 13, 14, 13, 16, 14, 210, 13

- Find the sample mean

$$\bar{X} = \sum_{i=1}^9 \frac{13 + 18 + 13 + 14 + 13 + 16 + 14 + 210 + 13}{9} = 36$$

- Find the sample median

1 Order the data first: 13, 13, 13, 13, 14, 14, 16, 18, 210

## Example: Resistant (Robust) Statistics

Suppose we have the following list of values: 13, 18, 13, 14, 13, 16, 14, 210, 13

- Find the sample mean

$$\bar{X} = \sum_{i=1}^9 \frac{13 + 18 + 13 + 14 + 13 + 16 + 14 + 210 + 13}{9} = 36$$

- Find the sample median

1 Order the data first: 13, 13, 13, 13, 14, 14, 16, 18, 210

2 Compute the sample size  $n$  and identify (or compute) the median value

## Example: Resistant (Robust) Statistics

Suppose we have the following list of values: 13, 18, 13, 14, 13, 16, 14, 210, 13

- Find the sample mean

$$\bar{X} = \sum_{i=1}^9 \frac{13 + 18 + 13 + 14 + 13 + 16 + 14 + 210 + 13}{9} = 36$$

- Find the sample median

1 Order the data first: 13, 13, 13, 13, 14, 14, 16, 18, 210

2 Compute the sample size  $n$  and identify (or compute) the median value



## Example: Resistant (Robust) Statistics

Suppose we have the following list of values: 13, 18, 13, 14, 13, 16, 14, 210, 13

- Find the sample mean

$$\bar{X} = \sum_{i=1}^9 \frac{13 + 18 + 13 + 14 + 13 + 16 + 14 + 210 + 13}{9} = 36$$

- Find the sample median

- 1 Order the data first: 13, 13, 13, 13, 14, 14, 16, 18, 210
- 2 Compute the sample size  $n$  and identify (or compute) the median value
- 3  $n = 9 \Rightarrow$  the median is the 5th number, which is (still) 14

## Example cont'd

- Find the mode
  - ① Order the data first: 13, 13, 13, 13, 14, 14, 16, 18, 210

## Example cont'd

- Find the mode
  - ① Order the data first: 13, 13, 13, 13, 14, 14, 16, 18, 210

## Example cont'd

- Find the mode
  - ① Order the data first: 13, 13, 13, 13, 14, 14, 16, 18, 210
  - ② We have 3 13 and 2 14  $\Rightarrow$  13 is (still) the mode

## Example cont'd

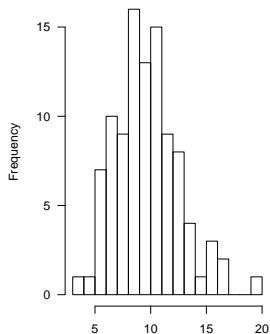
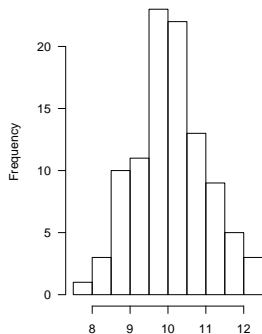
- Find the mode
  - ① Order the data first: 13, 13, 13, 13, 14, 14, 16, 18, 210
  - ② We have 3 13 and 2 14  $\Rightarrow$  13 is (still) the mode

## Example cont'd

- Find the mode
  - 1 Order the data first: 13, 13, 13, 13, 14, 14, 16, 18, 210
  - 2 We have 3 13 and 2 14  $\Rightarrow$  13 is (still) the mode

What is the take-home message?

# Measures of Spread



- **Measures:** Range, Variance/Standard Deviation, Interquartile range (IQR)

The **range** of a dataset is the difference between the largest and smallest values

$$\text{Range} = \text{Largest Value} - \text{Smallest Value}$$

- Compute the range of the following list of values: 13, 18, 13, 14, 13, 16, 14, 21, 13
- Compute the range of the following list of values: 13, 18, 13, 14, 13, 16, 14, 210, 13

**Question:** Is Range a robust statistic?



- The sample standard deviation (variance), denoted by  $s$  ( $s^2$ ), is a measure of the amount of variation of data.  $s$  ( $s^2$ ) can be used as the estimate of the population standard deviation (variance), denoted by  $\sigma$  ( $\sigma^2$ )
- $s$  is calculated in the following way:
  - 1 Calculate the sample mean  $\bar{X}$
  - 2 Calculate the deviation (from the sample mean) for each observation (i.e.,  $X_i - \bar{X}$ ,  $i = 1, \dots, n$ )
  - 3 Square each deviation and add them (i.e.,  $\sum_{i=1}^n (X_i - \bar{X})^2$ )
  - 4 Divide by  $n - 1$  and take the square root, that is,

$$s = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}}$$

## Example

- Compute  $s$  of the following list of values: 13, 18, 13, 14, 13, 16, 14, 21, 13
- Compute  $s$  of the following list of values: 13, 18, 13, 14, 13, 16, 14, 210, 13

**Question:** Is standard deviation a robust statistic?

## Interquartile range (IQR)

- $IQR = Q_3 - Q_1$ , where  $Q_1$  is the **Lower Quartile** (the median of the lower half of the data) and  $Q_3$  is the **Upper Quartile** (the median of the upper half of the data)
- Compute the IQR of the following list of values: 13, 18, 13, 14, 13, 16, 14, 21, 13
- Compute the IQR of the following list of values: 13, 18, 13, 14, 13, 16, 14, **210**, 13

**Question:** Is IQR a robust statistic?

## Summary

In this lecture, we learned

- Sampling Techniques
- Summarizing Categorical Data
- Summarizing Numerical Data

In next lecture we will learn

- How to construct a boxplot
- How to visualize numerical + categorical variables and numerical + numerical variables
- How to visualize time series, cross-sectional, and spatio-temporal Data sets