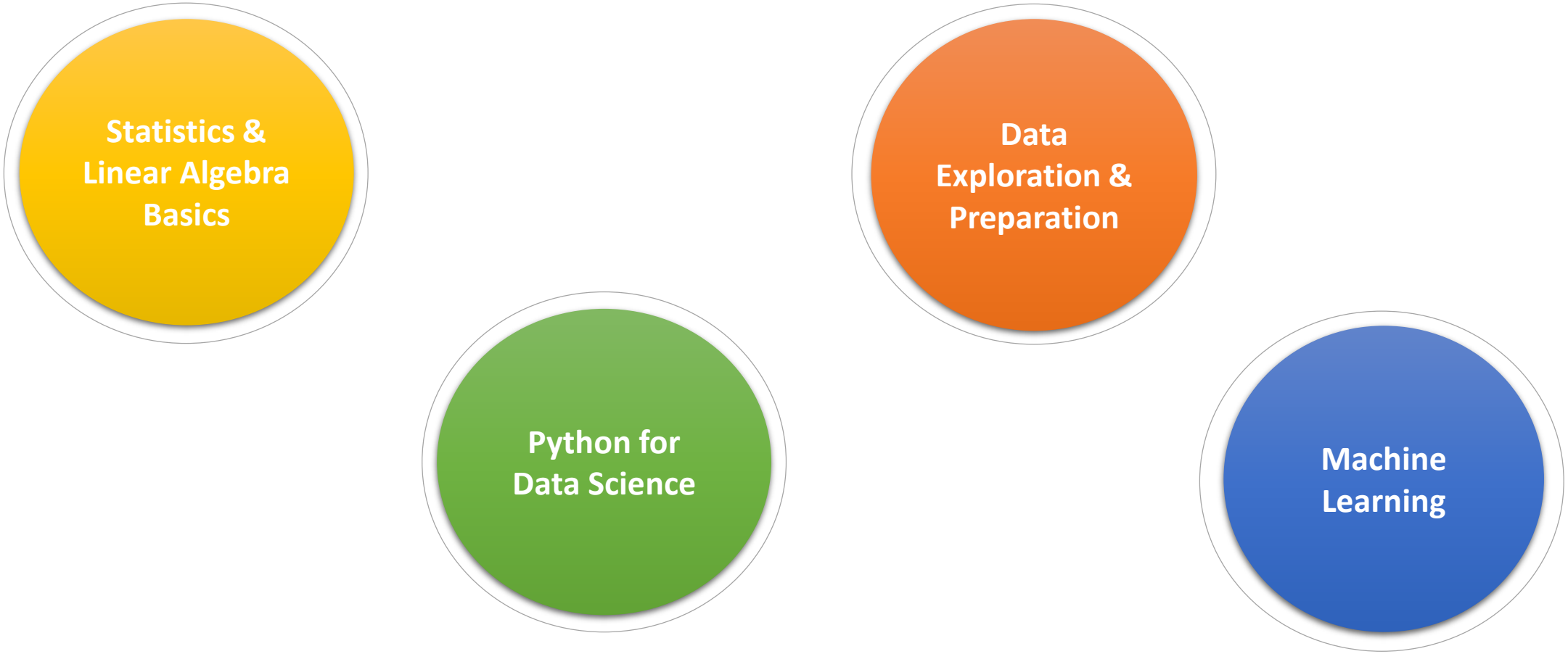
Data Science & Machine Learning

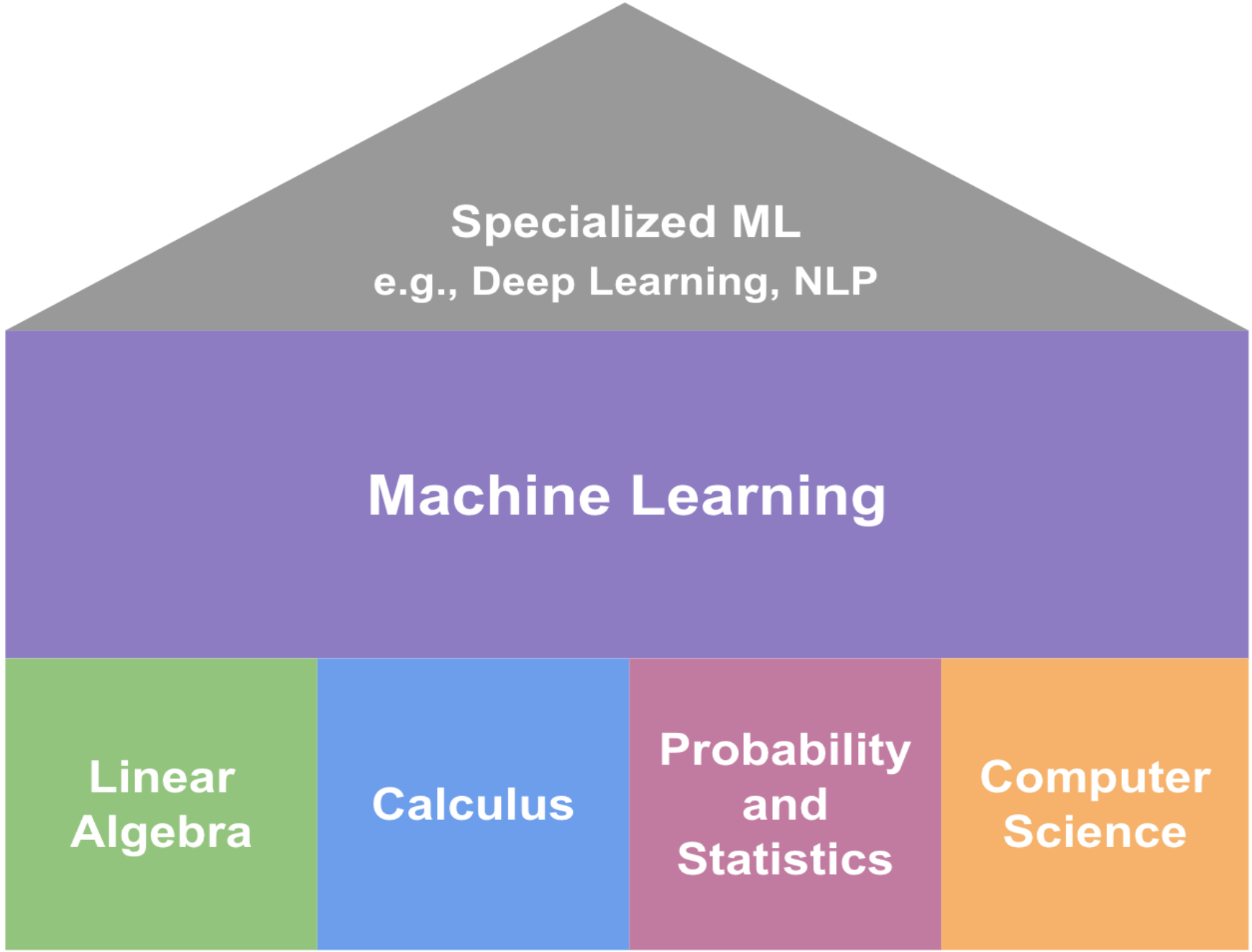
Eman Raslan

Course Schedule

Week #	Day1	Day2
1	24-Nov-2023	
2	1-Dec-2023	2-Dec-2023
3	8-Dec-2023	9-Dec-2023
4	15-Dec-2023	16-Dec-2023
5	22-Dec-2023	23-Dec-2023
6	29-Dec-2023	30-Dec-2023
7	5-Jan-2023	6-Jan-2023
8	12-Jan-2023	13-Jan-2023
	19-Jan-2023	
Project		

Course Agenda

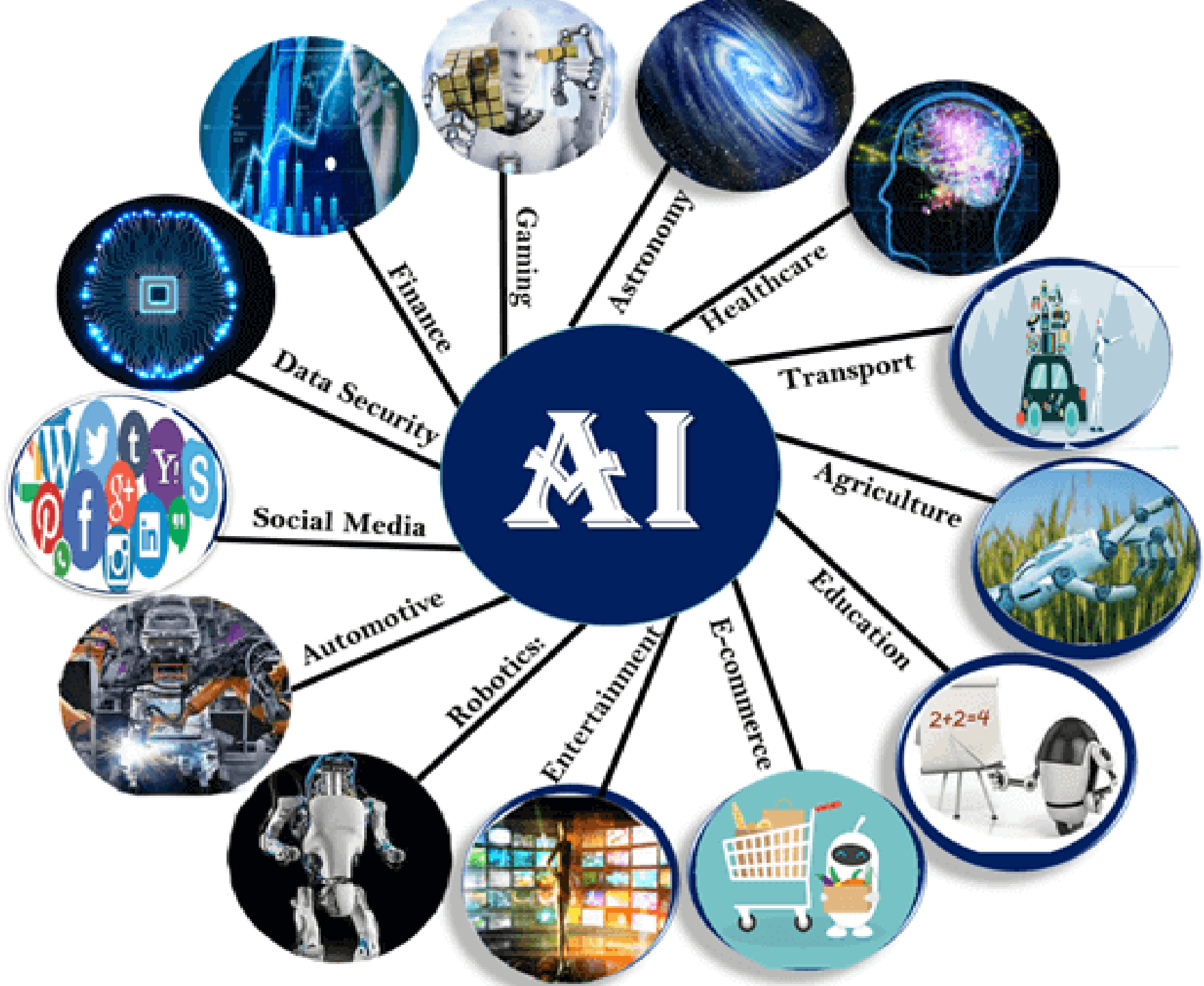



Specialized ML
e.g., Deep Learning, NLP
Machine Learning
Linear Algebra
Calculus
Probability and Statistics
Computer Science
JonKrohn.com

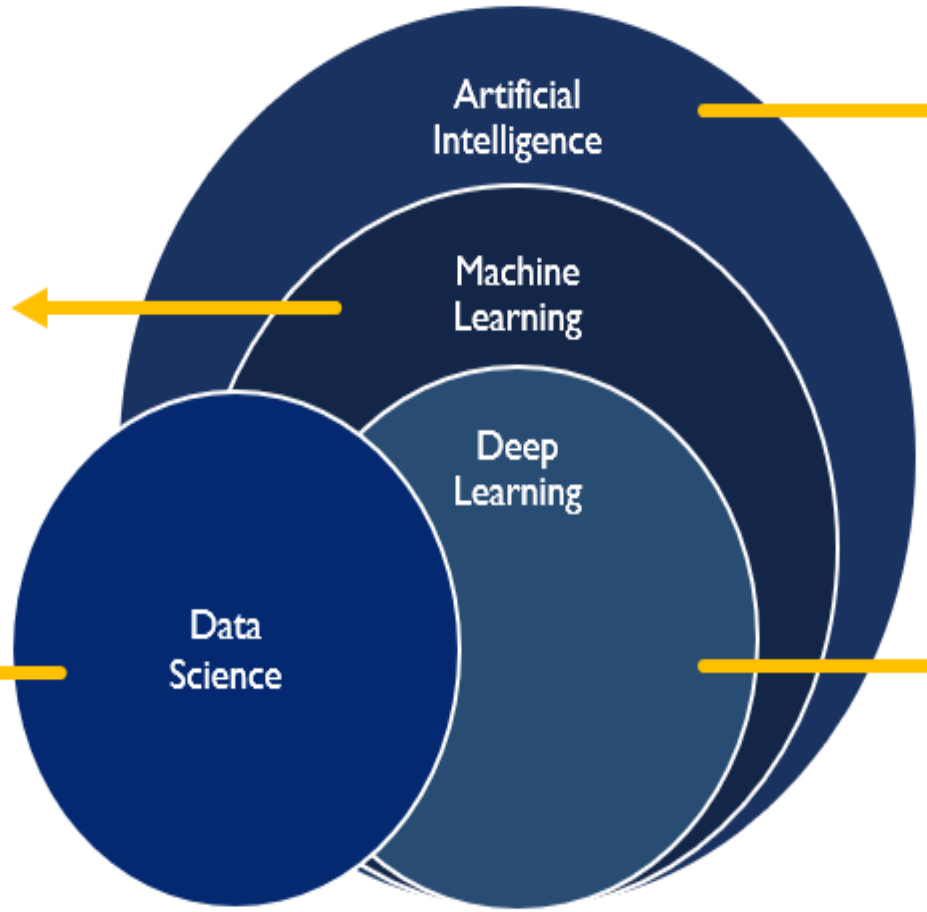
What is AI?

Types of Artificial Intelligence

Narrow AI	General AI
○ Application specific/ task limited	○ Perform general (human) intelligent action
○ Fixed domain models provided by programmers	○ Self-learns and reasons with its operating environment
○ Learns from thousands of labeled examples	○ Learns from few examples and/or from unstructured data
○ Reflexive tasks with no understanding	○ Full range of human cognitive abilities
○ Knowledge does not transfer to other domains or tasks	○ Leverages knowledge transfer to new domains and tasks
○ Today's AI	○ Future AI?


AI
Finance
Gaming
Astronomy
Healthcare
Data Security
Transport
Social Media
Agriculture
Automotive
Education
Robotics
Entertainment
E-commerce
2+2=4

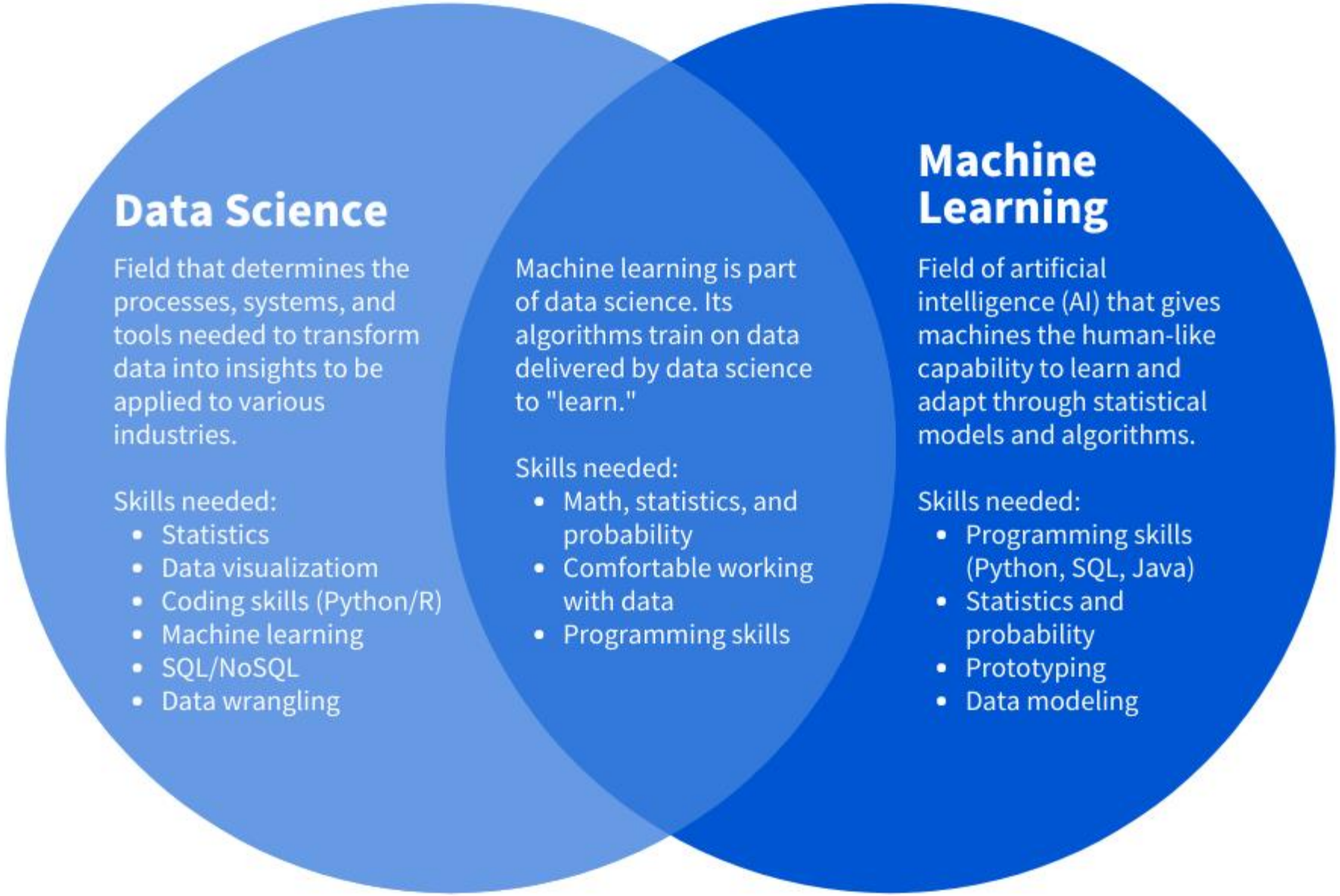
Machine Learning is a subset of AI, and builds a model based on training data to make predictions

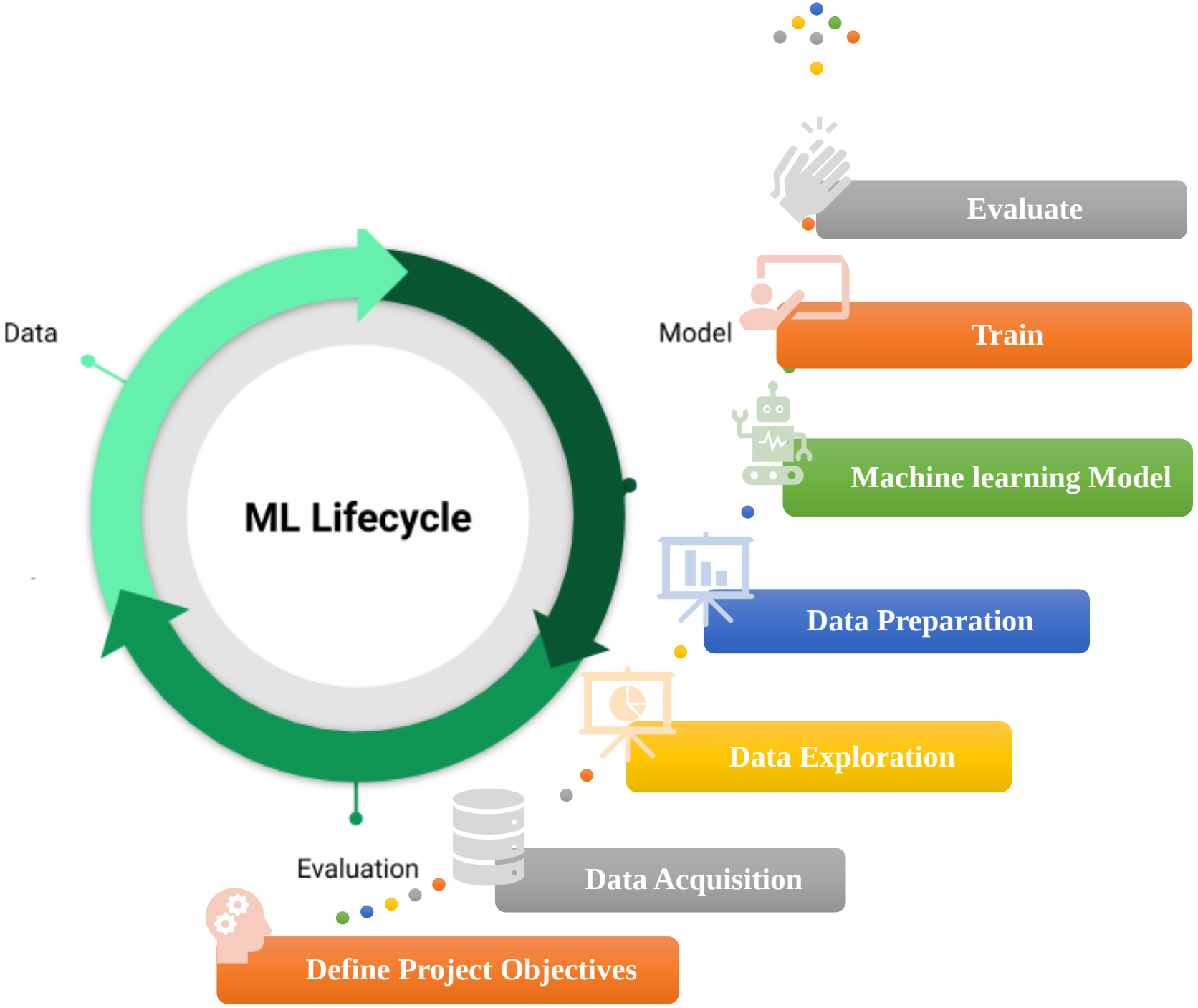


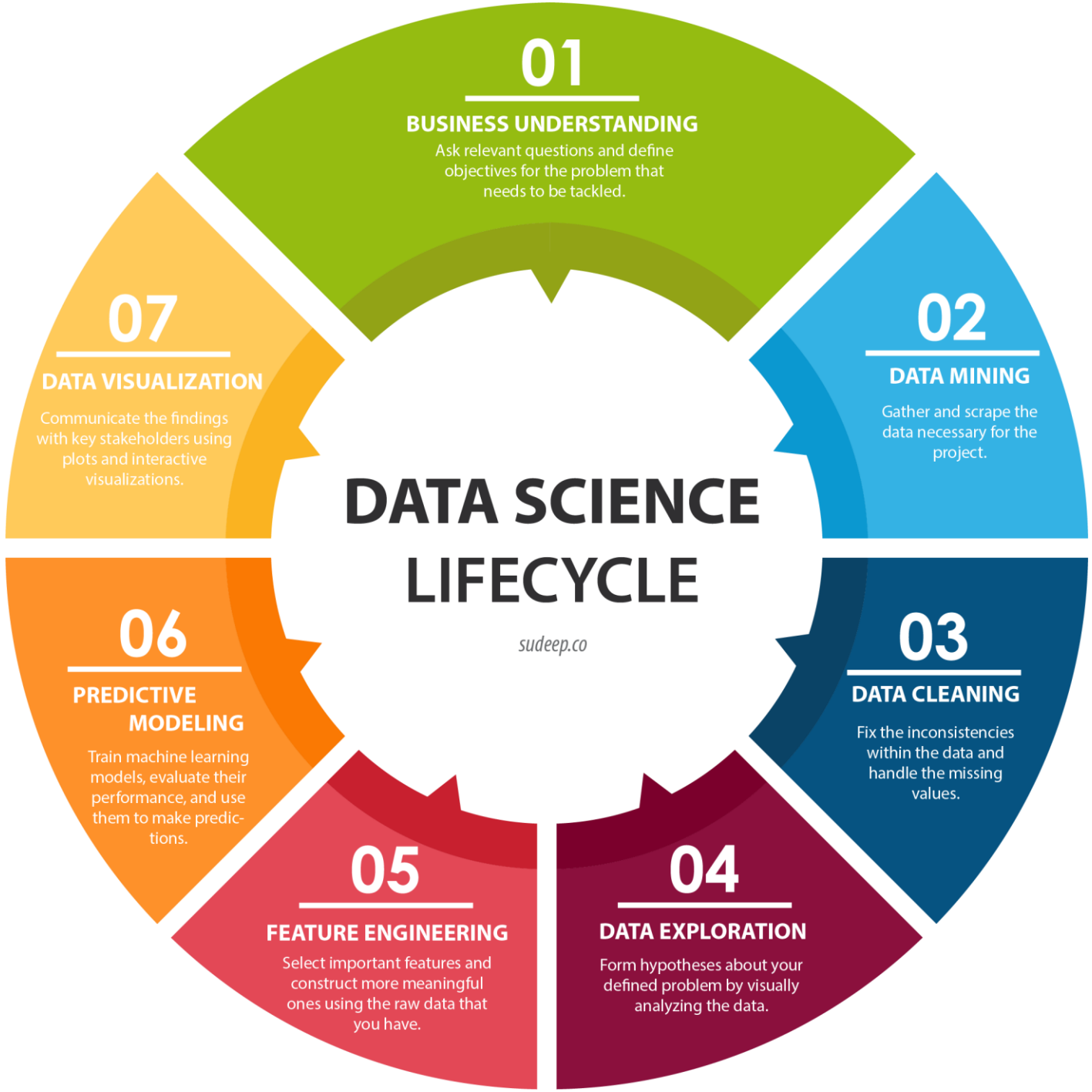
Artificial Intelligence means creating smart machines to mimic human behavior

Data Science is a subset of AI. It is an area of statistics, scientific methods, etc. to extract meaning and insights from data..

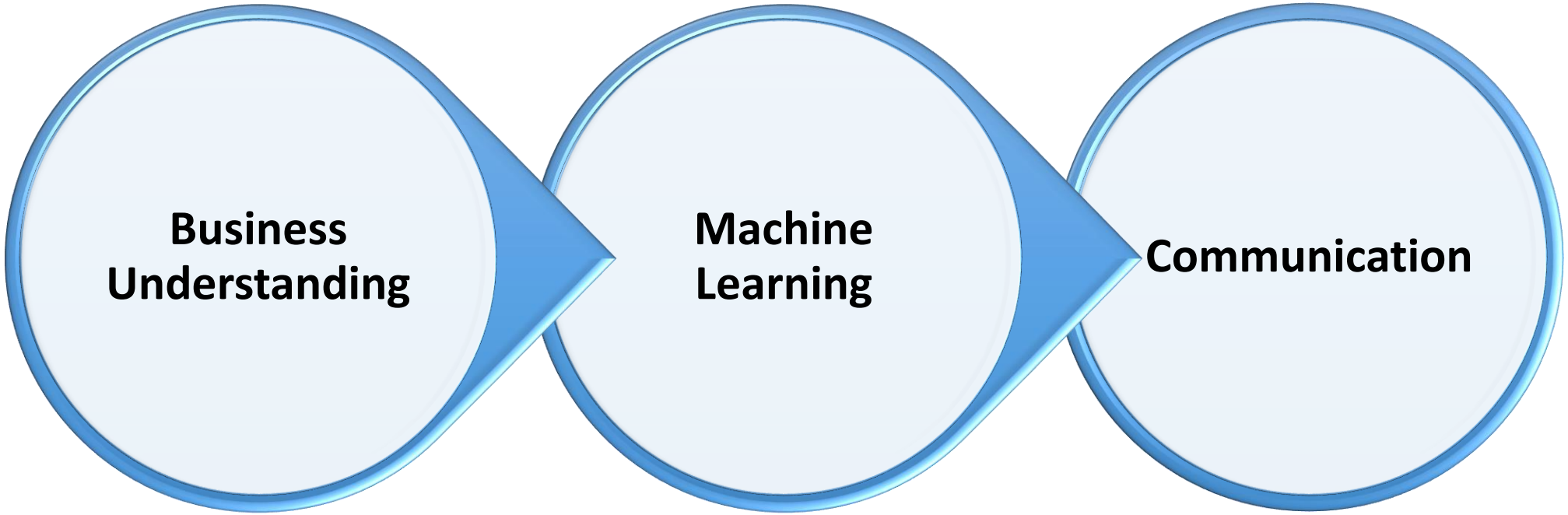
Deep learning is a subset of ML, a class of ML algorithms to solve complex problems.


Data Science
Field that determines the processes, systems, and tools needed to transform data into insights to be applied to various industries.
Skills needed:
• Statistics
• Data visualizatiom
• Coding skills (Python/R)
• Machine learning
• SQL/NoSQL
• Data wrangling
Machine learning is part of data science. Its algorithms train on data delivered by data science to "learn."
Skills needed:
• Math, statistics, and probability
• Comfortable working with data
• Programming skills
Machine Learning
Field of artificial intelligence (AI) that gives machines the human-like capability to learn and adapt through statistical models and algorithms.
Skills needed:
• Programming skills (Python, SQL, Java)
• Statistics and probability
• Prototyping
• Data modeling


Data
Model
ML Lifecycle
Evaluation
Evaluate
Train
Machine learning Model
Data Preparation
Data Exploration
Data Acquisition
Define Project Objectives


01
BUSINESS UNDERSTANDING
Ask relevant questions and define objectives for the problem that needs to be tackled.
02
DATA MINING
Gather and scrape the data necessary for the project.
03
DATA CLEANING
Fix the inconsistencies within the data and handle the missing values.
04
DATA EXPLORATION
Form hypotheses about your defined problem by visually analyzing the data.
05
FEATURE ENGINEERING
Select important features and construct more meaningful ones using the raw data that you have.
06
PREDICTIVE MODELING
Train machine learning models, evaluate their performance, and use them to make predictions.
07
DATA VISUALIZATION
Communicate the findings with key stakeholders using plots and interactive visualizations.
DATA SCIENCE LIFECYCLE
sudeep.co

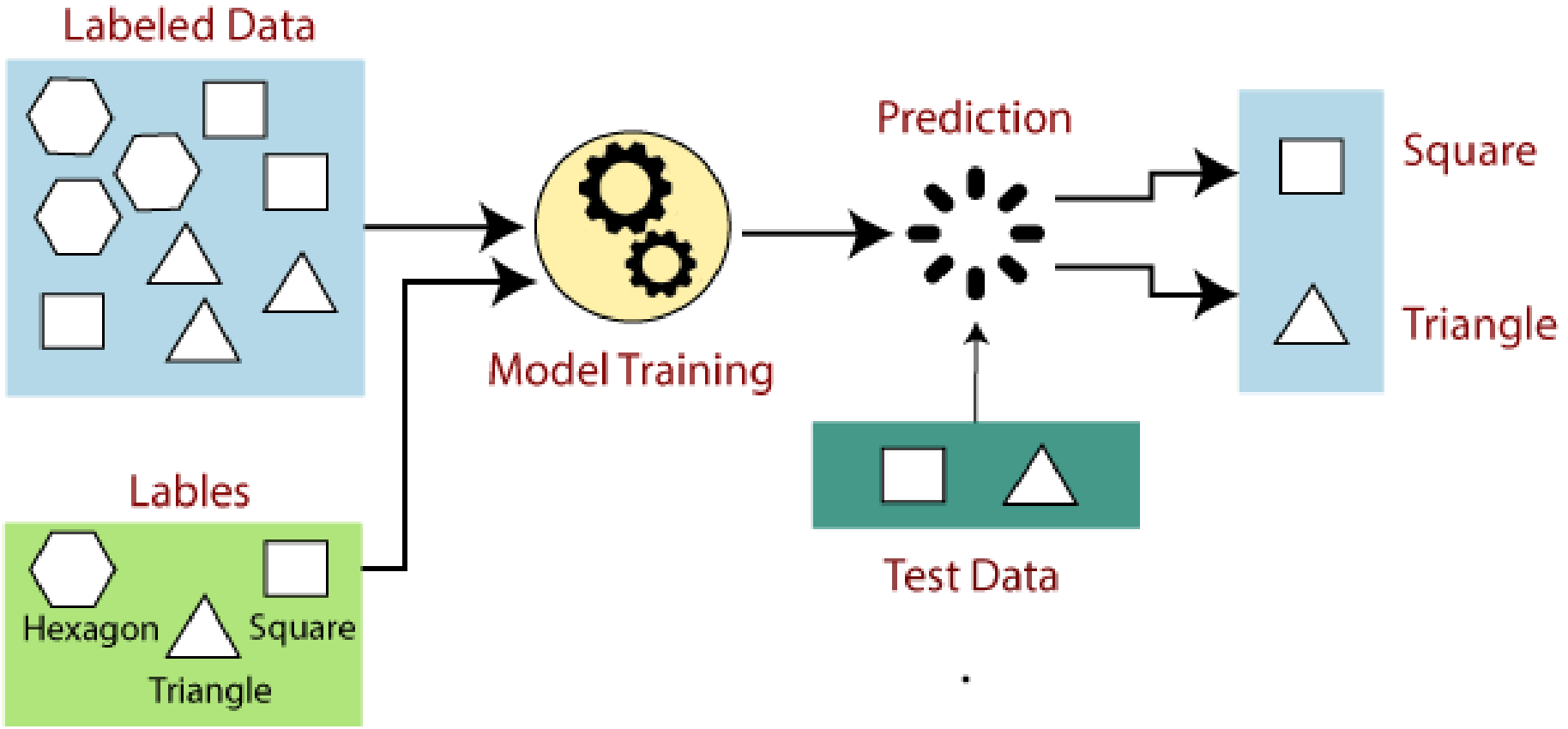
Data Science Process



Types of ML Systems

- Supervised learning
- Unsupervised learning
- Reinforcement learning
- Generative AI

Supervised Learning



Supervised Learning

enjoy algorithms

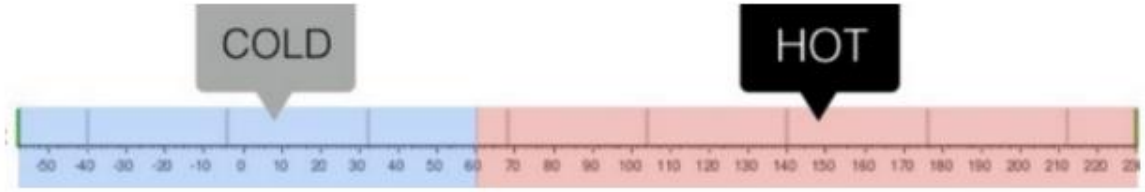
Regression

What will be the temperature tomorrow?

Fahrenheit

Classification

Will it be hot or cold tomorrow?

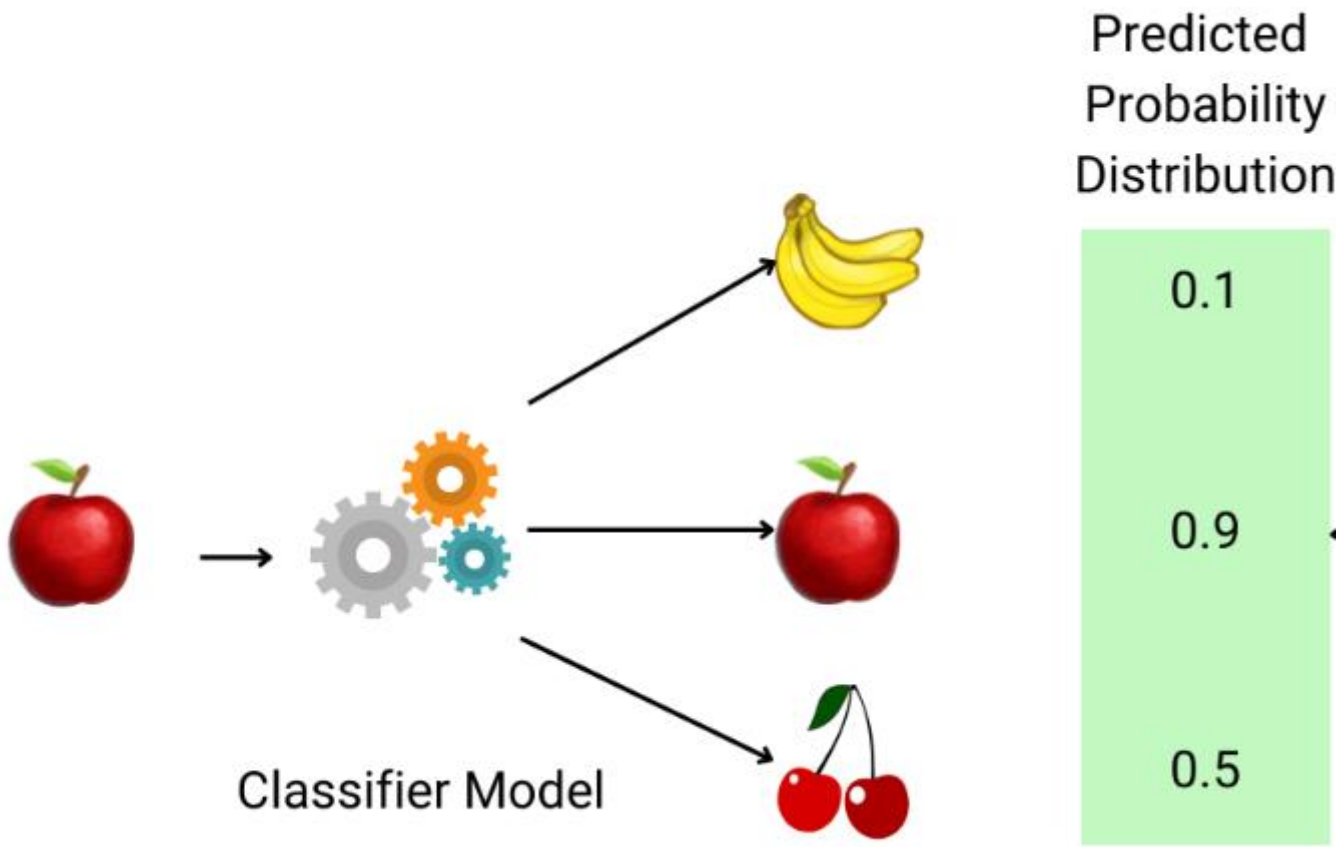


Fahrenheit

Regression

Scenario	Possible input data	Numeric prediction
Future house price	Square footage, zip code, number of bedrooms and bathrooms, lot size, mortgage interest rate, property tax rate, construction costs, and number of homes for sale in the area.	The price of the home.
Future ride time	Historical traffic conditions (gathered from smartphones, traffic sensors, ride-hailing and other navigation applications), distance from destination, and weather conditions.	The time in minutes and seconds to arrive at a destination.

Classification



Types of Classification

Binary Classification

Dog 0.9

Not Dog 0.1

Multiclass Classification

Dog 0.5

Cat 0.09

Bus 0.01

Plant 0.4

Multilabel Classification

Dog 0.8

Cat 0.2

Bus 0.04

Plant 0.7

Suppose you had a flight dataset with the following columns:

departure_airport	destination_airport	date	time_of_day	airline	fuel_price	coach_ticket_cost

If you wanted to predict the cost of a coach ticket, would you use regression or classification?

- [ ] Classification
- [ ] Regression

Suppose you had a flight dataset with the following columns:

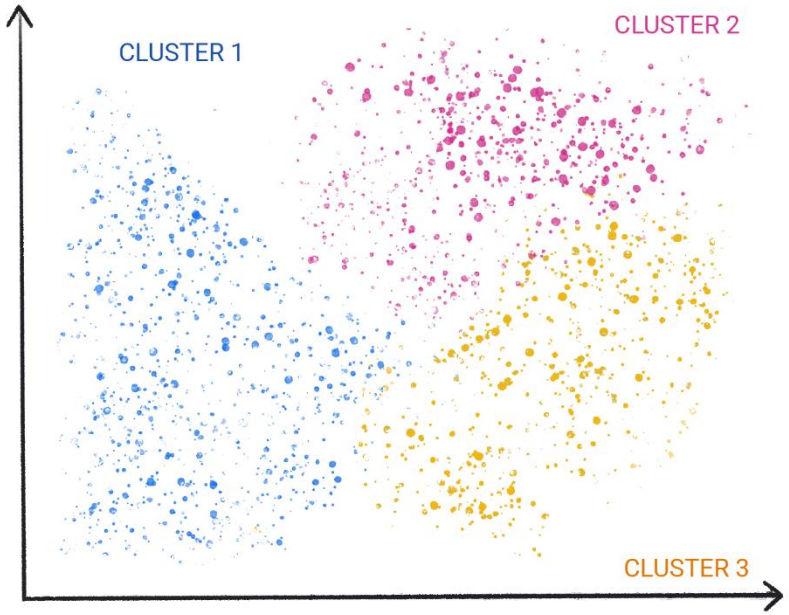
departure_airport	destination_airport	date	time_of_day	airline	fuel_price	coach_ticket_cost

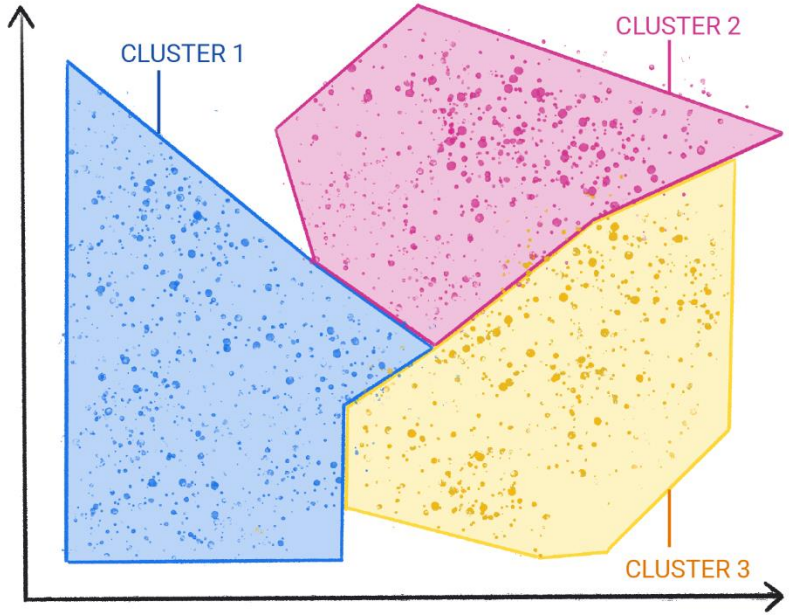
Based on the dataset and a little bit of work, would it be possible to create a model that classified the cost of a coach ticket as "high," "average," or "great deal."?

- [ ] No. It's not possible to create a classification model from the columns in the dataset.
- [ ] Yes. It's possible to create a classification model from this dataset.

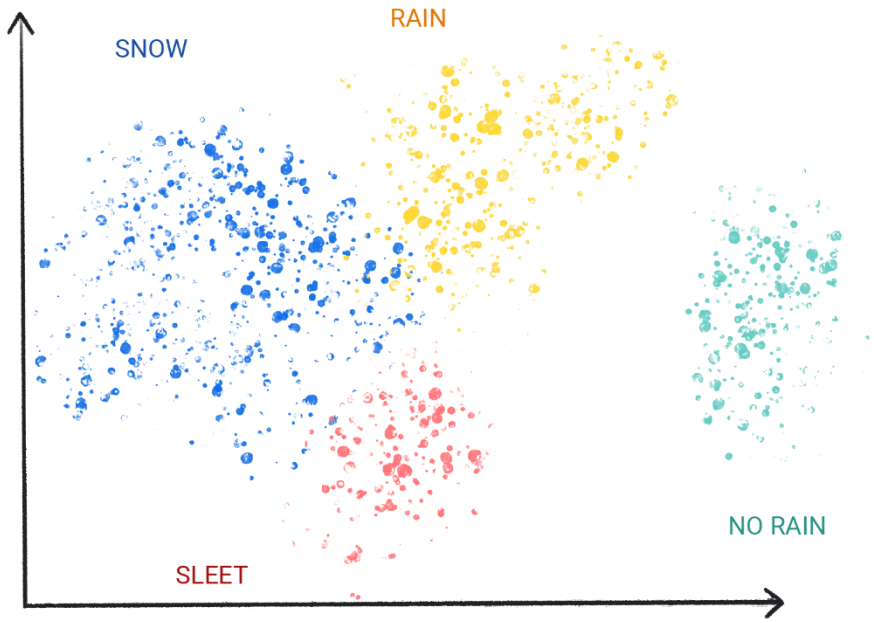
Unsupervised Learning

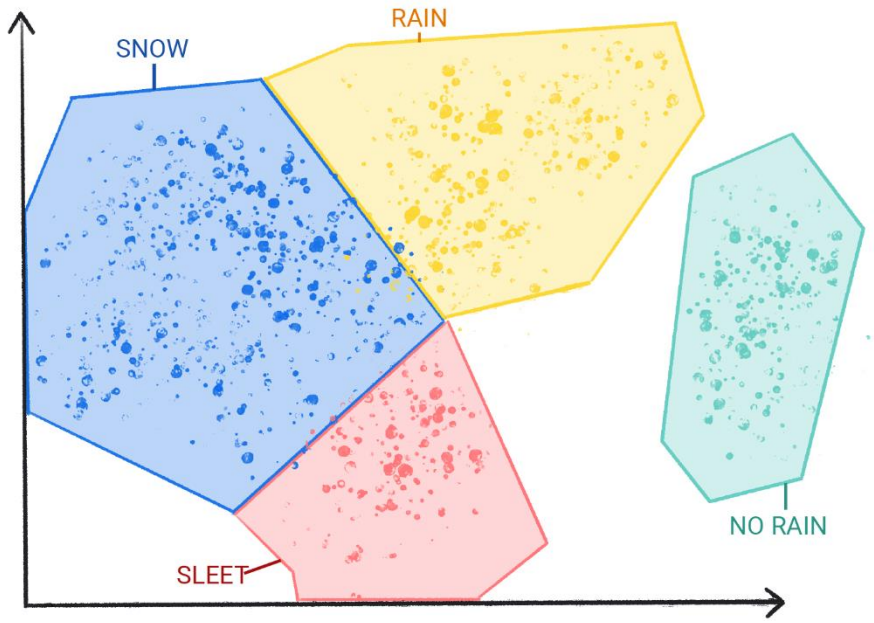
Unsupervised learning





Unsupervised learning





What distinguishes a supervised approach from an unsupervised approach?

- [ ] An unsupervised approach knows how to label clusters of data.
- [ ] A supervised approach is given data that contains the correct answer.
- [ ] A supervised approach typically uses clustering.

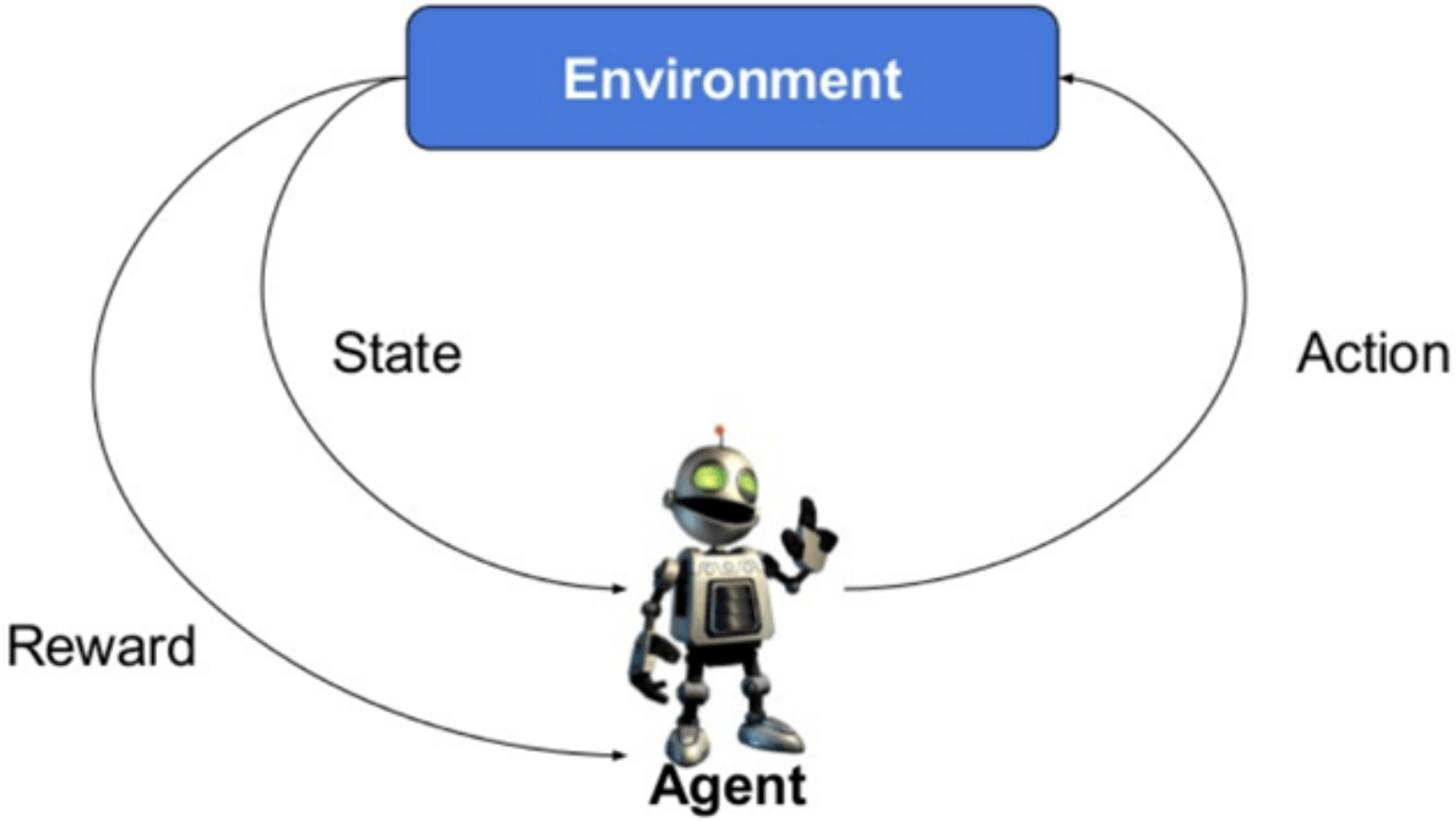
Suppose you had a dataset of users for an online shopping website, and it contained the following columns:

last_purchase_amt	years_since_signup	num_purchases	num_coupon_usage	education	marital_status	income

If you wanted to understand the types of users that visit the site, would you use supervised or unsupervised learning?

- [ ] Supervised learning because I'm trying to predict which class a user belongs to.
- [ ] Unsupervised learning.

Reinforcement learning



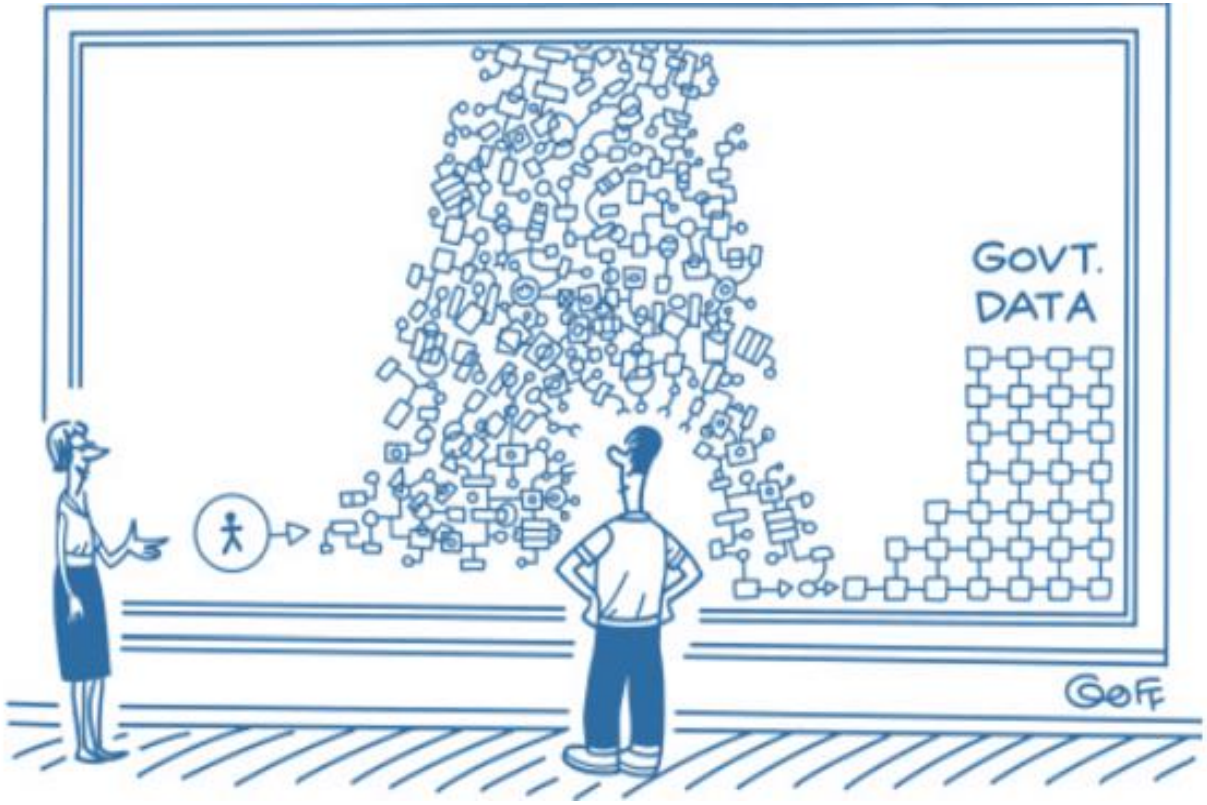
Generative AI

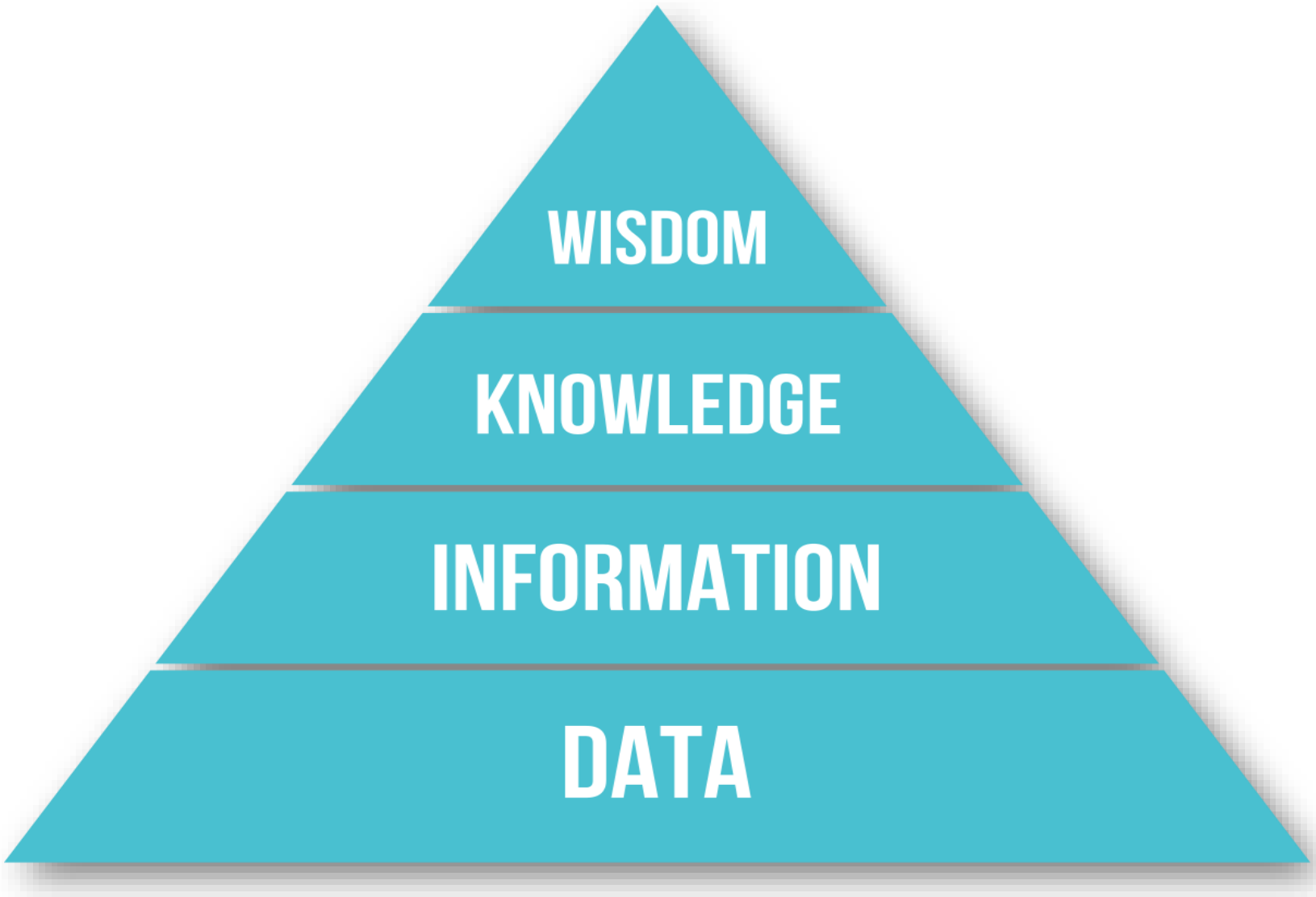
Text-to-image	An alien octopus floats through a portal reading a newspaper.	 Source: Imagen
Text-to-video	A photorealistic teddy bear is swimming in the ocean at San Francisco. The teddy bear goes under water. The teddy bear keeps swimming under the water with colorful fishes. A panda bear is swimming under water.	 Source: Phenaki
Text-to-code	Write a Python loop that loops over a list of numbers and prints the prime numbers.	

```
for number in numbers:
# Check if the number is prime.
is_prime = True
for i in range(2, number):
  if number % i == 0:
      is_prime = False
      break
```

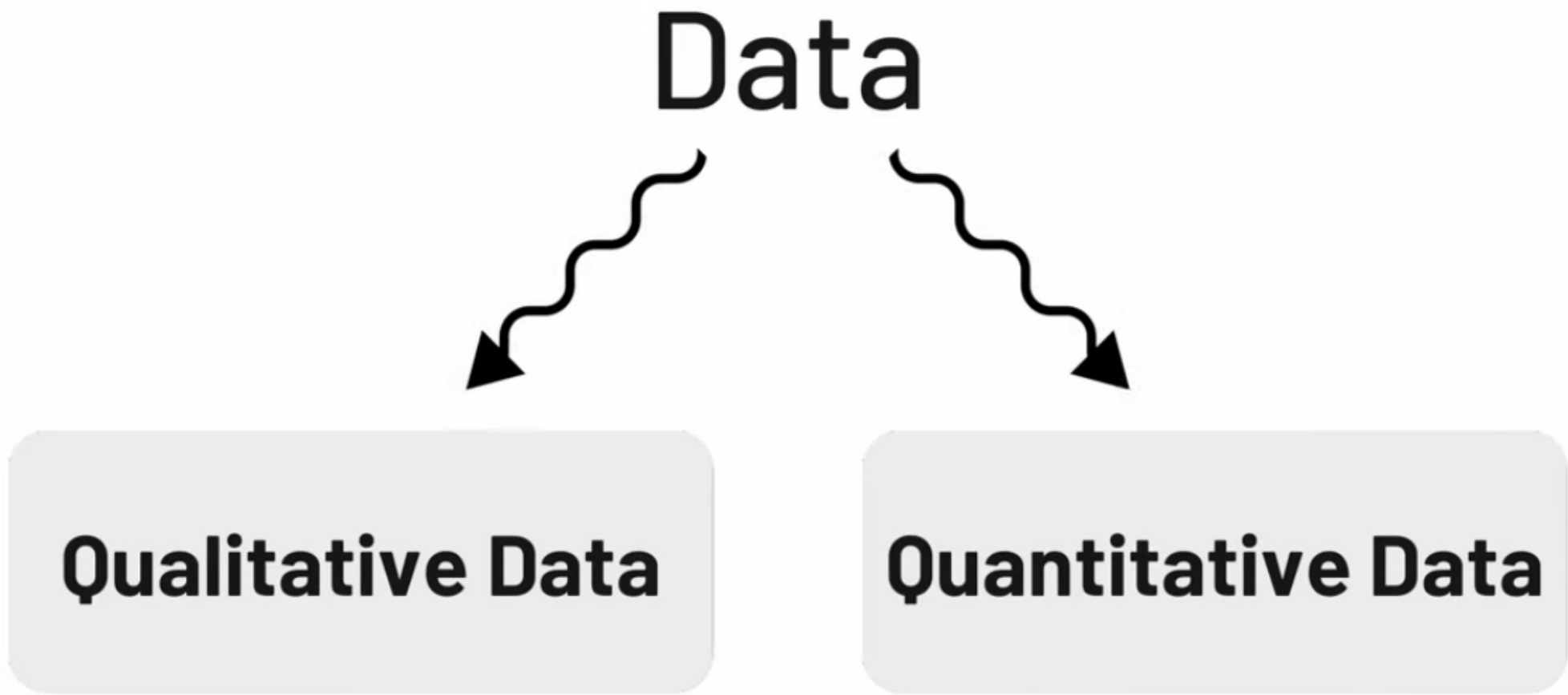
Data

Before trying anything, you should know your data!




WISDOM
KNOWLEDGE
INFORMATION
DATA

Variable Data Types



Qualitative Data (Categorical)

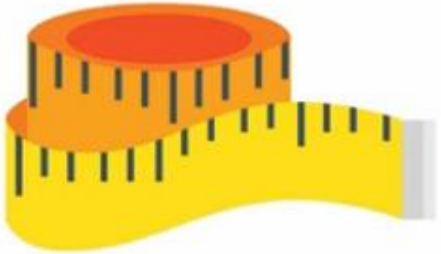
- Cannot be measured in the form of numbers
- Arithmetic operation

Quantitative Data (Numerical)

- Talks about the quantities
- Arithmetic Operations
- Can be measured in the form of numbers
- Measured using Measuring Devices

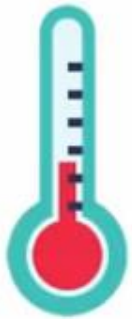
Weighing Machine

Ruler



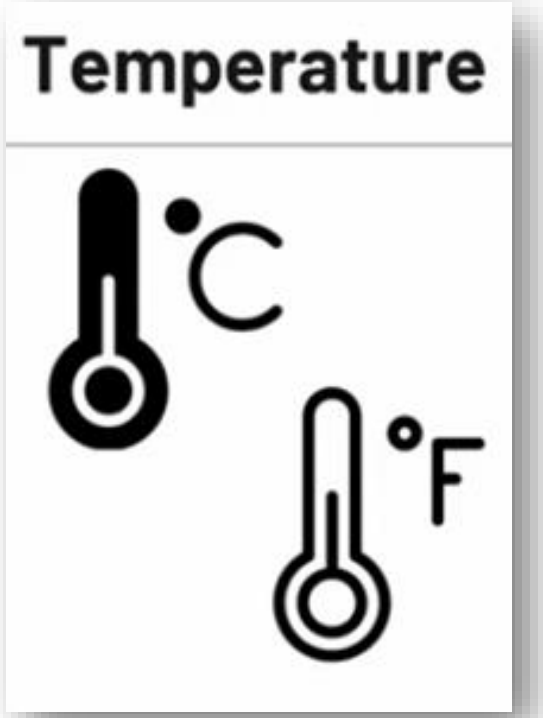
Measuring Tape

Stopwatch

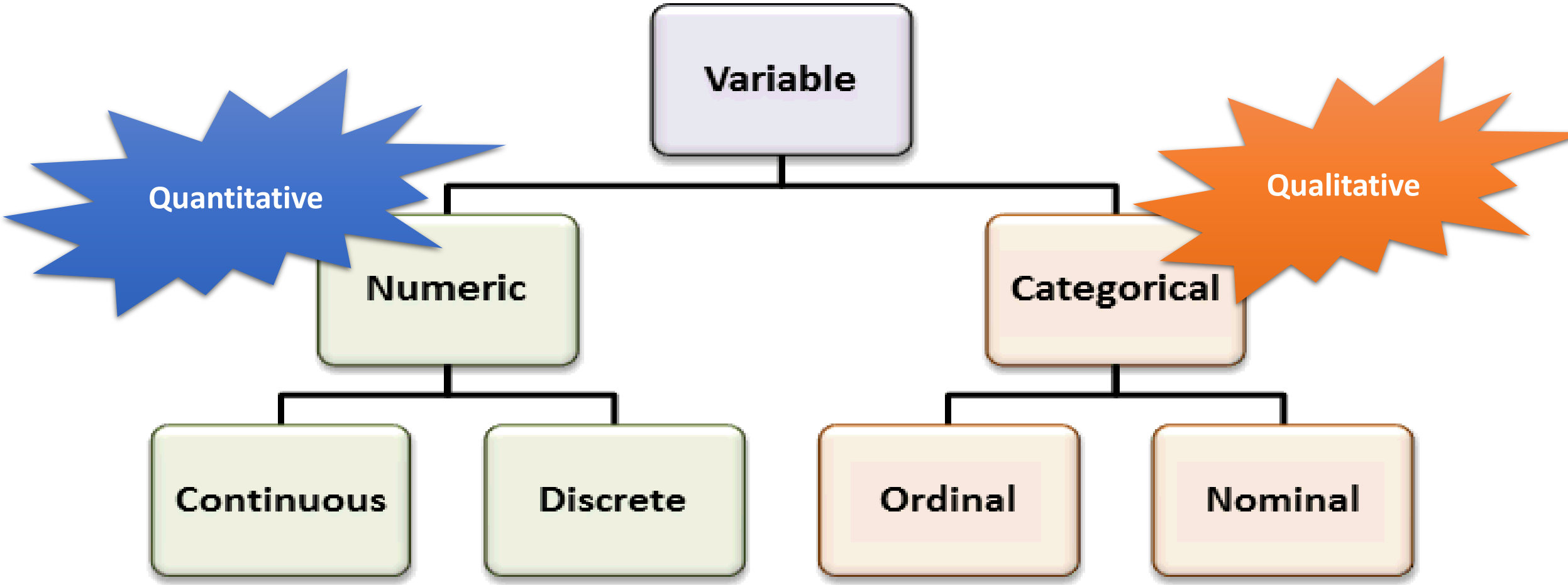


Thermometer

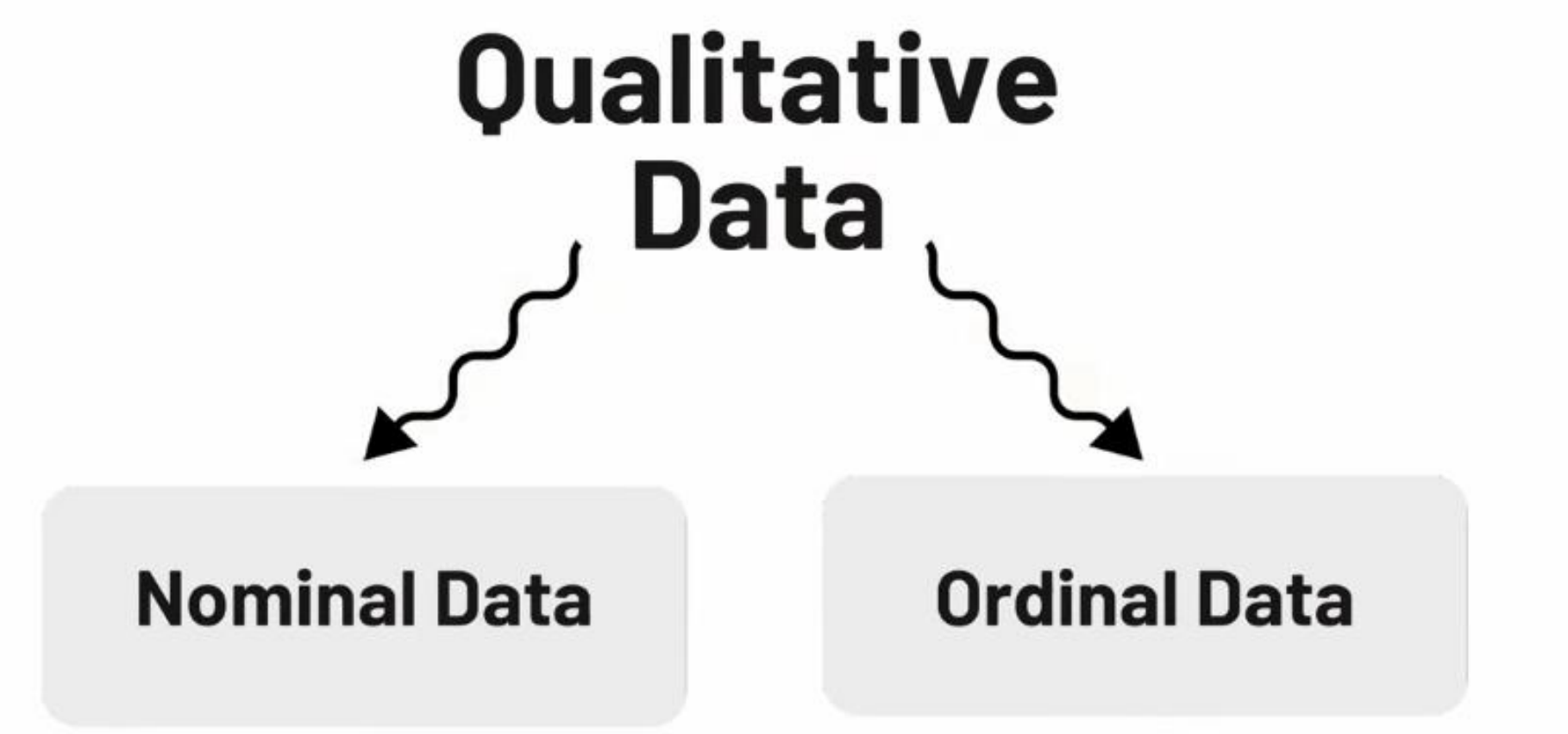
Quantitative Data (Numerical)



Variable Data Types

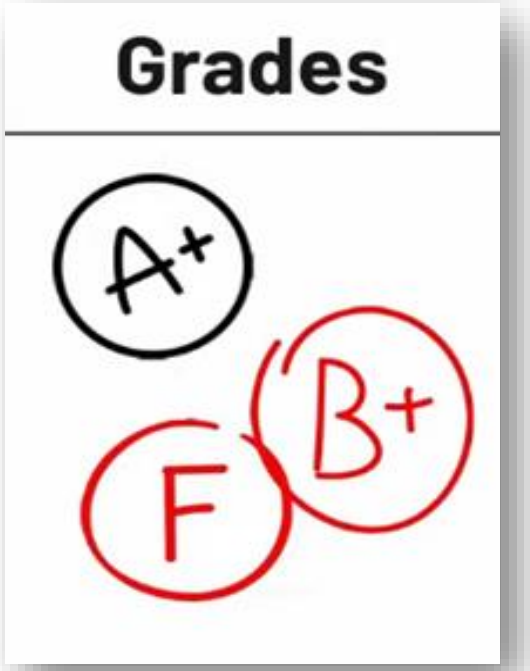


Qualitative Data (Categorical)

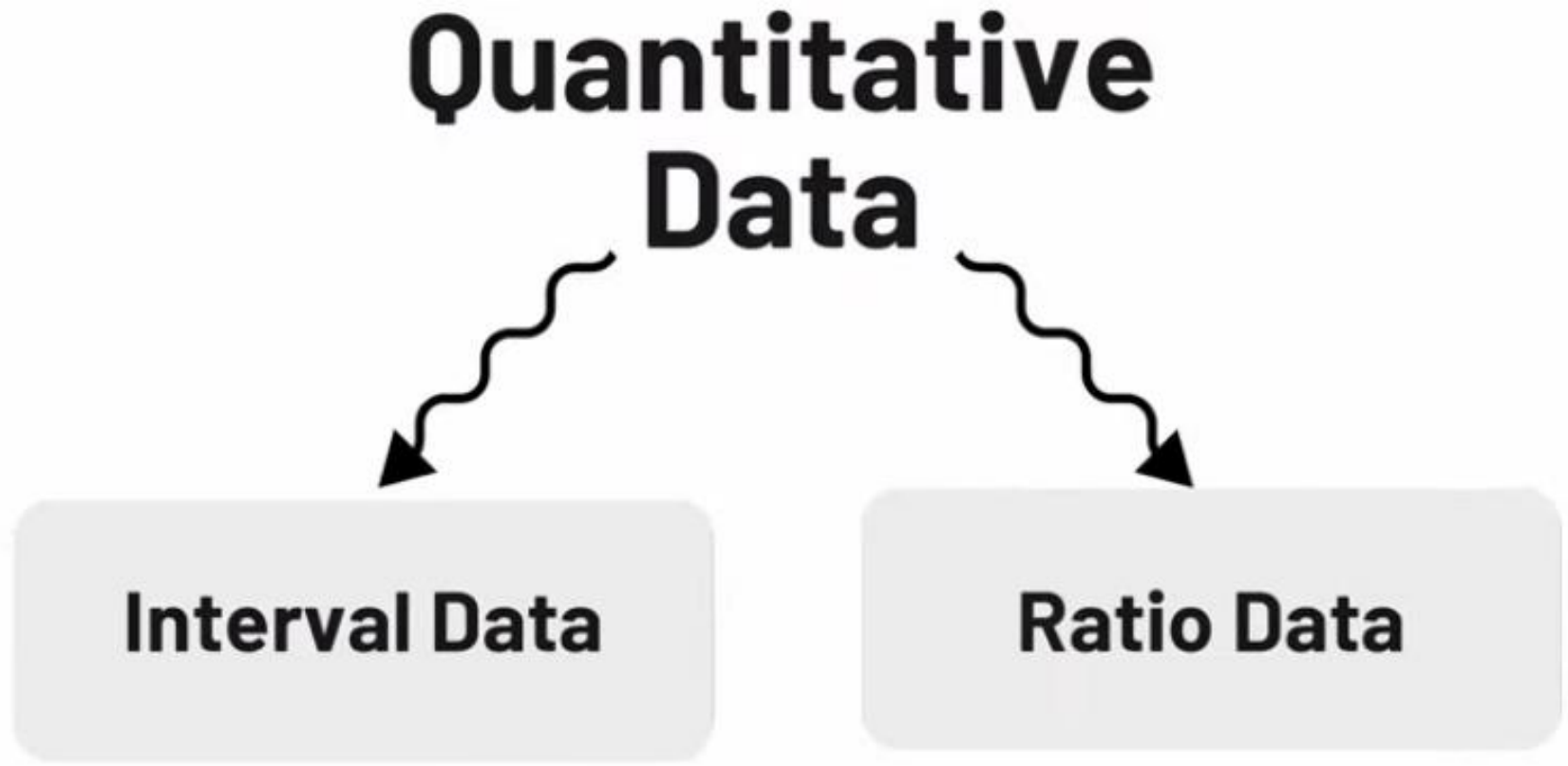


Nominal Data Examples

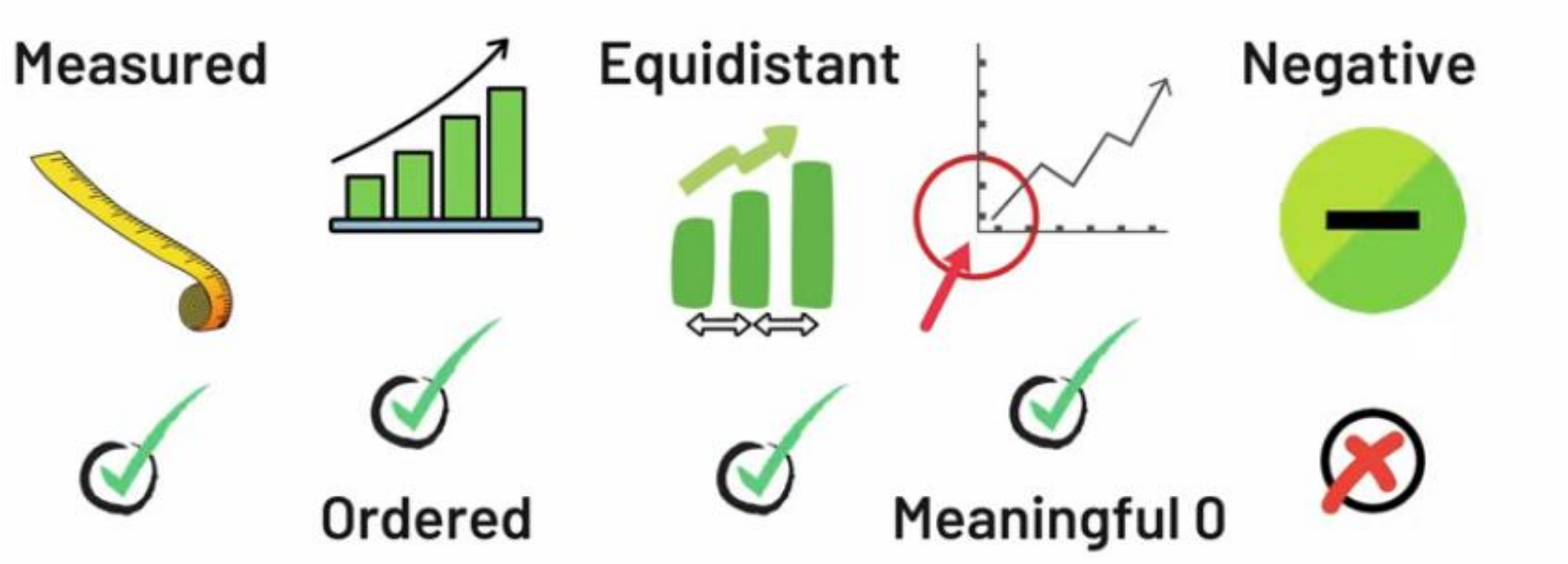
Ordinal Data Examples



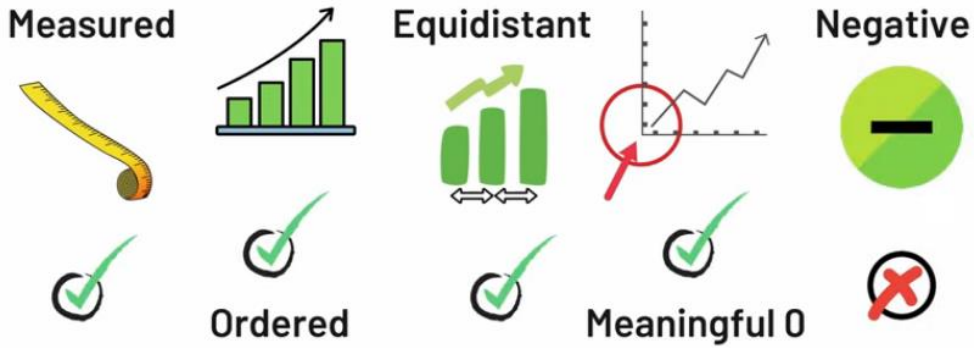
Quantitative Data (Numerical) Types



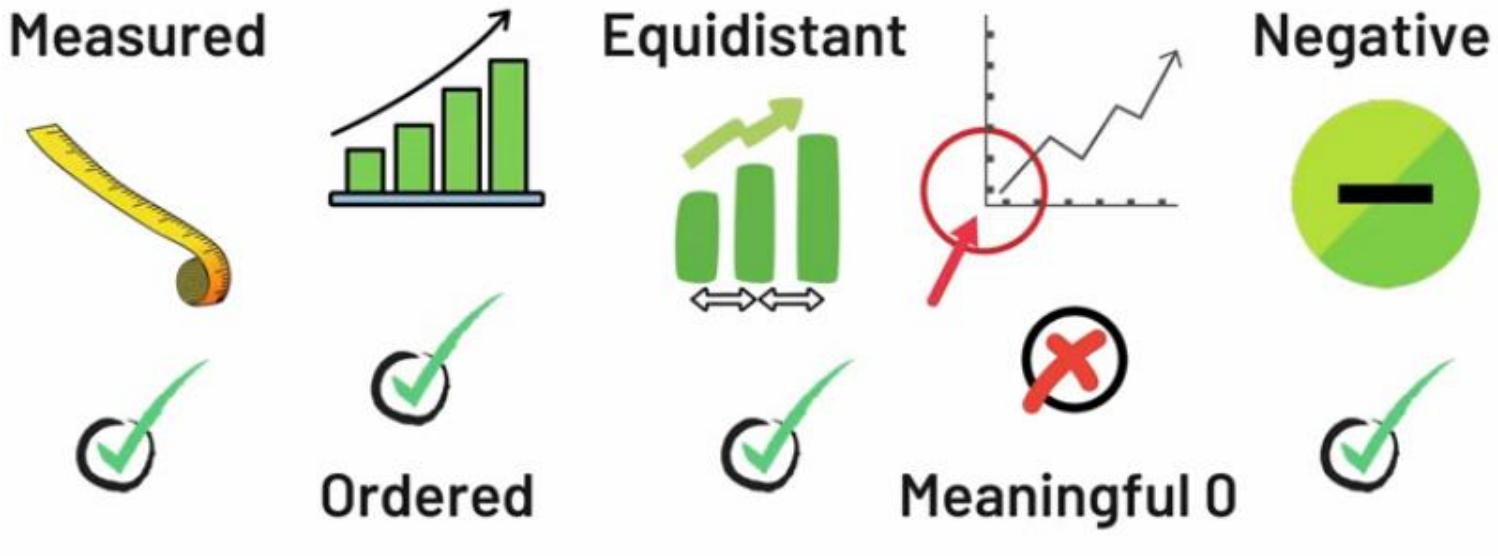
Ratio Data Properties



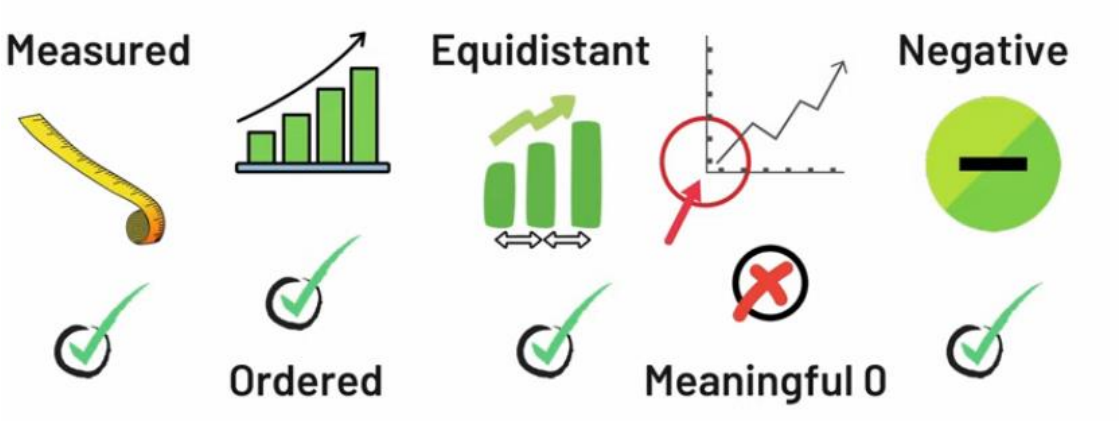
Ratio Data Example

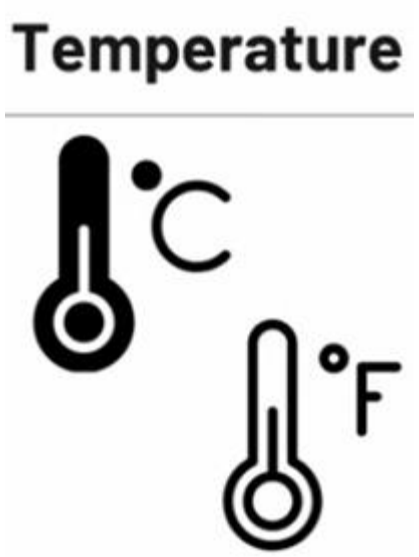


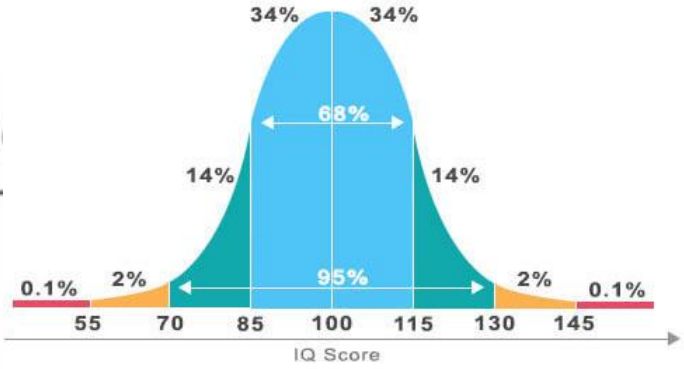
Interval Data Properties

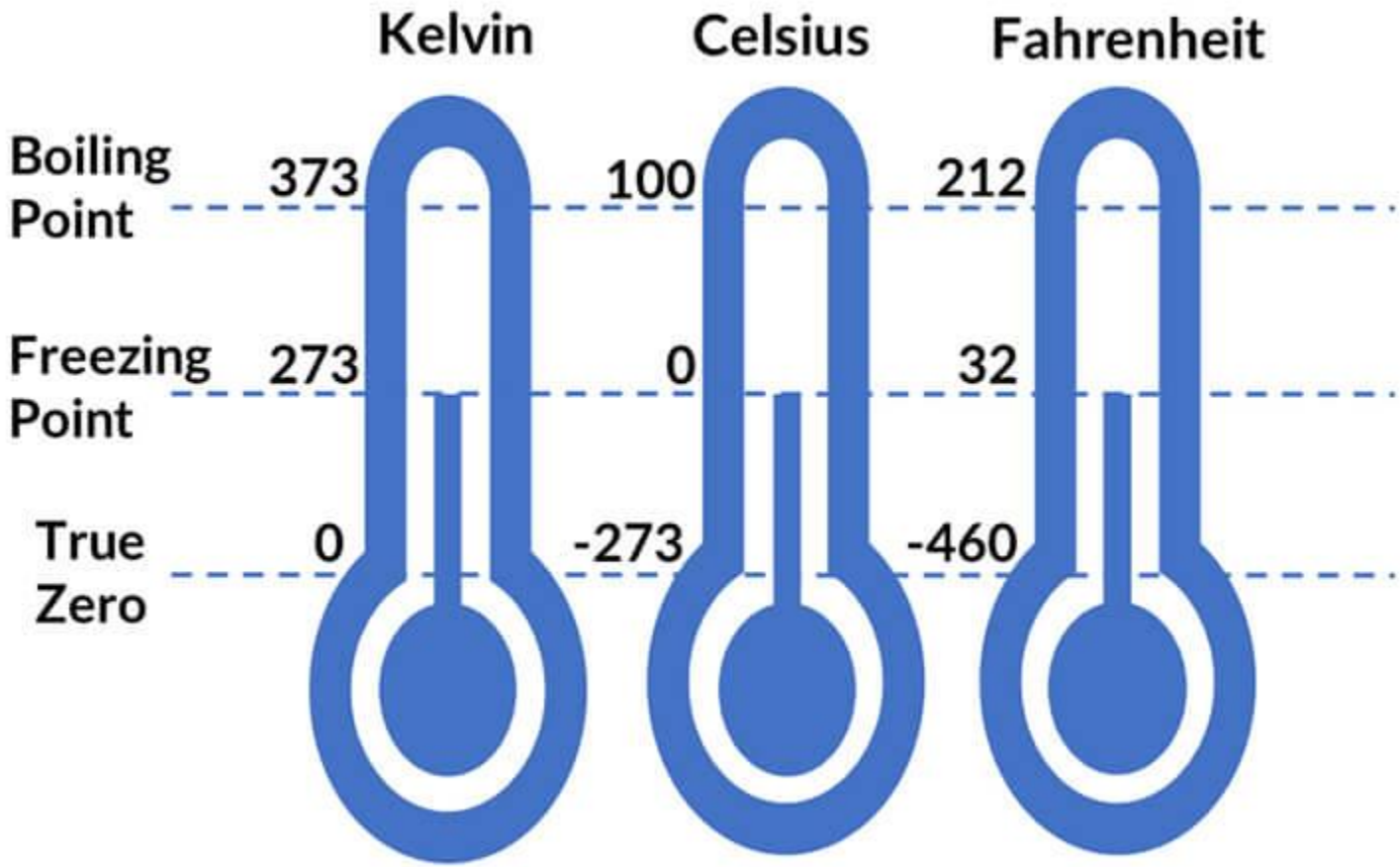


Interval Data Examples

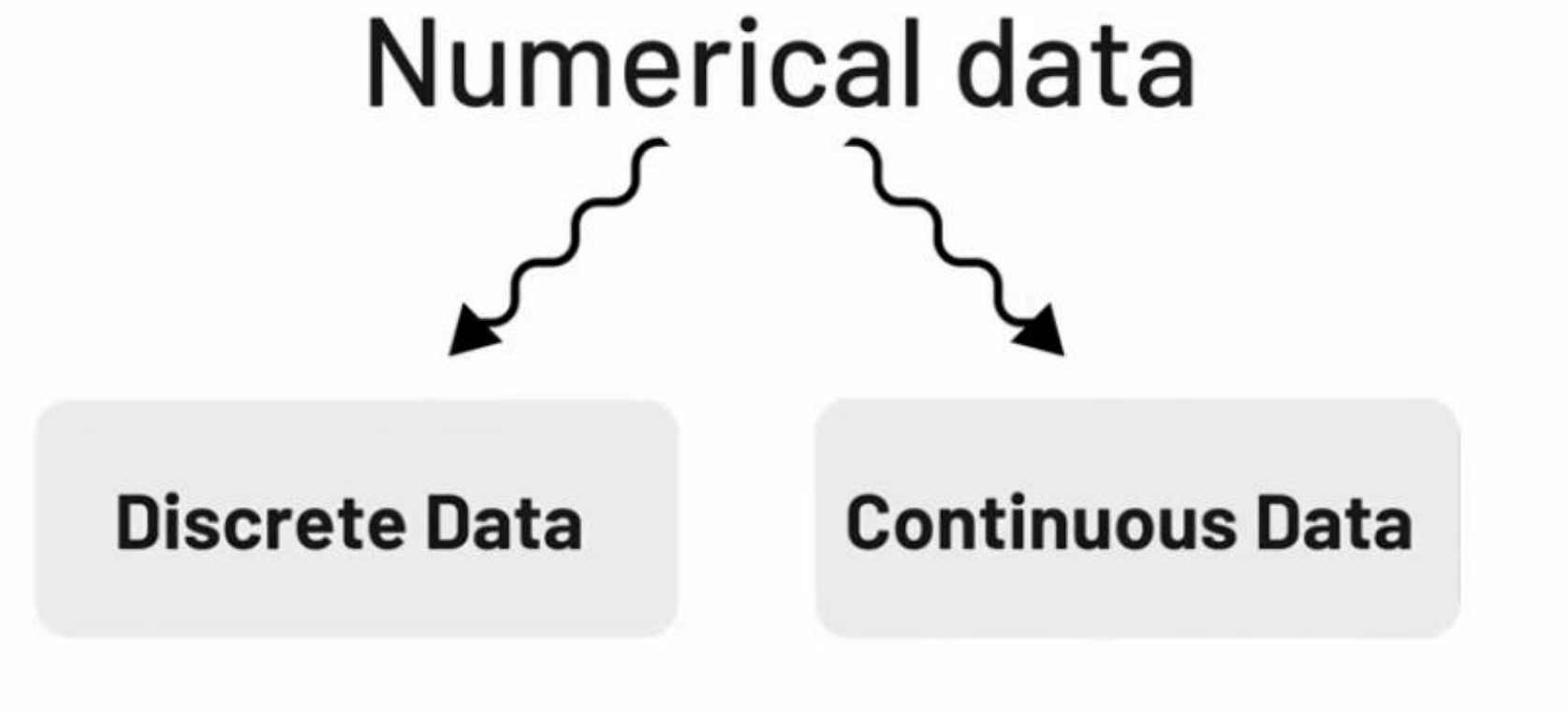




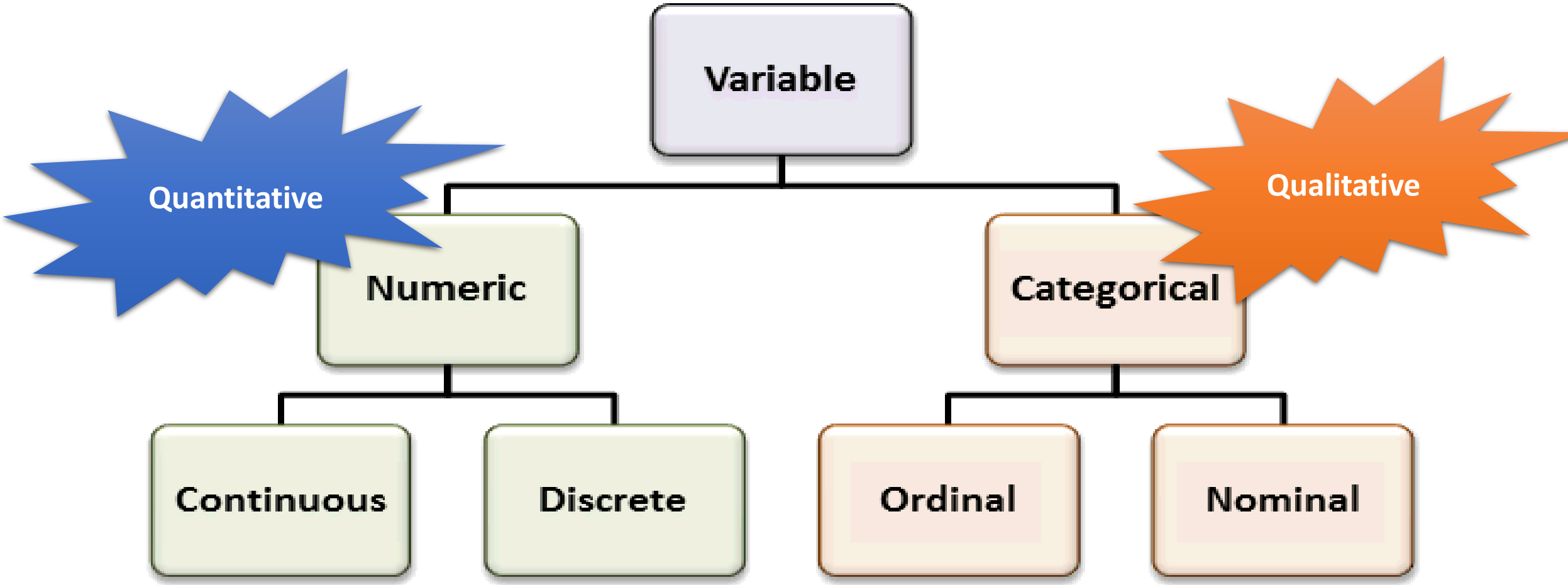



Kelvin
Celsius
Fahrenheit
Boiling Point
373
100
212
Freezing Point
273
0
32
True Zero
0
-273
-460

Discrete vs. Continuous Data



Variable Data Types

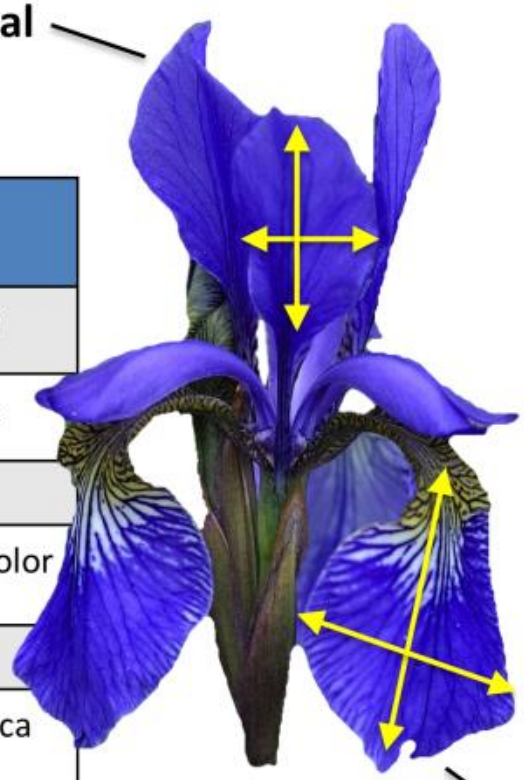


Dataset

Samples
(instances, observations)

	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
...					
50	6.4	3.5	4.5	1.2	Versicolor
...					
150	5.9	3.0	5.0	1.8	Virginica

Petal



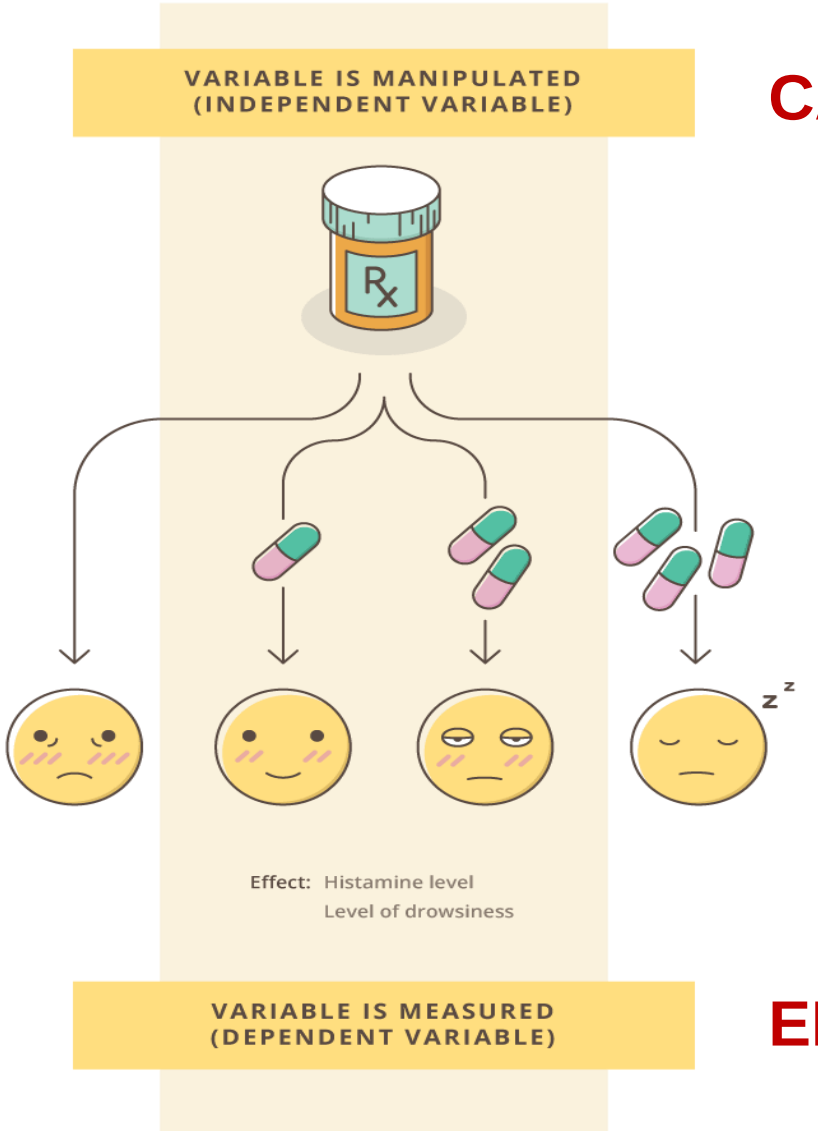
Sepal

Class labels
(targets)

Features
(attributes, measurements, dimensions)

IRIS Dataset

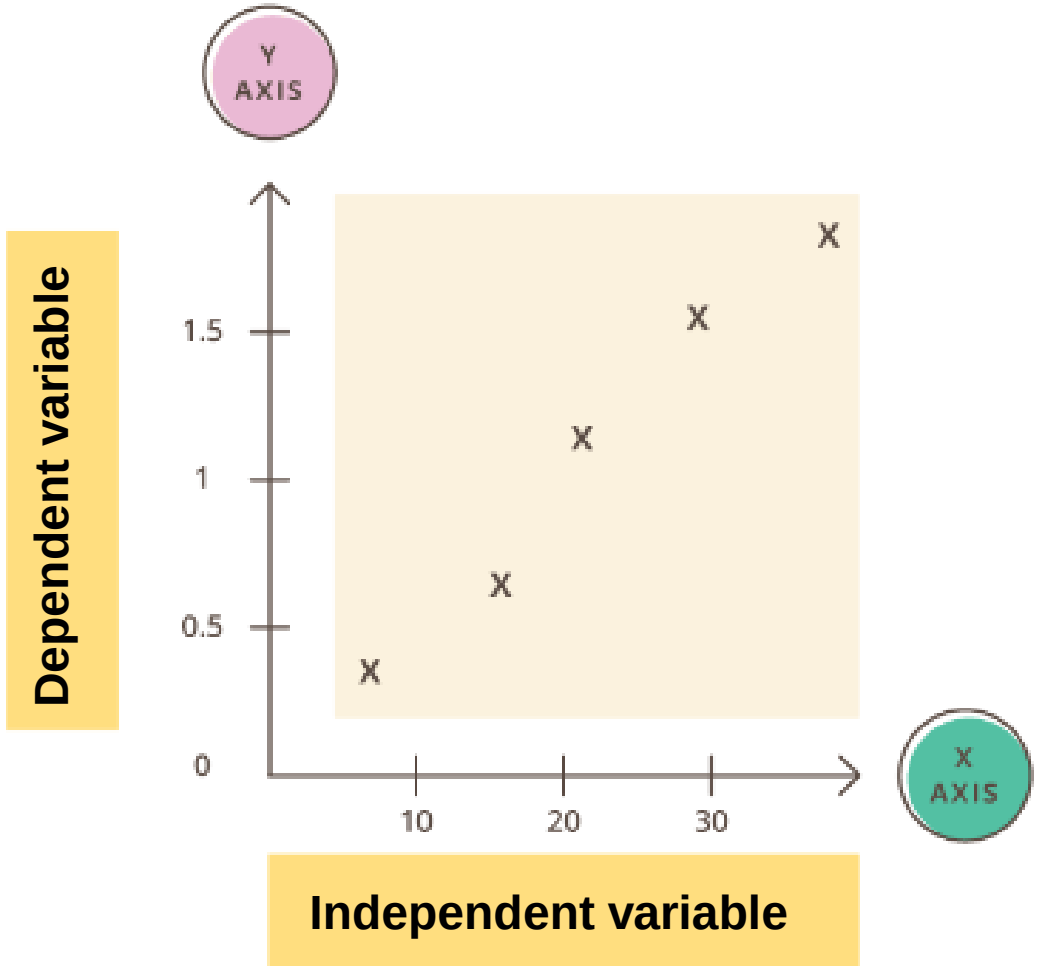
Independent and Dependent variable



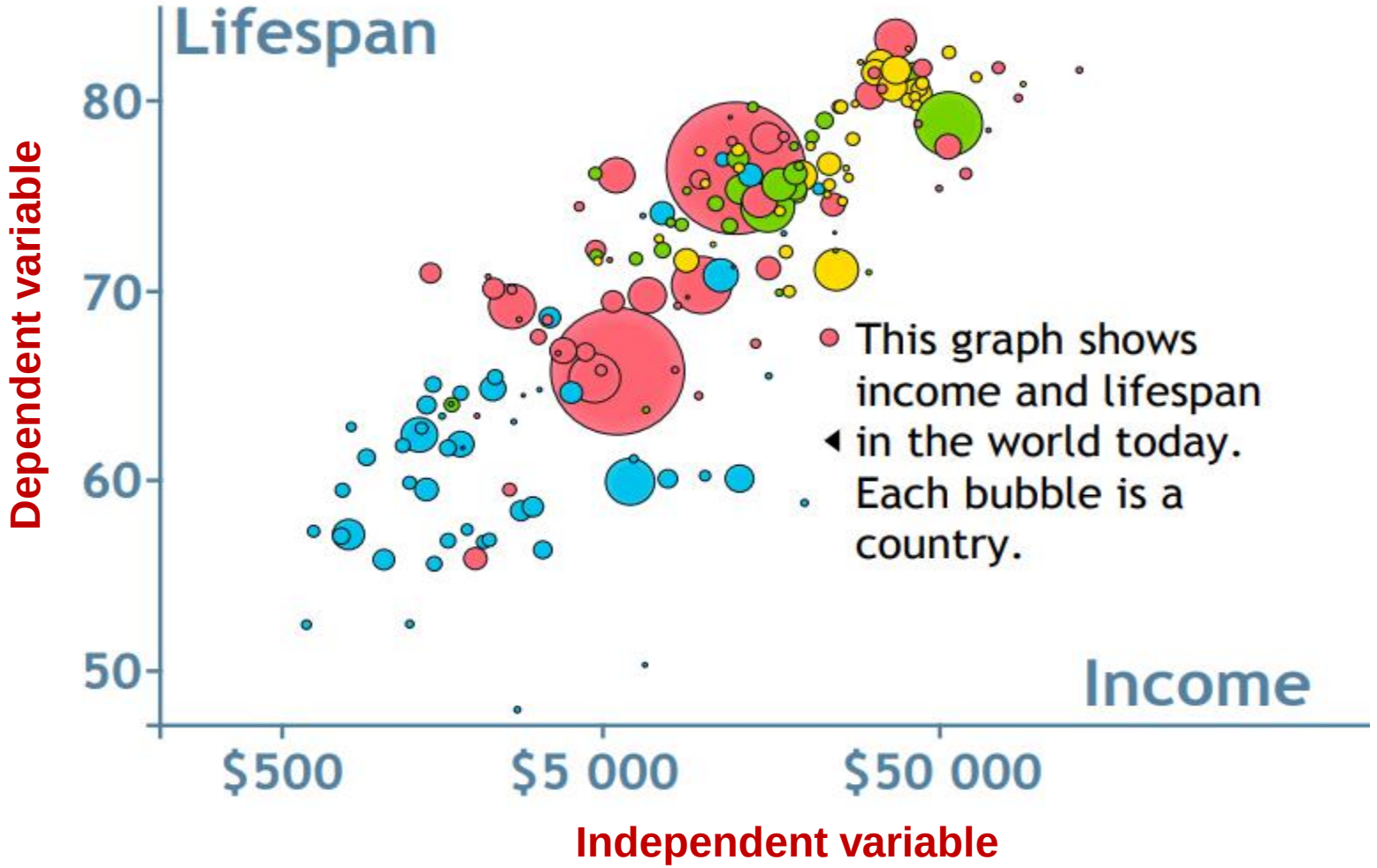
CAUSE

EFFECT

Independent and Dependent variable



Independent and Dependent variable

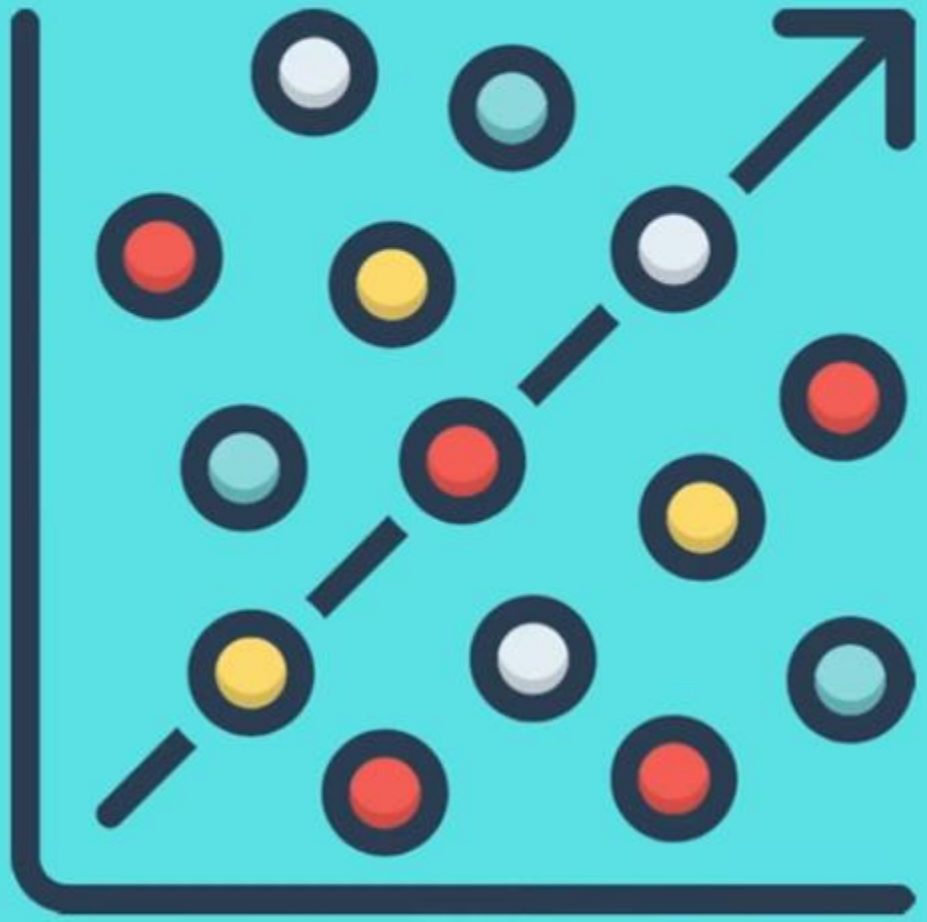


Exercise

id	price	bedrooms	sqft_living	floors	waterfront	condition	grade	yr_built	postalcode
7129300520	221900	3	1180	1	0	3	7	1955	98178
6414100192	538000	3	2570	2	0	3	7	1951	98125
5631500400	180000	2	770	1	0	3	6	1933	98028
2487200875	604000	4	1960	1	0	5	7	1965	98136
1954400510	510000	3	1680	1	0	3	8	1987	98074
7237550310	1.23E+06	4	5420	1	0	3	11	2001	98053
1321400060	257500	3	1715	2	0	3	7	1995	98003
2008000270	291850	3	1060	1	0	3	7	1963	98198

Why do you need to study different types of data?

Linear Regression



Continuous Data

Categorical Data

Statistics

What is Statistics?

Science of **collecting**, **summarizing**, **analyzing**, **interpreting**, and **presenting** data.

POPULATION	SAMPLE
Collection of all items of interest	A subset of the population
N	n
parameters	statistics

Population vs. Samples

Population

- It refers to the whole data set for the use case.
- It includes all groups which can be correlated with each other.
- Example: All the members of an online forum reading articles.

Samples

- It is a subset of the population.
- These are a random sample of data points.
- The process of determining the sample from population data is known as sampling.
- Example: A group of club members sample who read technical articles.

SAMPLE

Less time consuming

Less costly (cheaper)

SAMPLE

Canteen

RANDOMNESS

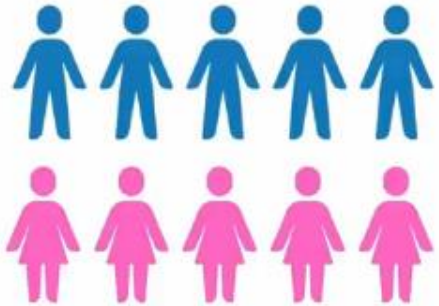
A random sample is collected when each member of the sample is chosen from the population strictly by chance.

REPRESENTATIVENESS

A representative sample is a subset of the population that accurately reflects the mem

Example

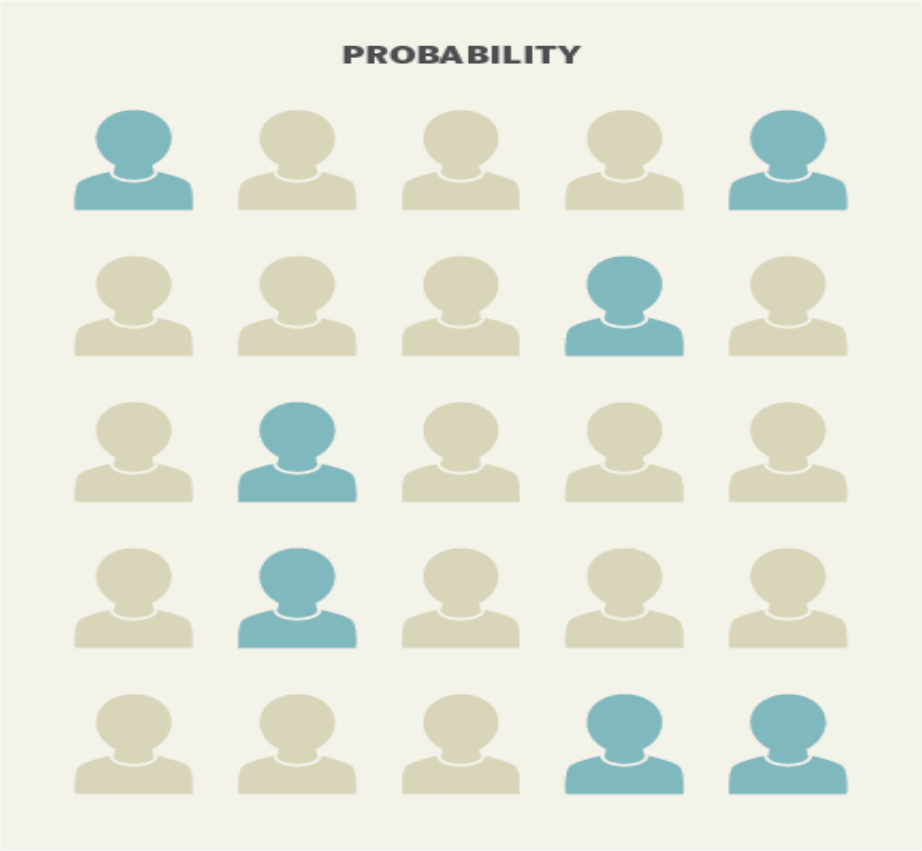
Students (Population)

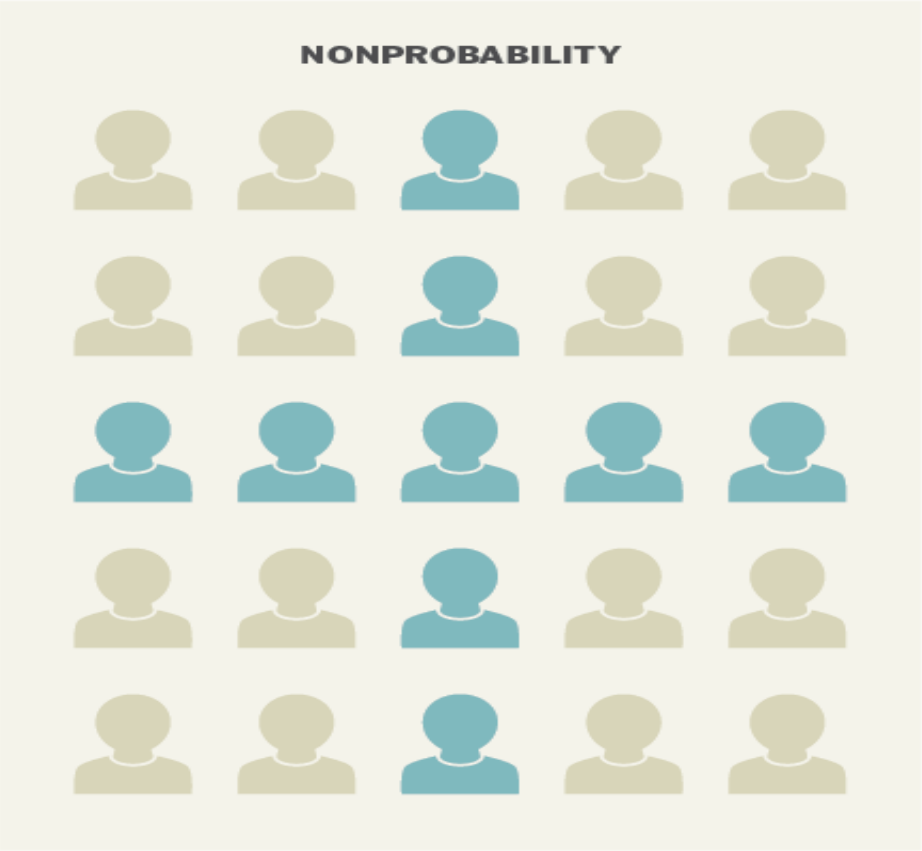


Sample Drawn

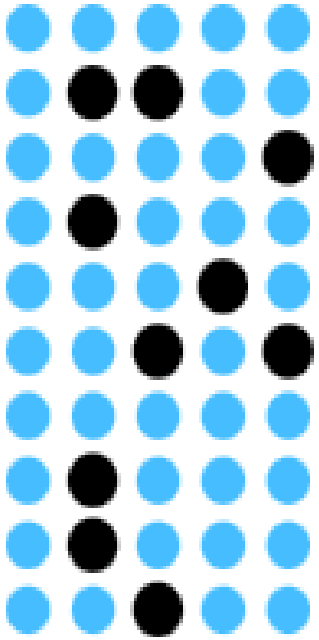
Data sampling methods

Data analysts can use either random methods (probability) to select data to include in a sample or their own judgment to determine what to include (nonprobability).



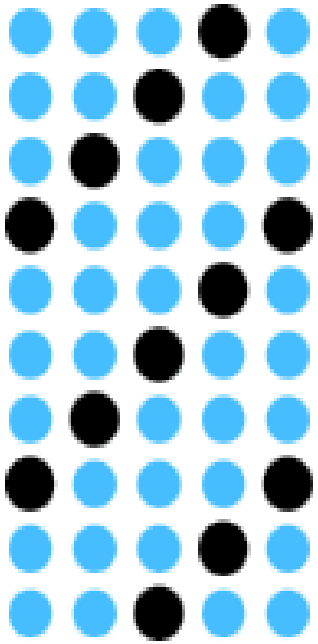


How do we choose what members of the population to sample?



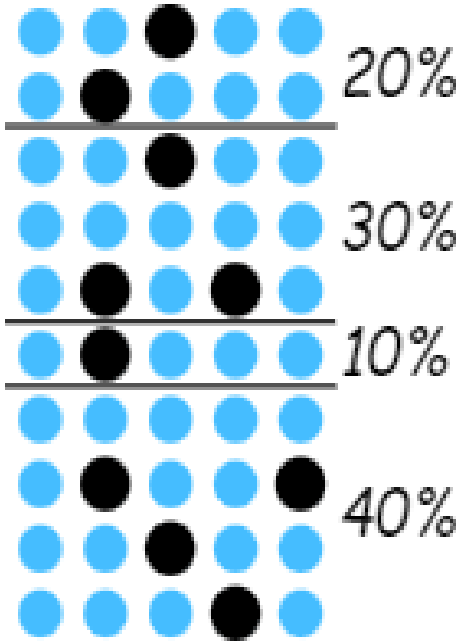
Random Sample

(pick randomly from list)



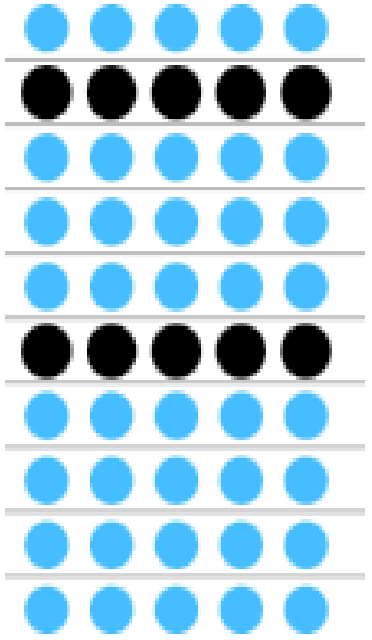
Systematic Sample

(such as every 4th)



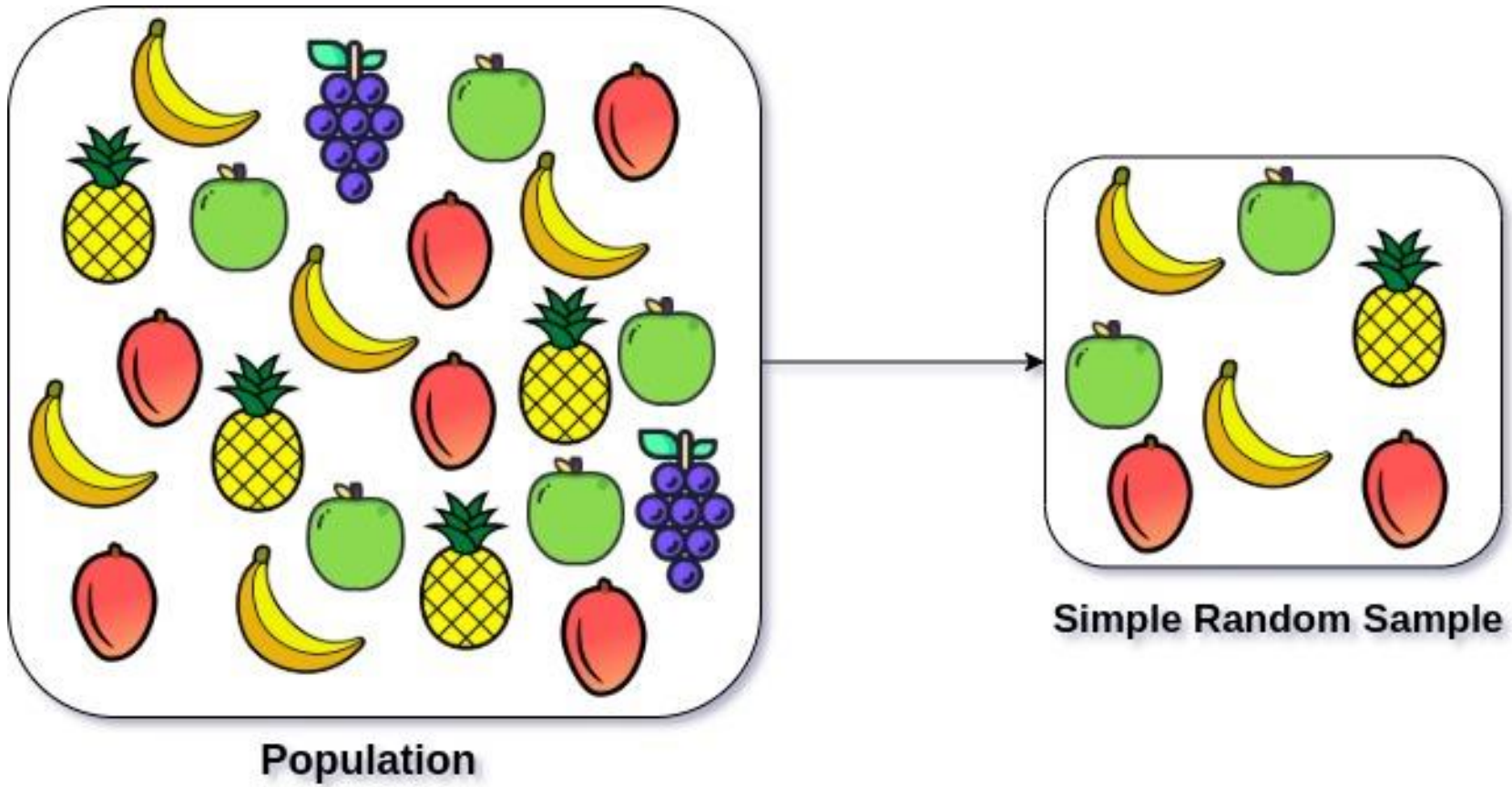
Stratified Sample

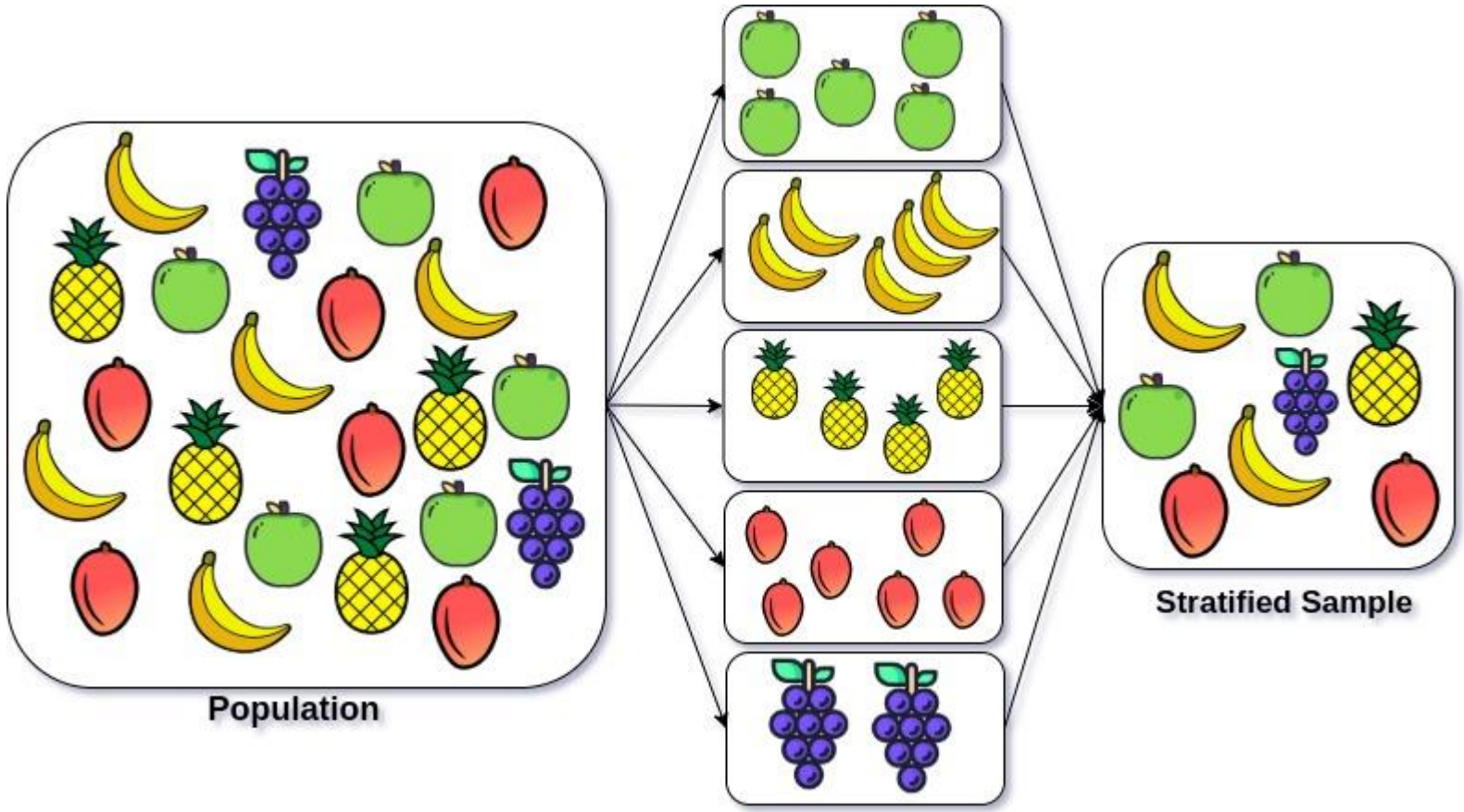
(randomly, but in ratio to group size)



Cluster Sample

(choose whole groups randomly)


Population
Simple Random Sample


Population
Stratified Sample

Exercise

You are trying to estimate the average valuation of start-ups in the USA. Imagine you are able to visit 200 start-ups in Silicon Valley in a random manner. What is a possible problem of your study?

- **The sample is not random.**
- **The sample is too small.**
- **The sample was not representative.**
- **The population is unknown.**

Statistical Analysis Types

Summarization

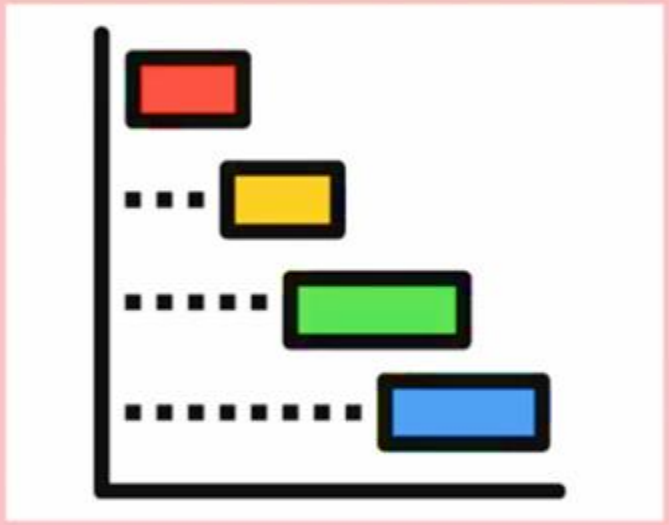
Generalization

	Descriptive Statistics	Inferential Statistics
Definition	Descriptive statistics is used to describe the characteristics of the population using a sample.	Inferential statistics uses various analytical tools to draw inferences about the population using samples.
Tools	Measures of central tendency and measures of dispersion.	Hypothesis testing and regression analysis.
Use	Organizes, describes and presents data in a meaningful way with the help of charts and graphs.	Tests, predicts, and compares data obtained from various samples.
Relevance	It is used to summarize known data in a way that can be used for further predictions and analysis.	It tries to use the summarized samples to draw conclusions about the population.

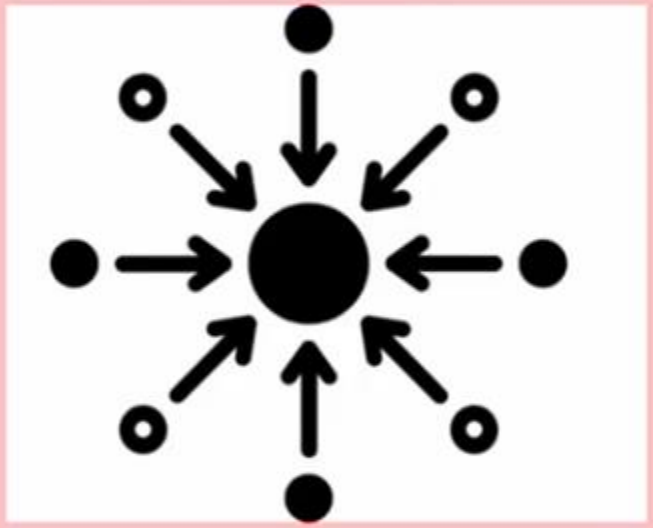
Descriptive Statistics

- It is concerned with describing, summarizing, and graphing the data.
- We do not reach a conclusion or a proof of hypothesis in it.

Types of Descriptive Statistics

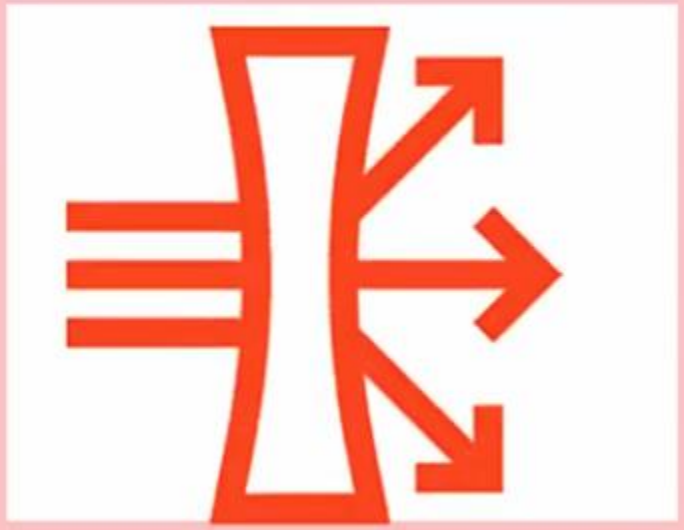


Distribution Frequency

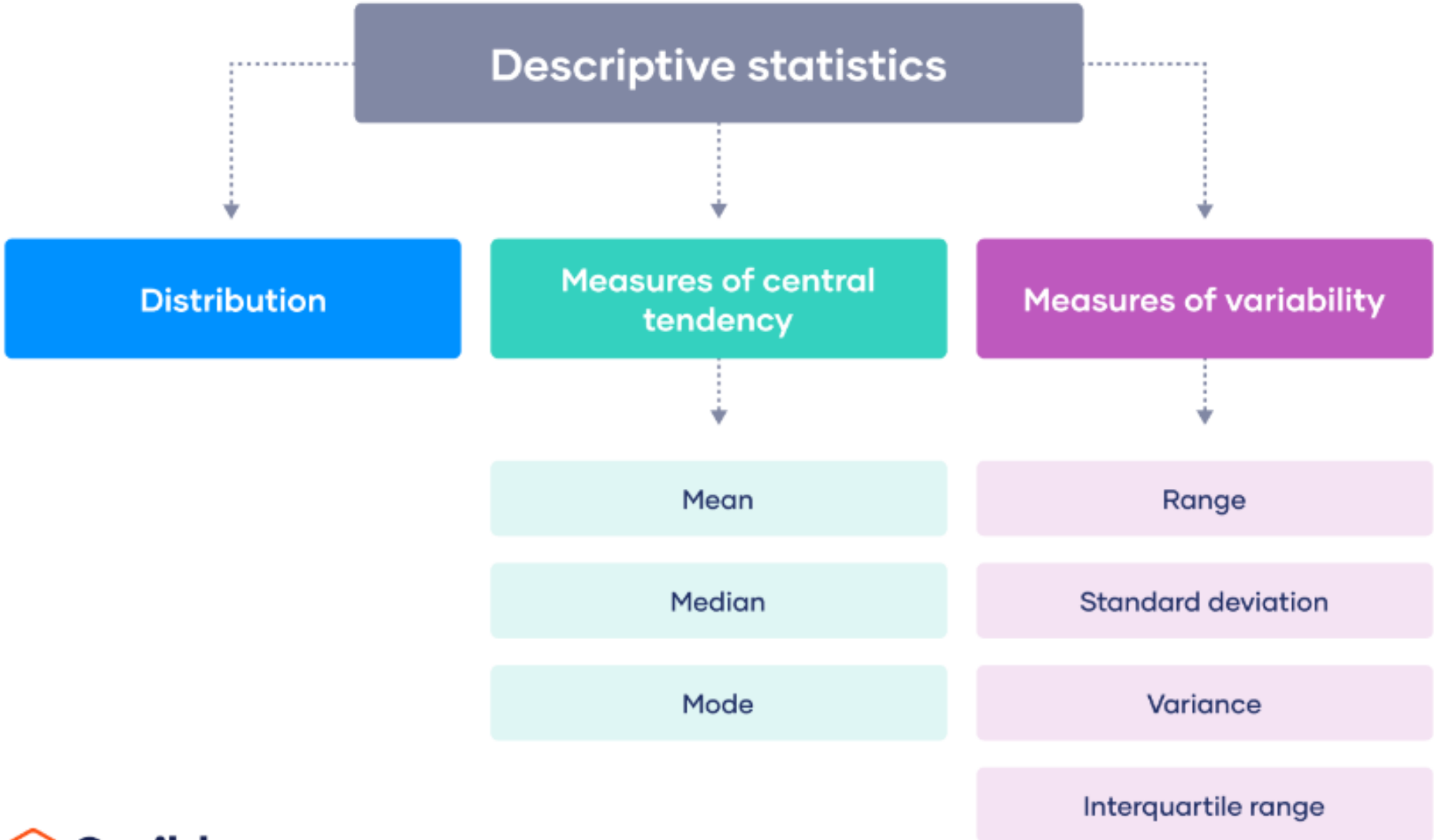


Measure of Central Tendency

- Mean
- Median
- Mode



Measure of Variability/Dispersion


Descriptive statistics
Distribution
Measures of central tendency
Measures of variability
Mean
Median
Mode
Range
Standard deviation
Variance
Interquartile range

Scribbr

Frequency Distribution

blood type for 20 patients

O	B	O	O	A	AB	AB	A	B	O
A	AB	A	O	B	B	O	AB	A	O

Blood Type	Frequency
A	5
AB	4
B	4
O	7
Total	20

Frequency Distribution

blood type for 20 patients

O	B	O	O	A	AB	AB	A	B	O
A	AB	A	O	B	B	O	AB	A	O

Blood Type	Frequency	Relative Frequency
A	5	5/20 = 0.25 25%
AB	4	4/20 = 0.20 20%
B	4	4/20 = 0.20 20%
O	7	7/20 = 0.35 35%
Total	**20**	1.00

Frequency Distribution

12	13	21	21	23	26	26	31	37	38	38	38	40
42	48	49	51	52	53	54	54	54	55	55	56	

ages

X	f	X	f	X	f	X	f
12	1	24	0	36	0	48	1
13	1	25	0	37	1	49	1
14	0	26	2	38	3	50	0
15	0	27	0	39	0	51	1
16	0	28	0	40	1	52	1
17	0	29	0	41	0	53	1
18	0	30	0	42	1	54	3
19	0	31	1	43	0	55	2
20	0	32	0	44	0	56	1
21	2	33	0	45	0		
22	0	34	0	46	0		
23	1	35	0	47	0		

Frequency Distribution

12	13	21	21	23	26	26	31	37	38	38	38	40
42	48	49	51	52	53	54	54	54	55	55	56	

ages

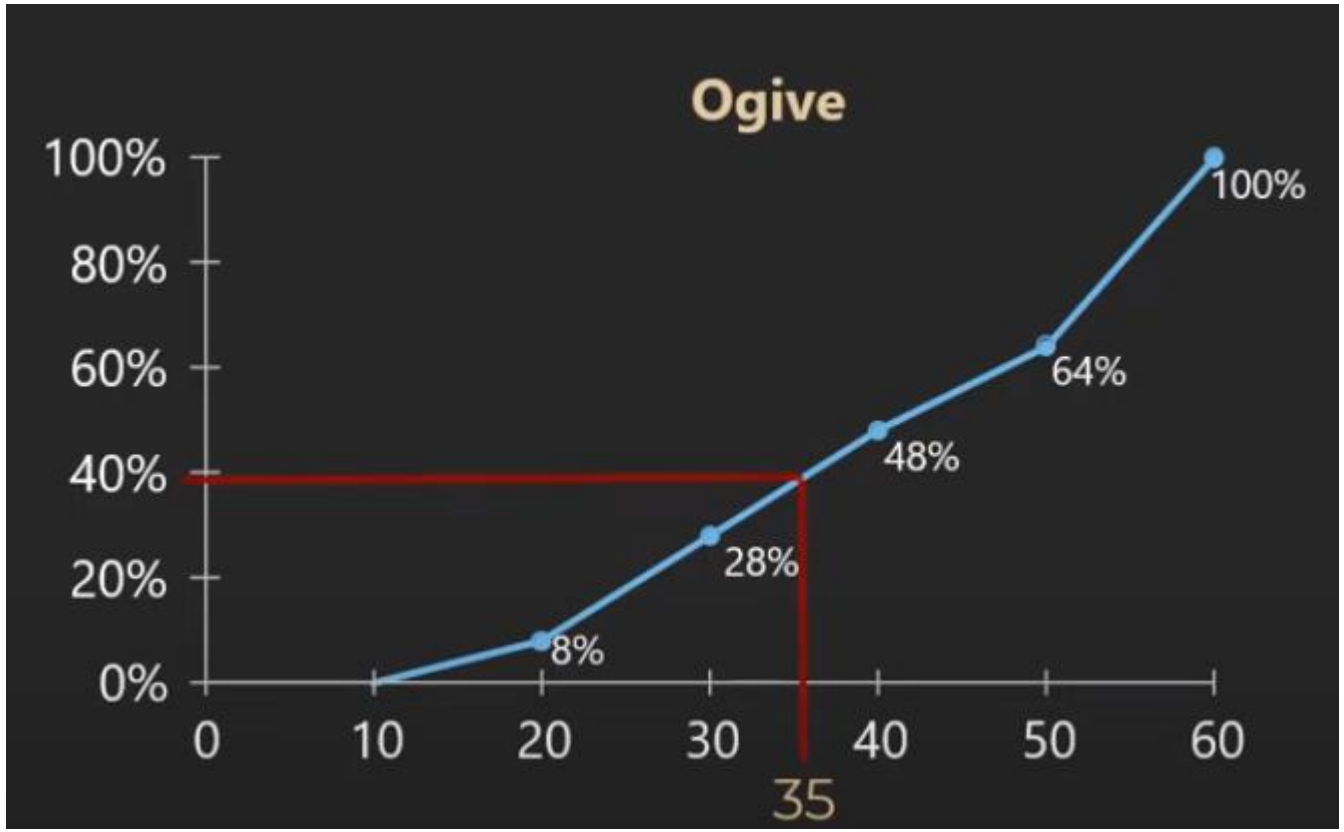
Interval	Frequency
10 to < 20	2
20 to < 30	5
30 to < 40	5
40 to < 50	4
50 to < 60	9
Total	25

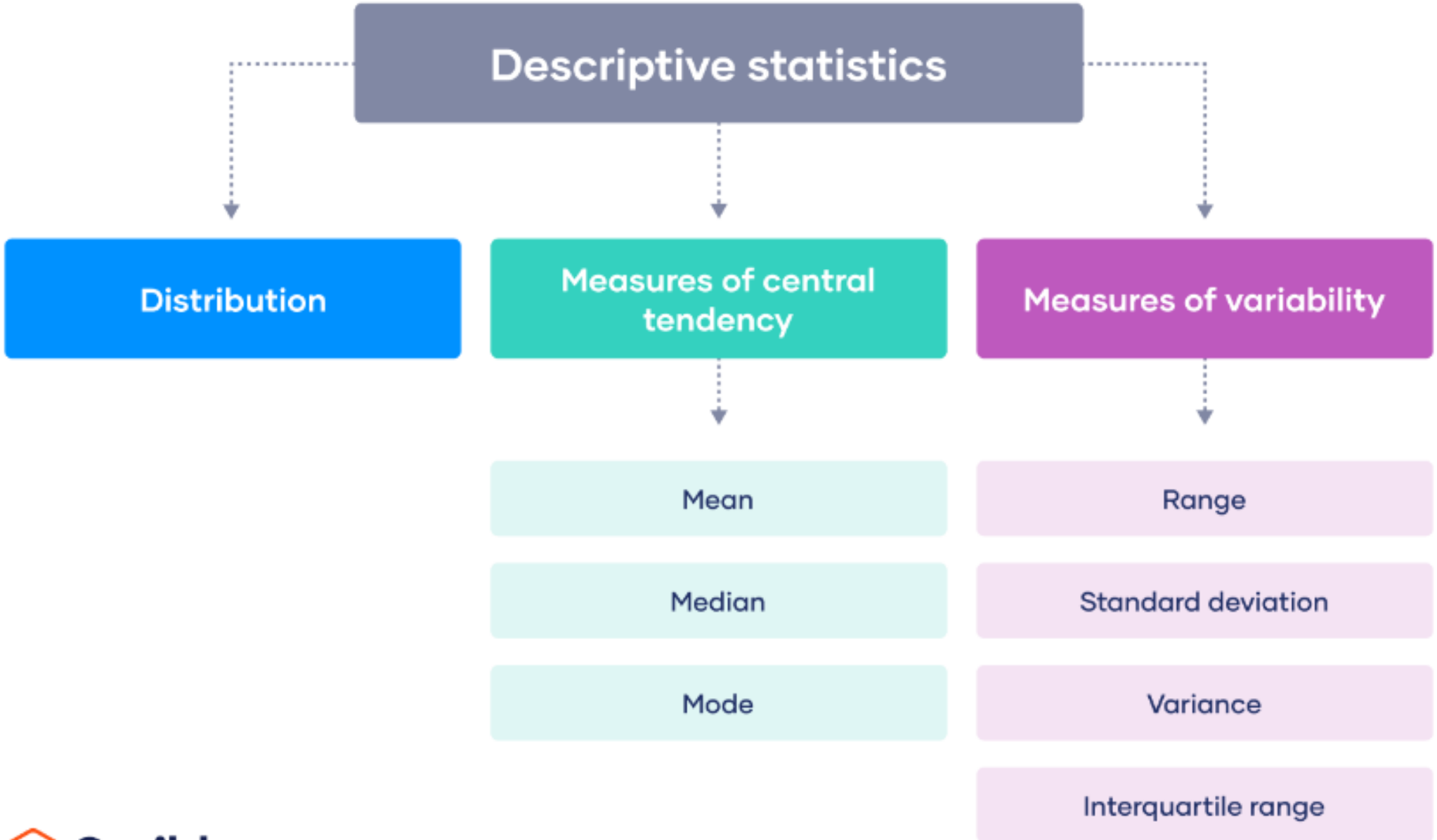
Frequency Distribution

Interval	Frequency	Cumulative Frequency	Cumulative Relative Frequency
10 to < 20	2	2	2/25 = 8%
20 to < 30	5	7	7/25 = 28%
30 to < 40	5	12	12/25 = 48%
40 to < 50	4	16	16/25 = 64%
50 to < 60	9	25	25/25 = 100%
Total	25		

Frequency Distribution

Interval	Frequency	Cumulative Frequency	Cumulative Relative Frequency
10 to < 20	2	2	2/25 = 8%
20 to < 30	5	7	7/25 = 28%
30 to < 40	5	12	12/25 = 48%
40 to < 50	4	16	16/25 = 64%
50 to < 60	9	25	25/25 = 100%
Total	25		




Descriptive statistics
Distribution
Measures of central tendency
Measures of variability
Mean
Median
Mode
Range
Standard deviation
Variance
Interquartile range

Scribbr

Mean

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

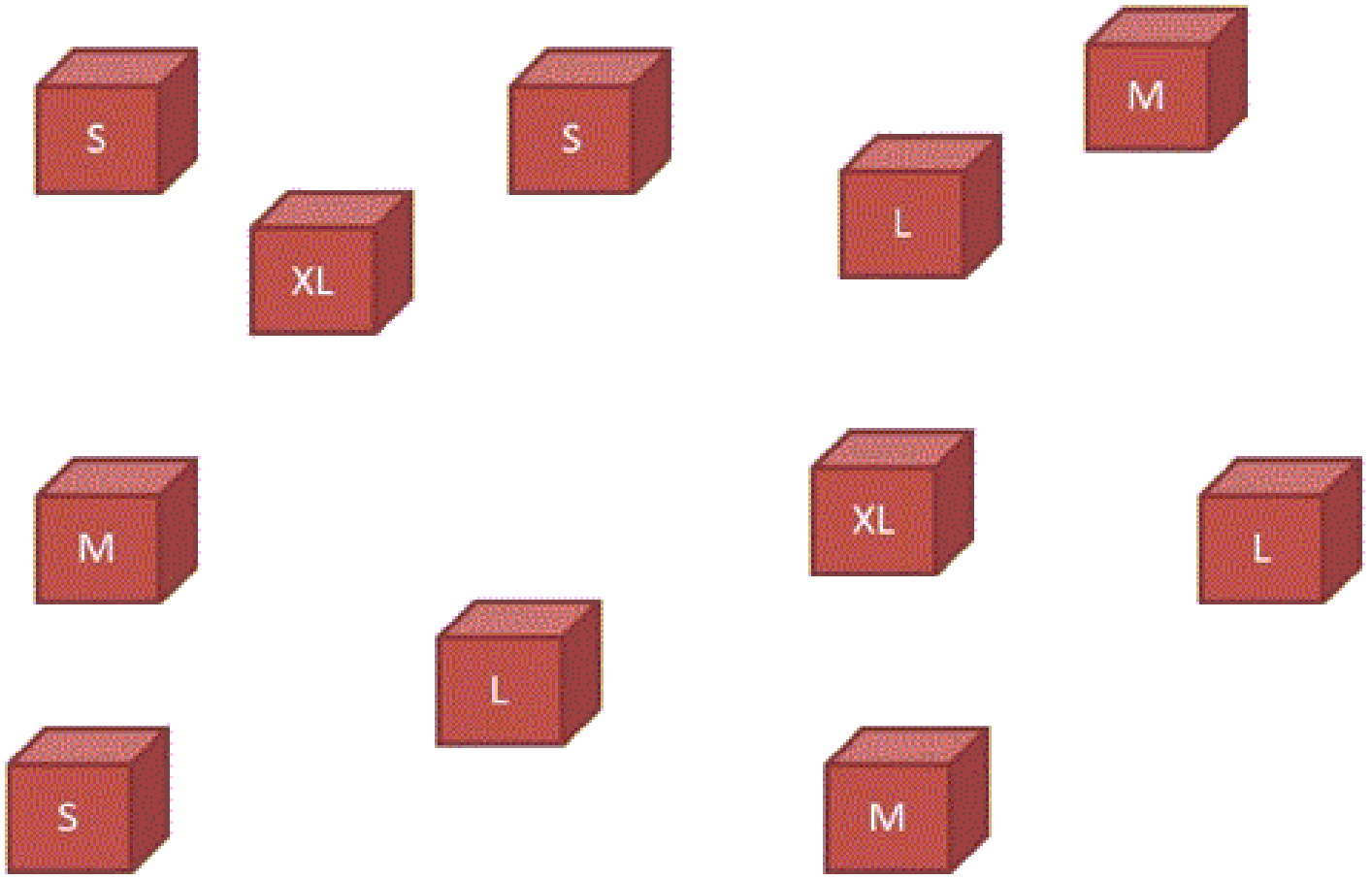
15, 10, 19, 19, 7, 11, 15, 19, 20, 12, 17, 18

$$\text{Mean} = \frac{\text{Sum of the numbers}}{\text{How many numbers}} \qquad \frac{182}{12}$$

Mean = 15.17

Visualisation of the Median

Which item(s) is/are in the middle?



Median

[3, 4, 5, 6, 7]

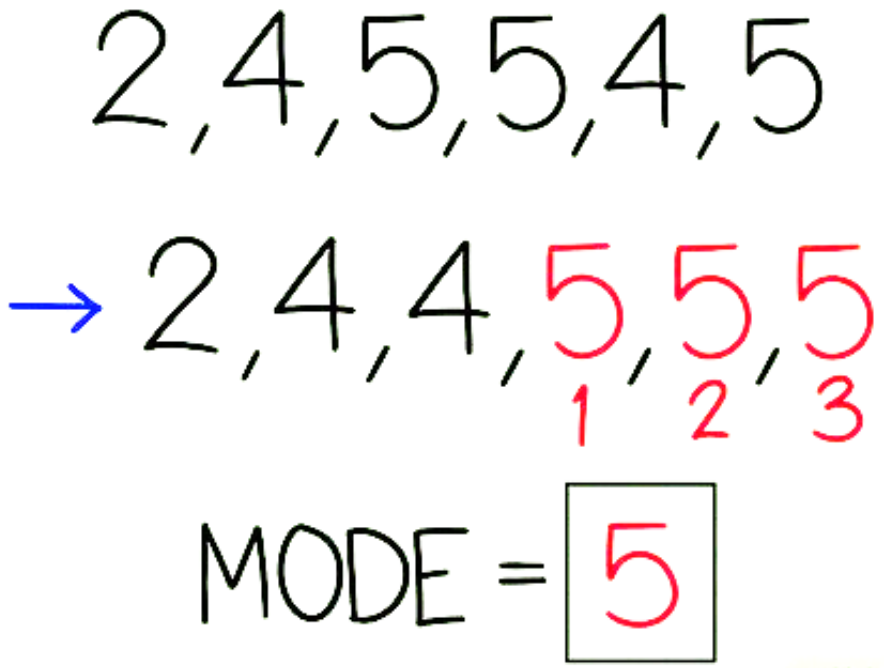
Median
(Odd number of data)

[3, 4, 5, 6, 7, 8]

(5 + 6) / 2 = 5.5

Median
(Even number of data)

Mode



Exercise

Background You have a sample of 11 people and their personal annual income.
Task 1 Calculate the mean, median and mode
Task 2 Try to interpret on the numbers you got

	Annual income
$	62,000.00
$	64,000.00
$	49,000.00
$	324,000.00
$	1,264,000.00
$	54,330.00
$	64,000.00
$	51,000.00
$	55,000.00
$	48,000.00
$	53,000.00

Task 1:

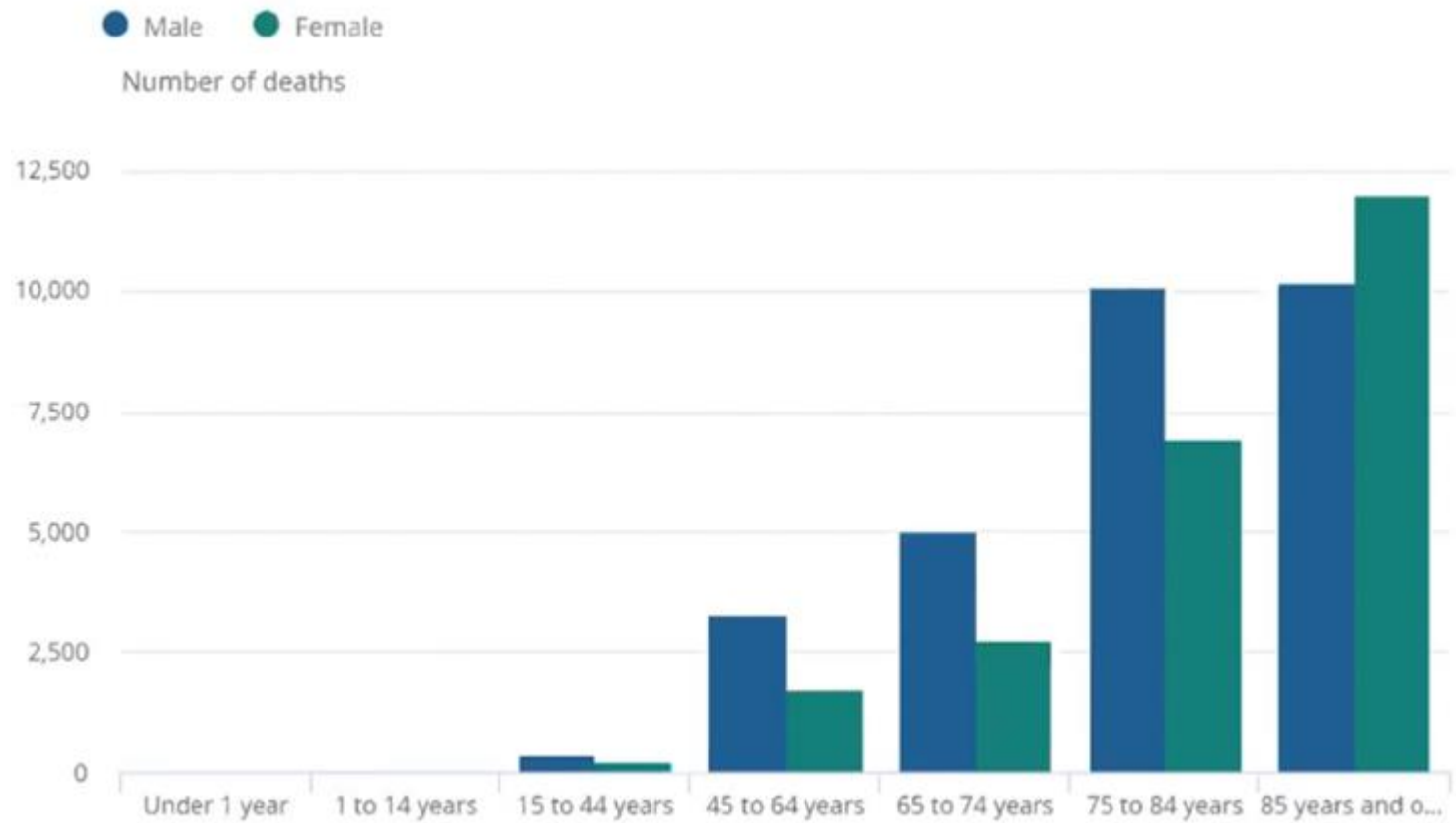
	Annual income
Mean	$ 189,848.18
Median	$ 55,000.00
Mode	$ 64,000.00

Which measure is best?

Task 2: Income is a very interesting topic. There is extreme variability in the income of different individuals.
Generally, most of the people gravitate around a certain salary.
Moreover, in most countries there is a minimum salary, therefore most data points are constrained between the minimum salary and some number.
Finally, there are certain individuals that are earn much more than others. They are the outliers.
Usually, whenever we have research on income, we use the **median** income, instead of the mean income.
Income is an example where averages are meaningless. You should be aware that the correct measure to use depends on the research that you are conducting.

If you would like to estimate the typical age of people who have died of Covid-19 in the UK, would you be more interested in the mean or median age?

1. Median ***2. Mean***



Variance

$$\sigma^2 = \frac{\Sigma (X - \mu)^2}{N}$$

- σ^2 = population variance
- Σ = sum of...
- X = each value
- μ = population mean
- N = number of values in the populatic

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

- s^2 = sample variance
- Σ = sum of...
- X = each value
- $\bar{x}$ = sample mean
- n = number of values in the sample

Data set					
46	69	32	60	52	41

Data set					
46	69	32	60	52	41

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

Mean ($\bar{x}$)
$\bar{x}$ = (46 + 69 + 32 + 60 + 52 + 41) ÷ 6 = **50**

Score	Deviation from the mean
46	46 − 50 = **-4**
69	69 − 50 = **19**
32	32 − 50 = **-18**
60	60 − 50 = **10**
52	52 − 50 = **2**
41	41 − 50 = **-9**

Squared deviations from the mean
$(-4)^2$ = 4 × 4 = **16**
19^2 = 19 × 19 = **361**
$(-18)^2$ = -18 × -18 = **324**
10^2 = 10 × 10 = **100**
2^2 = 2 × 2 = **4**
$(-9)^2$ = -9 × -9 = **81**

Sum of squares
16 + 361 + 324 + 100 + 4 + 81 = **886**

Variance
886 ÷ (6 − 1) = 886 ÷ 5 = **177.2**

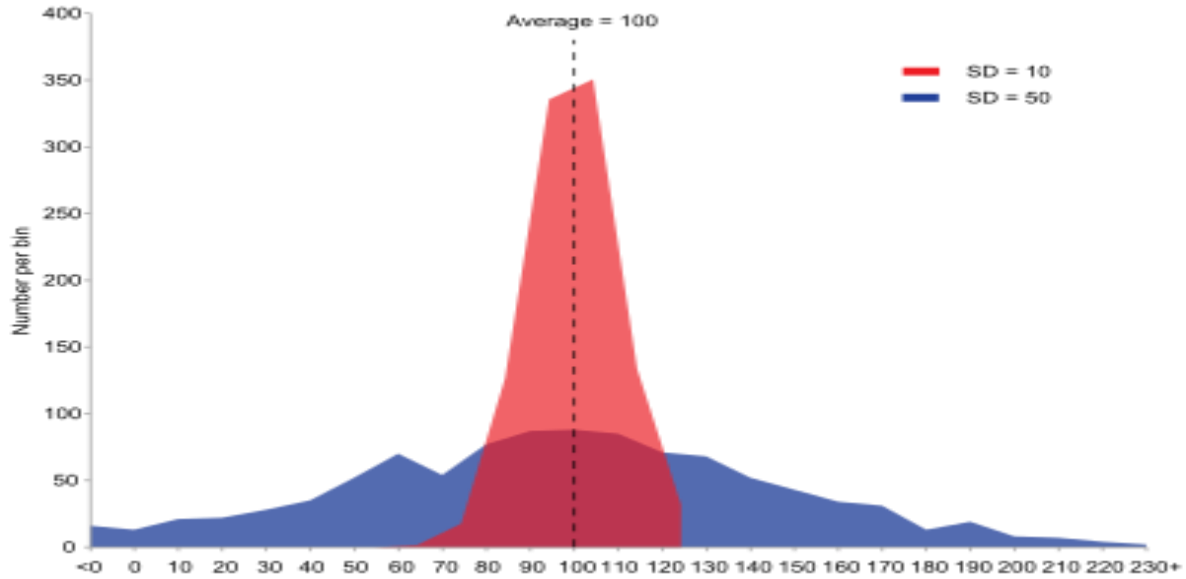
Standard Deviation

$$\sigma = \sqrt{\frac{\Sigma (X - \mu)^2}{N}}$$

- σ = population standard deviation
- ∑ = sum of...
- *X* = each value
- μ = population mean
- *N* = number of values in the population

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

- *s* = sample standard deviation
- ∑ = sum of...
- *X* = each value
- $\bar{x}$ = sample mean
- *n* = number of values in the sample



Exercise

Background You have the annual personal income of 11 people from the USA. You have the mean income from the exercise on mean, median and mode

Task 1 Decide whether you have to use sample or population formula for the variance

Task 2 Calculate the variance of their income

Task 3 Generally, what does this number tell you?

Annual income
$ 62,000.00
$ 64,000.00
$ 49,000.00
$ 324,000.00
$ 1,264,000.00
$ 54,330.00
$ 64,000.00
$ 51,000.00
$ 55,000.00
$ 48,000.00
$ 53,000.00

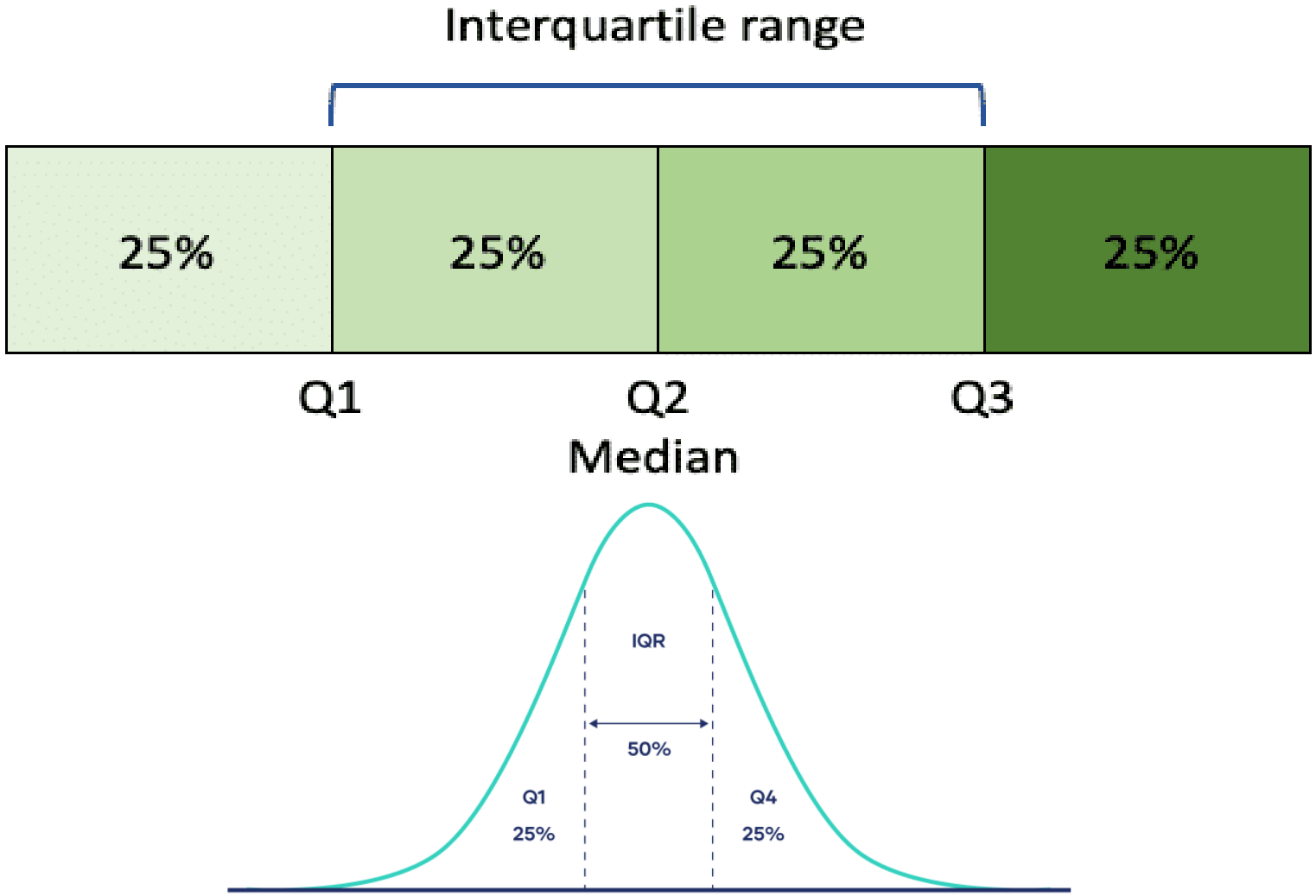
Mean $ 189,848.18

Task 1: The question is asking if this is a sample or a population. In other words, are those all the people in the US, receiving salaries? Obviously not. This is a sample, drawn from the population of all working people in the USA.

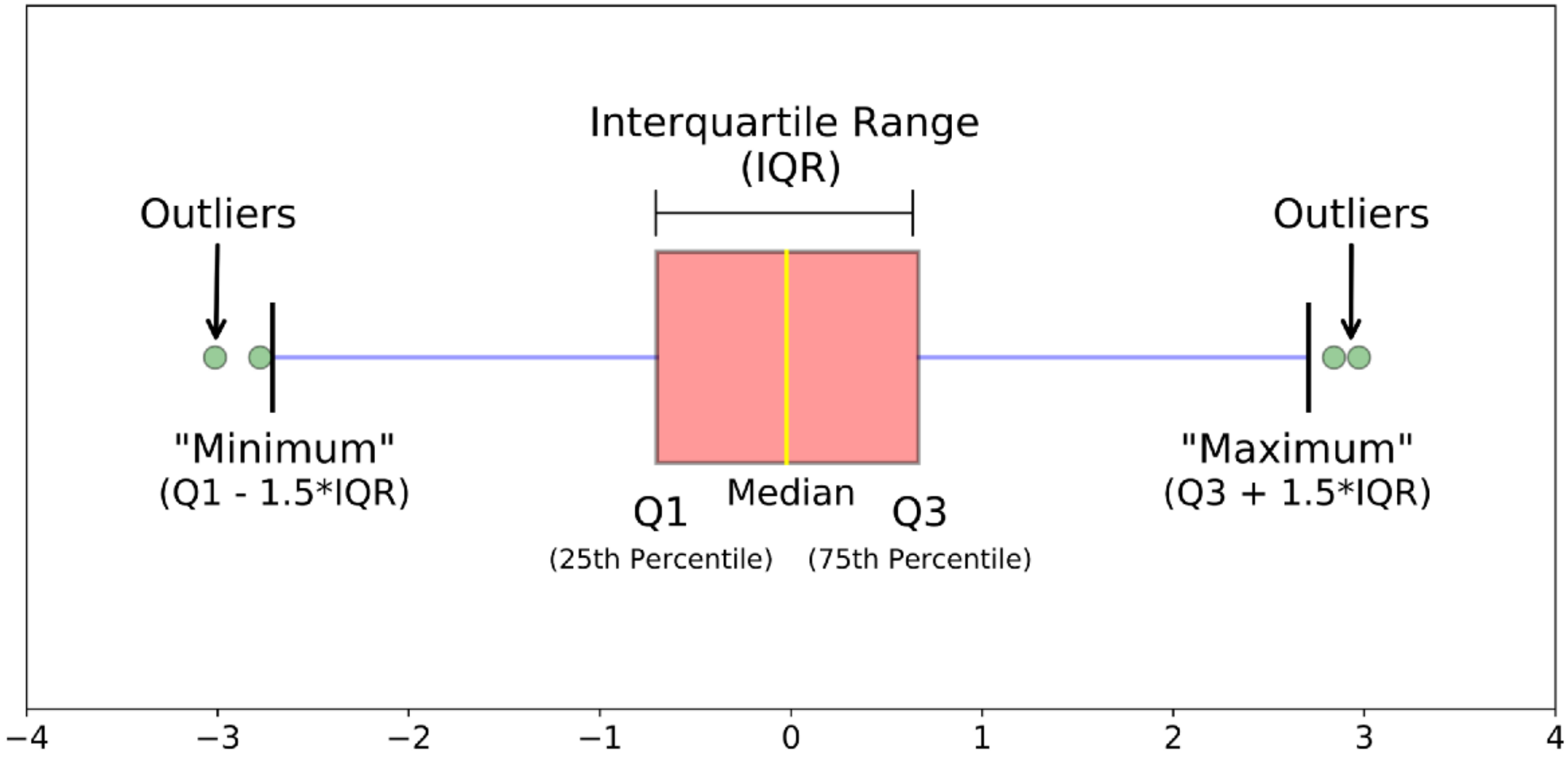
Task 2: Variance ^{2} 133,433,409,536.36

Task 3: There is great dispersion between the income of different people in the USA.

Interquartile Range (IQR)



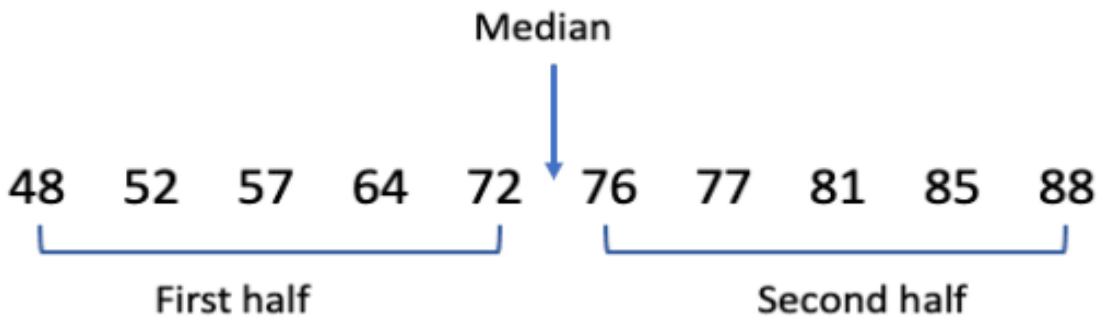
Box Plot


Interquartile Range
(IQR)
Outliers
Outliers
"Minimum"
(Q1 - 1.5*IQR)
"Maximum"
(Q3 + 1.5*IQR)
Q1
Median
Q3
(25th Percentile)
(75th Percentile)
−4
−3
−2
−1
0
1
2
3
4

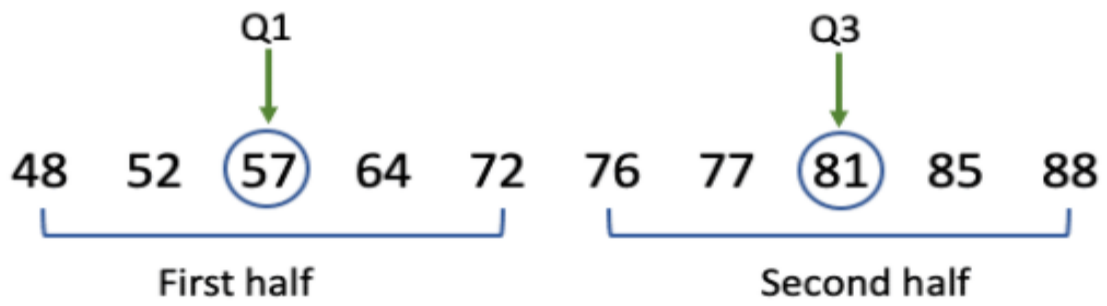
Step 1: Order your values from low to high.

48 52 57 64 72 76 77 81 85 88

Step 2: Locate the median, and then separate the values below it from the values above it.



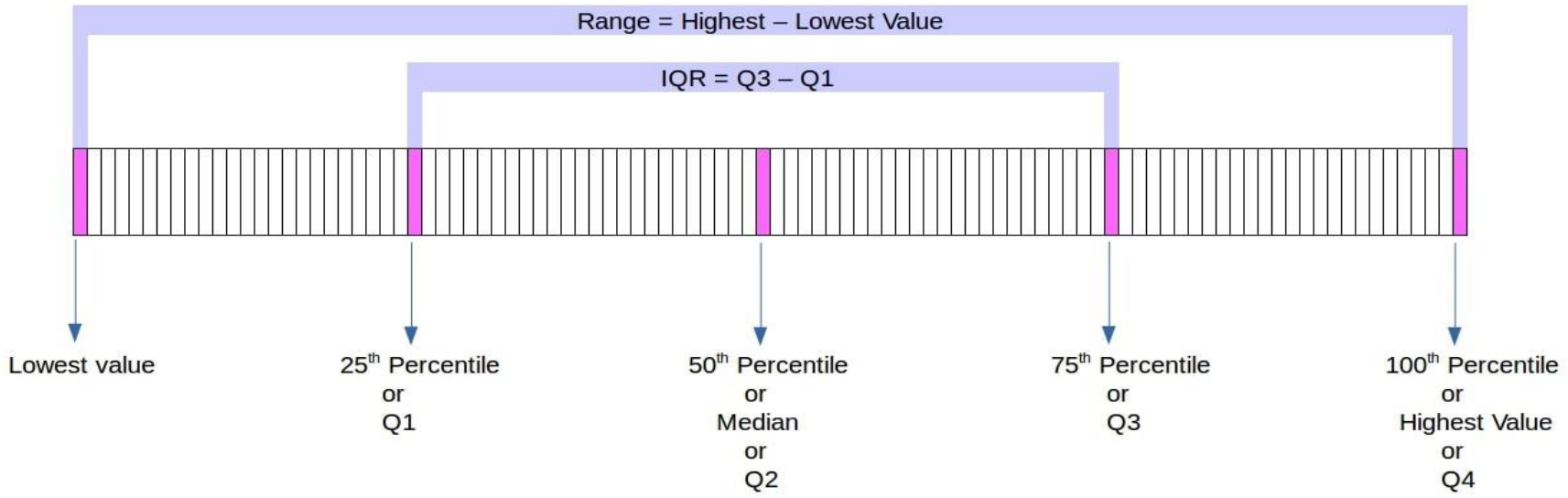
Step 3: Find Q1 and Q3.



Exercise

- Find Q1, Q2, Q3 for the following dataset. Identify any outlier, and draw a box plot

{5,40,42,46,48,50,50,50,52,52,55,56,58,75,102}



Solution

There are 15 values, arranged in increasing order. So, Q_2 is the 8^{th} data point, 50 .

Q_1 is the 4^{th} data point, 46 , and Q_3 is the 12^{th} data point, 56 .

The interquartile range IQR is $Q_3 - Q_1$ or $56 - 47 = 10$.

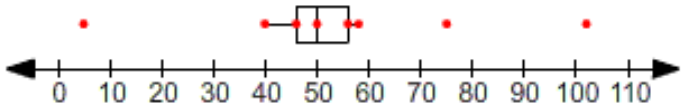
Now we need to find whether there are values less than $Q_1 - (1.5 \times \text{IQR})$ or greater than $Q_3 + (1.5 \times \text{IQR})$.

$$Q_1 - (1.5 \times \text{IQR}) = 46 - 15 = 31$$

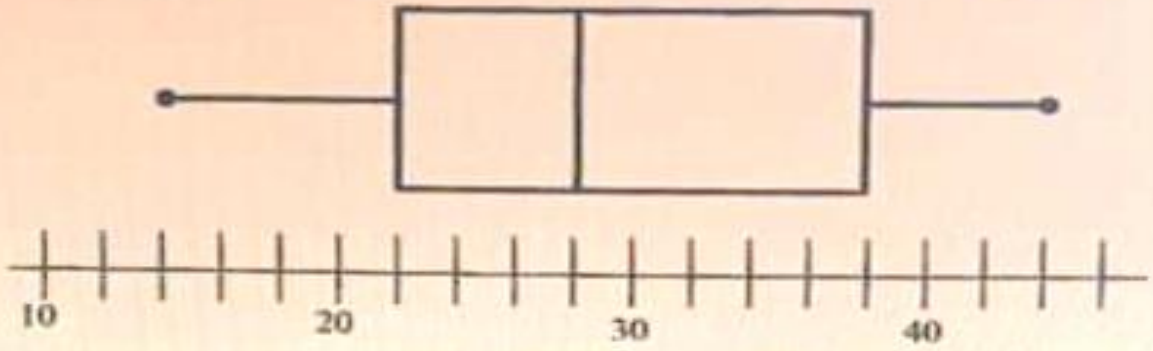
$$Q_3 + (1.5 \times \text{IQR}) = 56 + 15 = 71$$

Since 5 is less than 31 and 75 and 102 are greater than 71 , there are 3 outliers.

The box-and-whisker plot is as shown. Note that 40 and 58 are shown as the ends of the whiskers, with the outliers plotted separately.



Use the box plot to answer the questions below.



Amount of Time on the Phone

1) What is the median of the time spent on the phone?

a) 22 minutes b) 25 minutes c) 28 minutes d) 38 minutes

2) What is the difference between the range of the data and the Interquartile Range (IQR)?

a) 16 minutes b) 14 minutes c) 30 minutes d) 12 minutes

3) What percent of the people spent less than 22 minutes on the phone?

a) 25% b) 50% c) 75% d) not enough info

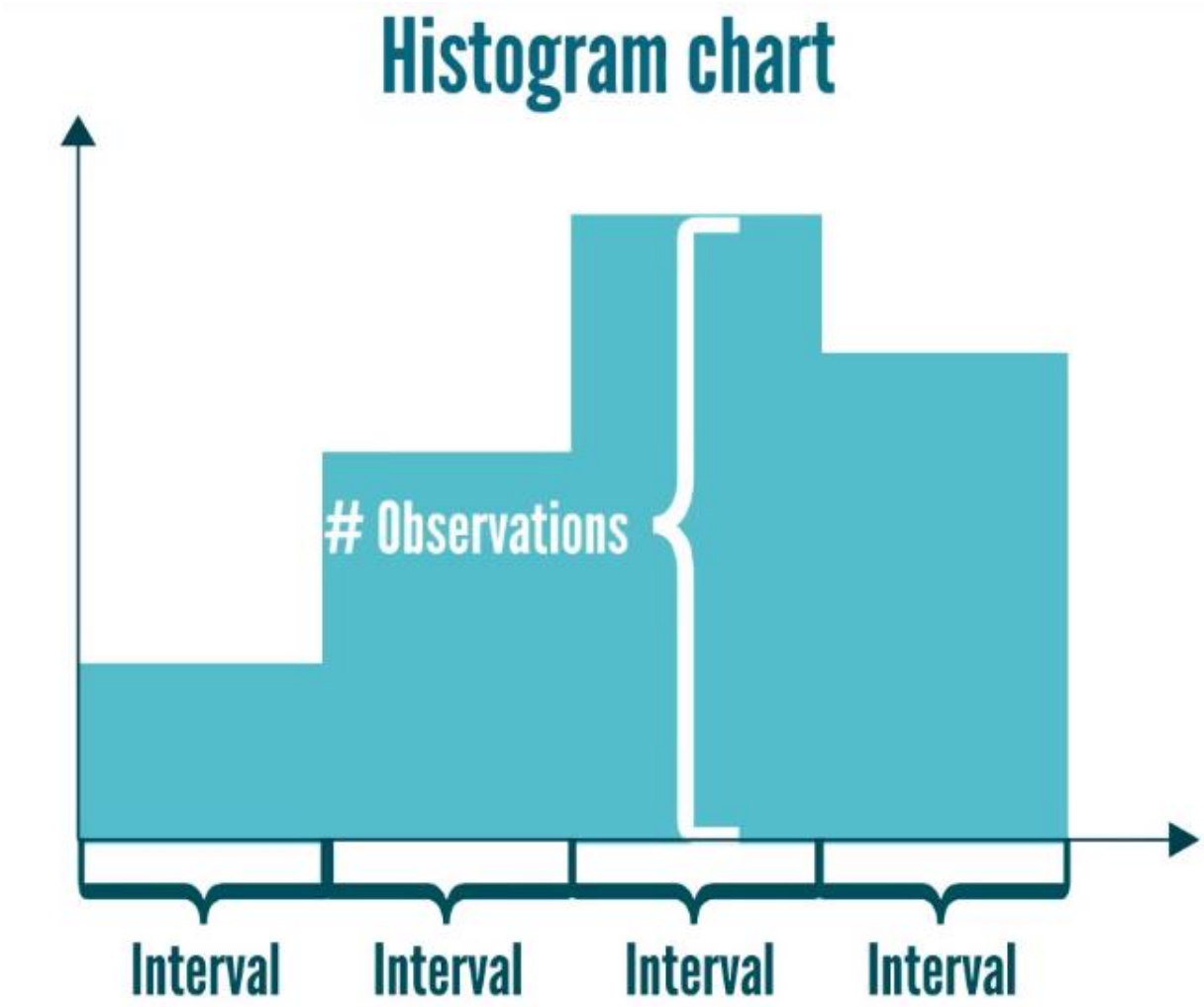
4) True or False? Less people spent 22-28 minutes than 28-38 minutes.

a) True b) False

5) What percent of the people spent less than 38 minutes on the phone?

a) 25% b) 50% c) 75% d) 100%

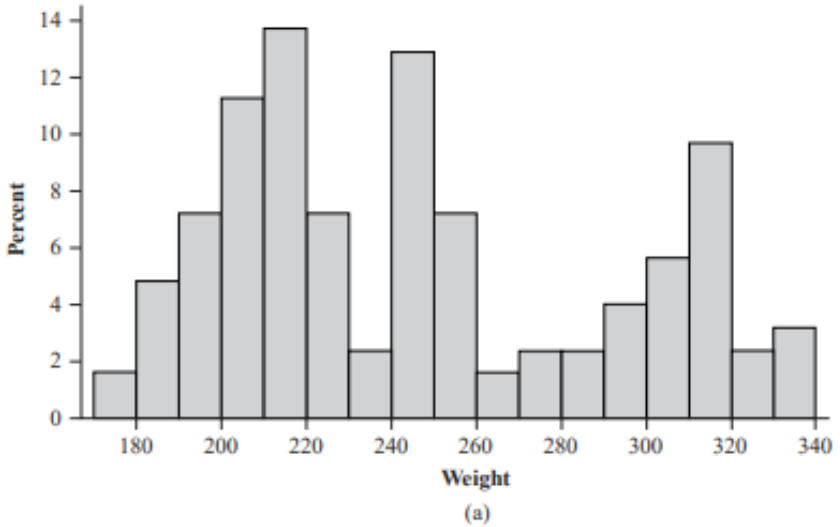
Histograms

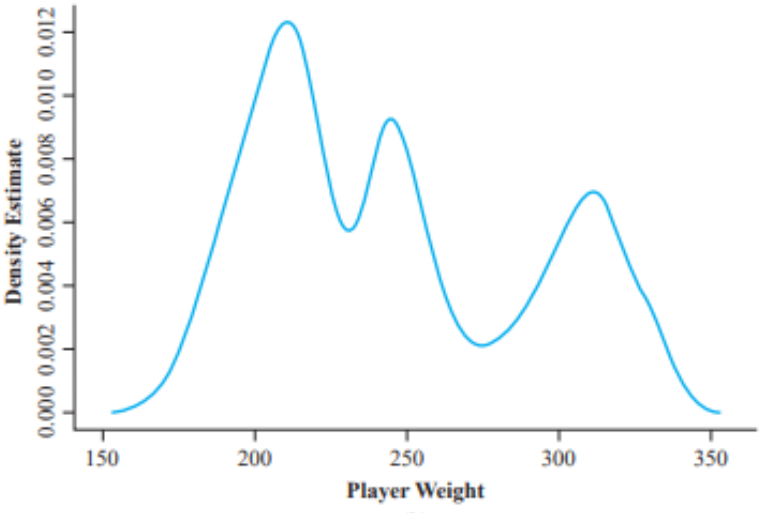


- Shows the distribution of a numeric variable
- The variable's range of values is split into intervals, represented by different bins
- The height of the bins shows the number of observations within an interval

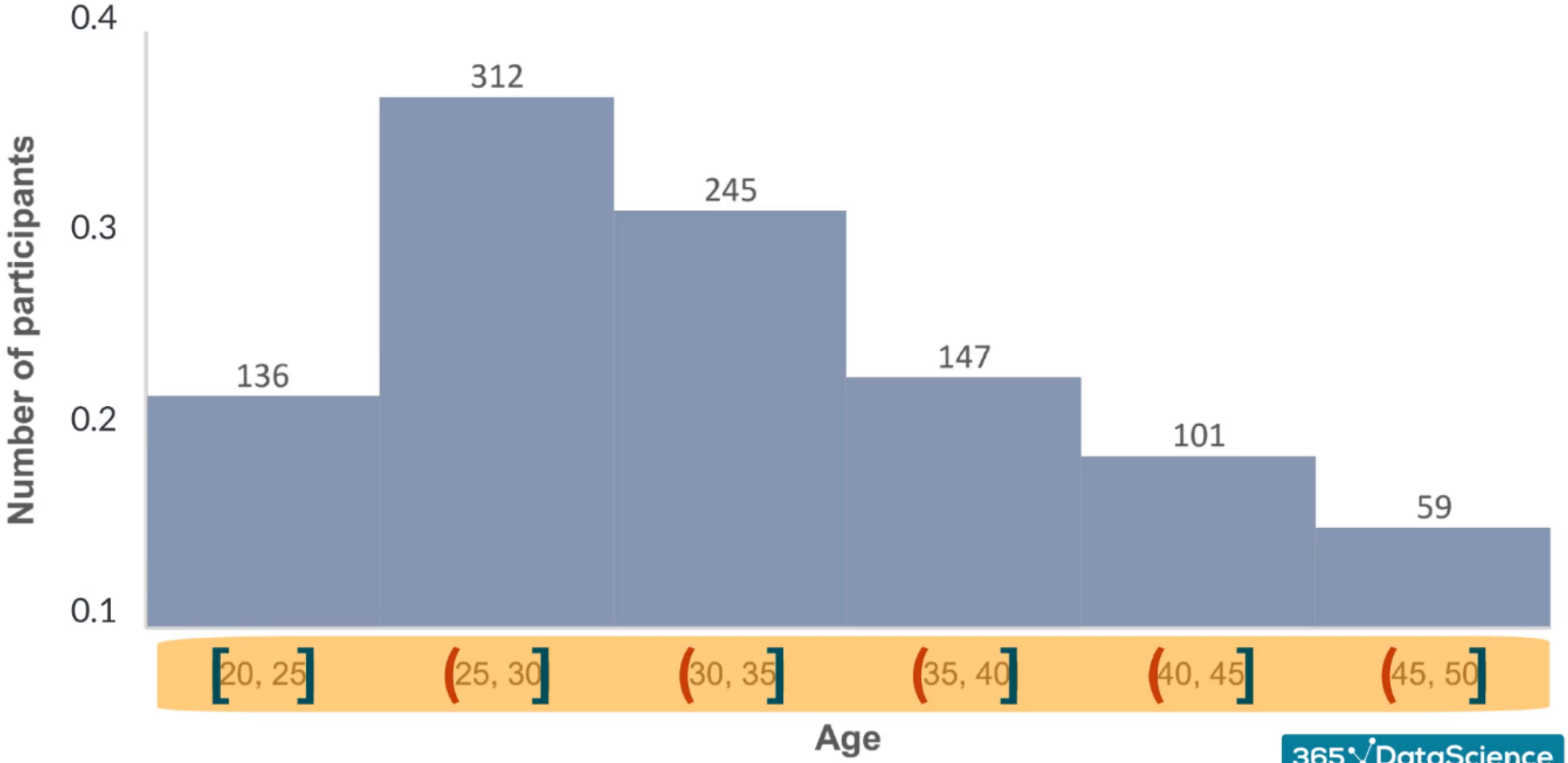
Histograms

- Histograms come in a variety of shapes.
- A unimodal histogram is one that rises to a single peak and then declines.
- A bimodal histogram has two different peaks
- A histogram with more than two peaks is said to be multimodal.





Age Distribution in Customers' Survey



Probability Basics

in fact…
Probability is the Bedrock of Machine Learning

- Classification models must predict a probability of class membership.
- Algorithms are designed using probability (e.g. Naive Bayes).
- Learning algorithms will make decisions using probability (e.g. information gain).
- Models are fit using probabilistic loss functions (e.g. log loss and cross entropy).
- Model hyperparameters are configured with probability (e.g. Bayesian optimization).
- Probabilistic measures are used to evaluate model skill (e.g. ROC).

Probability

- the study of randomness and uncertainty.
- In any situation in which one of a number of possible outcomes may occur, the discipline of probability provides methods for quantifying the chances, or likelihoods, associated with the various outcomes.
- Probability is simply how likely something is to happen.
- Whenever we're unsure about the outcome of an event, we can talk about the probabilities of certain outcomes—how likely they are.
- The analysis of events governed by probability is called **statistics**.

Probability

- Experiment: it is any activity or process whose outcome is subject to uncertainty
- The Sample Space of an Experiment: The sample space of an experiment, denoted by , is the set of all possible outcomes of that experiment.
- Events:An event is any collection (subset) of outcomes contained in the sample space .

Statistical experiment

Statistical experiments have three common features

1. Experiment has more than one possible outcome
2. Each outcome can be specified in advance
3. Outcome of the experiment depends on chance

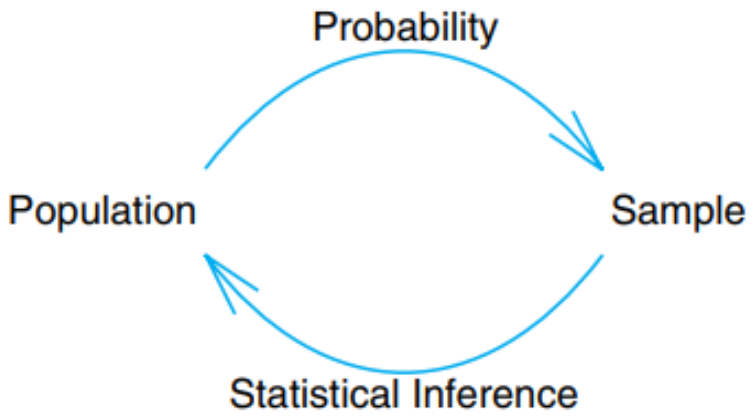
Probability

Statistical inference makes use of concepts in probability

Elements of probability allow us to quantify the strength or "confidence" in our conclusions

Probability provides the transition between descriptive statistics and inferential methods.

Elements in probability allow us to draw conclusions about characteristics of hypothetical data taken from the population, based on known features of the population



Probability

$$\text{Probability of an event} = \frac{\text{(\# of ways it can happen)}}{\text{(total number of outcomes)}}$$

- The probability of an event can only be between 0 and 1 and can also be written as a percentage.
- The probability of event A is often written as $P(A)$.
- If $P(A) > P(B)$, then event A has a higher chance of occurring than event B.
- If $P(A) = P(B)$, then events A and B are equally likely to occur.

1. You randomly draw a marble out of a bag that contains 20 total marbles. 12 of the marbles in the bag are blue.
What is P(draw a blue marble)

2. You roll a fair 6-sided die.
What is P(not 3)

3. The probability of Sandy flipping a coin twice and getting heads both times is 0.25. The probability of John's spinner stopping on the color red is 1/4.
Which of these events is more likely?

A. Sandy flips a coin twice and gets heads both times.
B. John's spinner stops on the color red.
C. Neither. Both events are equally likely.

- Anna and Charles have a bag that contains 1 black marble and 1 white marble. They are playing a game where they randomly select a marble out of the bag three times, with replacement.
- Anna thinks that the probability of getting at least two white marbles is greater than the probability of getting exactly two white marbles. Charles disagrees. He thinks that the two probabilities are equal.
- The sample space of possible outcomes is listed below. B represents a black marble, and W represents a white marble.
- **Who is correct, Anna or Charles?**

Rules of Probabilites

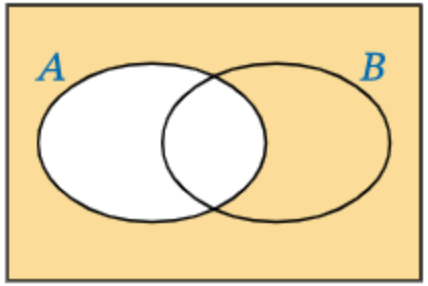
Rule 1. The probability of an event occurring is greater than or equal to 0 and less than or equal to 1:

$$0 \leq P(A) \leq 1.$$

Rule 2. The sum of the probabilities of all possible outcomes is equal to 1.

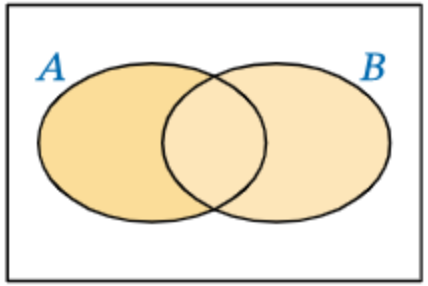
Rule 3. The probability of the complement, A', of event A, or "not A," is given by

$$P(A') = 1 - P(A).$$



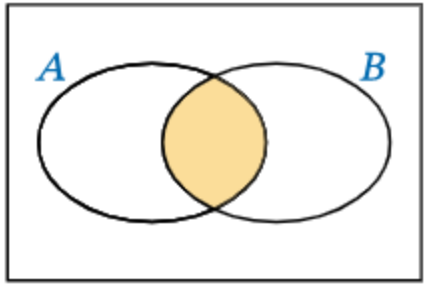
Rule 4. The probability of A or B occurring is called the "union" of A and B:

$$P(A \cup B).$$



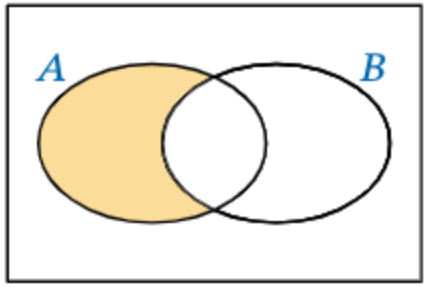
Rule 5. The probability of A and B occurring is called the "intersection" of A and B:

$$P(A \cap B).$$



Rule 6. The probability that event A occurs, but event B does not can be thought of as

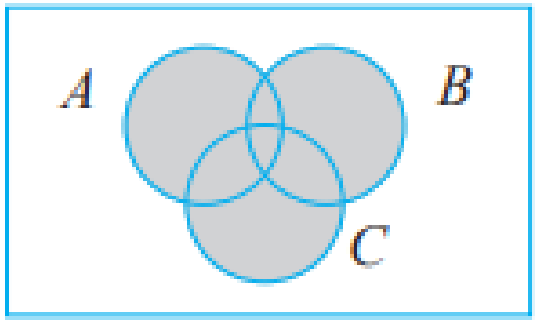
$$P(A \cap B') = P(A) - P(A \cap B).$$



Rules of Probabilites

For any three events A, B, and C,

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$$



Let X and Y be the following sets:

$X = \{15, 9, 11\}$

$Y = \{11, 9, 2\}$

Which of the following is the set $X \cap Y$?

Choose 1 answer:

- (A) $\{2, 9, 11, 15\}$
- (B) $\{2, 15\}$
- (C) $\{9, 11\}$
- (D) $\{\}$

Let X and Y be the following sets:

$X = \{4, 5, 7, 9\}$

$Y = \{\}$

Which of the following is the set $X \cap Y$?

Choose 1 answer:

- (A) $\{4, 5, 7, 9\}$
- (B) $\{7, 9\}$
- (C) $\{\}$
- (D) $\{4, 5\}$

Let X and Y be the following sets:

$X = \{\}$

$Y = \{2, 3, 5, 7, 11, 13\}$

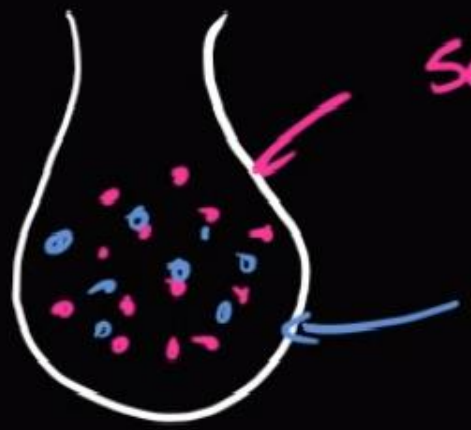
Which of the following is the set $X \setminus Y$?

Choose 1 answer:

- (A) $\{\}$
- (B) $\{2, 3, 5, 7, 11, 13\}$
- (C) $\{7, 11, 13\}$
- (D) $\{2, 3, 5\}$

Types of Probability

Experimental probability	Theoretical probability
6 appears 65 times. $P(\text{roll a } 6) = \frac{65}{100}$	6 appears. $P(\text{roll a } 6) = \frac{1}{6}$
	
Roll a die 100 times.	Roll a die once.



50 magenta

50 blue

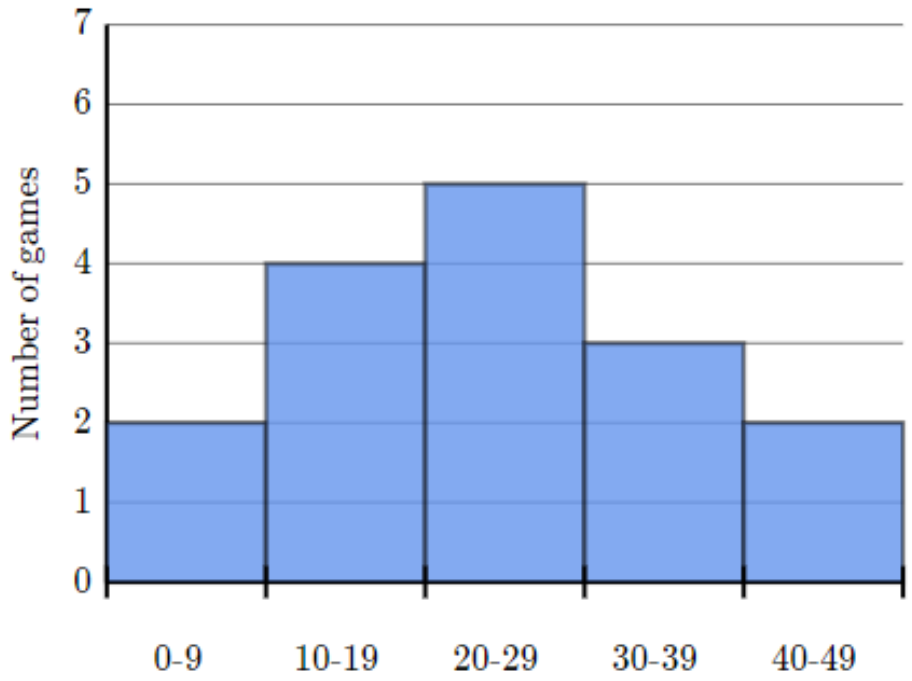
$P(\text{picking magenta}) = \frac{50}{100} = \frac{1}{2}$

After 10 experiments: 7 magenta 3 blue

After 10,000 experiments: 8,000 magenta 2,000 blue

Exp. prob. $= \frac{8,000}{10,000} = 80\%$

Coach Kelly documented the number of points the Ragin' Cajun football team scored each game this season. The following histogram summarizes the data.



Based on this data, what is a reasonable estimate of the probability that the Ragin' Cajun score 30 or more points in their next football game?

Choose the best answer.

Choose 1 answer:

(A) $\frac{3}{16}$

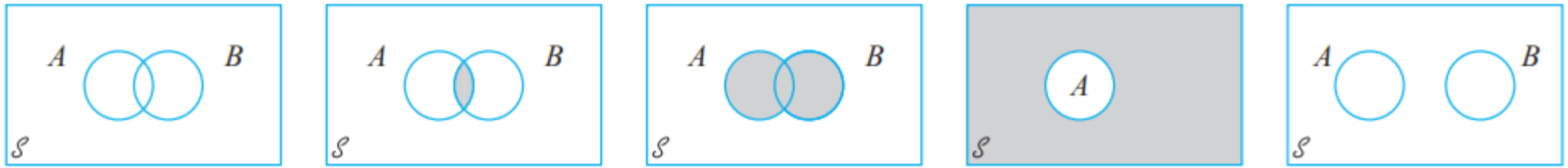
(B) $\frac{5}{16}$

(C) $\frac{5}{11}$

(D) $\frac{3}{11}$

Mutually Exclusive Events

- Let Ø denote the null event (the event consisting of no outcomes whatsoever).
- When $A \cap B = \emptyset$, A and B are said to be mutually exclusive or disjoint events.



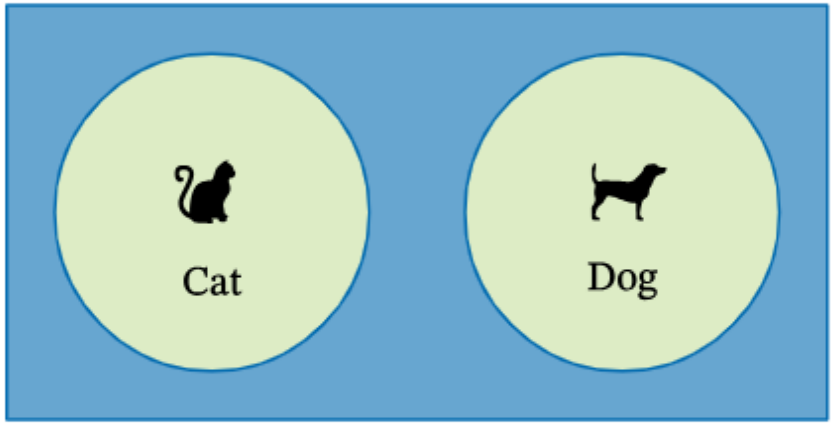
(a) Venn diagram of events A and B (b) Shaded region is $A \cap B$ (c) Shaded region is $A \cup B$ (d) Shaded region is A' (e) Mutually exclusive events

Figure 2.1 Venn diagrams

Mutually Exclusive Events

Mutually Exclusive disjoint events

An animal cannot be both a cat and a dog: "being a cat" and "being a dog" are mutually exclusive events.

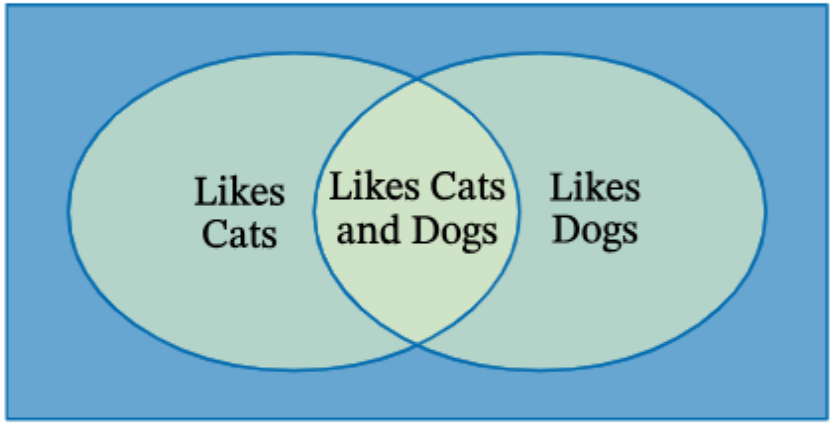


Rule 7. In general, for two mutually exclusive events:

$$P(\varnothing) = 0 \text{ where } \varnothing \text{ is the null event}$$
$$P(A \cap B) = 0,$$
$$P(A \cup B) = P(A) + P(B).$$

Non-Mutually Exclusive

A person may like both cats and dogs, so "likes cats" and "likes dogs" are not mutually exclusive events.



Rule 8. In general, for two non-mutually exclusive events:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

Mutually Exclusive Events

Addition Rule for Probability

Mutually Exclusive Events

$$P(X \cup Y) = P(X) + P(Y)$$

Non Mutually Exclusive Events

$$P(X \cup Y) = P(X) + P(Y) - P(X \cap Y)$$

Counting Techniques

- many experiments for which the effort involved in constructing such a list is prohibitive because N is quite large.

The Product Rule for Ordered Pairs

- If the first element or object of an ordered pair can be selected in n1 ways, and for each of these n1 ways the second element of the pair can be selected in n2 ways, then the number of pairs is n1xn2.

The Product Rule for Ordered Pairs

- A homeowner doing some remodeling requires the services of both a plumbing contractor and an electrical contractor.
- If there are 12 plumbing contractors and 9 electrical contractors available in the area, in how many ways can the contractors be chosen?
- If we denote the plumbers by P1, . . . , P12 and the electricians by Q1, . . . , Q9, then we wish the number of pairs of the form.
- With n1=12 and n2=9, the product rule yields N = (12)(9) = 108 possible ways of choosing the two types of contractors.

Permutations and Combinations

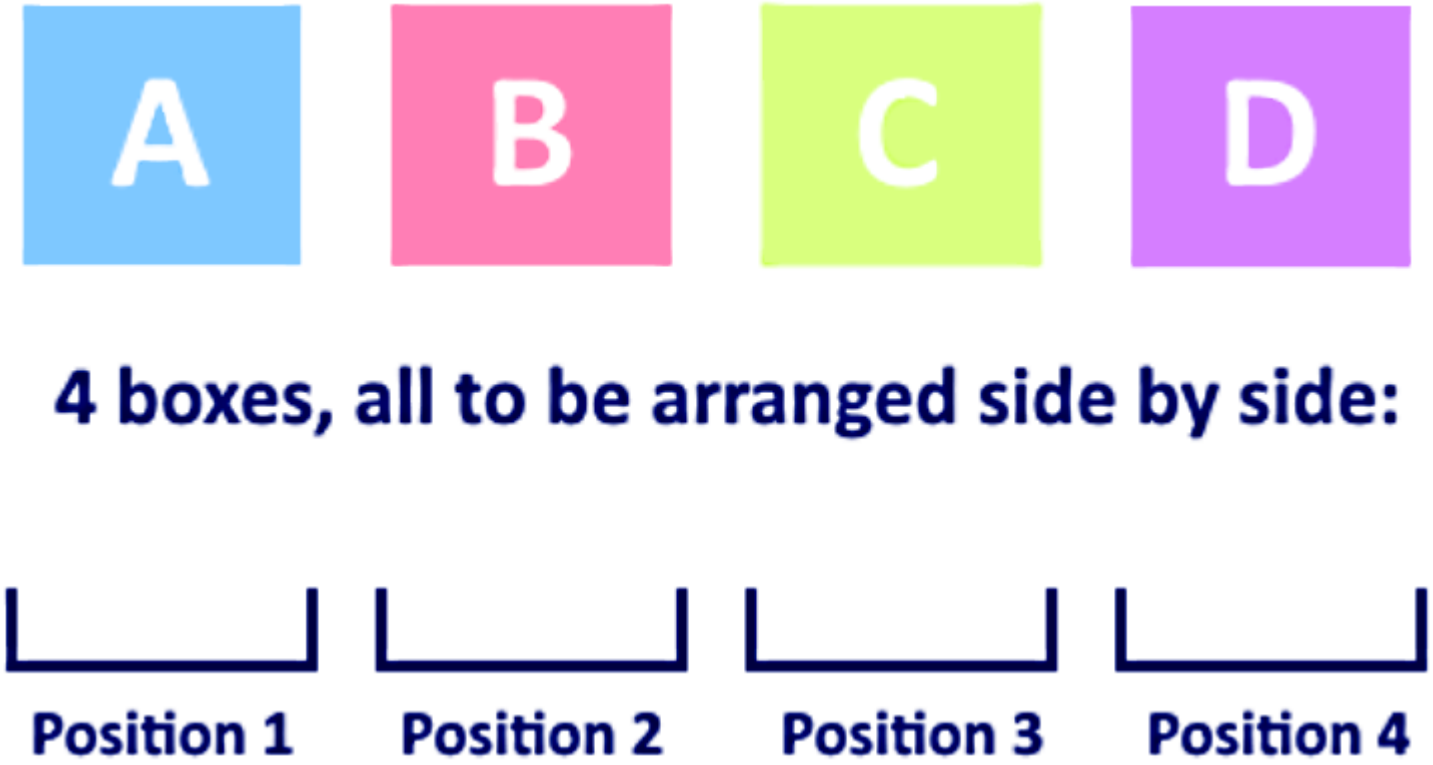
- An ordered subset is called a permutation. The number of permutations of size k that can be formed from the n individuals or objects in a group will be denoted by $P_{k\ n}$.

$$P_{k,\,n} = \frac{n!}{(n - k)!}$$

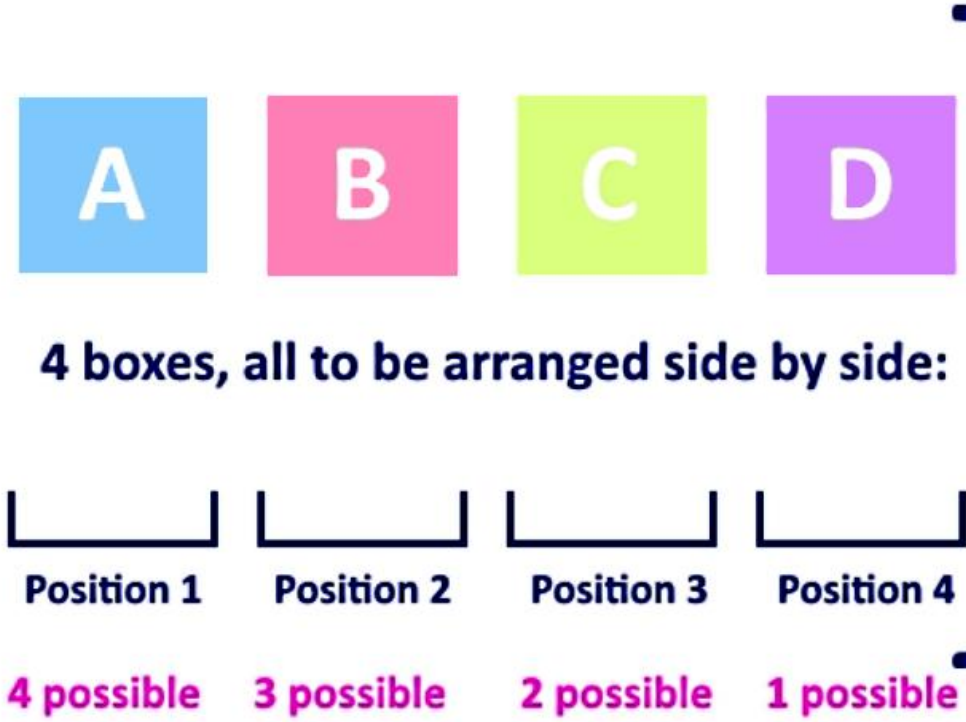
- An unordered subset is called a combination. One way to denote the number of combinations is $C_{k,n}$

$$\frac{n!}{k!(n - k)!}$$

Permutations and Combinations



Permutations and Combinations



Possible Arrangements:

A B C D	B A C D	C A B D	D A B C
A C B D	B A D C	C A D B	D A C B
A B D C	B C A D	C B A D	D B A C
A C D B	B C D A	C B D A	D B C A
A D B C	B D C A	C D B A	D C A B
A D C B	B D A C	C D A B	D C B A

$4 \times 3 \times 2 \times 1 = 4! \rightarrow$ 24 possible arrangements!

Permutations and Combinations

4 boxes, 2 of them (any of them) to be arranged; therefore 2 positions are needed:

Position 1 Position 2

4 possible 3 possible

n = 4

r = 2

Possible Arrangements:

A B	B A	C A	D A
A C	B C	C B	D B
A D	B D	C D	D C

$$4 \times 3 = \frac{n!}{(n-r)!} = 12 \rightarrow$$ 12 possible arrangements!

Permutations and Combinations

Using 2 out of 4 boxes in a set:

Possible arrangements: 12

A B B A C A D A
A C B C C B D B
A D B D C D D C

PERMUTATIONS

Possible selections of distinct items: 6

A B = B A B C = C B
A C = C A B D = D B
A D = D A C D = D C

COMBINATIONS

Exercise

Selecting 4 fruits out of 10 fruits

Selecting 4 fruits out of 10 fruits

$$^{n}C_{r} = \frac{n!}{r!(n-r)!}$$

$$^{10}C_{4} = C(n, r) = C(10, 4)$$

$$= \frac{10!}{(4!(10-4)!)}$$

$$= \frac{10!}{4! \times 6!}$$

$$= 210 \text{ ways}$$

Conditional Probability

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Conditional Probability

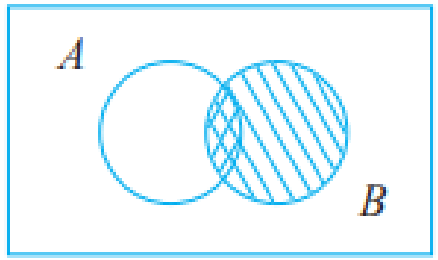
If the probability of an event B is affected by the occurrence of an event A, then we say that the probability of B is conditional on the occurrence of A. The notation for conditional probability is $P(B \mid A)$, read as "the probability of B given A."

The Probability of A given B

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

The Probability of A and B (intersection)

The Probability of B



Conditional Probability

- What is the probability that a selected person is a smoker given it is male?

	Male	Female	Total
Smoker	45	25	70
Nonsmoker	75	55	130
Total	120	80	200

  - We have to find **P(smoker | male)**
  - By using the probability of A given B formula, **P(A | B) = P(A∩B) / P(B)**.
  - Using this, we can write:

P(smoker | male) = P (smoker ∩ male) / P(male)
= (45/200) / (120/200)
= 45/120
= 9/24

The Multiplication Rule

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

The Multiplication Rule

$$P(A \cap B) = P(A \mid B) \cdot P(B)$$

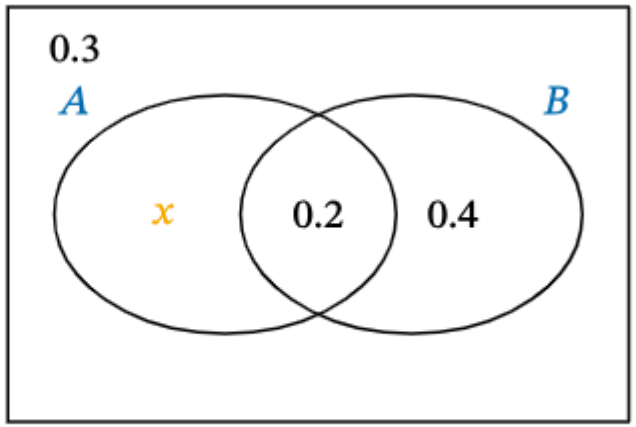
Example 4: Using Venn Diagrams to Calculate Dependent Probabilities

The figure shows a Venn diagram with some of the probabilities given for two events A and B.

1. Work out $P(A \cap B')$.
2. Work out $P(A)$.
3. Work out $P(B \mid A)$.

When given a Venn diagram with an unknown probability, it is usually a good idea to find any missing probabilities, if possible.

In this case, we only have one missing probability. Let's label this with the variable x.

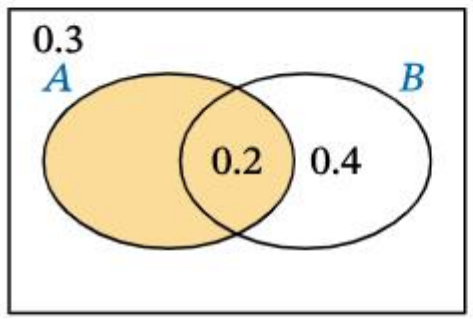


Example 4 (Continued)

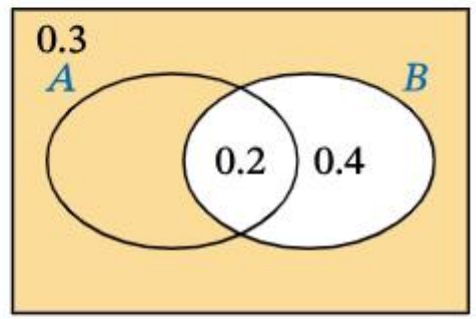
1. Work out $P(A \cap B')$.

As it turns out, finding the probability of the missing section aligns with part 1 of our question, since it is the intersection of $P(A)$ and $P(B')$.

$P(A)$

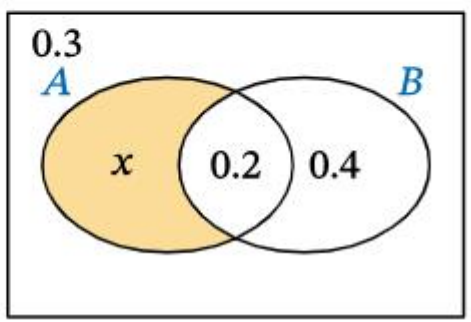


$P(B')$



intersection

$P(A \cap B')$



We can find this probability by recalling the sum of the total probabilities is equal to 1.

Since we have labelled the only unknown probability, we can equate all of the probabilities on our Venn diagram to 1:

$$x + 0.3 + 0.2 + 0.4 = 1$$

$$x = 1 - 0.9$$

$$x = 0.1$$ Solving for x.

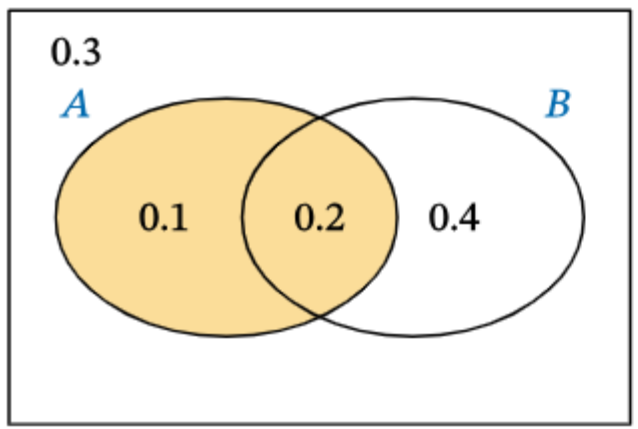
$P(A \cap B') = 0.1$ And remembering x is the probability of $P(A \cap B')$.

Example 4 (Continued)

2. Work out $P(A)$.

Part 2 of our question is a straightforward task, since we can use the information we found in part 1.

$P(A)$



It should be clear from our diagram that

$$P(A) = 0.1 + 0.2$$

$$P(A) = 0.3$$

Note: We could formally say this calculation is $P(A) = P(A \cap B') + P(A \cap B)$.

Example 4 (Continued)

3. Work out $P(B \mid A)$.

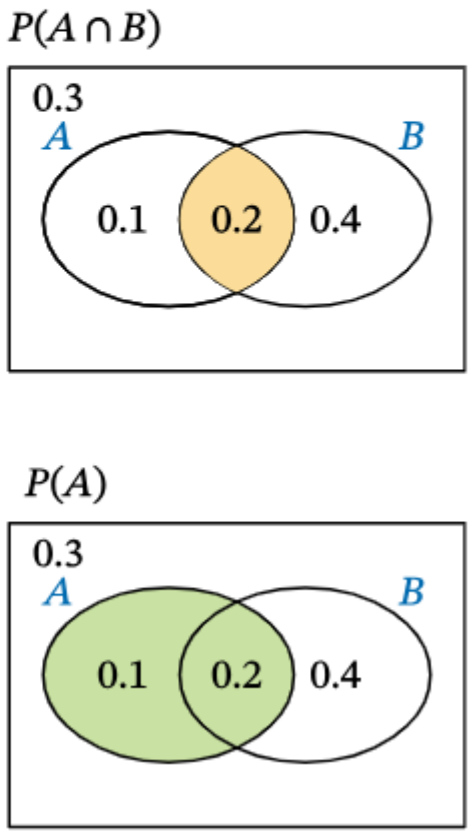
The final part of this question involves conditional probabilities, and we can use the familiar formula modified for $P(B \mid A)$. It might also be useful to pay attention to the Venn diagram representation of the terms used in our formula.

The value of $P(A \cap B)$ can be directly taken from the Venn diagram, and we have just found $P(A)$ during part 2 of the question.

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)}$$
$$= \frac{0.2}{0.3}$$
$$= \frac{2}{3} \approx 0.667$$

We substitute in these values and then simplify to reach an answer.

$$P(A \mid B) = \frac{P(A \cap B)}{P(A)}$$



Independent and Dependent Events

An important idea in probability theory is the independence of events. Two events A and B are

- **independent** if the fact that A occurs does not affect the probability of B occurring,
- **dependent** if the fact that A occurs does affect the probability of B occurring.

Independent

If a coin is tossed repeatedly, each throw is an independent event. What happened in a previous throw does not affect the result of the next throw.

Dependent

We have already seen an example of a dependent event. If two cards are taken from a pack and the second card is taken without the first card being replaced, then probabilities for the second card are affected by the first card.

Independent and Dependent Events (Continued)

For **Independent** events, $P(A \cap B) = P(A) \times P(B)$.
But if we rearrange the formula for conditional probability, we find that

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

$$P(A \cap B) = P(A \mid B) \times P(B).$$

Combining these two equations gives us the following:

Rule 9 (For Independent Events Only)

$$P(A \mid B) \times P(B) = P(A) \times P(B)$$

$$P(A \mid B) = P(A)$$

and we can follow a similar method for $P(B \mid A)$ to find

$$P(B \mid A) = P(B)$$

This gives us a useful tool to check the independence of events using conditional probabilities.

If both of these equations are true, we can say that events A and B are independent.

Probability

Multiplication Rule

Independent Events

$$P(X \cap Y) = P(X) \bullet P(Y)$$

Dependent Events

$$P(X \cap Y) = P(Y) \bullet P(X \mid Y)$$

Bayes' Theorem

$$P(X \mid Y) = \frac{P(X \cap Y)}{P(Y)}$$

Example 3: Determining Whether Two Events Are Independent Using Conditional Probability

Suppose $P(A) = \frac{2}{5}$ and $P(B) = \frac{3}{7}$. The probability that event A occurs and event B also occurs is $\frac{1}{5}$.

$P(A \cap B)$

Calculate $P(A \mid B)$, and then evaluate whether events A and B are independent.

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

The question has given us all the information we need to calculate conditional probability using the formula.

$$= \frac{\frac{1}{5}}{\frac{3}{7}}$$

$$= \frac{1}{5} \times \frac{7}{3}$$

$$= \frac{7}{15}$$

Substituting in the values we have and simplifying, we can find the conditional probability of A given B.

Example 3 (Continued)

If A and B are independent events, then $P(A \mid B) = P(A)$ and $P(B \mid A) = P(B)$.

$$P(A \mid B) = \frac{7}{15}$$

We have just found $P(A \mid B)$ and the question gives us $P(A)$.

$$P(A) = \frac{2}{5} = \frac{6}{15}$$

This can be expressed with a denominator of 15 for direct comparison of the two fractions.

We have now shown that

$$P(A \mid B) \neq P(A).$$

Since one of the conditions for independence is not true, we conclude that A and B are not independent events.

Key Points

- If the probability of an event B is affected by the occurrence of an event A, then we say that the probability of B is conditional on the occurrence of A. The notation for conditional probability is $P(B \mid A)$, read as the probability of B given A.
- For two events, A and B, we can use Venn diagrams, and the conditional probability formula to help work out the conditional probability of A given B. The formula is given by

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}.$$

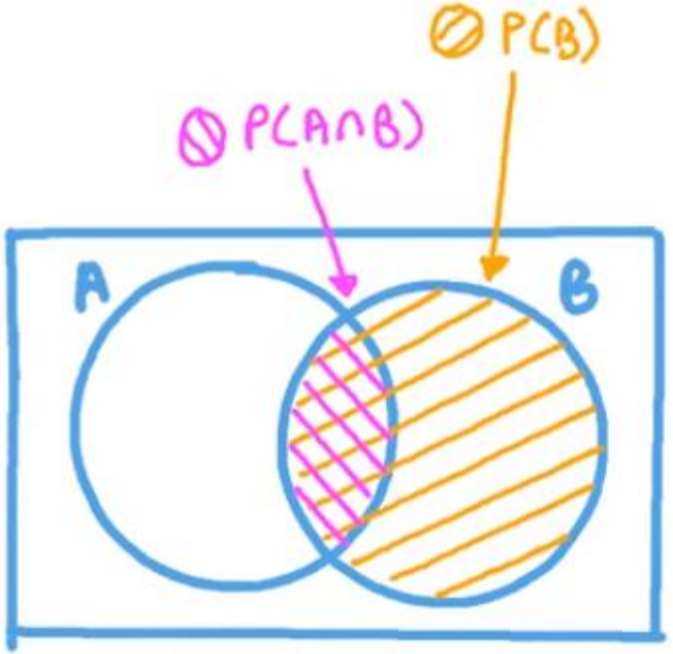
- We can use conditional probabilities to help us determine whether two events are dependent or independent.
- Events A and B are independent if

$$P(A \mid B) = P(A)$$

and

$$P(B \mid A) = P(B).$$

CONDITIONAL PROBABILITY



$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

If $P(A|B) = P(A)$, and $P(B|A) = P(B)$, then events A and B are INDEPENDENT

BAYES' THEOREM

THE PROBABILITY OF A HYPOTHESIS, **H** CONDITIONAL ON A NEW PIECE OF EVIDENCE, **E**

$$P(H \mid E) = \frac{P(E \mid H)\, P(H)}{P(E)}$$

PROBABILITY OF THE EVIDENCE GIVEN THE HYPOTHESIS

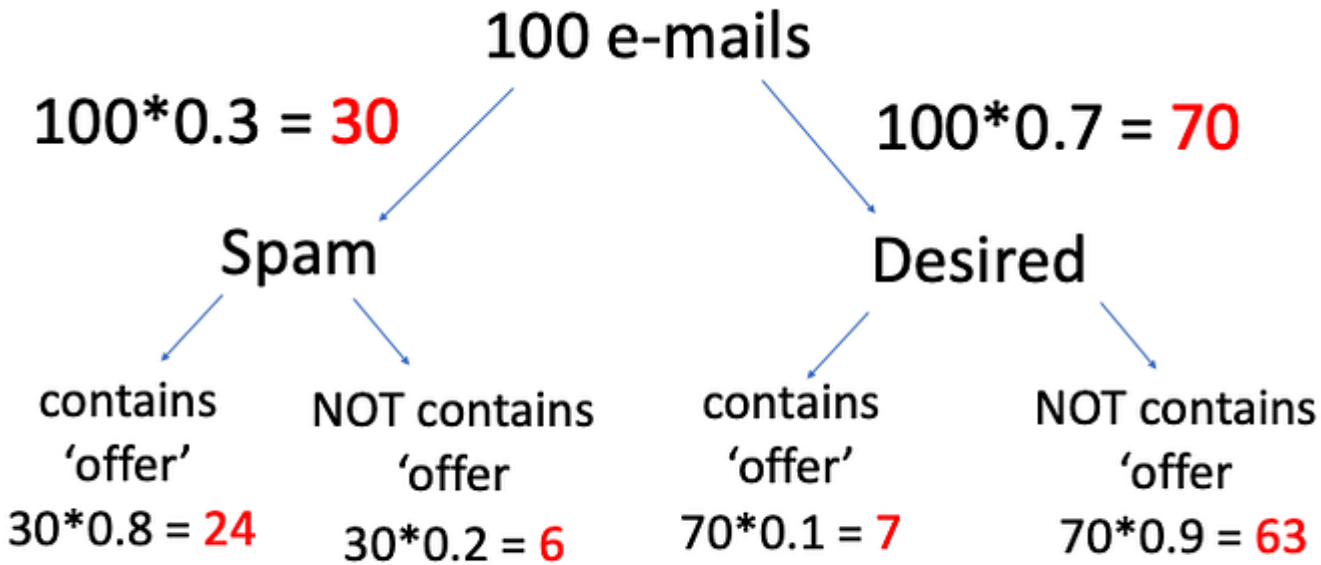
THE PRIOR PROBABILITY OF THE HYPOTHESIS

THE PRIOR PROBABILITY OF THE EVIDENCE

Problem 1

- Let's work on a simple **NLP problem with Bayes Theorem**. By using **NLP**, I can detect **spam e-mails** in my inbox.
- Now, I assume that I received **100 e-mails**.
- Assume that the word **'offer'** occurs in **80%** of the **spam** messages in my account. Also, let's assume **'offer'** occurs in **10%** of my **desired** e-mails. If **30%** of the received e-mails are considered as a **spam**, and I will receive a new message which contains '**offer**', what is the probability that it is **spam**?

$$P(spam|contains\ offer) = \frac{P(contains\ offer|spam) * P(spam)}{P(contains\ offer)}$$



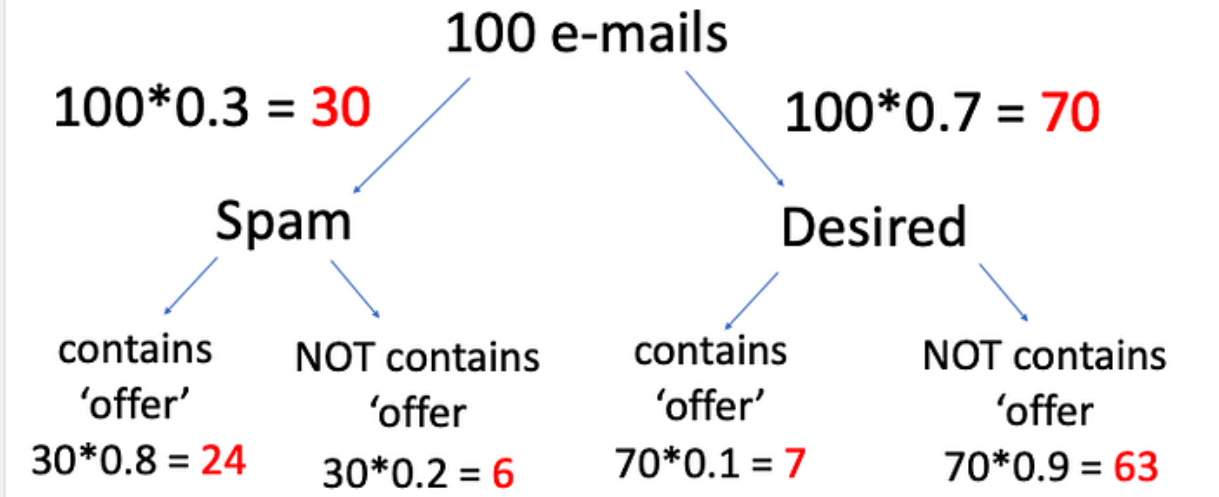
Solution with Bayes' Equation:

A = Spam

B = Contains the word 'offer'

$$P(spam|contains\ offer) = \frac{P(contains\ offer|spam) * P(spam)}{P(contains\ offer)}$$

$$P(spam|contains\ offer) = \frac{P(contains\ offer|spam) * P(spam)}{P(contains\ offer)}$$



P(contains offer|spam) = 0.8 (given in the question)

P(spam) = 0.3 (given in the question)

Now we will find the probability of e-mail with the word 'offer'. We can compute that by adding 'offer' in spam and desired e-mails. Such that;

P(contains offer) = 0.3*0.8 + 0.7*0.1 = 0.31

$$P(spam|contains\ offer) = \frac{0.8 * 0.3}{0.31} = 0.774$$

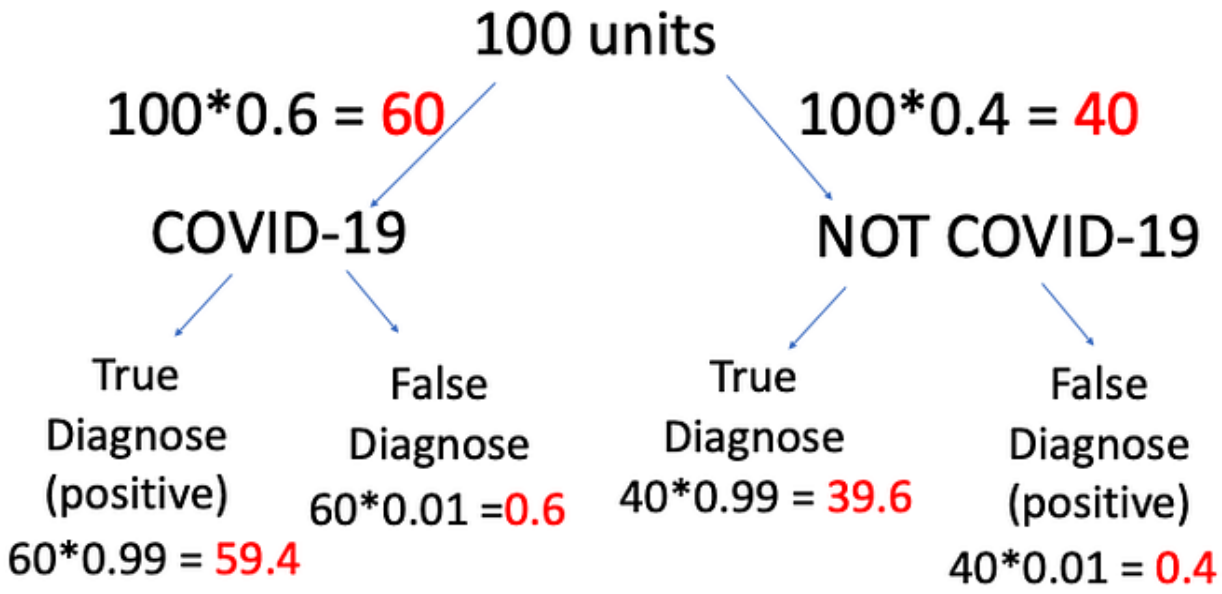
Problem 2

As you know, Covid-19 tests are common nowadays, but some results of tests are not true. Let's assume; a diagnostic test has **99%** accuracy and **60%** of all people have Covid-19. If a patient tests **positive**, what is the probability that they actually have the **disease**?

$$P(covid19|positive) = \frac{P(positive|covid19) * P(covid19)}{P(positive)}$$

Problem 2

As you know, Covid-19 tests are common nowadays, but some results of tests are not true. Let's assume; a diagnostic test has **99%** accuracy and **60%** of all people have Covid-19. If a patient tests **positive**, what is the probability that they actually have the **disease**?



With Bayes';

$$P(covid19|positive) = \frac{P(positive|covid19) * P(covid19)}{P(positive)}$$

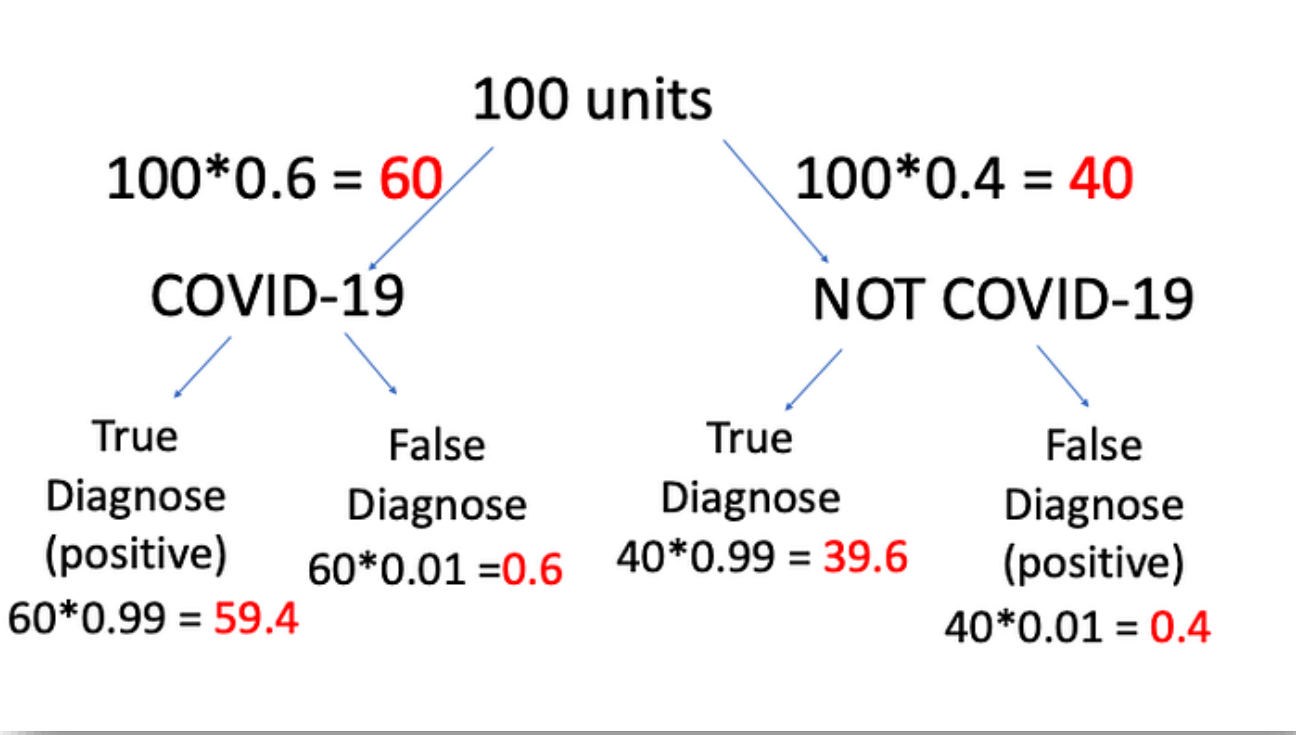


image by author

P(positive|covid19) = 0.99

P(covid19) = 0.6

P(positive) = 0.6*0.99+0.4*0.01=0.598

$$P(covid19|positive) = \frac{0.99 * 0.6}{0.598} = 0.993$$

Naïve Bayes Example

Outlook	Temp	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

$P(x \mid c) = P(Sunny \mid Yes) = 3/9 = 0.33$

Frequency Table		Play Golf	
		Yes	No
Outlook	Sunny	3	2
	Overcast	4	0
	Rainy	2	3

Likelihood Table		Play Golf		
		Yes	No	
Outlook	Sunny	3/9	2/5	5/14
	Overcast	4/9	0/5	4/14
	Rainy	2/9	3/5	5/14
		9/14	5/14	

$P(x) = P(Sunny) = 5/14 = 0.36$

$P(c) = P(Yes) = 9/14 = 0.64$

Posterior Probability: $P(c \mid x) = P(Yes \mid Sunny) = 0.33 \times 0.64 \div 0.36 = 0.60$

$P(x \mid c) = P(Sunny \mid No) = 2/5 = 0.4$

Frequency Table		Play Golf	
		Yes	No
Outlook	Sunny	3	2
	Overcast	4	0
	Rainy	2	3

		Play Golf		
		Yes	No	
Outlook	Sunny	3	2	5
	Overcast	4	0	4
	Rainy	2	3	5
		9	5	14

$P(x) = P(Sunny) = 5/14 = 0.36$

$P(c) = P(No) = 5/14 = 0.36$

Posterior Probability: $P(c \mid x) = P(No \mid Sunny) = 0.40 \times 0.36 \div 0.36 = 0.40$

Frequency Table

		Play Golf	
		Yes	No
Outlook	Sunny	**3**	**2**
	Overcast	**4**	**0**
	Rainy	**2**	**3**

		Play Golf	
		Yes	No
Humidity	High	**3**	**4**
	Normal	**6**	**1**

		Play Golf	
		Yes	No
Temp.	Hot	**2**	**2**
	Mild	**4**	**2**
	Cool	**3**	**1**

		Play Golf	
		Yes	No
Windy	False	**6**	**2**
	True	**3**	**3**

Likelihood Table

		Play Golf	
		Yes	No
Outlook	Sunny	**3/9**	**2/5**
	Overcast	**4/9**	**0/5**
	Rainy	**2/9**	**3/5**

		Play Golf	
		Yes	No
Humidity	High	**3/9**	**4/5**
	Normal	**6/9**	**1/5**

		Play Golf	
		Yes	No
Temp.	Hot	**2/9**	**2/5**
	Mild	**4/9**	**2/5**
	Cool	**3/9**	**1/5**

		Play Golf	
		Yes	No
Windy	False	**6/9**	**2/5**
	True	**3/9**	**3/5**

Outlook	Temp	Humidity	Windy	Play
Rainy	Cool	High	True	?

$$P(Yes \mid X) = P(Rainy \mid Yes) \times P(Cool \mid Yes) \times P(High \mid Yes) \times P(True \mid Yes) \times P(Yes)$$

$$P(Yes \mid X) = 2/9 \times 3/9 \times 3/9 \times 3/9 \times 9/14 = 0.00529 \rightarrow 0.2 = \frac{0.00529}{0.02057 + 0.00529}$$

$$P(No \mid X) = P(Rainy \mid No) \times P(Cool \mid No) \times P(High \mid No) \times P(True \mid No) \times P(No)$$

$$P(No \mid X) = 3/5 \times 1/5 \times 4/5 \times 3/5 \times 5/14 = 0.02057 \rightarrow 0.8 = \frac{0.02057}{0.02057 + 0.00529}$$

Random Variables

Ordinary variable

What color is the ball?

- Color is a variable
- Not a random variable

Random variable

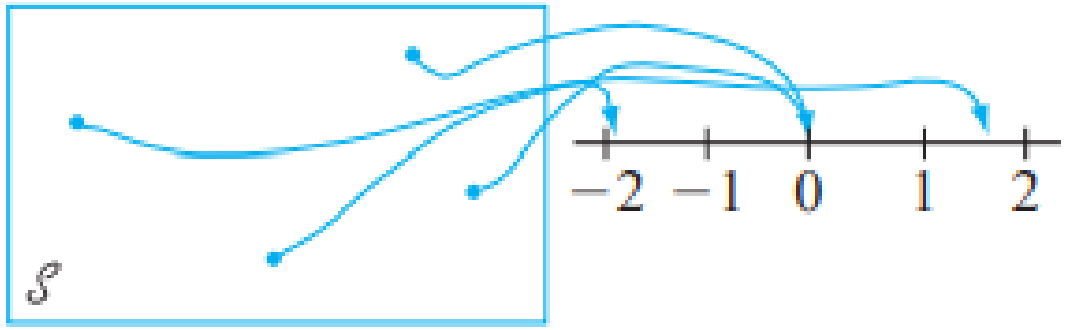
What color is the ball?

- Color is based on chance
- Color is random variable

Random Variables

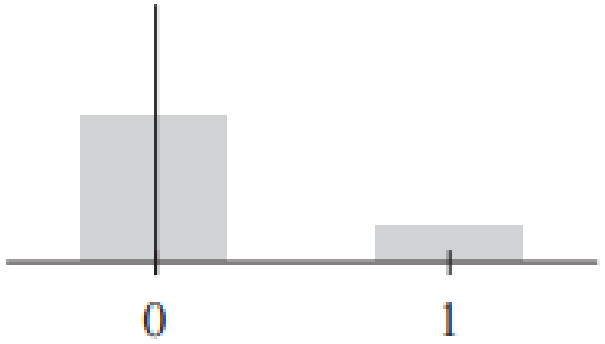
- Whether an experiment yields qualitative or quantitative outcomes, methods of statistical analysis require that we focus on certain numerical aspects of the data (such as a sample proportion, mean, or standard deviations).
- For a given sample space of some experiment, a random variable (rv) is any rule that associates a number with each outcome in space.
- There are two fundamentally different types of random variables—discrete random variables and continuous random variables.

$$p(0) = \text{the probability of the } X \text{ value } 0 = P(X = 0)$$
$$p(1) = \text{the probability of the } X \text{ value } 1 = P(X = 1)$$



Discrete random variables

- Any random variable whose only possible values are 0 and 1 is called a Bernoulli random variable.



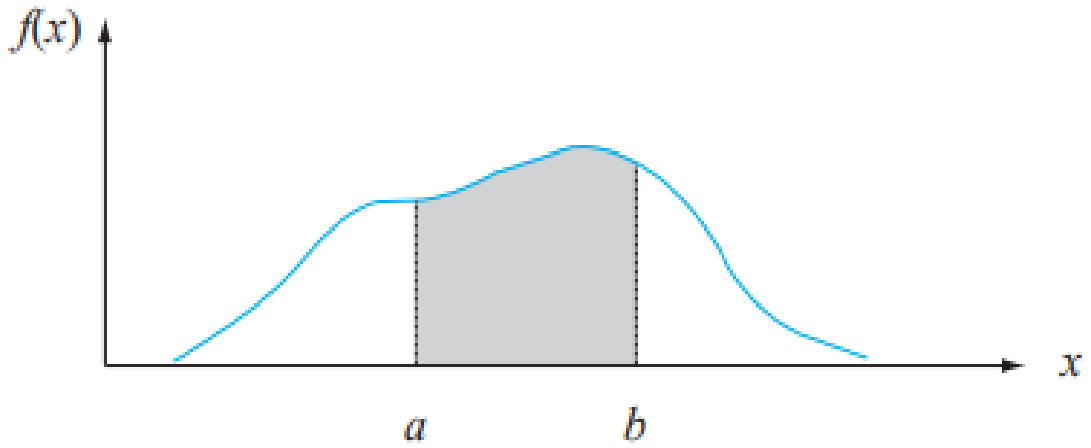
Continuous random variables

Problem

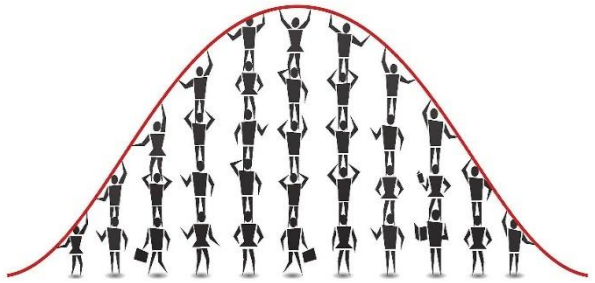
- Randomly choose one teenager
- Find age of that person
- Discrete or continuous?

Solution

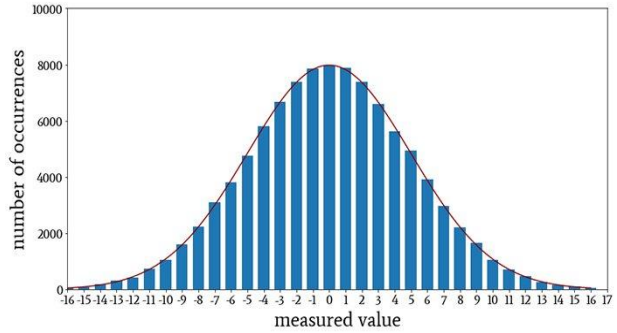
- Age can be any value
- Age is a continuous variable



$P(a \leq X \leq b)$ = the area under the density curve between a and b

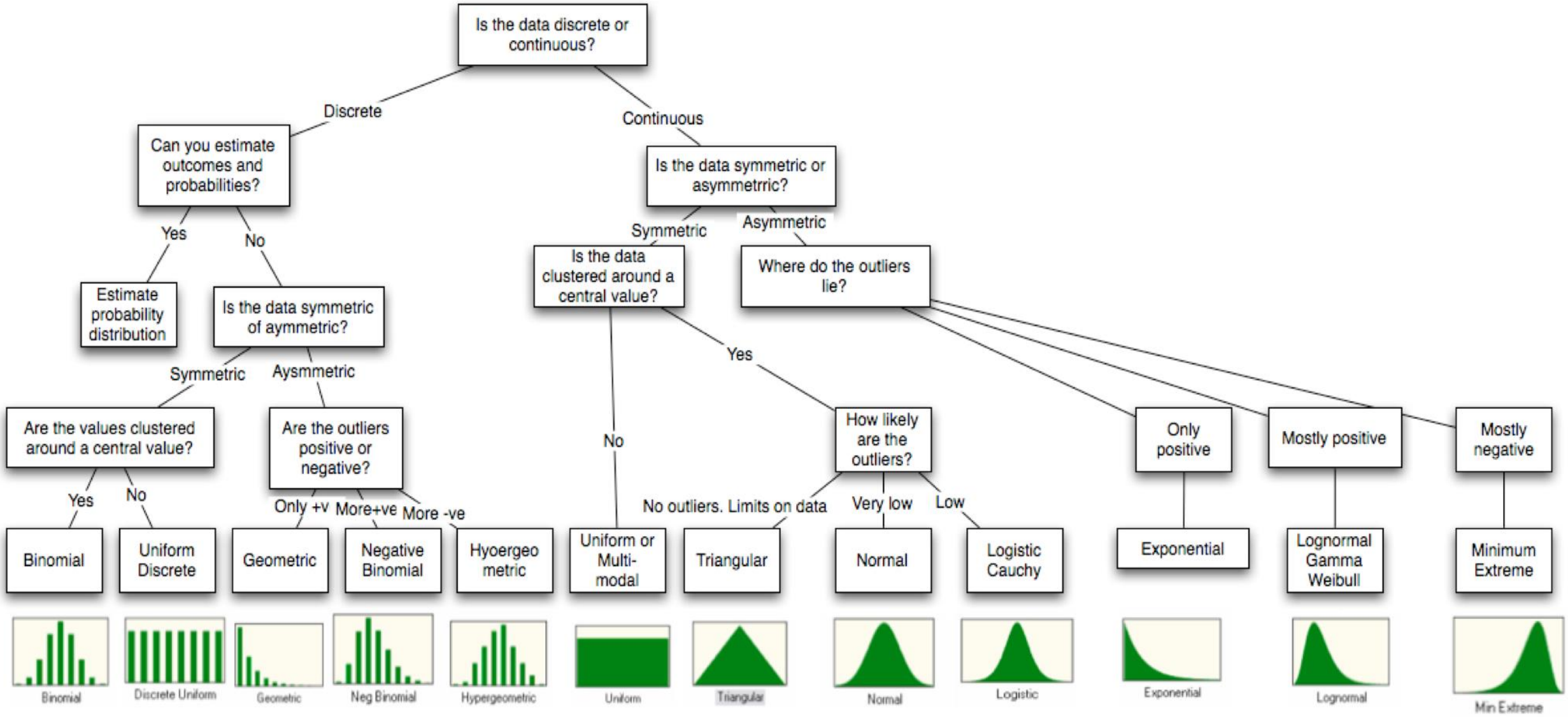


ProbabilityDistribution



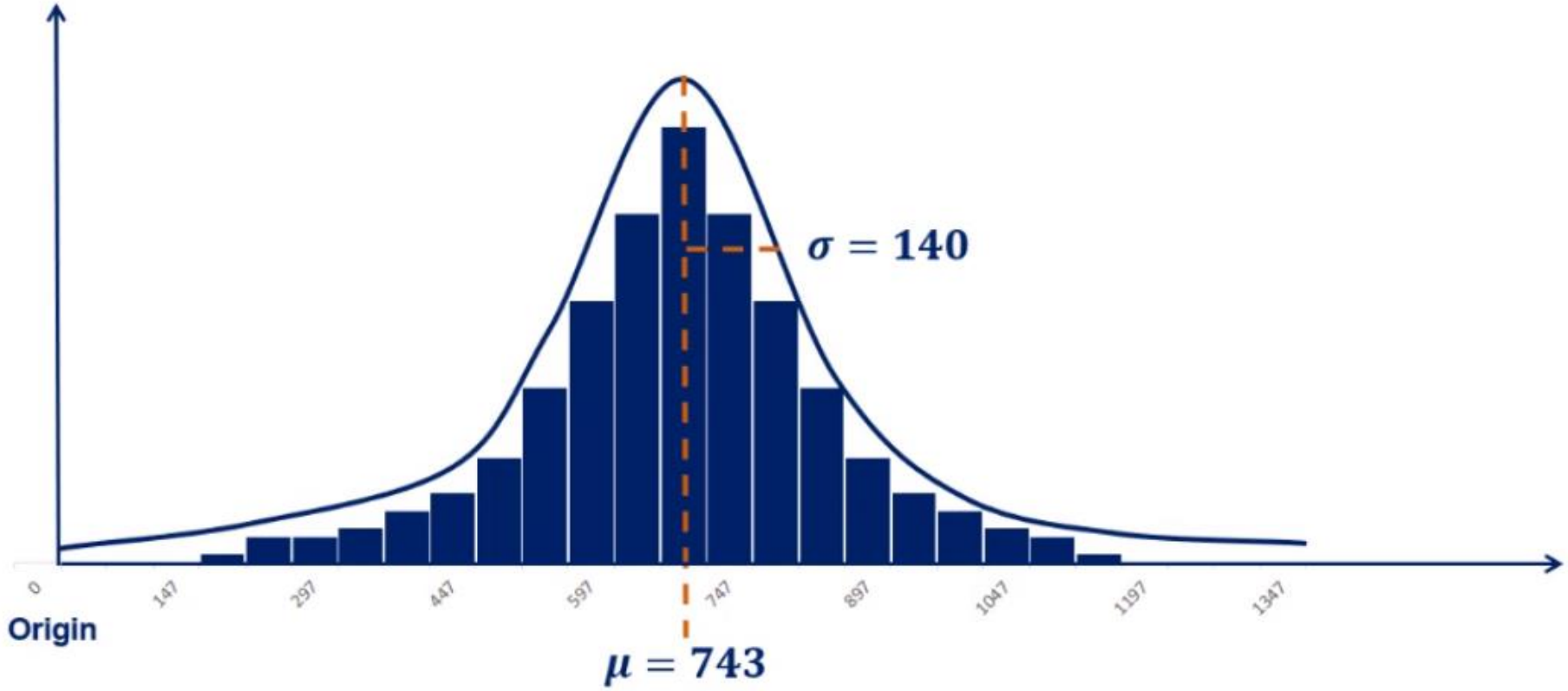
- A probability distribution is a way to represent the possible values and the respective probabilities of a random variable.
- There are two types of probability distributions:
  - Discrete probability distribution for discrete random variables.
  - Continuous probability distribution for continuous random variables.
- We can directly calculate probabilites of a discrete random variable, X = x, as the proportion of times the x value occurs in the random process.
- Probabilites of a continuous random variable taking on a specific value (e.g. Y = y) are not directly measureable. Instead, we calculate the probability as the proportion of times $y \in [a,b]$
- Probability mass functions (pmf) are used to describe discrete probability distributions. While probability density functions (pdf) are used to describe continuous probability distributions.
- By assuming a random variable follows an established probability distribution, we can use its derived pmf/pdf and established principles to answer questions we have about the data.

Distribution

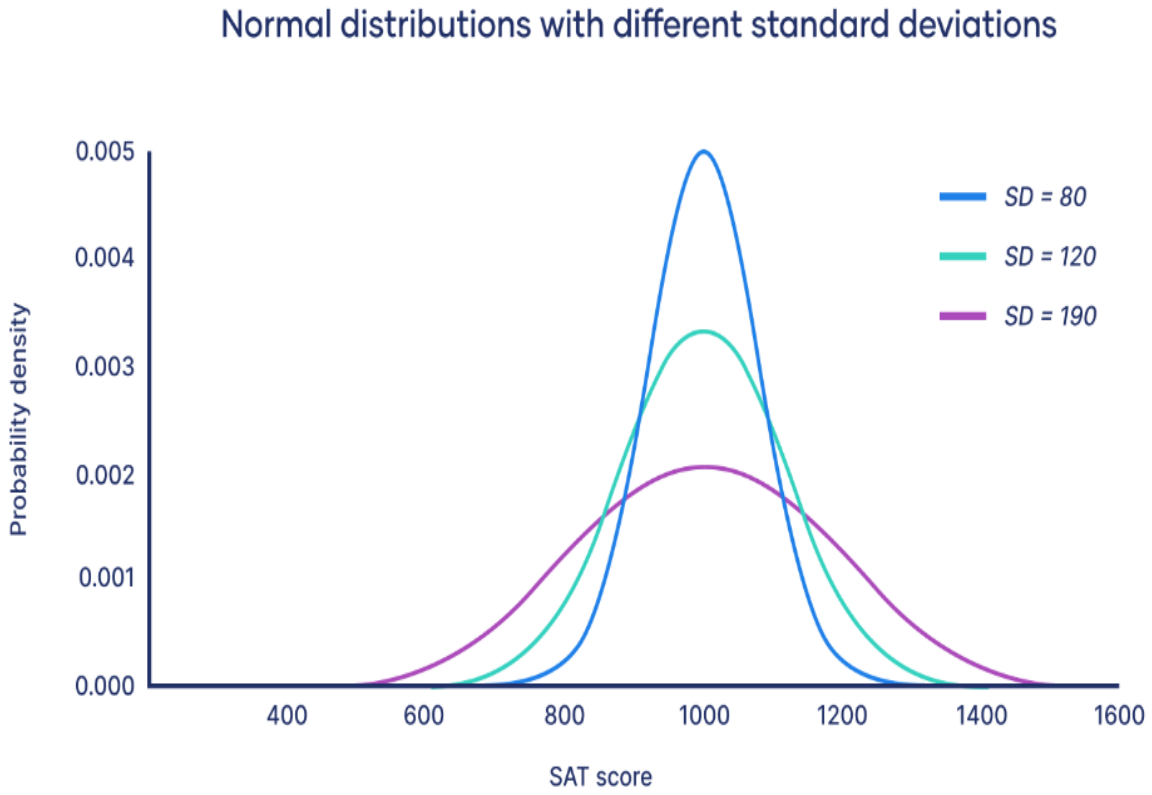


Normal Distribution

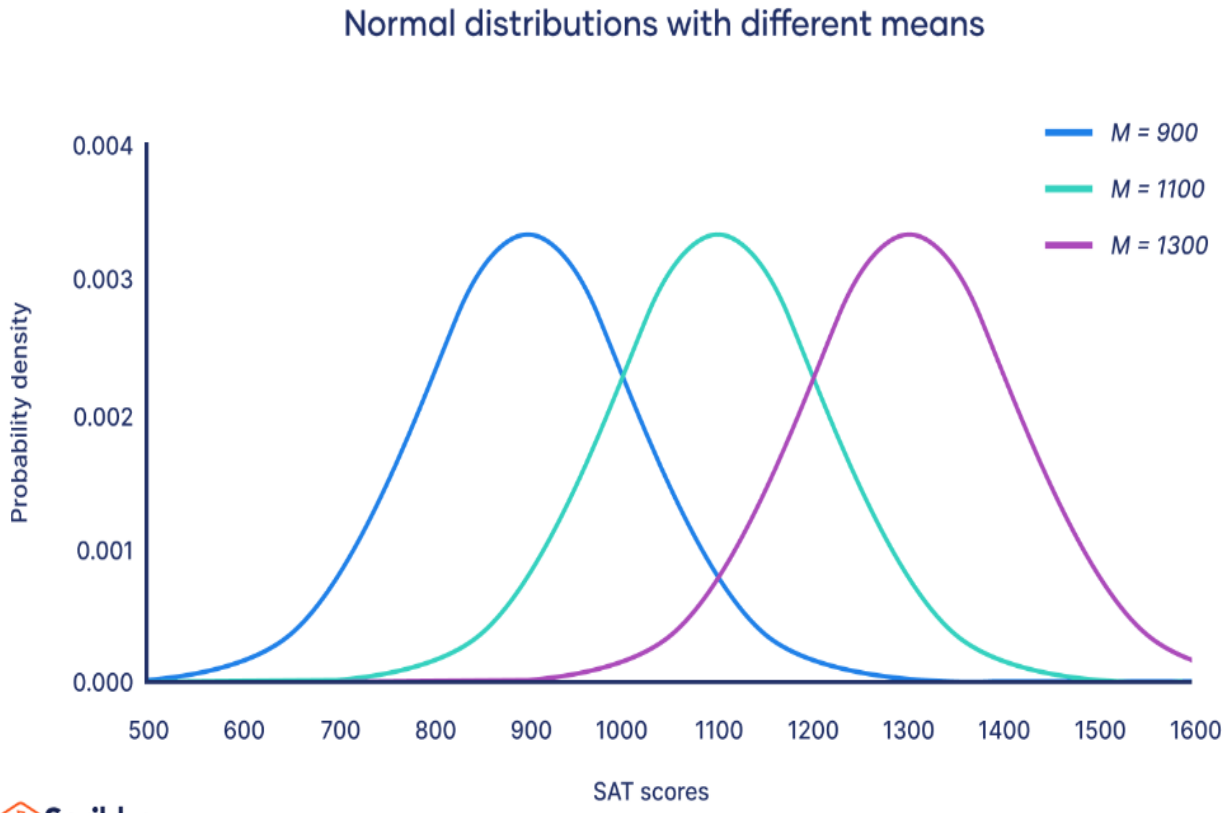
- In a normal distribution, data is **symmetrically** distributed with no skew. When plotted on a graph, the data follows a **bell shape**, with most values clustering around a central region and tapering off as they go further away from the center.



Normal Distribution

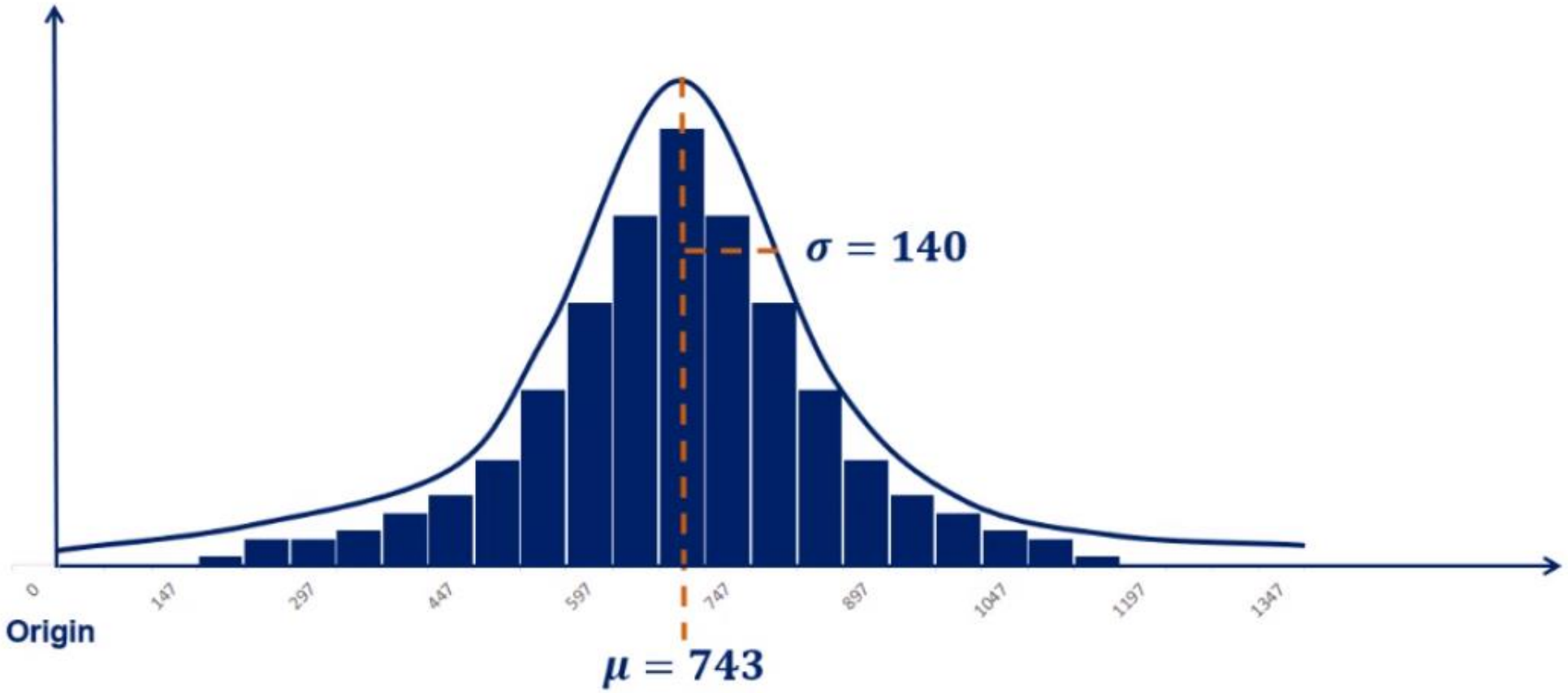


Scribbr

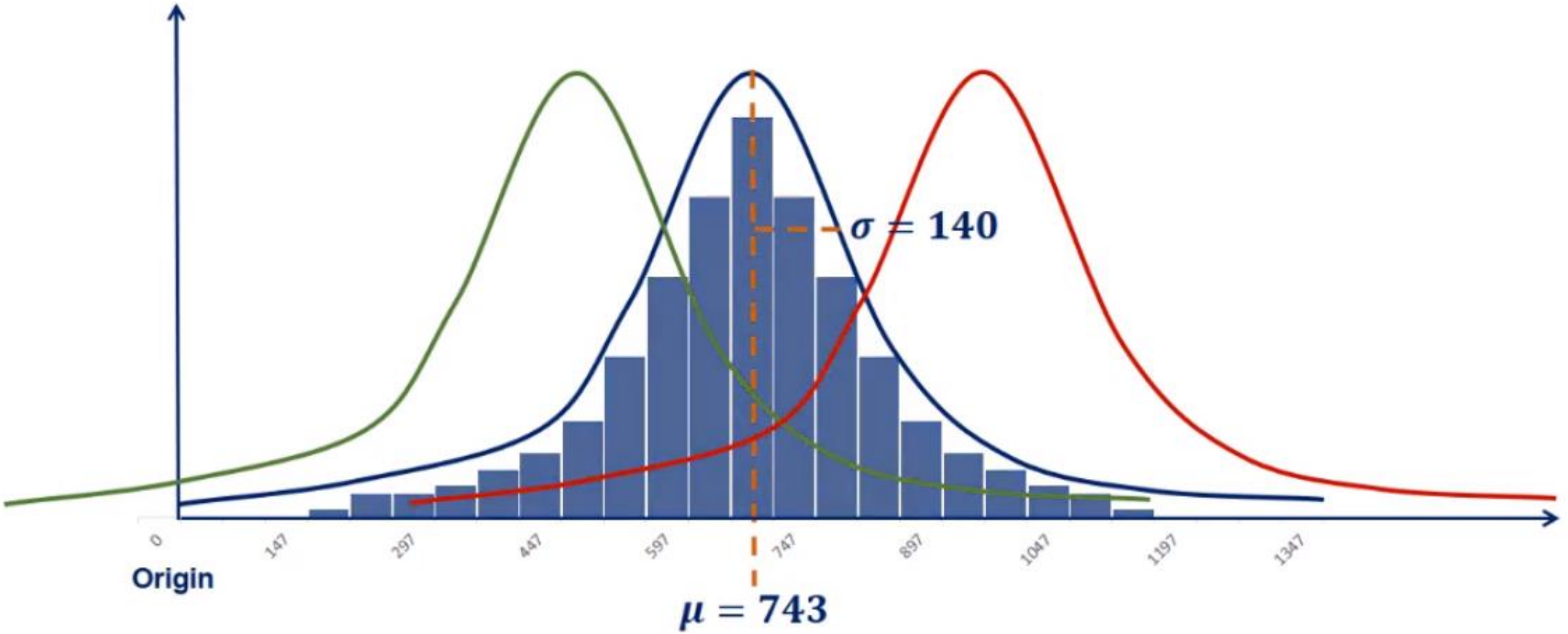


Scribbr

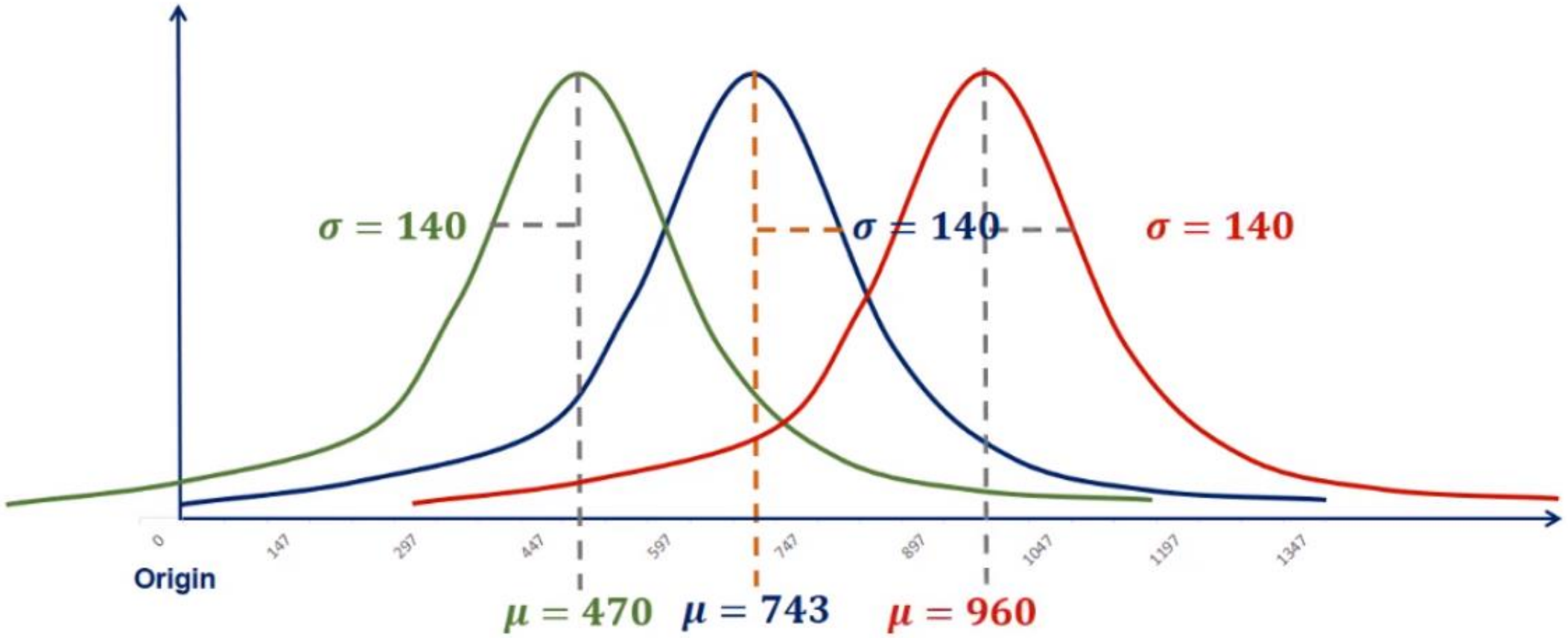
Normal Distribution



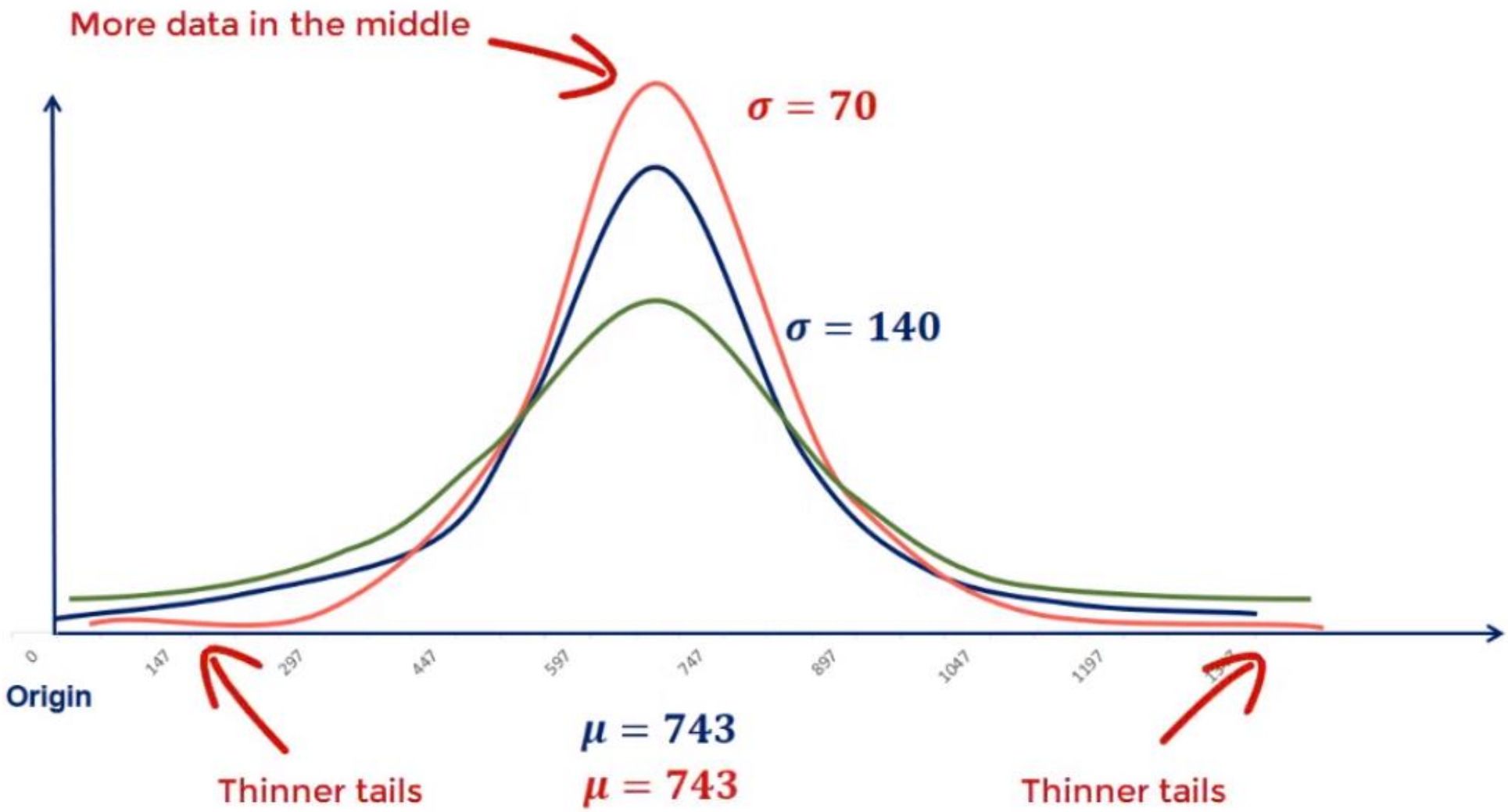
Normal Distribution



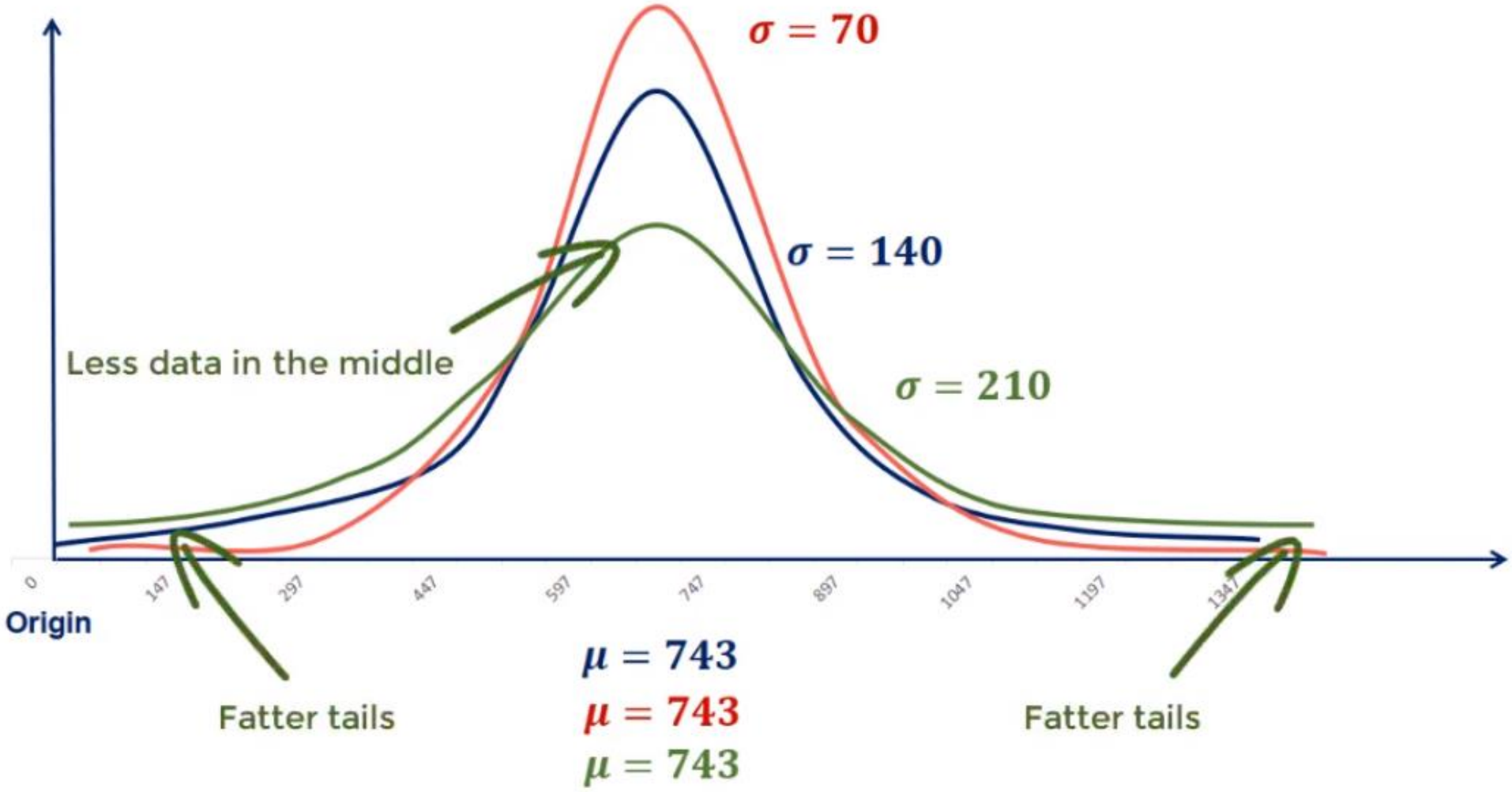
Normal Distribution



Normal Distribution

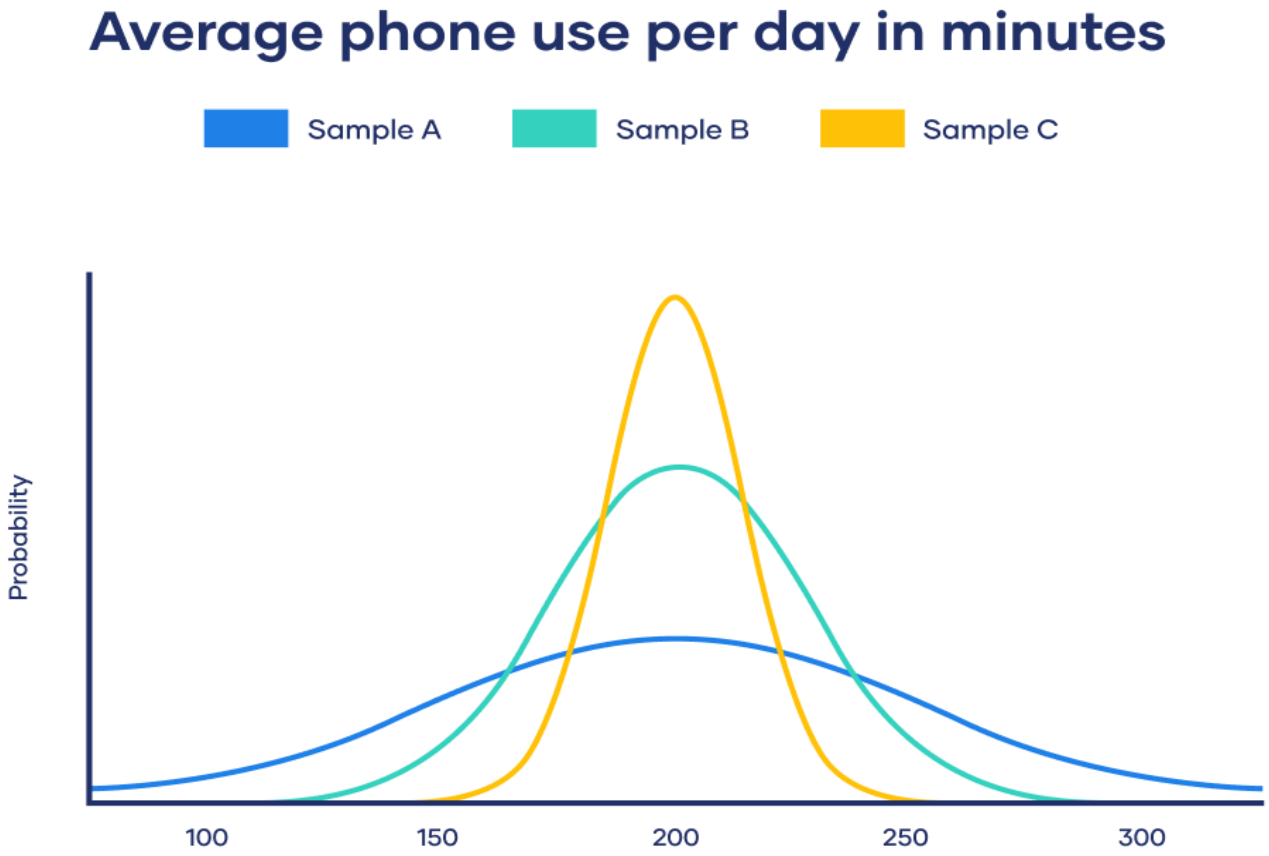


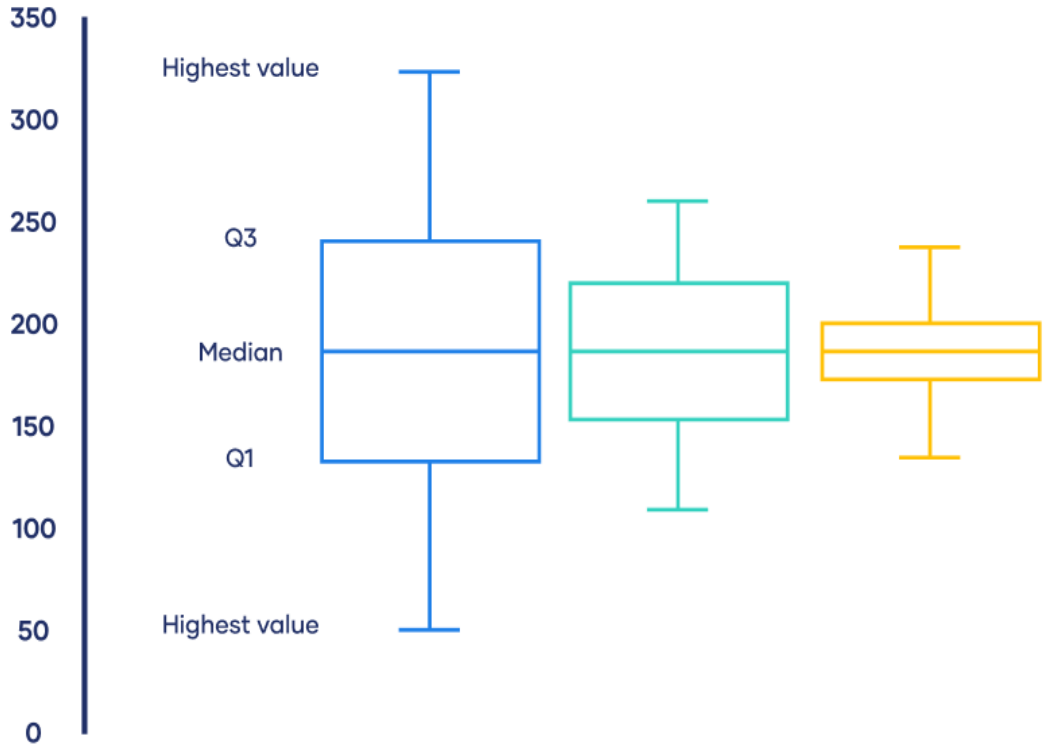
Normal Distribution



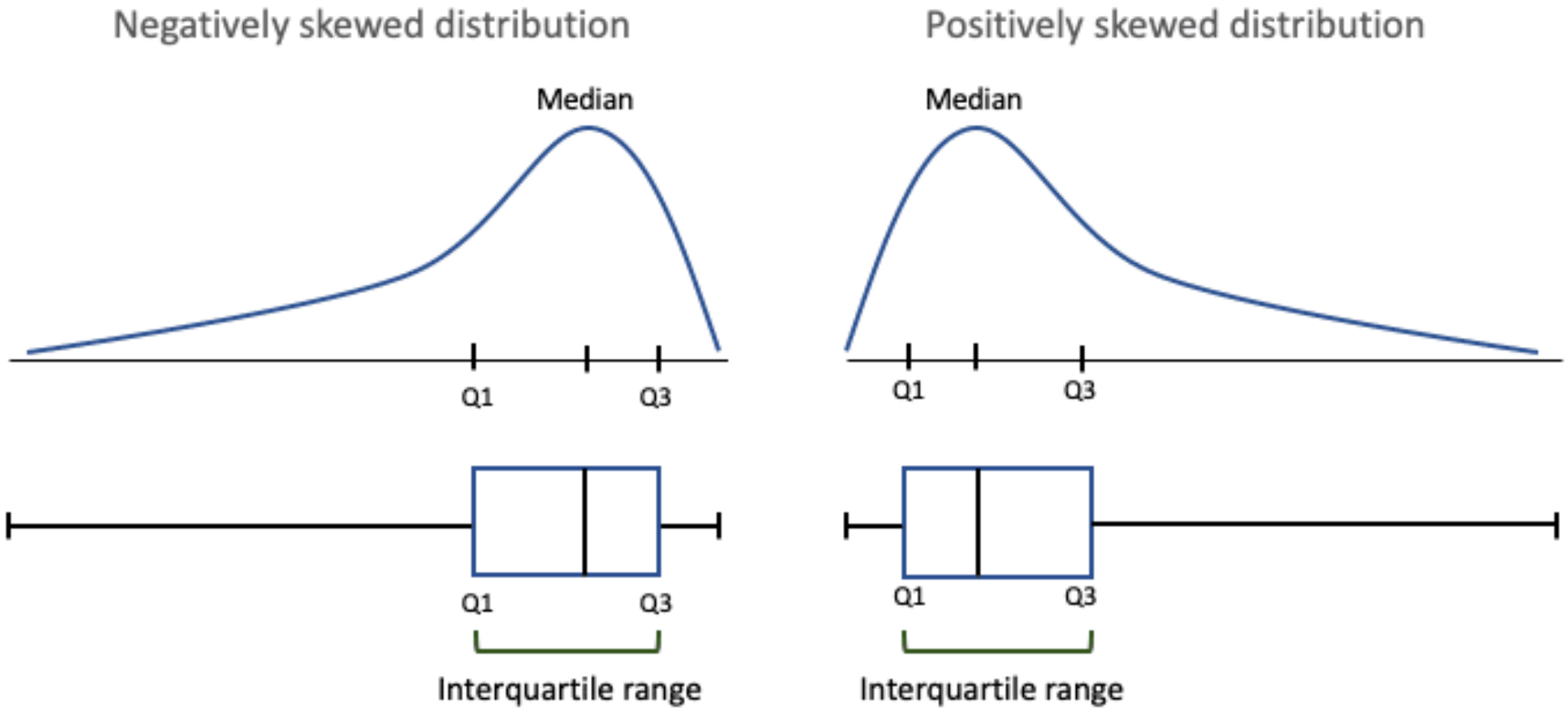
Time spent on phones Daily

- Sample A: high school students,
- Sample B: college students,
- Sample C: adult full-time employees.





Measuring Skewness

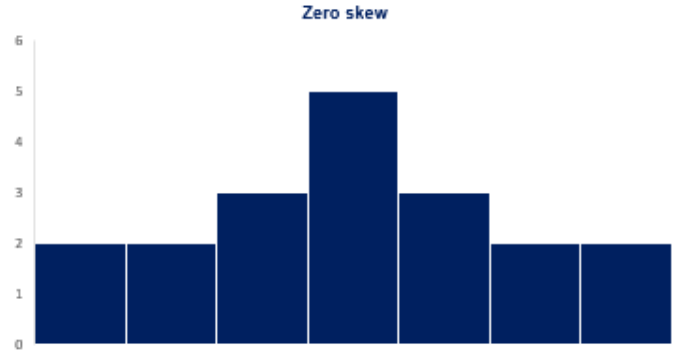


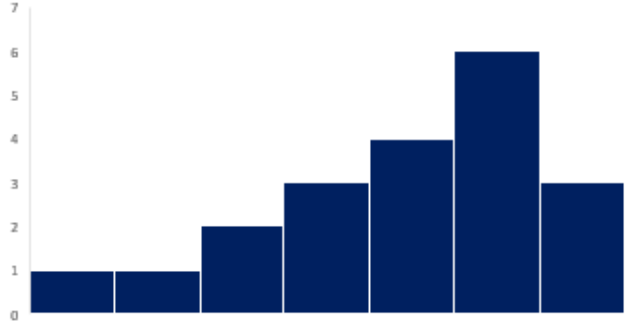
Measuring Skewness

Positive (right)				
Dataset 1		Interval	Frequency	
1		0 to 1	4	
1		1 to 2	6	
1		2 to 3	4	
1		3 to 4	2	
2		4 to 5	2	
2		5 to 6	0	
2		6 to 7	1	
2				
2				
2				
3		**Mean**	**Median**	**Mode**
3		**2.79**	**2.00**	**2.00**
3				
3				
4				
4				
5				
5				
7				

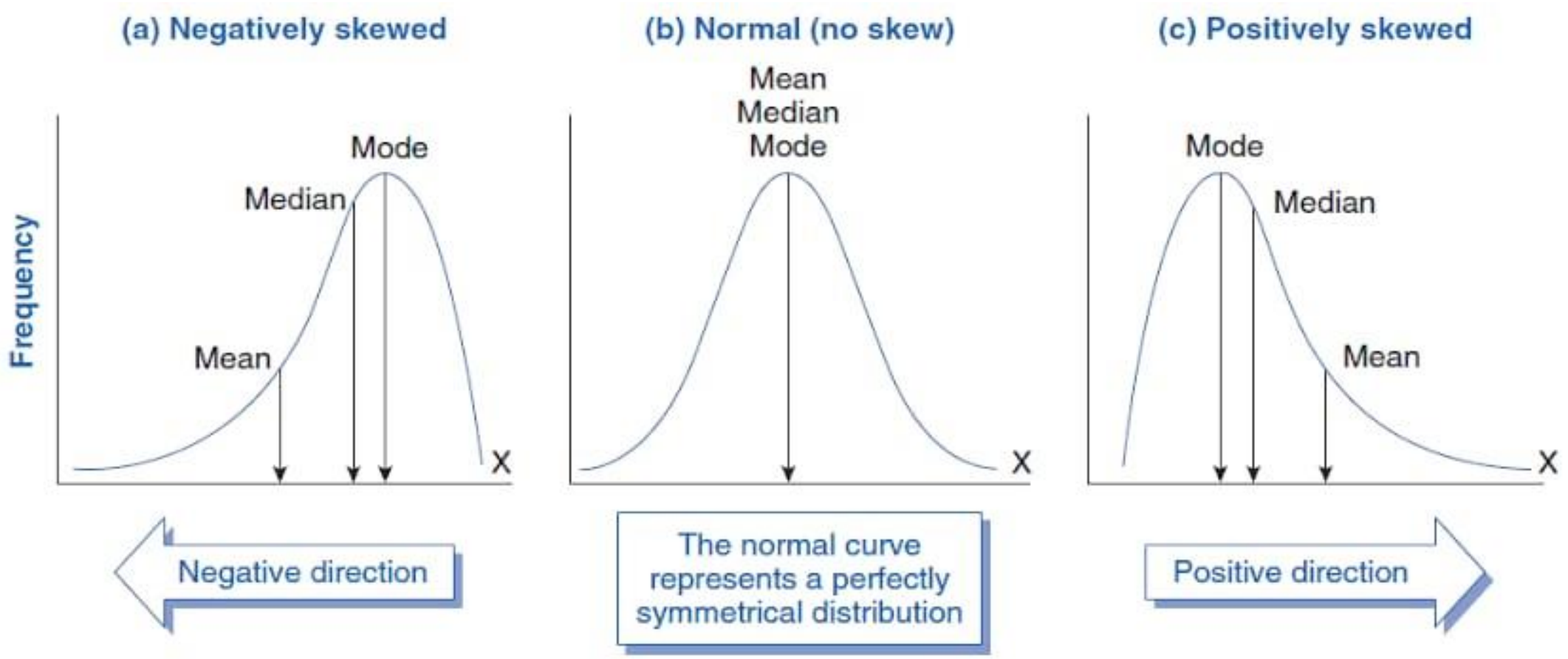
Zero (no skew)				
Dataset 2		Interval	Frequency	
1		0 to 1	2	
1		1 to 2	2	
2		2 to 3	3	
2		3 to 4	5	
3		4 to 5	3	
3		5 to 6	2	
3		6 to 7	2	
4				
4				
4				
4		**Mean**	**Median**	**Mode**
4		**4.00**	**4.00**	**4.00**
5				
5				
5				
6				
6				
7				
7				

Negative (left)				
Dataset 3		Interval	Frequency	
1		0 to 1	1	
2		1 to 2	1	
3		2 to 3	2	
3		3 to 4	3	
4		4 to 5	4	
4		5 to 6	6	
4		6 to 7	3	
5				
5				
5				
5		**Mean**	**Median**	**Mode**
6		**4.90**	**5.00**	**6.00**
6				
6				
6				
6				
6				
7				
7				
7				





Measuring Skewness

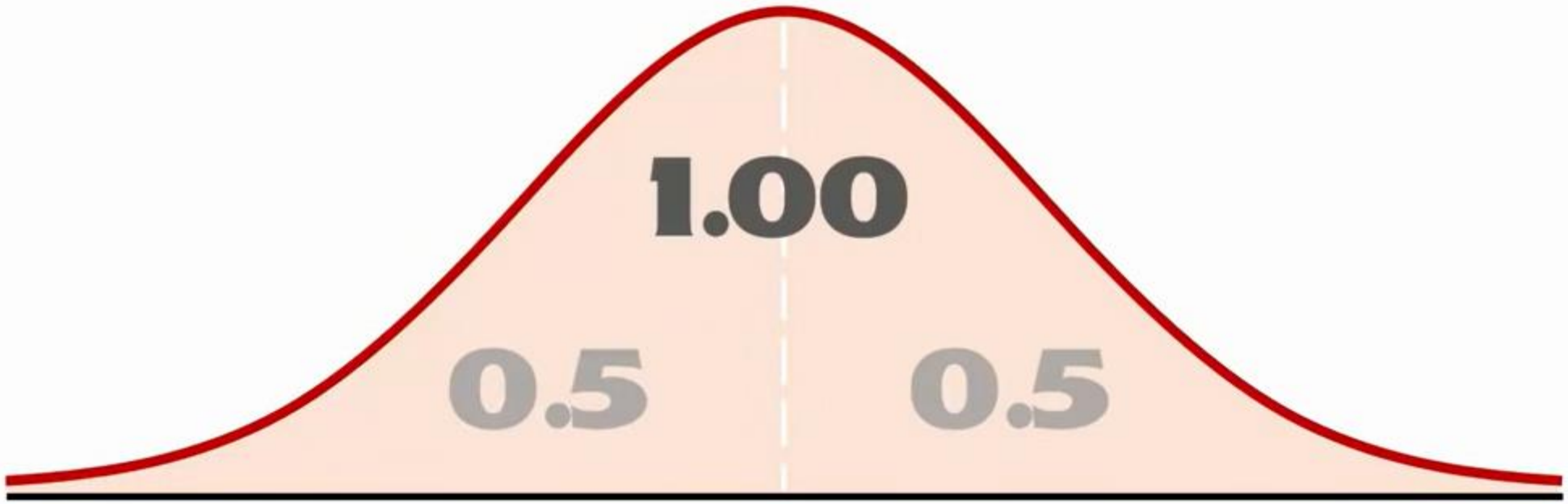


Mean<Median<Mode

Mean=Median=Mod

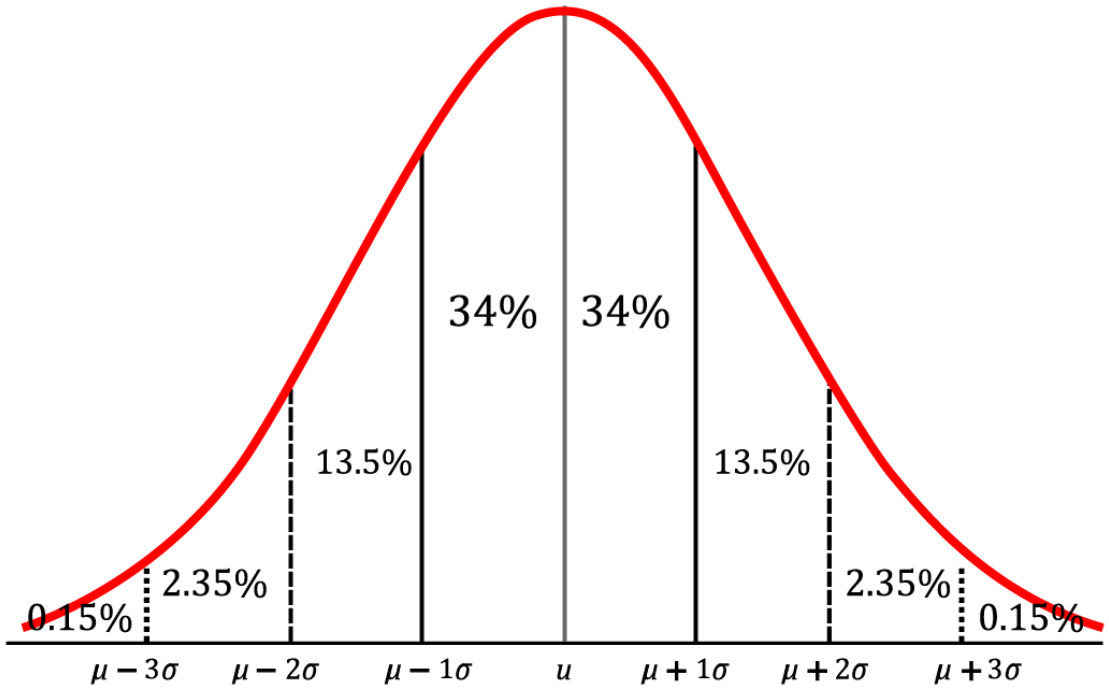
Mean>Median>Mode

Normal Distributions
Calculating Probabilities



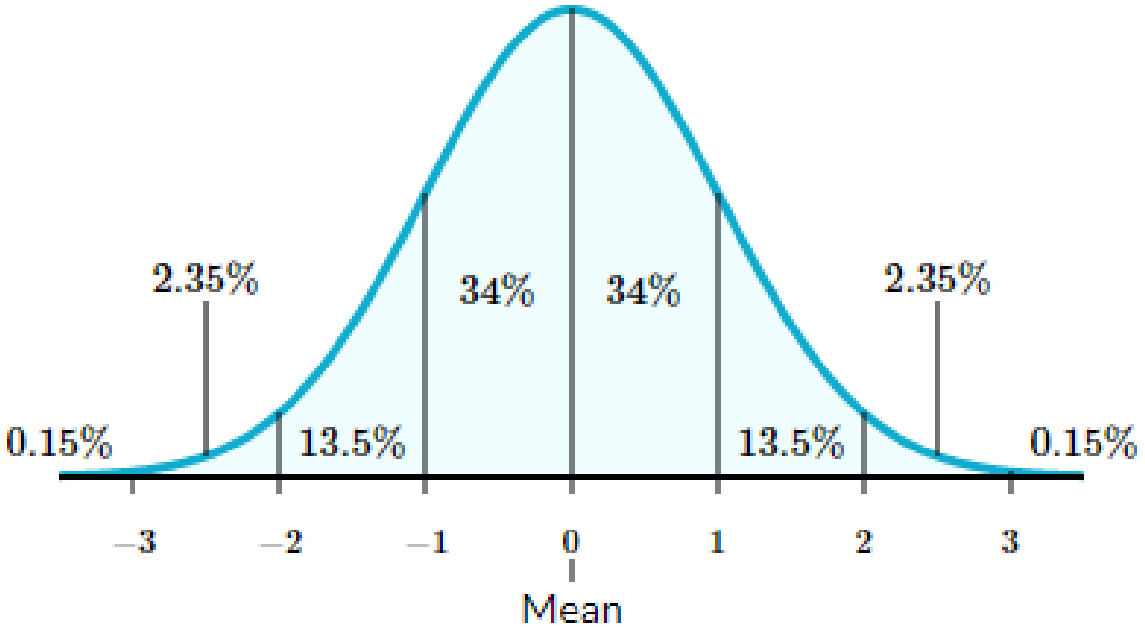
What is the Empirical Rule?

- The empirical rule, or the **68-95-99.7 rule**, tells you where most of the values lie in a normal distribution:
  - Around **68%** of values are within **1 standard deviation** of the mean.
  - Around **95%** of values are within **2 standard deviations** of the mean.
  - Around **99.7%** of values are within **3 standard deviations** of the mean.



What is the Empirical Rule?

- The lifespans of gorillas in a particular zoo are normally distributed. The average gorilla lives 20.8 years; the standard deviation is 3.1 years.
- **estimate the probability of a gorilla living less than 23.9 years.**

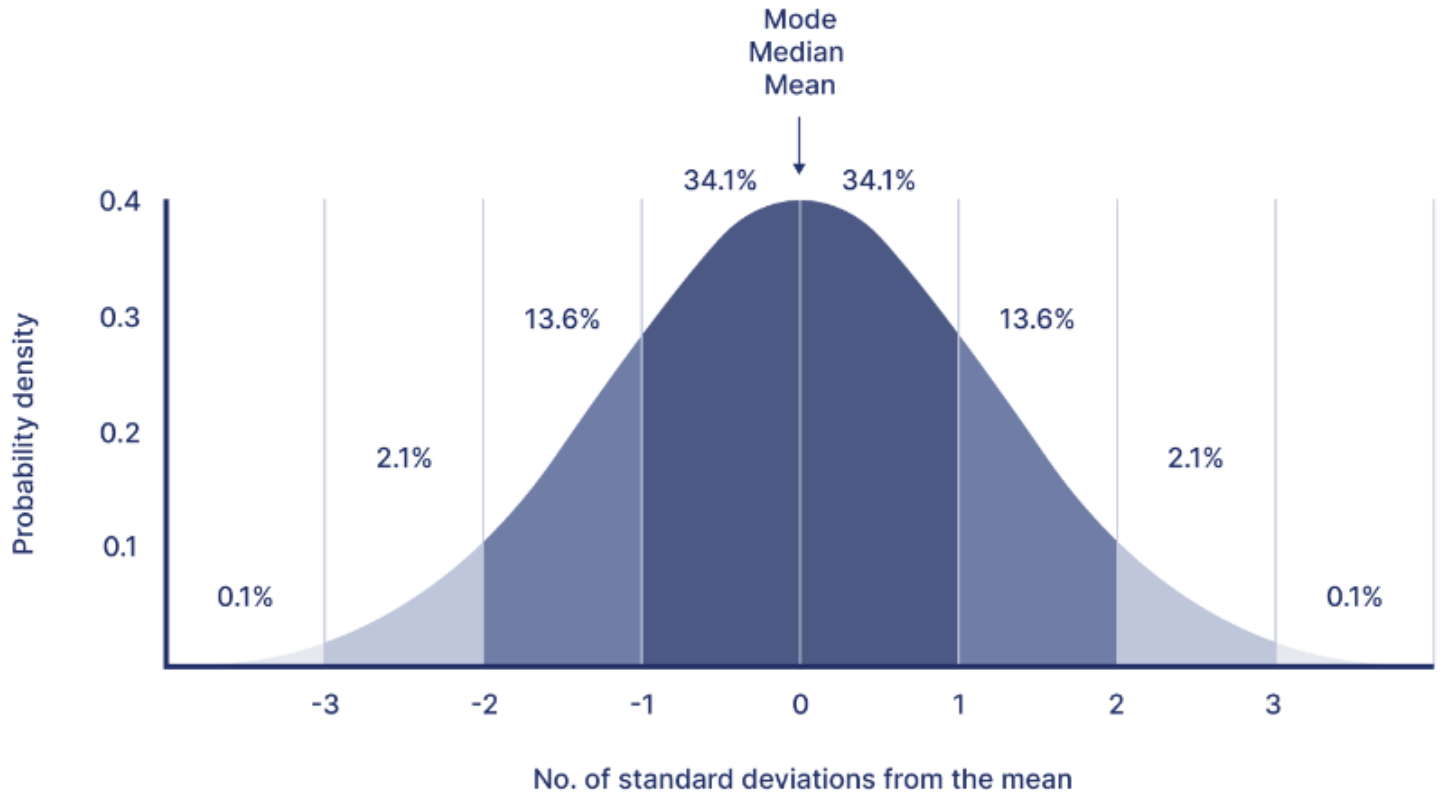


What is the Empirical Rule?

- A set of middle school students' heights are normally distributed with a mean of 150 centimesters and a standard deviation of 20 centimeters. Darnell is a middle school school student with a height of 161.4 centimeters. What proportion of student heights are lower than Darnell's height?

Standard Normal Distribution

- Also called the ***Z*-distribution**, is a special normal distribution where the **mean** is **0** and the **standard deviation** is **1**.



Important Points

- All the **standard Normal Distributions** are **Normal Distributions** **but all normal distributions** are **not Standard Normal Distributions**
- All Normal Distributions are symmetric in nature but all symmetric distributions are not normal.

Standard Normal Distribution

- We **convert** normal distributions into the standard normal distribution for several reasons:
  - To find the probability of observations in a distribution falling above or below a given value.
  - To find the probability that a sample mean significantly differs from a known population mean.
  - To compare scores on different distributions with different means and standard deviations.
  - To Normalize scores for statistical decision.

STANDARD NORMAL DISTRIBUTION

z-distribution

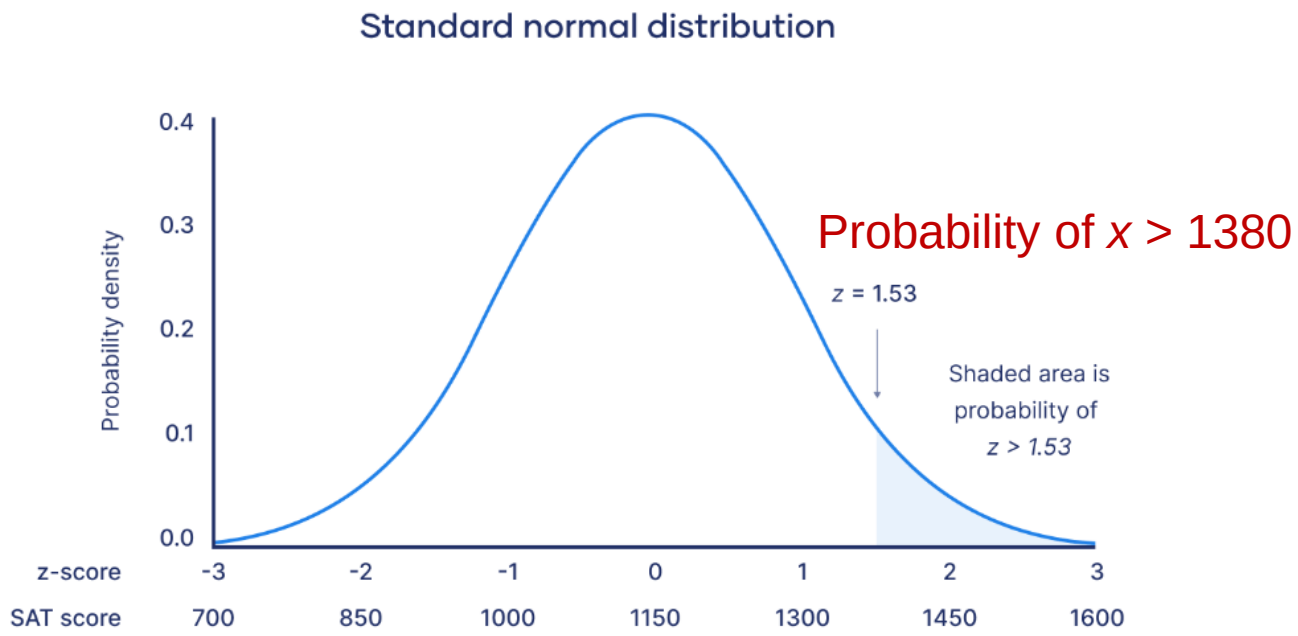
z-score = *number of standard deviations from the mean*

$$z = \frac{x - \mu}{\sigma}$$

positive *z* score means that your *x* value is **greater** than the mean.
negative *z* score means that your *x* value is **less** than the mean.
***z* score of zero** means that your *x* value is **equal** to the mean.

Example

- You collect SAT scores from students in a new test preparation course. The data follows a **normal distribution** with a mean score (*M*) of 1150 and a standard deviation (*SD*) of 150. You want to find the probability that SAT scores in your sample exceed 1380.

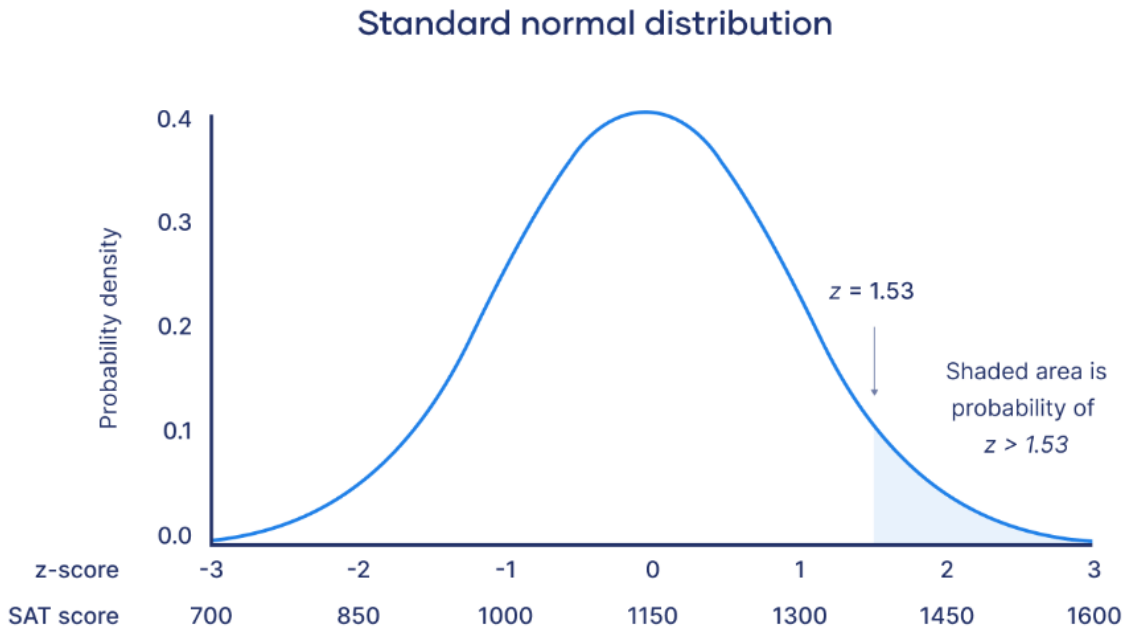


Formula	Calculation
$z = \frac{x - \mu}{\sigma}$	$z = \frac{1\,380 - 1\,150}{150}$ $z = 1.53$

Example

For a ***z*-score** of 1.53, the ***p*-value** is 0.937.

This is the probability of SAT scores being 1380 or less (93.7%), and it's the area under the curve left of the shaded area.

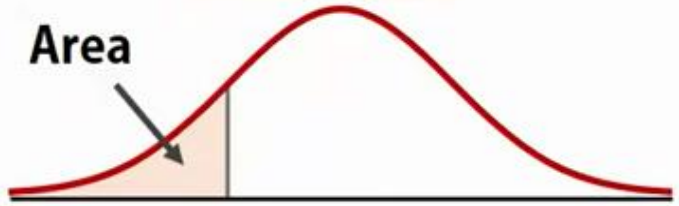


To find the shaded area, you take away 0.937 from 1, which is the total area under the curve.

Probability of $x > 1380 = 1 - 0.937 =$ **0.063**

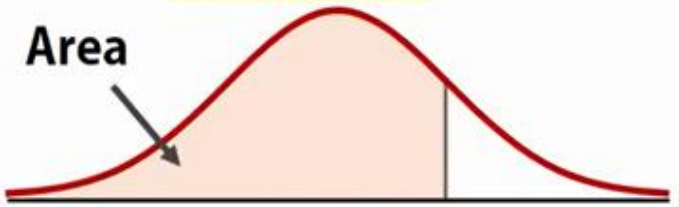
That means it is likely that only 6.3% of SAT scores in your sample exceed 1380.

Negative *z*



Less Than *Cumulative*

Positive *z*



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
-0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.

You take the **GATE examination** and score 500. The mean score for the GATE is 390 and the standard deviation is 45. How well did you score on the test compared to the average test taker?

Raw score/observed value = X = 500
Mean score = μ = 390
Standard deviation = σ = 45
By applying the formula of z-score,

Z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07
1.7	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994	0.9608	0.96164
1.8	0.96407	0.96485	0.96562	0.96638	0.96712	0.96784	0.96856	0.96926
1.9	0.97128	0.97193	0.97257	0.9732	0.97381	0.97441	0.975	0.97558
2	0.97725	0.97778	0.97831	0.97882	0.97932	0.97982	0.9803	0.98077
2.1	0.98214	0.98257	0.983	0.98341	0.98382	0.98422	0.98461	0.985
2.2	0.9861	0.98645	0.98679	0.98713	0.98745	0.98778	0.98809	0.9884
2.3	0.98928	0.98956	0.98983	0.9901	0.99036	0.99061	0.99086	0.99111
2.4	0.9918	0.99202	0.99224	0.99245	0.99266	0.99286	0.99305	0.99324
2.5	0.99379	0.99396	0.99413	0.9943	0.99446	0.99461	0.99477	0.99492
2.6	0.99534	0.99547	0.9956	0.99573	0.99585	0.99598	0.99609	0.99621
2.7	0.99653	0.99664	0.99674	0.99683	0.99693	0.99702	0.99711	0.9972

What is the probability that a student scores between 350 and 400 (with a mean score μ of 390 and a standard deviation σ of 45)?

Min score = X1 = 350
Max score = X2 = 400

- By applying the formula of z-score,

z1 = (X1 – μ) / σ
z1 = (350 – 390) / 45
z1 = -40 / 45 = -0.88

z2 = (X2 – μ) / σ
z2 = (400 – 390) / 45
z2 = 10 / 45 = 0.22

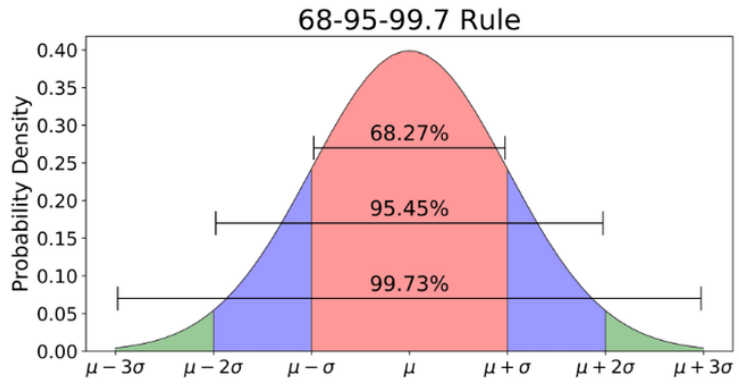
- Since z1 is negative, we will have to look at a negative Z-Table and find that p1, the first probability, is **0.18943**.
- z2 is positive, so we use a positive Z-Table which yields a probability p2 of **0.58706**.
- The final probability is computed by subtracting p1 from p2:

p = p2 – p1
p = 0.58706 – 0.18943 = 0.39763

- The probability that a student scores between 350 and 400 is **39.763%**

Interpretation of Z-score

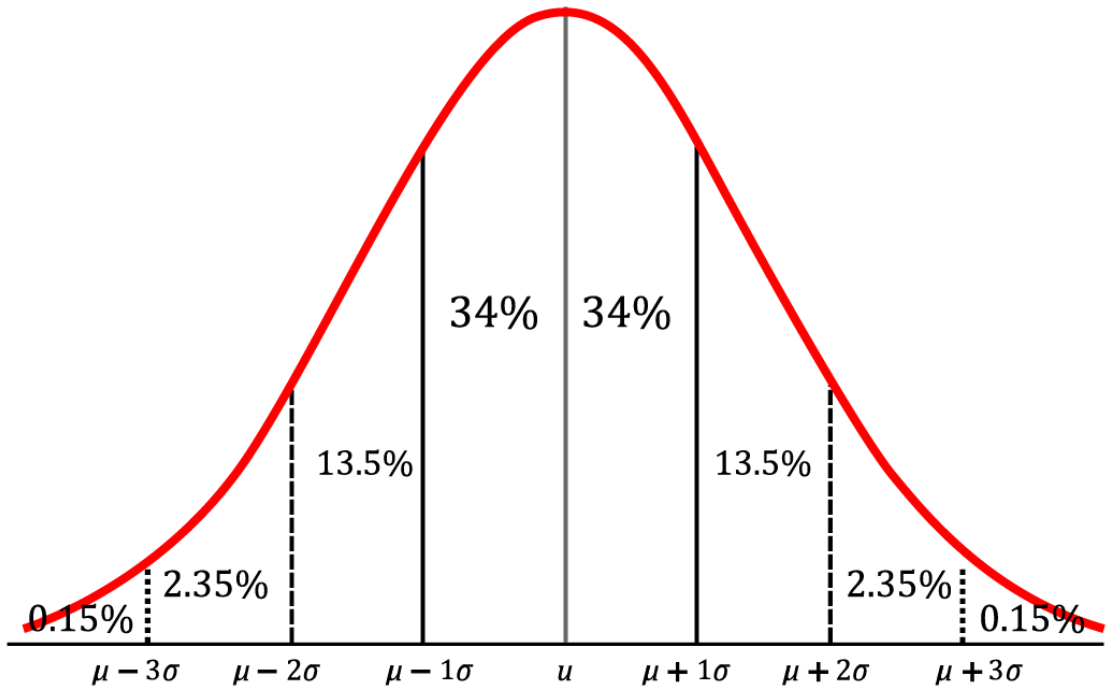
- An element having a z-score of less than 0 represents that the element is less than the mean.
- An element having a z-score greater than 0 represents that the element is greater than the mean.
- An element having a z-score equal to 0 represents that the element is equal to the mean.
- An element having a z-score equal to 1 represents that the element is 1 standard deviation greater than the mean; a z-score equal to 2, 2 standard deviations greater than the mean, and so on.
- An element having a z-score equal to -1 represents that the element is 1 standard deviation less than the mean; a z-score equal to -2, 2 standard deviations less than the mean, and so on.
- If the number of elements in a given set is large, then about 68% of the elements have a z-score between -1 and 1; about 95% have a z-score between -2 and 2; about 99% have a z-score between -3 and 3. This is known as the Empirical Rule or the 68-95-99.7.

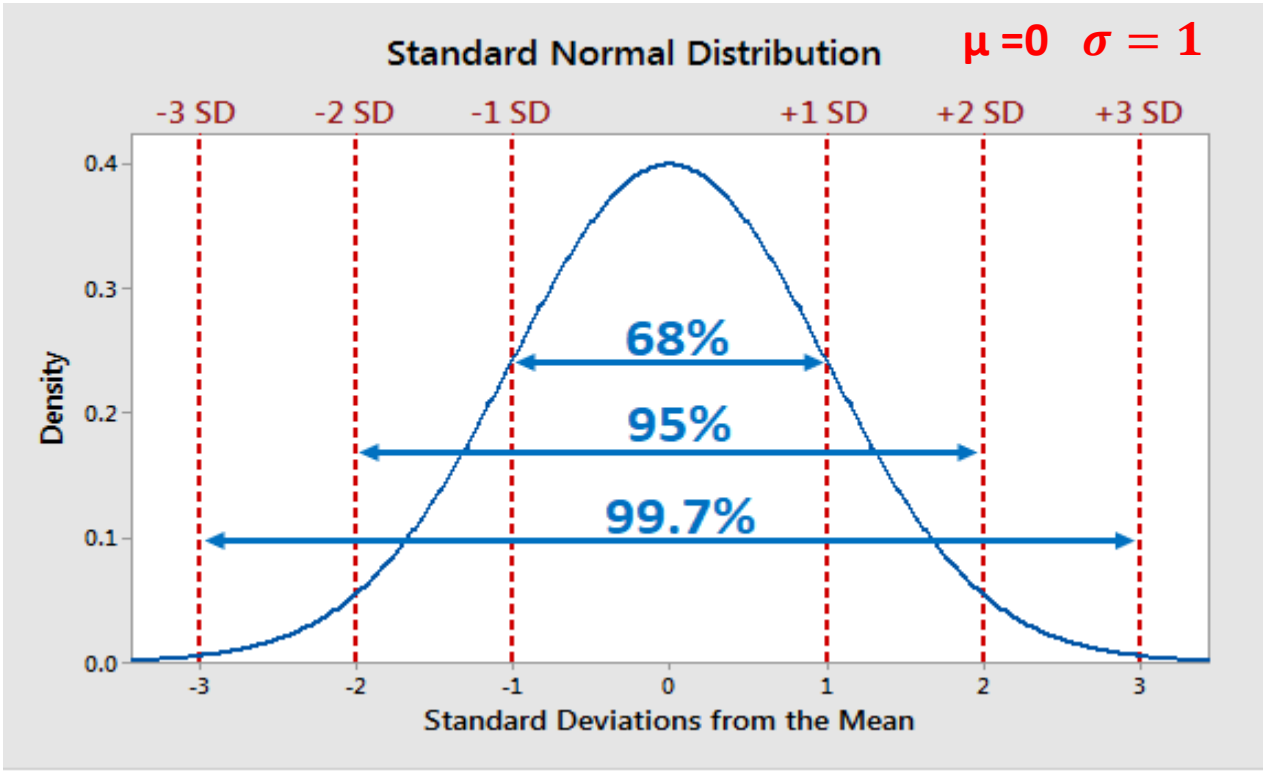


The 68-95-99.7 Rule for a Normal Distribution

Empirical Rule

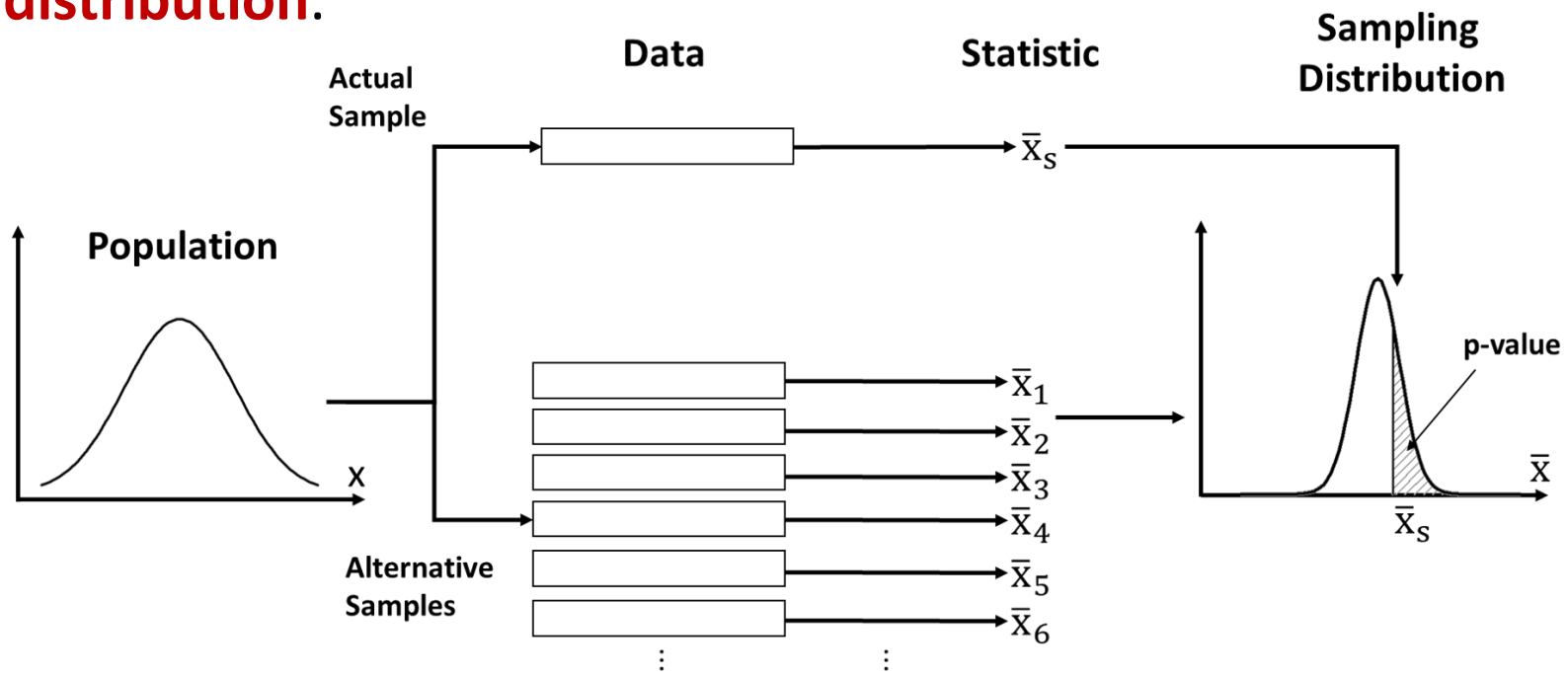
- The empirical rule, or the **68-95-99.7 rule**, tells you where most of the values lie in a normal distribution:
  - Around **68%** of values are within **1 standard deviation** of the mean.
  - Around **95%** of values are within **2 standard deviations** of the mean.
  - Around **99.7%** of values are within **3 standard deviations** of the mean.





Sampling Distribution

- Suppose that we draw all possible samples of size *n* from a given population.
- Suppose further that we compute a statistic (e.g., a mean, proportion, standard deviation) for each sample. The probability distribution of this statistic is called a **sampling distribution**.



Sampling Distribution

- A Sampling Distribution behaves much like a normal curve.
- The shape of the Sampling Distribution does not reveal anything about the shape of the population.

Sampling Distribution

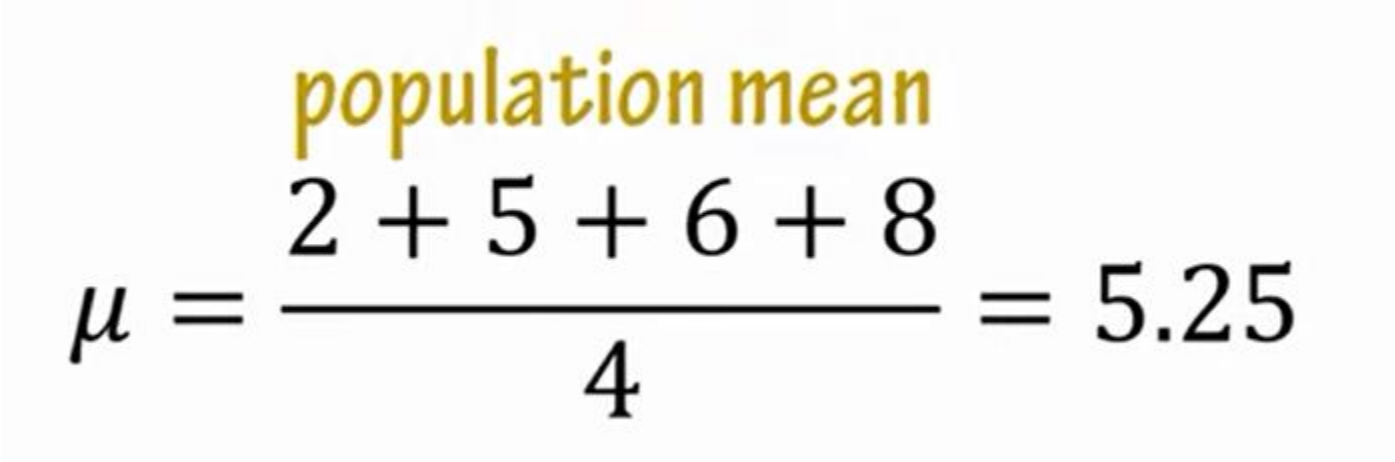
- Sampling Distribution helps to estimate the population statistic.

But how ?

- This will be explained using a very important theorem in statistics – **The Central Limit Theorem.**

Sampling Distribution

2, 5, 6, 8



Sampling Distribution

2, 5, 6, 8

$$\sigma = \sqrt{\frac{(2-5.25)^2 + (5-5.25)^2 + (6-5.25)^2 + (8-5.25)^2}{4}}$$

$$= 2.165$$ population standard deviation

Sampling Distribution

$\mu = 5.25$ **2, 5, 6, 8** $\sigma = 2.165$

all samples of n = 2

2	2	2	2	5	5	5	5	6	6	6	6	8	8	8	8
2	5	6	8	2	5	6	8	2	5	6	8	2	5	6	8

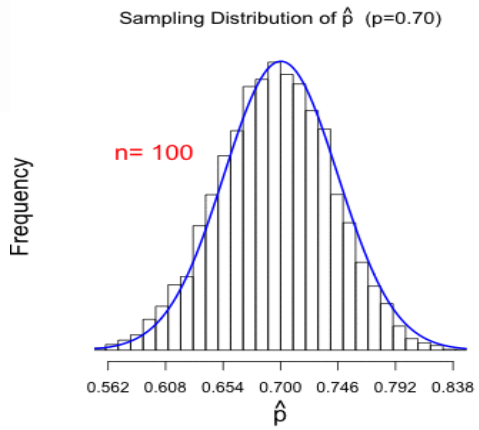
Sampling Distribution

$\mu = 5.25$ **2, 5, 6, 8** $\sigma = 2.165$

all samples of n = 2

2	2	2	2	5	5	5	5	6	6	6	6	8	8	8	8
2	5	6	8	2	5	6	8	2	5	6	8	2	5	6	8
$\bar{x}$	2	3.5	4	5	3.5	5	5.5	6.5	4	5.5	6	7	5	6.5	7	8

distribution of sample means



Sampling Distribution

$\mu = 5.25$ **2, 5, 6, 8** $\sigma = 2.165$

all samples of n = 2

	2	2	2	2	5	5	5	5	6	6	6	6	8	8	8	8
	2	5	6	8	2	5	6	8	2	5	6	8	2	5	6	8
$\bar{x}$	2	3.5	4	5	3.5	5	5.5	6.5	4	5.5	6	7	5	6.5	7	8

$$\mu_{\bar{x}} = 5.25 = \mu$$

$$\sigma_{\bar{x}} = 1.531 = \frac{2.165}{\sqrt{2}} = \frac{\sigma}{\sqrt{n}}$$

Sampling Distribution

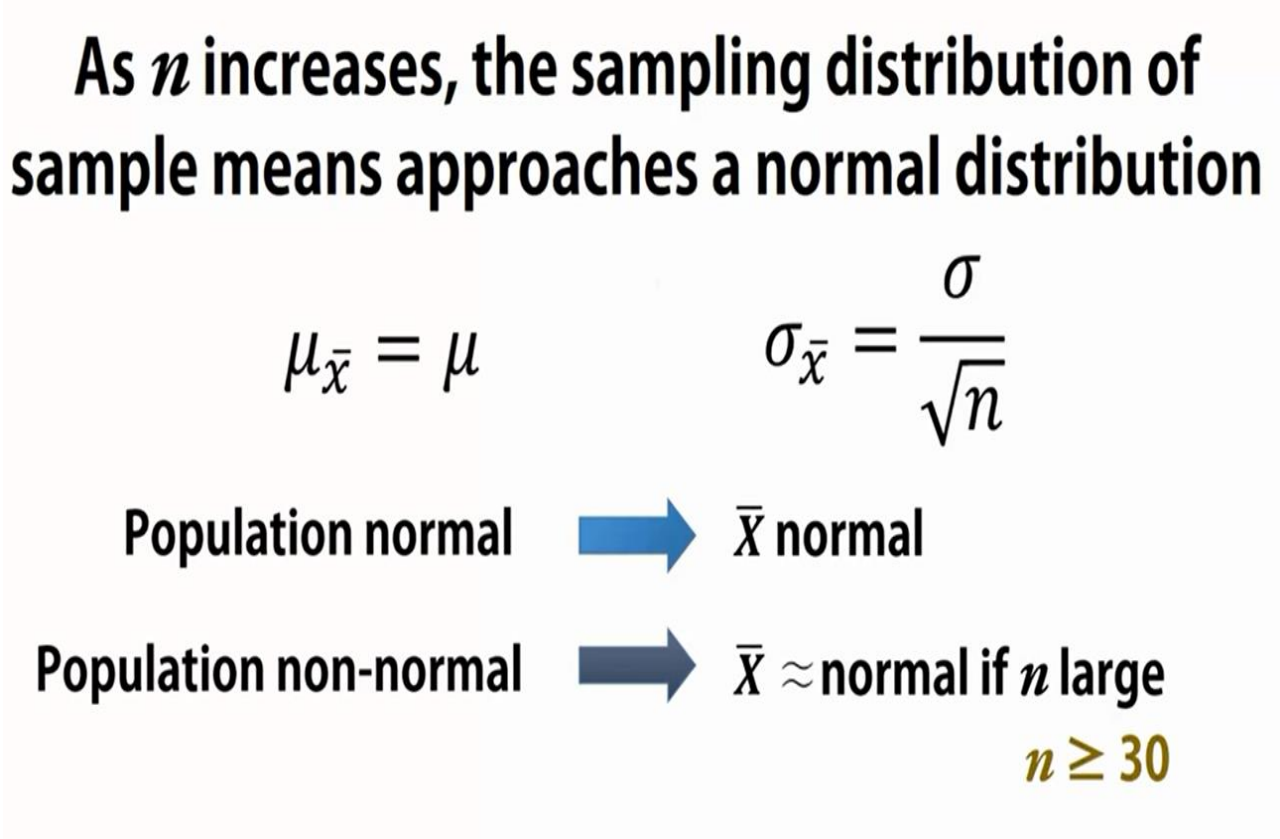
$\mu = 5.25$ **2, 5, 6, 8** $\sigma = 2.165$

	2 2	2 5	2 6	2 8	5 2	5 5	5 6	5 8	6 2	6 5	6 6	6 8	8 2	8 5	8 6	8 8
$\bar{x}$	2	3.5	4	5	3.5	5	5.5	6.5	4	5.5	6	7	5	6.5	7	8

$$\mu_{\bar{x}} = \mu \qquad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Central Limit Theorem

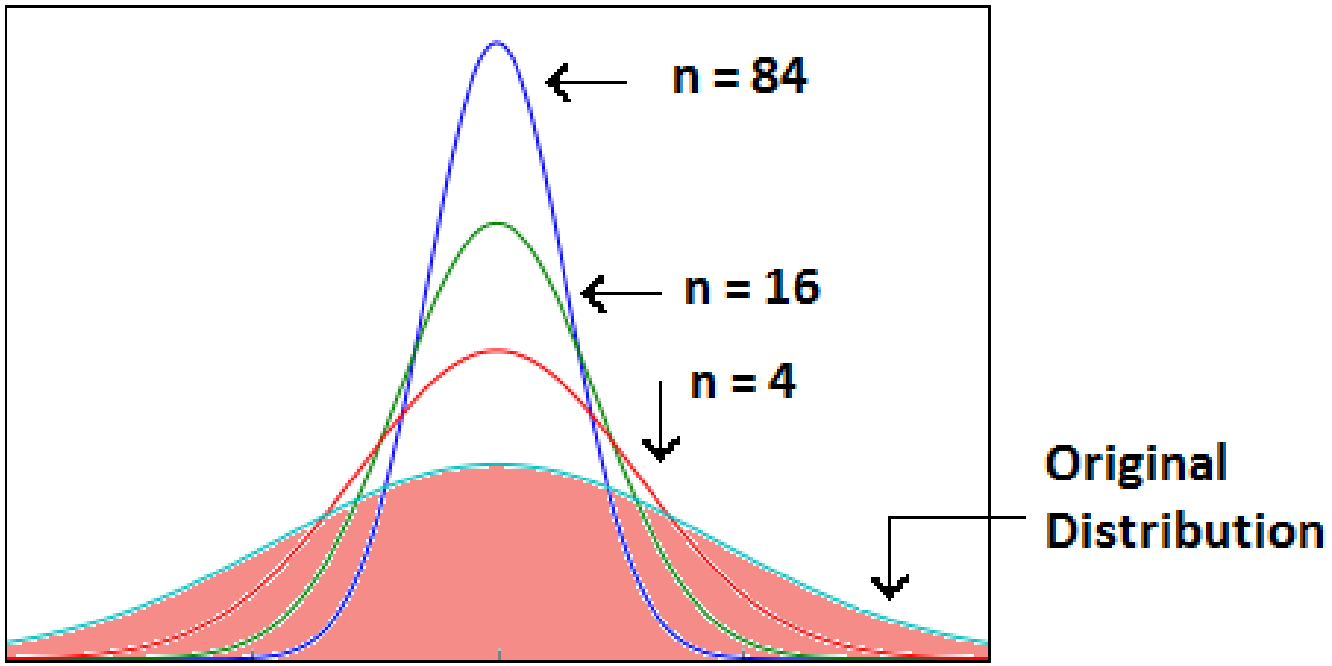
- It states that when plotting a sampling distribution of means, the **mean** of sample means will be equal to the population mean. And the sampling distribution will approach a normal distribution with **variance** equal to **σ/√n** where **σ** is the standard deviation of population and n is the sample size.



Central Limit Theorem

https://www.geogebra.org/m/kUmJeEwx

1. Central Limit Theorem holds true irrespective of the type of distribution of the population.
2. Now, we have a way to estimate the population mean by just making repeated observations of samples of a fixed size.
3. Greater the sample size, lower the standard error and greater the accuracy in determining the population mean from the sample mean.



P-value

P-values are **tail probabilities** calculated from the **sampling distribution** of a sample-based statistic.

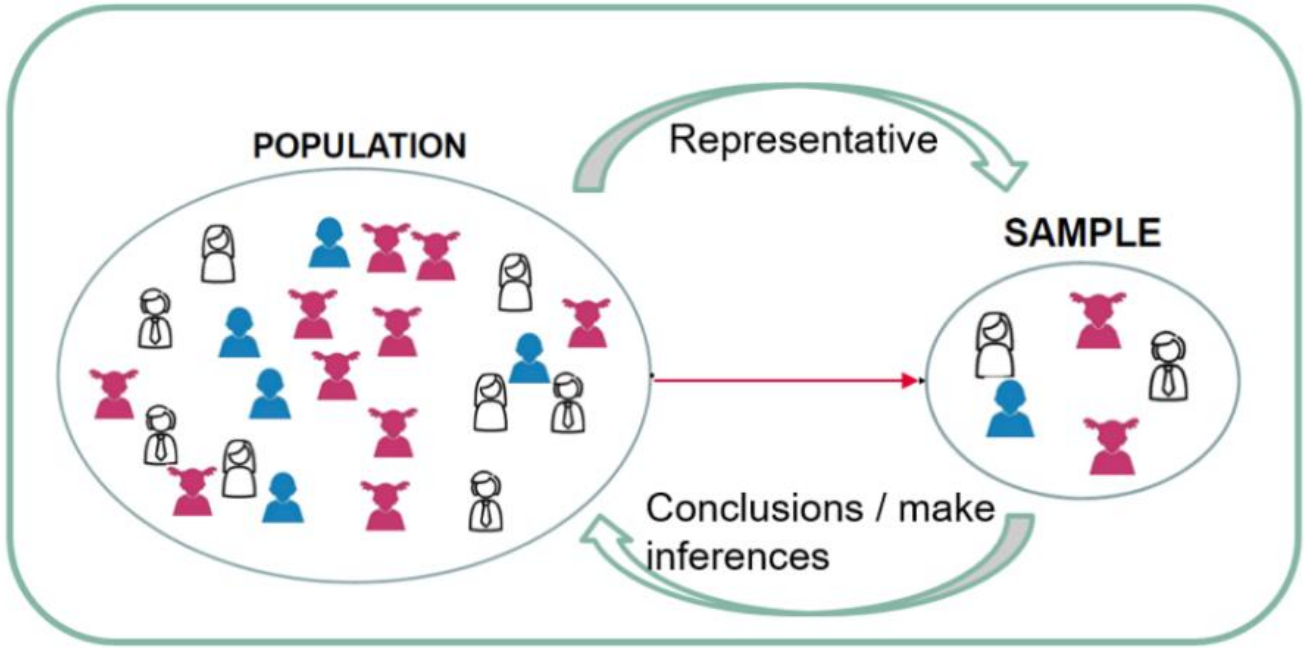
Inferential Statistics

50% of the **US population** cannot live without chocolate every day.

Nine out of **ten** people **love** chocolate.

Inferential Statistics

- **Inferential statistics** allow you to make ***inferences*** based on a data set.



Types of Inferences

Estimation

- We estimate the value of population parameter using a sample

Testing

- Do test to help us make a decision about a population parameter

Regression

- Make predictions or forcasting about statistics

Types of Estimates

- **Estimating parameter:** Predicting / Coming up with a summarized value using your sample data for the population
- There are two important types of estimates you can make about the population:
  - A **point estimate** is a single value estimate of a parameter. For instance, a sample mean is a point estimate of a population mean.
  - An **interval estimate** gives you a range of values where the parameter is expected to lie. A **confidence interval** is the most common type of interval estimate.

estimate the average age of students

sample mean age, $\bar{x} = 23$

Point Estimate of μ

Point Estimate

A single value, statistic, computed from sample data, and used to estimate a population parameter

range of values

Confidence Interval

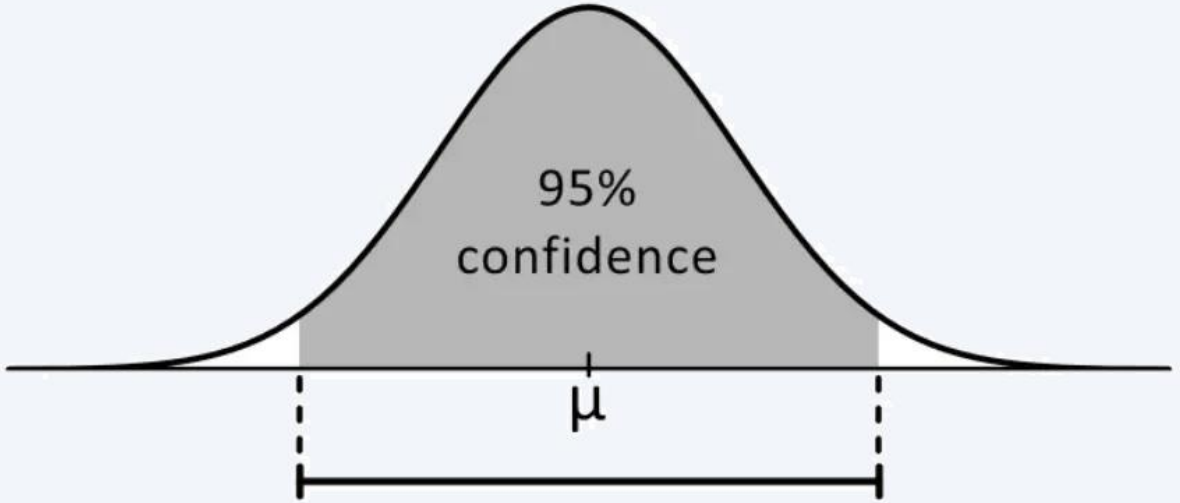
The confidence interval is a type of **interval estimate** from the **sampling distribution** which gives a **range of values** in which the **population statistic** may **lie**.

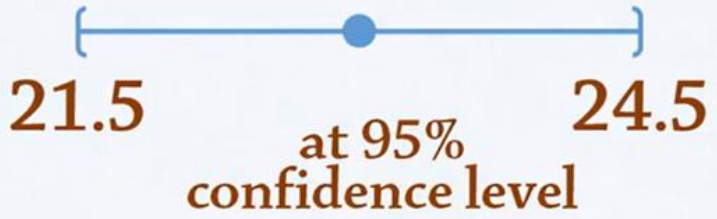
Confidence Interval

23 ± 1.5 → Margin of Error E

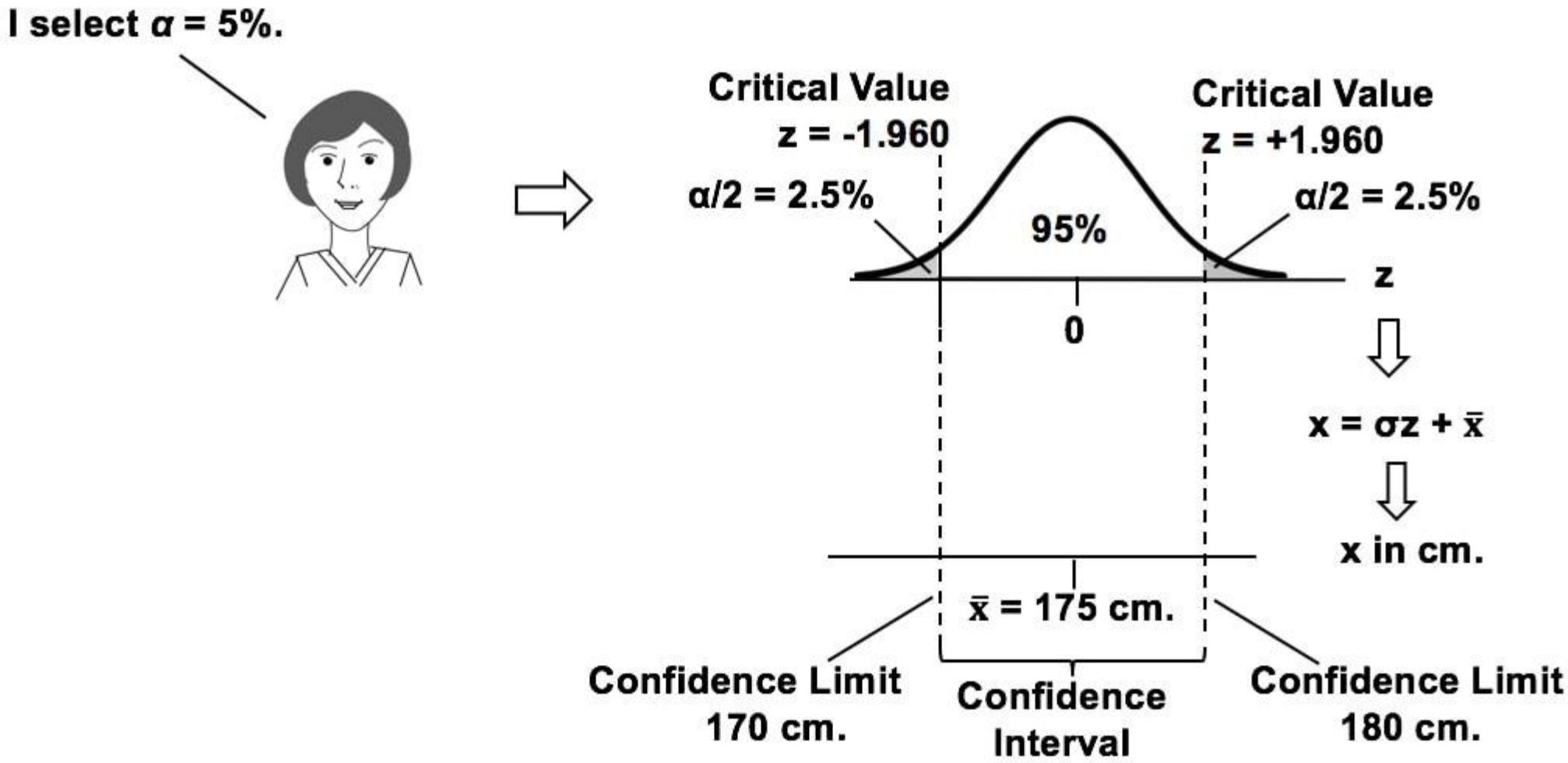
$$\bar{x} - E < \mu < \bar{x} + E$$

Confidence Level





we are 95% confident that the population mean lies between 21.5 and 24.5



Confidence Interval = Point Estimate ± Margin of Error

Confidence Interval = Sample Mean ± Margin of Error

where Margin of Error = Critical Value x Standard Error

Confidence Interval = Sample Mean ± Critical Value x Standard Error

where Standard Error means Standard Deviation of Sampling Distribution of Sample means.

We can calculate Standard Error as:

Standard Error = Standard Deviation of Population / $\sqrt{sample\ size}$

Confidence Interval = Sample Mean ± $\frac{\text{Critical Value x Pop. Std. Dev}}{\sqrt{sample\ size}}$

We are given with the Standard Deviation of population weight to be 4 Kgs, and a sample of 100 people is chosen and their average weight is 62 kgs. What is the Standard Error, Margin of Error and Confidence interval for Confidence Level 95%?

$$\text{Confidence Interval} = \text{Sample Mean} \pm \text{Critical Value} \times \frac{\text{Pop. Std. Dev}}{\sqrt{\text{sample size}}}$$

Standard Error = 4 / Square_root(100) = 4 / 10 = 0.4

For **95%** confidence level, critical value is **1.96**

Margin of Error = 1.96 x Standard Error

Margin of Error = 0.784

Confidence Interval = Sample mean +- Margin of Error

Lower bound of Confidence Interval = 62 - 0.784 = 61.216

Upper bound of Confidence Interval = 62 + 0.784 = 62.784

Let us write what is given:

Standard Deviation of population = 4

Sample Size = 100

Sample Mean = 62

Confidence Level = 95

Hence the confidence interval will be [61.216, 62.784] with 95% of confidence level.

This means that there is **95% probability** that the **estimated confidence interval** [61.216, 62.784] will contain the **true population mean**.

Confidence Level

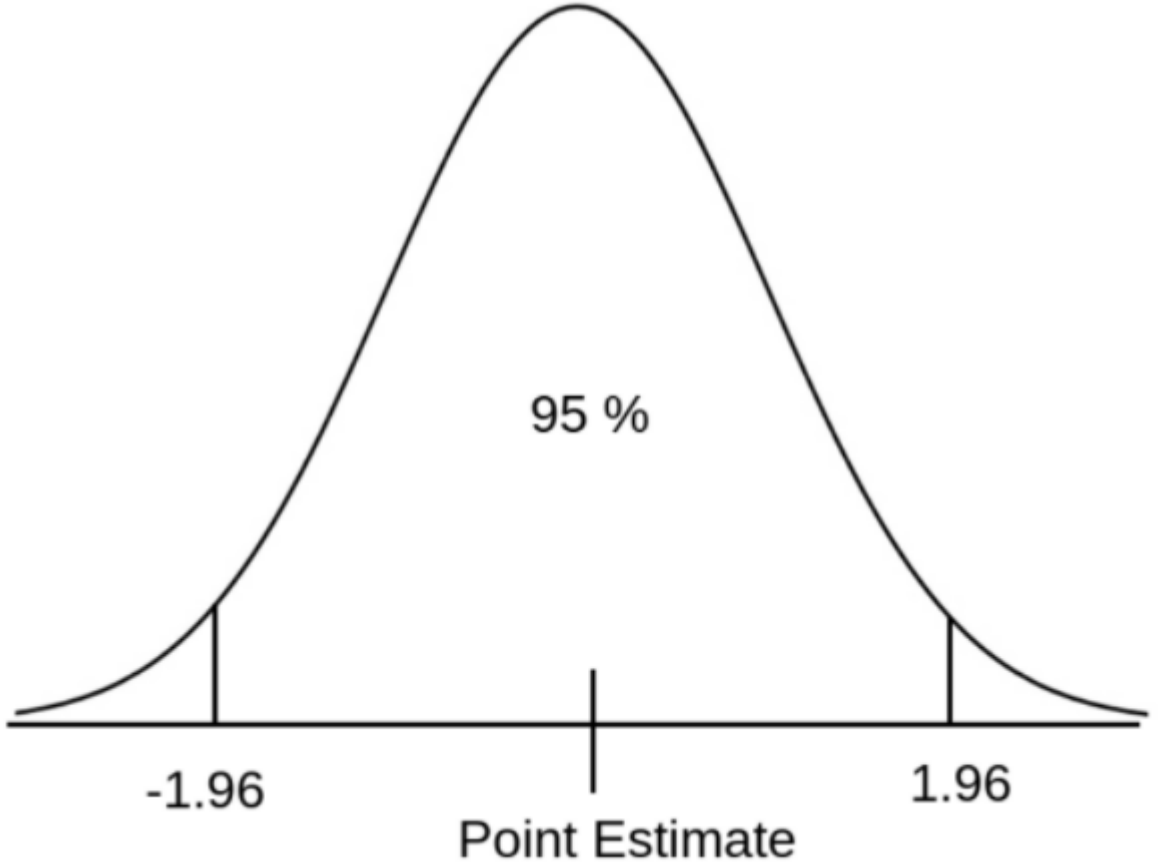
% of confidence intervals that we expect to contain the population parameter

90%	95%	99%
1.64	1.96	2.57
	most common	

With a 95% confidence level, 95% of all sample means will be expected to lie within a confidence interval of ± 1.96 standard errors of the sample mean.

$$\bar{X} - (1.96 \times SE) < \mu < \bar{X} + (1.96 \times SE)$$

$$SE = \frac{s}{\sqrt{n}}$$



- The 95% confidence level is the middle portion in the above figure and the leftover section on the left and right side is 5% combined which is 2.5% on each side.
- Hence when we are referring to the **z-distribution table**, we need to look at the value 95 + 2.5 = 97.5 to get the critical value of 95% **confidence level**.

Look at the row and column value for the cell **0.975**.
Row value is 1.9
Column value is 0.06
Hence Critical value will be 1.9 + 0.06 = 1.96 to get 95% confidence level

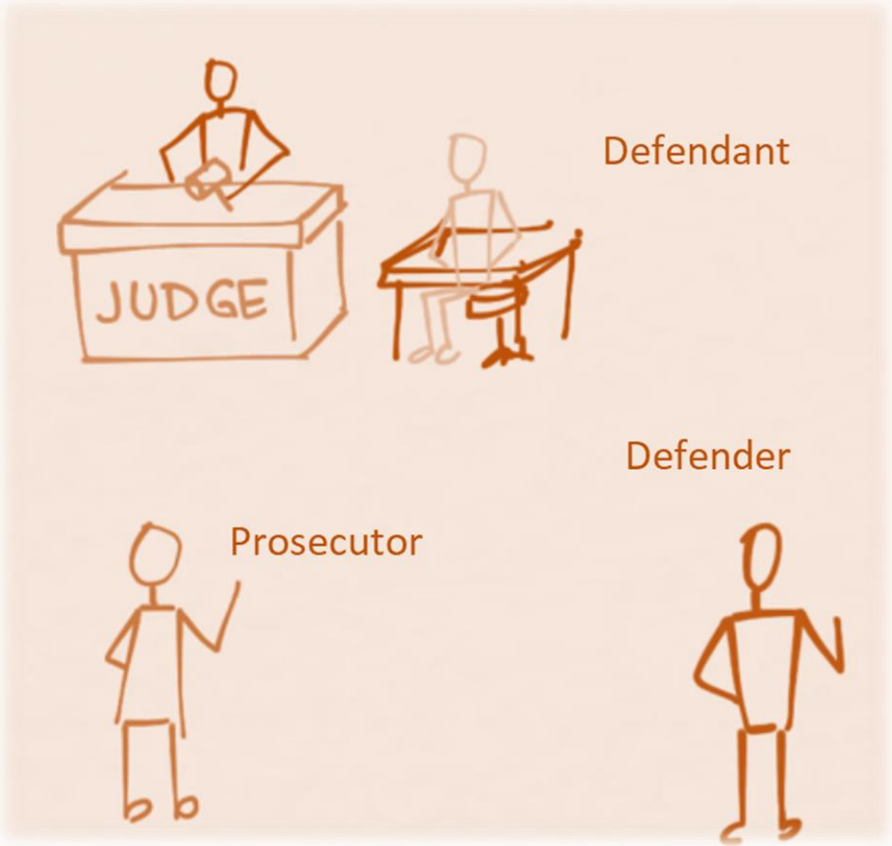
The probability for a z- score **below** −1.96 is 2.5%, and similarly for a z- score **above** +1.96; added together this is 5%.

1.96 is used because the 95% confidence interval has only 2.5% on **each** side.

z	+ 0.00	+ 0.01	+ 0.02	+ 0.03	+ 0.04	+ 0.05	+ 0.06	+ 0.07	+ 0.08	+ 0.09
0.0	0.50000	0.50399	0.50798	0.51197	0.51595	0.51994	0.52392	0.52790	0.53188	0.53586
0.1	0.53983	0.54380	0.54776	0.55172	0.55567	0.55962	0.56360	0.56749	0.57142	0.57535
0.2	0.57926	0.58317	0.58706	0.59095	0.59483	0.59871	0.60257	0.60642	0.61026	0.61409
0.3	0.61791	0.62172	0.62552	0.62930	0.63307	0.63683	0.64058	0.64431	0.64803	0.65173
0.4	0.65542	0.65910	0.66276	0.66640	0.67003	0.67364	0.67724	0.68082	0.68439	0.68793
0.5	0.69146	0.69497	0.69847	0.70194	0.70540	0.70884	0.71226	0.71566	0.71904	0.72240
0.6	0.72575	0.72907	0.73237	0.73565	0.73891	0.74215	0.74537	0.74857	0.75175	0.75490
0.7	0.75804	0.76115	0.76424	0.76730	0.77035	0.77337	0.77637	0.77935	0.78230	0.78524
0.8	0.78814	0.79103	0.79389	0.79673	0.79955	0.80234	0.80511	0.80785	0.81057	0.81327
0.9	0.81594	0.81859	0.82121	0.82381	0.82639	0.82894	0.83147	0.83398	0.83646	0.83891
1.0	0.84134	0.84375	0.84614	0.84849	0.85083	0.85314	0.85543	0.85769	0.85993	0.86214
1.1	0.86433	0.86650	0.86864	0.87076	0.87286	0.87493	0.87698	0.87900	0.88100	0.88298
1.2	0.88493	0.88686	0.88877	0.89065	0.89251	0.89435	0.89617	0.89796	0.89973	0.90147
1.3	0.90320	0.90490	0.90658	0.90824	0.90988	0.91149	0.91308	0.91466	0.91621	0.91774
1.4	0.91924	0.92073	0.92220	0.92364	0.92507	0.92647	0.92785	0.92922	0.93056	0.93189
1.5	0.93319	0.93448	0.93574	0.93699	0.93822	0.93943	0.94062	0.94179	0.94295	0.94408
1.6	0.94520	0.94630	0.94738	0.94845	0.94950	0.95053	0.95154	0.95254	0.95352	0.95449
1.7	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994	0.96080	0.96164	0.96246	0.96327
1.8	0.96407	0.96485	0.96562	0.96638	0.96712	0.96784	0.96856	0.96926	0.96995	0.97062
1.9	0.97128	0.97193	0.97257	0.97320	0.97381	0.97441	0.97500	0.97558	0.97615	0.97670
2.0	0.97725	0.97778	0.97831	0.97882	0.97932	0.97982	0.98030	0.98077	0.98124	0.98169
2.1	0.98214	0.98257	0.98300	0.98341	0.98382	0.98422	0.98461	0.98500	0.98537	0.98574
2.2	0.98610	0.98645	0.98679	0.98713	0.98745	0.98778	0.98809	0.98840	0.98870	0.98899

Hypothesis Testing

Hypothesis testing is a form of statistical inference that uses data from a **sample** to draw conclusions about a **population parameter**.

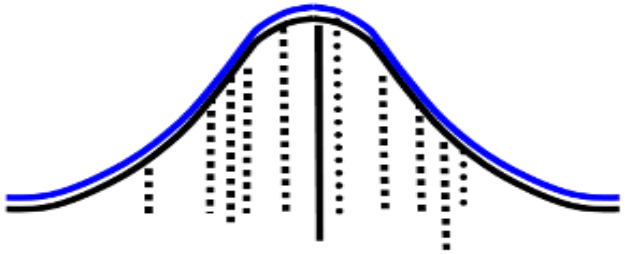


Hypothesis Testing

Do smokers weigh the same as non-smokers?

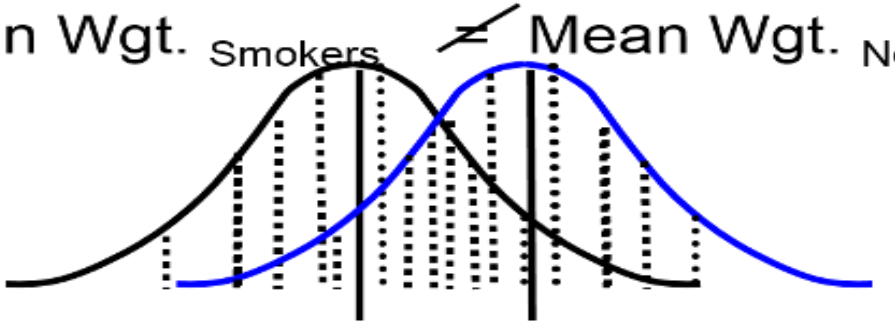
Null Hypothesis (H_0): the average weight does *not* differ

H_0: Mean Wgt. $_{smokers}$ = Mean Wgt. $_{Non\text{-}smokers}$



Alternative Hypothesis (H_A): the average weights differ

H_A: Mean Wgt. $_{Smokers}$ $\neq$ Mean Wgt. $_{Non\text{-}smokers}$



Steps in Hypothesis Testing

- **Step 1**: State your null and alternate hypothesis
- **Step 2**: Collect data
- **Step 3**: Perform a statistical test
- **Step 4**: Decide whether to reject or fail to reject your null hypothesis
- **Step 5**: Present your findings

	Null hypotheses (H_0)	Alternative hypotheses (H_a)
Definition	A claim that there is **no effect** in the population.	A claim that there is **an effect** in the population.
Also known as	H_0	H_a H_1
Typical phrases used	• No effect • No difference • No relationship • No change • Does not increase • Does not decrease	• An effect • A difference • A relationship • A change • Increases • Decreases
Symbols used	Equality symbol (=, ≥, or ≤)	Inequality symbol (≠, <, or >)
$p \leq \alpha$	Rejected	Supported
$p > \alpha$	Failed to reject	Not supported

Excercise

- You want to test whether there is a relationship between gender and height. Based on your knowledge of human physiology, you formulate a hypothesis that men are, on average, taller than women.

To test this hypothesis, you restate it as:

H_0: Men are, on average, not taller than women.
H_a: Men are, on average, taller than women.

Example: In recent years, the mean age of all college students in city X has been 23. A random sample of 42 students revealed a mean age of 23.8. Suppose their ages are normally distributed with a population standard deviation of $\sigma = 2.4$. Can we infer at $\alpha = 0.05$ that the population mean age has changed? $n = 42 \quad \bar{x} = 23.8 \quad \sigma = 2.4 \quad \alpha = 0.05$

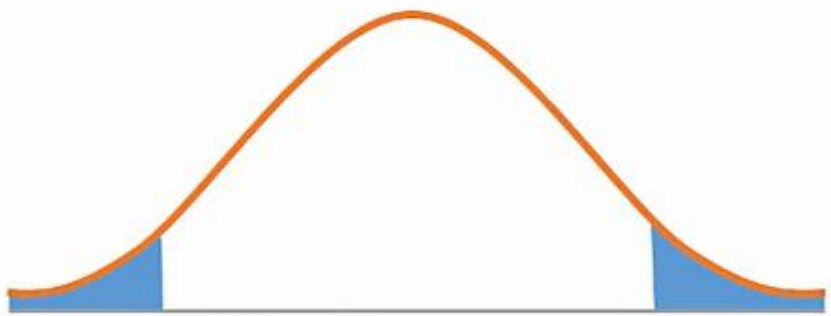
Example: In recent years, the mean age of all college students in city X has been 23. A random sample of 42 students revealed a mean age of 23.8. Suppose their ages are normally distributed with a population standard deviation of σ = 2.4. Can we infer at α = 0.05 that the population mean age has changed? $n = 42 \quad \bar{x} = 23.8 \quad \sigma = 2.4 \quad \alpha = 0.05$

2-tailed test

$H_0: \mu = 23$

$H_1: \mu \neq 23$

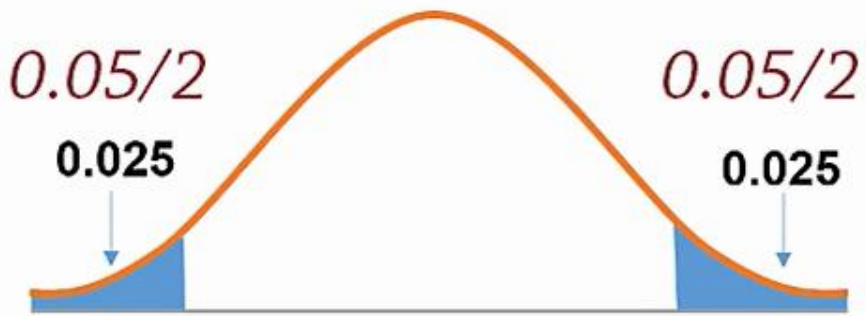
$\alpha = 0.05$



Example: In recent years, the mean age of all college students in city X has been 23. A random sample of 42 students revealed a mean age of 23.8. Suppose their ages are normally distributed with a population standard deviation of σ = 2.4. Can we infer at α = 0.05 that the population mean age has changed? $n = 42 \quad \bar{x} = 23.8 \quad \sigma = 2.4 \quad \alpha = 0.05$

$$H_0: \mu = 23$$
$$H_1: \mu \neq 23$$

$$\alpha = 0.05$$

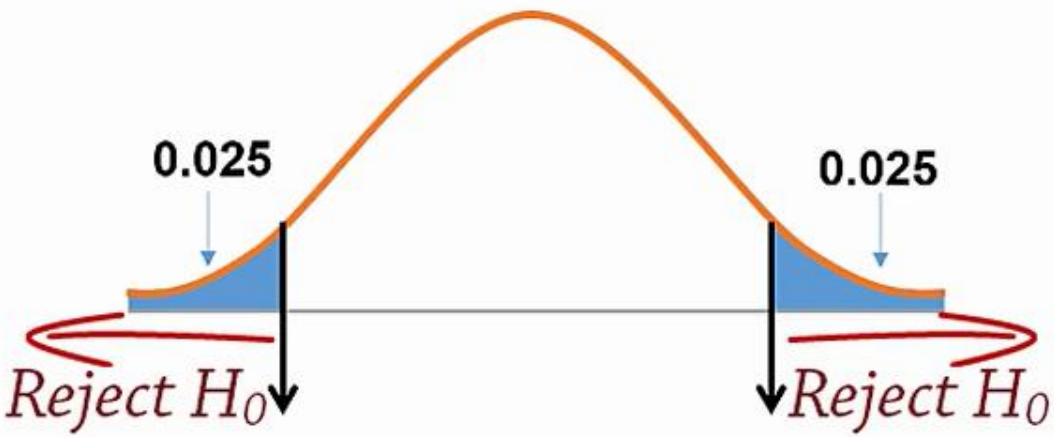


Example: In recent years, the mean age of all college students in city X has been 23. A random sample of 42 students revealed a mean age of 23.8. Suppose their ages are normally distributed with a population standard deviation of σ = 2.4. Can we infer at α = 0.05 that the population mean age has changed?

$$n = 42 \quad \bar{x} = 23.8 \quad \sigma = 2.4 \quad \alpha = 0.05$$

$$H_0: \ \mu = 23$$
$$H_1: \ \mu \neq 23$$

$$\alpha = 0.05$$



Test of μ

z-test

or

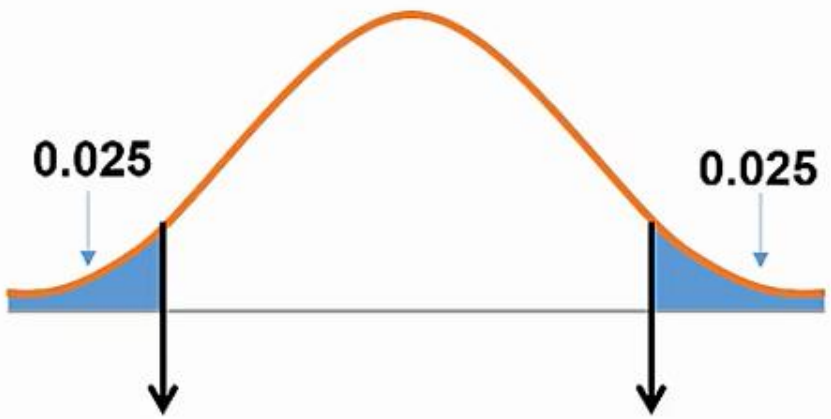
t-test

Example: In recent years, the mean age of all college students in city X has been 23. A random sample of 42 students revealed a mean age of 23.8. Suppose their ages are normally distributed with a population standard deviation of σ = 2.4. Can we infer at α = 0.05 that the population mean age has changed?

$$n = 42 \quad \bar{x} = 23.8 \quad \sigma = 2.4 \quad \alpha = 0.05$$

$$H_0: \ \mu = 23$$
$$H_1: \ \mu \neq 23$$

$$\alpha = 0.05$$



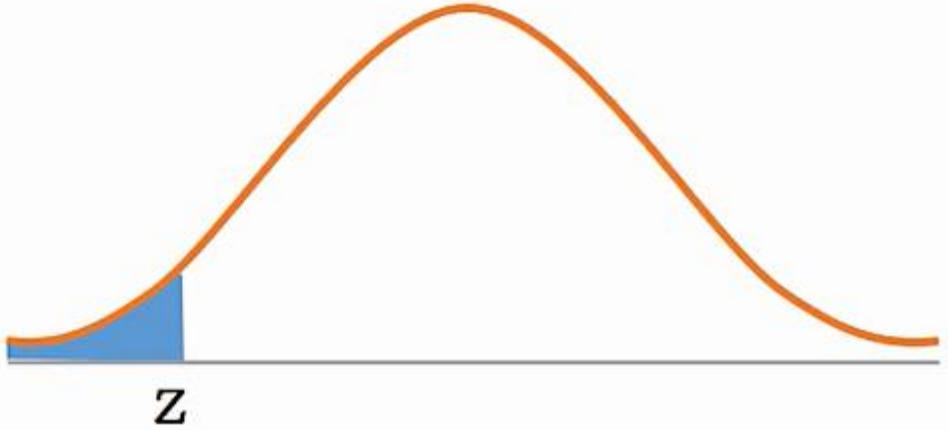
σ *known* $\quad$ $\boldsymbol{z-test}$

σ *unknown* $\quad$ $\boldsymbol{t-test}$

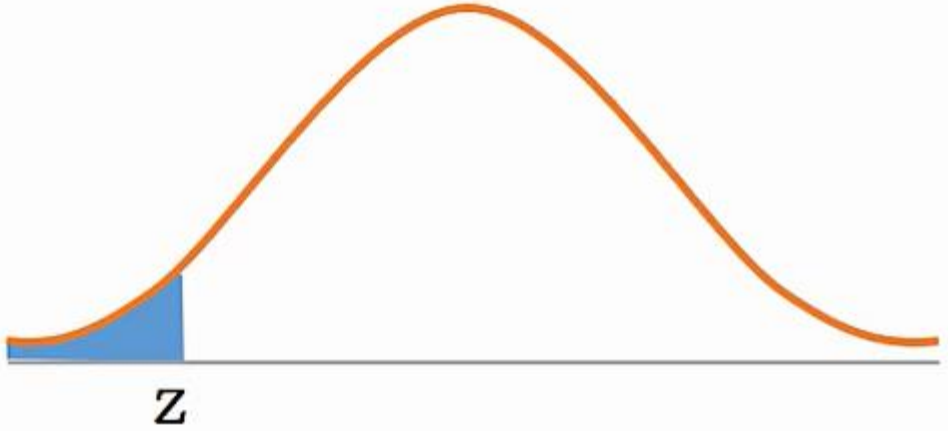
Central Limit Theorem

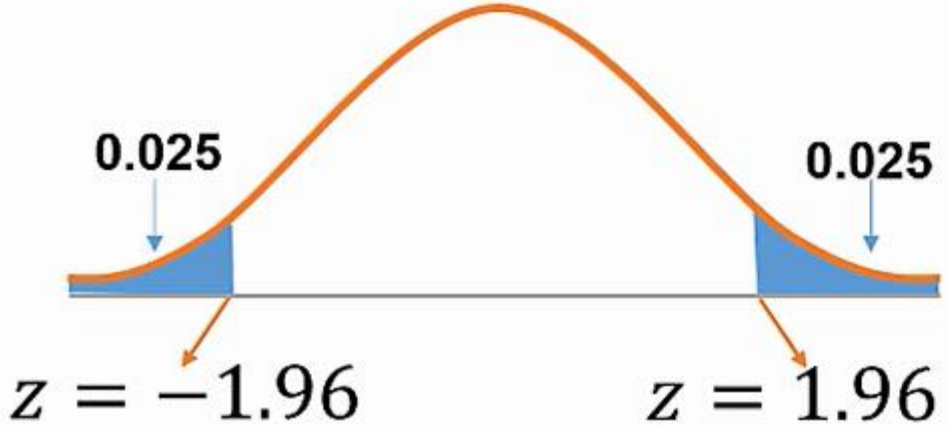
large samples $\quad$ $\boldsymbol{z-test}$

$n \geq 30$

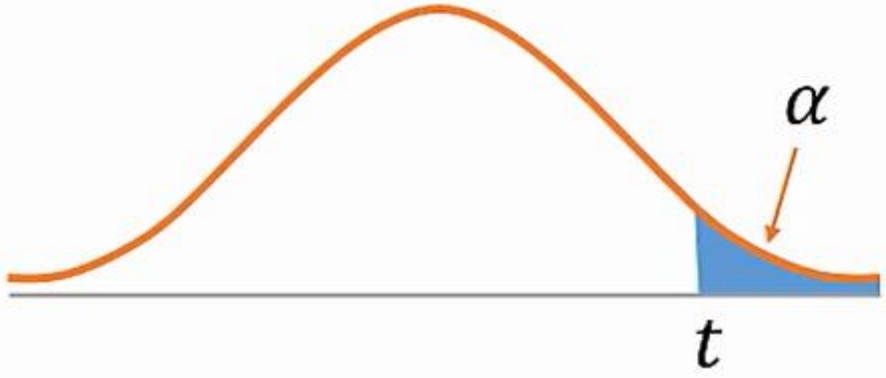


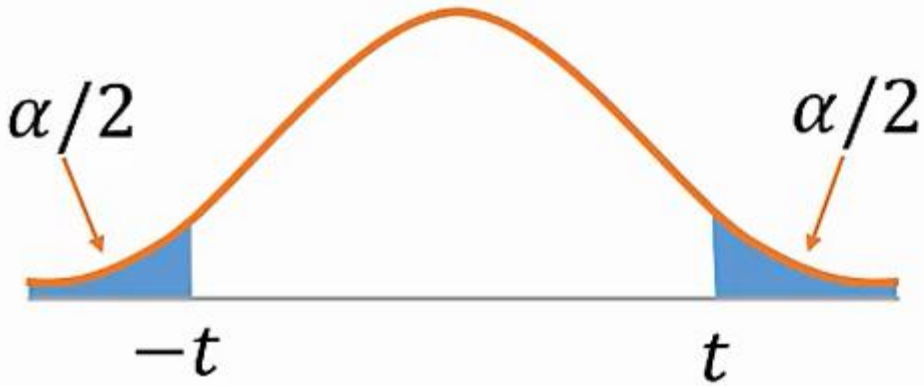
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559





z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559





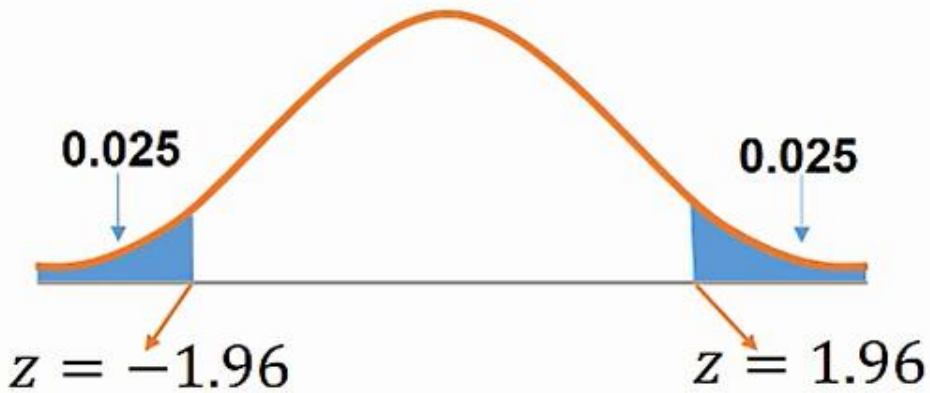
1 tail, α	0.10	0.05	0.025	0.01	0.005
2 tails, α	0.20	0.10	0.05	0.02	0.01
df 1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
Z ∞	**1.282**	**1.645**	**1.960**	**2.326**	**2.576**

Example: In recent years, the mean age of all college students in city X has been 23. A random sample of 42 students revealed a mean age of 23.8. Suppose their ages are normally distributed with a population standard deviation of σ = 2.4. Can we infer at α = 0.05 that the population mean age has changed?

$n = 42 \quad \bar{x} = 23.8 \quad \sigma = 2.4 \quad \alpha = 0.05$

$$H_0: \mu = 23$$
$$H_1: \mu \neq 23$$

$\alpha = 0.05$



Reject H_0 if $z < -1.96$ or $z > 1.96$

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{23.8 - 23}{2.4/\sqrt{42}} = 2.16$$

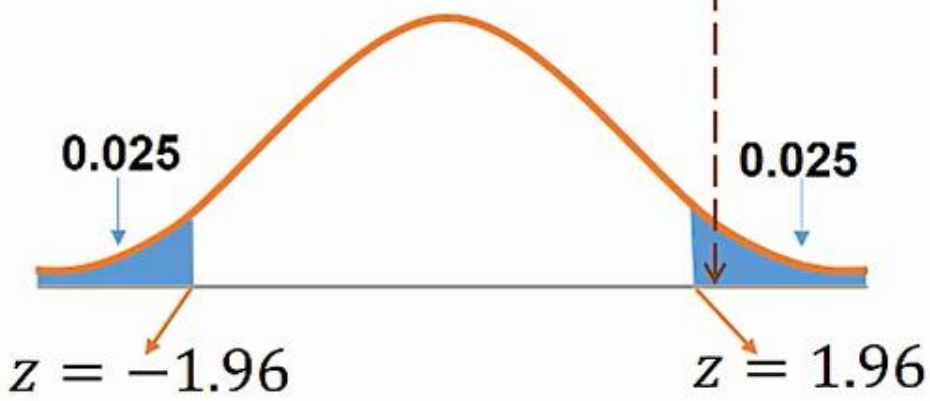
Example: In recent years, the mean age of all college students in city X has been 23. A random sample of 42 students revealed a mean age of 23.8. Suppose their ages are normally distributed with a population standard deviation of σ = 2.4. Can we infer at α = 0.05 that the population mean age has changed?

$n = 42 \quad \bar{x} = 23.8 \quad \sigma = 2.4 \quad \alpha = 0.05$

$H_0: \mu = 23$ ✗

$H_1: \mu \neq 23$ ✓

$\alpha = 0.05$



Reject H_0 if $z < -1.96$ or $z > 1.96$

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{23.8 - 23}{2.4/\sqrt{42}} = 2.16$$

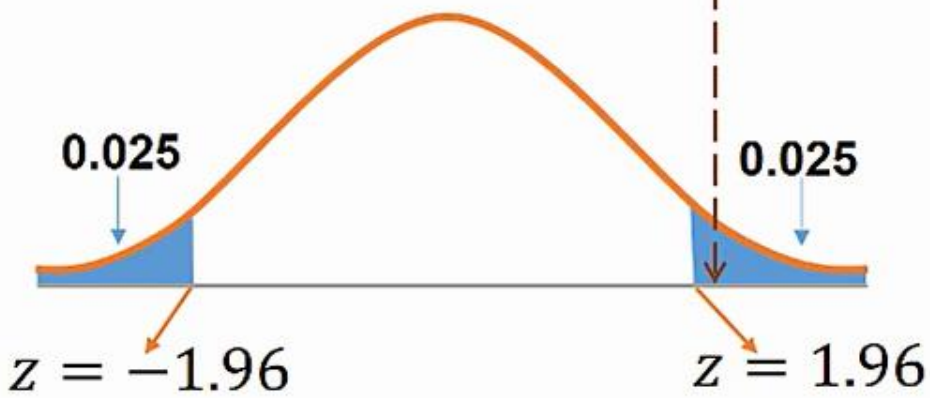
Since $z = 2.16 > 1.96$, reject H_0.

Example: In recent years, the mean age of all college students in city X has been 23. A random sample of 42 students revealed a mean age of 23.8. Suppose their ages are normally distributed with a population standard deviation of σ = 2.4. Can we infer at α = 0.05 that the population mean age has changed?

$n = 42 \quad \bar{x} = 23.8 \quad \sigma = 2.4 \quad \alpha = 0.05$

$$H_0: \mu = 23$$
$$H_1: \mu \neq 23$$

$\alpha = 0.05$



Reject H_0 if $z < -1.96$ or $z > 1.96$

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{23.8 - 23}{2.4/\sqrt{42}} = 2.16$$

Since $z = 2.16 > 1.96$, reject H_0.

There is enough evidence that the mean age has changed.

at $\alpha = 0.05$

Example 2: In recent years, the mean age of all college students in city X has been 23. A random sample of 42 students revealed a mean age of 23.8. Suppose their ages are normally distributed with a population standard deviation of σ = 2.4. Can we infer at α = 0.02 that the population mean age has changed?

$n = 42 \quad \bar{x} = 23.8 \quad \sigma = 2.4 \quad \alpha = 0.02$

$H_0: \quad \mu = 23$

$H_1: \quad \mu \neq 23$

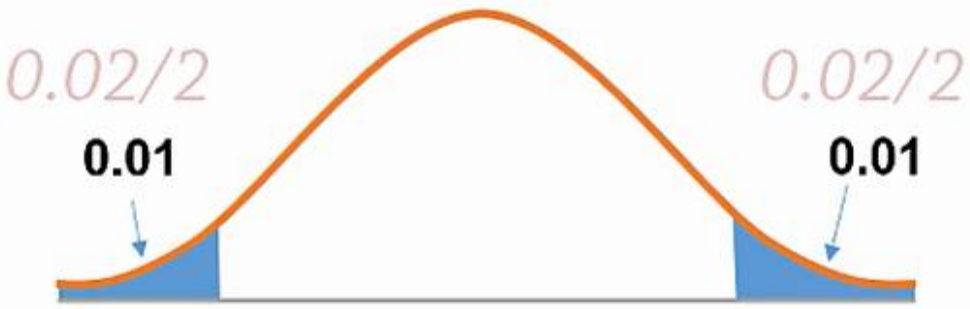
Example 2: In recent years, the mean age of all college students in city X has been 23. A random sample of 42 students revealed a mean age of 23.8. Suppose their ages are normally distributed with a population standard deviation of σ = 2.4. Can we infer at α = 0.02 that the population mean age has changed?

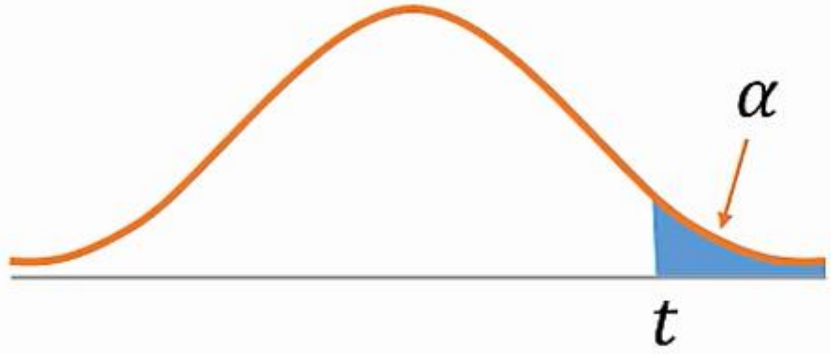
$n = 42 \quad \bar{x} = 23.8 \quad \sigma = 2.4 \quad \alpha = 0.02$

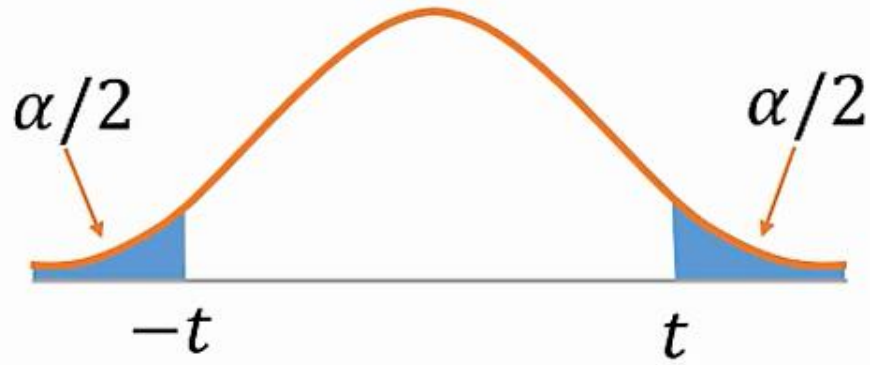
$H_0: \quad \mu = 23$

$H_1: \quad \mu \neq 23$

$\alpha = 0.02$







1 tail, α	0.10	0.05	0.025	0.01	0.005
2 tails, α	0.20	0.10	0.05	0.02	0.01
df 1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
Z ∞	**1.282**	**1.645**	**1.960**	**2.326**	**2.576**

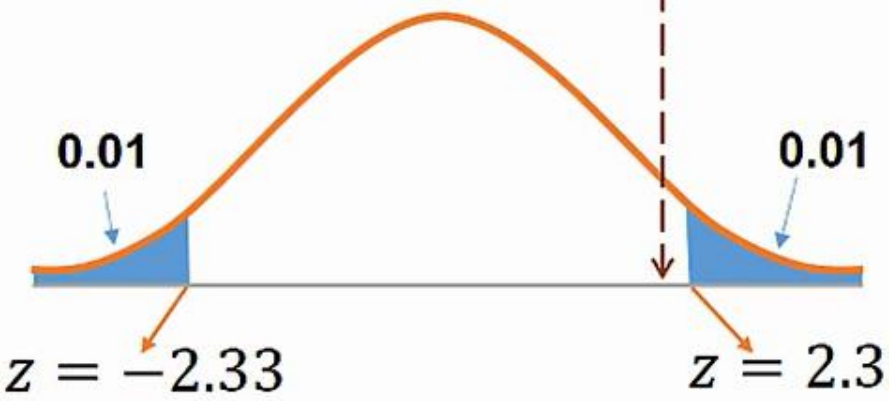
Example 2: In recent years, the mean age of all college students in city X has been 23. A random sample of 42 students revealed a mean age of 23.8. Suppose their ages are normally distributed with a population standard deviation of σ = 2.4. Can we infer at α = 0.02 that the population mean age has changed?

$n = 42 \quad \bar{x} = 23.8 \quad \sigma = 2.4 \quad \alpha = 0.02$

$H_0: \quad \mu = 23$

$H_1: \quad \mu \neq 23$

$\alpha = 0.02$



Reject H_0 if $z < -2.33$ or $z > 2.33$

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{23.8 - 23}{2.4/\sqrt{42}} = 2.16$$

Since $z = 2.16 < 2.33$, fail to reject H_0.

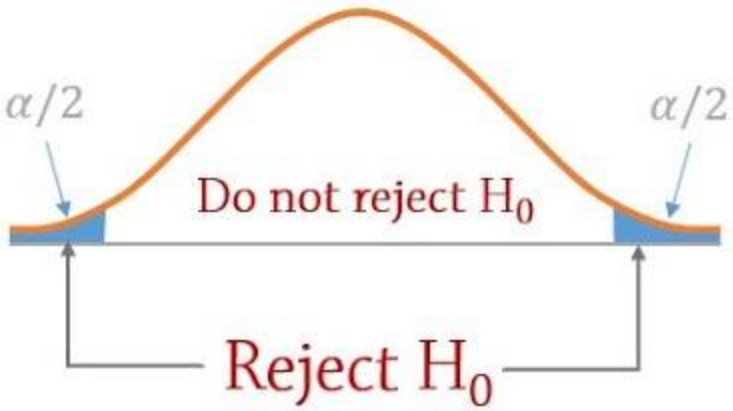
There is not enough evidence that the mean age has changed.

Hypothesis Testing

Two-tailed

$H_0: \mu = 23$

$H_1: \mu \neq 23$

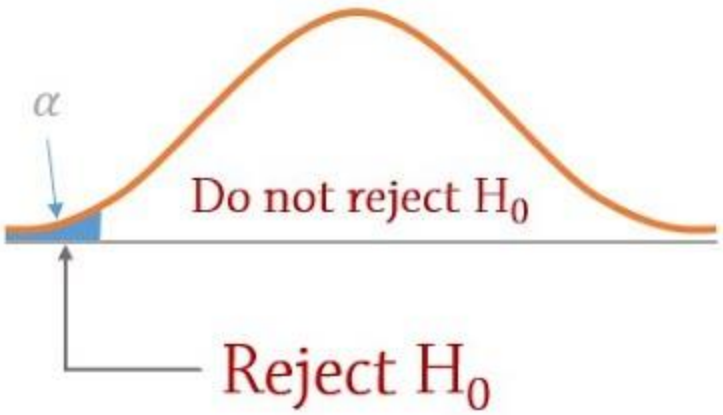


One-tailed

Left-tailed

$H_0: \mu \geq 23$

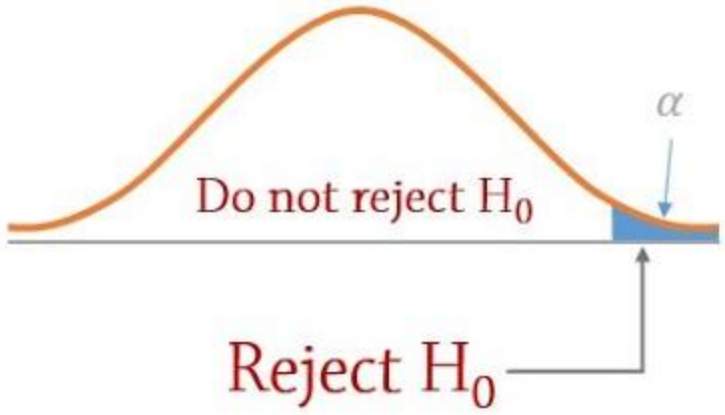
$H_1: \mu < 23$



Right-tailed

$H_0: \mu \leq 23$

$H_1: \mu > 23$



One-tailed Example

- Suppose it is up to you to determine if a certain state receives significantly more public school funding (per student) than the USA average. You know that the USA mean public school yearly funding is $6300 per student per year, with a standard deviation of $400. Next, suppose you collect a sample **(n = 100)** and determine that the sample mean for **New Jersey** (per student per year) is $8801. Use the **z-test** and the correct **Ho** and **Ha** to run a hypothesis test to determine if New Jersey receives significantly **more** funding for public school education (per student per year).

- **Step 1: Set up your hypothesis**
  Hypothesis: The mean per student per year funding in New Jersey is significantly greater than the average per student per year funding over the entire USA.
- **Step 2: Create Ho and Ha**

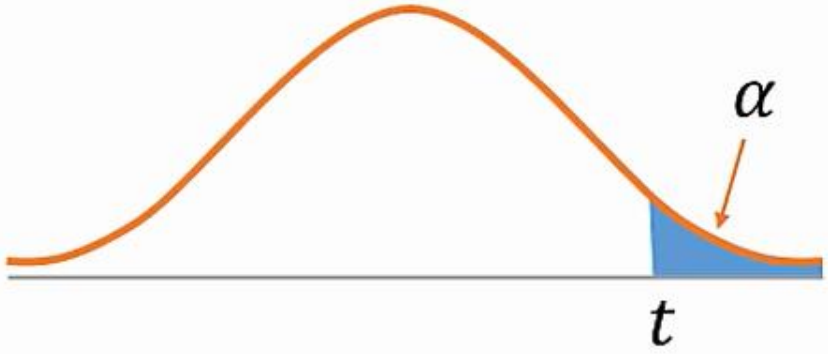
  Ho: mean per student per year funding for New Jersey = mean per student per year funding for the USA

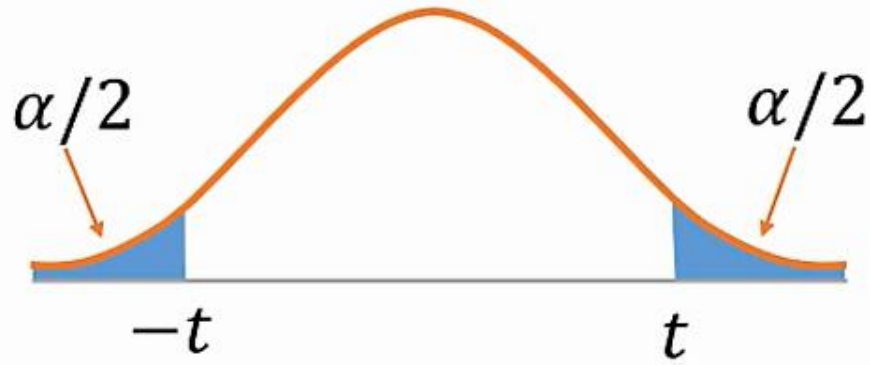
  Ha: mean per student per year funding for New Jersey > mean per student per year funding for the USA
- **Step 3: Calculate the z-test statistic**
  Now, calculate the test statistic.

z = (sample mean – population mean) / [population standard deviation/sqrt(n)]
z = (8801 – 6300) / [400/sqrt(100)]
z = 2501 / [400/10]
z = 2501 / [40]
z = 62.5

- **Step 4: Using the z-table**, determine the **rejection regions** for you z-test.
- In our case, we will use alpha = .05
- This is ONE-TAILED test, therefore the rejection region is any z-test value greater than the critical z value for a one-tailed test with alpha = .05.

  The critical value for one-tailed z-test at alpha = .05 is 1.645.
- Therefore, the rejection region is any value GREATER than 1.645.





1 tail, α	0.10	0.05	0.025	0.01	0.005
2 tails, α	0.20	0.10	0.05	0.02	0.01
df 1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
Z ∞	**1.282**	**1.645**	**1.960**	**2.326**	**2.576**

- **Step 5: Create a conclusion**
  Our z-test result is 62.5.
  This is very large!
  62.5 is MUCH LARGER than 1.645 and so the result of the z test is INSIDE the rejection region.
- **Conclusion:**
  The funding for **New Jersey** public schools (per student per year) is **significantly GREATER than** the average funding per student per year for the **USA**.


KEEP IT UP