Artificial Intelligence Algorithms and Mathematics

CSCN 8000

Statistics

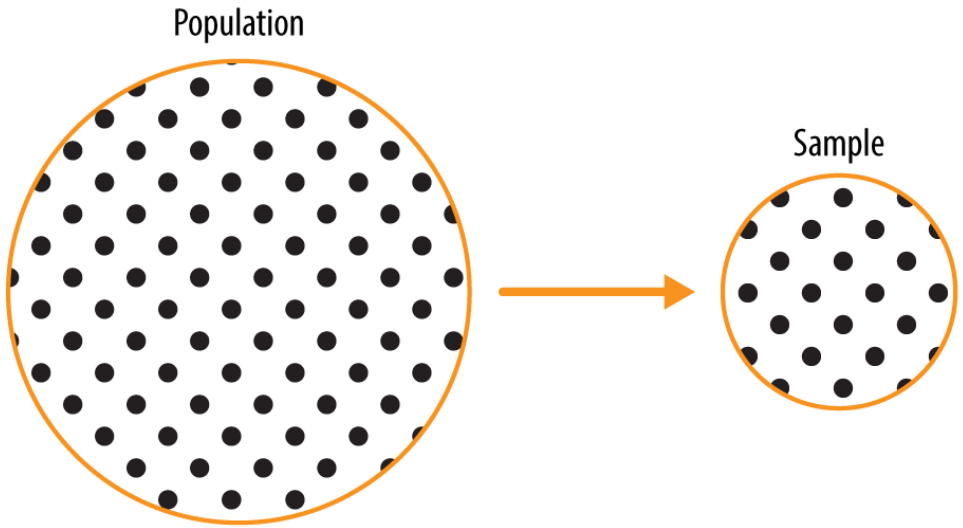
- Statistics
  - Descriptive Vs. Inferential statistics
- Population and Sample
- Sampling
- Measure of Central Tendency
- Measure of Dispersion
- Normal Distribution
- Z Score

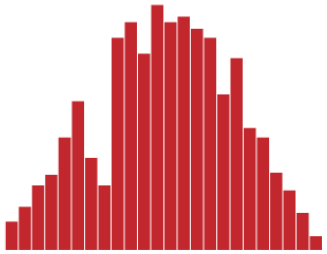
Statistics

- Science of collecting, organizing and analyzing the data.
- Descriptive Statistics:
  - Used to summarize and describe the main features of a dataset.
  - They provide a clear and concise overview of the data, helping to understand its characteristics.
  - Examples: Mean, median, mode, etc.
- Inferential Statistics:
  - Make inferences or draw conclusions about a population based on a sample of data.
  - Examples: Hypothesis testing, confidence intervals, etc.

Population and Sample

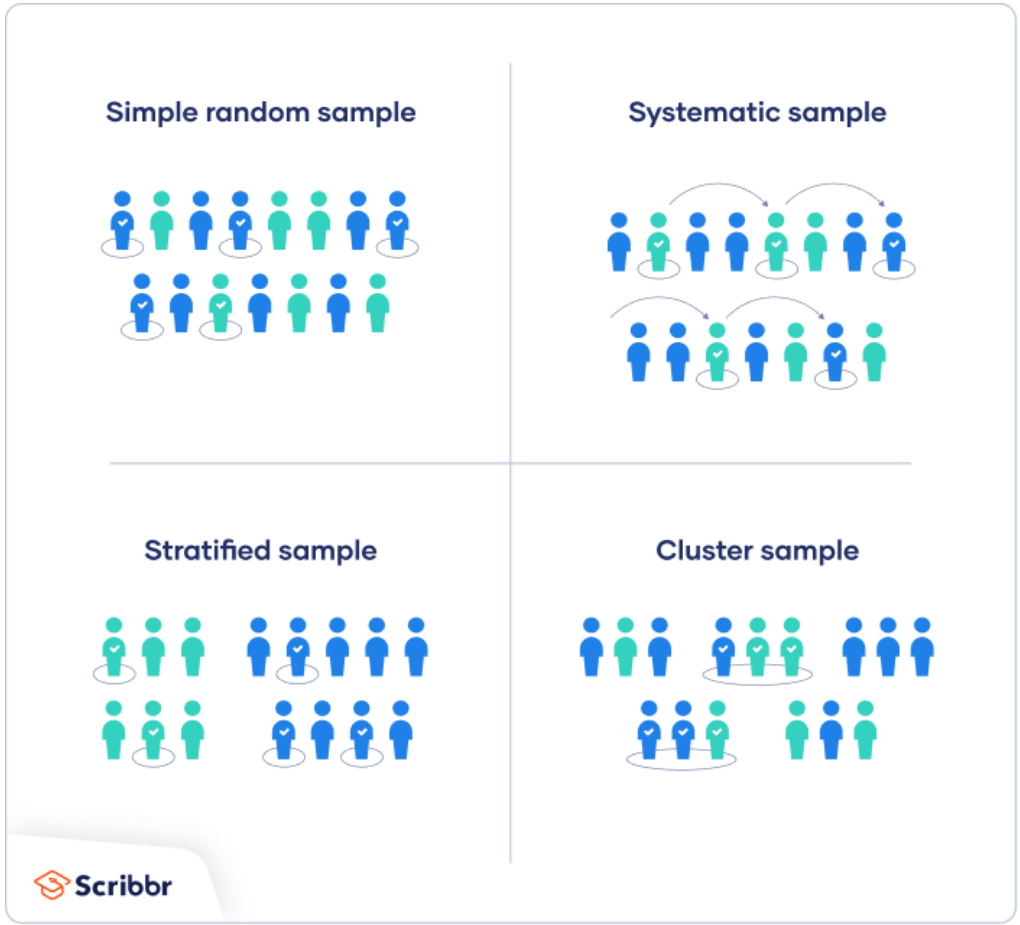
Population – Entire Data (N)
Sample – Small set of data (n)





How to generate the samples?

- Simple Random Sampling:
  - Every member of the population has an equal chance of being selected for sample.
  - Best representation of population.
- Stratified Sampling:
  - The population is split into nonoverlapping groups then randomly sample from each group.
  - Beneficial for handling imbalanced datasets.
- Systematic Sampling:
  - From the population, we select the nth individual.
- Convenience Sampling:
  - From the population, we select the sample data that is expertise in the specific domain.



Measures of Central Tendency

- Central Tendency – the measure used to determine the center of distribution of data.
- Mean(μ) is the average of the population (N) or sample(n).

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

  - Examples:
    - $S = \{1,1,2,2,3,4,5,1,2,1,3\}, \quad \mu(S) = \frac{25}{11} = 2.3$
    - $S = \{1,1,2,2,3,4,5,1,2,1,3,100\}, \quad \mu(S) = \frac{350}{11} = 11.3$
  - Drawback: Affected by outliers

Measures of Central Tendency

Name of the Person	Earning
Sean	30000
Carl	15000
John	20000
Ranny	15250
Dansh	17500
Tony	15500
James	10000
Lisa	15000

Mean: 13,8250

Name of the Person	Earning
Sean	30000
Carl	15000
John	20000
Ranny	15250
Dansh	17500
Tony	15500
James	10000
Lisa	150000

Mean: 1,638,250 → Outlier

Measures of Central Tendency

- Median: It is the middle value in a sorted dataset.
  - Sort the numbers ascendingly
  - Find the center number
    - If odd number of samples → Take single center number as median.
    - If even number of samples → Take average of the two middle numbers.
- Examples:
  - $S = \{1,1,1,1,2,2,2,3,3,4,5\}$,
    - $median(S) = 2$
  - $S = \{1,1,1,1,2,2,2,3,3,4,5,100\}$,
    - $median(S) = \frac{2+2}{2} = \mathbf{2}$
- Pros: Not affected by outliers

Measures of Central Tendency

Name of the Person	Earning
Sean	30000
Carl	15000
John	20000
Ranny	15250
Dansh	17500
Tony	15500
James	10000
Lisa	15000

Sorting

Name of the Person	Earning
James	10000
Carl	15000
Ranny	15250
Tony	15500
Dansh	17500
John	20000
Sean	30000
Lisa	150000

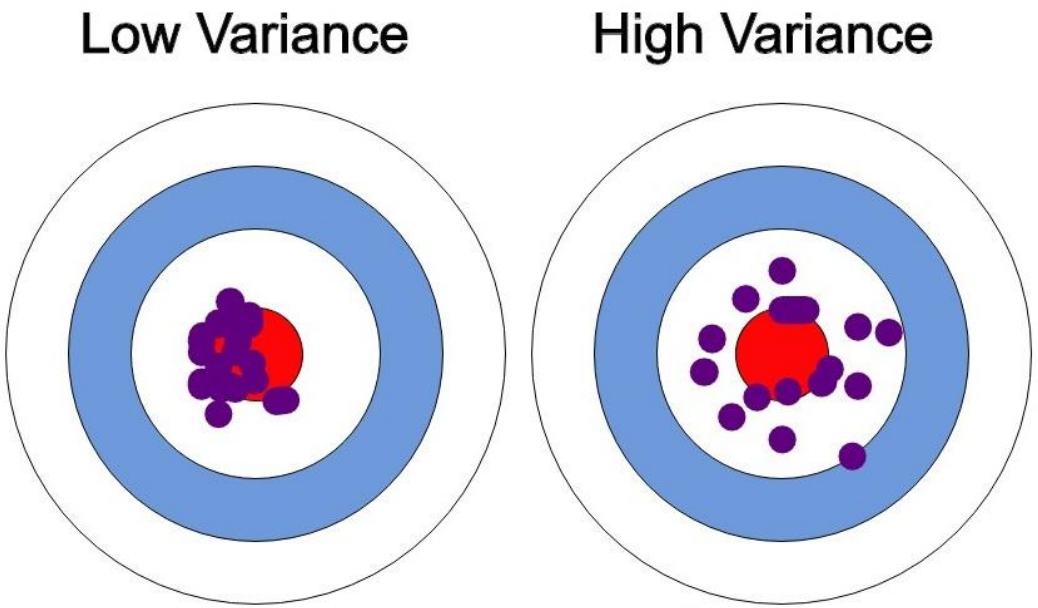
Median: 16,500

Measures of Central Tendency

- Mode: the value that occurs most frequently in a dataset.
- Examples:
  - $S = \{1,1,2,7,3,4,5,1,8,1,10\}$,
    - $mode(S) = 1 \rightarrow Unimodal$
  - $S = \{1,1,2,2,3,4,5,1,2,1,6\}$
    - $mode(S) = \{1,2\} \rightarrow Multimodal$
- Usage:
  - Could fill the missing values in a column using the mode value.

Measures of Dispersion

- Dispersion – How well spread the data points are?
  - Variance
  - Standard Deviation



Measures of Dispersion

- Variance: Variance is the average of the squared differences from the Mean.
  - Population Variance - $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$
  - Sample Variance - $s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu)^2$ → Bessel's Correction
- Example:
  - $S = \{1,2,2,3,4,6\}$
    - $\mu = 3$
    - $s^2 = \frac{(1-3)^2}{5} + \frac{(2-3)^2}{5} + \ \ldots = 3.2$

Measures of Dispersion

- Standard Deviation: the square root of the variance. It provides a measure of the average distance between each data point and the mean.

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

- Example:
  - $S = \{1,2,2,3,4,6\}$
    - $\mu = 3$
    - $s^2 = \frac{(1-3)^2}{5} + \frac{(2-3)^2}{5} + \ \ldots = 3.2$
    - $\sigma = \sqrt{3.2} = 1.78$
- What if we want to study how ***two*** variables vary compared to each other?

Covariance

- **Covariance:** It indicates the direction of the *linear relationship* between two variables. It tells us whether both variables tend to increase or decrease at the same time.
  - Positive Covariance: the two variables tend to move in the same direction: as one increases, the other tends to increase, and vice-versa.
  - Negative Covariance: the two variables tend to move in opposite directions: as one increases, the other tends to decrease, and vice versa.
  - Zero Covariance: No linear relationship exists.
- Main Drawback:
  - It can range from negative infinity to positive infinity.
  - The magnitude (how large the relationship is) is hard to interpret because it depends on the units of the variables.

$$Cov(x,y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Correlation

- **Correlation:** is a standardized version of covariance that measures the ***strength*** and ***direction*** of the linear relationship between two variables. It provides a value between -1 and 1:
  - A correlation of 1 means there's a perfect positive linear relationship (as one variable increases, the other does too).
  - A correlation of -1 means there's a perfect negative linear relationship (as one variable increases, the other decreases).
  - A correlation of 0 means there is no linear relationship between the variable.
- Formulation:
  - $Corr(x, y) = \frac{Cov(x,y)}{\sigma_x \sigma_y}$

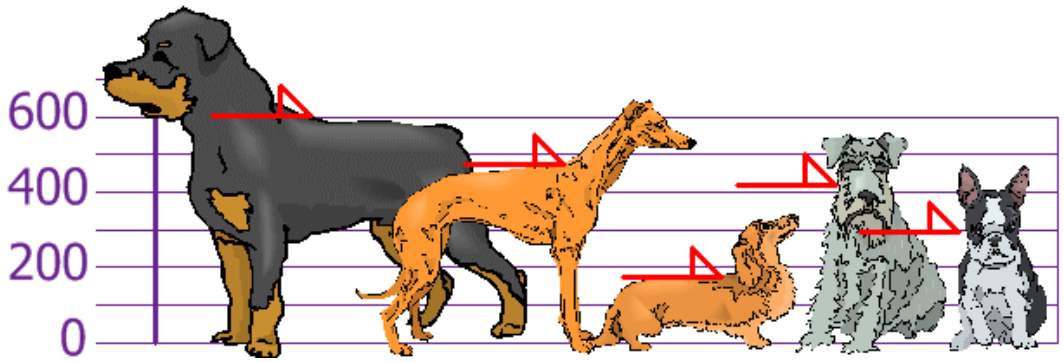
Hours Studied (x)	Exam Score (y)
1	2
2	4
3	5
4	4
5	5

$$Cov(x, y) = ?$$
$$Corr(x, y) = ?$$

$$Cov(x, y) = -0.5$$
$$Corr(x, y) = -0.28$$

Example

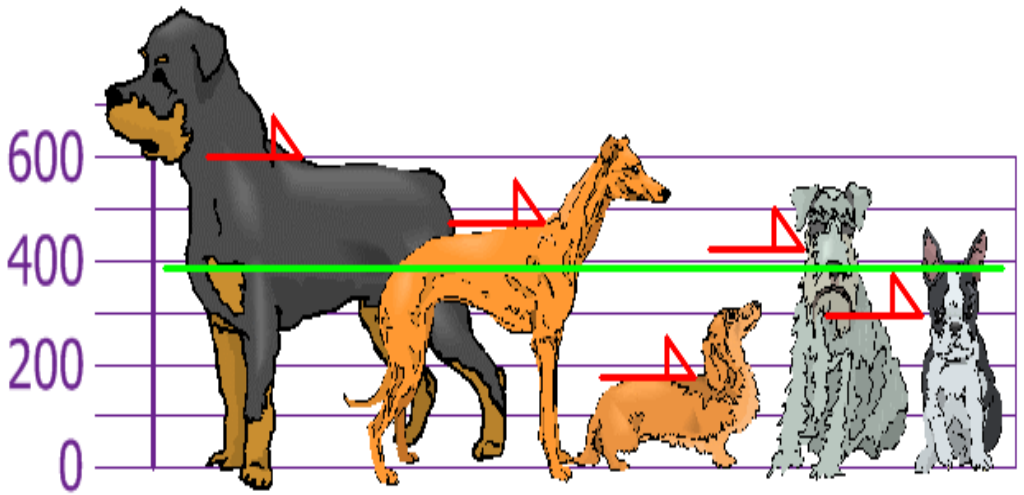
- You and your friends have just measured the heights of your dogs (in millimeters). The heights (at the shoulders) are: $H = [600\ mm, 470\ mm, 170\ mm, 430\ mm\ and\ 300\ mm]$



- Find out the Mean, the Variance, and the Standard Deviation.

Example: Mean

- $\mu(H) = \dfrac{600 + 470 + 170 + 430 + 300}{5}$
- $\mu(H) = \dfrac{1970}{5} = 394$



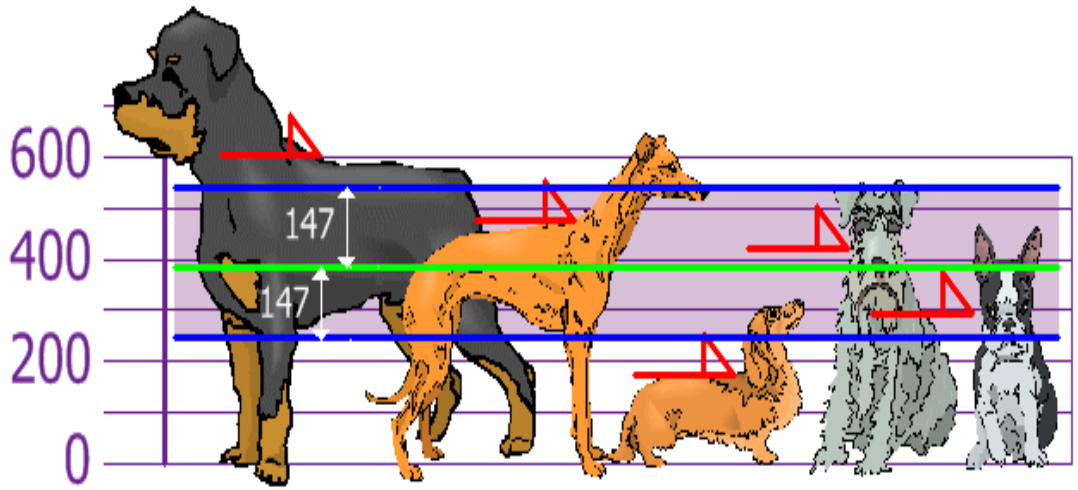
Example: Variance

- $\sigma^2 = \frac{206^2 + 76^2 + (-224)^2 + 36^2 + (-94)^2}{5}$
- $\sigma^2 = \frac{42436 + 5776 + 50176 + 1296 + 8836}{5}$
- $\sigma^2 = \frac{108520}{5} = 21704$

- $\sigma^2 = \frac{206^2 + 76^2 + (-224)^2 + 36^2 + (-94)^2}{5}$
- $\sigma^2 = \frac{42436 + 5776 + 50176 + 1296 + 8836}{5}$
- $\sigma^2 = \frac{108520}{5} = 21704$

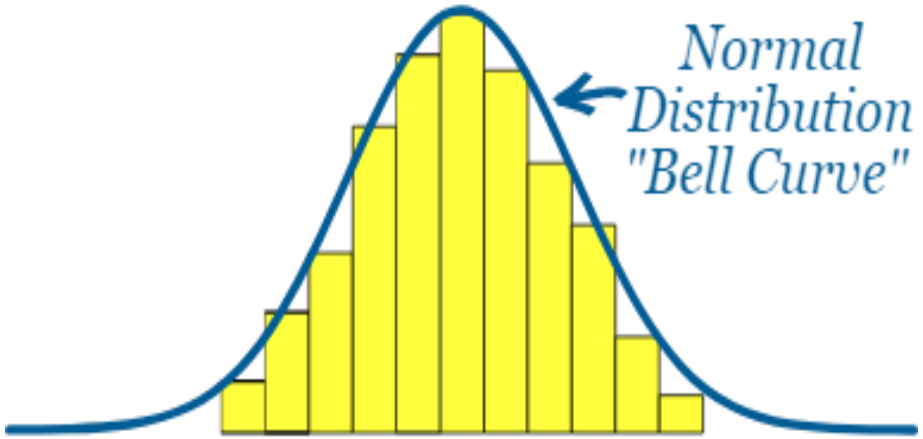
Example: Standard Deviation

- $\sigma = \sqrt{\sigma^2} = \sqrt{21704} = 147\ mm$



Normal distribution

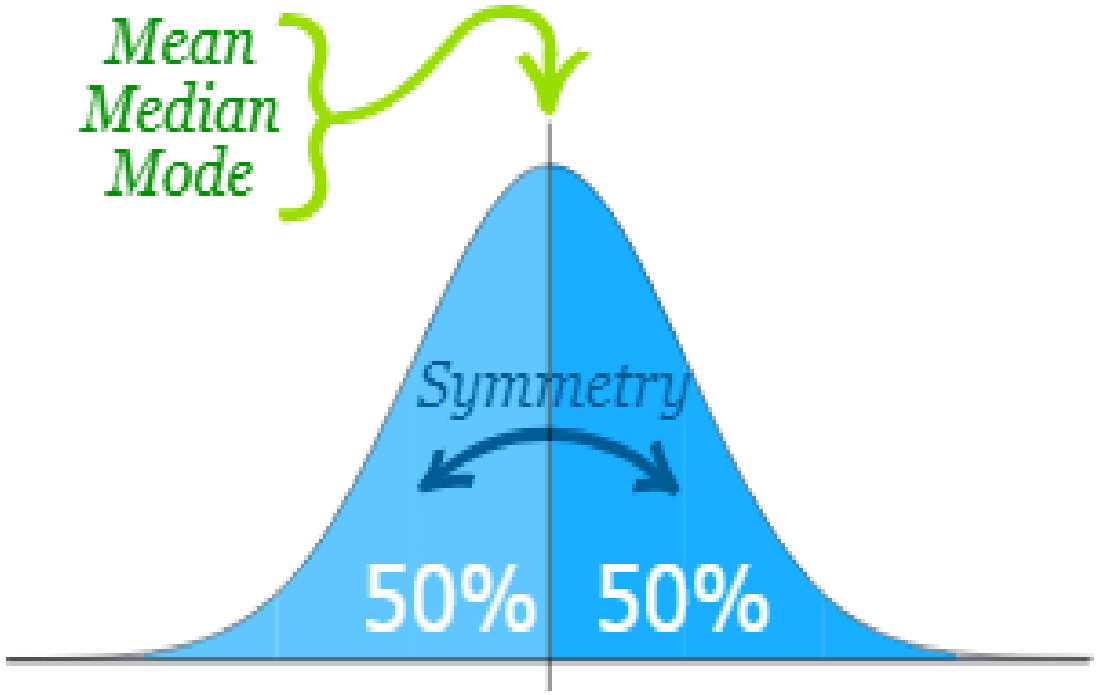
- Data tends to be around a central value with no bias left or right, and it gets close to a "Normal Distribution" like this:



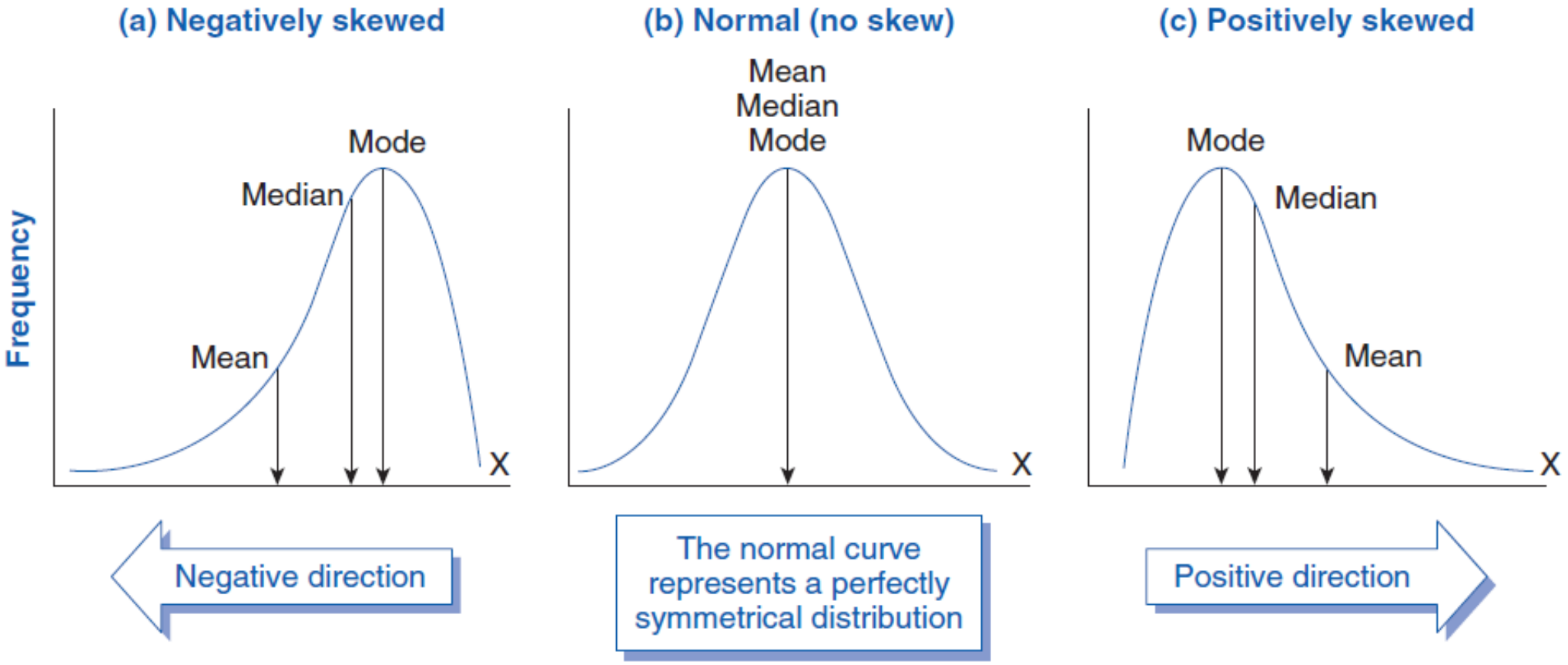
The blue curve is a Normal Distribution. The yellow histogram shows some data that follows it closely, but not perfectly. It is often called a "Bell Curve"

Normal Distribution

- The Normal Distribution has:
  - mean = median = mode
  - symmetry about the center
  - 50% of values less than the mean
  - and 50% greater than the mean

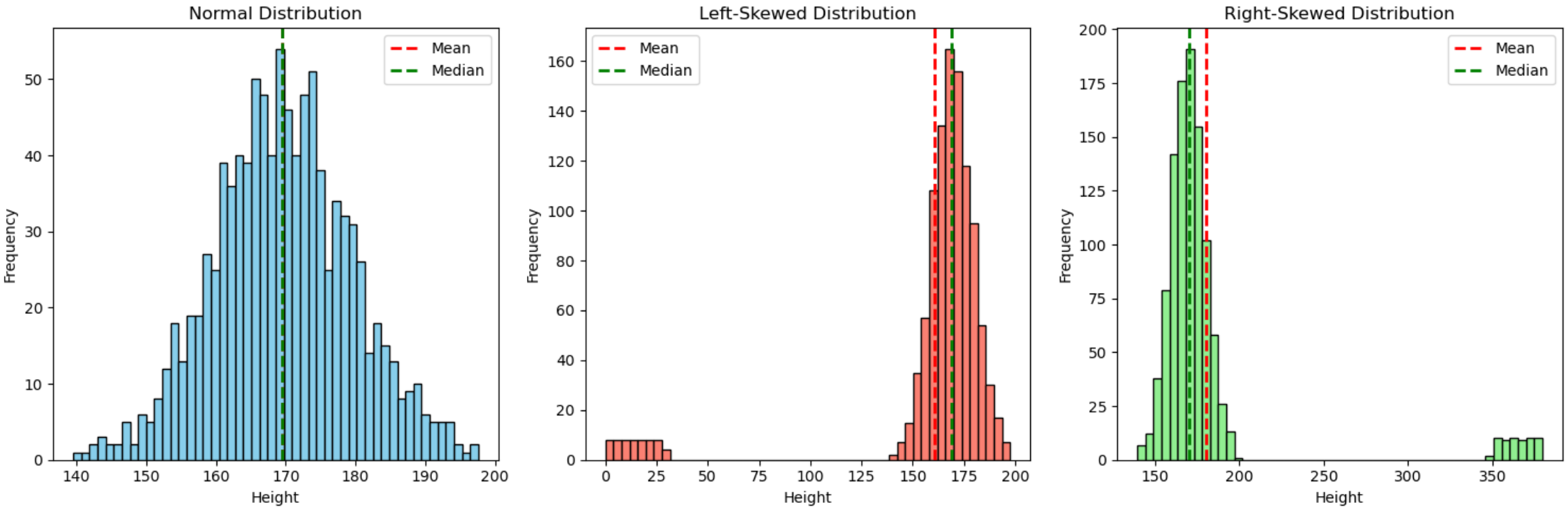


Normal vs Skewed Distributions



- In Skewed distributions the ***median*** is usually utilized instead of the ***mean*** as a representative center of tendency.

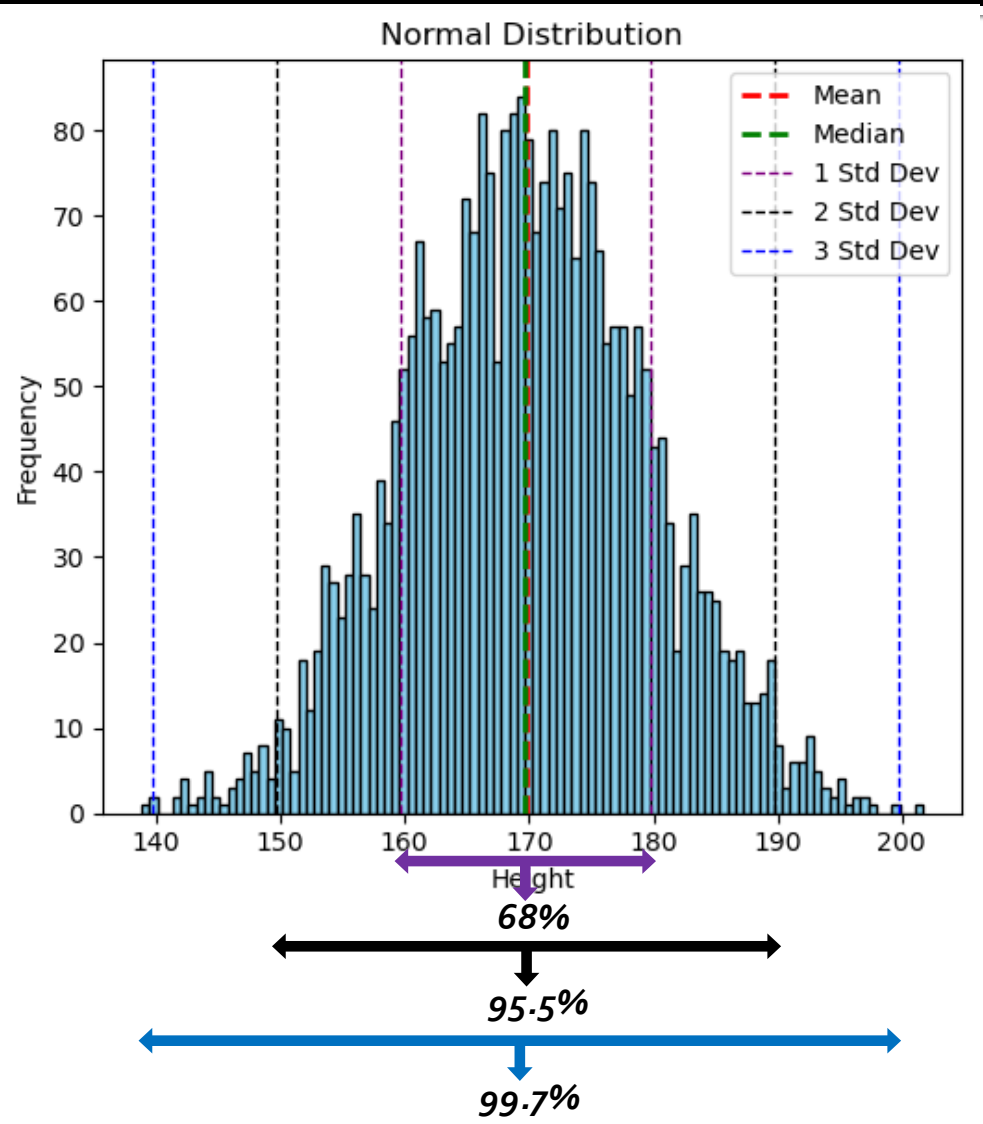
Normal vs Skewed Distributions: Example



- For this Height feature in a diagnosis prediction dataset, outliers caused skewness of distribution.
- The median is least affected by this compared to the mean.

Gaussian /Normal Distribution

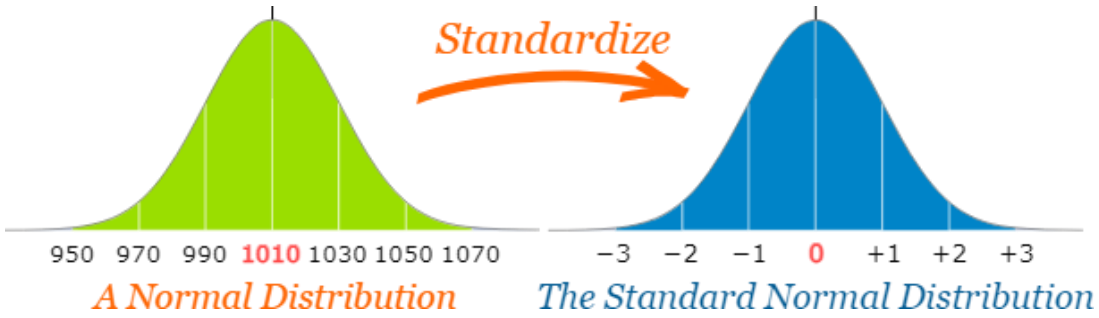
- In a normal distribution:
  - 68% of all values are within 1 standard deviation from mean
  - 95.5% of all values are within 2 standard deviations from mean
  - 99.7% of all values are within 3 standard deviations from mean



Z-score / Standard Score

- The number of standard deviations from the mean is also called the "Standard Score", "sigma" or "z-score".
- To convert a value to a Standard Score ("z-score"):
  - first subtract the mean,
  - then divide by the Standard Deviation
  - And doing that is called "Standardizing":
- Formula:

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



Z-score: Example

- Convert the following dataset values to Z-score representation.

Name of the Person	Earning
James	10000
Carl	15000
Ranny	15250
Tony	15500
Dansh	17500
John	20000
Sean	30000
Lisa	150000

$\mu = 34156.25, \sigma = 44117.55$

Name of the Person	Earning
James	-0.548
Carl	-0.434
Ranny	-0.429
Tony	-0.423
Dansh	-0.378
John	-0.321
Sean	-0.094
Lisa	2.626

Outlier Removal: Z-Score

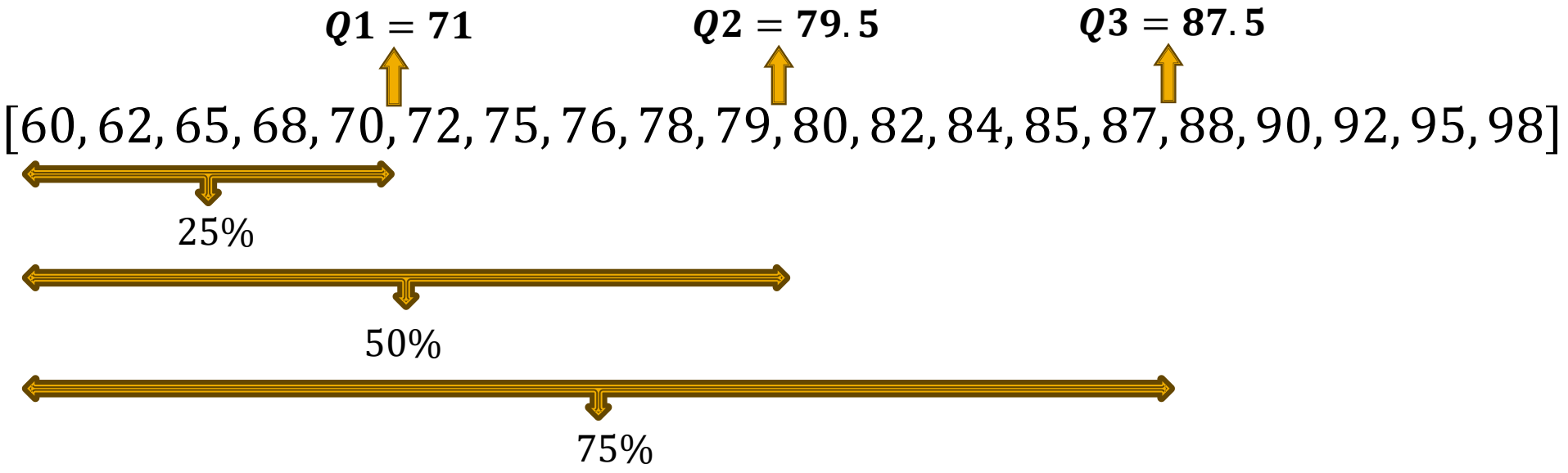
- Data points that have a z-score above a certain threshold τ could be labeled as outliers.
- Steps:
  - Normalize the feature to Z-score representation.
  - Based on threshold τ, any value larger than τ is an outlier
  - Remove the outliers.
- τ usually is set to a number between 2 and 3, but it could be any arbitrary value.

Percentiles and Quartiles

- Assume we have a dataset of 20 student exam scores.

$$[75, 80, 65, 90, 85, 70, 88, 92, 78, 68, 95, 60, 72, 82, 98, 76, 84, 62, 79, 87]$$

- Let's start by sorting the values ascendingly.

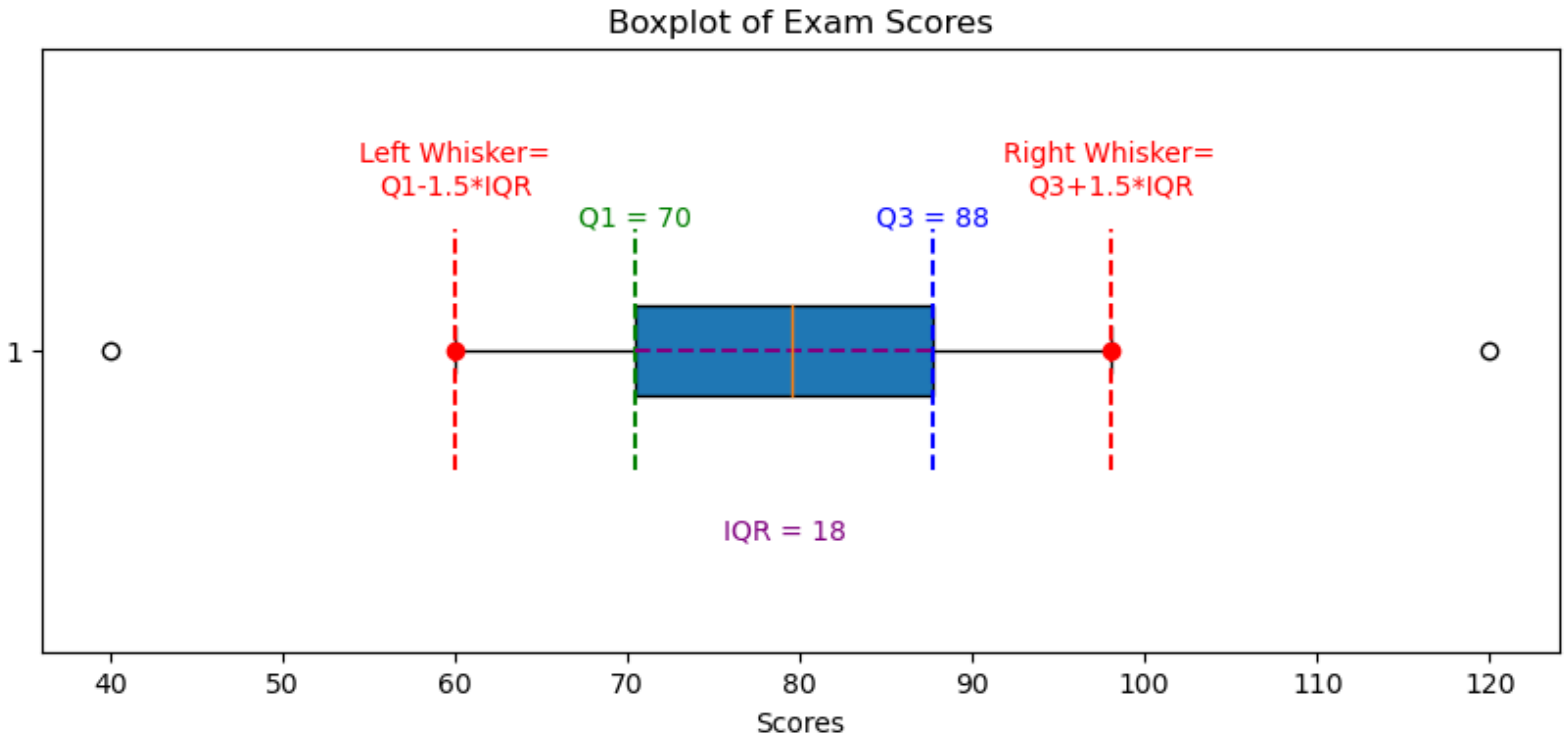


Percentiles and Quartiles

- Quartiles divide a dataset into four equal parts. The three quartiles, denoted as Q1, Q2, and Q3, represent the following:
  - **Q1 (First Quartile)**: This is the value below which 25% of the data falls. It is the median of the lower half of the data set.
  - **Q2 (Second Quartile or Median)**: This is the value below which 50% of the data falls. It is the median of the entire dataset.
  - **Q3 (Third Quartile)**: This is the value below which 75% of the data falls. It is the median of the upper half of the data set.
- Percentiles are similar to quartiles but are more general. Quartiles specifically divide data into four parts, while percentiles divide data into n equal parts, where n is any whole number.

Plotting Quartiles: Boxplots

$$[40, 60, 62, 65, 68, \mathbf{70}, 72, 75, 76, 78, \mathbf{79}, \mathbf{80}, 82, 84, 85, 87, \mathbf{88}, 90, 92, 95, 98, 120]$$

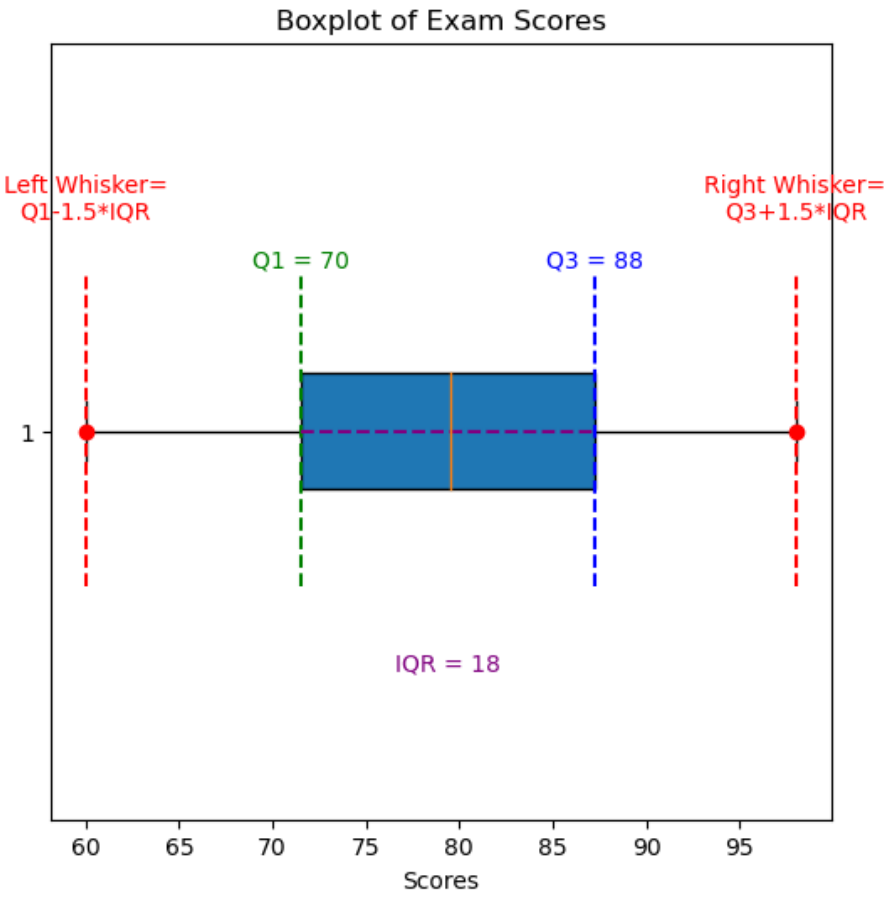


$$Left\ Whisker = Q1 - 1.5 * IQR$$

$$Right\ Whisker = Q3 + 1.5 * IQR$$

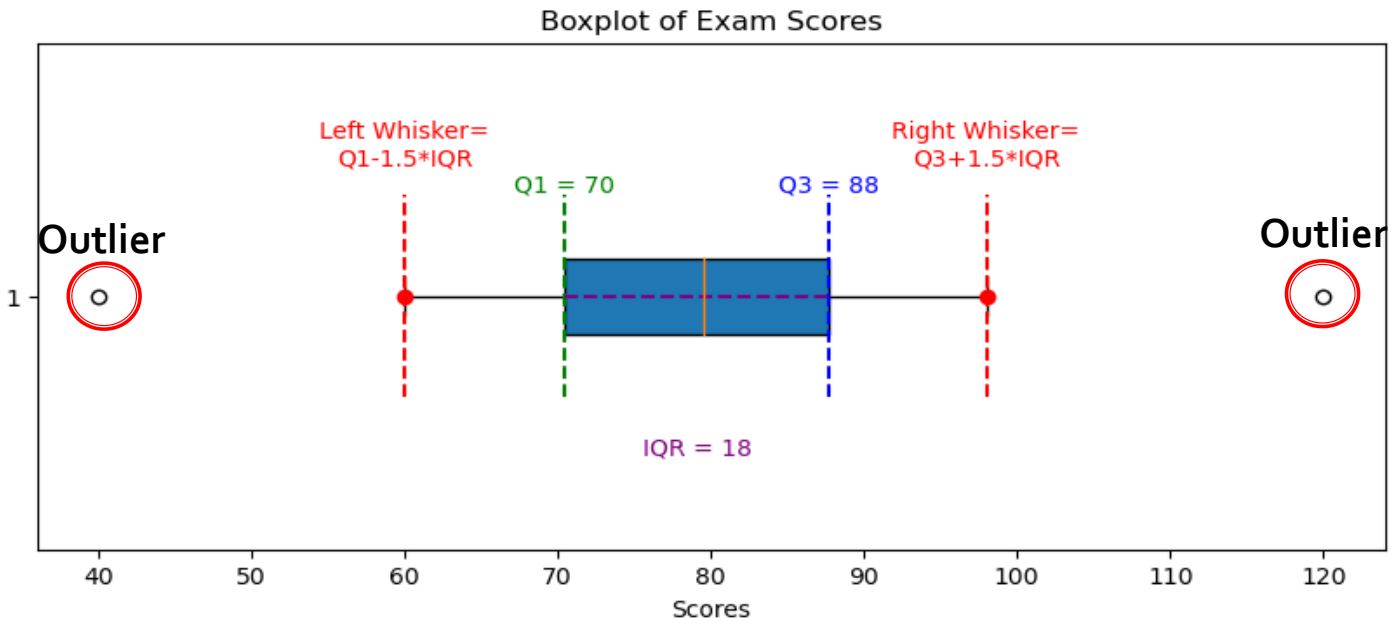
Plotting Quartiles: Boxplots

- **Box Plot:** is a graphical representation of a dataset's distribution. It displays a summary of a set of data values including the minimum, first quartile, median, third quartile, and maximum.
- **Left Whisker:** It represents the lower bound of the data within 1.5 times the interquartile range (IQR) below the first quartile (Q1).
- **Right Whisker:** It represents the upper bound of the data within 1.5 times the IQR above the third quartile (Q3).
- **Inter-Quartile Range (IQR):** it represents the spread of the middle 50% of the data. It is calculated as the difference between the third quartile (Q3) and the first quartile



Outlier Removal: Boxplots

- Data points that have values x above the right whisker or below the left whisker → labeled as outliers.
- In other words:
  - If $x > Q3 + 1.5 * IQR \; or \; x < Q1 - 1.5 * IQR$ → Outlier

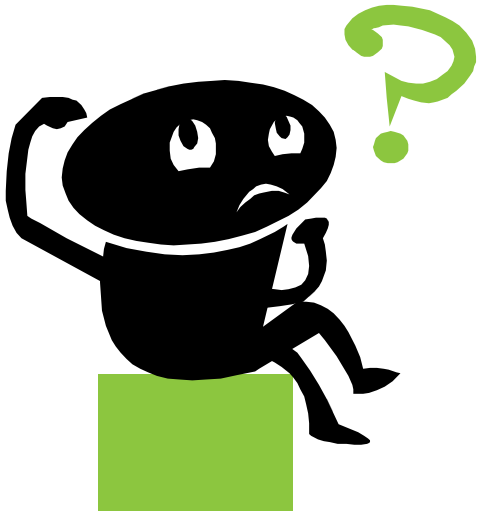


References

- https://learning.oreilly.com/library/view/practical-statistics-for/9781491952955/ch06.html
- https://www.mathsisfun.com/data/standard-deviation.html

Thank you!

- Any questions?



CONESTOGA

Thank You

Youssef Abdelkareem

yabdelkareem@conestogac.on.ca

**** Credits for Dr. Anasuya Bhima for assisting in the slides structure***