Stat 154: Elementary Statistics

Jongyun Jung

Minnesota State University, Mankato

Ch 2: **Descriptive Statistics**

jongyun.jung@mnsu.edu

January 25, 2019

Overview

1. Measures of Central Tendency
   - Arithmetic Average or Mean
   - Unbiased Estimator
   - Mean of Grouped Frequency Data
   - Median
   - Mode
   - Midrange
   - Relation between Mean, Median and Mode
2. Measures of Variation
   - Range
   - Mean Absolute Deviation (MAD)
   - Variance
   - Standard Deviation
   - Coefficient of Variation
   - Chebyshev's Inequality or Theorem
   - Empirical Rule
   - Z-Score

Arithmetic Average or Mean

- There are three different means: arithmetic mean, geometric mean and harmoic mean. But we only consider **arithmetic mean** in this course.
- Let us consider $x_1, x_2, ..., x_N$ as the population data having N measurements. Then the population mean is the average of the N measurements denoted as μ,

$$\mu = \frac{x_1 + x_2 + ... + x_n}{N} = \frac{\sum x}{N}$$

- where $\sum$ (upper case sigma) stands for the sum.

Arithmetic Average or Mean

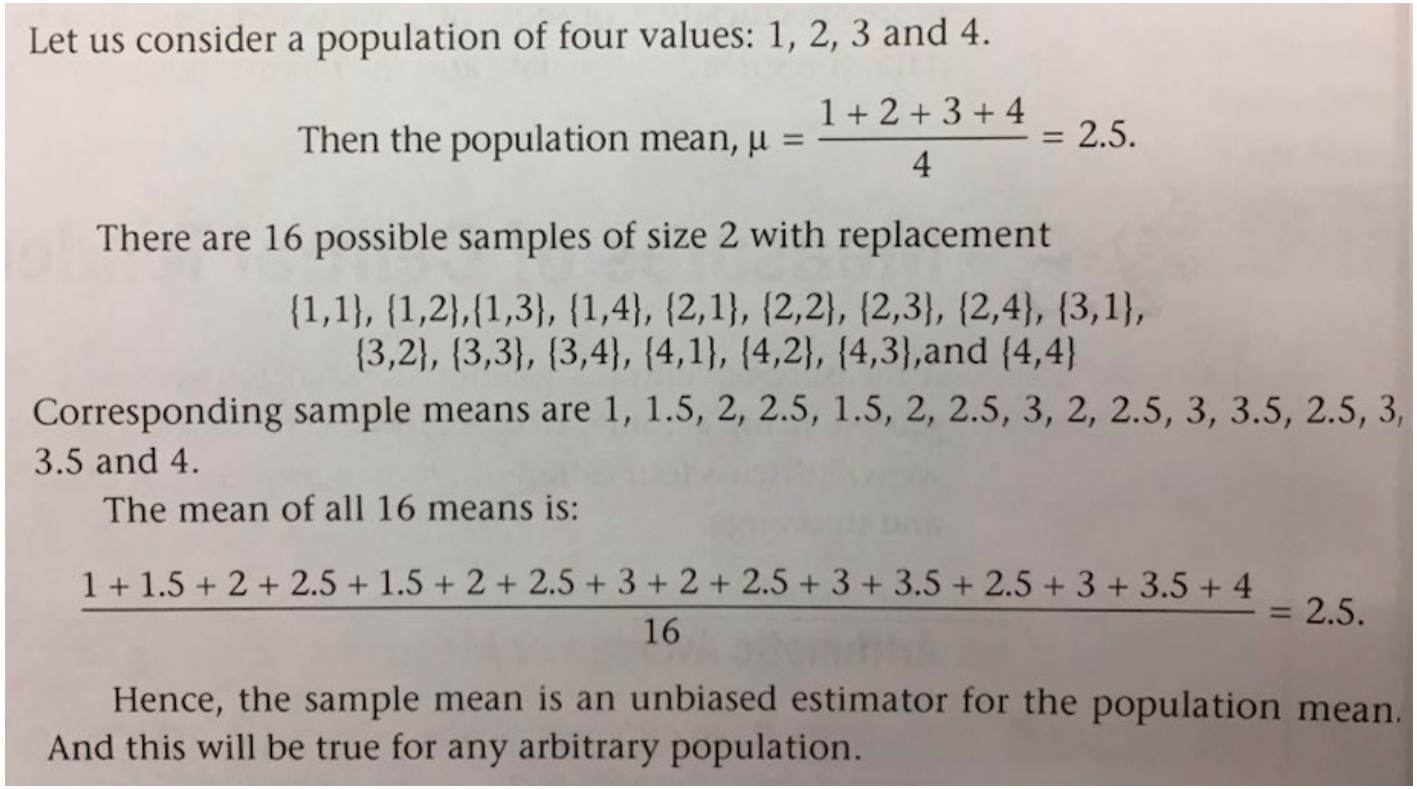
- Let us consider $x_1, x_2, ..., x_N$ as the sample data having N measurements. Then the sample mean is the average of the N measurements denoted as μ,

$$\overline{X} = \frac{x_1 + x_2 + ... + x_n}{n} = \frac{\sum x}{n}$$

- The sample mean $\overline{X}$ is a sample statistic and can product different values depending on the sample selected.
- The sample mean $\overline{X}$ is an **unbiased estimator** of the population mean μ.

Unbiased Estimator

- Let us consider θ(theta) as a population parameter and $\widehat{\theta}$ is an unbiased estimator for θ if the mean of all possible $\widehat{\theta}$'s is exactly θ.



Let us consider a population of four values: 1, 2, 3 and 4.

Then the population mean, $\mu = \frac{1+2+3+4}{4} = 2.5$.

There are 16 possible samples of size 2 with replacement

{1,1}, {1,2},{1,3}, {1,4}, {2,1}, {2,2}, {2,3}, {2,4}, {3,1},
{3,2}, {3,3}, {3,4}, {4,1}, {4,2}, {4,3},and {4,4}

Corresponding sample means are 1, 1.5, 2, 2.5, 1.5, 2, 2.5, 3, 2, 2.5, 3, 3.5, 2.5, 3, 3.5 and 4.

The mean of all 16 means is:

$$\frac{1+1.5+2+2.5+1.5+2+2.5+3+2+2.5+3+3.5+2.5+3+3.5+4}{16} = 2.5.$$

Hence, the sample mean is an unbiased estimator for the population mean. And this will be true for any arbitrary population.

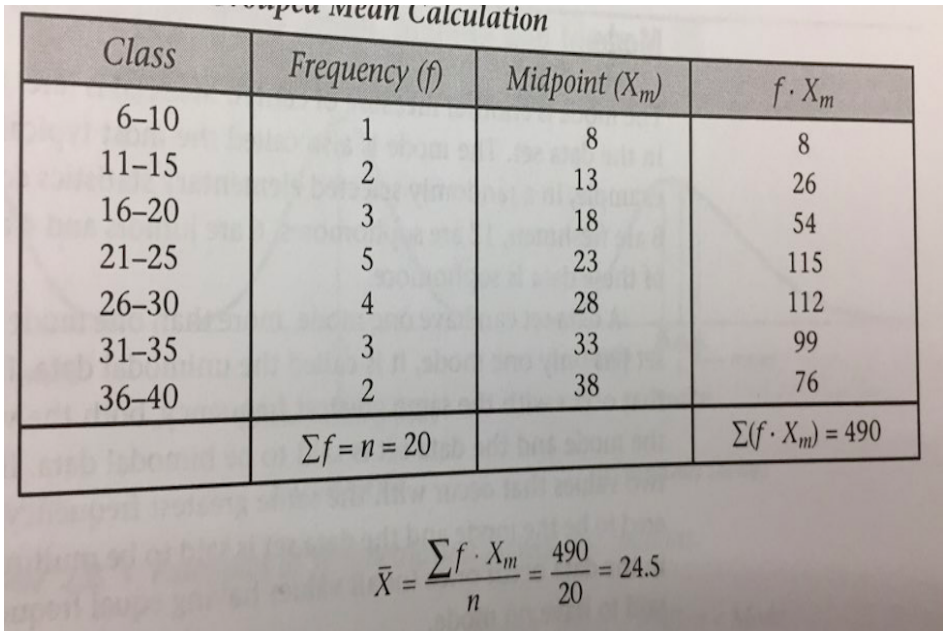
Figure: Unbiased Estimator

Mean of Grouped Frequency Data

- The mean in cases of grouped data is found using the following formula:

$$\overline{X} = \frac{\sum f \times X_m}{n}$$

- where X_m is the class midpoint. $x_m = \dfrac{LowerLimit + UpperLimit}{2}$.



Grouped Mean Calculation

Class	Frequency (f)	Midpoint (X_m)	$f \cdot X_m$
6–10	1	8	8
11–15	2	13	26
16–20	3	18	54
21–25	5	23	115
26–30	4	28	112
31–35	3	33	99
36–40	2	38	76
	$\Sigma f = n = 20$		$\Sigma(f \cdot X_m) = 490$

$$\bar{X} = \frac{\sum f \cdot X_m}{n} = \frac{490}{20} = 24.5$$

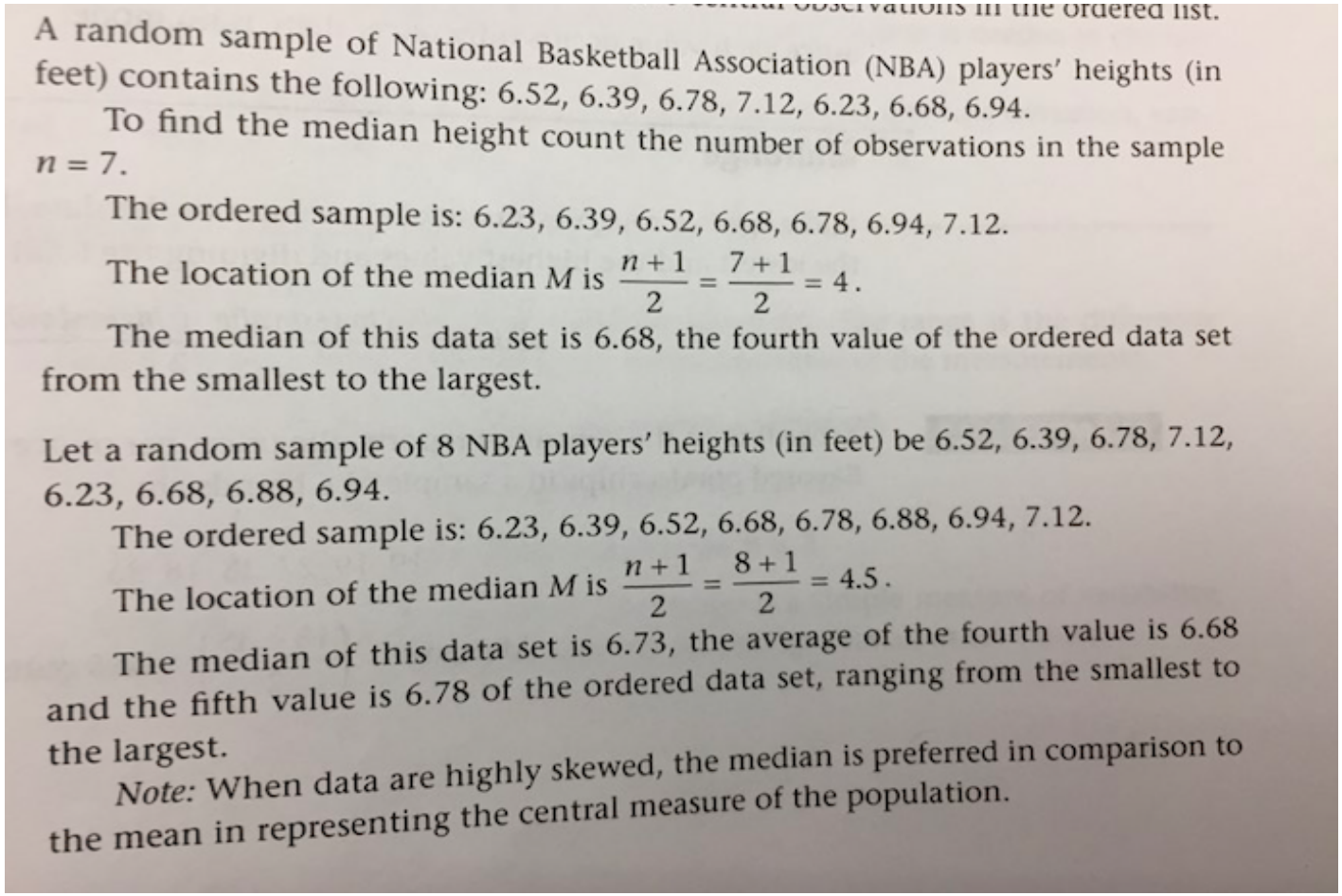
Figure: Grouped Mean Calculation

Median

- The median can be for a population or for a sample. We will consider only for sample in this course.
- Median M is the midpoint of the distribution meaning: half of the observation are smaller than the median and the other half are larger than the median.
- How to find the median?
  1. Arrange the observations in order, from the smallest to the largest.
  2. Find $\frac{n+1}{2}$ which is the position of the median where n is the number of observations in the data. Then locate the $\frac{n+1}{2}th$ positioned measurement in the ordered list, which is the median.
  3. If the number of observations n is odd, the median M is the central observation in the ordered list. If the number of the observations n is even, the median M is the mean of the two central observations in the ordered list.

Median

- Median Example



A random sample of National Basketball Association (NBA) players' heights (in feet) contains the following: 6.52, 6.39, 6.78, 7.12, 6.23, 6.68, 6.94.

To find the median height count the number of observations in the sample $n = 7$.

The ordered sample is: 6.23, 6.39, 6.52, 6.68, 6.78, 6.94, 7.12.

The location of the median M is $\frac{n+1}{2} = \frac{7+1}{2} = 4$.

The median of this data set is 6.68, the fourth value of the ordered data set from the smallest to the largest.

Let a random sample of 8 NBA players' heights (in feet) be 6.52, 6.39, 6.78, 7.12, 6.23, 6.68, 6.88, 6.94.

The ordered sample is: 6.23, 6.39, 6.52, 6.68, 6.78, 6.88, 6.94, 7.12.

The location of the median M is $\frac{n+1}{2} = \frac{8+1}{2} = 4.5$.

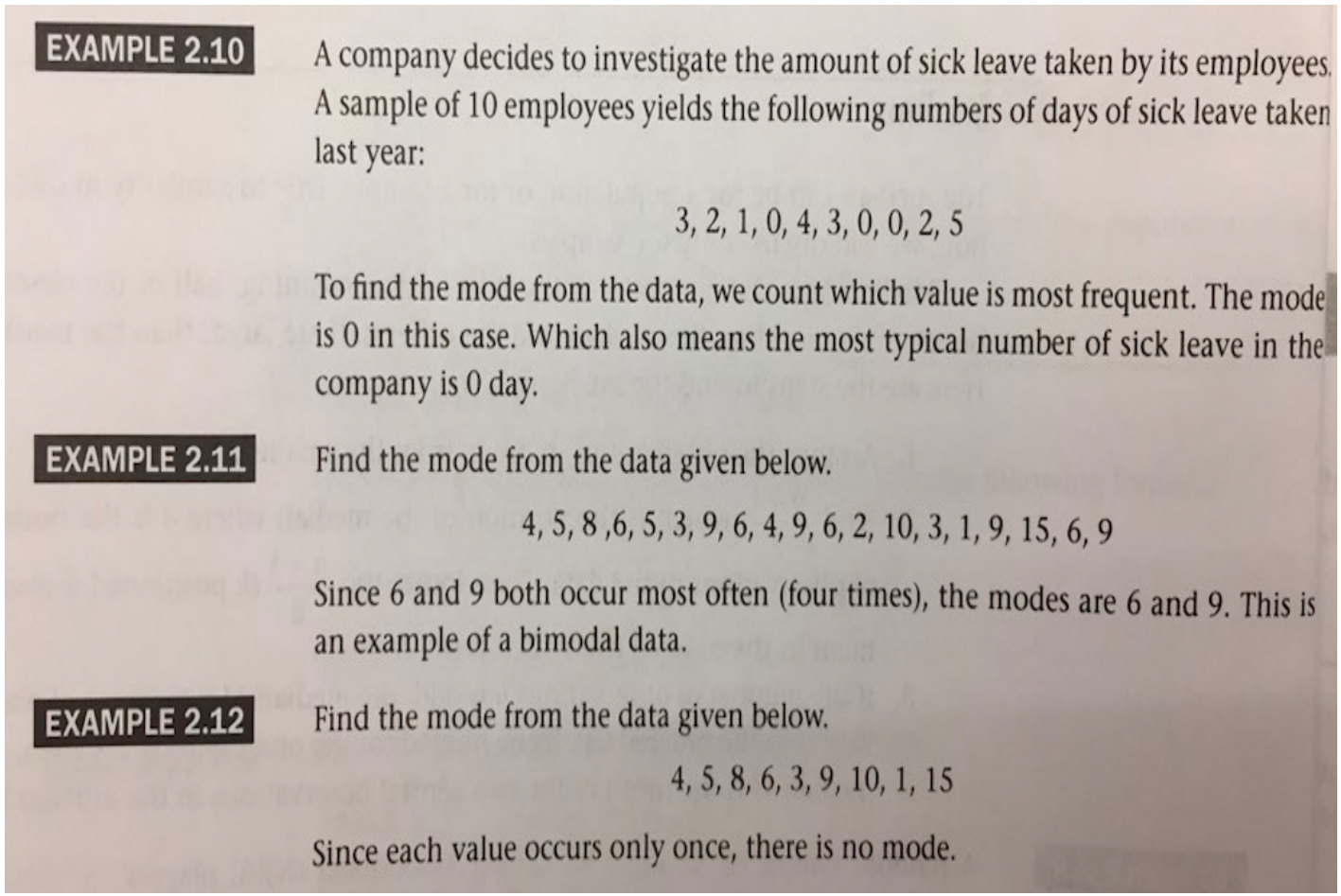
The median of this data set is 6.73, the average of the fourth value is 6.68 and the fifth value is 6.78 of the ordered data set, ranging from the smallest to the largest.

Note: When data are highly skewed, the median is preferred in comparison to the mean in representing the central measure of the population.

Figure: Example 2-8 and 2-9

Mode

- Mode is another measure of center.
- Mode(s) is (are) the most frequent value(s) in the data set.
- Also, it is called most frequent value in the data set.



EXAMPLE 2.10 A company decides to investigate the amount of sick leave taken by its employees. A sample of 10 employees yields the following numbers of days of sick leave taken last year:

3, 2, 1, 0, 4, 3, 0, 0, 2, 5

To find the mode from the data, we count which value is most frequent. The mode is 0 in this case. Which also means the most typical number of sick leave in the company is 0 day.

EXAMPLE 2.11 Find the mode from the data given below.

4, 5, 8 ,6, 5, 3, 9, 6, 4, 9, 6, 2, 10, 3, 1, 9, 15, 6, 9

Since 6 and 9 both occur most often (four times), the modes are 6 and 9. This is an example of a bimodal data.

EXAMPLE 2.12 Find the mode from the data given below.

4, 5, 8, 6, 3, 9, 10, 1, 15

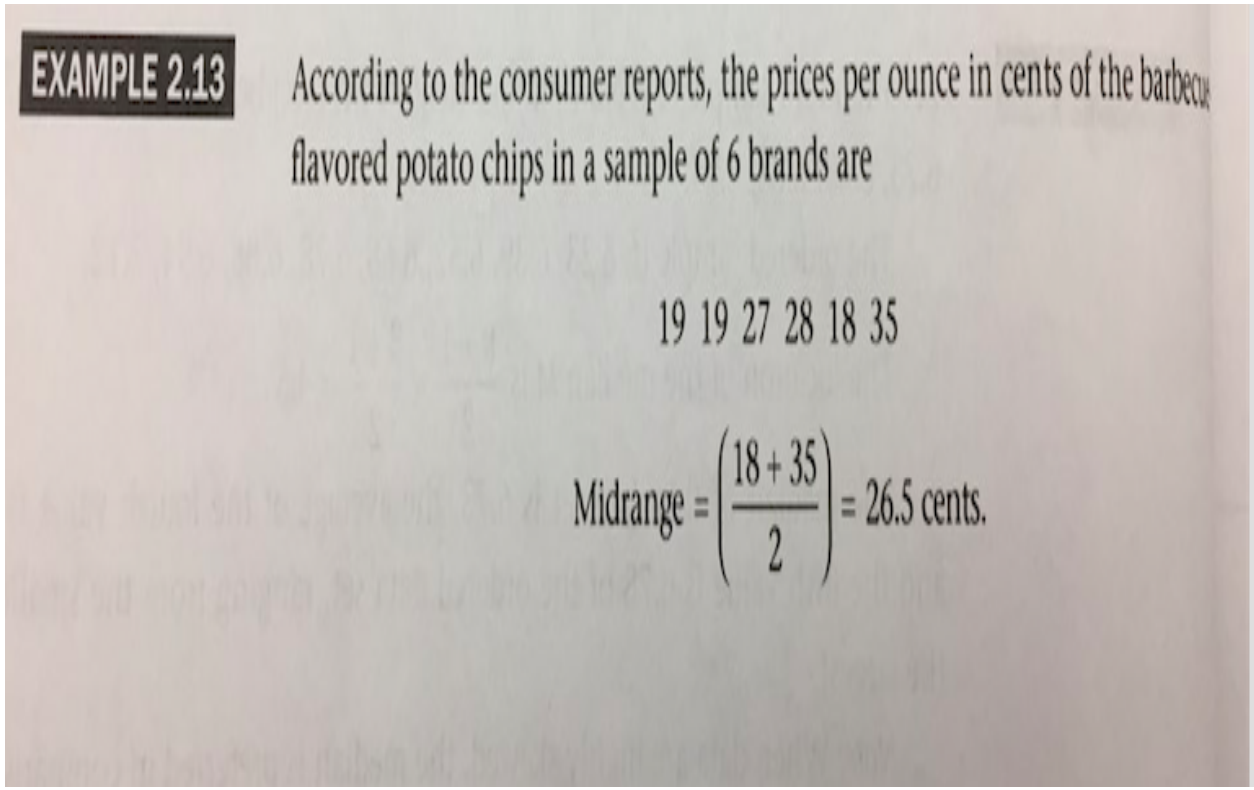
Since each value occurs only once, there is no mode.

Figure: Mode

Midrange

- Midrange is a rough estimate of the center of the data.
- It is computed by as below:

$$midrange = \frac{lowestvalue + highestvalue}{2}$$



EXAMPLE 2.13 According to the consumer reports, the prices per ounce in cents of the barbecue flavored potato chips in a sample of 6 brands are

19 19 27 28 18 35

$$\text{Midrange} = \left(\frac{18+35}{2}\right) = 26.5 \text{ cents.}$$

Figure: Midrange

Relation between Mean, Median and Mode

- When the distribution is negatively skewed, $Mean < Median < Mode$.
- When the distribution is positively skewed, $Mode < Median < Mean$.
- When the distribution is symmetric or bell shaped, $Mean = Median = Mode$.

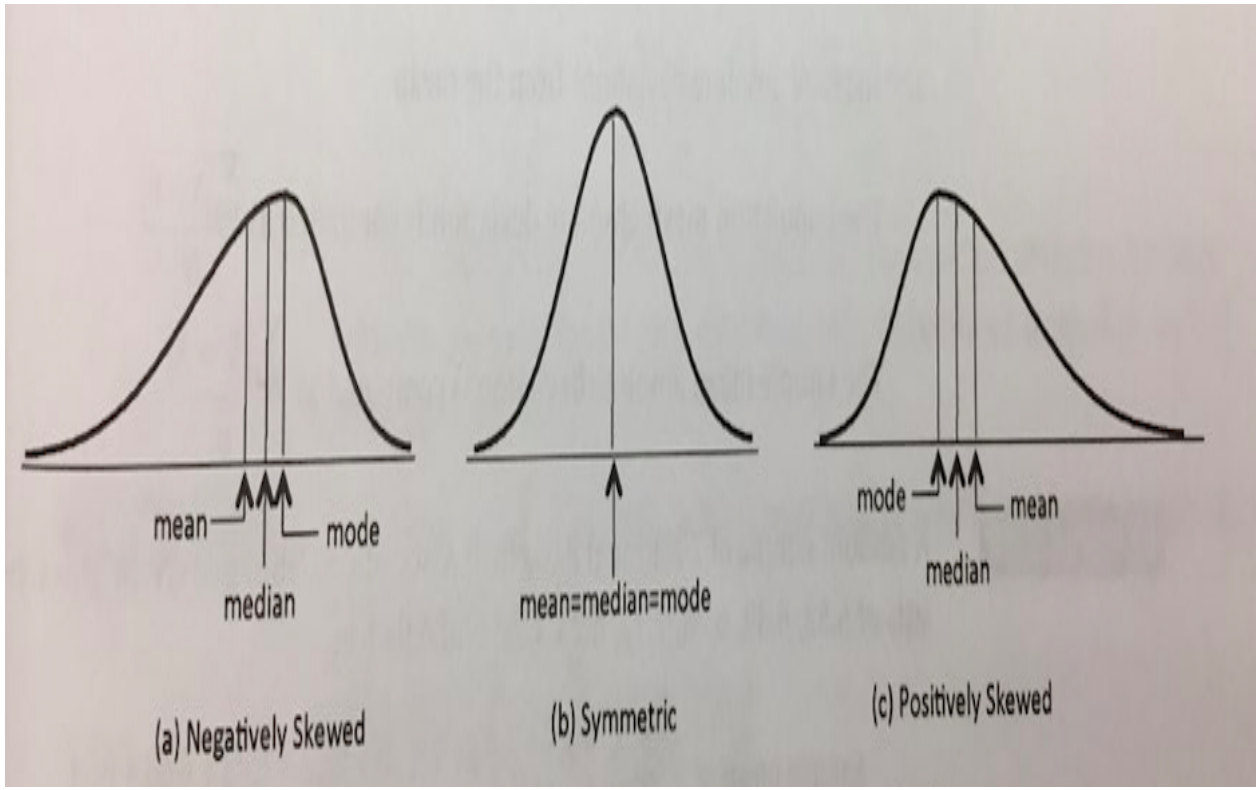


Figure: Examples of symmetric and skewed distributions

Range

- Measures of central tendency give us measure of where the center of a set of data is, but this is not sufficient to characterize the data.
- The Range is the simplest measure of variation. The range is the difference between the maximum value and the minimum value of the measurements.

Range = Maximum - Minimum

Data Set A: 50, 60, 70, 80, 90

Data Set B: 69, 69, 70, 71, 71

Mean Absolute Deviation (MAD)

- The mean of absolute deviations from the mean.
- The population mean absolute deviation is computed as below

$$\frac{\sum |x - \mu|}{N}$$

- The sample mean absolute deviation is computed as below

$$\frac{\sum |x - \overline{X}|}{N}$$

EXAMPLE 2.15 A random sample of 7 National Basketball Association (NBA) players' heights co sists of 6.52, 6.39, 6.78, 7.12, 6.23, 6.68 and 6.94 feet.

$$\text{Sample mean, } \bar{X} = \frac{\sum X}{n} = \frac{6.52 + 6.39 + 6.78 + 7.12 + 6.23 + 6.68 + 6.94}{7}$$

$$= \frac{46.66}{7} = 6.67.$$

$$\text{Mean absolute deviation, } MAD = \frac{\sum |X - \bar{X}|}{n} =$$

$$\frac{1}{7}\left(\begin{array}{l}|6.52 - 6.67| + |6.39 - 6.67| + |6.78 - 6.67| + |7.12 - 6.67| + \\ |6.23 - 6.67| + |6.68 - 6.67| + |6.94 - 6.67|\end{array}\right) =$$

$$\frac{0.15 + 0.28 + 0.11 + 0.45 + 0.44 + 0.01 + 0.27}{7} = \frac{1.71}{7} = 0.2443.$$

As a function of absolute value, it is not a continuous function and hence di not get popularity as some mathematical analyses are complex in nature. Hence the following measure, the variance, is developed using a quadratic function.

Variance & Standard Deviation

- The variance is the measure of the spread of data around the mean.

$$\sigma^2 = \frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + ... + (x_N - \mu)^2}{N} = \frac{\sum(x - \mu)^2}{N}$$

- a sample variance is denoted by s^2 and computed as

$$s^2 = \frac{(x_1 - \overline{X})^2 + (x_2 - \overline{X})^2 + ... + (x_N - \overline{X})^2}{N} = \frac{\sum(x - \overline{X})^2}{n - 1}$$

- The computational formula for a sample variance s^2 is

$$s^2 = \frac{n\sum X^2 - (\sum X)^2}{n(n-1)}$$

Variance & Standard Deviation

EXAMPLE 2.16

The total amount of snowfall in Minnesota in 2011 is estimated from a sample of 10 locations given below.

Location	*Inches*
ANDOVER	19.0
MINNETONKA	18.0
CHASKA	17.5
HUTCHINSON	16.5
PLYMOUTH	17.0
DEEPHAVEN	16.0
HILLTOP	14.5
CHANHASSEN	14.0
WACONIA	14.0
BUFFALO	13.5

Find the sample variance and the sample standard deviation using both the main and shortcut formulas.

Using the main formula,

$$s^2 = \frac{\sum (X - \bar{X})^2}{n - 1}$$

First find the sample mean,

$$\bar{X} = \frac{\sum X}{n} = \frac{19.0 + 18.0 + 17.5 + \cdots + 13.5}{10} = \frac{160.0}{10} = 16.0$$

The table below shows the calculation of the sample variance.

TABLE 2.6 • *Computations for the Variance*

X	$X - \bar{X}$	$(X - \bar{X})^2$
19.0	3	9
18.0	2	4
17.5	1.5	2.25
16.5	0.5	0.25
17.0	1	1
16.0	0	0
14.5	−1.5	2.25
14.0	−2	4
14.0	−2	4
13.5	−2.5	6.25
	$\sum (X - \bar{X}) = 0$	$\sum (X - \bar{X})^2 = 33$

Variance & Standard Deviation

TABLE 2.7 • *The Variance Using the Shortcut Formula*

X	X^2
19.0	361
18.0	324
17.5	306.25
16.5	272.25
17.0	289
16.0	256
14.5	210.25
14.0	196
14.0	196
13.5	182.25
$\Sigma X = 160$	$\Sigma X^2 = 2593$

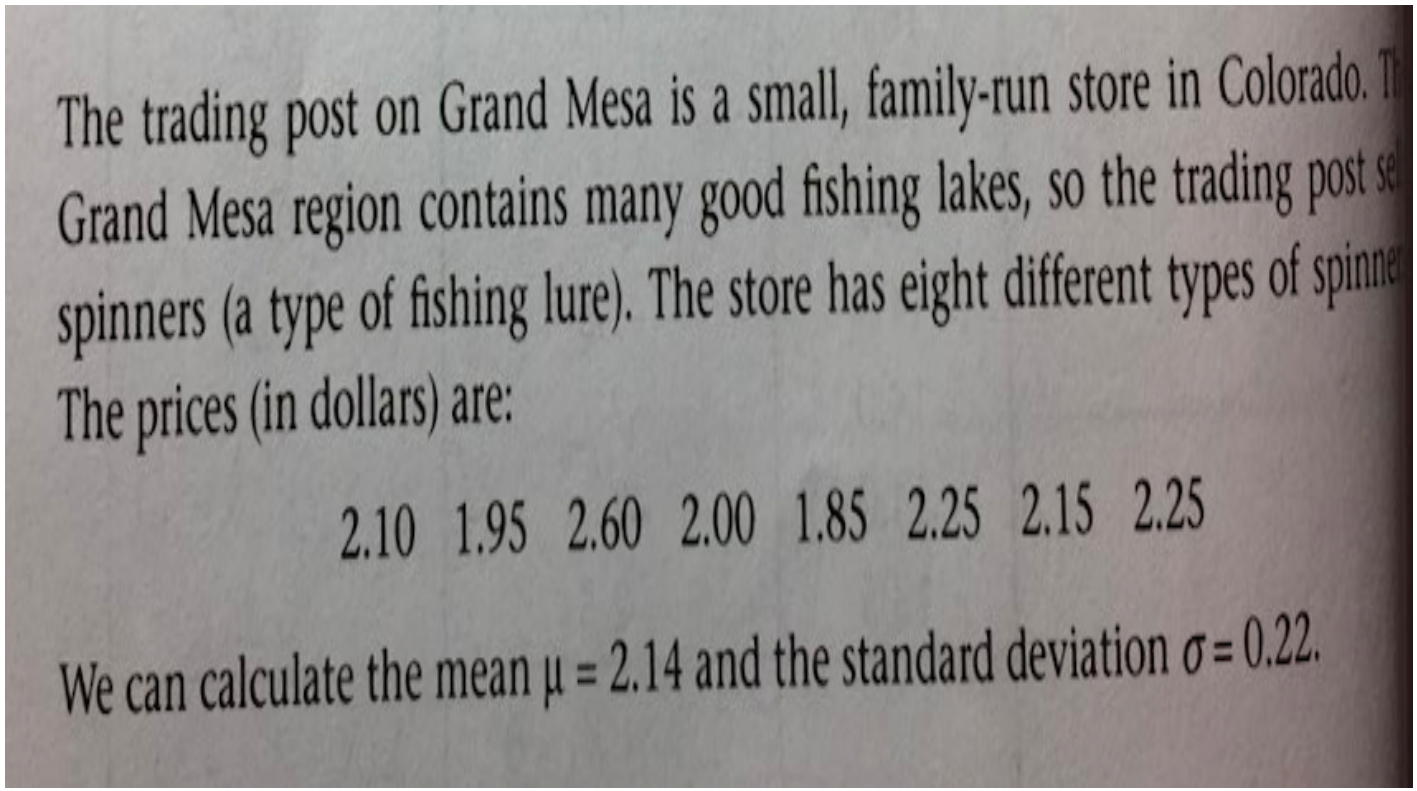
The sample variance using the shortcut formula,

$$s^2 = \frac{n\sum X^2 - \left(\sum X\right)^2}{n(n-1)} = \frac{10 \times 2593 - (160)^2}{10 \times 9} = 3.667,$$

Coefficient of Variation

- We use the coefficient of variation (CV) in order to compare the variation in two or more different distributions when the unit of measurement is not the same.

$$CV = \frac{\text{Standard deviation}}{mean} \times 100\%$$


The trading post on Grand Mesa is a small, family-run store in Colorado. Th
Grand Mesa region contains many good fishing lakes, so the trading post sel
spinners (a type of fishing lure). The store has eight different types of spinne
The prices (in dollars) are:

2.10 1.95 2.60 2.00 1.85 2.25 2.15 2.25

We can calculate the mean $\mu = 2.14$ and the standard deviation $\sigma = 0.22$.

Figure: Coefficient of Variation

Chebyshev's Inequality or Theorem

- For a population, Chebyshev's inequality applies to any distribution, which state that for $k > 1$, at least $(1 - \frac{1}{k^2} \times 100\%)$ measurements will fall between $\mu - k\sigma$ and $\mu + k\sigma$. Or at least $(1 - \frac{1}{k^2} \times 100\%)$ of the measurements will fall within the k standard deviation of the mean.
- For a sample, Chebyshev's inequality applies to any number of data, which state that for $k > 1$, at least $(1 - \frac{1}{k^2} \times 100\%)$ measurements will fall between $\overline{X} - ks$ and $\overline{X} + ks$. Or at least $(1 - \frac{1}{k^2} \times 100\%)$ of the measurements will fall within the k standard deviation of the mean.

Chebyshev's Inequality or Theorem

Let the mean price of houses in a certain neighborhood be \$50,000 and the standard deviation \$10,000. Find the price range for which at least 75% of the houses will sell.

Using Chebyshev's theorem, 75% of the data values will fall within 2 standard deviations of the mean, that is, in the interval $(\bar{X} - 2s, \bar{X} + 2s)$.

Here, $\bar{X} - 2s = 50000 - 2(10000) = 30000$ and $\bar{X} + 2s = 50000 + 2(10000) = 70000$.

Hence, at least 75% of the houses that will be sold in the area will be within the price range of \$30,000 to \$70,000.

Figure: Chebyshev's Inequality or Theorem

Empirical Rule

- When a distribution is symmetric and bell shaped, approximately 68 % of the data values will fall between 1 standard deviation of the mean or within $(\overline{X} - s, \overline{X} + s)$, approximately 95 % of the data values will fall within 2 standard deviations of the mean or within $(\overline{X} - 2s, \overline{X} + 2s)$, and approximately 99.7 % (or almost all) of the data values will fall within 3 standard deviations of the mean or within $(\overline{X} - 3s, \overline{X} + 3s)$.
- Note that for population measurements the corresponding intervals are $(\mu - \sigma, \mu + \sigma), (\mu - 2\sigma, \mu + 2\sigma), (\mu - 3\sigma, \mu + 3\sigma)$, respectively.

Empirical Rule

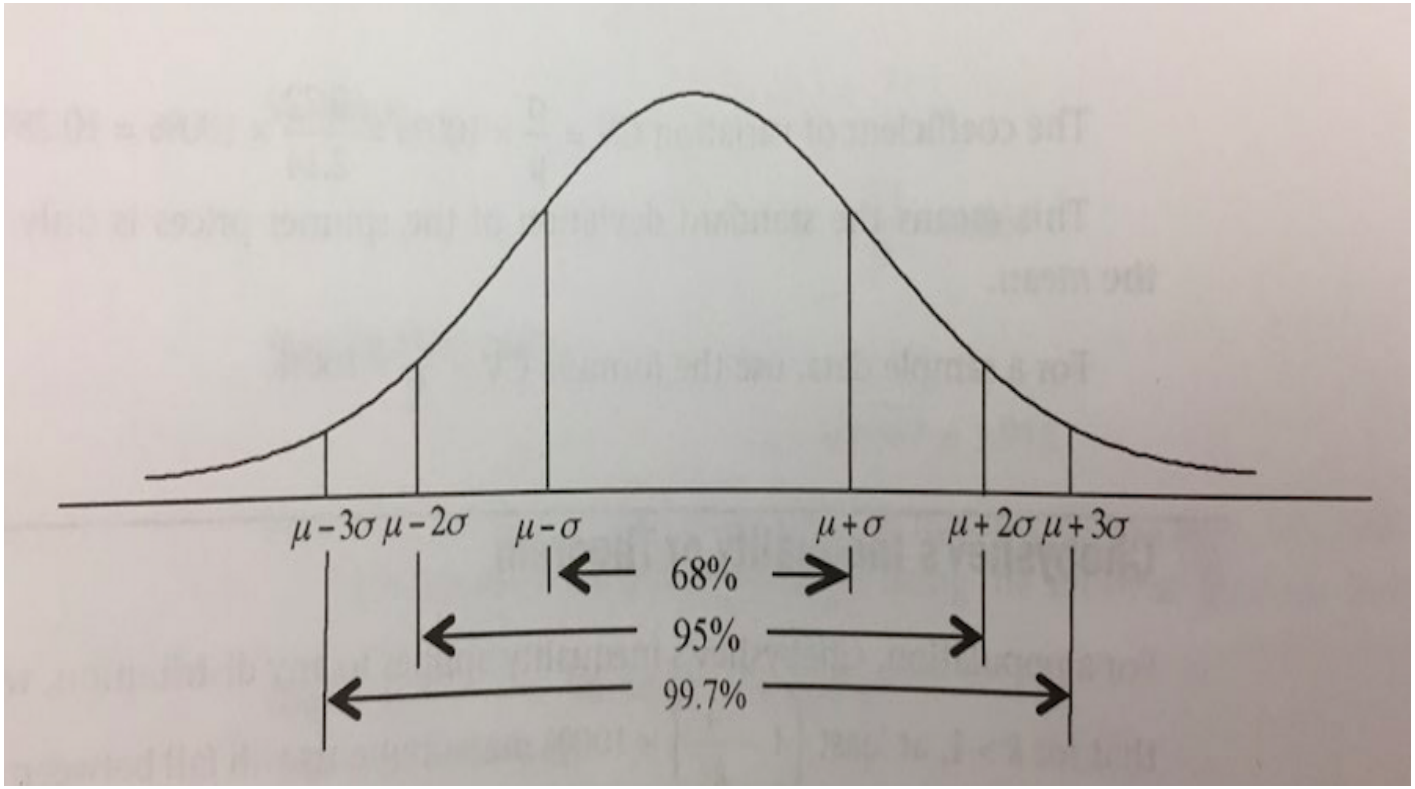


Figure: Empirical Rule

Empirical Rule

EXAMPLE 2.19

A random sample of 7 National Basketball Association (NBA) players' heights feet) consists of 6.52, 6.39, 6.78, 7.12, 6.23, 6.68, 6.94.

Sample mean,

$$\bar{X} = \frac{\sum X}{n} = \frac{6.52 + 6.39 + 6.78 + 7.12 + 6.23 + 6.68 + 6.94}{7}$$

$$= \frac{46.66}{7} = 6.67.$$

Sample variance,

$$s^2 = \frac{n\sum X^2 - \left(\sum X\right)^2}{n(n-1)},$$

where

$$\sum X^2 = (6.52)^2 + (6.39)^2 + \cdots + (6.94)^2 = 311.6042.$$

Then

$$s^2 = \frac{7(311.6042) - (46.66)^2}{7(7-1)} = 0.0970.$$

Sample standard deviation,

$$s = \sqrt{s^2} = \sqrt{0.0970} = 0.3114.$$

Assuming that the distribution is bell shaped, the empirical rule is applied follows. Approximately 68% of the measurements for the respective populati are between 6.67 − 0.3114 = 6.3586 and 6.67 + 0.3114 = 6.9814, i.e., betwe (6.3586, 6.9814), which is the one standard deviation of the mean interv Approximately 95% of the measurements are between 6.67 − 2(0.3114) = 6.0 and 6.67 + 2(0.3114) = 7.2928, i.e., between (6.0472, 7.2928), which is the standard deviation of the mean interval; and almost all measurements, or 99.7 are between 6.67 − 3(0.3114) = 5.7358 and 6.67 + 3(0.3114) = 7.6042, i.e., betwe (5.7358, 7.6042), which is the three standard deviation of the mean interval.

EXAMPLE 2.20

IQ scores have a bell-shaped di

Z - Score

- A standardized score is also called a $z-score$ for an observation is obtained by subtracting the mean from that observation and then dividing the result by the standard deviation.
- The population $z-score$ is

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

- The sample $z-score$ is

$$Z = \frac{X - \overline{X}}{s}$$

- Note:
  1. The $z-score$ measures the number of standard deviations in which a data value falls above or below the mean.
  2. If the $z-score$ is positive, the actual value is greater than the mean.
  3. If the $z-score$ is negative, the actual value is lower than the mean.
  4. If the $z-score$ is zero, the actual value is the same as the mean.

Z - Score

EXAMPLE 2.21 The human body temperature has a mean of 98.2°F and a standard deviation of 0.62°F. Convert 97.5°F to a *z-score*.

$$z\text{-score, } Z = \frac{X - \bar{X}}{s} = \frac{97.5 - 98.2}{0.62} = -1.129.$$

EXAMPLE 2.22 A student scored 85 on a test where the standard deviation was 3. The student's z score was 2.33. Find the test average

$$\text{If } Z = \frac{X - \bar{X}}{s}, \text{ then } \bar{X} = X - Z * s = 85 - 3 * 2.33 = 78.01.$$

Figure: Z - Score

References

Mezbahur Rahman, Deepak Sanjel, Han Wu. Statistics Introduction, Revised Printing

KendallHunt