Artificial Intelligence (Machine Learning & Deep Learning) [Course]

Week 2 - Day 4 n Day 5

[Descriptive Statistics and Probability]

[See examples / code in GitHub code repository]

It is not about Theory, it is 20% Theory and 80% Practical – Technical/Development/Programming [Mostly Python based]

Theoretical Background

Descriptive Statistics

What Are Descriptive Statistics?

Descriptive statistics are brief informational coefficients that summarize a given data set, which can be either a representation of the entire population or a sample of a population. Descriptive statistics are broken down into measures of central tendency and measures of variability (spread). Measures of central tendency include the mean, median, and mode, while measures of variability include standard deviation, variance, minimum and maximum variables, kurtosis, and skewness.

*****Details about various calculation above, later in course

References:

https://www.investopedia.com/terms/d/descriptive_statistics.asp
https://corporatefinanceinstitute.com/resources/data-science/descriptive-statistics/

Theoretical Background

Probability

Probability is simply how likely something is to happen.

Probability means possibility. It is a branch of mathematics that deals with the occurrence of a random event.

Example 1:
There are 6 pillows in a bed, 3 are red, 2 are yellow and 1 is blue. What is the probability of picking a yellow pillow?

Ans: The probability is equal to the number of yellow pillows in the bed divided by the total number of pillows, i.e. 2/6 = 1/3.

Example 2:Flipping a coin:

$$P(H) = \frac{1}{2} = 50\%$$

References:

https://www.khanacademy.org/math/statistics-probability/prob
library/basic-theoretical-probability/a/probability-the-basics
https://byjus.com/maths/probability/
https://www.cuemath.com/data/probability/

Theoretical Background

Data and its types (structured, Unstructured)

Properties	Structured data	Unstructured data
Format examples	• CSV • Excel	• audio files (WAV, MP3, OGG) • PDF documents • images (JPEG, PNG, etc.)
Sources examples	• online forms • point-of-sale (POS) systems • online transaction processing (OLTP) systems	• emails • social media posts • multimedia files • IoT outputs
Nature of data	Quantitative	Qualitative
Databases	Relational (SQL)	Non-relational (NoSQL)
Storage for analytics use	Warehouses and data lakehouses	Data lakes and data lakehouses
Specialists to handle data	Business analysts, software engineers, data analysts	Data scientists, data engineers, data analysts
Main benefits	Easy to search and analyze, doesn't require much space	Easy to collect and store
Main challenges	All data must fit predefined schema	Difficult to search and analyze

References: https://www.altexsoft.com/blog/structured-unstructured-da
https://www.ibm.com/think/topics/structured-vs-unstructur

Measures of Central Tendency

A measure of central tendency is a single value that attempts to describe a set of data by identifying the central position within that set of data.

Mean (Arithmetic)

The mean is equal to the sum of all the values in the data set divided by the number of values in the data set.

Median

The middle score for a set of data that has been arranged in order of magnitude.

Mode

The most frequent score in our data set.

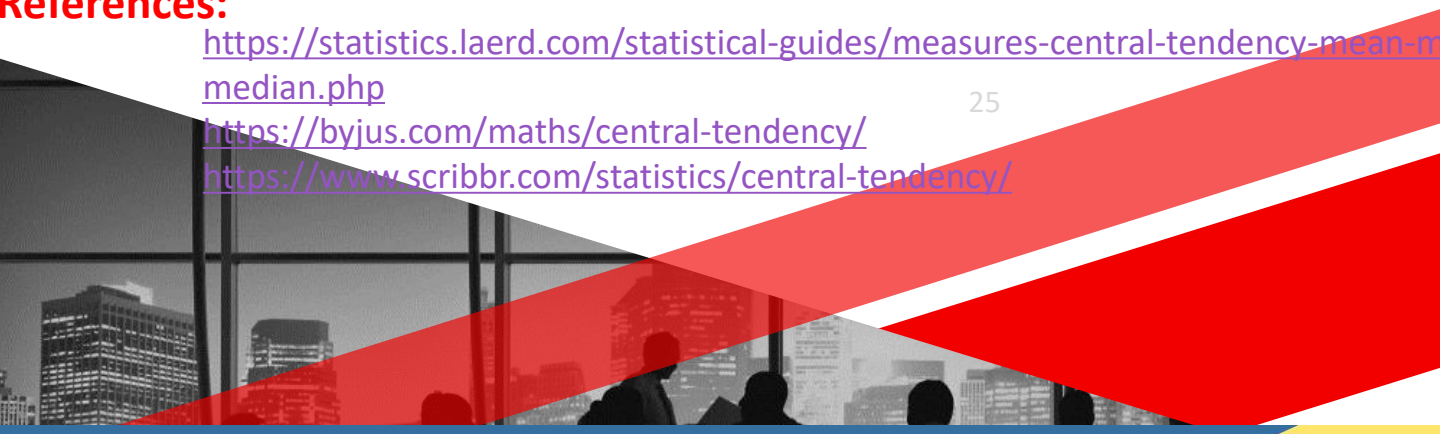
Example:

65 55 89 56 35 14 56 55 87 45 92

Mean:59 , Median: 56 , Mode: 56 , 55

References:

https://statistics.laerd.com/statistical-guides/measures-central-tendency-mean-m
median.php
https://byjus.com/maths/central-tendency/
https://www.scribbr.com/statistics/central-tendency/



Measures of Position

Measures of position give a range where a certain percentage of the data fall. The measures we consider here are percentiles and quartiles.

Variance

The average squared deviation from the mean of the given data set

$$\sum_1^n \frac{(X_i - \bar{X})^2}{n}$$

Standard Deviation

The square root of the variance gives the ”standard deviation”

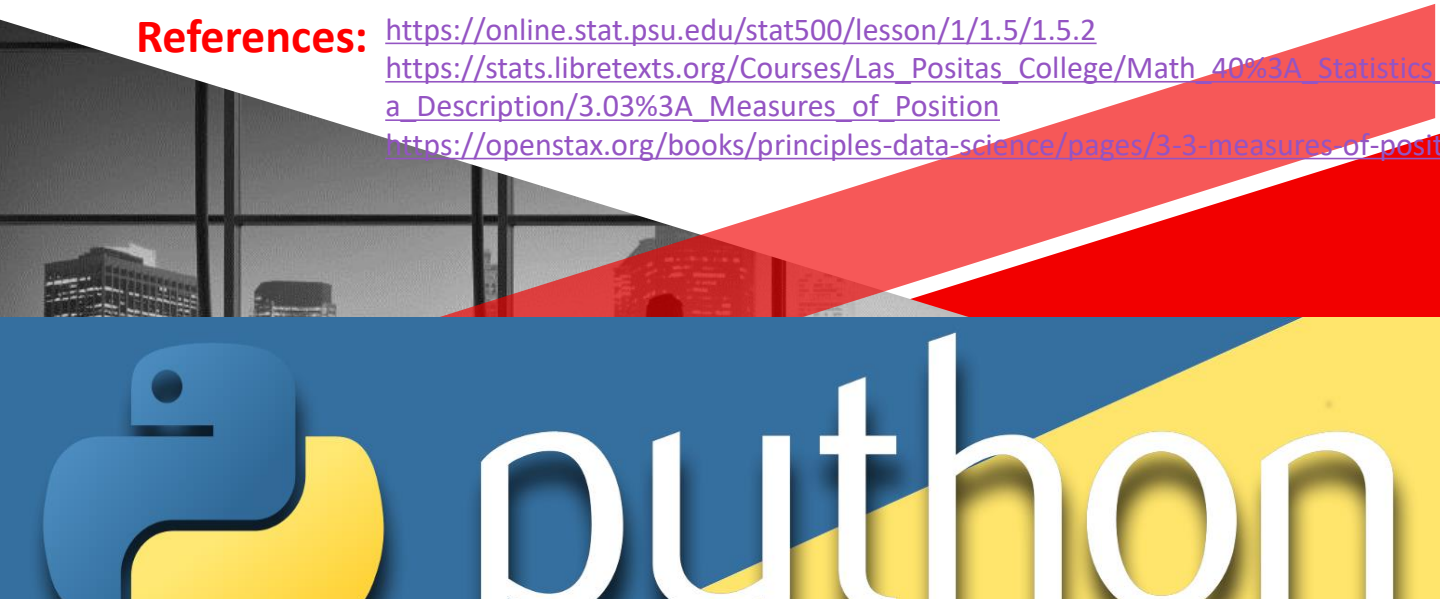
$$S.D. = \sqrt{Variance} = \sigma$$

Coefficient of Variation

The ratio of the standard deviation to the mean of the data set

$$(S.D. / Mean) * 100$$

References: https://online.stat.psu.edu/stat500/lesson/1/1.5/1.5.2
https://stats.libretexts.org/Courses/Las_Positas_College/Math_40%3A_Statistics_a_Description/3.03%3A_Measures_of_Position
https://openstax.org/books/principles-data-science/pages/3-3-measures-of-posit



Measures of Dispersion

Measures of dispersion are non-negative real numbers that help to gauge the spread of data about a central value.

Quartiles

Quartiles are numbers that separate the data into quarters.

first quartile is at position n+14, second quartile (i.e. the median) is at position 2(n+1)4, and the third quartile is at position 3(n+1)4.

Percentiles

Percentiles provide a way to assess and compare the distribution of values and the position of a specific data point in relation to the entire dataset by indicating the percentage of data points that fall below it.

$$\text{Percentile} = \frac{\text{number of data values below the measurement}}{\text{total number of data values}} \times 100\% = \frac{n}{N} \times 100\%$$

z-score

The z-score is a measure of the position of an entry in a dataset that makes use of the mean and standard deviation of the data.

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

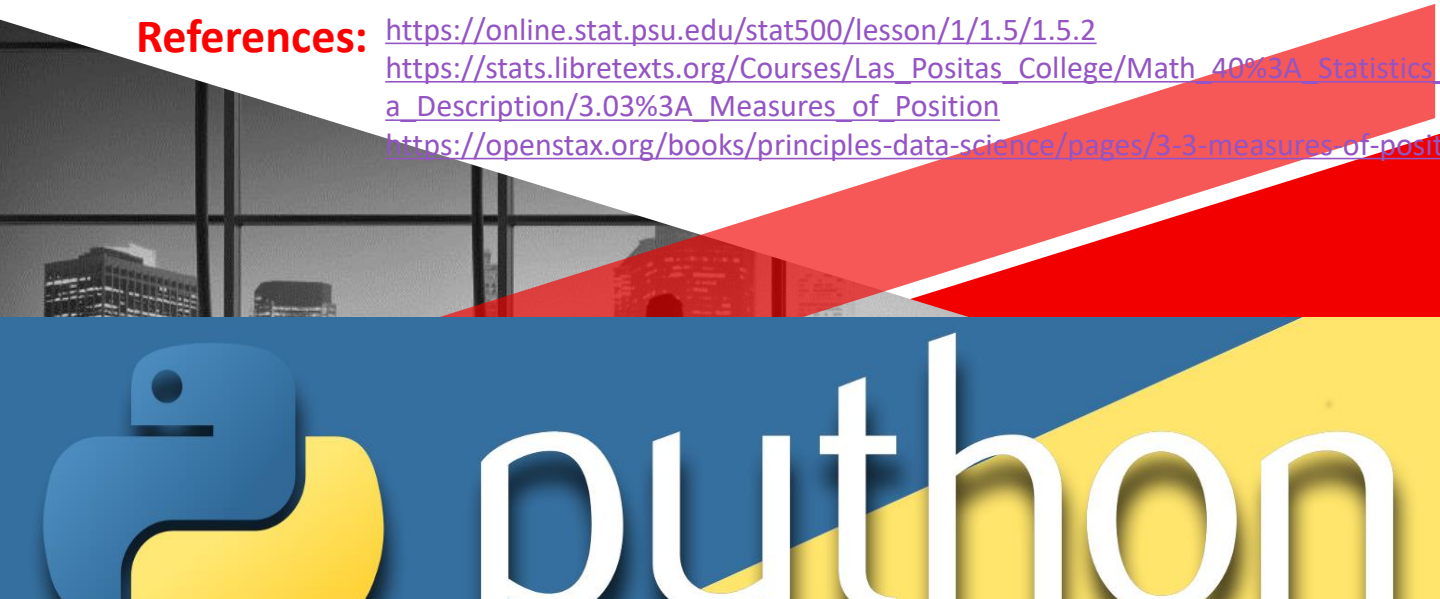
Where:

x is the measurement

μ is the mean

σ is the standard deviation

References: https://online.stat.psu.edu/stat500/lesson/1/1.5/1.5.2
https://stats.libretexts.org/Courses/Las_Positas_College/Math_40%3A_Statistics_a_Description/3.03%3A_Measures_of_Position
https://openstax.org/books/principles-data-science/pages/3-3-measures-of-posit



Measures of Position

Measures of position give a range where a certain percentage of the data fall. The measures we consider here are percentiles and quartiles.

Variance

The average squared deviation from the mean of the given data set

$$\sum_1^n \frac{(X_i - \bar{X})^2}{n}$$

Standard Deviation

The square root of the variance gives the ”standard deviation”

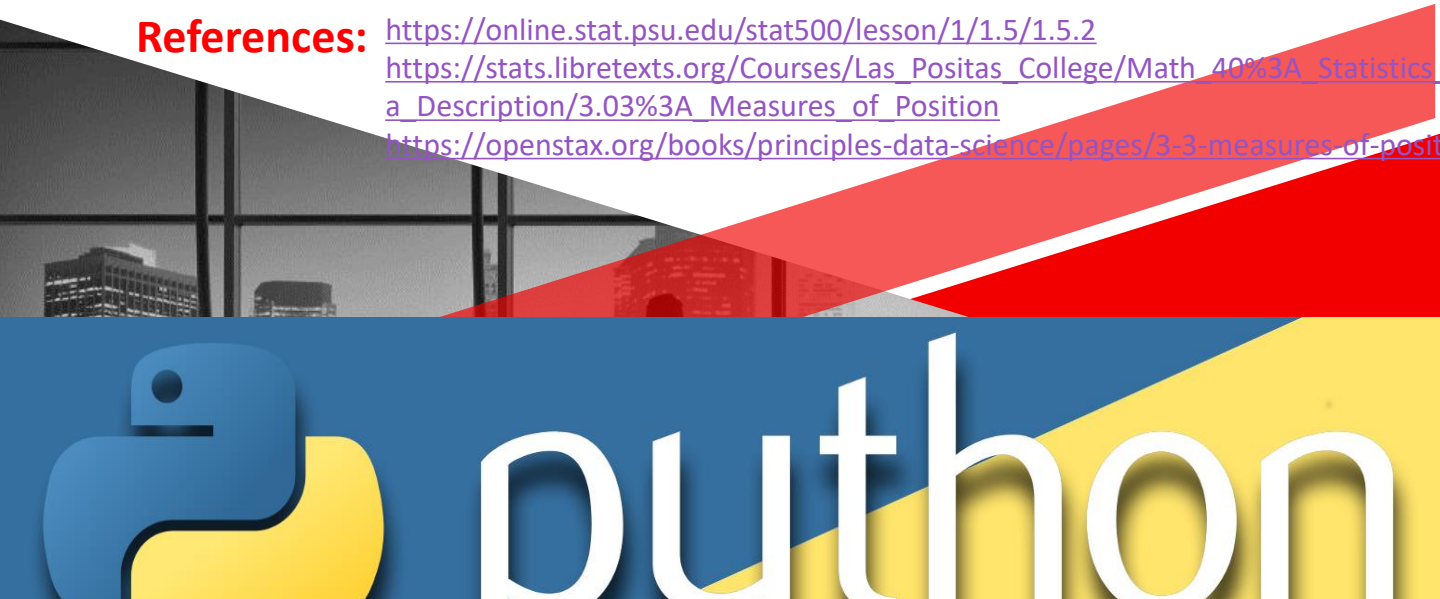
$$S.D. = \sqrt{Variance} = \sigma$$

Coefficient of Variation

The ratio of the standard deviation to the mean of the data set

$$(S.D. / Mean) * 100$$

References: https://online.stat.psu.edu/stat500/lesson/1/1.5/1.5.2
https://stats.libretexts.org/Courses/Las_Positas_College/Math_40%3A_Statistics_a_Description/3.03%3A_Measures_of_Position
https://openstax.org/books/principles-data-science/pages/3-3-measures-of-posit



Thank you - for listening and participating

- ❑**Questions / Queries**
- ❑**Suggestions/Recommendation**
- ❑**Ideas…..?**

Shahzad Sarwar
Cognitive Convergence
https://cognitiveconvergence.com
shahzad@cognitiveconvergence.com
voice: +1 4242530744 (USA) +92-3004762901 (Pak)