20-07-2025

Agenda: Statistics - II

Recap - Statistics - I

- Statistics (Definition, why, when)
- Measure of central tendency
  - Mean
  - Median
  - Mode + outliers
- Measure of Dispersion
  - variance → standard deviation
- keywords of statistics
  - population, sample, variable, parameter, etc...

Types of Data:

→ int, string, float, bool, list, set, dictionary, datetime, series etc. (python + pandas + ...)

merge to create categories → numeric, bodean, string, datetime

Data types of statistics:

- Continuous (float, numeric, range)

Data that can take on any values in an interval.

- Discrete (integer, count)

Data that can take on only integers values, such as counts.

- categorical (classification problems, groups) ← Nominal

- Binary (logical, boolean, indicter) →

Special case of categorical data with just two categories 0/1, false/true

A
B
A B
O

- Ordinal

Categorical data that has an explicit ordering.

ex- small < medium < large

- A > B > C > D

gender →

- Male
- Female
- other

- Example of Continuous data - Height of a person + Weight of an object + Temperature + Distance traveled + Time taken to complete a task

- Example of Discreet data - Number of website visitors per day+ Number of customers served at a restaurant + Number of errors in a software program + Word count in a document + Number of items sold in a store

- Example of Ordinal data - Education level (high school, bachelor's degree, master's degree) + Movie rating (1 star, 2 stars, 3 stars, 4 stars, 5 stars) Shirt size (small, medium, large, extra large) + Customer service rating (poor, fair, good, excellent) + Level of agreement (strongly disagree, disagree, neutral, agree, strongly agree)

- Example of Nominal data - Hair color (blonde, brown, black, red) + Product type (laptop, smartphone, tablet) + Survey responses (yes/no, agree/disagree) + Country of origin (US, China, France)

Measurement of central tendency

- Mean
- Median
- Mode

Measurement of Dispersion (spread)

- variance
- standard deviation
- IQR (Interquartile Range)

(1) variance

If population:

$$\text{population variance} = \frac{\sum (x_i - \bar{x})^2}{n}$$

$\bar{x} \rightarrow$ mean

if sample:

$$\text{sample variance} = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

population variance $\rightarrow \sigma^2$

sample variance $\rightarrow s^2$

population data : [2, 4, 6, 8]

→ mean = 5

→ $(2-5)^2 + (4-5)^2 + (6-5)^2 + (8-5)^2 = 20$

→ $\frac{20}{n} = \frac{20}{4} = 5$

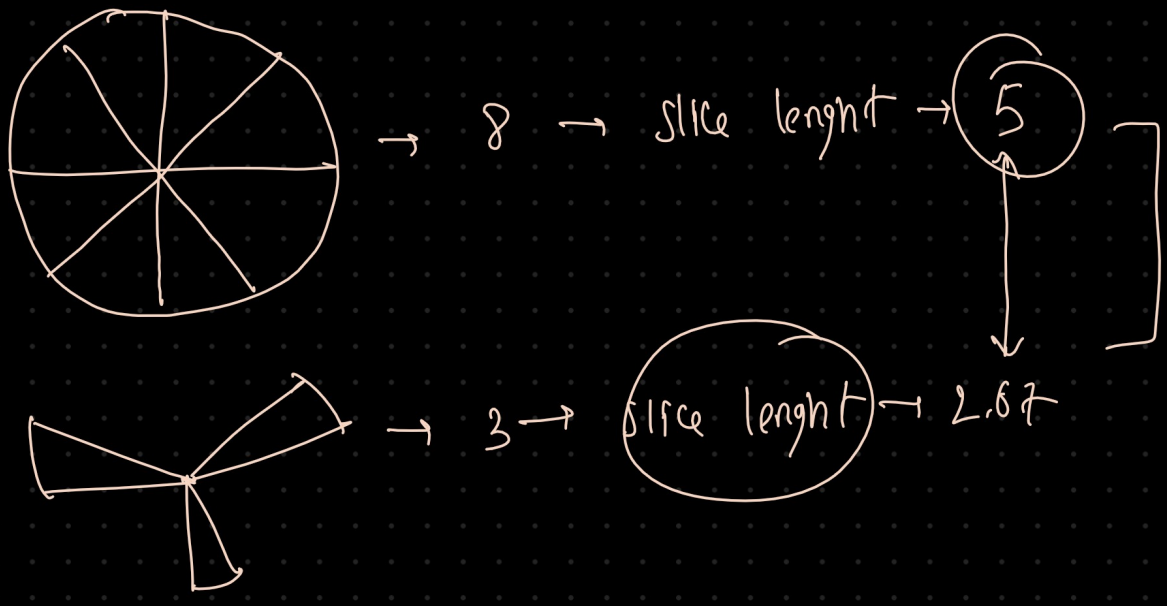
sample data = [2, 4, 6]

→ mean = 4

→ $(2-4)^2 + (4-4)^2 + (6-4)^2 = 8$

→ $\frac{8}{2} = 4$

$2 \rightarrow (n-1)$



ML model

A → 2024 - 2025 → ML → working
SD → 300

B → 2025 - July 2025 → ML → difference

IQR (InterQuartile Range)

IQR = Q3 − Q1

→ Jan, Feb, Mar, April, May, June, July, Aug, Sept, Oct, Nov, Dec

Q1 | Q2 | Q3 | Q4

0 – 25% . 26 – 50% . 51 – 75% . 75 – 100% .

25th percentile

Score = [60, 70, 75, 80, 85, 90, 95] →

Q1 → position

= 0.25 × [n+1] = → 70 →

→ [px (n+1) = 2 (exclusive)]

→ → [px (n-1) +1 (inclusive)] → 72.5 →

[60, 70, 75, 80, 85, 90, 95]

$p = 0.25 \times (n-1) + 1 \rightarrow 2.5$

2 & 3