Module 6 Basic Statistical Indicators

Chapter Three

Describing Data: Numerical Measures

GOALS

When you have completed this chapter, you will be able to:

ONE

Calculate the arithmetic mean, median, mode, weighted mean, and the geometric mean.

TWO

Explain the characteristics, uses, advantages, and disadvantages of each measure of location.

THREE

Identify the position of the arithmetic mean, median, and mode for both a symmetrical and a skewed distribution.

Chapter Three

Describing Data: Numerical Measures

FOUR

Compute and interpret the range, the mean deviation, the variance, and the standard deviation of ungrouped data.

FIVE

Explain the characteristics, uses, advantages, and disadvantages of each measure of dispersion.

SIX

Understand Chebyshev's theorem and the Empirical Rule as they relate to a set of observations.

The **Arithmetic Mean** is the most widely used measure of location and shows the central value of the data.

It is calculated by summing the values and dividing by the number of values.

The major characteristics of the mean are:

- **It requires the interval scale.**
- **All values are used.**
- **It is unique.**
- **The sum of the deviations from the mean is 0.**

For ungrouped data, the Population Mean is the sum of all the population values divided by the total number of population values:

$$\mu = \frac{\sum X}{N}$$

where

- μ is the population mean
- N is the total number of observations.
- X is a particular value.
- Σ indicates the operation of adding.

A **Parameter** is a measurable characteristic of a population.

The Banda family owns four cars. The following is the current mileage on each of the four cars.

56,000

42,000

23,000

73,000

Find the mean mileage for the cars.

$$\mu = \frac{\sum X}{N} = \frac{56{,}000 + \ldots + 73{,}000}{4} = 48{,}500$$

For ungrouped data, the **Sample Mean** is the sum of all the sample values divided by the number of sample values:

$$\overline{X} = \frac{\Sigma X}{n}$$

where *n* is the total number of values in the sample.

A **statistic** is a measurable characteristic of a sample.

- A sample of five chief executives received the following bonus last year ($000):

14.0,
15.0,
17.0,
16.0,
15.0

$$\overline{X} = \frac{\Sigma X}{n} = \frac{14.0 + \ldots + 15.0}{5} = \frac{77}{5} = 15.4$$

Properties of the Arithmetic Mean

1. Every set of interval-level and ratio-level data has a mean.
2. All the values are included in computing the mean.
3. A set of data has a unique mean (not multiple ones).
4. The mean is affected by unusually large or small data values.
5. The arithmetic mean is the only measure of location where the sum of the deviations of each value from the mean is zero.

Consider the set of values: 3, 8, and 4. The **mean** is 5. Illustrating the fifth property

$$\Sigma(X-\bar{X})=\left[(3-5)+(8-5)+(4-5)\right]=0$$

The **Weighted Mean** of a set of numbers $X_1, X_2, ..., X_n$, with corresponding weights $w_1, w_2, ..., w_n$, is computed from the following formula:

$$\overline{X}_w = \frac{(w_1X_1 + w_2X_2 + ... + w_nX_n)}{(w_1 + w_2 + ...w_n)}$$

During a one hour period on a hot Saturday afternoon a Kantemba boy Chris served fifty drinks. He sold five drinks for $0.50, fifteen for $0.75, fifteen for $0.90, and fifteen for $1.10. Compute the weighted mean of the price of the drinks.

$$\bar{X}_w = \frac{5(\$0.50)+15(\$0.75)+15(\$0.90)+15(\$1.15)}{5+15+15+15}$$

$$= \frac{\$44.50}{50} = \$0.89$$

Practical exercise

- A company has 150 employees. Out of them, 10 are in senior managerial positions, 25 are middle levels mangers, 40 are in supervisory positions, 65 are mechanical hands, and the rest 10 belong to line staff. The accounts book of the company shows their monthly bill as given below. Find the average monthly bill of the company.

Payroll statistics

Employee's category	Monthly bill(K)	
Snr managers (A)	120,000	
Middle level managers (B)	150,000	
Supervisory staff (C)	180,000	
Mechanical Hands(D)	195,000	
Line Staff(E)	25,000	

The **Median** is the midpoint of the values after they have been ordered from the smallest to the largest.

There are as many values above the median as below it in the data array.

For an even set of values, the median will be the arithmetic average of the two middle numbers and is found at the (n+1)/2 ranked observation.

Furthermore, for an even set of values, the median is not necessarily a value in the characteristic being measured.

The ages for a sample of five PanAfric workers are:

21, 25, 19, 20, 22.

Arranging the data in ascending order gives:

19, 20, 21, 22, 25.

Thus the median is 21.

The heights of four trade union researchers, in inches, are: 76, 73, 80, 75.

Arranging the data in ascending order gives:

73, 75, 76, 80

Thus the median is 75.5.

The median is found at the (n+1)/2 = (4+1)/2 =2.5th data point.

Properties of the Median

- There is a unique median for each data set.
- It is not affected by extremely large or small values and is therefore a valuable measure of location when such values occur.
- It can be computed for ratio-level, interval-level, and ordinal-level data.

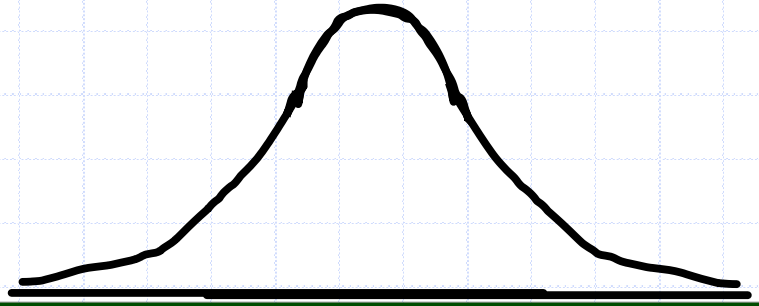
The **Mode** is another measure of location and represents the value of the observation that appears most frequently.

Example 6: The exam scores for ten students are: 81, 93, 84, 75, 68, 87, 81, 75, 81, 87. Because the score of 81 occurs the most often, it is the mode.

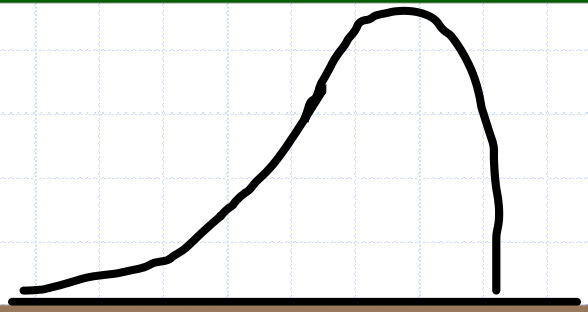
Data can have more than one mode. If it has two modes, it is referred to as bimodal, three modes, trimodal, and the like.

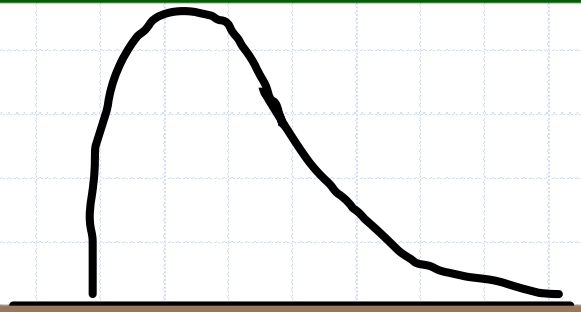
(Can data have NO mode?)

Symmetric distribution: A distribution having the same shape on either side of the center



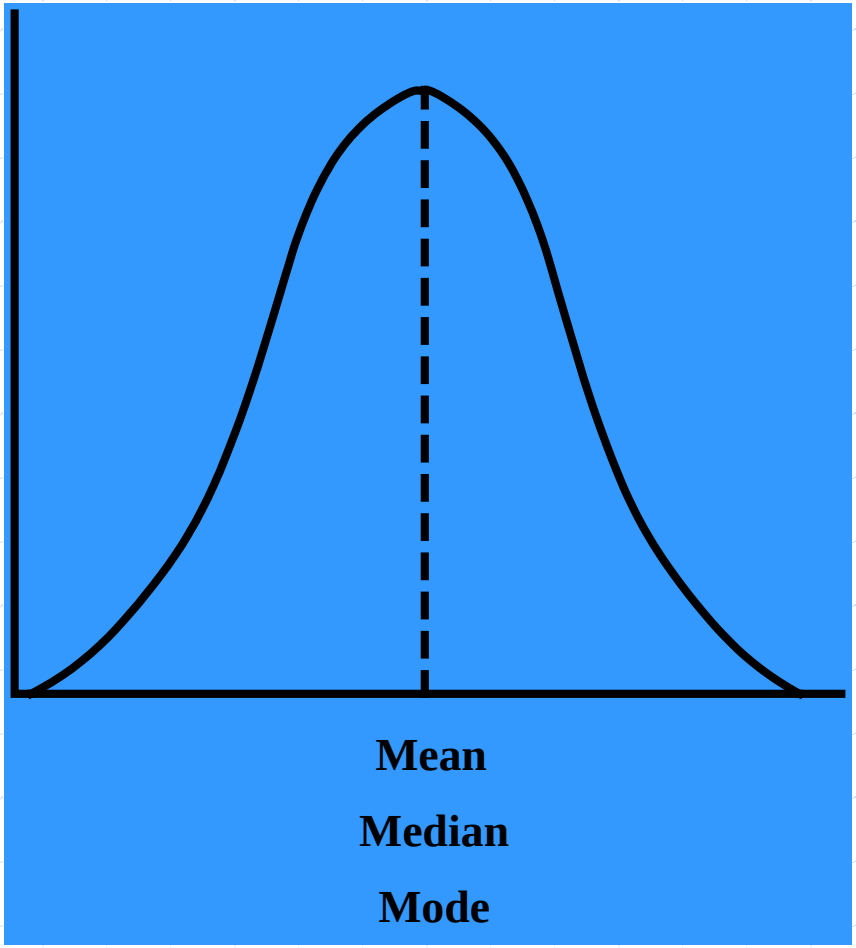
Skewed distribution: One whose shapes on either side of the center differ; a nonsymmetrical distribution.





Can be positively or negatively skewed, or bimodal

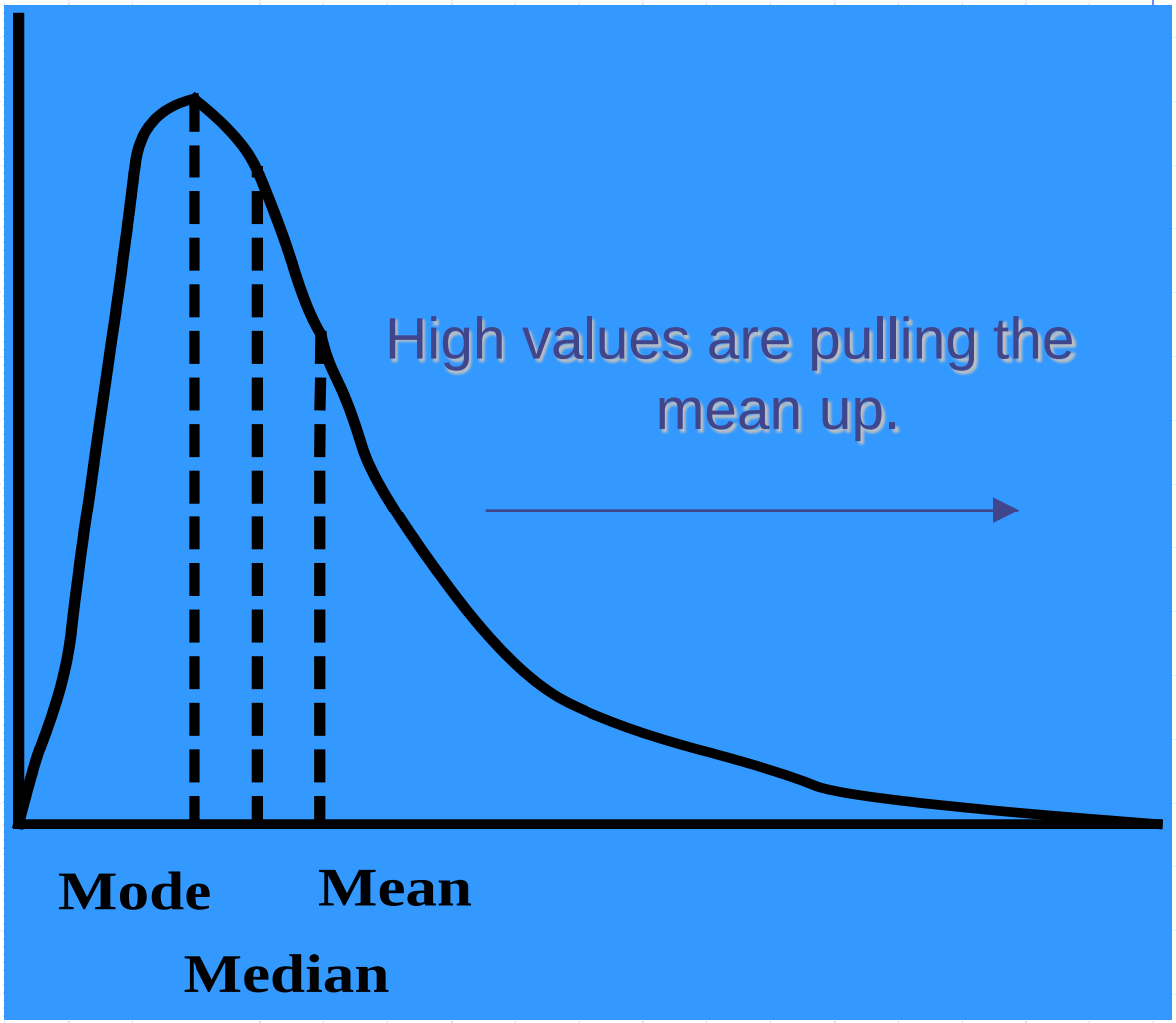
Zero skewness **Mean = Median = Mode**



The Relative Positions of the Mean, Median, and Mode: Symmetric Distribution

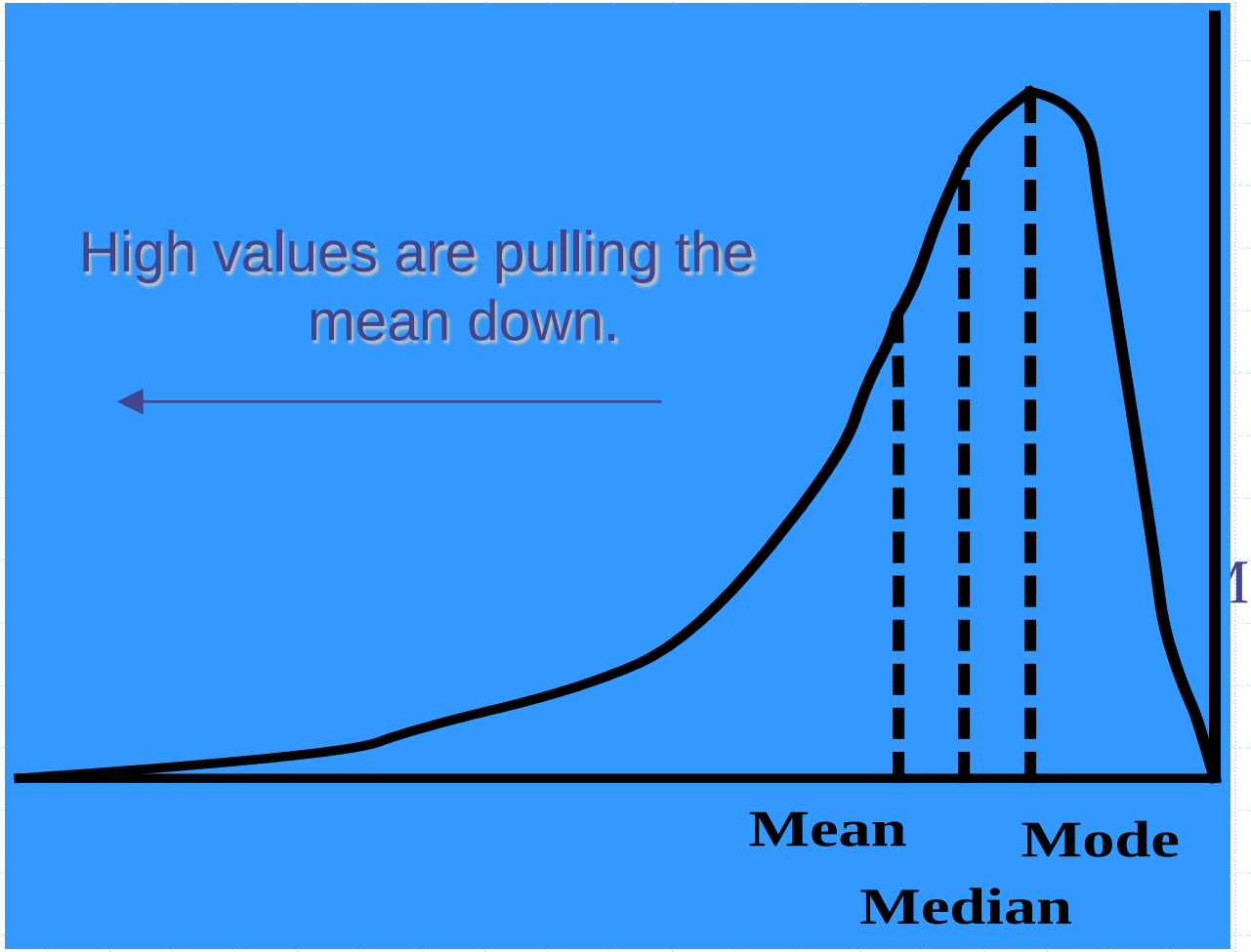
- **Positively skewed**: Mean and median are to the right of the mode.

Mean>Median>Mode



The Relative Positions of the Mean, Median, and Mode: Right Skewed Distribution

Negatively Skewed: Mean and Median are to the left of the Mode.



Mean<Median<Mode

The Relative Positions of the Mean, Median, and Mode: Left Skewed Distribution

The **Geometric Mean** (*GM*) of a set of n numbers is defined as the *nth* root of the product of the *n* numbers. The formula is:

$$GM = \sqrt[n]{(X_1)(X_2)(X_3)\ldots(X_n)}$$

The geometric mean is used to average percents, indexes, and relatives.

The interest rate on three bonds were 5, 21, and 4 percent.

The arithmetic mean is (5+21+4)/3 =10.0.

The geometric mean is

$$GM = \sqrt[3]{(5)(21)(4)} = 7.49$$

The *GM* gives a more conservative profit figure because it is not heavily weighted by the rate of 21percent.

Practical Exercise: Suppose you are offered two different pay raises
Which is better???

- You earn $3,000 per month as a starting salary, and you are offered two different alternative pay raises.

1. 10% this year, 20% next year
2. 15% this year, 15% next year
3. 20% this year, 10% next year
4. None of the above

The arithmetic mean of all BUT #4 is 15%, but is one better than the others?

Which to choose???

	Salary	Raise	Salary after Raise	
Period 1	3000	1.1	3300	300
Period 2	3300	1.2	3960	660
				960
Period 1	3000	1.15	3450	450
Period 2	3450	1.15	3967.5	517.5
				967.5
Period 1	3000	1.2	3600	600
Period 2	3600	1.1	3960	360
				960

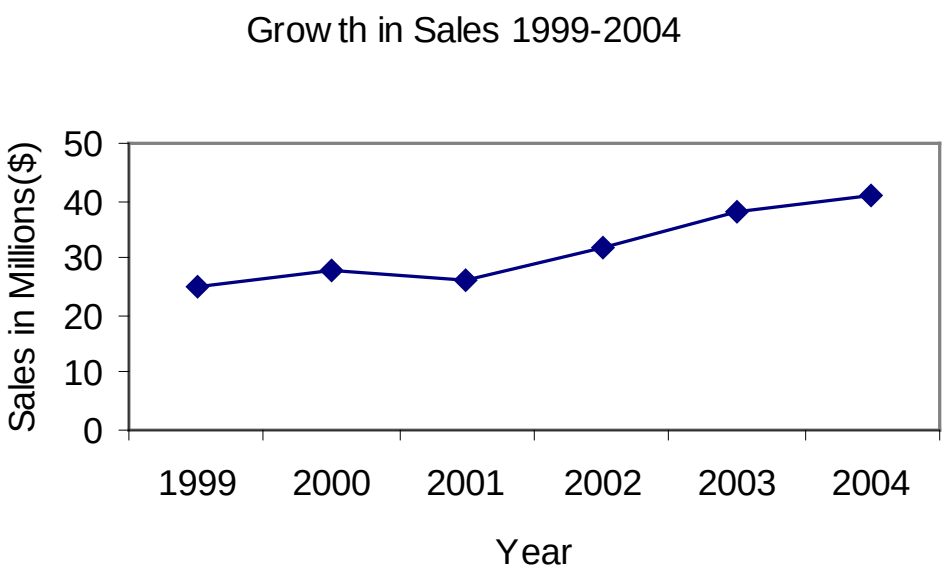
The only time the Geometric mean and the arithmetic mean are the same is when all the factors are the same.

$$GM = \sqrt{(1.1)(1.2)} = 1.14891 \text{ not } 1.15$$

Base		Average Raise	$ Raise
$	3,000.00	0.148912529	446.74
$	3,446.74	0.148912529	513.26
			960.00

Just as a check. Yep,15% is best.

Another use of the geometric mean is to determine the percent increase in sales, production or other business or economic series from one time period to another.



$$GM = \sqrt[n]{\frac{(\text{Value at end of period})}{(\text{Value at beginning of period})}} - 1$$

$$GM = \sqrt[5]{\frac{(41)}{(25)}} - 1 \quad = 10.4\%$$

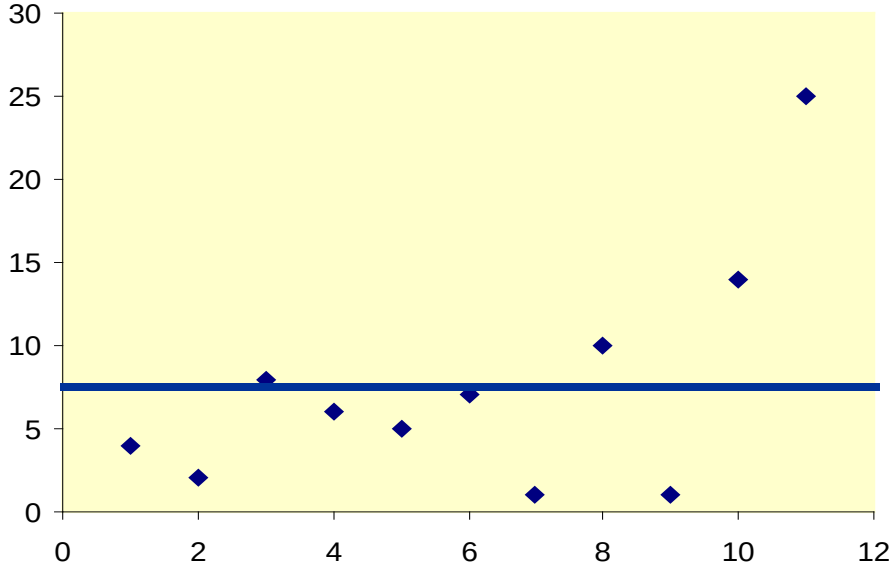
The total number of females enrolled in Kenyan colleges increased from 755,000 in 1992 to 835,000 in 2000. That is, the geometric mean rate of increase is 1.27%.

$$GM = \sqrt[8]{\frac{835,000}{755,000}} - 1 = .0127$$

So, keep in mind

- There are several measures of central tendency…
  - Mean
  - Median
  - Mode
  - Weighted mean
  - Geometric mean
- Know which are affected by single large values and which are not.
- Know which is appropriate for a given variable.
  - E.g., % changes, or anything that compounds requires the GM.

Dispersion refers to the spread or variability in the data.



Measures of dispersion include the following:

range, mean deviation, variance, and standard deviation.

Range = Largest value – Smallest value

The Range

- The Range is affected by extreme values (large or small).
- It does not use all the variable information.
- It's probably the simplest measure of dispersion and also the easiest to calculate.

- In Excel, you can calculate the range as follows:

= Max()-Min()

Where the "()" contains the list of data, usually the start and end points on a column.

The following represents the current year's Return on Equity of the 25 companies in an investor's portfolio.

-8.1	3.2	5.9	8.1	12.3
-5.1	4.1	6.3	9.2	13.3
-3.1	4.6	7.9	9.5	14.0
-1.4	4.8	7.9	9.7	15.0
1.2	5.7	8.0	10.3	22.1

Highest value: 22.1

Lowest value: -8.1

Range = Highest value – lowest value
= 22.1-(-8.1)
= 30.2

Mean Deviation

The arithmetic mean of the absolute values of the deviations from the arithmetic mean. A.K.A. MAD

The main features of the mean deviation are:

- All values are used in the calculation.
- It is not unduly influenced by large or small values.
- The absolute values are difficult to manipulate.

$$MD = \frac{\Sigma \left| X - \overline{X} \right|}{n}$$

The weights of a sample of crates containing books for the bookstore (in kgs) are:

103, 97, 101, 106, 103

Find the mean deviation.

$\overline{X} = 102$

The mean deviation is:

$$MD = \frac{\Sigma|X - \overline{X}|}{n} = \frac{|103-102| + \ldots + |103-102|}{5}$$

$$= \frac{1+5+1+4+1}{5} = 2.4$$

Variance: the arithmetic mean of the squared deviations from the mean.

Standard deviation: The square root of the variance.

The major characteristics of the Population Variance are:

- Always positive
- All values are used in the calculation.
- The units are awkward, the square of the original units.

Population Variance formula:

$$\sigma^2 = \frac{\Sigma (X - \mu)^2}{N}$$

X is the value of an observation in the population

μ is the arithmetic mean of the population

N is the number of observations in the population

Population Standard Deviation formula:

$$\sigma = \sqrt{\sigma^2}$$

Example of Variance and St. Dev.

Suppose we are trying to get an idea about how much our power bill varies each month. The following chart has ALL of our bills for the 20 months we have lived in our house. (let's assume it's a population)

	Amount for Power Bll	(X-Xbar)^2
	54	0.302
	48	42.903
	58	11.903
	50	20.703
	25	873.203
	47	57.003
	75	418.203
	46	73.102
	60	29.703
	70	238.703
	67	155.003
	68	180.903
	39	241.803
	35	382.203
	56	2.103
	66	131.103
	33	464.403
	62	55.503
	65	109.203
	67	155.003
Average	54.55	182.148 Pop Var

$$\frac{\sum (X-\mu)^2}{N}$$

$$\sqrt{\frac{\sum (X-\mu)^2}{N}}$$

Population SD 13.4962

The variance is 182.15 dollars squared

Sample variance (s^2)

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

Sample standard deviation (s)

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

The hourly wages earned by a sample of five workers are:

$7, $5, $11, $8, $6.

Find the sample variance and standard deviation.

$$\bar{X} = \frac{\Sigma X}{n} = \frac{37}{5} = 7.40$$

$$s^2 = \frac{\Sigma\left(X - \bar{X}\right)^2}{n-1} = \frac{(7-7.4)^2 + \ldots + (6-7.4)^2}{5-1}$$

$$= \frac{21.2}{5-1} = 5.30$$

$$s = \sqrt{s^2} = \sqrt{5.30} = 2.30$$

Chebyshev's theorem: For any set of observations, REGARDLESS OF THE UNDERLYING DISTRIBUTION, the minimum proportion of the values that lie within *k* standard deviations of the mean is at least:

$$1 - \frac{1}{k^2}$$

- where *k* is any constant greater than 1.

What **Chebyshev's theorem** means

- **Chebyshev's theorem** says that we can use a sample and from ANY population, no matter how weird it's distributed, and know something about how close our sample statistics are from our population parameters.

For example…

- If we now NOTHING about the shape of the distribution from which we draw a sample…

- **Chebyshev's theorem says that we do know that**

$$1-\frac{1}{2^2} \text{ or } 75\%$$

Lie within 2 standard deviations of the mean.

- If the sample is drawn from a symmetrically distributed population, we can be even more precise.

Empirical Rule: For any symmetrical, bell-shaped distribution:

- About 68% of the observations will lie within 1*s* the mean
- About 95% of the observations will lie within 2*s* of the mean
- Virtually all the observations will be within 3*s* of the mean

Bell - Shaped Curve showing the relationship between σ and μ.