Summarization of data
(Measures of central location)

Professor Narjis A-H Ajeel

Learning objectives

At the end of the lecture you should be able to:

1. Calculate and interpret the following measures of central location:
   -arithmetic mean
   -median
   -mode
   -geometric mean
2. Choose and apply the appropriate measure of central location.

Summarization of Data

Two types of summary measures are used when describing data (distribution):

1. **Measures of central location**
2. **Measures variability or spread**

Measures of Central Location

- Measures of central location are numbers that tend to cluster around the “middle” of a set of values.
- They are also known as measures of central tendency

1.Measures of Central Location

- They are numbers that tend to cluster around the "middle" of a set of values.
- They are also known as measures of **central tendency**

1.The mean (arithmetic mean)

The mean is the average of all the data values in a distribution

$$\text{Mean} = \frac{\text{Sum of the observations}}{\text{Number of observations}}$$

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

Example (1)

The reported time on the Internet of 10 adults are 0, 7, 12, 5, 33, 14, 8, 0, 9, 22 hours/week. Find the mean time on the Internet.

$$\bar{x} = \frac{\sum x}{n} = \frac{0+7+\ldots+22}{10} = 11$$

Population mean = μ (mu)

The Mean of Grouped Data

- The mean of a sample of data organized in a frequency distribution is computed by the following formula:

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

➢ Where:

- $\sum fx$ is the sum of the product of *X* times the frequency
- Σf is the sum of the frequencies = *n* (the sample size)

Example (2):
Calculation of the mean number of previous pregnancies of a group of women attending the antenatal clinic:

Number of Previous pregnancies	Frequency (f)	fx
0	18	0
1	32	32
2	27	54
3	17	51
4	6	24
	$\sum f=100$	$\sum fx=161$

$\bar{x} = \frac{\sum fx}{n}$ $=\sum fx/\sum f=161/100=1.61$

Example 3:Calculation of the mean age of females.

Age(years)	No. of females(f)	Midpoint (X)	fX
10-	1	12.5	12.5
15-	7	17.5	122.5
20-	14	22.5	630
25-	28	27.5	770
30-	41	32.5	1332.5
35-	30	37.5	1125
40-	15	42.5	637.5
45-	18	47.5	855
50-	10	52.5	525
55-60	2	57.5	115
Total	∑f=n=166		∑fX=6125

—Mean = ∑ f x / ∑ f = 6125 /166 = 36.898 years

Where:

- *X* is the class midpoint
- Σf*X*, is the sum of the product of class interval midpoint times the class frequency
- Σ*f* is the sum of the class frequencies = *n* (the sample size)

Properties of the mean

1. The most commonly used measure of central location
2. Uses every value (uses all observations)
3. For each set of data there is only one mean.
4. Influenced (affected) by extreme values (high and low)

2. The Median

The Median is the "middle" value when the observations are arranged in ascending or descending order.

Steps for finding the median

1. Arrange observations in ascending or descending order

2. Find the position of the median:

$$\text{Median position} = \frac{n + 1}{2}$$

- If n is odd, the median is the middle observation
- If n is even, the median is the average of the two middle observations

Example

Find the median of the time on the internet for the 10 adults of the example

$$\text{Median position} = \frac{10+1}{2} = 5.5$$

Even number of observations

8.5

0, 0, 5, 7, [8, 9], 12, 14, 22, 33

Suppose only 9 adults were
Sampled exclude, for example,
the longest time (33)

Median position $= \frac{9+1}{2}$

$= 5$

Odd number of observations

0, 0, 5, 7, [8] , 9, 12, 14, 22

The Median of Grouped Data

- **The median of a sample of data organized in a frequency distribution is computed by the following formula:**

$$Median = L + \frac{n/2 - CF}{f}.w$$

- **Where:**
  - ***L* is the lower limit of the median class**
  - ***n* is the sample size**
  - ***CF* is the cumulative frequency preceding the median class**
  - ***f* is the frequency of the median class**
  - **w is the width of the class interval.**

Example 1.

Number of Previous pregnancies	Frequency (f)	Cumulative frequency
0	18	18
1	30	48
2	29	77
3	17	94
4	6	100
	∑f=100	

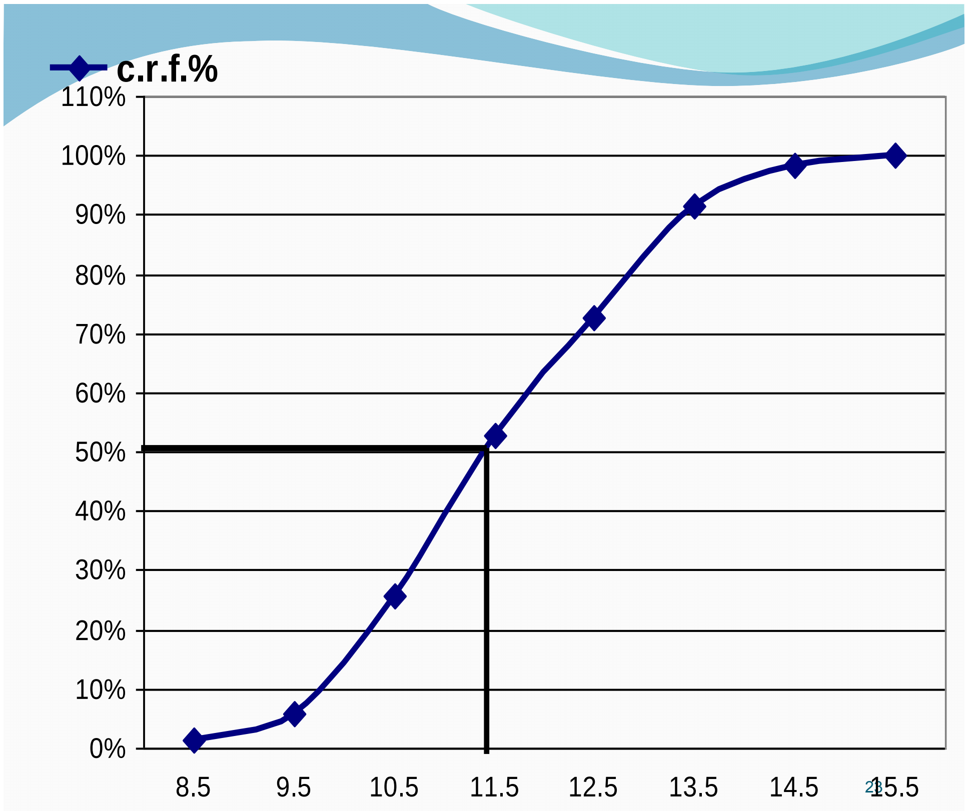
19

- ☐ ***The position of the median=n+1/*2=50.5**
- ☐ **Therefore the median is= 2, since observations 50 and 51 are =2.**

Example 2.

Age(years)	No. of females(f)	Cumulative frequency
10-	1	1
15-	7	8
20-	14	22
25-	28	50
30-	41	91
35-	30	121
40-	15	136
45-	18	154
50-	10	164
55-60	2	166
Total	∑f=n=166	

- ☐ **The median class is 30 – , since it contains the 83rd and the 84th values (*n+1/*2=83.5).**
- ☐ **From the table, *L* = 30, *n* = 166, *f* =41 , *CF* =50, w=5.**
- ☐ **Thus, the median** $= 30 + \left(\frac{166/2 - 50}{41}\right)(5) = 34.024\,years$

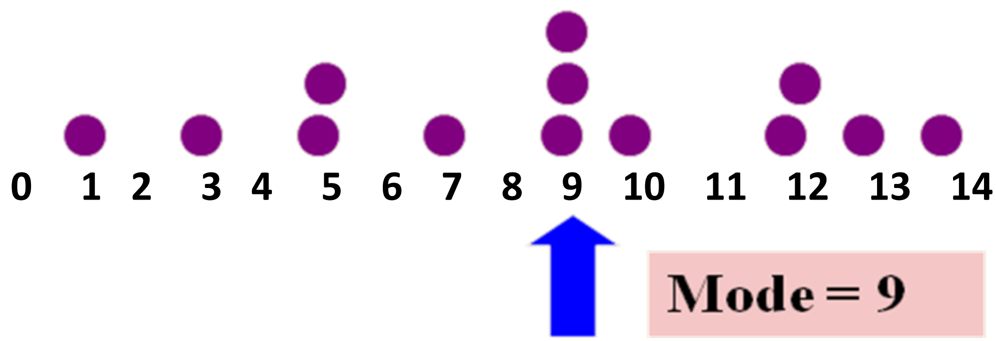


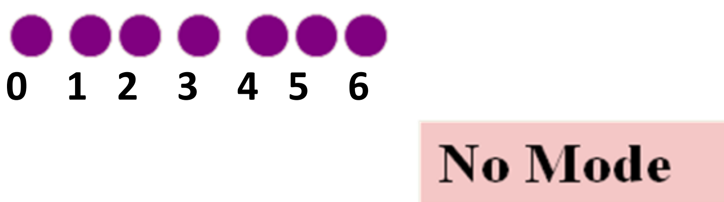
Properties of the median

1. It divides the observations into two equal halves (50% of the observations above and 50% below the median)
2. Uses only one or two values
3. For each set of data there is only one median.
4. It is not affected by extreme values

3. The Mode

The mode is the most frequently occurring observation (value) .



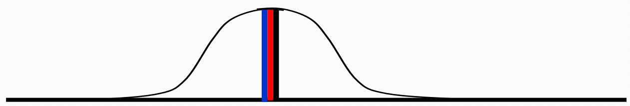


Properties of the mode

1. It is not affected by extreme values
2. A set of data may have no mode , one mode , two modes or more .

Relationship among Mean, Median, and Mode

- If a distribution is symmetrical, the mean, median and mode coincide



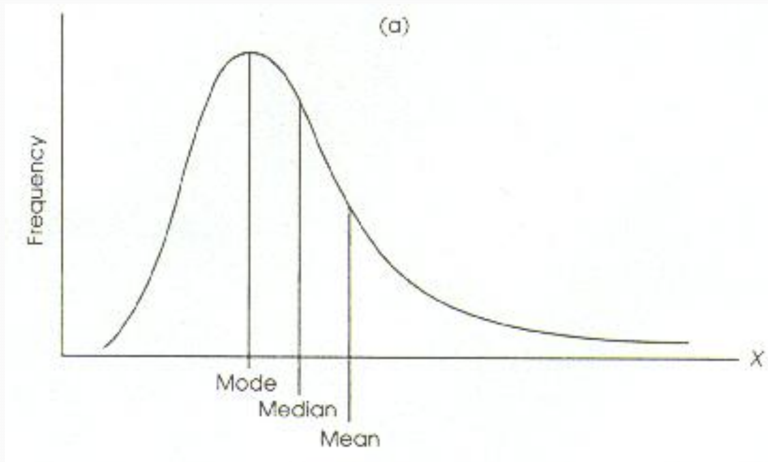
- If a distribution is non symmetrical, and skewed to the left or to the right, the three measures differ.

Types of distributions

Positively (Right) skewed distribution

Positively skewed: Mean and Median are to the right of the Mode

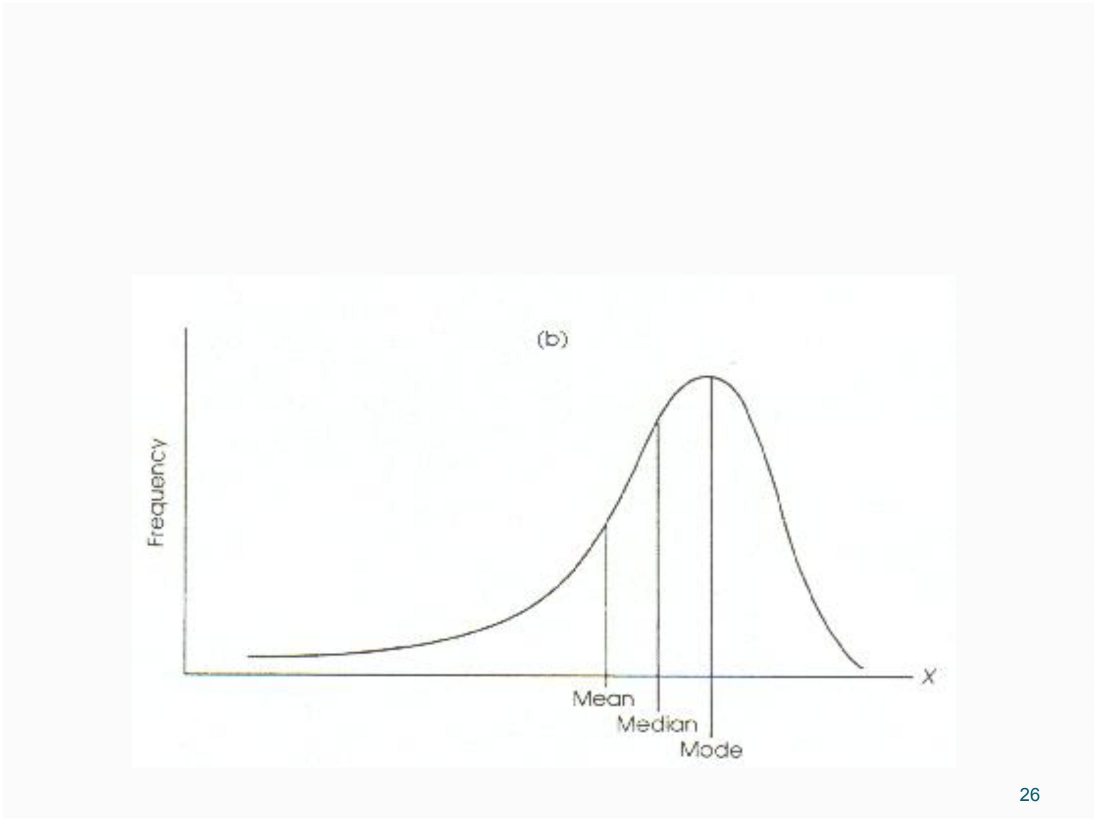
— **Mode < Median < Mean**



Negatively (Left) skewed distribution

Negatively skewed: Mean and Median are to the left of the Mode

Mean < Median < Mode



The Geometric Mean:

Definition: The geometric mean (GM) of a set of *n* numbers is defined as the *nth* root of the product of *n* numbers. The formula for the geometric mean is given by:

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

i

Geometric mean is also defined as the antilog of the mean of the logs x_i

The geometric mean:

-.Take the logarithm of each score

- Average the log values (mean of the logs).

- Calculate the antilog.

$$Geometricmean = anti\log\frac{\log X_1 + \log X_2 + \ldots + \log X_n}{n}$$

Sample	**E. coli. counts**	**Log**
1	160	2.20
2	700	2.85
3	60	1.78
4	12000	3.51
Average		**2.58**5

- **Antilog**
  = 10^2.585
  = 764.1
  - The geometric mean is 764.1 cells/ 100 ml