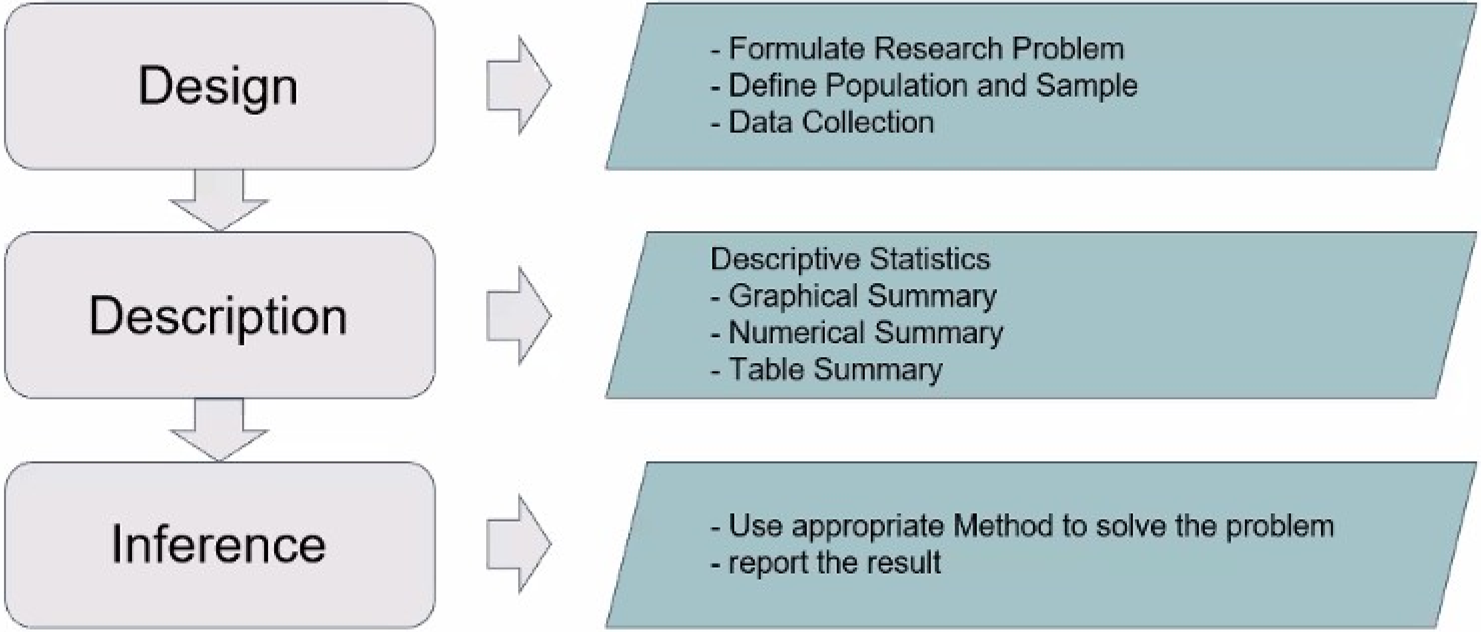
Step by Step Statistics



Design

- Formulate Research Problem
- Define Population and Sample
- Data Collection

Description

Descriptive Statistics

- Graphical Summary
- Numerical Summary
- Table Summary

Inference

- Use appropriate Method to solve the problem
- report the result

Type of Statistic

- **Descriptive Statistic**

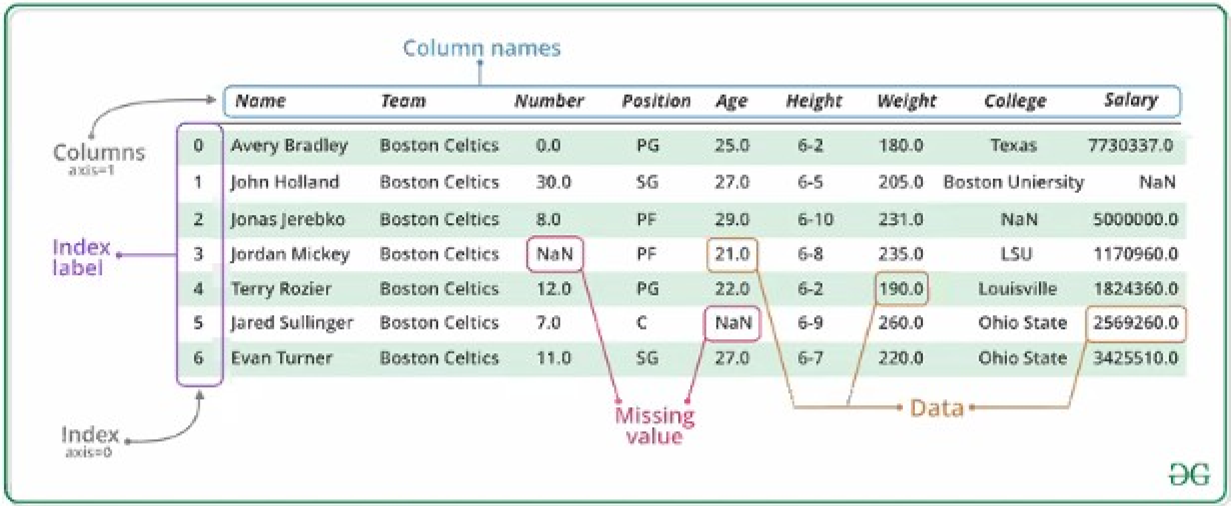
  Branch of statistic to summarize and describe the data. It consist of methods for organizing and summarizing information.

- **Inferential Statistic**

  Branch of statistic to use the data sample to make an inference about a population. It consist of methods for drawing and measuring the reliability of conclusion based on the sample from the population.

Data

Individual units of information



	Name	Team	Number	Position	Age	Height	Weight	College	Salary
0	Avery Bradley	Boston Celtics	0.0	PG	25.0	6-2	180.0	Texas	7730337.0
1	John Holland	Boston Celtics	30.0	SG	27.0	6-5	205.0	Boston Uniersity	NaN
2	Jonas Jerebko	Boston Celtics	8.0	PF	29.0	6-10	231.0	NaN	5000000.0
3	Jordan Mickey	Boston Celtics	NaN	PF	21.0	6-8	235.0	LSU	1170960.0
4	Terry Rozier	Boston Celtics	12.0	PG	22.0	6-2	190.0	Louisville	1824360.0
5	Jared Sullinger	Boston Celtics	7.0	C	NaN	6-9	260.0	Ohio State	2569260.0
6	Evan Turner	Boston Celtics	11.0	SG	27.0	6-7	220.0	Ohio State	3425510.0

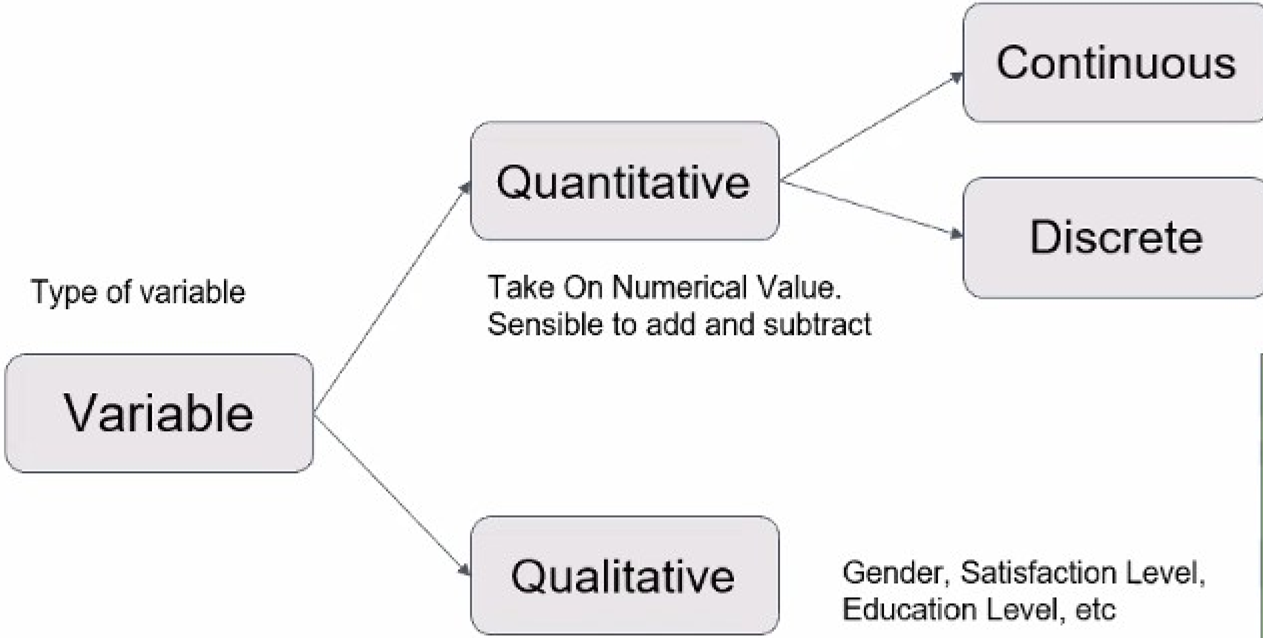
Observation/Unit (players)

Variables

Variables

A characteristic that varies from one person or thing to another is called a variable. Ex: Height, Weight, Eye Color, etc.

Type of variable



Percentage, height, weight, age. etc

Number of family member, Number of Child, Number of pets in a household, etc

Take On Numerical Value. Sensible to add and subtract

Gender, Satisfaction Level, Education Level, etc

Observation belong to one set of distinct category

Why do we care about type of variable ?

In Practice Data Analysis Depends on Type of Variable, example:

- should we use classification or regression when predicting whether a customer will leave or not
- should we use average to describe Number of pets in a household, etc

Scale of Measurement

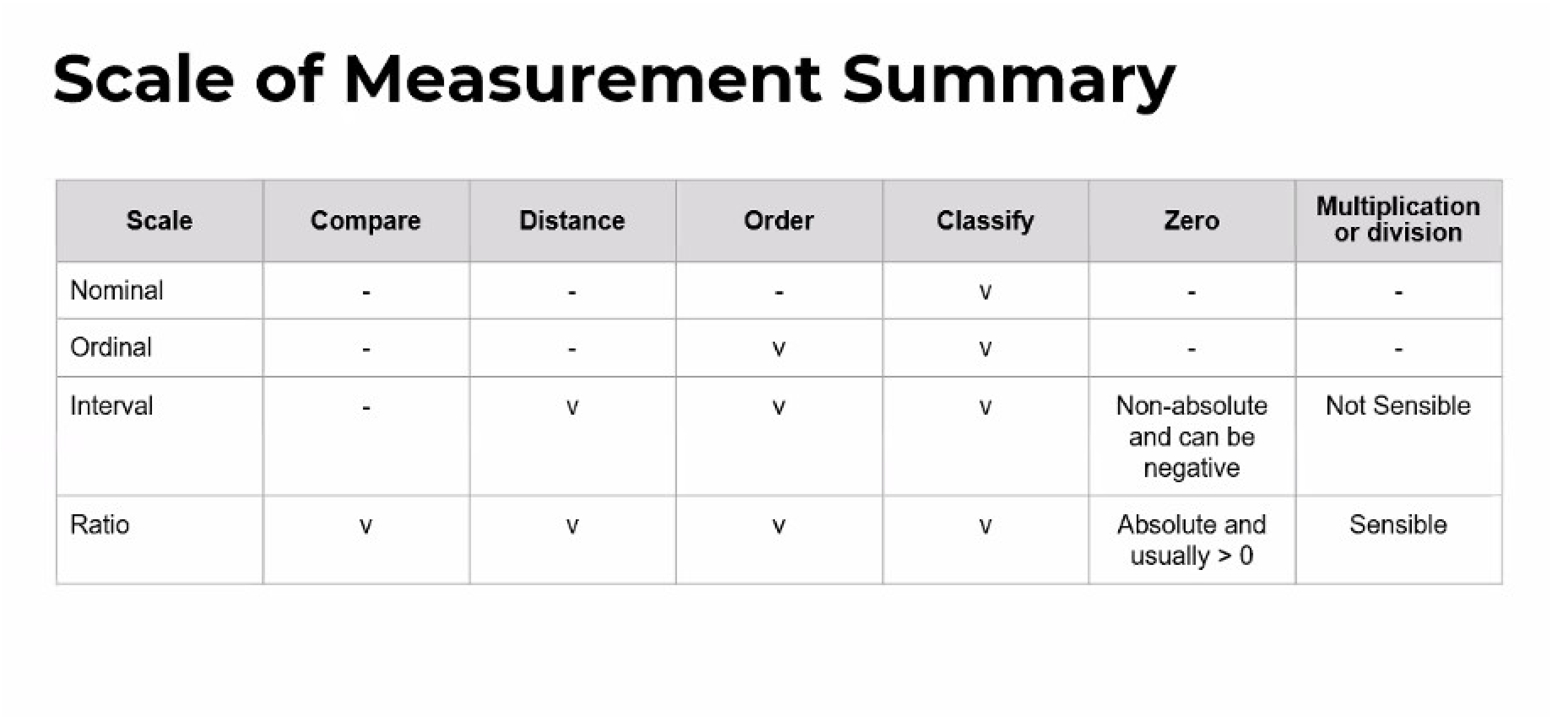
- **Nominal**: Qualitative variables that have two or more categories, but did not have intrinsic order. Ex: Type of fruit (Apple, Banana, Grape), Type of properties(Commercial or Residential), Own a house (Yes or No). Special type of **Nominal** scale when only has two category is **Dichotomous or Binary**.

- **Ordinal**: Qualitative variables that have two or more categories, but the categories could be ranked or ordered. Ex: Satisfaction level (Satisfied, Normal, Not Satisfied), Education Level (SD, SMP, SMA).

Scale of Measurement

- **Interval**: Quantitative variable which their central characteristic could be measured along continuum value and have numerical value. Multiplication or division are not sensible. Ex: Temperature, difference between 20C and 30C is the same as difference between 30C and 40C and 40C is not as hot as 2 x 20C.

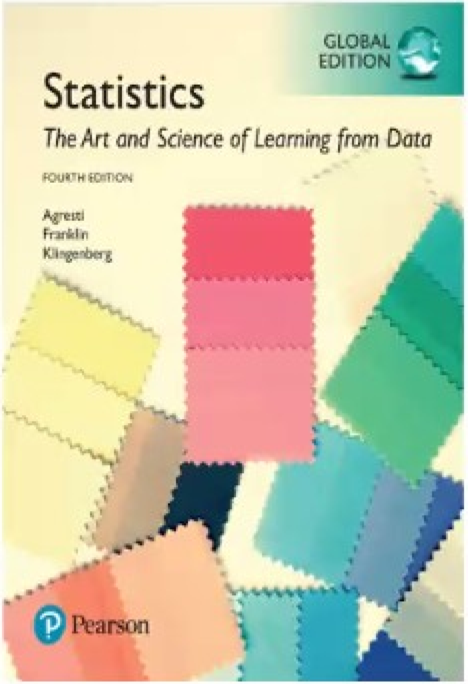
- **Ratio**: Interval variables, but with the added condition that 0 (zero) of the measurement indicates that there is none of that variable. Multiplication or division are sensible. Values from ratio variable usually is greater than 0. So, temperature measured in degrees Celsius or Fahrenheit is not a ratio variable because 0C does not mean there is no temperature. Ex: Weight 0 kg meaning the unit doesn’t exist, 2 x 4 kg is equal to 8 kg. However, temperature measured in Kelvin is a ratio variable as 0 Kelvin (often called absolute zero) indicates that there is no temperature whatsoever.

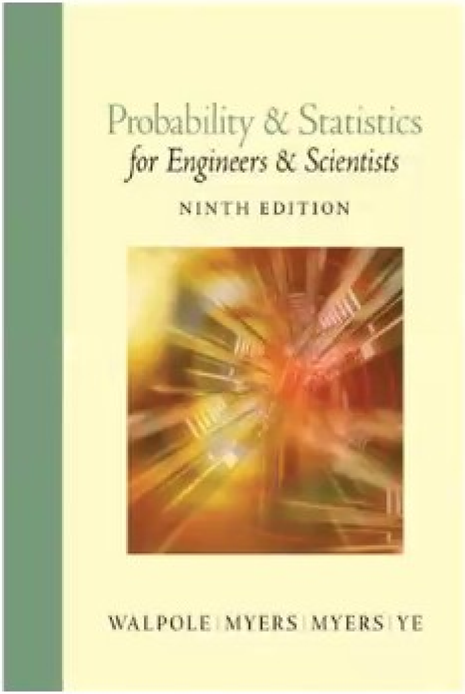


Scale of Measurement Summary

Scale	Compare	Distance	Order	Classify	Zero	Multiplication or division
Nominal	-	-	-	v	-	-
Ordinal	-	-	v	v	-	-
Interval	-	v	v	v	Non-absolute and can be negative	Not Sensible
Ratio	v	v	v	v	Absolute and usually > 0	Sensible

Reference


GLOBAL EDITION
Statistics
The Art and Science of Learning from Data
FOURTH EDITION
Agresti
Franklin
Klingenberg
Pearson


Probability & Statistics
for Engineers & Scientists
NINTH EDITION
WALPOLE | MYERS | MYERS | YE

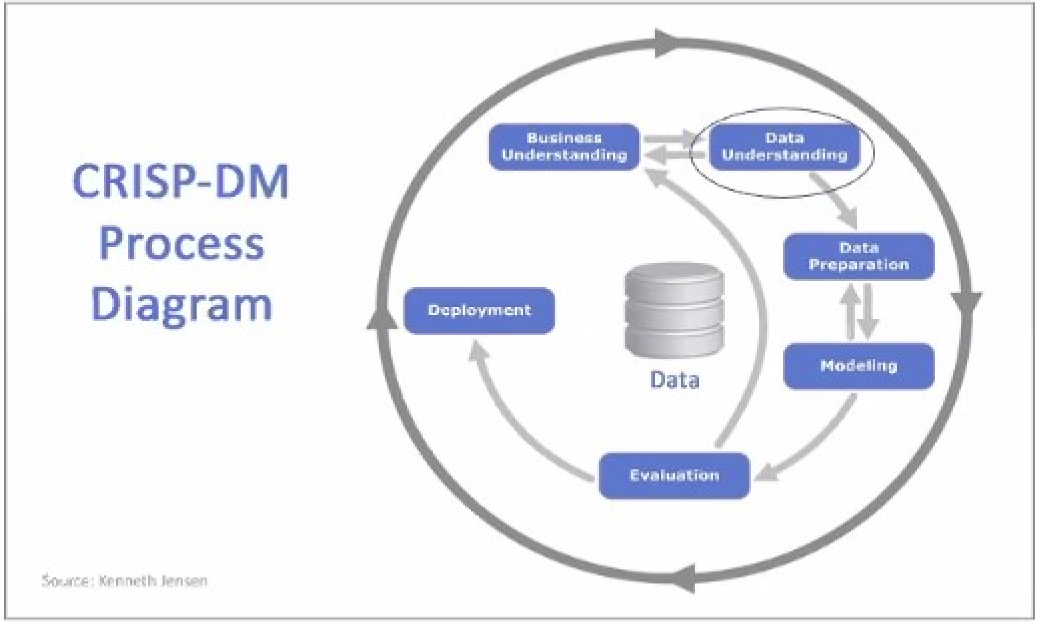
SESSIONS 4

Design Thinking of Statistics

Data Science Program

Outline

- Experimental and Observational Study
- Population and Sample
- Sampling
- Randomness
- Experimenting



Source: Kenneth Jensen

Design Thinking

What aspect that we design in statistics?

- □ Type of Study
- □ Population and sample
- □ Randomness
- □ Sampling
- □ Experimental

Experimental Study

- A researcher conducts an experimental study, or more simply, an experiment, by assigning subjects to certain experimental conditions and then observing outcomes on the response variable (or variables).
- The experimental conditions, which correspond to assigned values of the explanatory variable, are called treatments.
- Example: **A/B Testing** of new web design to increase the conversion rate.

Observational Study

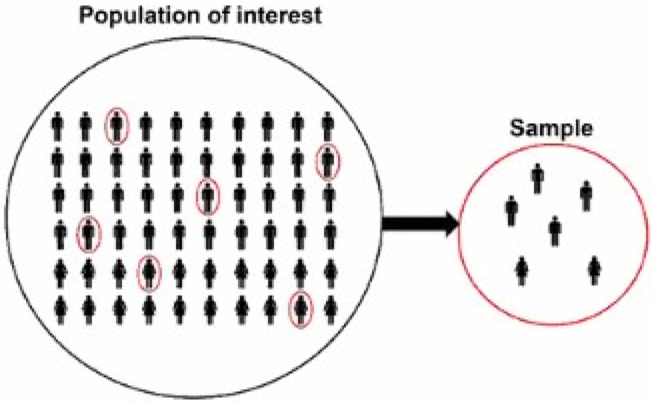
In an observational study, the researcher observes values of the response variable and explanatory variables for the sampled subjects, without anything being done to the subjects (such as imposing a treatment).

Example: **Sample survey** for quick count.

Population & Sample

Population and Sample

- Population can be characterized as the set of individual persons or objects in which an investigator is primarily interested during the research.
- Set of individual or object observed as representation of the population is called sample
- Sample should represent the population.



Population and Sample

- Population always represent the target of an investigation. We learn about the population from the samples
- **Finite Population** is a population that could be physically listed. Ex:
  - Student at Purwadhika
  - Chair at the classroom
- **Hypothetical Population** is a population that was abstract and arise from the phenomenon under consideration. Ex:
  - Factory producing light bulb, if the factory keep the same equipment, using the same produce method, and raw materials. The bulb produced could be consider as hypothetical population

Sampling

Sampling

- **Sample** should **represent** the **population** of interest
- If possible, define the sampling frame (population that could be physically listed).
- A method to choose appropriate sample is called **sampling**
- Sample obtained randomly
- Be cautious of Sampling bias

Randomness

- Notice that young people have higher proportion suffer from heart disease 88.8%.
- Is that make sense that younger people have tend to have higher risk of suffering from heart disease ?

Let's analyze data gathered from a hospital.

	Heart Disease	No Heart Disease	Total
Age > 50	57.8 %	42.2 %	100%
Age < 50	88.8 %	11.2 %	100%

Randomness

- Data from table are **not collected randomly**. The reason why younger people have higher proportion of heart disease may be because their awareness to check their health is lower than older people.
- **Younger people** only **check** only **if** they start to **feel** something **wrong** with their body while **older people** because of their **awareness** regardless there is something wrong with their body or not.
- **Data** is still part of population but **not representative**.

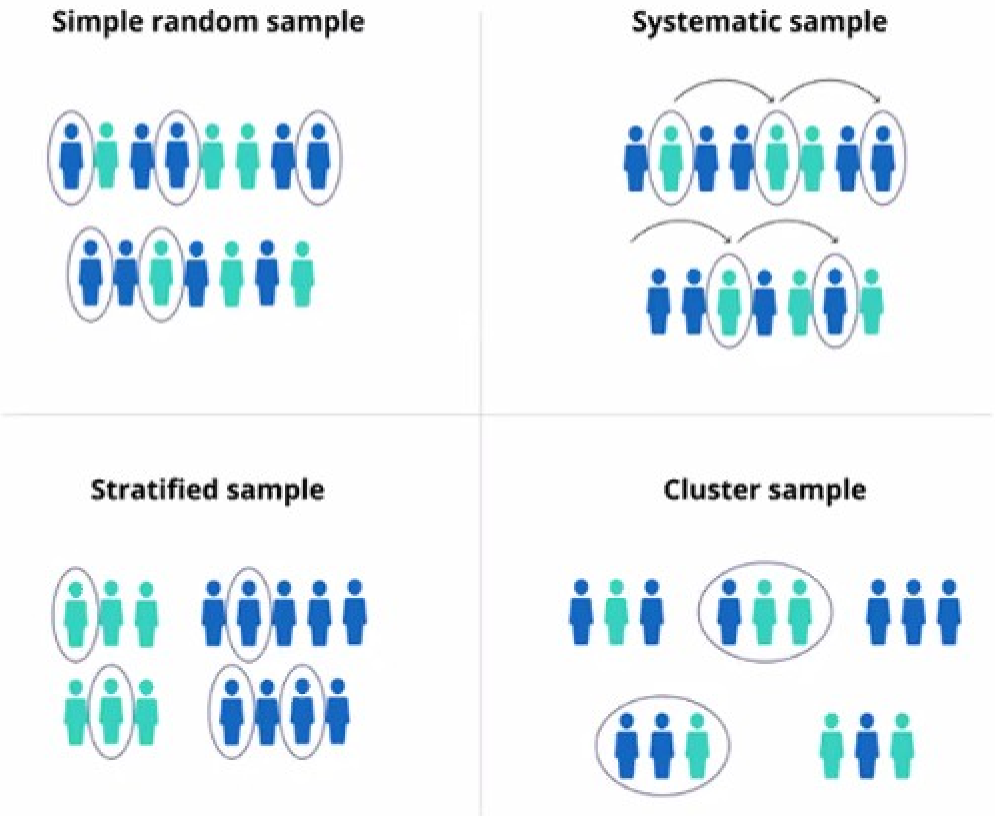
Why do we need sample ?

Things to Consider

- Resource
- Time
- Cost

- When a doctor want to test your blood, Should he/she take all of your blood ?
- When your Mom is cooking, should she eat all the cook or eat only small part of it ?

How to take sample ?



Simple random sample

Systematic sample

Stratified sample

Cluster sample

Experimenting

Experimenting

Key Parts of a Good Experiment

- A good experiment has a control comparison group, randomization in assigning experimental units to treatments, and blinding.
- The experimental units are the subjects—the people, animals, or other objects to which the treatments are applied.
- The treatments are the experimental conditions imposed on the experimental units. One of these may be a control (for instance, either a placebo or an existing treatment) that provides a basis for determining whether a particular treatment is effective. The treatments correspond to values of an outcome.
- Randomly assign the experimental units to the treatments. This tends to balance the comparison groups with respect to lurking variables (covariate).
- Replication of studies (sample size) increases confidence in the conclusions.

Example of An Experimentation

Amazon want to test whether their new app design help the company to increase the conversion rate.

1. Six months ago, the company randomly selected 240 newly signed up users and assigned 122 of them to the control and each 118 to the new designs.
2. Control group went to current design (layout A)
3. The 240 users represent just a small portion of the app's total users (sample).
4. Subject on this experiment is people.
5. By Assigning randomly, we minimize effect of covariate. (ex gender, age, job, etc) for each group.
6. In the end we want to compare the conversion rate.

	Layout A	Layout B
Visitors	122	118
Customers	22	25
Conversion %	18.0%	21.2%

Purwadhika
Startup and Coding School

Reference

https://towardsdatascience.com/data-science-you-need-to-know-a-b-testing-f2f12aff619a

https://towardsdatascience.com/data-science-fundamentals-a-b-testing-cb371ceecc27

https://www.niagahoster.co.id/blog/ab-testing-adalah/

https://vwo.com/blog/ab-testing-examples/

https://www.scribbr.com/methodology/sampling-methods/

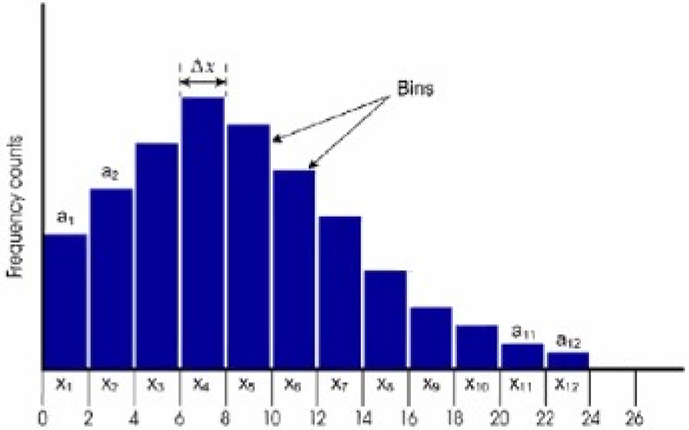
SESSIONS 4

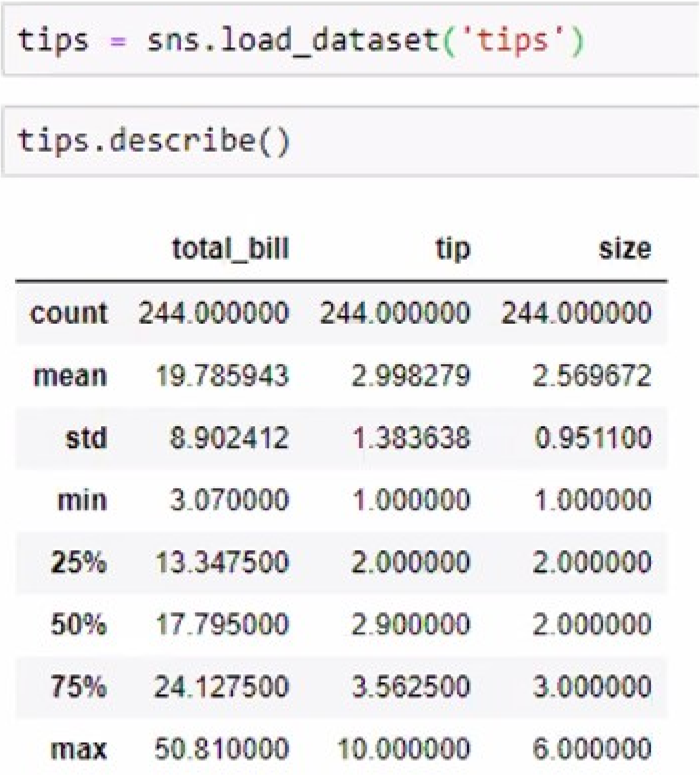
Descriptive Statistics: Numerical and Table Summary

Data Science Program

Descriptive Statistic

- Descriptive statistic includes the construction of graphs, charts, tables, and calculation of various descriptive measures such as averages, variation, and percentile.

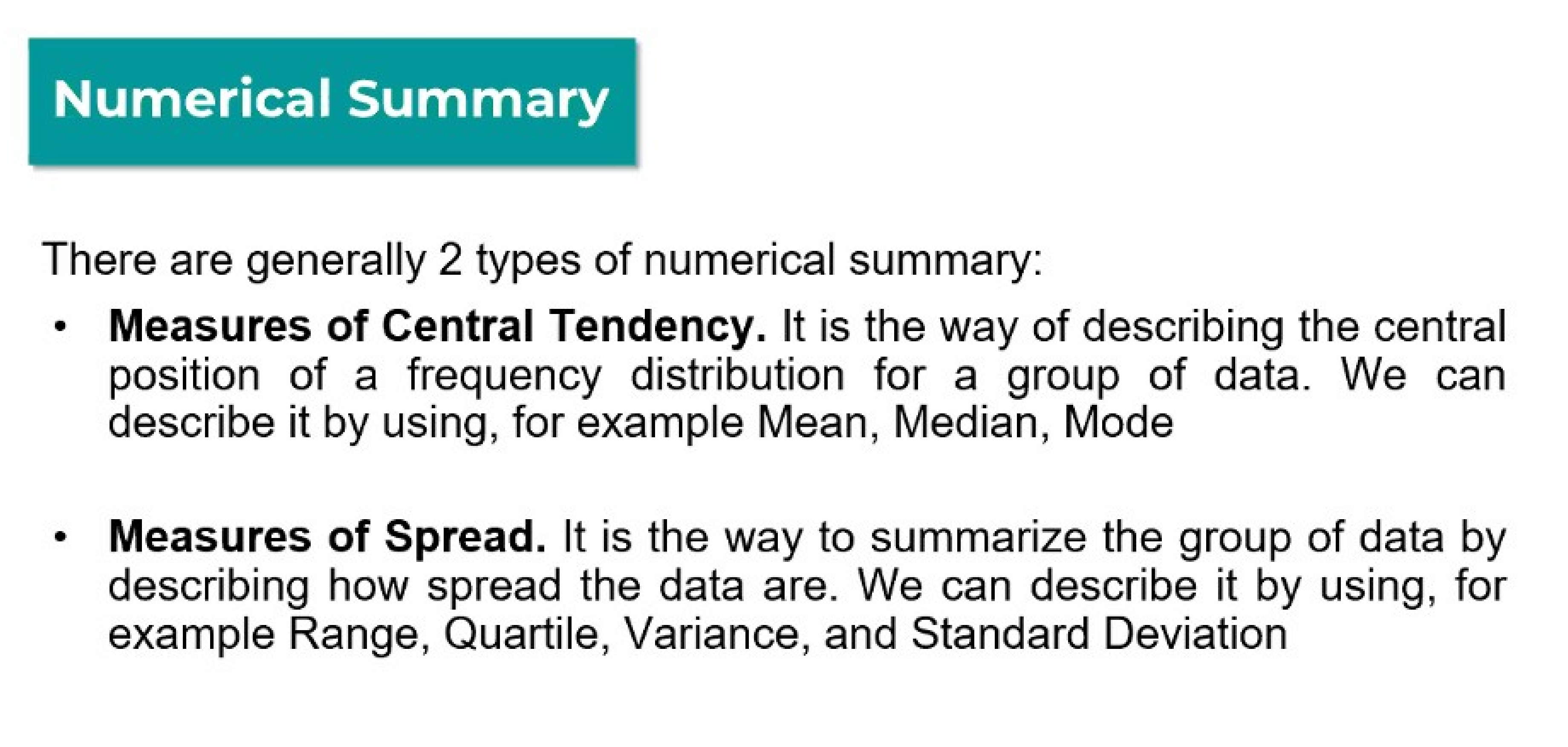




```
tips = sns.load_dataset('tips')
```

```
tips.describe()
```

	total_bill	tip	size
count	244.000000	244.000000	244.000000
mean	19.785943	2.998279	2.569672
std	8.902412	1.383638	0.951100
min	3.070000	1.000000	1.000000
25%	13.347500	2.000000	2.000000
50%	17.795000	2.900000	2.000000
75%	24.127500	3.562500	3.000000
max	50.810000	10.000000	6.000000



Numerical Summary

There are generally 2 types of numerical summary:

- **Measures of Central Tendency.** It is the way of describing the central position of a frequency distribution for a group of data. We can describe it by using, for example Mean, Median, Mode

- **Measures of Spread.** It is the way to summarize the group of data by describing how spread the data are. We can describe it by using, for example Range, Quartile, Variance, and Standard Deviation

Measure of Central Tendency

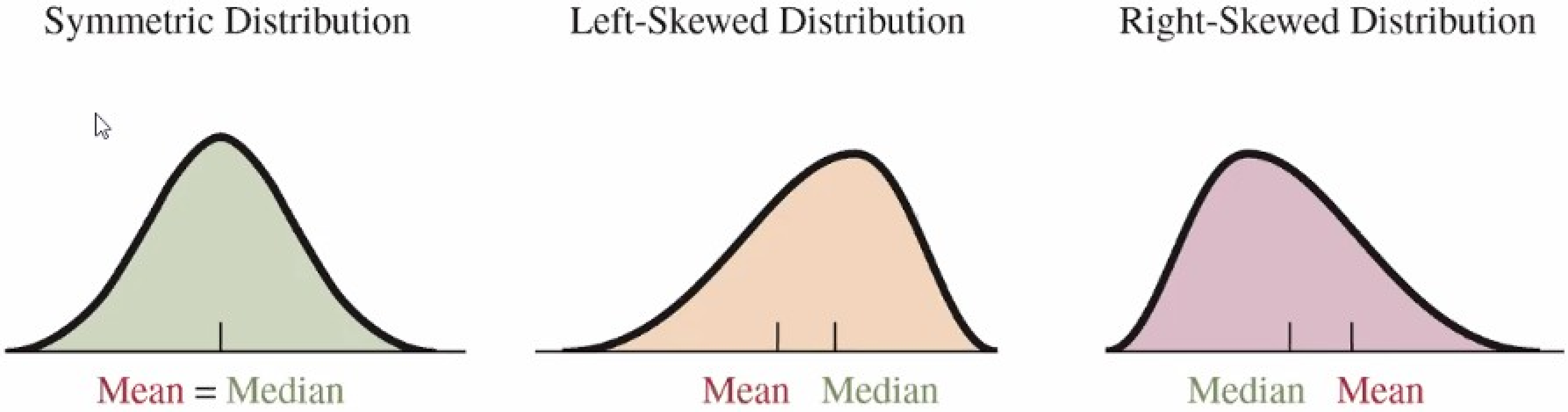
There are three most common way to describe the central measurement of the frequency distribution:

- **Mode**: Value of a qualitative or a countable quantitative variable where the frequency is occurring the most.
- **Median**: The middle value in the ordered list. If the number of observation is odd, then the sample median is the observed value exactly in the middle. If the number of observation is even, then the sample median is the number halfway between the two middle observed values in the ordered list. In either case, the sample median position is at n+1/2 when n is the number of observation
- **Mean**: The sum of observed values in a data divided by the number of observations. The most commonly used measure of center for quantitative variable.

Which measurement to choose?

- **Mode** should be used when calculating the measure of center for the qualitative variable
- **Mean** is the proper measure of center if dealing with the quantitative variable with symmetric distribution (often bell shaped)
- **Median** is the good choice if the quantitative variable have a skewed distribution. We do not used mean in this case, because mean could be highly influenced by an observation that falls far from the rest of the data (outlier)
- It should be noted that this measurement assume that the sample measurement is corresponding to the population measures of center, which are unknown. The sample measurement can be used to estimate this unknown parameter.

Median vs Mean



Symmetric Distribution

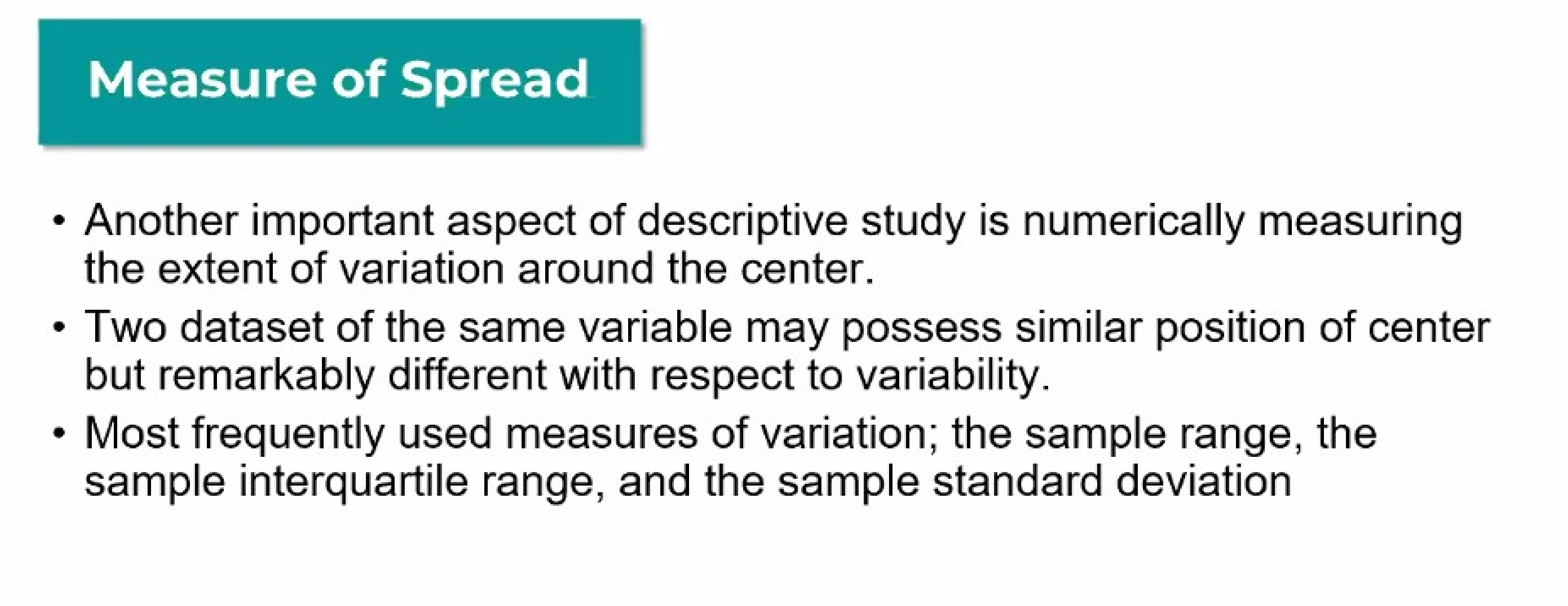
Left-Skewed Distribution

Right-Skewed Distribution

Measure of Central Tendency Example

Patient	Gender	Age
Andrew	Male	22
Jacob	Male	22
Ros	Female	23
Andersen	Male	23
Lina	Female	29
Robert	Male	24
Jack	Male	27
Annie	Female	28

- **Mode**
  - Gender: Male
  - Age: 22 and 23
- **Mean**
  - Age
    (22 + 22 + … + 28) / 8 = 24.75
- **Median**
  Age: 22, 22, 23, **23, 24**, 27, 28, 29,
  Median = 23.5



Measure of Spread

- Another important aspect of descriptive study is numerically measuring the extent of variation around the center.
- Two dataset of the same variable may possess similar position of center but remarkably different with respect to variability.
- Most frequently used measures of variation; the sample range, the sample interquartile range, and the sample standard deviation

Range

- The sample range is obtained by computing the difference between the largest observed value of the variable and the smallest one

Range = Max – Min

- Range is overly sensitive to extreme value

Standard Deviation

- The sample standard deviation is the most frequently used measure of variability. It can be considered as a kind of average of the absolute deviations of observed values from the mean of the variable in question.

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

S = Sample Standard Deviation
X_i = Each value of dataset
$\bar{x}$ = Mean of the dataset
N = Sample size

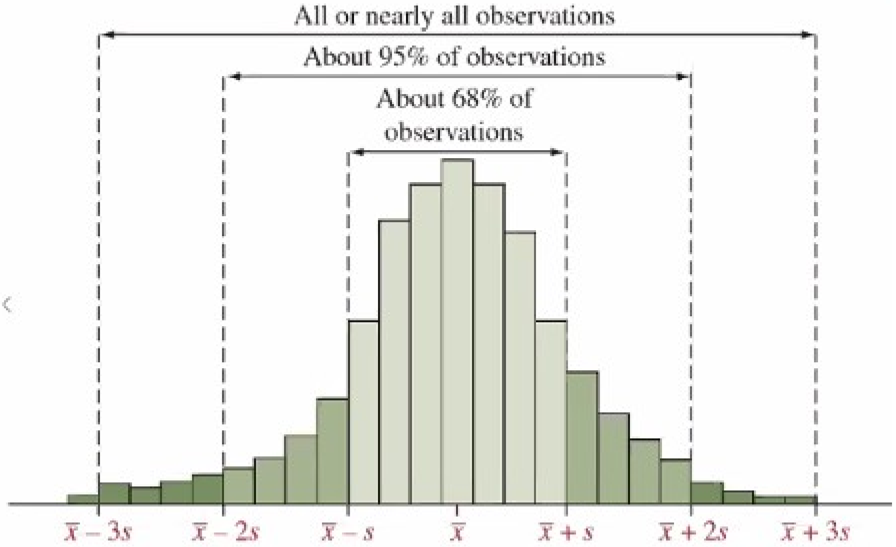
- Since Standard Deviation defined by the sample mean, it is preferred measure of variation if the mean is used as the measure of center(ex: Symmetric Distribution)

Standard Deviation Properties

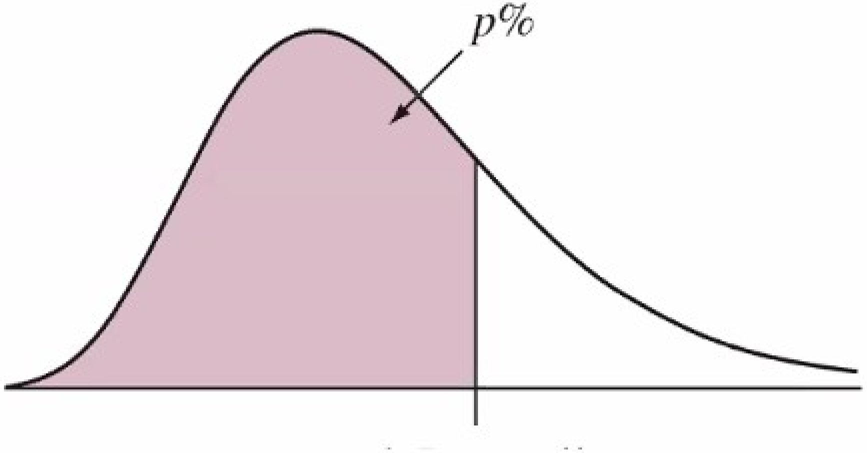
- The more variation in the observed value, the larger the standard deviation for the variable observed.
- Standard Deviation is greatly affected by a few extreme observation (usually outlier).

When a distribution of Data is bell shaped, then approximately:

- 68% of observation fall between $\bar{x}$ - *s* and $\bar{x}$ + *s*
- 95% of observation fall between $\bar{x}$ - 2*s* and $\bar{x}$ + 2*s*
- 99.7% of observation fall between $\bar{x}$ - 3*s* and $\bar{x}$ + 3*s*
- *s* ≈ Range / 4 = (Max – Min) / 4

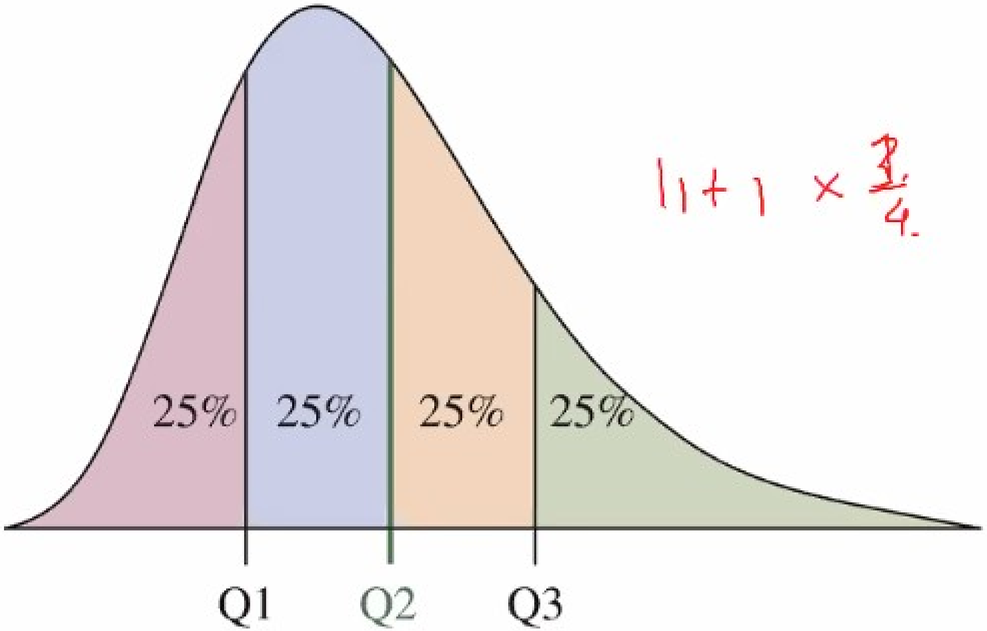


Percentile



The ***p*th Percentile** is a value such that *p* percent of the observation fall below or at that value.

Quartile



11 + 1 × 3/4.

Quartile is a special case of **percentile**.

- Q1 is percentile 25
- Q2 is percentile 50 or also known as median
- Q3 is percentile 75

Q3 = percenti 75

1 2 3 4 5 6 7 8 9 10

Q1 = percentile 25th.

Q2 = percentile 50th.

percentile.

percentile = .. th.

IQR and Outlier

Interquartile range (IQR) is the distance between Q1 and Q2:

IQR = Q3 - Q1

IQR can be used as replacement for standard deviation when data is normally distributed because standard deviation is very sensitive to outliers

S = 1.34898 x IQR

IQR is used to detect the potential outlier

An observation considered outlier if the value

- below Q1 - 1.5 x IQR
- above Q3 + 1.5 x IQR

Measure of Spread Example

Patient	Gender	Age
Andrew	Male	22
Jacob	Male	22
Ros	Female	23
Andersen	Male	23
Lina	Female	29
Robert	Male	24
Jack	Male	27
Annie	Female	28

- Range
  - Age
    29 - 22 = 27
- Standard Deviation
  - Age
    s = 2.63391
- Q2
  22, 22, 23, **23, 24**, 27, 28, 29
  23.5
- Q1 22.75
- Q3 27.25
- IQR 4.5

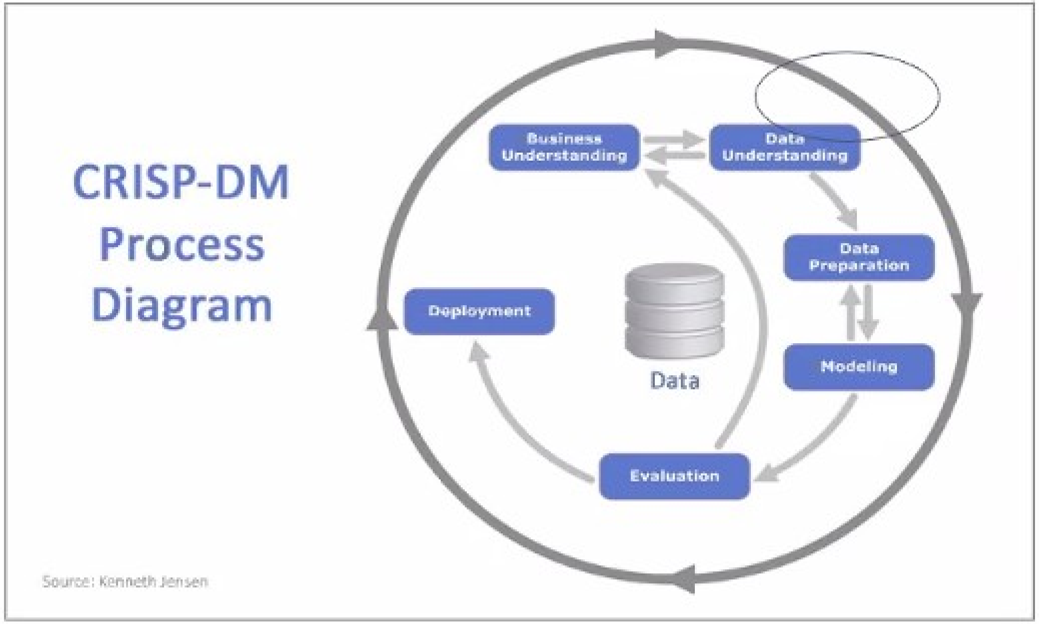
SESSIONS 4

Descriptive Statistics: Graphical Summary

Data Science Program

Outline

- Frequency table
  - For numerical
  - For categorical
- Cross tabulation
- Graphical Summary



CRISP-DM Process Diagram

Source: Kenneth Jensen

Frequency Table

Frequency Table for categorical variable

Day	Visitor Count
Saturday	87
Sunday	76
Thursday	62
Friday	19

Frequency Table for numerical variable

Tip Range ($)	Visitor Count
0 - 2.5	108
2.5 - 4	95
4 - 5.5	29
5.5 - 7	9

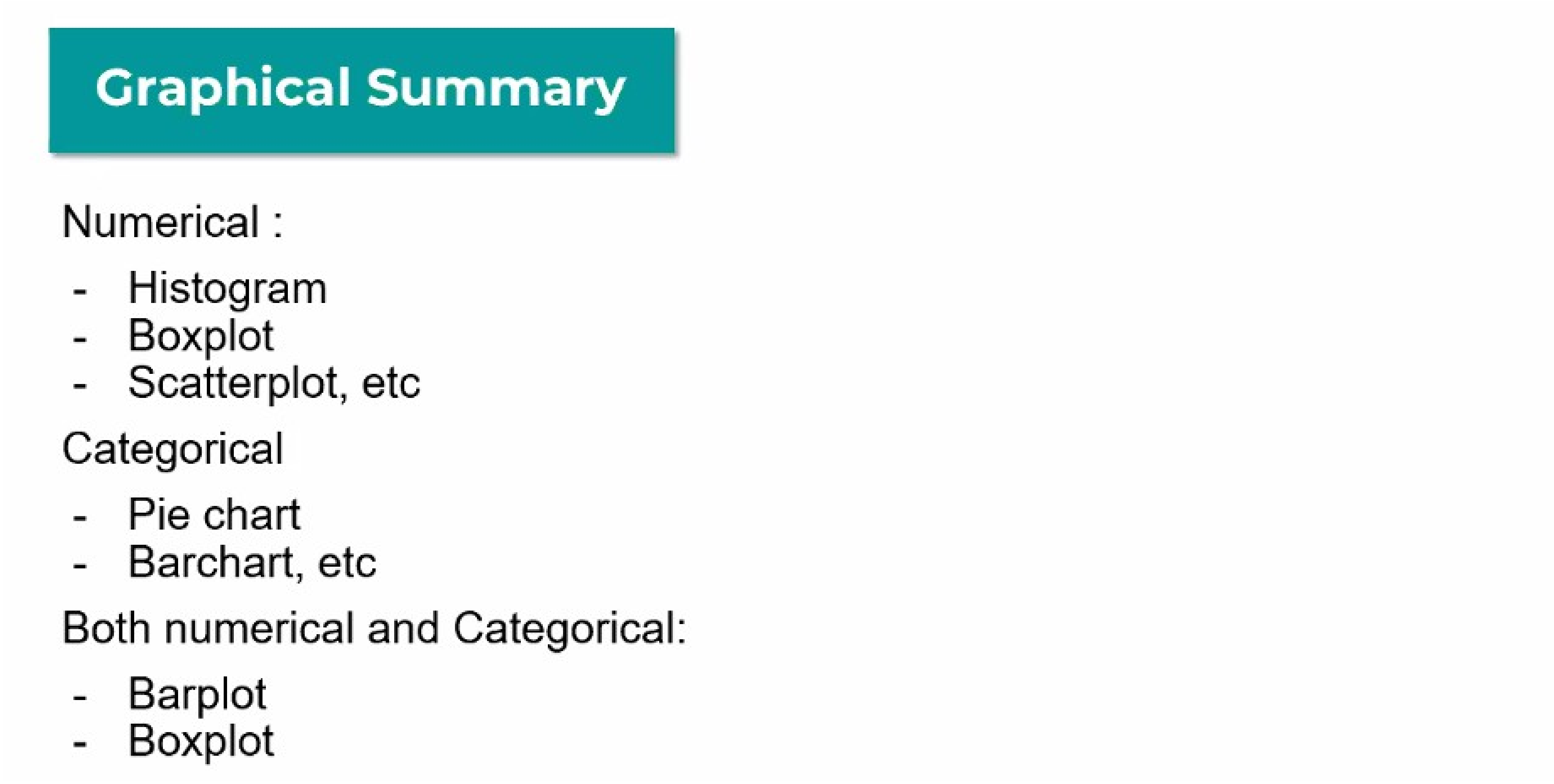
Cross Tabulation / Contingency Table

Cross Tabulation : Frequency

Day	Visitor Count (Male)	Visitor Count (Female)
Saturday	87	32
Sunday	76	9
Thursday	62	28
Friday	19	18

Cross Tabulation : Percentage

Day	Visitor Count (Male)	Visitor Count (Female)	Total
Saturday	73.1	26.9%	100%
Sunday	89.4%	10.6%	100%
Thursday	68.8%	31.1%	100%
Friday	51.3%	48.6%	100%



Graphical Summary

Numerical :

- Histogram
- Boxplot
- Scatterplot, etc

Categorical

- Pie chart
- Barchart, etc

Both numerical and Categorical:

- Barplot
- Boxplot