AI - Chapter 3.4

Simple Statistical Concepts

Instructor: Saptarshi Jana

Learning Objectives

After studying this chapter, students will be able to:[a]

- Understand the importance of context in mathematical operations
- Learn about different measures of Central Tendency in statistics
- Learn about different measures of dispersion in statistics

[a] Unit 3: Mathematics in AI

January 2, 2026

1 Statistical Concepts

1.1 Introduction to Statistics

Statistics provides the mathematical framework for collecting, analyzing, interpreting, and presenting data. In the context of artificial intelligence and data science, statistical methods are indispensable for extracting meaningful insights from raw information and making data-driven decisions.

Statistics encompasses:

- Data collection methodologies
- Classification and organization techniques
- Presentation and visualization methods
- Analysis and interpretation procedures
- Quantitative reasoning and inference

1.2 Central Tendency

Central tendency refers to statistical measures that identify a single representative value describing the "center" of a dataset. These measures help summarize large amounts of data with a single meaningful number.

The three primary measures of central tendency are:

1. Mean (arithmetic average)
2. Median (middle value)
3. Mode (most frequent value)

1.2.1 Mean

The **mean** represents the arithmetic average of all values in a dataset, calculated by summing all observations and dividing by the total count.

Formula for ungrouped data:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n}$$

Where:

- $\bar{x}$ = sample mean
- x_i = individual values
- n = number of observations

Example 1: Calculate the mean height of students: 160, 165, 170, 175, 180 cm

$$\begin{aligned}\bar{x} &= \frac{160+165+170+175+180}{5} \\ &= \frac{850}{5} = 170 \text{ cm}\end{aligned}$$

Formula for grouped data:

$$\bar{x} = \frac{\sum f_i x_i}{\sum f_i} = \frac{\sum fm}{N}$$

Where:

- f_i = frequency of each class
- x_i or m = midpoint of each class
- N = total frequency

Example 2: Grouped Data
Student exam scores:

Score Range	Frequency	Midpoint	f × m
40-50	5	45	225
50-60	12	55	660
60-70	18	65	1170
70-80	10	75	750
80-90	5	85	425
Total	50		3230

$$\bar{x} = \frac{3230}{50} = 64.6$$

Properties of Mean:

- Sensitive to extreme values (outliers)
- Can be positive, negative, or zero
- Uses all data points in calculation
- Best for normally distributed data

1.2.2 Median

The **median** is the middle value in an ordered dataset, dividing it into two equal halves. Half the observations fall below the median and half above it.

For ungrouped data:
Case 1: Odd number of observations (n is odd)

$$\text{Median} = x_{\frac{n+1}{2}}$$

Example: Dataset: 23, 29, 35, 41, 47, 53, 61

Number of values: $n = 7$ (odd)
Position: $\frac{7+1}{2} = 4$th value
Median = 41
Case 2: Even number of observations (n is even)

$$\text{Median} = \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}$$

Example: Dataset: 15, 20, 25, 30, 35, 40
Number of values: $n = 6$ (even)
Median = $\frac{25+30}{2} = \frac{55}{2} = 27.5$
For grouped data:

$$\text{Median} = L + \left(\frac{\frac{N}{2} - CF}{f}\right) \times h$$

Where:

- L = lower boundary of median class
- N = total frequency
- CF = cumulative frequency before median class
- f = frequency of median class
- h = class width

Example: Grouped Data

Class	**Frequency**	**Cumulative Frequency**
10-20	6	6
20-30	14	20
30-40	22	42
40-50	16	58
50-60	8	66
Total	66	

$\frac{N}{2} = \frac{66}{2} = 33$
Median class is 30-40 (CF = 42 is first to exceed 33)

$$\text{Median} = 30 + \left(\frac{33 - 20}{22}\right) \times 10 = 30 + 5.91 = 35.91$$

Advantages of Median:

- Not affected by extreme values
- Suitable for skewed distributions
- Easy to understand and calculate
- Appropriate for ordinal data

1.2.3 Mode

The **mode** is the value that appears most frequently in a dataset. A distribution may have one mode (unimodal), two modes (bimodal), multiple modes (multimodal), or no mode at all.

For ungrouped data:

Simply identify the most frequent value(s).

Example 1: Unimodal Dataset: 7, 9, 12, 12, 12, 15, 18, 21

Mode = 12 (appears 3 times)

Example 2: Bimodal Dataset: 3, 5, 5, 7, 9, 9, 11

Modes = 5 and 9 (both appear twice)

Example 3: No mode Dataset: 2, 4, 6, 8, 10, 12

No mode (all values appear once)

For grouped data:

The modal class is identified as having the highest frequency. The mode can be calculated using:

$$\text{Mode} = L + \left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2}\right) \times h$$

Where:

- L = lower boundary of modal class
- f_1 = frequency of modal class
- f_0 = frequency of class before modal class
- f_2 = frequency of class after modal class
- h = class width

Example:

Temperature Range (°C)	Frequency
15-20	4
20-25	8
25-30	15
30-35	11
35-40	3

Modal class: 25-30 (highest frequency = 15)

$$\text{Mode} = 25 + \left(\frac{15 - 8}{2(15) - 8 - 11}\right) \times 5 = 25 + \left(\frac{7}{11}\right) \times 5 = 25 + 3.18 = 28.18$$

When to Use Each Measure

- **Mean:** Best for symmetric distributions without outliers; provides precise average
- **Median:** Ideal for skewed distributions or when outliers are present; robust to extreme values
- **Mode:** Useful for categorical data; identifies most common value; can reveal distribution patterns

2 Measures of Dispersion

2.1 Understanding Dispersion

While central tendency measures describe the center of a dataset, **dispersion** measures indicate how spread out or scattered the data points are around the central value. Dispersion helps us understand the variability and consistency within data.

Dispersion quantifies the extent to which data values differ from the average. High dispersion indicates data points are widely scattered; low dispersion suggests data points cluster closely around the center.

2.2 Variance

Variance is a fundamental measure quantifying how far individual data points deviate from the mean. It represents the average of squared deviations from the mean.

Formula for population variance:

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{N}$$

Formula for sample variance:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

Where:

- σ^2 = population variance
- s^2 = sample variance
- x_i = individual data points
- μ = population mean
- $\bar{x}$ = sample mean
- N = population size
- n = sample size

Step-by-step calculation:

Example: Calculate variance for: 8, 12, 16, 20, 24

Step 1: Calculate mean

$$\bar{x} = \frac{8 + 12 + 16 + 20 + 24}{5} = \frac{80}{5} = 16$$

Step 2: Calculate deviations from mean

$$\begin{aligned} (8 - 16) &= -8 \\ (12 - 16) &= -4 \\ (16 - 16) &= 0 \\ (20 - 16) &= 4 \\ (24 - 16) &= 8 \end{aligned}$$

Step 3: Square the deviations

$$\begin{aligned} (-8)^2 &= 64 \\ (-4)^2 &= 16 \\ (0)^2 &= 0 \\ (4)^2 &= 16 \\ (8)^2 &= 64 \end{aligned}$$

Step 4: Sum squared deviations and divide

$$\text{Sum} = 64 + 16 + 0 + 16 + 64 = 160$$

$$s^2 = \frac{160}{5} = 32$$

Properties of Variance:

- Always non-negative ($\sigma^2 \geq 0$)
- Variance of zero indicates all values are identical
- Units are squared (e.g., if data is in cm, variance is in cm$^{\mathbf{2}}$)
- Sensitive to outliers due to squaring

2.3 Standard Deviation

Standard deviation is the square root of variance, providing a dispersion measure in the same units as the original data. It's more interpretable than variance for practical applications.

Formula:

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{N}}$$

For the previous example:

$$s = \sqrt{32} \approx 5.66$$

Example: Complete Calculation

Dataset: 45, 50, 55, 60, 65, 70, 75

Step 1: Mean

$$\bar{x} = \frac{45 + 50 + 55 + 60 + 65 + 70 + 75}{7} = \frac{420}{7} = 60$$

Step 2: Deviations and squares

Value	**Deviation**	**Squared**
45	-15	225
50	-10	100
55	-5	25
60	0	0
65	5	25
70	10	100
75	15	225
Total		700

Step 3: Variance

$$s^2 = \frac{700}{7} = 100$$

Step 4: Standard deviation

$$s = \sqrt{100} = 10$$

2.4 Interpretation and Applications

Understanding Standard Deviation

Low standard deviation:

- Data points cluster tightly around the mean
- High consistency and predictability
- Example: Manufacturing quality control with tight tolerances

High standard deviation:

- Data points spread widely from the mean
- High variability and diversity
- Example: Income distribution in a population

2.5 Applications in AI and Machine Learning

Feature Scaling: Standard deviation is crucial for normalizing features in machine learning:

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

This z-score standardization ensures features have mean 0 and standard deviation 1.

Outlier Detection: Values beyond $\mu \pm 3\sigma$ are often considered outliers (assuming normal distribution).

Model Evaluation: Standard deviation of errors helps assess model reliability and prediction consistency.

Error Analysis: Measuring variability in model predictions across different test sets indicates robustness.

Key Concepts Summary

Chapter Recap

Statistical Concepts:

- Central tendency (mean, median, mode) summarizes data with representative values
- Mean: arithmetic average, sensitive to outliers
- Median: middle value, robust to extreme values
- Mode: most frequent value, useful for categorical data
- Variance and standard deviation measure data spread around the center
- These concepts are fundamental for data analysis and AI applications

3 Practice Exercises

1. Calculate the mean, median, and mode for: 12, 15, 15, 18, 21, 21, 21, 24, 27, 30
2. What is the relationship between variance and standard deviation? Why is standard deviation often preferred?
3. Explain when you would use median instead of mean as a measure of central tendency.
4. Calculate the variance and standard deviation for: 5, 10, 15, 20, 25
5. How do outliers affect mean and median differently? Which is more robust?