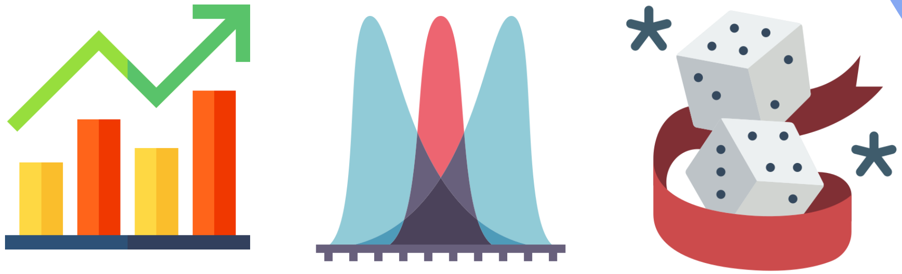


Data Visualization

Mohamad Irwan Afandi

WHAT IS STATISTICS?

Analyzing
Read the data and try to solve problems with it

ORGANIZING
Arrange into structured whole or order

INTERPRETING
Understand (translate) the data and get new information from it

COLLECTING
How to get enough data for better analysis

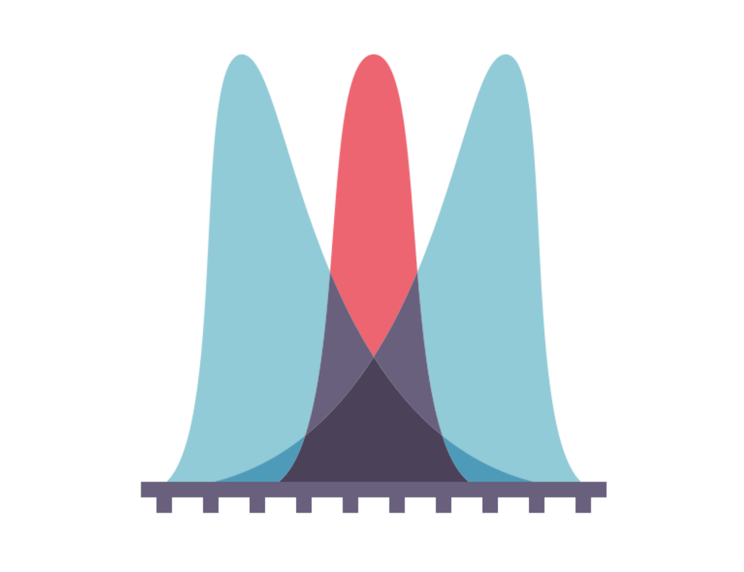
PRESENTING
Inform that insight to the client

DESCRIPTIVE STATISTICS

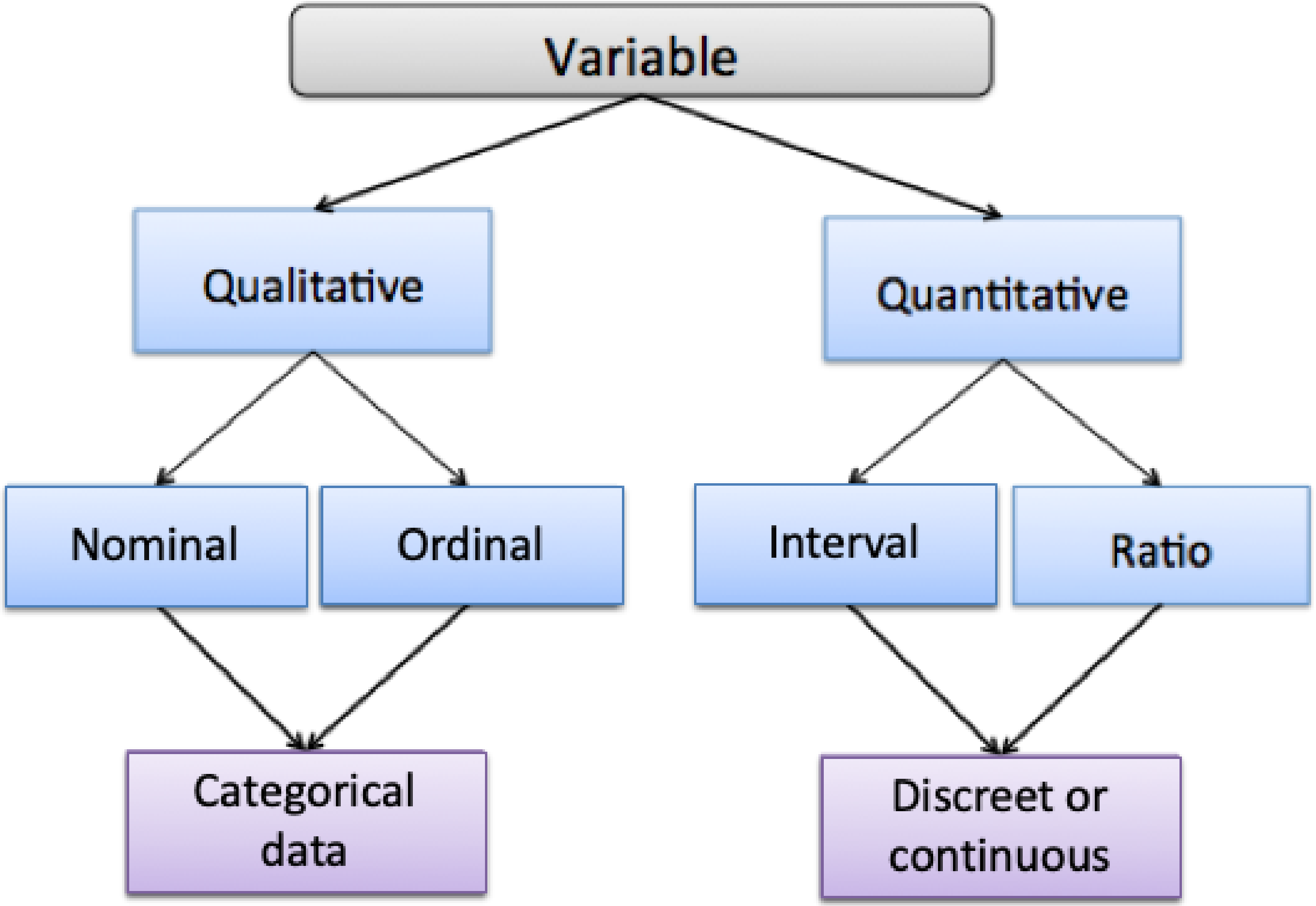
In Descriptive Statistics your are ***describing, organizing, summarizing and presenting*** your data (population), either through ***numerical calculations or using visualization*** like graphs or tables.

INFERENTIAL STATISTICS

Inferential Statistics are produced by ***more complex mathematical calculations***, and allow us to infer ***trends, make assumptions and predictions*** about a population ***based on a study of a sample*** taken from it.



TYPES OF VARIABLES



EXPLANATION

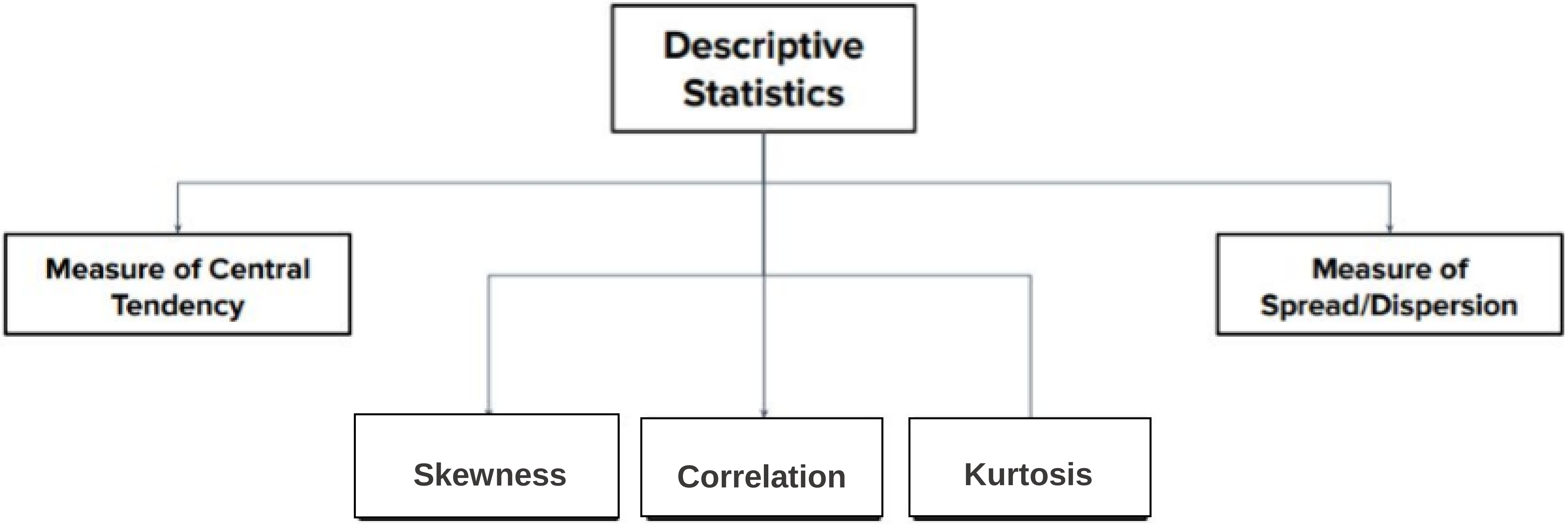
Qualitative : the data can not be computed. Just like a label from the data

- *Nominal (can't be sorted) like : gender, marital status, country, etc.*
- *Ordinal (can be shorted or ranked) like : scale (easy, medium, expert), agreement (disagree, neutral and agree)*

Quantitative : can apply computation in the data. There are 2 type of qualitative data : discrete and continuous.

- *Interval data (integers) : temperature (30-40 C), annual income (65B – 120B)*
- *Ratio : height, weight, temperature in kelvin*

Descriptive Statistics



Measure of Central Tendency

Central tendency is a ***central or typical value for a distribution (location of distribution)*** *The most common measures of central tendency are the arithmetic* ***mean, the median and the mode****.*

Example case

We have numerical data : ***8, 4, 8, 17, 2, 10, 15, 17, 9, 20, 25, 17, 13***

Calculate mean, median and mode

First you have to short the data: ***2, 4, 8, 8, 9, 10, 13, 15, 17, 17, 17, 20, 25***

The number of data (N) ***13 data,*** *and the total of data (sum) is* ***165***

1

Mean (average)

Mean = (sum) / N
= 165 / 13
= ***12.69***

2

Median (middle point)

If N is odd, the median is at (N+1) / 2
If N is even, the median use this formula
((N/2)+((N/2)+1))/2

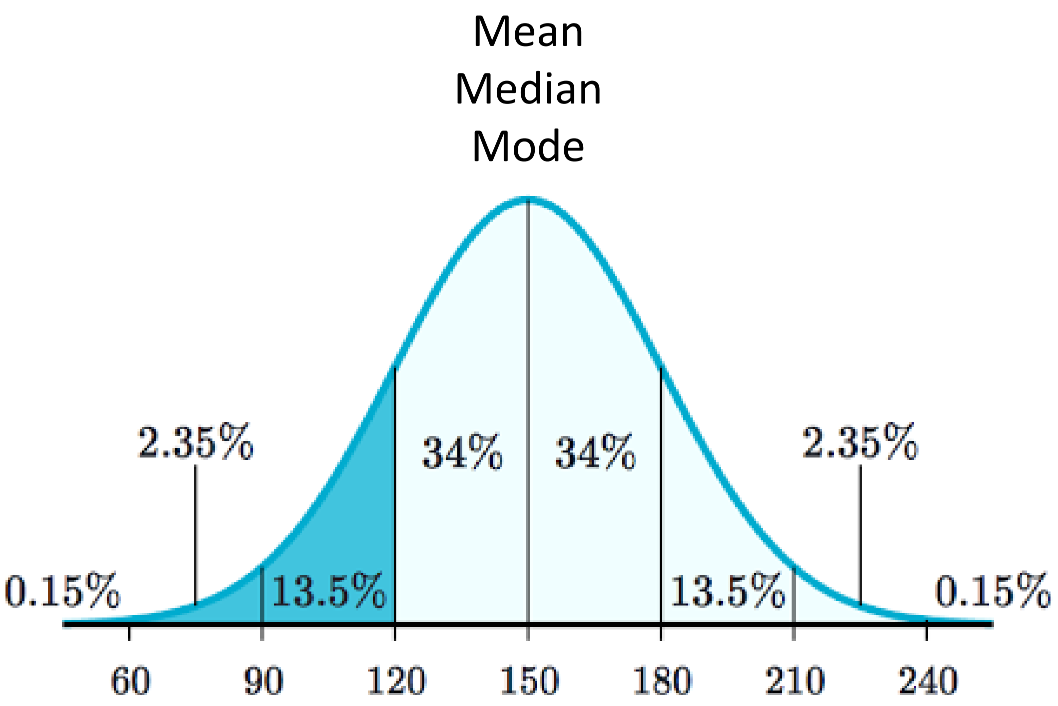
In this case median is at (13+1)/2 = ***7***
2, 4, 8, 8, 9, 10, 13, 15, 17, 17, 17, 20, 25
The answer is ***13***

3

Mode (appear frequently)

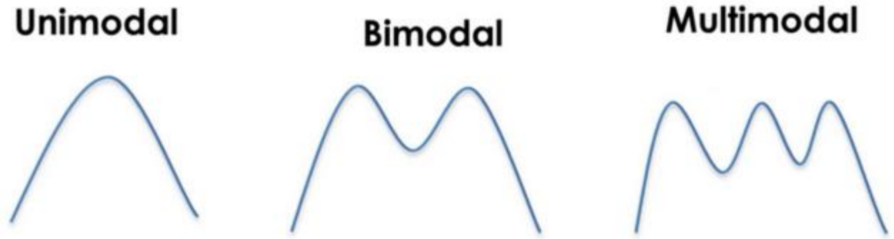
Mode from that data is ***17*** *with 3 frequency number*

But how if when there is more than one mode?
Just, ***select all the numbers.***
Ex : 17, 22, 35



Modality

The number if peaks in a distribution



Normal Distribution

the most important concepts in statistics since nearly **all *statistical tests require normally distributed data***. It basically describes how large samples of data look like when they are plotted.

if your ***data is not normally distributed***, you need to be very careful what statistical tests you apply to it since they could ***lead to wrong conclusions***.

The mean, median, and mode of a normal distribution **are equal.** The normal distribution usually use when normalize the data when we do **preprocessing.**

Measure of Spread and Dispersion

*Statistics that tell us about the **variability** in the data. You can use **range**, interquartile range (**IQR**), **standard deviation** and **variance***

1

Range

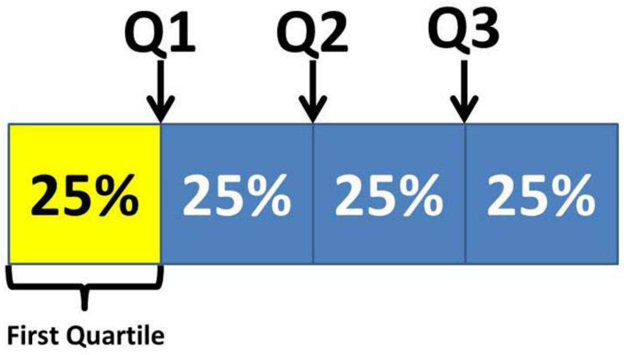
The different between the highest and the lowest data

Range = max(data) – min(data)

2

Interquartile Range (IQR)

Measure of variability based on dividing a dataset into quartiles



Boundary Limit:

- *Higher Outlier : Q3 + (1.5 * Q3)*
- *Lower Outlier : Q1 – (1.5 * Q1)*

3

Standard deviation and Variance

- *Standard deviation is a measure of the amount of variation or dispersion of a set of values.*
- *Variance is expectation of the squared deviation of a random variable from its mean*

For samples:

$$\text{variance} = s^2 = \frac{\Sigma (x - \bar{x})^2}{n - 1}$$

$$\text{standard deviation} = s = \sqrt{s^2}$$

Calculating Formula

$$s^2 = \frac{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}{n - 1}$$

For populations:

$$\text{variance} = \sigma^2 = \frac{\Sigma (x - \bar{x})^2}{n}$$

$$\text{standard deviation} = \sigma = \sqrt{\sigma^2}$$

Calculating Formula

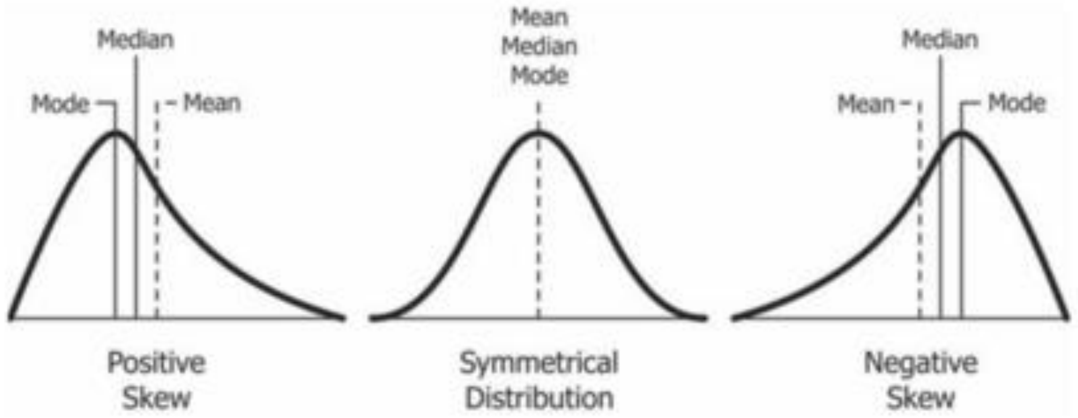
$$\sigma^2 = \frac{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}{n}$$

Skewness, Kurtosis and Correlation

1

Skewness

A measure of symmetry, or more precisely the lack of symmetry. There are 3 kinds of skewness : positive, symmetry and negative.

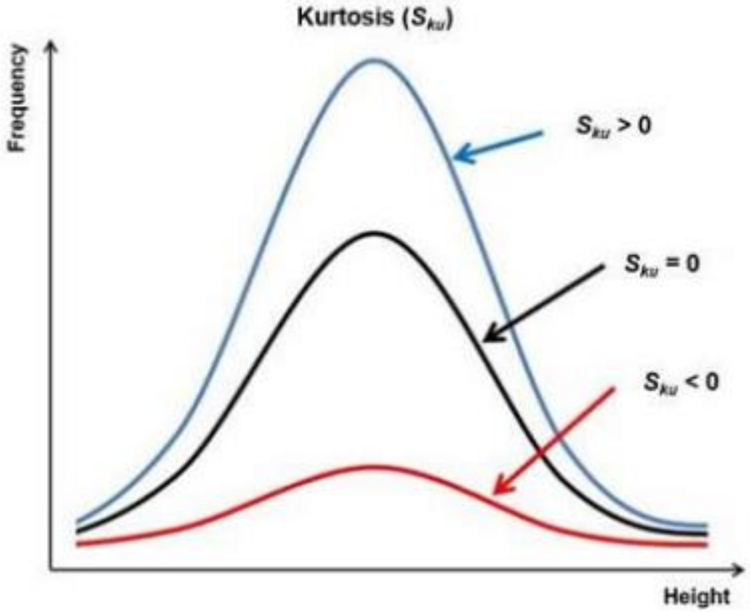


2

Kurtosis

*Whether your dataset is **heavy-tailed or light-tailed** compared to a normal distribution. Data sets with high kurtosis **have heavy tails and more outliers** and data sets with **low kurtosis tend to have light tails and fewer outliers.***

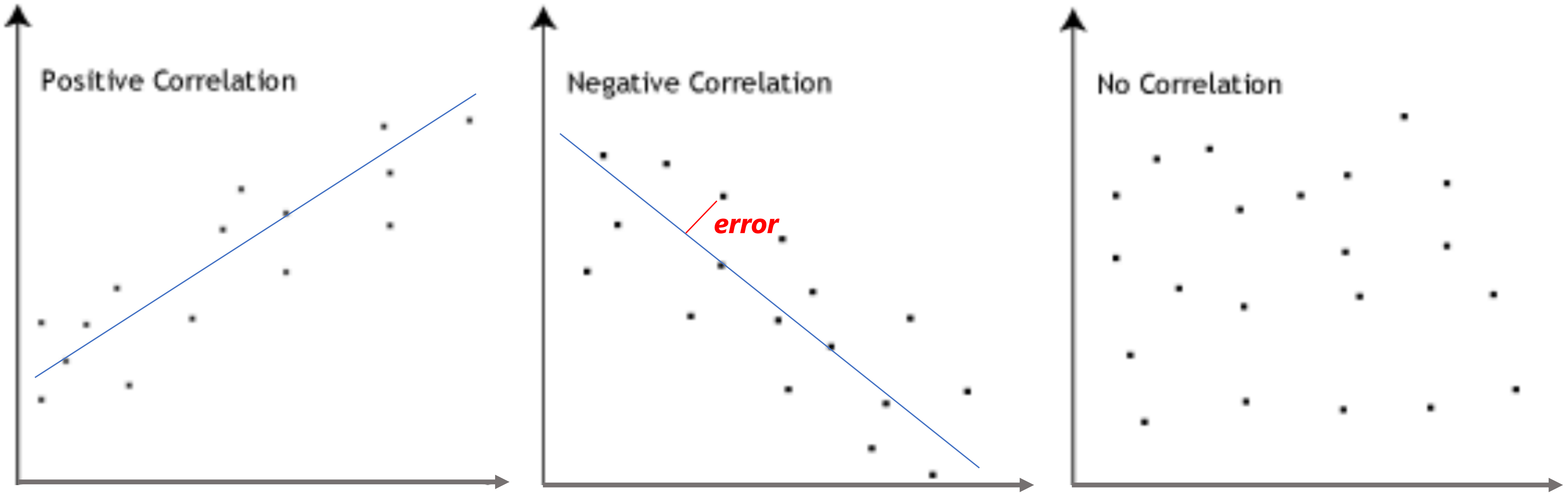
Kurtosis > 3 : leptokurtic (heavy-tailed)
Kurtosis == 0 : mesokurtic
Kurtosis < 3 : platykurtc (light-tailed)



3

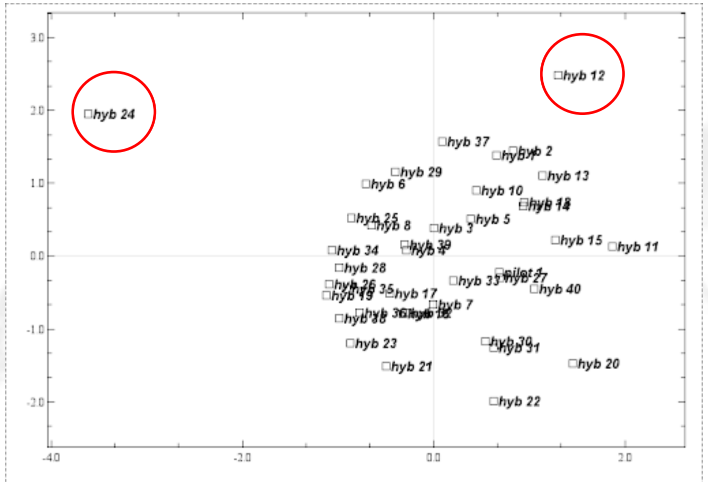
Correlation

Mutual relation / association between quantitative variable. It can help us to predicting one the quantity of x variable from y variable



Outlier

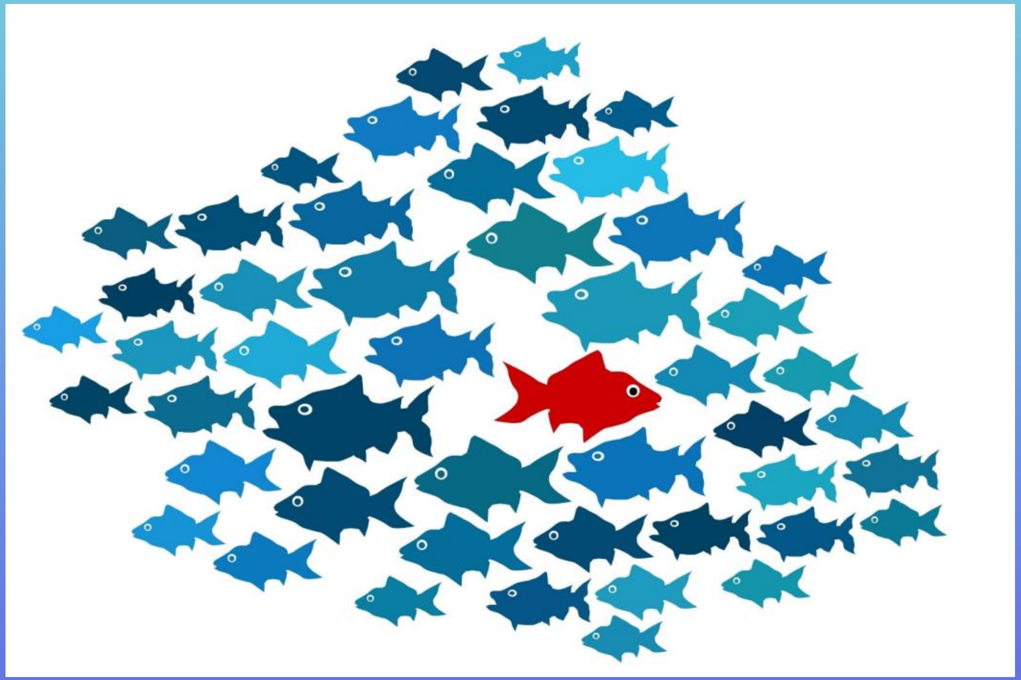
A data point that differs significantly from other observations.



Handle outliers using **IQR** method. Set the outlier value with higher outlier or lower outlier.

Anomalies

An *anomaly* is something that is unusual or unexpected; an abnormality.

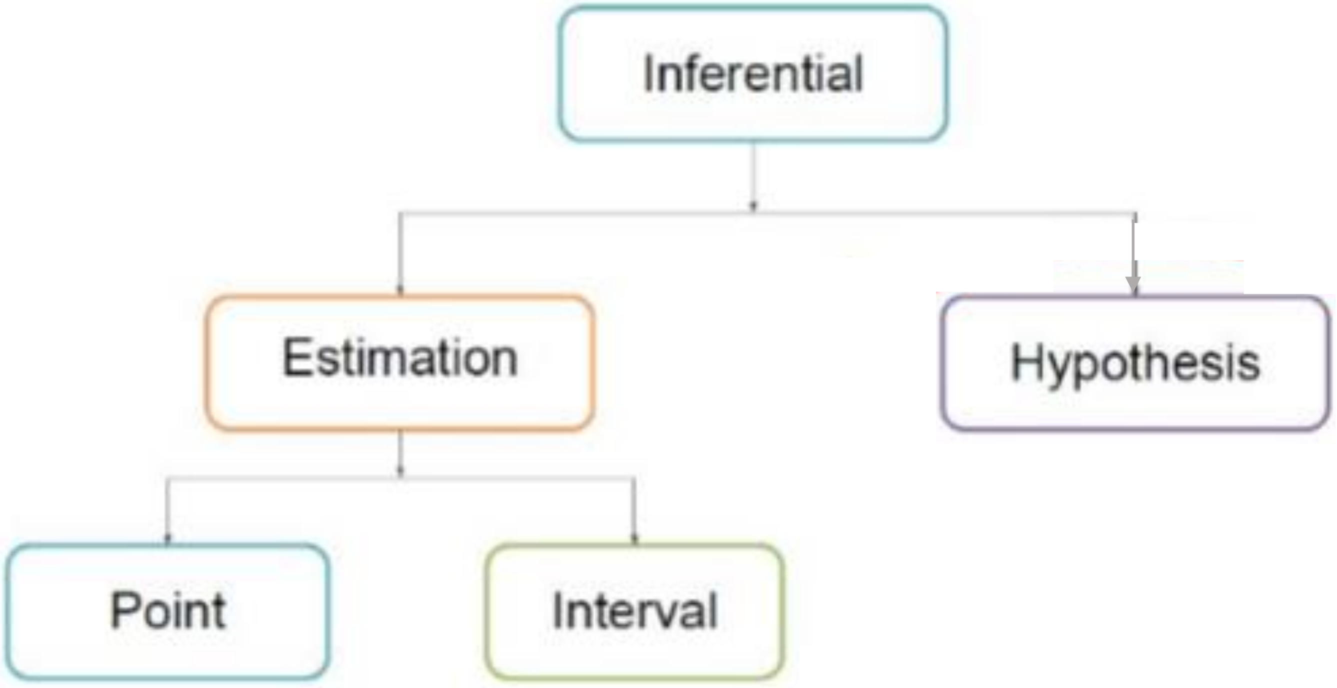


Handle anomaly data by **deleting** that data or replace the anomaly values by the **median** of data.

Inferential Statistics

Inferring the characteristics of a population when only a sample is given.

- ***Population*** *: all Banyuwangi's residents*
- ***Sample*** *: more than 25 years*



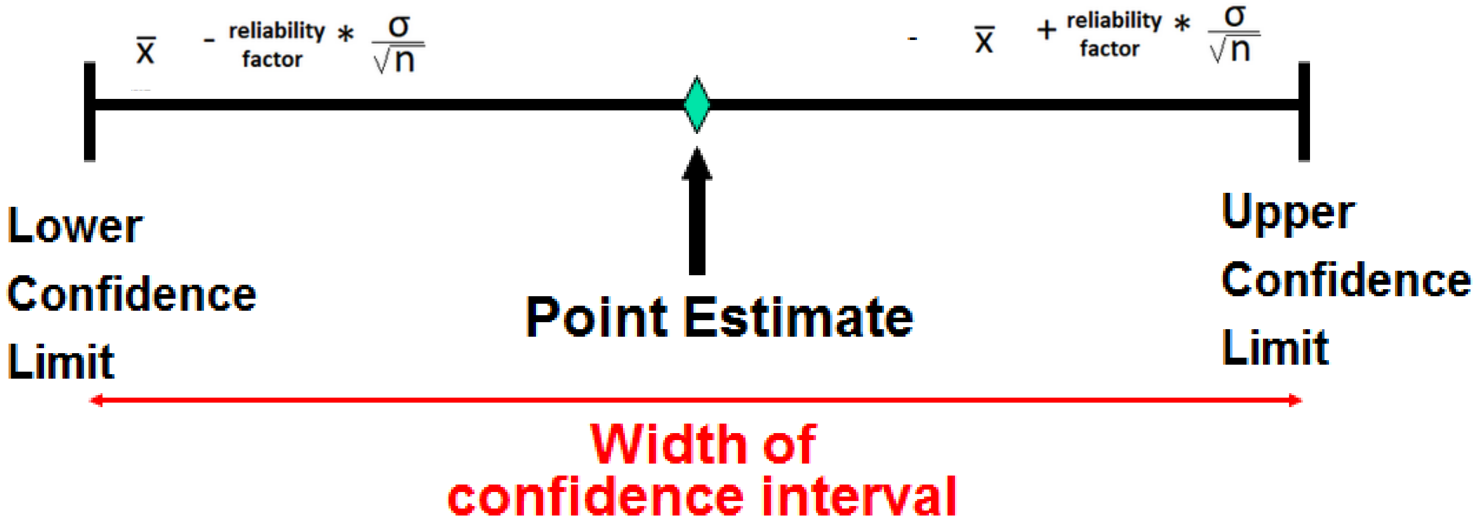
ESTIMATION

Point of Estimation is a ***single value used to estimate the population parameter***. For example, the sample mean x is a point estimate of the population **mean** µ

Data science salaries in Jakarta are around **8 million** / month

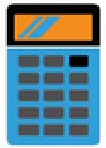
Interval Estimation is ***constructed around the point estimate***, and it is stated that this interval is likely to contain the corresponding population parameter

Data science salaries in Jakarta are around **6.5 - 9 million** / month



Steps of Confident Interval Estimation

1. *Find the **number of sample** and **point estimation** from the sample*
2. *Define **confident level** and table (Z (if sample more than 30) and t (sample less than 30))*
3. *Formula to get confident interval = point estimation ± margin of error*



$$\text{Confidence Interval} = \bar{X} \pm Z \times \frac{\sigma}{\sqrt{n}}$$

Case

Suppose we want to estimate average weight of an adult man in Jakarta. We take a random sample 1,000 men from a population of 1,000,000. We find that the average man in our sample weighs 75 kg, and the standard deviation of the sample is 20 kg. What is the 95% confidence interval.

Answer

N = 1000

Mean = 75

Confident level 95%

Z = 1 – alfa/2

= 1 – (1 - CI) / 2

= 1 – (1-0.95) / 2

= **0.975**

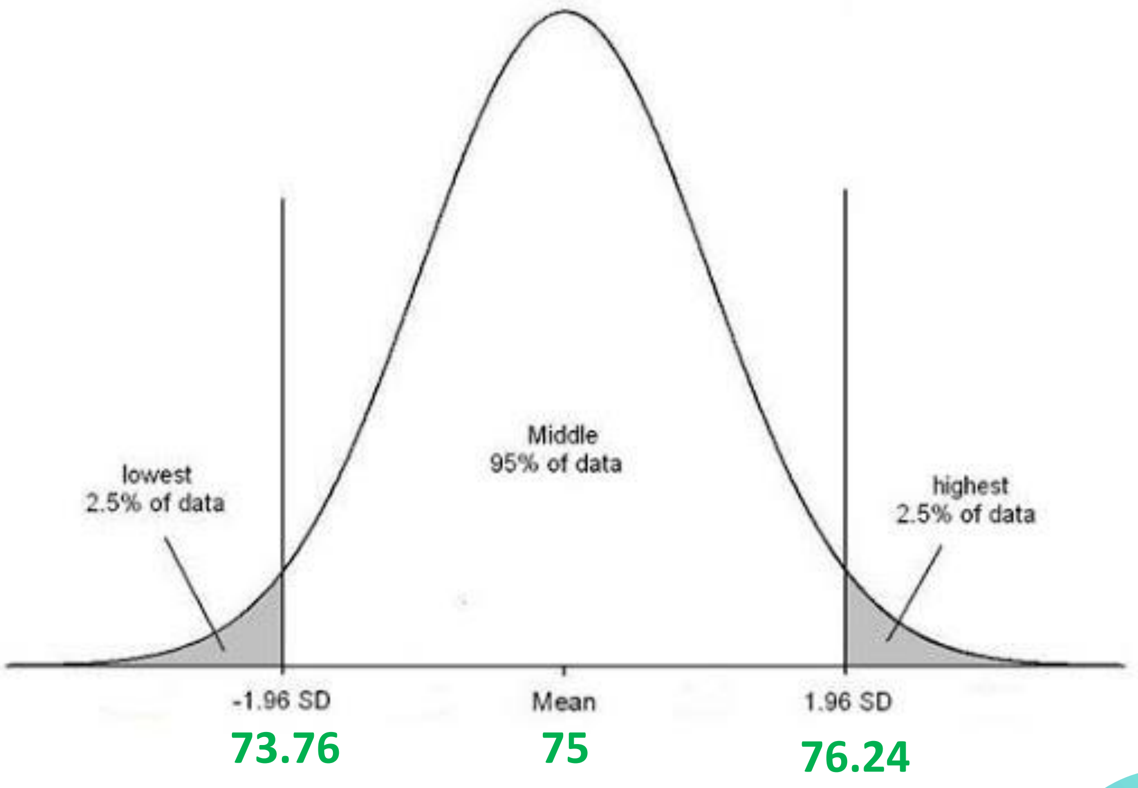
Z(0.975) = **1.96**

Confident Interval Estimation Steps

Find standard error = s / sqrt (n)
= 20 / sqrt(1000)
= 20 / 31.62
= **0.633**

Margin of error = Zval * SE
= 1.96 * 0.633
= **1.24**

So the confident interval estimation is **75 ± 1.24**
[73.76, 76.24]



HYPOTHESIS

Hypothesis are assumptions or statements about parameters. There are 2 type of hypothesis which has impact on z / t table.

- One side : x < 50 or x >= 50 (use : more than or less than word)
- Two side : 30 < x <70 (use : different word)

Hypothesis formulation:

Null Hypothesis (H0) : sample observations result purely from chance.

Alternative Hypothesis (H1/Ha) : is the hypothesis that sample observations influenced by some non - random cause

Possible Outcome:

1. Reject H0, means accept Ha
2. Accept H0

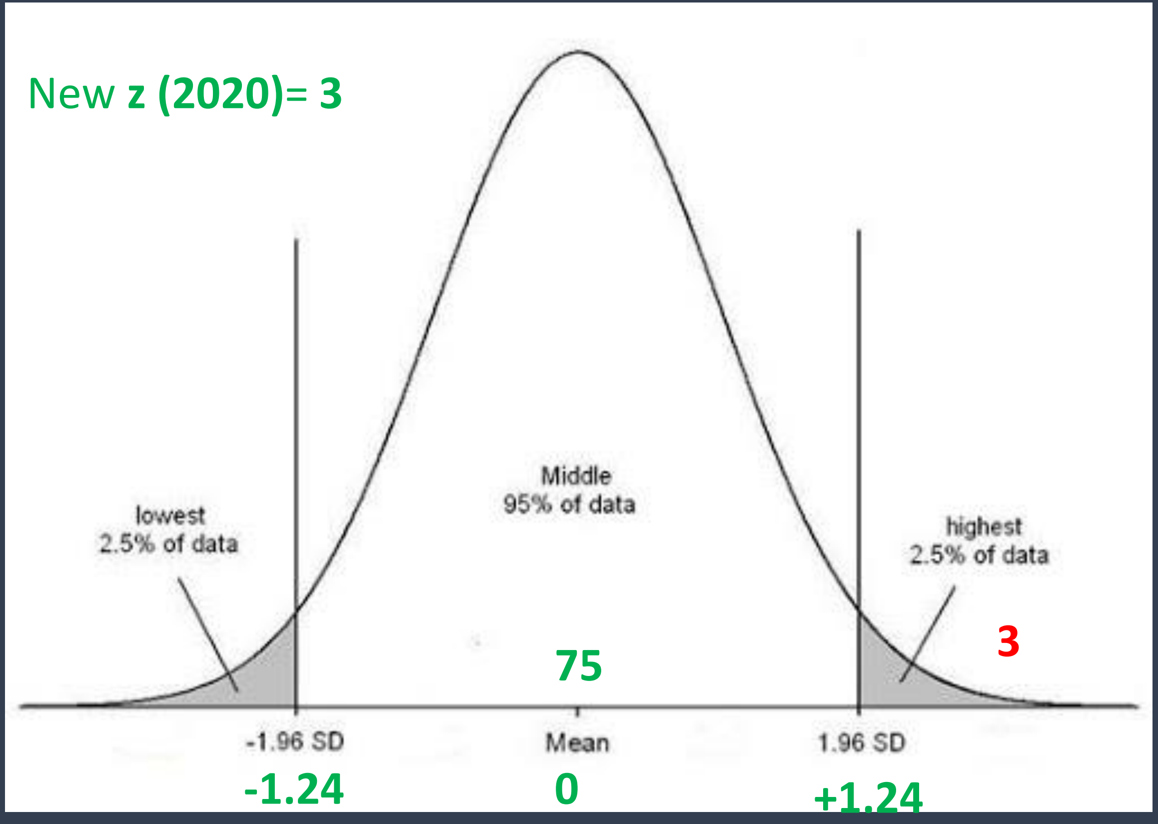
HYPOTHESIS

Steps for hypothesis testing with critical value:

- Set H0 and Ha
- Select the distribution to use
- Determining the rejection and non rejection regions
- Calculate the value of test statistic
- Make a decision

With the same question as before, add with this condition
Mean weight 2015 : 75kg
Mean weight 2020 : 67kg
can you conclude that the mean weight is different from 75kg in 2020?

So
H0 : 75
Ha != 75



Probability

Chance that something will happen or how likely it is that some event will happen

Probability

Data science use statistical inference to analyze trend from the data. While statistical inference uses probability distribution of data.

Simple Probability Formula

$$P(A) = \frac{Number\ of\ Favourable\ Outcome}{Total\ Number\ of\ Favourable\ Outcomes}$$

Example

The chances of the number 5 appearing on the dice.
P(5) = 1 / 6

Conditional Probability

the probability of an event given that another event has occurred.

Conditional Probability Formula

$$\underbrace{P(A|B)}_{\text{Probability of } A \text{ given } B} = \frac{\overbrace{P(A \cap B)}^{\text{Probability of } A \text{ and } B}}{\underbrace{P(B)}_{\text{Probability of } B}}$$

Example

what is the probability that the total of two dice will be greater than 7 given that the first dice is a 4?

P(B) = 6/36 = 1/6
P (A slice B) = 3/36
Result: = (3/36)/(1/6)
= **1/2**

	Die #2					
Die #1	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

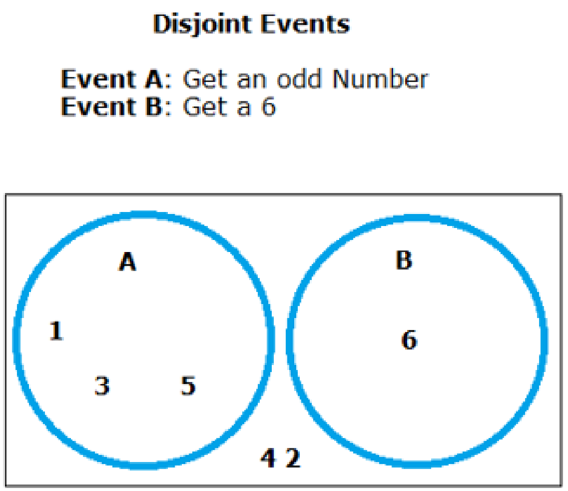
Disjoint and Overlapping Event

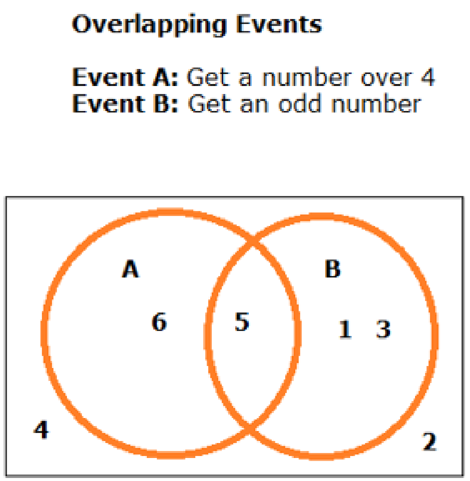
Disjoint

Events can not happen at the same time (mutually exclusive). Then the rule
P(A or B): P(A) + P(B)

Overlapping

Events that have outcomes in common. Then the rule
P(A or B): P(A) + P(B) – P(A and B)





Example

We have 9 balls with 3 colors

1 2 3
4 5 6
7 8 9

1. The chances of being selected are red or yellow balls

P(A or B) = P(A) + P(B)
= 3/9 + 1/9
= 4/9

P(A)
6 7 8

P(B)
9

2. The chances of being selected are have odd numbers or yellow balls

P(A or B) = P(A) + P(B) – P(A and B)
= 5/9 + 1/9 – 1/9
= 5/9

P(A)
1 3 5 7 9

P(B)
9

We don't have 2 yellow balls and the number is 9
Delete 1

Probability Distribution

List of possible event and the probabilities which they occur.
The rule is :

- *If the probabilities add, the result must be 1*
- *Must be mutually exclusive*
- *Each probability should be between 0 and 1*

Two dice are rolled at the same time

		Die #2					
		1	2	3	4	5	6
Die #1	1	2	3	4	5	6	7
	2	3	4	5	6	7	8
	3	4	5	6	7	8	9
	4	5	6	7	8	9	10
	5	6	7	8	9	10	11
	6	7	8	9	10	11	12

The distribution of the odds of the number of dice appearing

Sum	2	3	4	5	6	7
Probability	0.0278	0.0556	0.0833	0.1111	0.1389	0.1667

Sum	8	9	10	11	12
Probability	0.1389	0.1111	0.0833	0.0556	0.0278

Thank You

Don’t forget to say Alhamdulillah for today

Reference

- https://towardsdatascience.com/intro-to-descriptive-statistics-252e9c464ac9
- https://drive.google.com/file/d/1odxj-YkI3Uypdb7xBDP-enjDtgLtlMBj/view?usp=sharing
- https://www.khanacademy.org/math/probability/probability-geometry/probability-basics/a/probability-the-basics
- https://drive.google.com/file/d/19XJ8krutndXn50iNZdiVbDsrjAmhVELH/view?usp=sharing
- https://drive.google.com/file/d/1c3P8rbMUyv5ar_K-mipmzBP3WswPHp6B/view?usp=sharing
- https://www.statisticshowto.com/probability-and-statistics/hypothesis-testing/