Topic 4

Descriptive Statistics

Contents:

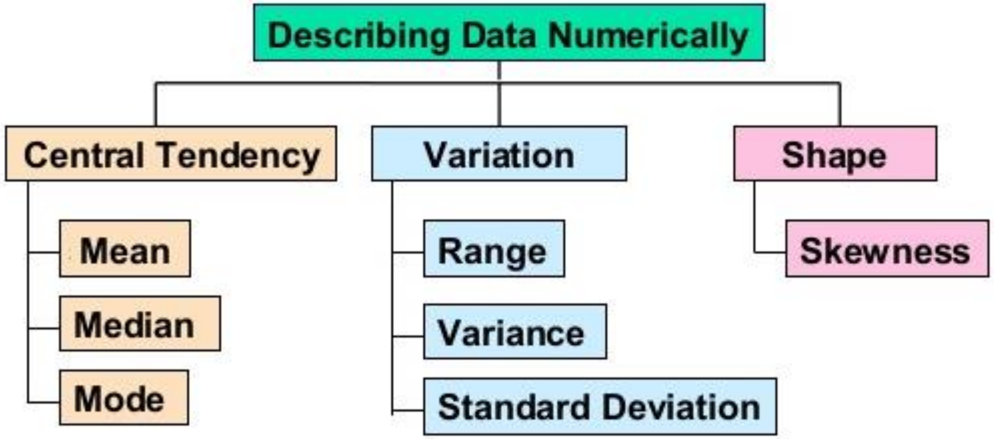
4.1 Introduction

4.2 Organizing data

4.3 Measurement of central tendency and dispersion

SUBTOPICS:

- 4.3.1 Measures of Central Tendency
- 4.3.2 Measures of Dispersion



Measures Summary		Raw Data			
		Ungrouped Data		Grouped Data	
		Sample	Population	Sample	Population
Measures of Central Tendency	Mean	$\bar{x} = \frac{\sum x}{n}$	$\mu = \frac{\sum x}{N}$	$\bar{x} = \frac{\sum mf}{\sum f}$	$\mu = \frac{\sum mf}{\sum f}$
	Median	1) Rank data in order 2) Position of Median = $\left(\frac{n+1}{2}\right)$ th 3) Find Median		1) Position of Median = $\frac{\sum f}{2}$ th 2) Median class 3) Median = $L + \left[\frac{\left[\frac{\sum f}{2}\right] - F_L}{f_m}\right]c$	
	Mode	1) Highest number of occurrences 2) Find Mode		1) Mode class (highest frequency) 2) Mode = $L + \left[\frac{f_m - f_B}{(f_m - f_B)+(f_m - f_A)}\right]c$	
Measures of Dispersion	Variance	$s^2 = \frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1}$	$\sigma^2 = \frac{\sum x^2 - \frac{(\sum x)^2}{N}}{N}$	$s^2 = \frac{\sum m^2 f - \frac{(\sum mf)^2}{\sum f}}{\sum f - 1}$	$\sigma^2 = \frac{\sum m^2 f - \frac{(\sum mf)^2}{\sum f}}{\sum f}$
	Standard Deviation	$s = \sqrt{s^2}$	$\sigma = \sqrt{\sigma^2}$	$s = \sqrt{s^2}$	$\sigma = \sqrt{\sigma^2}$

4.3.1 MEASURES OF CENTRAL TENDENCY

- A measure of central tendency is a measure that tells us where the middle of a bunch of data lies.
- There are three common measures of the central tendency, which are:

a) Mean

b) Mode

c) Median

4.3.1 MEASURES OF CENTRAL TENDENCY

4.3.1.1.1 Mean for Ungrouped Data

- Known as average

$$\text{Mean} = \frac{\text{Sum of all values}}{\text{Number of values}}$$

For **Sample** Data: $$\bar{x} = \frac{\sum x}{n}$$

For **Population** Data: $$\mu = \frac{\sum x}{N}$$

- The value of μ is constant but varies from sample to sample
- Outliers or extreme values are very small or very large relative to the majority of the values in a data set.

Example 1

Find the mean of the following sample of numbers:

28 36 49 20 17

Solution:

Number of values, $n = 5$

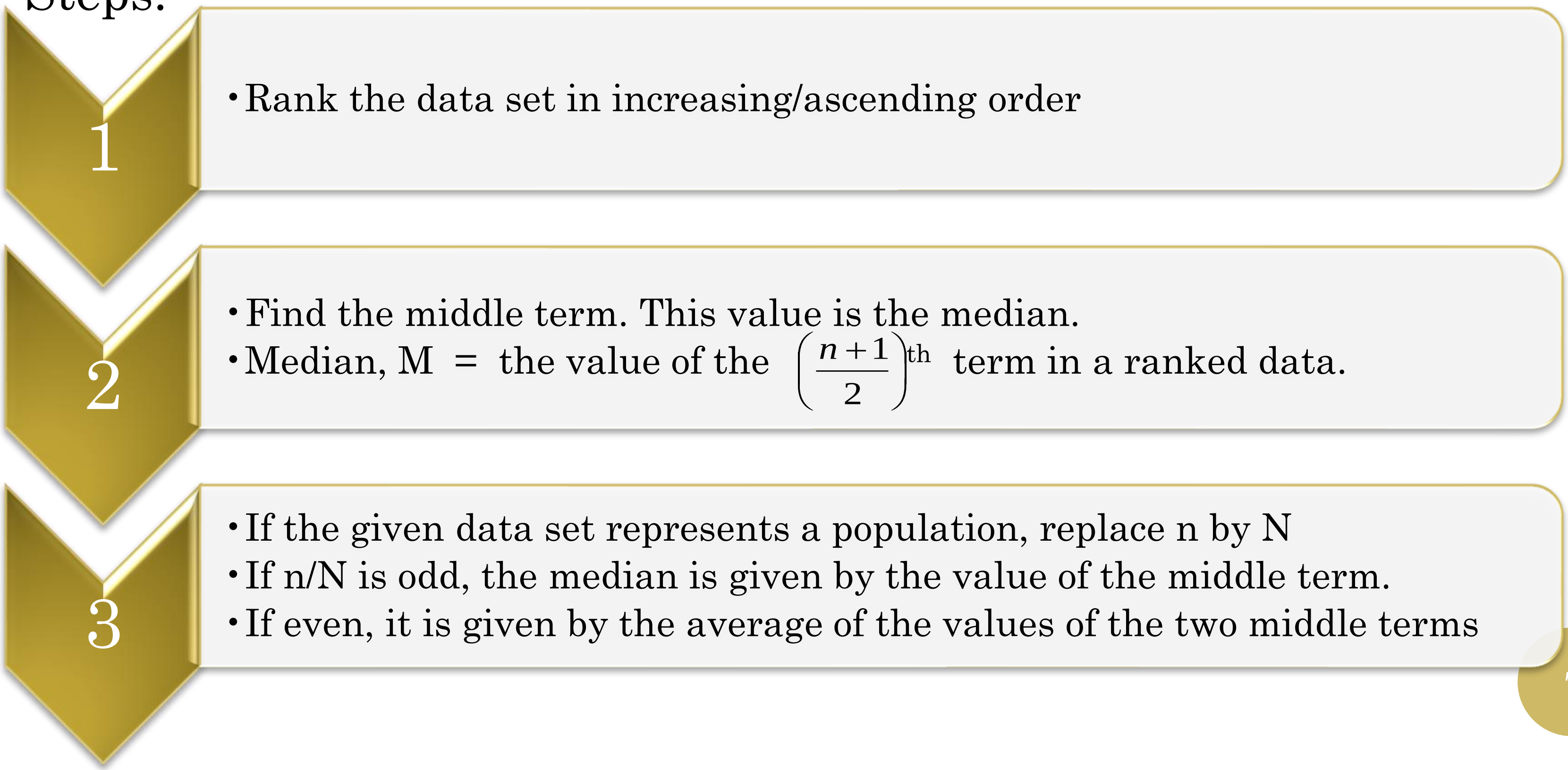
Thus, the mean, $\bar{x} = \frac{\sum x}{n} = \frac{28 + 36 + 49 + 20 + 17}{5} = 30$

4.3.1 MEASURES OF CENTRAL TENDENCY

4.3.1.1.2 Median for Ungrouped Data

- The value of the middle term in a data set that has been ranked in increasing/ascending order

Steps:



1

- Rank the data set in increasing/ascending order

2

- Find the middle term. This value is the median.
- Median, M = the value of the $\left(\frac{n+1}{2}\right)^{\text{th}}$ term in a ranked data.

3

- If the given data set represents a population, replace n by N
- If n/N is odd, the median is given by the value of the middle term.
- If even, it is given by the average of the values of the two middle terms

Example 2

Calculate the median of the following data:

250 300 180 240 290

Solution:

The data must be ranked in increasing order:

180 240 250 290 300

Thus, the median is 250

4.3.1 MEASURES OF CENTRAL TENDENCY

4.3.1.1.3 Mode for Ungrouped Data

- Mode is the value that has the **highest number of occurrence** or frequency in a data set
- **Uni**modal mode : **One** value occurring with the highest frequency
- **Bi**modal mode : **Two** values occurring with the highest frequency
- **Multi**modal mode: **More than 2** values.

Example 3

Find the mode of the following data

1 2 6 2 1 2 7 5

Solution: 2

4.3.1 MEASURES OF CENTRAL TENDENCY

4.3.1.2.1 Mean for Grouped Data

For **Sample** Data: $\bar{x} = \frac{\sum mf}{n} = \frac{\sum mf}{\sum f}$

For **Population** Data: $\mu = \frac{\sum mf}{N} = \frac{\sum mf}{\sum f}$

where m= the midpoint of the class $= \frac{\text{lower limit} + \text{upper limit}}{2}$

$\sum f =$ total frequency of the class

Example 4

The following table gives the frequency distribution of the age of all 100 staffs in FCI, MMU.

Age in years	Number of employees
20 to less than 30	45
30 to less than 40	24
40 to less than 50	21
50 to less than 60	10

Calculate the average age of all these 100 staffs.

Solution:

Age in years	f	m	mf
20 - < 30	45	25	1125
30 - < 40	24	35	840
40 - < 50	21	45	945
50 - < 60	10	55	550
	N=100		$\sum mf = 3460$

$$\mu = \frac{\sum mf}{N} = \frac{3460}{100} = 34.6$$

4.3.1 MEASURES OF CENTRAL TENDENCY

4.3.1.2.2 Median for Grouped Data

Steps:

1

- Determine the median class by calculating the position of the median, $w = \frac{\sum f}{2}$

2

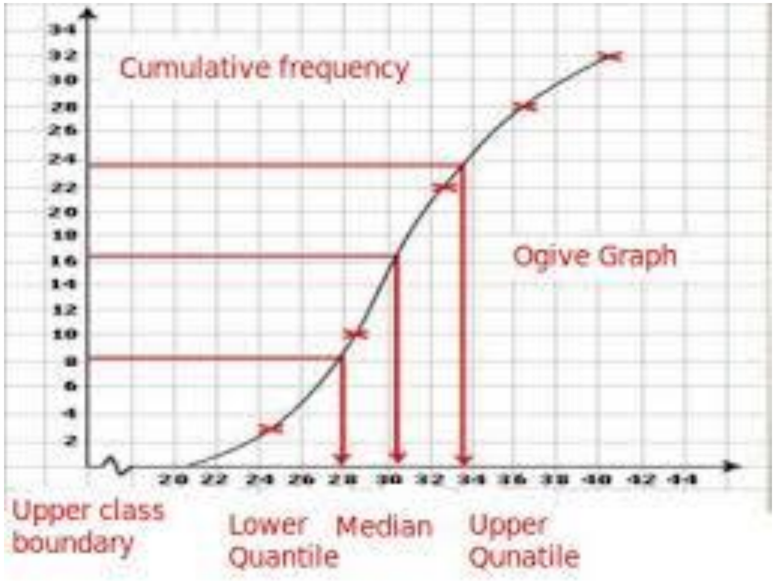
- Obtain median using the formula: $L + \left[\frac{\left[\frac{\sum f}{2} \right] - F_L}{f_m} \right] c$

where L = Lower boundary of median class
$\sum f$ = Total frequencies
F_L = Total frequencies for all classes before median class
f_m = Frequency of median class
c = Class width of median class

4.3.1 MEASURES OF CENTRAL TENDENCY

4.3.1.2.2 Median for Grouped Data

The median may also be estimated from an ogive.



4.3.1 MEASURES OF CENTRAL TENDENCY

4.3.1.2.3 Mode for Grouped Data

- Mode = $$L+\left[\frac{f_m-f_B}{(f_m-f_B)+(f_m-f_A)}\right]c=L+\left[\frac{\Delta_B}{\Delta_B+\Delta_A}\right]c$$

where

L = Lower boundary of the mode class

f_m = Frequencies of the mode class

f_B = Frequency of 1 class before the mode class

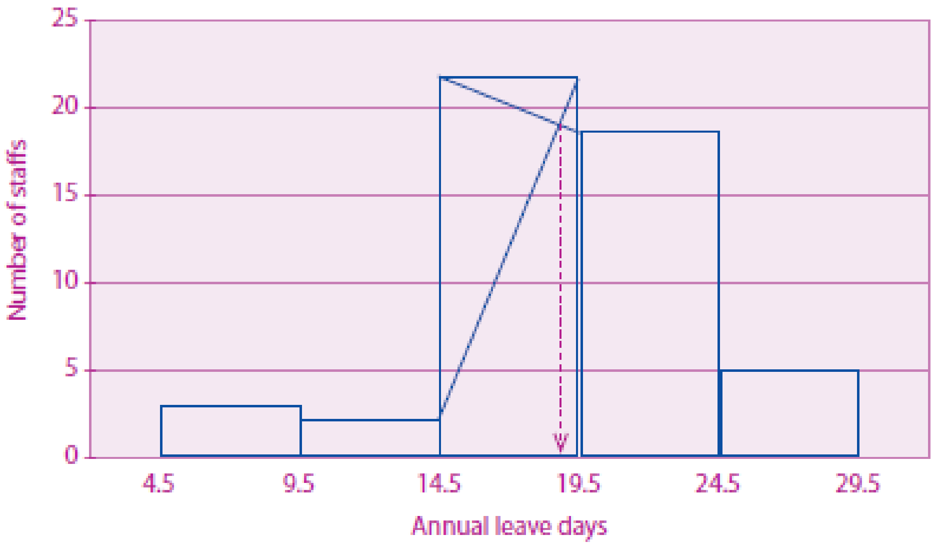
f_A = Frequency of 1 class after the mode class

c = Class width of the mode class

4.3.1 MEASURES OF CENTRAL TENDENCY

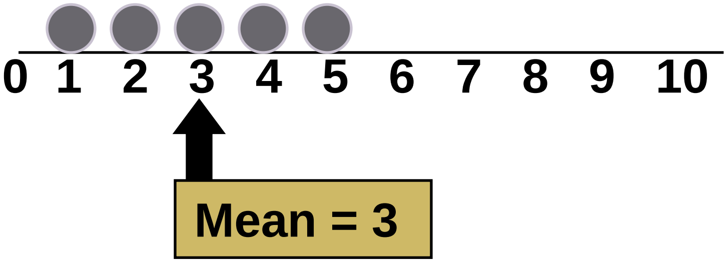
4.3.1.2.3 Mode for Grouped Data

The mode may also be estimated from a histogram.

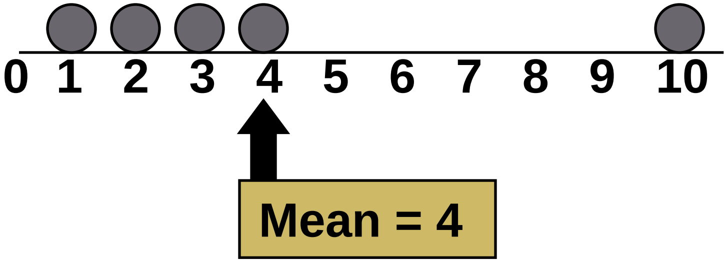


WHAT IS THE BEST MEASURE OF CENTRAL TENDENCY?

Consider this set of test score values:



$$\frac{1+2+3+4+5}{5} = \frac{15}{5} = 3$$



$$\frac{1+2+3+4+10}{5} = \frac{20}{5} = 4$$

The set on the left shows the actual scores. The set on the right shows what would happen if one of the scores was WAY out of range in regard to the other scores. Such a term is called an **outlier.**

With the outlier, the mean changed.

With the outlier, the median did NOT change.

The median is preferred in this situation as the value of the mean can be distorted by the outliers. However, it will depend on how influential the outliers are. If they do not significantly distort the mean then using the mean as the measure of central tendency will usually be preferred because it includes every value in of the data set as part of the calculation.

HOW DO I KNOW WHICH MEASURE OF CENTRAL TENDENCY TO USE?

MEAN

Use the mean to describe the middle of a set of data that *does not* have an outlier.

Advantages:

- Most popular measure in fields such as business, engineering and computer science.
- It is unique - there is only one answer.
- Useful when comparing sets of data.

Disadvantages:

- Affected by extreme values (outliers)

MEDIAN

Use the median to describe the middle of a set of data that *does* have an outlier.

Advantages:

- Extreme values (outliers) do not affect the median as strongly as they do the mean.
- Useful when comparing sets of data.
- It is unique - there is only one answer.

Disadvantages:

- Not as popular as mean.

MODE

Use the mode when the data is non-numeric or when asked to choose the most popular item.

Advantages:

- Extreme values (outliers) do not affect the mode.

Disadvantages:

- Not as popular as mean and median.
- Not necessarily unique - may be more than one answer
- When no values repeat in the data set, the mode is every value and is useless.
- When there is more than one mode, it is difficult to interpret and/or compare.

4.3.2 MEASURES OF DISPERSION

- A measure of dispersion is a measure that tells us how the data is spread out, or dispersed about the mean.
- There are three measures of the dispersion, which are:

a) range

b) variance

c) standard deviation

4.3.2 MEASURES OF DISPERSION

4.3.2.1.1 Range for Ungrouped Data

- The simplest measure of dispersion.
- It is the difference between the largest and the smallest value of observations in a data set.
- **Range = Largest value - Smallest value**
- Disadvantage: Influenced by outliers.

Example 5

The following data are the results of PMT0301 final examination for 6 students:

Students	Exam marks
Kelvin	80
Fatimah	67
Jason	10
Teoh	57
Justin	92
Mimi	79

Solution:

From the given data,

the highest value is 92 and

the lowest value is 10.

Thus, range = **92-10 = 82.**

4.3.2 MEASURES OF DISPERSION

4.3.2.1.2.1 Variance and Standard Deviation for Ungrouped Data

- Standard deviation is the most commonly used.
- It measures **how closely the values** of the data are gathered **around the mean**.
- If the value of the standard deviation is **small** in number, then it indicates that the spreads of the values of the observations are **relatively smaller** range around the mean.
- If the value of the standard deviation is **large** in number, then it indicates that the spreads of the values of the observations are **relatively bigger** range around the mean.

4.3.2 MEASURES OF DISPERSION

4.3.2.1.2 Variance and Standard Deviation for Ungrouped Data

For Population Data

Variance, $$\sigma^2 = \frac{\sum(x-\mu)^2}{N}$$

$$= \frac{\sum x^2 - \frac{(\sum x)^2}{N}}{N} \quad \text{or} \quad \frac{\sum x^2}{N} - \left(\frac{\sum x}{N}\right)^2$$

For Sample Data

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

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

Note:

☞ The population standard deviation $= \sqrt{\sigma^2} = \sigma$

☞ The sample standard deviation $= \sqrt{s^2} = s$

☞ Thus, the value of the standard deviation is **never negative**.

Example 6

The following data lists the numbers of hours spent per week by 5 randomly selected students to surf internet.

103 7 5 4

Calculate the variance and standard deviation for these data.

Solution:

Let x the numbers of hours spent per week to study Statistics course:

x	x^2
10	100
3	9
7	49
5	25
4	16
$\sum x = 29$	$\sum x^2 = 199$

$$s^2 = \frac{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}{n-1} = \frac{199 - \frac{29^2}{5}}{5-1} = \frac{199-168.2}{4} = 7.7$$

$$s = \sqrt{7.7} = 2.77$$

Example 7

Data below is the 2003 earnings (in millions of dollars) of ten celebrities. Find the variance and standard deviation for these data.

55	267	175	47	125
366	156	501	125	142

Solution: Let x the earnings (in millions of dollars) of ten celebrities:

x	x^2
55	3025
366	133956
267	71289
156	24336
175	30625
501	251001
47	2209
125	15625
125	15625
142	20164
$\sum x = 1959$	$\sum x^2 = 567855$

$$s^2 = \frac{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}{n-1} = \frac{567855 - \frac{1959^2}{10}}{10-1} = 20454.1$$

$$s = \sqrt{20454.1} = 143.02$$

4.3.2 MEASURES OF DISPERSION

4.3.2.1.2.2 Variance and Standard Deviation for Grouped Data

For Population Data

Variance, $$\sigma^2 = \frac{\sum m^2 f - \frac{(\sum mf)^2}{N}}{N}$$

For Sample Data

Variance, $$s^2 = \frac{\sum m^2 f - \frac{(\sum mf)^2}{n}}{n-1}$$

Note:

☞ The population standard deviation $= \sqrt{\sigma^2} = \sigma$

☞ The sample standard deviation $= \sqrt{s^2} = s$

☞ Thus, the value of the standard deviation is **never negative**.

Example 8

The following table gives the frequency distribution of the age of all 100 lecturers in a college. Calculate the variance and standard deviation for given data.

Age in years	Number of employees
20 to less than 30	45
30 to less than 40	24
40 to less than 50	21
50 to less than 60	10

Solution:

Age in yr.	F	*m*	*mf*	m^2	$m^2 f$
20 - < 30	45	25	1125	625	28125
30 - < 40	24	35	840	1225	29400
40 - < 50	21	45	945	2025	42525
50 - < 60	10	55	550	3025	30250
Sum			3460		130300

$$\sigma^2 = \frac{\sum m^2 f - \frac{(mf)^2}{N}}{N} = \frac{130300 - \frac{3460^2}{100}}{100} = \frac{130300 - 119716}{100} = 105.84$$

$$\sigma = \sqrt{105.84} = 10.29$$

The Shape of the Distribution and the Relationships among the Mean, Median and Mode

a) Skewed to the Left

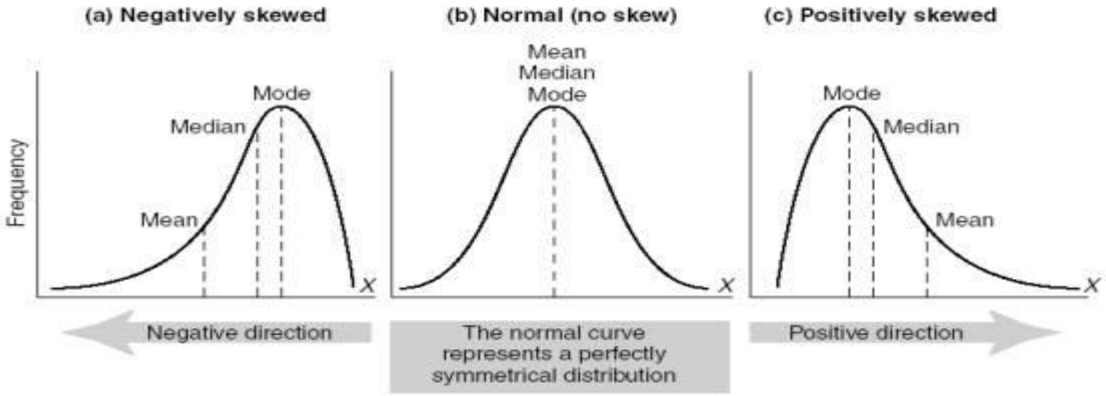
- Known as negatively skewed graph.
- Mean has the smallest value and the largest value is mode

b) Symmetry

- Mean, median and mode are identical (all have the same values).
- Lied at the center of the distribution

c) Skewed to the Right

- Known as positively skewed graph.
- Mean is the largest and the smallest value is mode


(a) Negatively skewed
(b) Normal (no skew)
(c) Positively skewed
Frequency
Mode
Median
Mean
X
Mean
Median
Mode
X
Mode
Median
Mean
X
Negative direction
The normal curve represents a perfectly symmetrical distribution
Positive direction