Measures of location and dispersion

Part 1
THE MEAN VALUES

CHAPTER QUESTIONS

1. Measures of location
2. Types of means
3. Measures of location for ungrouped data

- Arithmetic mean
- Harmonic mean
- Geometric mean
- Median and Mode

4. Measures of location for grouped data

- Arithmetic mean
- Harmonic mean
- Geometric mean
- Median and Mode

Measures of location and dispersion

- Properties to describe numerical data:
  - Central tendency
  - Dispersion
  - Shape
- Measures calculated for:
  - Sample data
    - Statistics
  - Entire population
    - Parameters

Measures of location and dispersion

Measures of location include:

- Arithmetic mean
- Harmonic mean
- Geometric mean
- Median
- Mode

Grouped and Ungrouped

UNGROUPED or raw data refers to data as they were collected, that is, before they are summarised or organised in any way or form

GROUPED data refers to data summarised in a frequency table

What is the mean?

- The **mean** - is a general indicator characterizing the typical level of varying trait per unit of qualitatively homogeneous population.

- Statistics derive the formula of the means of the formula of mean exponential:

$$\overline{X} = \sqrt[Z]{\frac{\sum X^Z}{n}}$$

- We introduce the following definitions
- - X-bar - the symbol of the mean
- X1, X2...Xn – measurement of a data value
- f- frequency of a data values;
- n – population size or sample size.

- There are the following types of mean:
- If $z = -1$ - the harmonic mean,
- $z = 0$ - the geometric mean,
- $z = +1$ - arithmetic mean,
- $z = +2$ - mean square,
- $z = +3$ - mean cubic, etc.

- The higher the degree of z, the greater the value of the mean. If the characteristic values are equal, the mean is equal to this constant.
- There is the following relation, called the rule the majorizing mean:

$$\bar{x}_{harm} \leq \bar{x}_{geom} \leq \bar{x}_{arith} \leq \bar{x}_{sq}$$

There are two ways of calculating mean:

- **for ungrouped data -**

  is calculated as a simple mean

- **for grouped data -**

  is calculated weighted mean

Types of means

Mean	Formula	
	for ungrouped data - **simple**	for grouped data – **weighted**
Harmonic mean ($xf = M$)	$\bar{x} = \dfrac{n}{\sum \frac{1}{x_i}}$	$\bar{x} = \dfrac{\sum M_i}{\sum \frac{M_i}{x_i}}$
Geometric mean	$\bar{x} = \sqrt[n]{\Pi(x_i)}$	$\bar{x} = \sqrt[\sum f_i]{\Pi(x_i^{f_i})}$
Arithmetic mean	$\bar{x} = \dfrac{\sum x_i}{n}$	$\bar{x} = \dfrac{\sum x_i f_i}{\sum f_i}$

Arithmetic mean

Arithmetic mean value is called the mean value of the sign, in the calculation of the total volume of which feature in the aggregate remains unchanged

Characteristics of the arithmetic mean

The arithmetic mean has a number of mathematical properties that can be used to calculate it in a simplified way.

1. If the data values (Xi) to reduce or increase by a constant number (A), the mean, respectively, decrease or increase by a same constant number (A)

$$\frac{\sum (x_i \pm A) f_i}{\sum f_i} = \frac{\sum x_i f_i}{\sum f_i} \pm \frac{A \sum f_i}{\sum f_i} = \bar{x} \pm A$$

- 2. If the data values (Xi) divided or multiplied by a constant number (A), the mean decrease or increase, respectively, in the same amount of time (this feature allows you to change the frequency of specific gravities - relative frequency):
- a) when divided by a constant number:

$$\frac{\sum \frac{x_i}{A} f_i}{\sum f_i} = \frac{\frac{1}{A}\sum x_i f_i}{\sum f_i} = \frac{1}{A}\bar{x} = \frac{\bar{x}}{A}$$

- b) when multiplied by a constant number:

$$\frac{\sum xAf_i}{\sum f_i} = \frac{A\sum x_i f_i}{\sum f_i} = A \cdot \bar{x}$$

- 3. **If the frequency divided by a constant number, the mean will not change:**

$$\frac{\sum x_i \frac{f_i}{A}}{\sum \frac{f_i}{A}} = \frac{\frac{1}{A}\sum x_i f_i}{\frac{1}{A}\sum f_i} = \frac{\sum x_i f_i}{\sum f_i} = \bar{x}$$

- 4. Multiplying the mean for the amount of frequency equal to the sum of multiplications variants on the frequency:
- If $$\bar{x} = \frac{\sum x_i f_i}{\sum f_i}$$
- then the following equality holds:

$$\bar{x} \sum f_i = \sum x_i f_i$$

5. The sum of the deviations of the number in a data value from the mean is zero:

$$\sum (x_i - \bar{x}) = 0$$

- If $\sum x_i f_i = \bar{x} \sum f_i$
- then $\sum x_i f_i - \bar{x} \sum f_i = 0$
- So $\sum x_i f_i - \bar{x} \sum f_i = \sum (x_i - \bar{x}) f_i = 0$

Measures of location for ungrouped data

- In calculating summary values for a data collection, the best is to find a central, or typical, value for the data.
- More important measures of central tendency are presented in this section:
- **Mean (simple or weighter)**
- **Median and Mode**

Measures of location for ungrouped data

ARITHMETIC MEAN

- This is the most commonly used measure.
- The arithmetic mean is a summary value calculated by summing the numerical data values and dividing by the number of values

$$\text{Sample mean} = \frac{\text{sum of sample observations}}{\text{number of sample observations}}$$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Sample size

Measures of location for ungrouped data

ARITHMETIC MEAN

- This is the most commonly used measure and is also called the mean.

Population mean = $\dfrac{\textbf{sum of observations}}{\textbf{number of observations}}$

Mean

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

X_i = observations of the population

$\sum$ = "the sum of"

Population size

Example - The sales of the six largest restaurant chains are presented in table

Company	Sales ($ million)
McDonald's	14.110
Burger King	5.590
Kentucky Fried Chicken	3.700
Hardee's	3.030
Wendy's	2.800
Pizza Hut	2.450

A mean sales amount of 5.280 $ million is computed using Equation of arithmetical mean simple

$$\bar{x} = \frac{14100 + 5590 + 3700 + 3030 + 2800 + 2450}{6} = 5280$$

MEDIAN for ungrouped data

- The <u>median</u> of a data is the middle item in a set of observation that are arranged in order of magnitude.
- The <u>median</u> is the measure of location most often reported for annual income and property value data.
- A few extremely large incomes or property values can inflate the mean.

Characteristics of the median

- **MEDIAN**
  - Every ordinal-level, interval-level and ratio-level data set has a median
  - The median is not sensitive to extreme values
  - The median does not have valuable mathematical properties for use in further computations
  - Half the values in data set is smaller than median.
  - Half the values in data set is larger than median.
  - Order the data from small to large.

Position of median

- If ***n*** is odd:
  - Median item number = $(n+1)/2$
- If ***n*** is even:
  - Calculate $(n+1)/2$
  - The median is the average of the values before and after $(n+1)/2$.

Example

- The median number of people treated daily at the emergency room of St. Luke's Hospital must be determined from the following data for the last six days: **25, 26, 45, 52, 65, 78**

  Since the data values are arranged from lowest to highest, the median be easily found. If the data values are arranged in a mess, they must rank.

  Median item number = *(6+1)*/2 =3,5

  Since the median is item 3,5 in the array, the third and fourth elements need to be averaged: (45+52)/2=48,5. Therefore, 48,5 is the median number of patients treated in hospital emergency room during the six-day period.

MODE for ungrouped data

– Is the observation in the data set that occurs the most frequently.
– Order the data from small to large.
– If no observation *repeats* there is no mode.
– If one observation occurs more frequently:
    • Unimodal
– If two or more observation occur the same number of times:
    • Multimodal
– Used for nominal scaled variables.
– The mode does not have valuable mathematical properties for use in future computations

Example – Given the following data sample:

2 5 8 –35 2 6 5 –4

The simple mean of the *sample* of nine measurements is given by:

$$\overline{x} = \frac{\sum_{i=1}^{9} x_i}{n}$$

$$= \frac{\boxed{2} + \boxed{5} + \boxed{8} + \boxed{-3} + \boxed{5} + \boxed{2} + \boxed{6} + \boxed{5} + \boxed{-4}}{\boxed{9}}$$

$$= \frac{26}{9} = 2,89$$

Example – Given the following data set:

2 5 8 –3 5 2 6 5 –4

The median of the *sample* of **nine** measurements is given by:

Odd number

–4 –3 2 2 5 5 5 6 8

1 2 3 4 5 6 7 8 9

Median item number =

$(n+1)/2$ = (9+1)/2 = 5^{th} measurement

Median = 5

Example

Given the following data set:

2 5 8 –35 2 6 5 –43

Determine the median of the ***sample*** of **ten** measurements.

Order the measurements

Even number

–4–32 2 3 5 5 5 6 8

1 2 3 4 5 6 7 8 9 10

***(n+1)/2* = (10+1)/2 = 5,5th measurement**

Median = (3+5)/2 = 4

Example

Given the following data set:

2 5 8 −3 5 2 6 5 −4

Determine the mode of the *sample* of nine measurements.

‘

−4 −3 2 2 5 5 5 6 8

Mode = 5

- **Unimodal**

Example

Given the following data set:

2 5 8 −3 5 2 6 5 −4 2

Determine the mode of the *sample* of ten measurements.

−4 −3 2 2 2 5 5 5 6 8

Mode = 2 and 5

- **Multimodal - bimodal**

Harmonic mean for ungrouped data

- Is used if M = const:

$$\bar{x} = \frac{\sum M}{\sum \frac{M}{x}} = \frac{nM}{M \sum \frac{1}{x}} = \frac{n}{\sum \frac{1}{x}}$$

- Harmonic mean is also called the simple mean of the inverse values .

Harmonic mean for ungrouped data

- <u>For example</u>:
- One student spends on a solution of task 1/3 hours, the second student – ¼ (quarter) and the third student 1/5 hours. Harmonic mean will be calculated:

$$\bar{x} = \frac{n}{\sum \frac{1}{x}} = \frac{1+1+1}{\frac{1}{1/3} + \frac{1}{1/5} + \frac{1}{1/4}} = \frac{3}{3+5+4} = \frac{3}{12} = 1/4\,(hour)$$

Geometric mean for ungrouped data

- This value is used as the average of the relations between the two values, or in the ranks of the distributions presented in the form of a geometric progression.

Geometric mean for ungrouped data

$$\bar{x} = \sqrt[n]{x_1 * x_2 * \ldots * x_n} = \sqrt[n]{\Pi(x_i)}$$

- Where Π – the multiplication of the data value (Xi).
- n – power of root

Geometric mean for ungrouped data

For example, the known data about the rate of growth of production

Year	2009	2010	2011	2012
Growth rate	1,24	1,39	1,31	1,15

Calculate the geometric mean. It is 127 percent:

$$\overline{X} = \sqrt[4]{1.24 * 1.39 * 1.31 * 1.15} = 1{,}27$$

Measures of location for grouped data

- **ARITHMETIC MEAN**
  - Data is given in a frequency table
  - Only an approximate value of the mean

$$\bar{x} = \frac{\sum f_i x_i}{\sum f_i}$$

where f_i = frequency of the i^{th} class interval

x_i = class midpoint of the i^{th} class interval

Example

There are data on seniority hundred employees in the table

Seniority, year (x)	The number of employees (f)	xf	
1	2	3	
9	**10**	**90**	
11	**10**	**110**	
13	**50**	**650**	
15	**20**	**300**	
17	**10**	**170**	
Total	**100**	**1320**	

- Average seniority employee is:

$$\bar{x} = \frac{\sum xf}{\sum f} = \frac{1320}{100} = 13{,}2\,year$$

Harmonic mean for grouped data

- Harmonic mean - is the reciprocal of the arithmetic mean. Harmonic mean is used when statistical information does not contain frequencies, and presented as

  xf = M.

Harmonic mean for grouped data

- Harmonic mean is calculated by the formula:

$$\bar{x} = \frac{\sum M_i}{\sum \frac{M_i}{x_i}}$$

- where M = xf

Example

There are data on hárvesting the apples by three teams and on average per worker

Number of teams	Harvesting the apples, kg	
	One worker (X)	Whole team (M)
1	800	2400
2	1200	9600
3	900	5600
Всего	x	17600

$$\bar{x} = \frac{17600}{\frac{2400}{800} + \frac{9600}{1200} + \frac{5600}{900}} = 1023(kg)$$

Geometric mean for grouped data

is calculated by the formula:

$$\overline{x} = \sqrt[\sum f_i]{\Pi(x_i^{f_i})} =$$

$$= \sqrt[\sum f_i]{(x_1^{f_1})*(x_2^{f_2})*\ldots*(x_2^{f_2})}$$

- Where f_i – frequency of the data value (Xi)

П – multiplication sign.

Geometric mean for grouped data

EXAMPLE

Year	2010	2011	2012	2013
Growth rate	1,24	1,24	1,31	1,31

Calculate the geometric mean. It is 127,5% percent:

$$\overline{X} = \sqrt[4]{1.24^2 * 1.31^2} = 1,275$$

Measures of location for grouped data

- **MEDIAN**
  - Data is given in a frequency table.
  - First cumulative frequency ≥ *n/2* will indicate the median class interval.
  - Median can also be determined from the ogive.

$$M_e = l_i + \frac{(u_i - l_i)\left(\frac{n}{2} - F_{i-1}\right)}{f_i}$$

where l_i = lower boundary of the median interval

u_i = upper boundary of the median interval

$F_{i\text{-}1}$ = cumulative frequency of interval foregoing median interval

f_i = frequency of the median interval

Measures of location for grouped data

- **MODE**
  - Class interval that has the largest frequency value will contain the mode.
  - Mode is the class midpoint of this class.
  - Mode must be determined from the histogram.

- Mode is calculated by the formula:

$$Mo = x_{Mo} + i\,\frac{f_{Mo} - f_{Mo-1}}{\left(f_{Mo} - f_{Mo-1}\right) + \left(f_{Mo} - f_{Mo+1}\right)}$$

- where x_{Mo} – lower boundary of the modal interval
- $i = x_{Mo} - x_{Mo+1}$ - difference between the lower boundary of the modal interval and upper boundary
- f_{Mo}, f_{Mo-1}, f_{Mo+1} – frequencies of the modal interval, of interval foregoing modal interval and of interval following modal interval

Measures of location for grouped data

Example – The following data represents the number of telephone calls received for two days at a municipal call centre. The data was measured per hour.

To calculate the **mean** for the *sample* of the 48 hours:

Determine the class midpoints

Number of calls	Number of hours f_i	x_i
[2–under 5)	3	3,5
[5–under 8)	4	6,5
[8–under 11)	11	9,5
[11–under 14)	13	12,
[14–under 17)	9	15,5
[17–under 20)	6	18,5
[20–under 23)	2	21,5
	$n = 48$	

Measures of location for grouped data

Example – The following data represents the number of telephone calls received for two days at a municipal call centre. The data was measured per hour.

$$\bar{x} = \frac{\sum f_i x_i}{\sum f_i}$$

$$= \frac{597}{48}$$

$$= 12,44$$

Average number of calls per hour is 12,44.

Number of calls	Number of hours f_i	x_i
[2–under 5)	3	3,5
[5–under 8)	4	6,5
[8–under 11)	11	9,5
[11–under 14)	13	12,5
[14–under 17)	9	15,5
[17–under 20)	6	18,5
[20–under 23)	2	21,5
***n* =**	**48**	

Measures of location for grouped data

Example – The following data represents the number of telephone calls received for two days at a municipal call centre. The data was measured per hour.

To calculate the for the *sample* **median** of the 48: hours:

determine the cumulative frequencies

Number of calls	Number of hours f_i	F
[2–under 5)	3	3
[5–under 8)	4	7
[8–under 11)	11	18
[11–under 14)	13	31
[14–under 17)	9	40
[17–under 20)	6	46
[20–under 23)	2	48
	$n = 48$	

$n/2 = 48/2 = 24$

The first cumulative frequency ≥ 24

Measures of location for grouped data

Example – The following data represents the number of telephone calls received for two days at a municipal call centre. The data was measured per hour.

Median

$$= l_i + \frac{(u_i - l_i)\left(\frac{n}{2} - F_{i-1}\right)}{f_i}$$

$$= 11 + \frac{(14-11)(24-18)}{13}$$

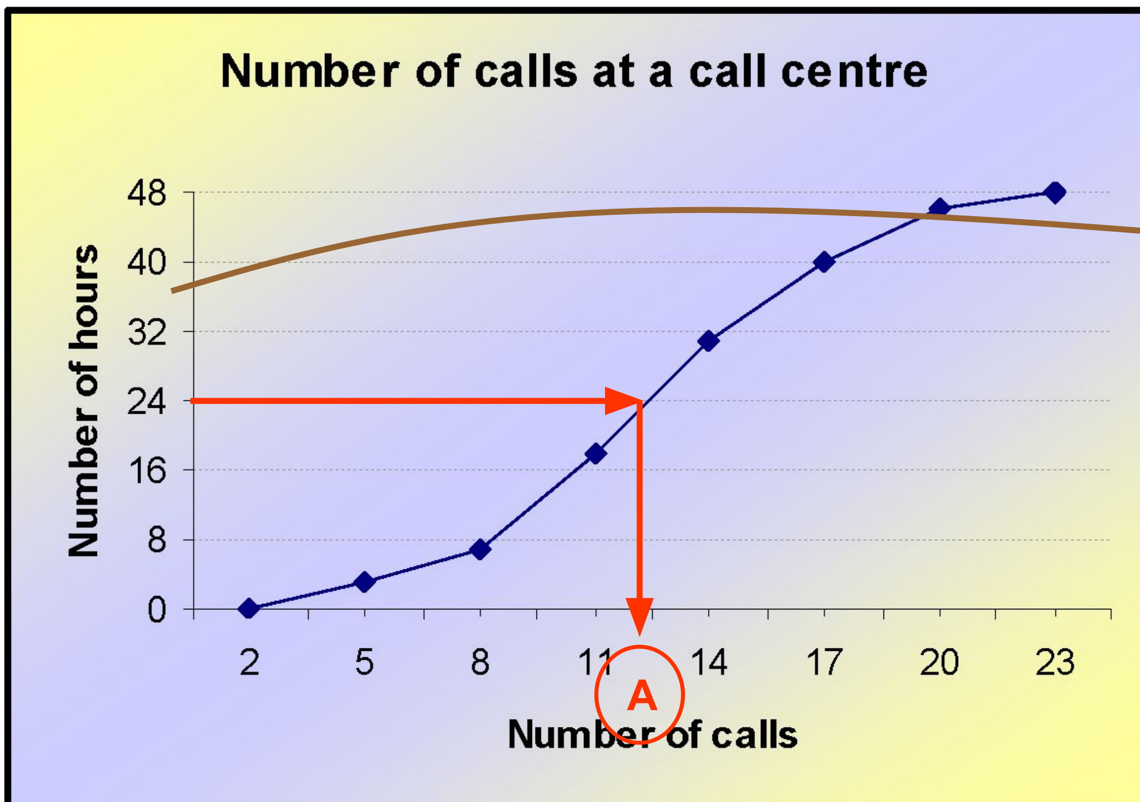
$$= 12,38$$

50% of the time less than 12,38 or 50% of the time more than 12,38 calls per hour.

Number of calls	Number of hours f_i	F
[2–under 5)	3	3
[5–under 8)	4	7
[8–under 11)	11	18
[11–under 14)	13	31
[14–under 17)	9	40
[17–under 20)	6	46
[20–under 23)	2	48
	$n = 48$	

Measures of location for grouped data

Example – The following data represents the number of telephone calls received for two days at a municipal call centre. The data was measured per hour.



The median can be determined form the ogive.

$n/2 = 48/2 = 24$

Median = 12,4
Read at A.

Measures of location for grouped data

Example – The following data represents the number of telephone calls received for two days at a municipal call centre. The data was measured per hour.

To calculate the for the *sample* **mode** of the 48 hours

The modal interval

The highest frequency

Number of calls	Number of hours f_i
[2–under 5)	3
[5–under 8)	4
[8–under 11)	11
[11–under 14)	13
[14–under 17)	9
[17–under 20)	6
[20–under 23)	2
$n = 48$	

MODE

- We substitute the data into the formula:

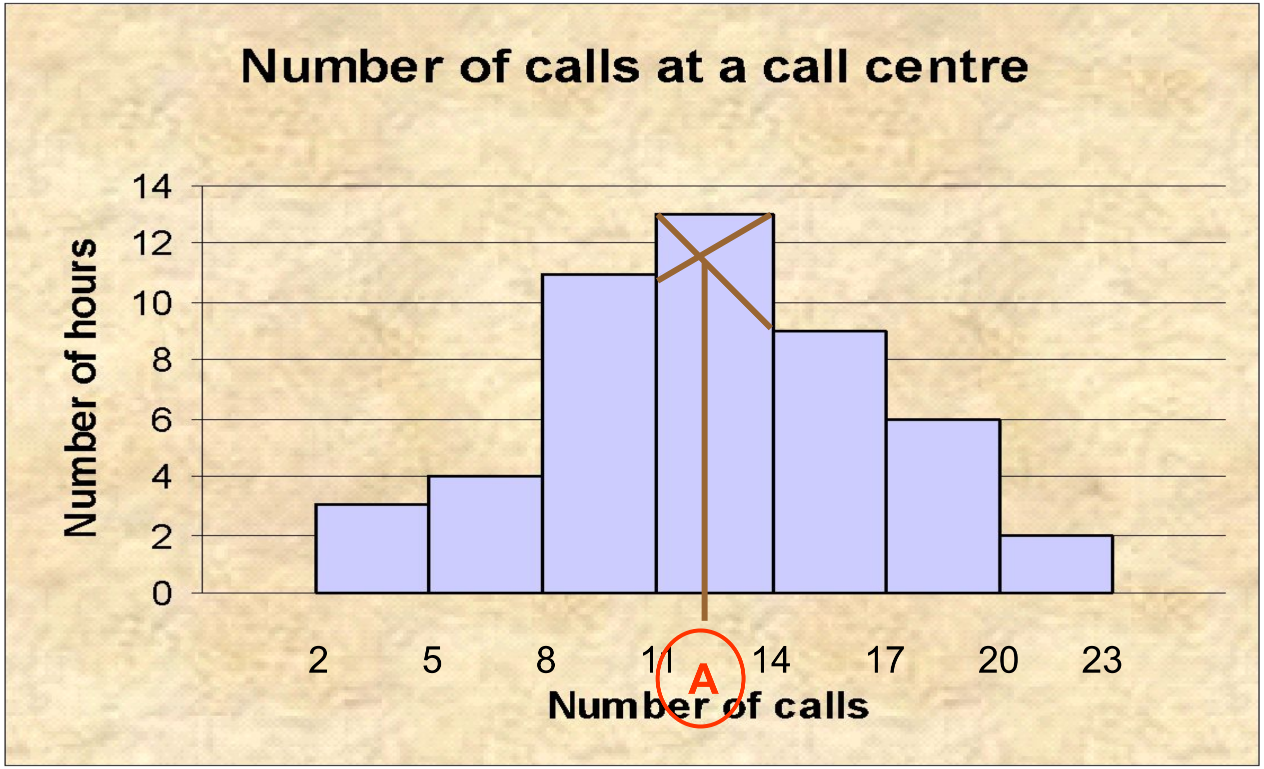
$$Mo = x_{Mo} + i \frac{f_{Mo} - f_{Mo-1}}{(f_{Mo} - f_{Mo-1}) + (f_{Mo} - f_{Mo+1})} =$$

$$= 11 + (14 - 11) \cdot \frac{13 - 11}{13 - 11 + 13 - 9} = 12{,}3$$

- Mo = 12,3
- So, the most frequent number of calls per hour = 12.3

Measures of location for grouped data

Example – The following data represents the number of telephone calls received for two days at a municipal call centre. The data was measured per hour.



The mode can be determined form the histogram.

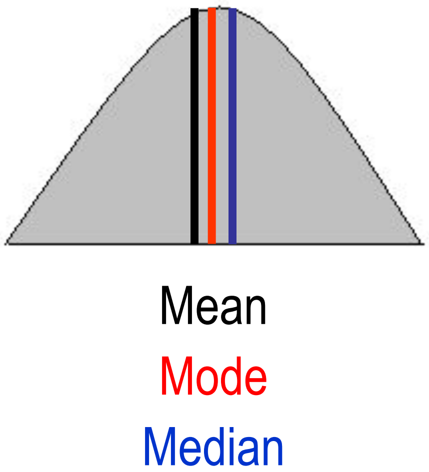
Mode = 12,3

Read at A.

Measures of location for grouped data

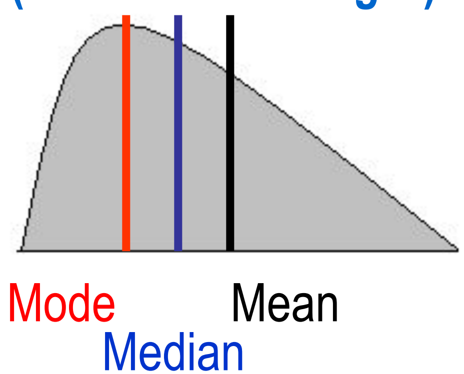
Relationship between mean, median, and mode

- If a distribution is symmetrical:
  - the mean, median and mode are the same and lie at centre of distribution

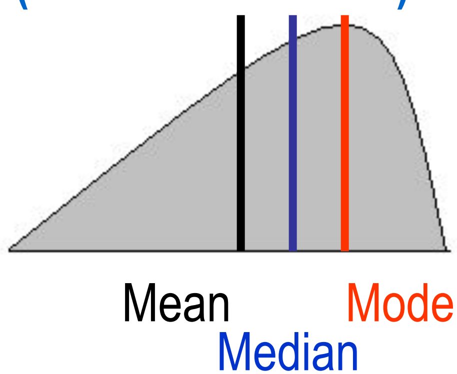


- If a distribution is non-symmetrical:
  - skewed to the left or to the right
  - three measures differ

A positively skewed distribution (*skewed to the right*)

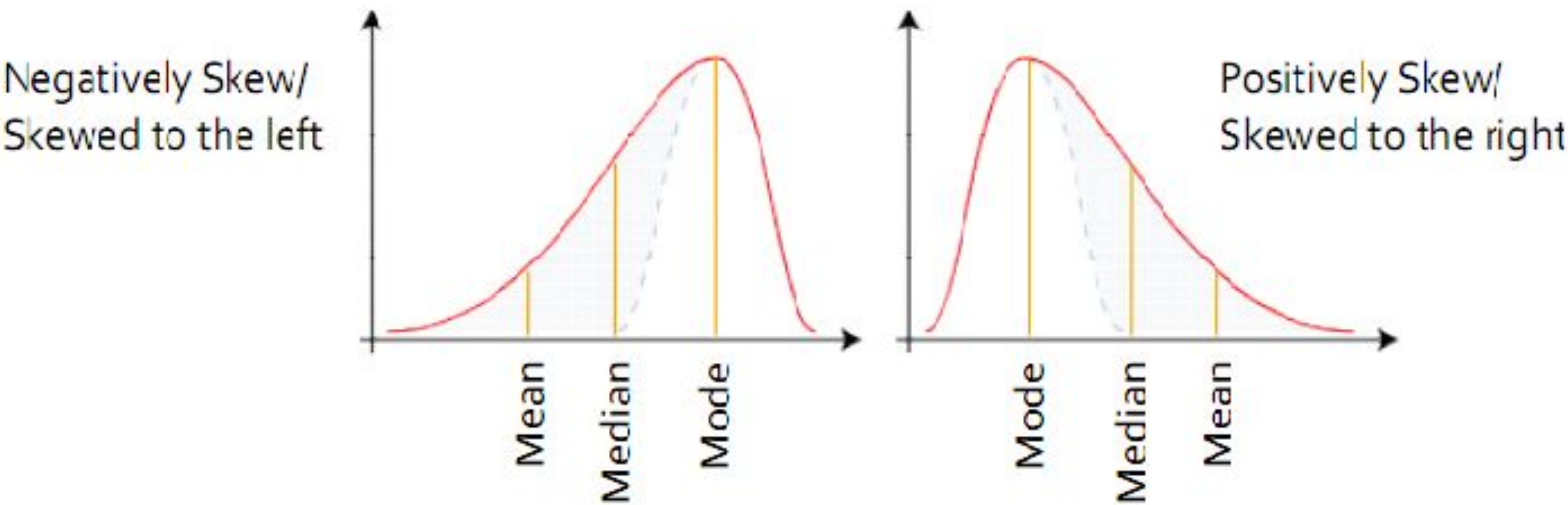


A negatively skewed distribution (*skewed to the left*)



Skewness

is a measure of the asymmetry of the probability distribution of a real-valued random variable



Sk = 3 (mean – median) / standard deviation

EXAMPLE

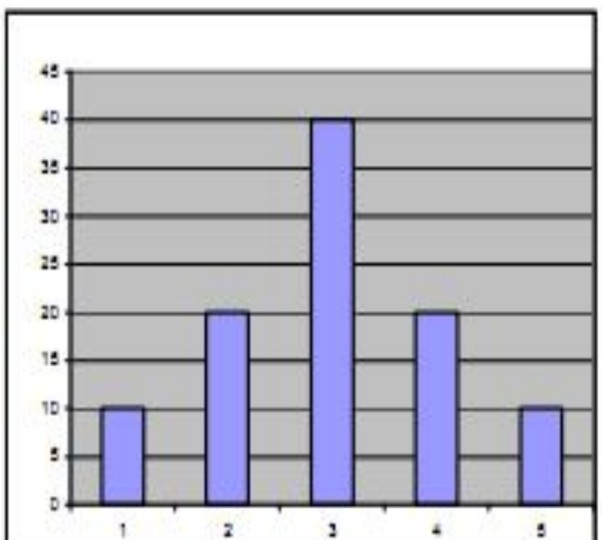
- Consider a study of the hourly wage rates in three different companies, For simplicity, assume that they employ the same number of employees: 100 people.

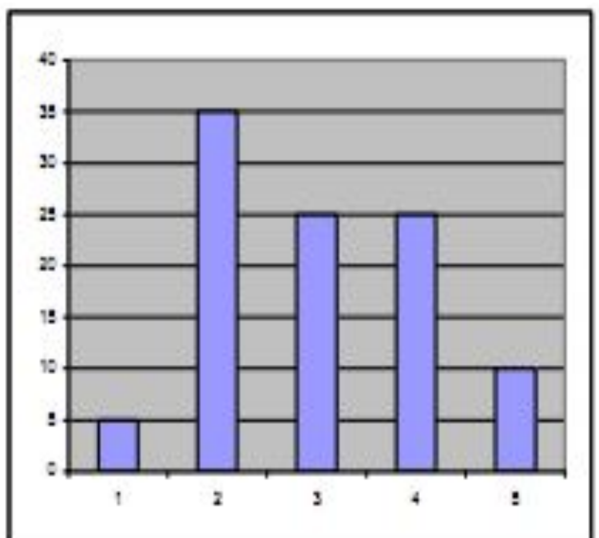
Percentage of employees

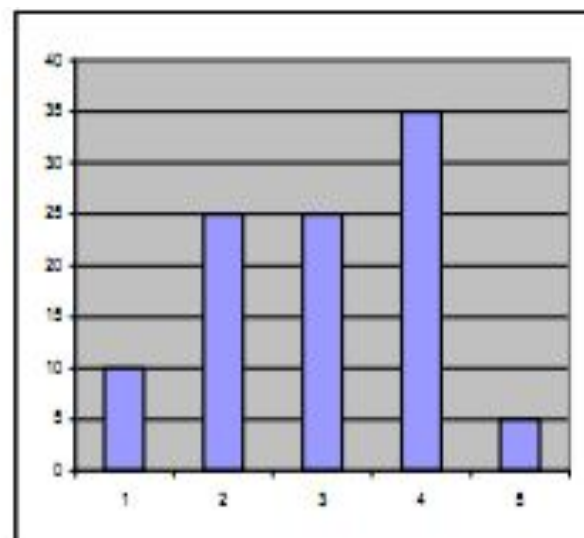
Hourly wages	Comp. I	Comp. II	Comp. III
10-20	10	5	10
20-30	20	35	25
30-40	40	25	25
40-50	20	25	35
50-60	10	10	5
Total:	100	100	100
Average:	35	35	35
Variance:	120	120	120

- So we have three 100-element samples, which have the same average value (35) and the same variability (120). But these are different samples. The diversity of these samples can be seen even better when we draw their histograms.

- The histogram for company I (left chart) is <u>symmetric</u>. The histogram for company II (middle chart) is <u>right skewed</u>. The histogram for company III (right chart) is <u>left skewed</u>. It remains for us to find a way of determining the type of asymmetry (skewness) and “distinguishing” it from symmetry.







POSITIONAL CHARACTERISTICS

- Knowing the median, modal and average values enables us to resolve the problem regarding the symmetry of the distribution of the sample. Hence,

\- For ***symmetrical distributions:***

x = Me = Mo ,

\- For ***right skewed distributions***:

x > Me > Mo

\- For ***left skewed distributions***:

x < Me < Mo .

- **We obtain the following relevant indicators (measures) of asymmetry:**
- **Index of skewness:** $\overline{X} - Mo$; $\overline{X} - Me$

- **Standardized skewness ratio:**

$$As = \frac{\overline{X} - Mo}{\sigma} \qquad As = \frac{\overline{X} - Me}{\sigma}$$

- **Coefficient of asymmetry**

$$As = m_3 / \sigma^3 = \frac{\frac{1}{n} \cdot \sum (X_j - \overline{X})^3}{\sigma^3}$$

Example

Years of service	Workers (f)	Calculation						
		X_i	xf	$\Sigma f=F$	$\left\|x-\bar{x}\right\|$	$\left\|x-\bar{x}\right\|*f$	$\left(x-\bar{x}\right)^2$	$\left(x-\bar{x}\right)^2 f$
6-10 10-14 14-18 18-22	15 30 45 10							
Total:	100							

Years of service	Workers (f)	Calculation			
		Midpoint X_i	xf	$\Sigma f=F$	
6-10	15	8			
10-14	30	12			
14-18	45	16			
18-22	10	20			
Всего:	100	14			

Years of service	Workers (f)	Calculation			
		Midpoint X_i	xf	$\Sigma f = F$	
6-10	15	8	120	15	
10-14	30	12	360	45	
14-18	45	16	720	90	
18-22	10	20	200	100	
Всего:	100	14	1400	x	

- **The weighted arithmetic mean**

$$\bar{x} = \frac{\sum x_i f_i}{\sum f_i}$$

$$\bar{x} = \frac{1400}{100} = 14\,years$$

- **The median**

$$Me = li + (Ui - li)\frac{n/2 - F_{i-1}}{f_i}$$

$$Me = 14 + (18 - 14)\frac{50 - 45}{45} = 14.4$$

- **The mode**

$$Mo = x_{Mo} + i\frac{f_{Mo} - f_{Mo-1}}{\left(f_{Mo} - f_{Mo-1}\right) + \left(f_{Mo} - f_{Mo+1}\right)}$$

$$Mo = 14 + (18 - 14)\frac{45 - 30}{\left(45 - 30\right) + \left(45 - 10\right)} = 15{,}2$$