PHOENICIA UNIVERSITY

College of Arts and Sciences

Numerical Descriptive Statistics

Chapter 3 — Measuring the Center, Spread, and Position of Data

Stat 201 – Statistics

1 Lesson Roadmap

Chapter 2 summarized data visually. This chapter summarizes data numerically — using single values that describe where data is centered, how spread out it is, and where individual values sit relative to the rest.

1

Central tendency

Mean, median, and mode — where data is centered

2

Dispersion

Range, variance, standard deviation, and z-scores

3

Grouped data

Estimating the mean and variance from a frequency table

4

Measures of position

Quartiles and the interquartile range (IQR)

5

Boxplots

Visualizing the five-number summary and outliers

6

Shape of a distribution

How mean, median, and mode relate under skew

2 Parameters vs. Statistics

Every numerical measure we study — mean, median, variance, quartiles — can be computed at two levels, and the vocabulary changes depending on which one we mean.

Parameter

A numerical measure calculated from the entire population. Its value is fixed — there is only one true population mean, one true population variance, and so on.

Statistic

The same kind of measure calculated from a sample. Its value varies from sample to sample (sample-to-sample variability), but a well-chosen sample statistic estimates the population parameter.

This chapter develops formulas for both ungrouped data (a plain list of values) and grouped data (values organized into a frequency table).

3 Central Tendency: The Mean

The mean (arithmetic average) is the sum of all values divided by the number of values. It is the most widely used measure of central tendency.

$$\mu = \Sigma x / N$$

Population mean

$$\bar{x} = \Sigma x / n$$

Sample mean

Example

Seven students' exam scores: 72, 85, 90, 68, 77, 95, 82.

$$\bar{x} = (72+85+90+68+77+95+82) / 7 = 569 / 7 \approx 81.29$$

The average score for this group of students is about 81.29.

4 The Mean Is Sensitive to Outliers

A major shortcoming of the mean is that it can be pulled strongly toward extreme values. When a data set contains an outlier, the mean may no longer represent a "typical" value well.

Example — Monthly salaries (in $) at a small company

Without the CEO's salary

3200, 3400, 3100, 3300, 3600

Mean = 16600 / 5 = $3,320

With the CEO's salary added

3200, 3400, 3100, 3300, 3600, 52000

Mean = 68600 / 6 = $11,433.33

One outlier more than tripled the mean, even though five of the six salaries did not change. When outliers are present, the median or a trimmed mean often gives a more representative measure of the center.

5 Central Tendency: The Median

The median is the middle value of a data set once it has been ranked in increasing order. It is not affected by extreme values — a key advantage over the mean.

Odd number of values

Data: 4, 8, 15, 16, 23

The middle value is the median:

Median = 15

Even number of values

Data: 4, 8, 15, 16, 23, 42

Average the two middle values:

Median = (15+16)/2 = 15.5

Trade-off: the median ignores the actual size of most values — only their rank matters — so it is a less "information-rich" measure than the mean when there are no outliers.

6 Central Tendency: The Mode

The mode is the value that occurs most frequently in a data set. Unlike the mean or median, a data set can have no mode, one mode, or several modes.

No mode

All values are distinct — nothing repeats.

Unimodal

Exactly one value repeats most often.

Bimodal

Two values are tied for the highest frequency.

Multimodal

Three or more values are tied for the highest frequency.

Example

Data set: 2, 3, 3, 5, 7, 7, 9 → both 3 and 7 occur twice, more than any other value → the data set is bimodal, with modes 3 and 7.

7 Mean, Median, and Mode Under Skew

The relative position of the mean, median, and mode reveals the shape of a distribution.

Symmetric

Mean = Median = Mode

All three coincide at the center of the distribution.

Skewed Right

Mode < Median < Mean

A long right tail pulls the mean above the median and mode.

Skewed Left

Mean < Median < Mode

A long left tail pulls the mean below the median and mode.

Rule of thumb: the mean is always pulled in the direction of the longer tail, because it is sensitive to every value — including the extreme ones.

8 Dispersion: The Range

Two data sets can share the same mean yet look very different — one tightly clustered, one widely spread. Measures of dispersion (also called spread or variability) capture this difference.

The range is the simplest measure: the difference between the largest and smallest values.

$$R = \text{Maximum} - \text{Minimum}$$

Example

Daily high temperatures (°C) over one week: 12, 15, 9, 22, 18, 7, 25

$$R = 25 - 7 = 18°\text{C}$$

Limitation: the range depends only on the two extreme values, ignoring everything in between — so, like the mean, it is easily distorted by a single outlier.

9 Variance and Standard Deviation

Variance and standard deviation improve on the range because they use every value in the data set, not just the extremes.

$$\sigma^2 = \Sigma(x-\mu)^2 / N$$

Population variance

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

Sample variance

Three steps: (1) subtract the mean from each value, (2) square and sum the deviations, (3) divide by N or n−1. Standard deviation is the square root of variance.

Example — Points scored by 5 players on a team

Data: 4, 8, 6, 5, 3 → mean $\bar{x} = 5.2$

x	4	8	6	5	3
$(x-\bar{x})^2$	1.44	7.84	0.64	0.04	4.84

$$s^2 = 14.8 / 4 = 3.7 \qquad s = \sqrt{3.7} \approx 1.92$$

10 Standardized Z-Scores

A z-score measures how many standard deviations a value lies from the mean — useful for comparing values from different distributions or scales.

$$z = (x - \mu) / \sigma$$

A positive z-score means the value lies above the mean; a negative z-score means it lies below the mean.

Example

A class has mean exam score $\mu = 75$ and standard deviation $\sigma = 8$. A student scores $x = 88$.

$$z = (88 - 75) / 8 = 1.625$$

This student's score is 1.625 standard deviations above the class mean — a strong result relative to the rest of the class.

11 Mean and Variance for Grouped Data

When data is organized into a frequency table (grouped into class intervals), individual values are no longer known — only how many observations fall in each class. We approximate each class by its midpoint, m.

$$\bar{x} = \Sigma(f \cdot m) \,/\, n$$

Grouped mean (f = class frequency, m = class midpoint)

$$s^2 = \Sigma f(m - \bar{x})^2 \,/\, (n-1)$$

Grouped sample variance

Example — Commute times of 20 employees

Class (min)	0–10	10–20	20–30	30–40
Frequency f	3	8	6	3
Midpoint m	5	15	25	35

$$\bar{x} = [3(5)+8(15)+6(25)+3(35)] \,/\, 20 = 375 \,/\, 20 = 18.75 \text{ min}$$

12 Measures of Position: Quartiles & IQR

Quartiles divide a ranked data set into four equal parts. Q1 (25th percentile), Q2 (the median), and Q3 (75th percentile) are the three cut points.

$$IQR = Q3 - Q1$$

Interquartile range — a spread measure resistant to outliers

Example

Ranked data (n = 11): 2, 5, 7, 8, 10, 12, 15, 18, 20, 21, 25

Q1 = 7

Median of the lower half
(2, 5, 7, 8, 10)

Q2 = 12

Median of the full data set

Q3 = 20

Median of the upper half
(15, 18, 20, 21, 25)

$$IQR = 20 - 7 = 13$$

13 Boxplots and the Five-Number Summary

A boxplot displays a data set using five values: the minimum, Q1, the median, Q3, and the maximum. Fences built from the IQR flag unusually extreme values as outliers.

Lower fence = Q1 − 1.5·IQR

Upper fence = Q3 + 1.5·IQR

Example — continuing the data set from the previous slide

Min = 2, Q1 = 7, Median = 12, Q3 = 20, Max = 25, IQR = 13

Lower fence = 7 − 19.5 = −12.5

Upper fence = 20 + 19.5 = 39.5

Since the minimum (2) and maximum (25) both fall inside these fences, there are no outliers in this data set — the whiskers extend to 2 and 25.

Key Takeaways

- The mean uses every value and is easy to compute, but it is highly sensitive to outliers.
- The median is resistant to outliers and is the better center measure for skewed data.
- The mode identifies the most frequent value and can apply to categorical data, but a data set may have zero, one, or several modes.
- Range, variance, and standard deviation measure spread; variance and standard deviation use every value and are the most informative.
- Z-scores standardize values so they can be compared across different means and scales.
- Quartiles, the IQR, and boxplots describe position and spread while staying resistant to outliers — and help flag outliers directly.