Lecture 03: Central Tendency and Variability

The earliest use of statistics?

In the Peloponnesian war in the 5th century BC, Athenians asked a number of soldiers to count the number of bricks on the wall of Platea and used the most frequent value (the mode) to estimate the height of the wall.

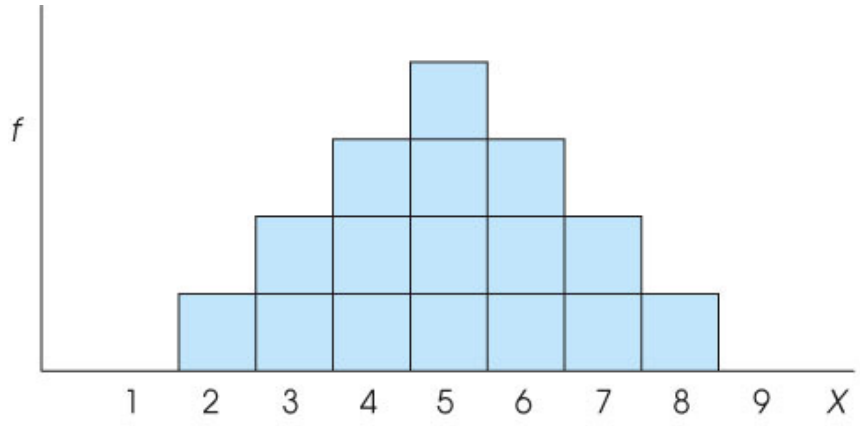
Central Tendency

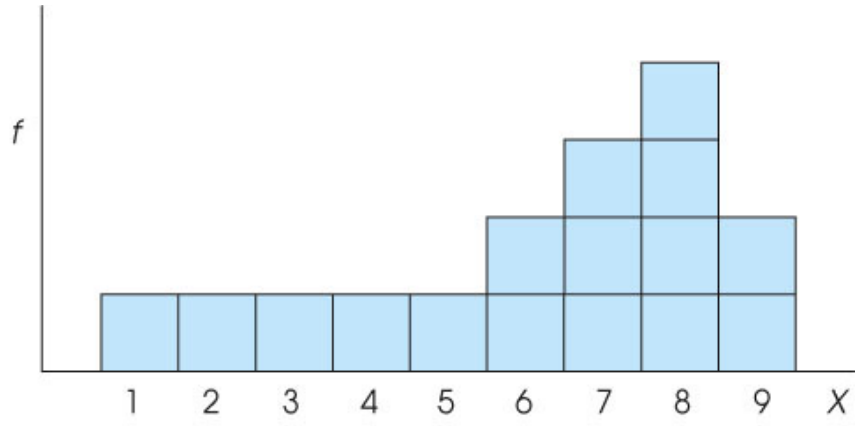
- In general terms, **central tendency** is a statistical measure that determines a single value that accurately describes the center of the distribution and represents the entire distribution of scores.
- The goal of central tendency is to identify **the single value** that is the best representative for the entire set of data.

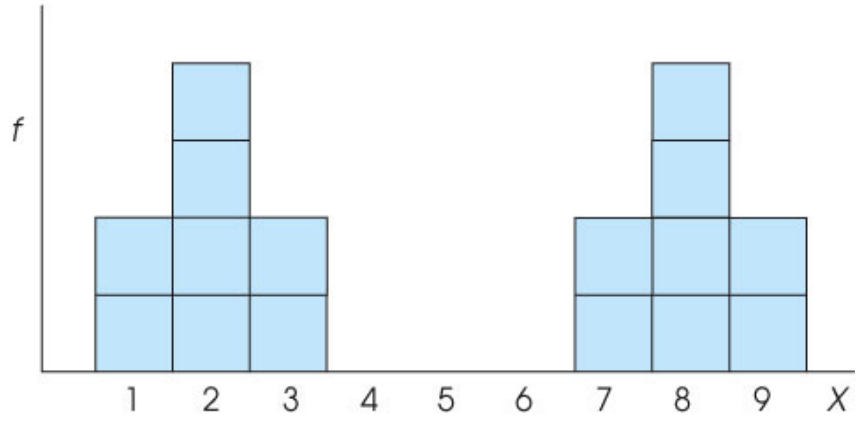
Central Tendency (cont.)

- By identifying the “average score”, central tendency allows researchers to **summarize** or condense a large set of data into a single value.
- Thus, central tendency serves as a descriptive statistic because it allows researchers to describe or present a set of data in a very simplified, concise form.
- In addition, it is possible to **compare** two (or more) sets of data by simply comparing the average score (central tendency) for one set versus the average score for another set.

When is good for using the mean?







The Mean, the Median, and the Mode

- It is essential that central tendency be determined by an objective and well-defined procedure so that others will understand exactly how the "average" value was obtained and can duplicate the process.
- No single procedure always produces a good, representative value. Therefore, researchers have developed *three* commonly used techniques for measuring central tendency: the mean, the median, and the mode.

The Mean

- The mean is the most commonly used measure of central tendency.
- Computation of the mean requires scores that are numerical values measured on an interval or ratio scale.
- The mean is obtained by computing the sum, or total, for the entire set of scores, then dividing this sum by the number of scores.

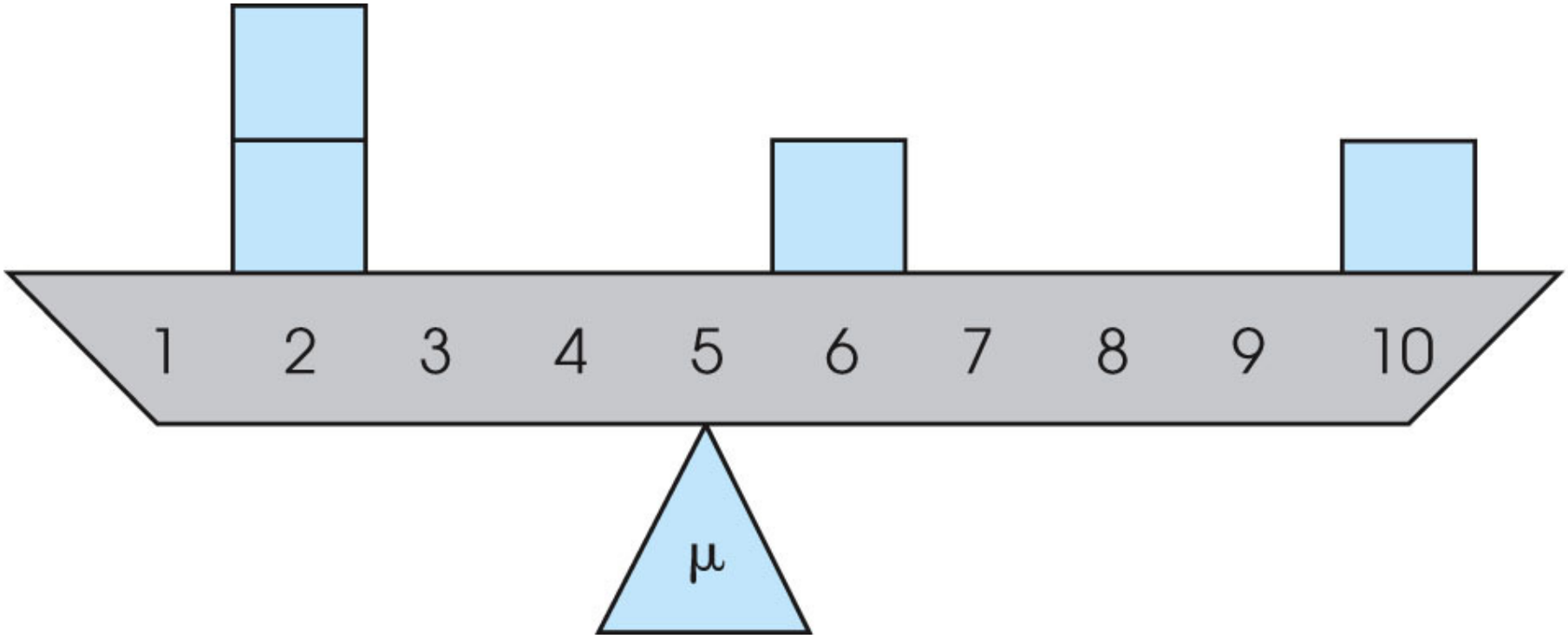
population mean $\mu = \frac{\Sigma X}{N}$

sample mean $= M = \frac{\Sigma X}{n}$

Alternative Definitions for the Mean

Conceptually, the mean can also be defined as:

1. The mean is the amount that each individual receives when the total (ΣX) is divided equally among all N individuals.
2. The mean is the balance point of the distribution because the sum of the distances below the mean is exactly equal to the sum of the distances above the mean.



The frequency distribution shown as a seesaw balanced at the mean.

The Weighted Mean

$$\text{overall mean} = M = \frac{\Sigma X \text{ (overall sum for the combined group)}}{n \text{ (total number in the combined group)}}$$

$$= \frac{\Sigma X_1 + \Sigma X_2}{n_1 + n_2}$$

Example

First Sample	Second Sample	Combined Sample
$n = 12$	$n = 8$	$n = 20\ (12 + 8)$
$\Sigma X = 72$	$\Sigma X = 56$	$\Sigma X = 128\ (72 + 56)$
$M = 6$	$M = 7$	$M = 6.4$

The Mean is Influenced by Every Score in the Distribution

- Because the calculation of the mean involves every score in the distribution, changing the value of any score will change the value of the mean.
- Modifying a distribution by discarding scores or by adding new scores will usually change the value of the mean.
- To determine how the mean will be affected for any specific situation you must consider: 1) how the number of scores is affected, and 2) how the sum of the scores is affected.

Properties of the Mean

- If a constant value is added to every score in a distribution, then the same constant value is added to the mean.
- If every score is multiplied by a constant value, then the mean is also multiplied by the same constant value.

Example of addition

Participant	No Alcohol	Moderate Alcohol
A	4	5
B	2	3
C	3	4
D	3	4
E	2	3
F	3	4
	$\Sigma X = 17$	$\Sigma X = 23$
	$M = 2.83$	$M = 3.83$

Example of multiplication

Original Measurement in Inches	Conversion to Centimeters (Multiply by 2.54)
10	25.40
9	22.86
12	30.48
8	20.32
11	27.94
$\Sigma X = 50$	$\Sigma X = 127.00$
$M = 10$	$M = 25.40$

When the Mean Won't Work

- Although the mean is **the most commonly used** measure of central tendency, there are situations where the mean does not provide a good, representative value, and there are situations where you cannot compute a mean at all.
- When a distribution contains a few **extreme** scores (or is very **skewed**), the mean will be pulled toward the extremes (displaced toward the tail). In this case, the mean will not provide a "central" value.
- With data from a **nominal** scale it is impossible to compute a mean, and when data are measured on an **ordinal** scale (ranks), it is usually inappropriate to compute a mean.

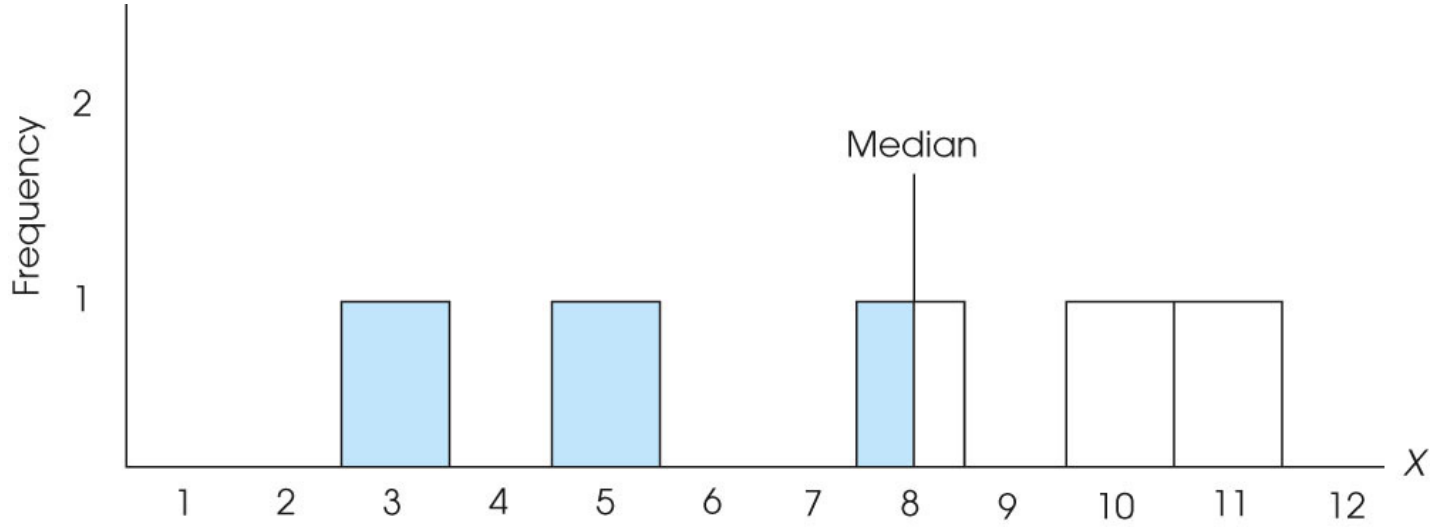
The Median

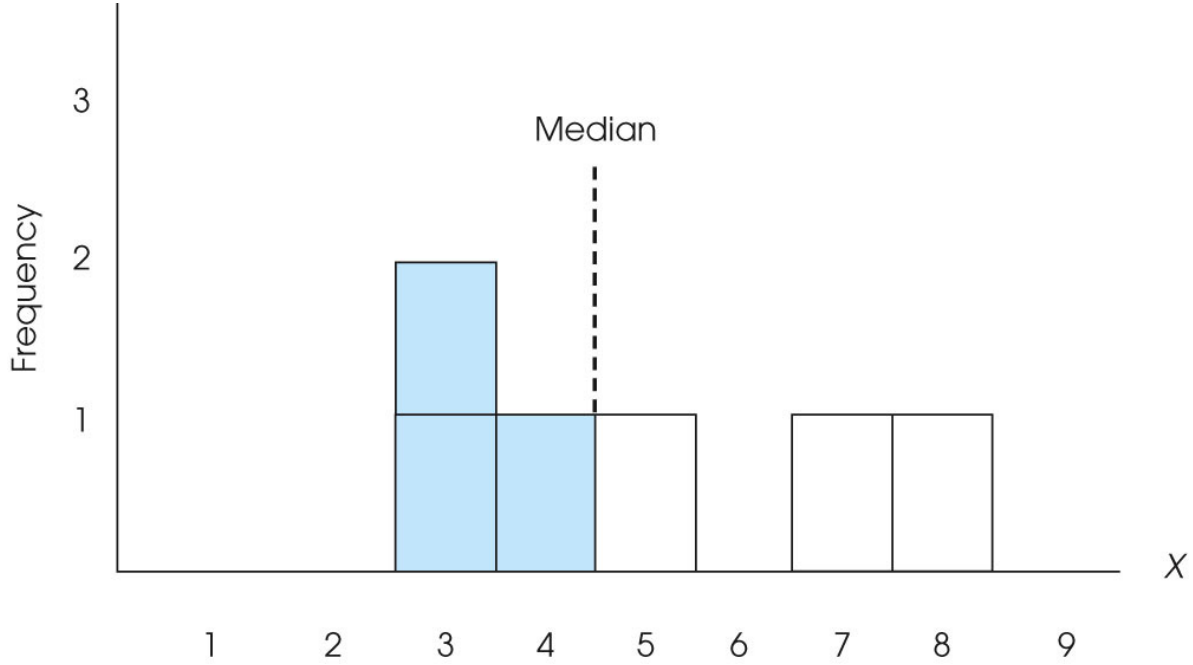
- If the scores in a distribution are listed in order from smallest to largest, the median is defined as the midpoint of the list.
- The median divides the scores so that 50% of the scores in the distribution have values that are **equal to or less than** the median.
- Computation of the median requires scores that can be placed in rank order (smallest to largest) and are measured on an **ordinal, interval, or ratio scale**.

Finding the Median

Usually, the median can be found by a simple counting procedure:

1. With an odd number of scores, list the values in order, and the median is the middle score in the list.
2. With an even number of scores, list the values in order, and the median is **half-way between the middle two scores**.

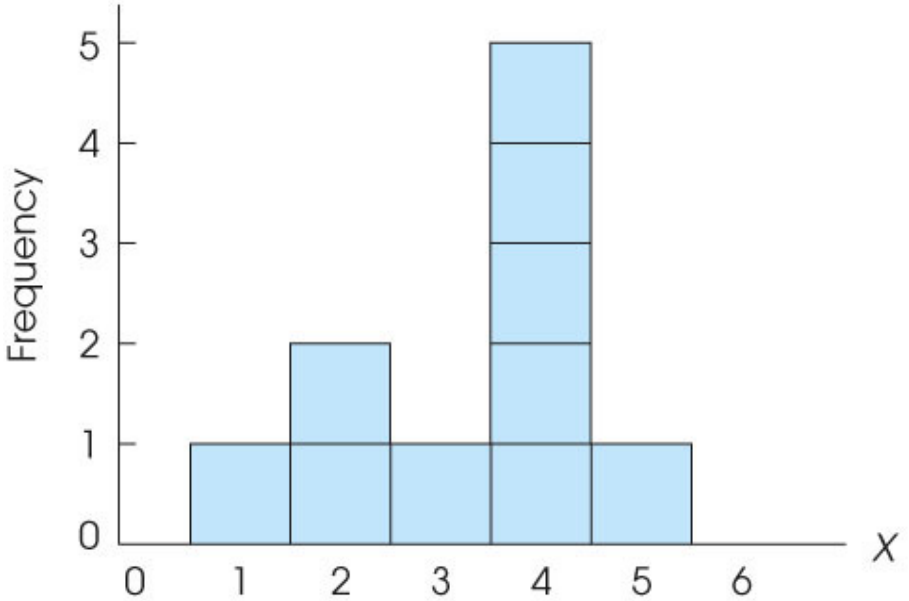

Frequency
2
1
Median
X
1
2
3
4
5
6
7
8
9
10
11
12

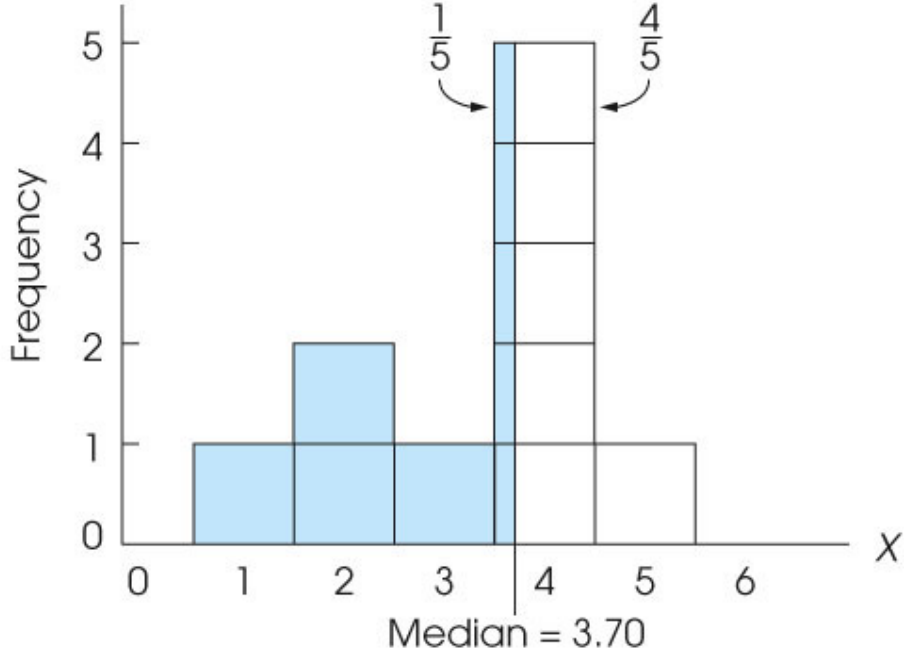

Frequency
3
2
1
Median
X
1
2
3
4
5
6
7
8
9

Finding the Precise Median for a Continuous Variable

- If the scores are measurements of a continuous variable, it is possible to find the median by **first placing the scores in a frequency distribution histogram** with each score represented by a box in the graph.
- Then, draw a vertical line through the distribution so that exactly half the boxes are on each side of the line. The median is defined by the location of the line.

$$\text{fraction} = \frac{\text{number needed to reach 50\%}}{\text{number in the interval}}$$





For grouped data. The median value is interpolated.

Properties of the Median

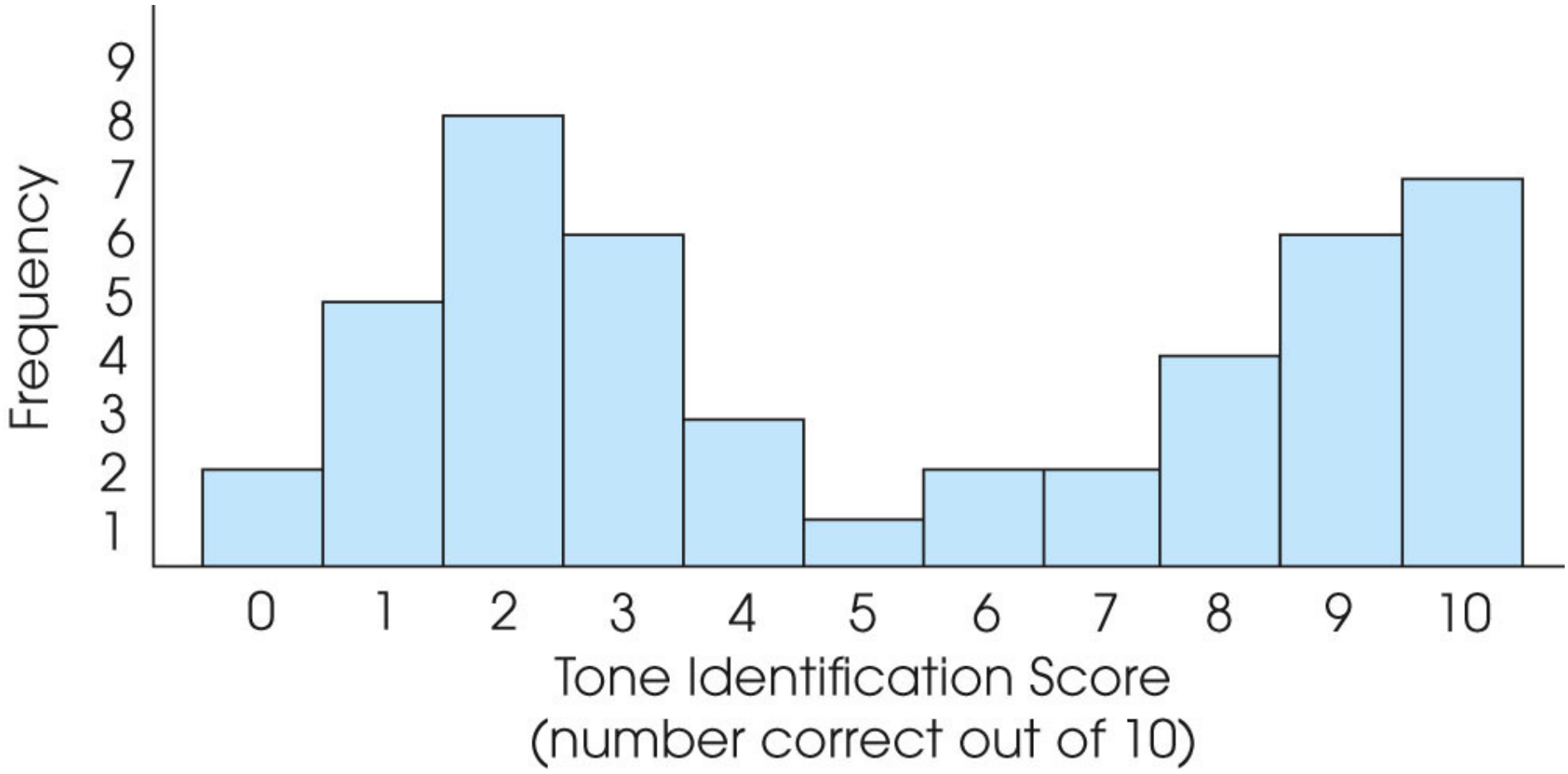
- One advantage of the median is that it is relatively unaffected by **extreme scores**.
- Thus, the median tends to stay in the "center" of the distribution even when there are a few extreme scores or when the distribution is very skewed. In these situations, the median serves as a good alternative to the mean.

The Mode

- The mode is defined as the most frequently occurring **category** or **score** in the distribution.
- In a frequency distribution graph, the mode is the category or score corresponding to the peak or high point of the distribution.
- The mode can be determined for data measured on any scale of measurement: **nominal, ordinal, interval, or ratio**.
- The primary value of the mode is that it is the only measure of central tendency that can be used for data measured on a **nominal scale.** In addition, the mode is often used as a supplemental measure of central tendency that is reported along with the mean or the median.

Bimodal Distributions

- It is possible for a distribution to have more than one mode. Such a distribution is called **bimodal**. (Note that a distribution can have only one mean and only one median.)
- In addition, the term "mode" is often used to describe a peak in a distribution that is not really the highest point. Thus, a distribution may have a ***major mode*** at the highest peak and a ***minor mode*** at a secondary peak in a different location.

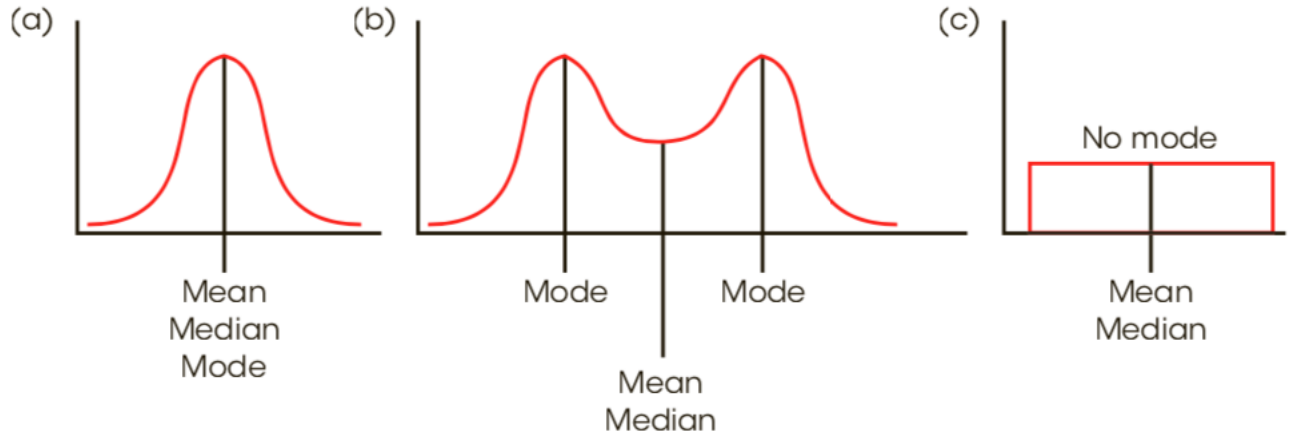

Frequency
9
8
7
6
5
4
3
2
1
0
1
2
3
4
5
6
7
8
9
10
Tone Identification Score
(number correct out of 10)

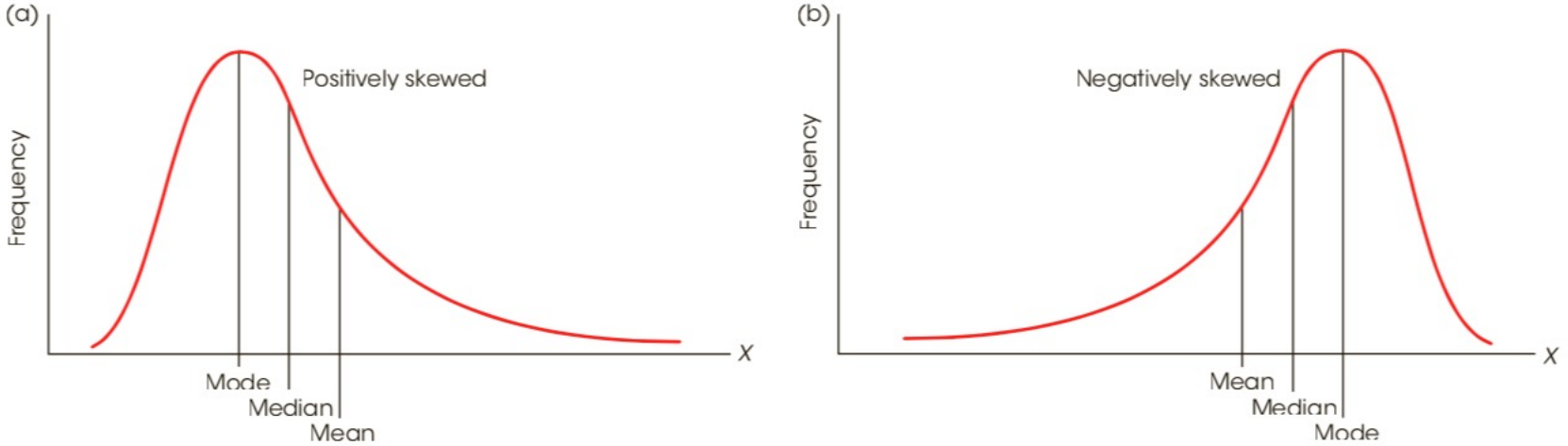
Central Tendency and the Shape of the Distribution

- Because the mean, the median, and the mode are all measuring central tendency, the three measures are often systematically related to each other.
- In a symmetrical distribution, for example, the mean and median will always be equal.

Central Tendency and the Shape of the Distribution (cont.)

- If a symmetrical distribution has only one mode, the mode, mean, and median will all have the same value.
- In a skewed distribution, the mode will be located at the peak on one side and the mean usually will be displaced toward the tail on the other side.
- The median is usually located between the mean and the mode.

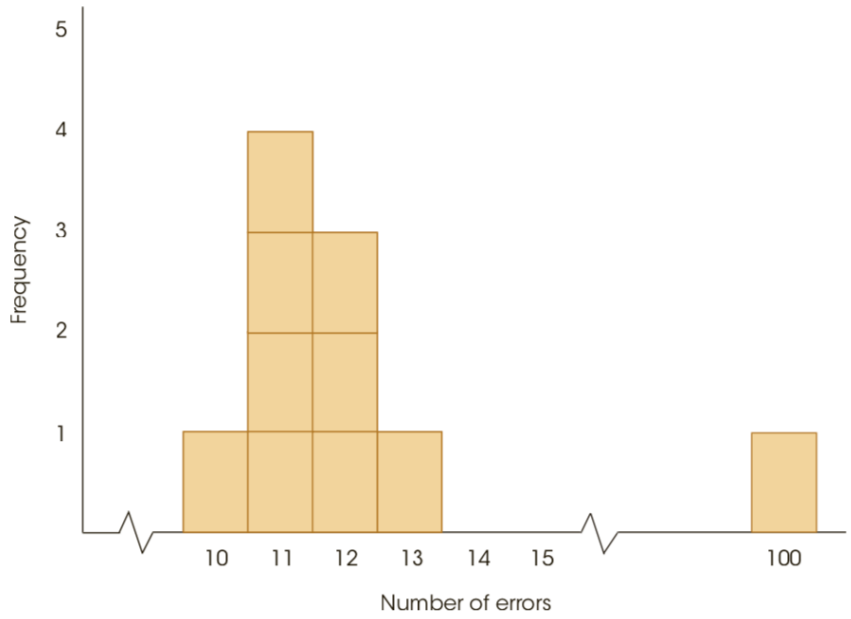

(a)
Mean
Median
Mode
(b)
Mode
Mode
Mean
Median
(c)
No mode
Mean
Median


(a)
Positively skewed
Frequency
X
Mode
Median
Mean
(b)
Negatively skewed
Frequency
X
Mean
Median
Mode

When to Use the Median

- Extreme Scores or Skewed Distributions
- Undetermined Values
- Open-Ended Distributions
- Ordinal Scale (no good for mean)

Person	Time (Min.)
1	8
2	11
3	12
4	13
5	17
6	Never finished



Number of Pizzas (X)	f
5 or more	3
4	2
3	2
2	3
1	6
0	4

When to Use the Mode

- **Nominal Scales**: only the mode is available for central tendency.
- **Discrete Variables**: the mode is always applicable; mean is applicable when the variable is numeric; median is applicable when the variable is ordered or numeric.
- **Describing Shape**

Reporting Central Tendency in Research Reports

- In manuscripts and in published research reports, the sample mean is identified with the letter M. However, there is no standardized notation for reporting the median or the mode.
- In research situations where several means are obtained for different groups or for different treatment conditions, it is common to present all of the means in a single graph.

Examples of reporting central tendency

In text

The treatment group showed fewer errors ($M = 2.56$) on the task than the control group ($M = 11.76$).

The median number of errors for the treatment group was 8.5, compared to a median of 13 for the control group.

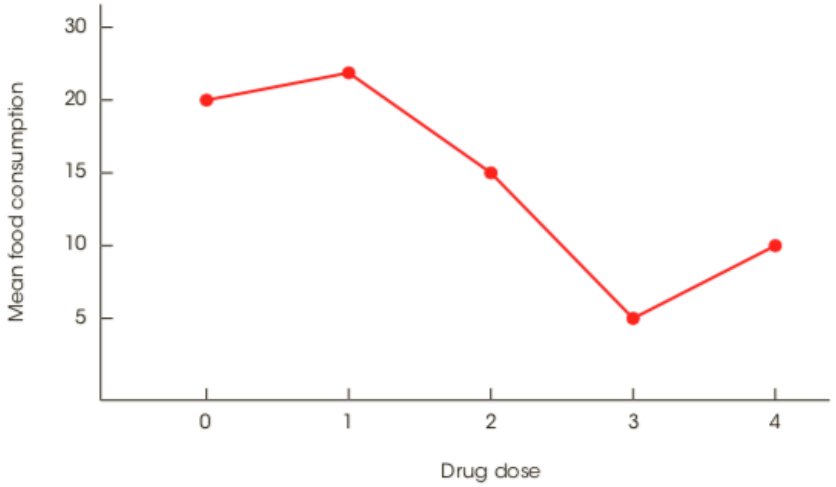
As table

The mean number of errors made on the task for treatment and control groups according to age.

	Treatment	Control
Older Adults	1.45	8.36
Younger Adults	3.83	14.77

Reporting Central Tendency in Research Reports (cont.)

- The different groups or treatment conditions are listed along the horizontal axis and the means are displayed by a bar or a point above each of the groups.
- The height of the bar (or point) indicates the value of the mean for each group. Similar graphs are also used to show several medians in one display.



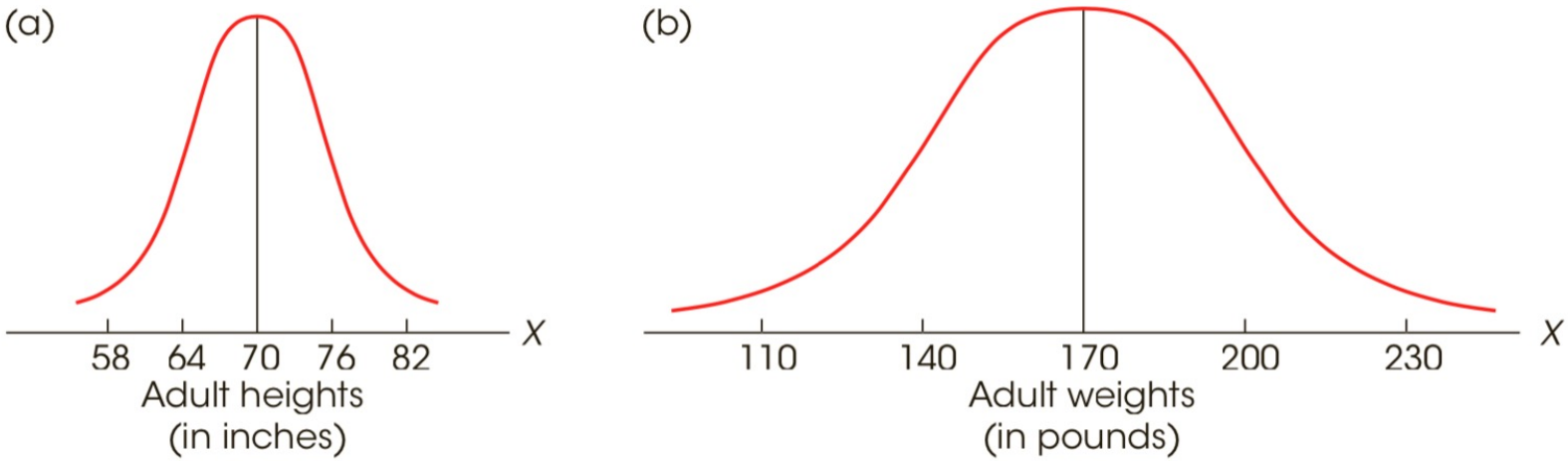
Variability

Central Tendency and Variability

- Central tendency describes the central point of the distribution, and variability describes how the scores are scattered around that central point.
- Together, central tendency and variability are the two primary values that are used to describe a distribution of scores.

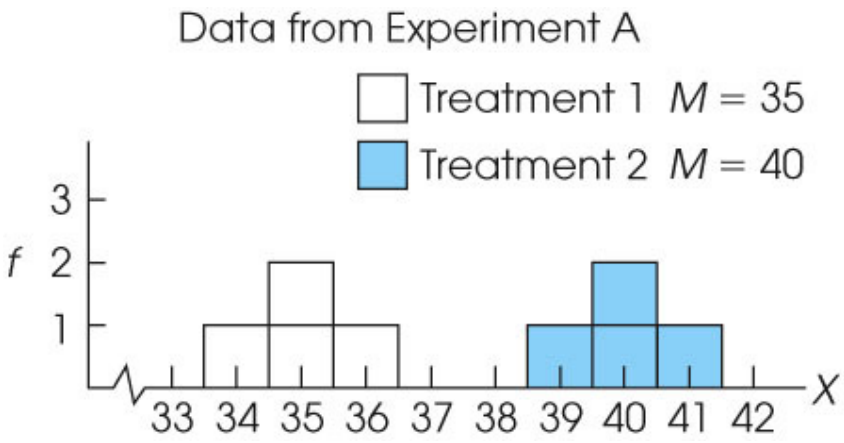
Variability

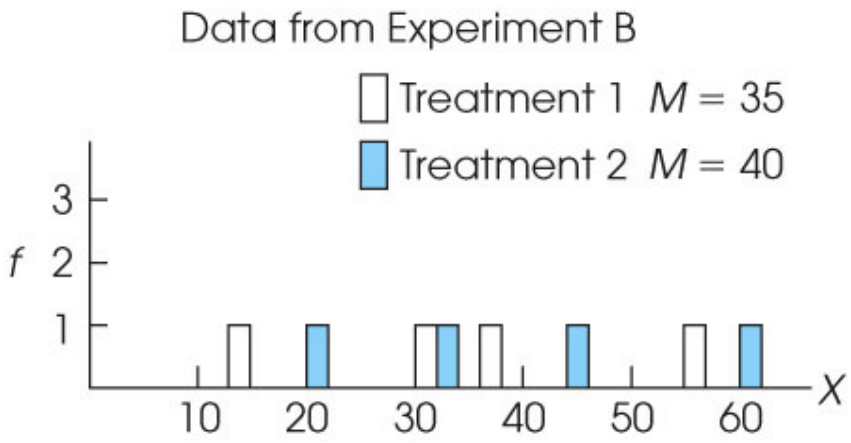
- **Variability** provides a quantitative measure of the differences between scores in a distribution and describes the degree to which the scores are spread out or clustered together.



Variability (cont.)

- Variability serves both as a descriptive measure and as an important component of most inferential statistics.
- As a descriptive statistic, variability measures the degree to which the scores are spread out or clustered together in a distribution.
- In the context of inferential statistics, variability provides a measure of how accurately any individual score or sample represents the entire population.
- **When the population variability is small, all of the scores are clustered close together and any individual score or sample will necessarily provide a good representation of the entire set.**
- On the other hand, when variability is large and scores are widely spread, it is easy for one or two extreme scores to give a distorted picture of the general population.


Data from Experiment A
Treatment 1 M = 35
Treatment 2 M = 40
f
3
2
1
33 34 35 36 37 38 39 40 41 42
X


Data from Experiment B
Treatment 1 M = 35
Treatment 2 M = 40
f
3
2
1
10 20 30 40 50 60
X

Measuring Variability

- Variability can be measured with
  - the range
  - the interquartile range
  - the standard deviation/variance.
- In each case, variability is determined by measuring *distance*.

The Range

- The **range** is the total distance covered by the distribution, from the highest score to the lowest score (using the upper and lower real limits of the range).

Alternative definitions

$$\text{range} = X_{\text{max}} - X_{\text{min}}$$

(for a sample, normal case)

$$\text{range} = \text{URL for } X_{\text{max}} - \text{LRL for } X_{\text{min}}$$

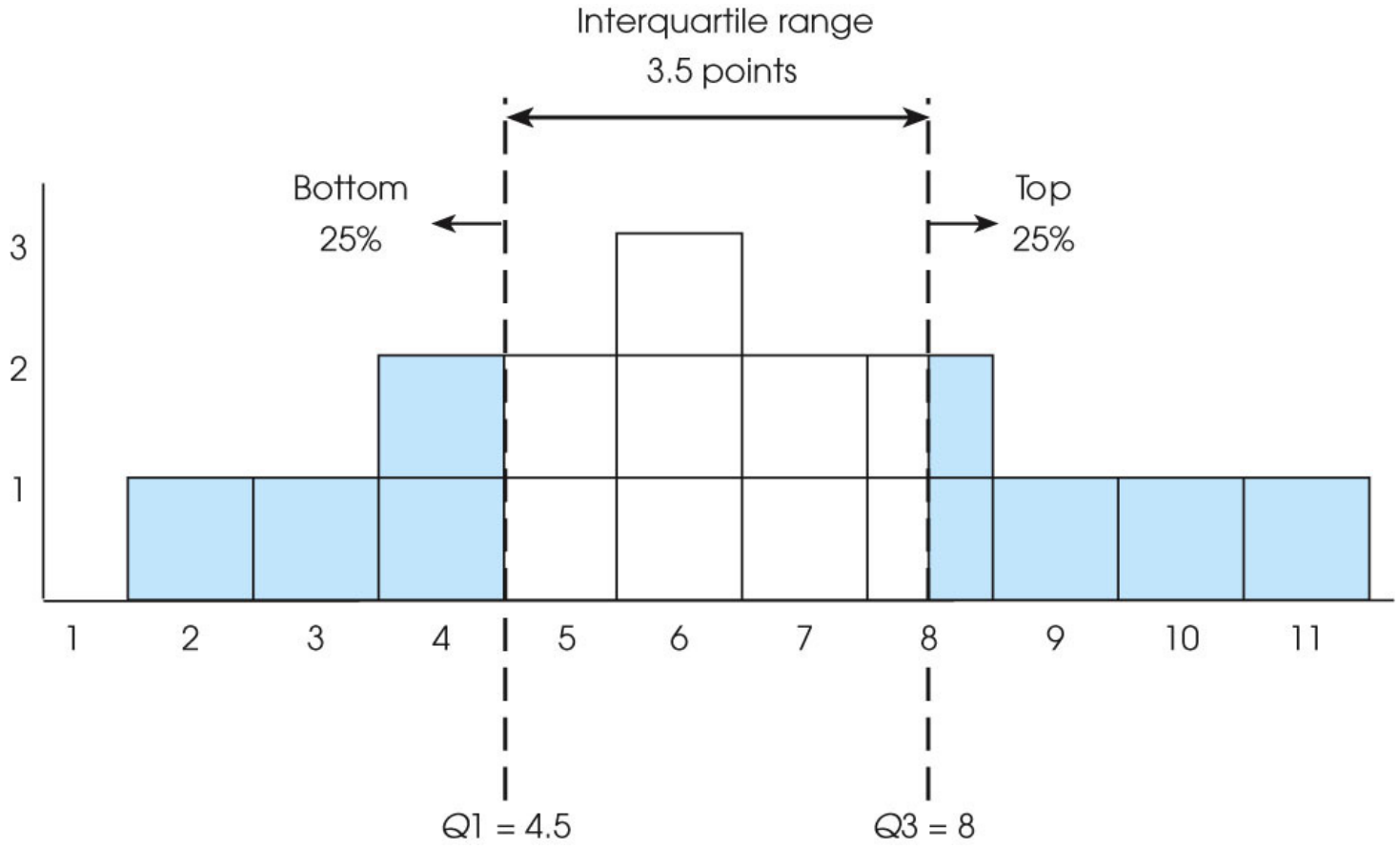
(Continuous variables **with certain precision or rounding**: upper/lower real limit)

$$\text{range} = X_{\text{max}} - X_{\text{min}} + 1.$$

(Integers)

The Interquartile Range

- The **interquartile range** is the distance covered by the middle 50% of the distribution (the difference between Q1 and Q3).

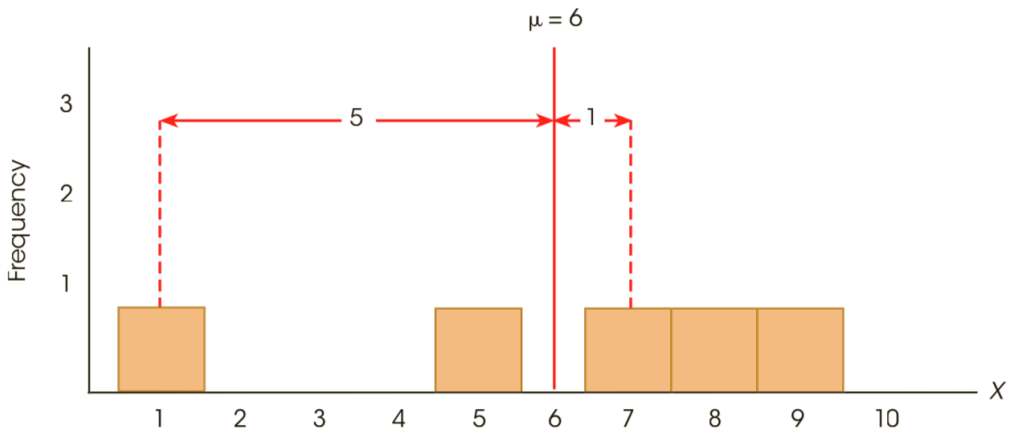


Variance and Standard Deviation

- A **deviation** or **deviation score** is the difference between a score and the mean, and is calculated as:

  deviation = $X - \mu$

- **Variance** equals the mean of the squared deviations. Variance is the average squared distance from the mean.
- **Standard deviation** is the square root of the variance and provides a measure of the standard, or average distance from the mean.



Example

Score X	Deviation $X - \mu$	Squared Deviation $(X - \mu)^2$	
1	−5	25	
9	3	9	
5	−1	1	
8	2	4	
7	1	1	
		40	= the sum of the squared deviations

For this set of $N = 5$ scores, the squared deviations add up to 40. The mean of the squared deviations, the variance, is $\frac{40}{5} = 8$, and the standard deviation is $\sqrt{8} = 2.83$. ■

The Sum of Squared Deviations (*SS*)

- *SS*, or **sum of squares**, is the sum of the squared deviation scores.

Definitional formula: $SS = \Sigma (X - \mu)^2$

Score X	Deviation $X - \mu$	Squared Deviation $(X - \mu)^2$	
1	−1	1	$\Sigma X = 8$
0	−2	4	$\mu = 2$
6	+4	16	
1	−1	1	
		22	$\Sigma (X - \mu)^2 = 22$

Computational formula: $SS = \Sigma X^2 - \frac{(\Sigma X)^2}{N}$

X	X^2
1	1
0	0
6	36
1	1
$\Sigma X = 8$	$\Sigma X^2 = 38$

$$SS = \Sigma X^2 - \frac{(\Sigma X)^2}{N}$$
$$= 38 - \frac{(8)^2}{4}$$
$$= 38 - \frac{64}{4}$$
$$= 38 - 16$$
$$= 22$$

Population versus Sample Variance and Standard Deviation

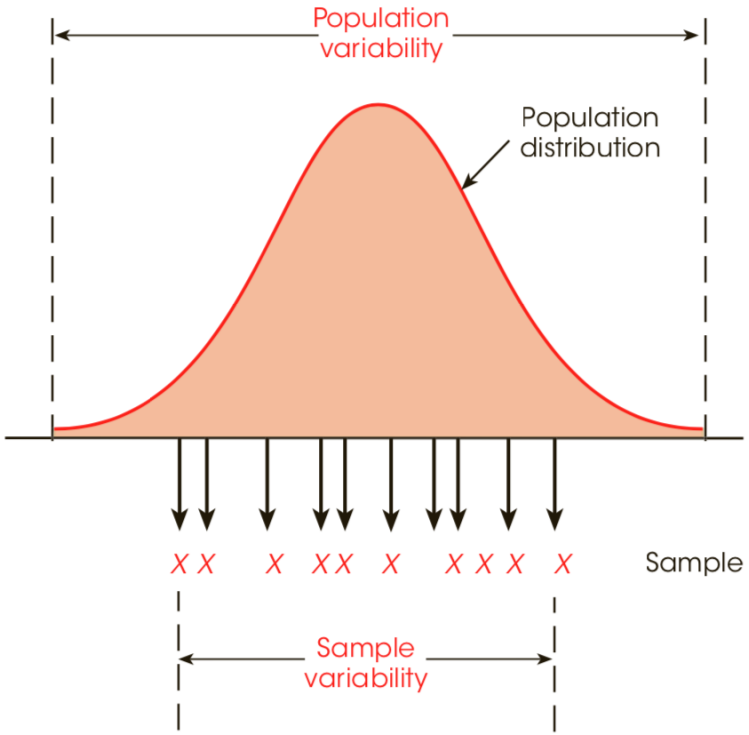
- **Population variance** is represented by the symbol $\boldsymbol{\sigma^2}$ and equals the mean squared distance from the mean. Population variance is obtained by dividing the sum of squares (*SS*) by *N*.
- **Population standard deviation** is represented by the symbol $\boldsymbol{\sigma}$ and equals the square root of the population variance.

$$\text{population standard deviation} = \sigma = \sqrt{\sigma^2} = \sqrt{\frac{SS}{N}}$$

$$\text{population variance} = \sigma^2 = \frac{SS}{N}$$

A Problem with Sample Variability

- The variability for the scores in the sample is on average smaller than the variability for the scores in the population.



Population versus Sample Variance and Standard Deviation

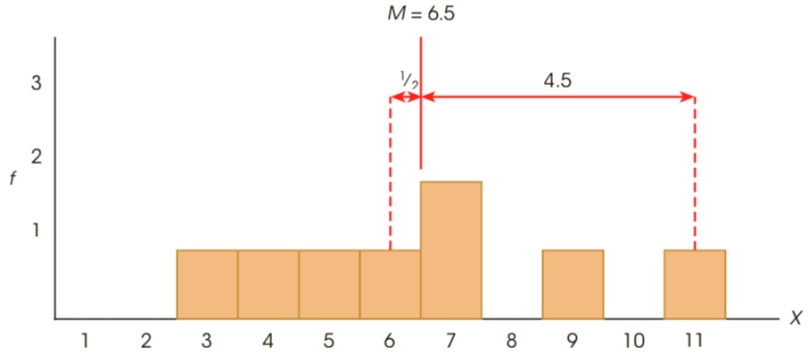
- **Sample variance** is represented by the symbol s^2 and equals the mean squared distance from the mean. Sample variance is obtained by dividing the sum of squares (*SS*) by *n*–1.
- **Sample standard deviation** is represented by the symbol ***s*** and equals the square root of the sample variance.

$$\text{Definitional formula: } SS = \Sigma(X - M)^2$$

$$\text{Computational formula: } SS = \Sigma X^2 - \frac{(\Sigma X)^2}{n}$$

$$\text{sample variance} = s^2 = \frac{SS}{n - 1}$$

$$\text{sample standard deviation} = s = \sqrt{s^2} = \sqrt{\frac{SS}{n - 1}}$$



Scores X	Squared Scores X^2
4	16
6	36
5	25
11	121
7	49
9	81
7	49
3	9
$\Sigma X = 52$	$\Sigma X^2 = 386$

Using the two sums,

$$SS = \Sigma X^2 - \frac{(\Sigma X)^2}{n} = 386 - \frac{(52)^2}{8}$$
$$= 386 - 338$$
$$= 48$$

the sum of squared deviations for this sample is $SS = 48$. Continuing the calculations,

$$\text{sample variance} = s^2 = \frac{SS}{n-1} = \frac{48}{8-1} = 6.86$$

Finally, the standard deviation is

$$s = \sqrt{s^2} = \sqrt{6.86} = 2.62$$

Sample Variability and Degrees of Freedom

- For a sample of *n* scores, the **degrees of freedom**, or ***df***, for the sample variance are defined as $df = n - 1$. The degrees of freedom determine the number of scores in the sample that are independent and free to vary.

Calculating the value of *M* places a restriction on the variability of the scores in the sample: with a sample of n scores, the first $n - 1$ scores are free to vary, but the final score is restricted. As a result, the sample is said to have $n - 1$ degrees of freedom.

X	A sample of $n = 3$ scores with a mean of $M = 5$.
2	
9	
—	← What is the third score?

Sample Variability and Degrees of Freedom (cont.)

$$\text{mean} = \frac{\text{sum}}{\text{number}}$$

$$s^2 = \frac{\text{sum of squared deviations}}{\text{number of scores free to vary}} = \frac{SS}{df} = \frac{SS}{n-1}$$

Sample Variance as an Unbiased Statistic

- A sample statistic is **unbiased** if the average value of the statistic is equal to the population parameter. (The average value of the statistic is obtained from all the possible samples for a specific sample size, *n*.)
- A sample statistic is **biased** if the average value of the statistic either underestimates or overestimates the corresponding population parameter.
- Sample variance (i.e. divided by $n-1$) is an unbiased estimate of the variance of the population.

Illustration: “divided by n – 1” instead of “divided by n” is unbiased

Sample	First Score	Second Score	Sample Statistics		
			Mean M	Biased Variance (Using n)	Unbiased Variance (Using $n - 1$)
1	0	0	0.00	0.00	0.00
2	0	3	1.50	2.25	4.50
3	0	9	4.50	20.25	40.50
4	3	0	1.50	2.25	4.50
5	3	3	3.00	0.00	0.00
6	3	9	6.00	9.00	18.00
7	9	0	4.50	20.25	40.50
8	9	3	6.00	9.00	18.00
9	9	9	9.00	0.00	0.00
		Totals	36.00	63.00	126.00

Properties of the Standard Deviation

- Adding a constant to each score *does not change* the standard deviation.
- Multiplying each score by a constant causes the standard deviation to be *multiplied by the same constant*.

- (How about variance?)

The Mean and Standard Deviation as Descriptive Statistics

- If you are given numerical values for the mean and the standard deviation, you should be able to construct a visual image (or a sketch) of the distribution of scores.
- As a general rule, about 70% of the scores will be within one standard deviation of the mean, and about 95% of the scores will be within two standard deviations of the mean.

Reporting the Standard Deviation

Children who viewed the violent cartoon displayed more aggressive responses ($M = 12.45$, $SD = 3.7$) than those who viewed the control cartoon ($M = 4.22$, $SD = 1.04$).

TABLE 4.2

The number of aggressive behaviors for male and female adolescents after playing a violent or nonviolent video game.

	Type of Video Game	
	Violent	Nonviolent
Males	$M = 7.72$ $SD = 2.43$	$M = 4.34$ $SD = 2.16$
Females	$M = 2.47$ $SD = 0.92$	$M = 1.61$ $SD = 0.68$

Review questions for central tendency and variability

- How are different measures of central tendency and variability defined?
- What are their properties?
- When to use which?
- What are the differences between population and sample variances (standard deviations)?

Manual calculation and R exercises

- Compute the **Mean, Mode, Median, Range, Standard Deviation, and Variance** for a sample of scores.
- mean(), median(), range(), var(), sd()
- How to compute mode in R?