MEDICAL UNIVERSITY – PLEVEN
FACULTY OF PUBLIC HEALTH

DEPARTMENT OF PUBLIC HEALTH SCIENCES
CENTRE FOR DISTANT LEARNING

LECTURE No3

DESCRIPTIVE STATISTICS FOR QUANTITATIVE DATA. MEASURES OF CENTRAL TENDENCY

Assoc. Prof. G. Grancharova, MD. PhD

Plan of the lecture

Part 1. Introduction

Part 2. Measures of central tendency.

Part 3. Measures of location: quantiles and percentiles

Part 1. Introduction

There are two basic methods of summarization:
numerical and graphical.

The objective of the numerical approach is to convert masses of numbers (raw data) into meaningful **summary statistics** (indices), reduced to a single number, that convey information about the average (typical) degree of a given variable and the degree to which observations differ (the degree of dispersion or spread).

After the collection of raw data they should be organized and presented in a meaningful way.

Frequency distributions give a general picture of the pattern of the observations but sets of measurements cannot be adequately described only by the values of all individual measurements.

For many purposes, the overall summary of a group's characteristics is of utmost importance.

The process of summarization is based on **two main characteristics of quantitative data:**

1. **Firstly,** this is the individual variability of observations in any set of measurements.

2. **Secondly,** despite the individual fluctuations, the values of the most quantitative variables tend to some **typical "middle" level** (central point or the most characteristic value) around which all the values are distributed. Measures of such distribution are referred to as **measures of central tendency.**

The central tendency is due to determining factors and causes inherent in all cases of a given sample or population while the variability or dispersion is due to specific factors which may occur in some cases but may be absent in others.

Part 2

MEASURES OF CENTRAL TENDENCY

BASIC TERMS:

An array (a distribution) of a set of numbers is simply those numbers in ordered sequence from the lowest to the highest.

Each array (a distribution) has the following basic components:

x - each individual raw score in a sample or in a population;

n - the number of cases in a sample;

N - the number of cases in a population;

f - frequency (the number of observations with the same value);

range - the difference between the largest and the smallest value in an array;

Σx - the sum of all values in a sample or in a population

Before presenting the specific measures of central tendency, it is important to know **the shape of the distribution** and the dispersion of the scores in order to interpret the data correctly.

The three most commonly used measures of central tendency are:

the arithmetic mean,

the median, and

the mode.

Arithmetic mean

It is denoted by:

$\bar{x}$ – for a sample and by

μ - for a population.

How to compute the arithmetic mean depends on the way on which the initial data are presented (raw or grouped data), and on the number of cases (statistical units).

Arithmetic mean

In ungrouped data – small number of observations (n<30)

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

Example: The age of 10 primi-birth women is:
18, 21, 23, 23, 25, 27, 27, 28, 30, 33.

$\overline{\mathbf{x}}$ = 255/10 = 25.5

In grouped data:

$$\bar{x} = \frac{\Sigma x.f}{\Sigma f}$$

Example: The values of birth height of 100 newborns:

Height in sm – x	Frequency f	x.f
46	2	92
47	6	282
48	7	336
49	20	980
50	30	1500
51	20	1020
52	8	416
53	5	265
54	2	108
	$\Sigma f = 100$	$\Sigma x.f = 4999$

$$\bar{x} = \frac{\Sigma x.f}{\Sigma f}$$

$\bar{x} = 4999:100=$
$= 49.99 \sim 50.0$

In an interval array

with an equal interval length:

$$\bar{x} = \frac{\Sigma\, c.f}{\Sigma\, f}$$

c is the middle of each interval

In the above example the data can be regrouped in 3 intervals with a length of 3 cm:

Height cm – *x*	**Frequency** *f*	*c*	*c.f*	$\bar{x} = \frac{5000}{100} = 50$
46 - 48	15	47	705	
49 - 51	70	50	3500	
52 - 54	15	53	795	
	$\Sigma f = 100$		$\Sigma c.f = 5000$	

Characteristics of the mean

1. It is the most widely used measure of central tendency. It substitutes by one single number all the individual values of a given variable and describe its typical level in a data set.

2. For **normal or roughly symmetric distributions the mean is the best measure of central tendency.**

3. **In skewed distributions**, the mean can be misleading since it can be greatly influenced by scores in the tail. In such cases **the median is more informative.**

4. The mean can be affected by the presence of a small number of **outliers** (e.g. values that are different from the rest units) that can distort the mean. We can eliminate such extreme values and compute a new mean, which will be more typical.

Such method is based on the **criterion U - the ratio of the difference between the outlier and the mean and the standard deviation s.**

The computed criterion U is then compared with the table of critical values of u_t and if $u \geq u_t$, the extreme value x_i is discarded as unusual.

The mean is calculated without the discarded outlier.

5. The sum of the deviations of the scores in the distribution from the mean always is equal to zero because half of the distribution is above and half is below the mean.

6. If to each value of the frequency distribution the same number is added or subtracted, then the mean is increasing or decreasing by the same number.

7. The mean is not a “real” value and this makes the acceptance and interpretation of the data sometimes more difficult – e.g., a mean number of children in a sample might be 2.4, or an average number of limbs in a sample is 3.997.

Median

The median (M_e) is the measure of central tendency which can be identified or determined by an inspection.

It is a value that divides the array of observation in two equal part, e.g. it is the middle value.

The procedure to identify the median is to:

1. rearrange all observations from the smallest to the largest in an ordered series (all data values should be listed even though some values may repeat more than once);

2. Then we must determine whether the number of cases is odd or even;

- when it is odd, the median is the value in the middle;
- when it is even, the median is just a halfway of values of the two middle observations.

Characteristics of the median

1. The median is usually a realistic value, or measured in half-units (when the number of observations is even).

2. The median is more robust towards outliers (extreme scores). This makes the median a better measure than the mean for highly skewed distributions.

Example for outliers: 10 individuals who have been tested HIV positive reported the following number of sexual contacts in a 6-month period:

2 4 4 6 7 8 10 12 15 93

The mean value of 16.1 (Σx=161) is higher than that reported by 9 of the 10 individuals and yet far below that reported by the 10th individual. Such a mean in not, in any sense, typical or representative of any one in the study group.

In a situation like this, the median value may well be more informative.

Source: Thomas H. Hassard. Undetstandung biostatistics, Mosby Year Book, 1991, p.6

3. The median does not include all the individual values of a variable. So, it reflects only one value in odd number of cases or two values in even number of cases.

4. The median is preferred measure of central tendency when:

- the lowest and highest values of a quantitative variable are far off of the rest values;
- there is uncertainty in some values;
- it is not possible to determine the exact shape of the distribution or when the distribution is highly skewed;
- when the number of cases is small.

MODE

The mode (M_o) is the observation in an array with the highest frequency of occurrence. Its meaning is obvious and it is determined by an inspection of a frequency distribution.

Although it is common for most distributions to contain exactly one mode (as in a normal distribution and large homogeneous samples), it is possible for more than one mode to exist.

A distribution having one mode is called unimodal.

A distribution having two modes is called bimodal.

CHARACTERISTICS OF THE MODE

1. The mode is a quick and easy method of determining the most popular score at a glance.

2. The mode is the only measure of central tendency that can be used with nominal data.

3. The mode is the weakest measure of central tendency as compared to the mean and median. This is true because, in some cases, the mode may be the lowest or the highest value in the distribution.

4. Many distributions have more than one mode and they are called multimodal.

5. The mode has a true meaning and this is very important in medicine and public health. For example, it is more important to determine which group has higher risk for some disease, e.g. to determine the mode in the age distribution instead of calculating the mean age of persons with the disease.

Comparison of measures of central tendency

The mean is the most stable. If repeated samples were drawn from a given population, the means would vary or fluctuate less than the modes or medians. Because of its stability, the mean is the most reliable estimate of the central tendency of the population.

The mean is the most widely used because it takes every score into account.

The mean is the most efficient measure of central tendency for normal distributions and it is not appropriate for highly skewed distributions.

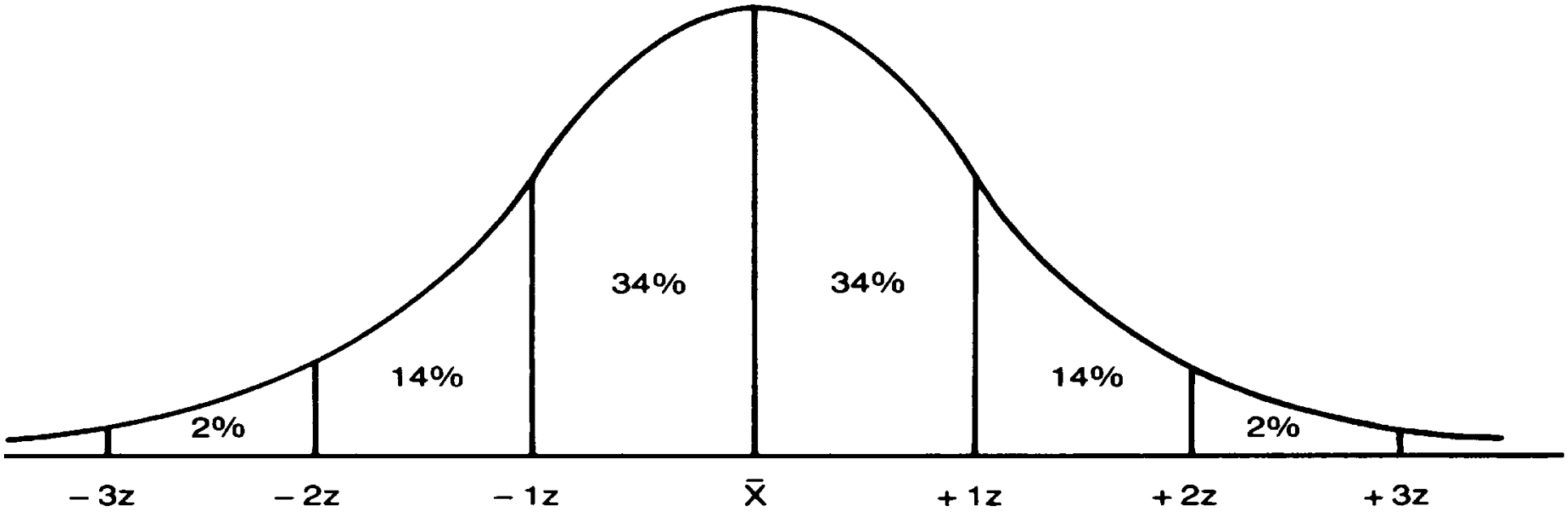
The median is useful because its meaning is clear and it is more efficient than the mean in highly-skewed distributions. However, it ignores many scores and is generally less efficient than the mean, the trimean, and trimmed means.

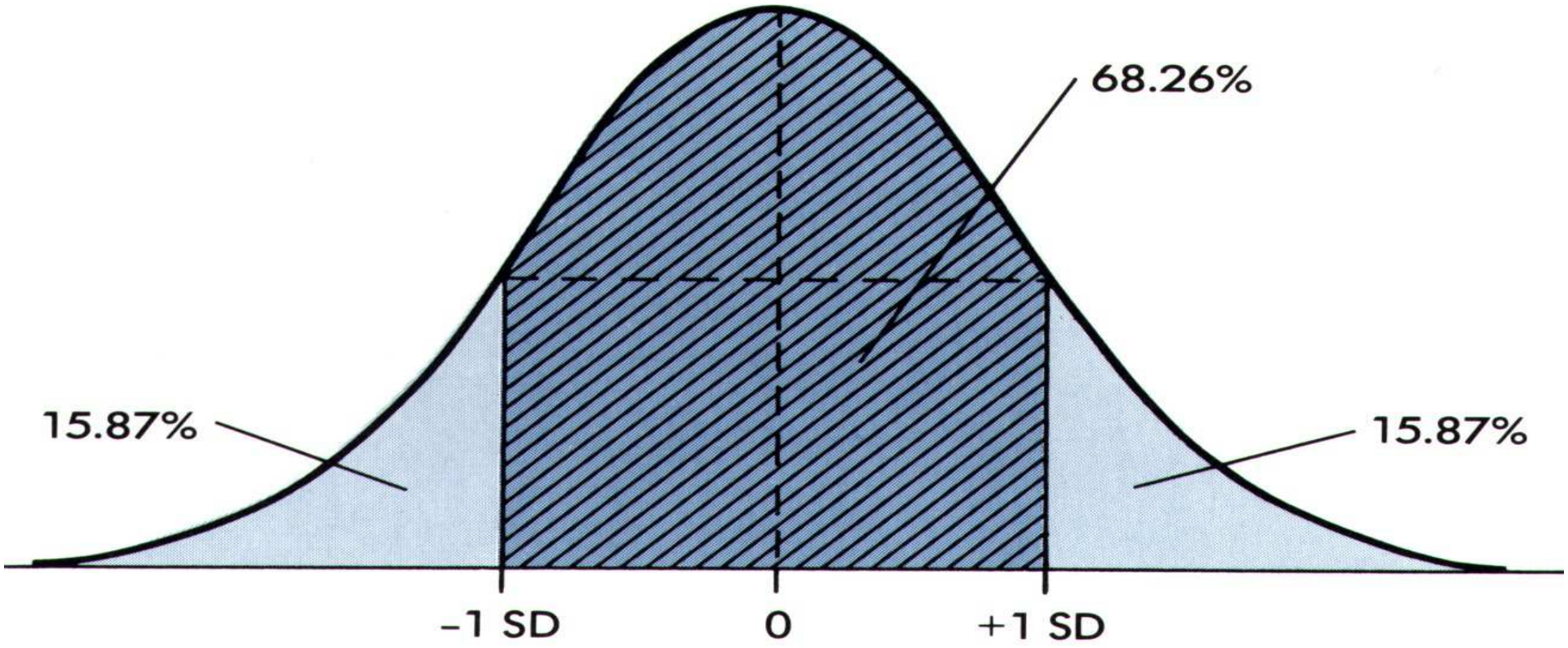
The mode can be informative but should almost never be used as the only measure of central tendency since it is highly susceptible to sampling fluctuations.

The level of measurement is very important to determine the appropriate index of central tendency:

- the mode is appropriate for nominal scales;
- the median is appropriate for ordinal scales;
- the mean is appropriate for interval and ratio scales.

When a **distribution is symmetric and unimodal, the mean, the median and the mode – coincide**.


34%
34%
14%
14%
2%
2%
– 3z
– 2z
– 1z
X̄
+ 1z
+ 2z
+ 3z

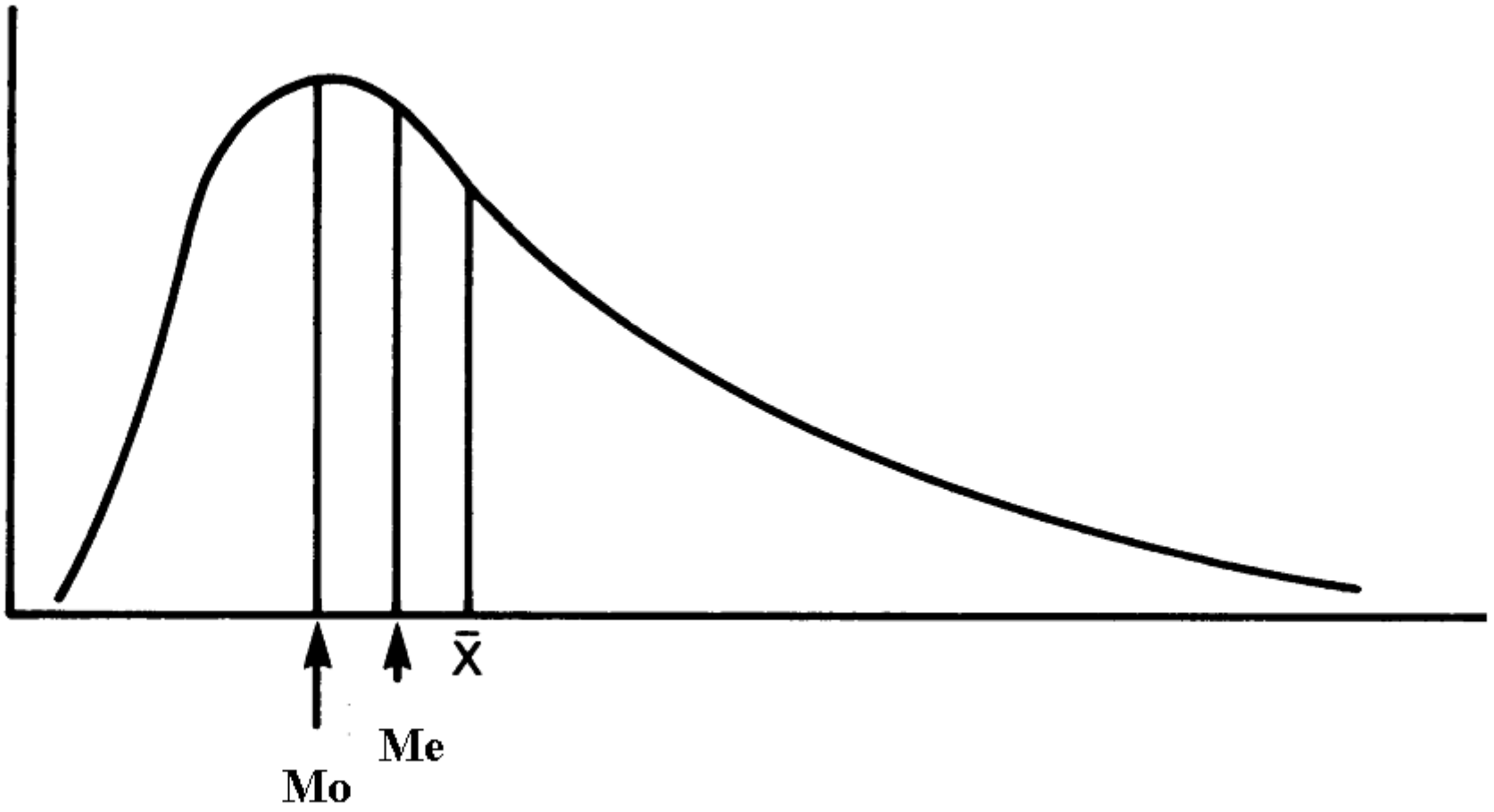

68.26%
15.87%
15.87%
−1 SD
0
+1 SD

In skewed distributions, the values of the mode, median, and mean differ.

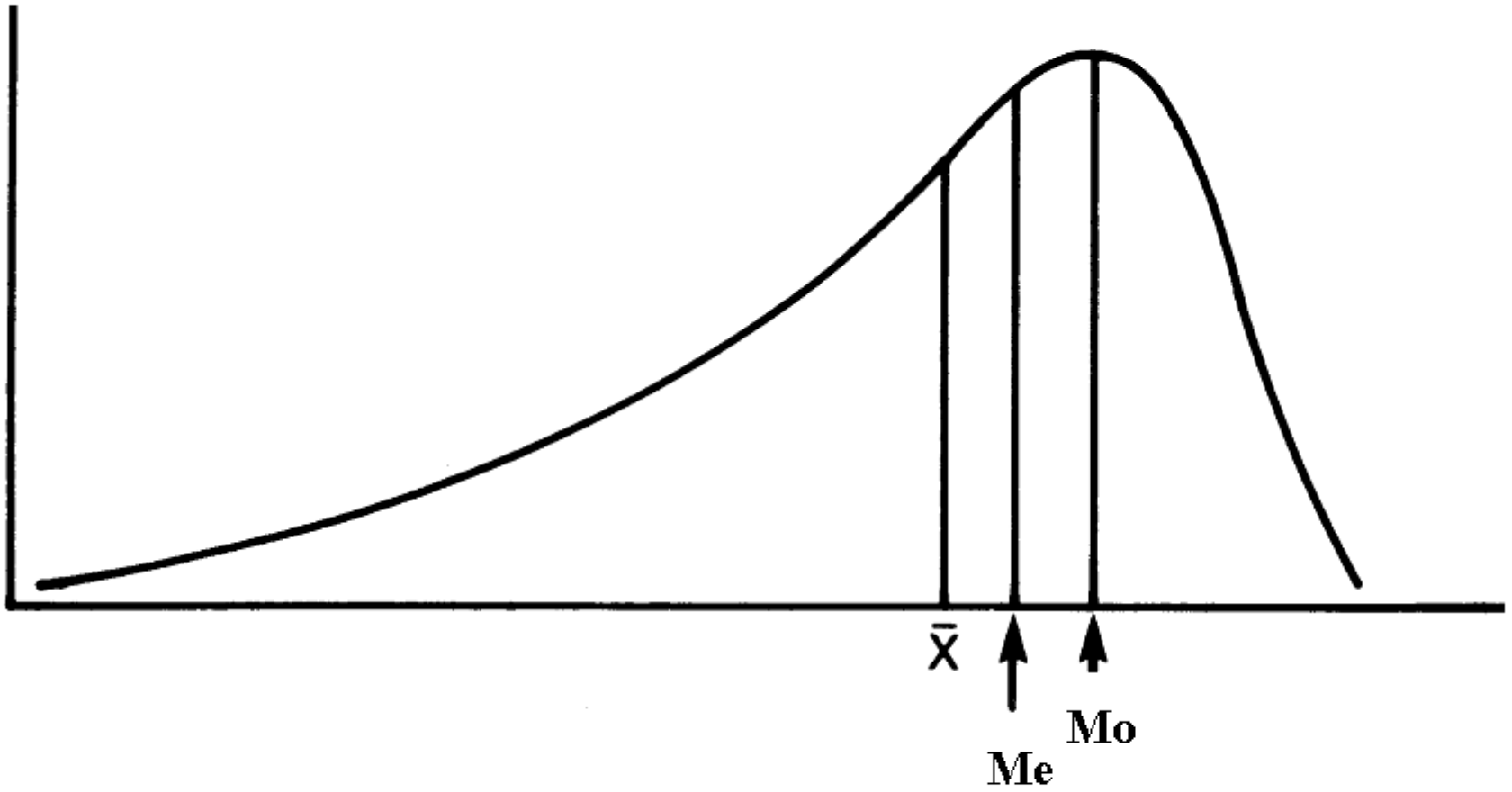
The mean is always pulled in the direction of the long tail and it is higher than the median and mode.

Thus, **in positively skewed distributions**

mode<median<mean


$\bar{X}$
Me
Mo

In negatively skewed distributions with most of the scores being high and with some scores spreading out towards the lower end of the distribution, e.g. the tail is directed to the left or negative side of the distribution the mean is lower than the median and mode – mean<median<mode


$\bar{X}$
Me
Mo

Part 3

MEASURES OF LOCATION: QUANTILES AND PERCENTILES

Quantiles (Q)

Quantiles are special measures of location - points that divide the ordered series of data (from the lowest to the highest value) into subgroups of equal size.

They mark the boundaries between consecutive subgroups in ordered series of data (an array).

Types of quantiles (Q)

There are several types of quantiles:
- **terciles** divide the distribution into three equal subgroups (called thirds);
- **quartiles** - into 4 subgroups (quarters);
- **quintiles** -into 5 subgroups (fifths);
- **deciles** - into 10 subgroups (tenths);
- **centiles** – into 100 parts (hundredths)

Quantile/s	Number of equal parts of a distribution	Number of quantiles of this type
Median	2	1
Quartiles	4	3
Deciles	10	9
Percentiles	100	99

Estimation of quantiles

Quantiles are usually identified or determined.

The procedure of identifying quantiles is as follows:

First, we need to rearrange all observations in an ordered series from the lowest to the highest value.

Second, we must determine whether the number of cases is odd or even.

Example:

Let's have an observation on the age at first birth for a sample of 10 mothers:

1. Raw data on the array are as follows:
21, 23, 27, 30, 18, 23, 33, 23, 27, 28

2. Now let's rearrange all observations according to the magnitude of a value of a variable we are observing in an ordered series of data from the lowest to the highest value:
18, 21, 23, 23, 25, 27, 27, 28, 30, 33

3. We need to determine if the number of cases is odd or even.
In this example the number of cases is even.

4. We locate the central two observations (5^{th} = 25 and 6^{th} = 27)

18, 21, 23, 23, 25, 27, 27, 28, 30, 33

Afterwards we sum the values of these two units and divide the sum by 2.

25+27/2 = 26

So, the median is just a halfway of values of the two middle observations.

5. Then we can repeat exactly the same procedure in the lower half and in the upper half of the ordered series to locate the quantiles dividing the ordered series in four equal parts (quartiles).
Numbers in green represent the other two quartiles,

18, 21, **23**, 23, 25, 27, 27, **28**, 30, 33

6. We can repeat the procedure until we divide the distribution in wanted number of equal parts.

Use of quantiles

Quantiles are used in description of both - the central tendency and the dispersion of a distribution they are describing.

Median (the quantile diving raw data in 2 equal parts) is used as a measure of central tendency in skewed distributions.

Quartiles are used for quick estimation of the degree of dispersion in an array.

Quartiles - observations that divide the distribution into four equal parts. There are 3 quartiles - **Q_1, Q_2 and Q_3.**

Example: If we have an array of 23 cases: the first quartile Q_1 is the 6^{th} observation; Q_2 is the 12^{th} observation, and Q_3 is equal to the 18^{th} observation.

Percentiles

Percentiles (also called centiles) - points that divide an array into 100 equal parts. There are 99 percentiles, denoted as P_1, P_2, ... P_{25},...P_{50},...., P_{75}, ...P_{99}.

Characteristics of percentiles

1. A percentile tells us the relative position of a given observation.

2. It allows us to compare scores on tests that have different means and standard deviations (e.g., the 10^{th} percentile exceeds 10% and is exceeded by 90% of the observations, the 75^{th} percentile exceeds 75% of the data, etc.)

Use of percentiles

Percentiles are used **to establish the reference limits of normality** in clinical and other areas of investigation.

The establishment of “normal ranges” of values for health data permits the selection of appropriate action in medical practice or allows for accurate estimate of many clinical and laboratory indicators.

For this purpose usually seven main percentiles are used:

P_3, P_{10}, P_{25}, P_{50}, P_{75}, P_{90} and P_{97} –

to form the upper and lower limits of seven reference groups of population.

Percentiles have an advantage as compared to the other methods of determining "normal" values as they **are applicable to any form of distribution** (not only to normal distribution).

Comparison between different types of quantiles

P_{25} corresponds to Q_1

P_{50} corresponds to Q_2 and to the median

P_{75} is equal to Q_3

Taking into account that the median (M_e) is the second quartile (Q_2), then the median of the lower half of the data gives the first quartile (Q_1), and similarly, the median of the upper half of the data gives the third quartile (Q_3).

Important!

I am asking very kindly the representatives from groups 7, 8, 10, 4 and 13 to come to me after the lecture.

Test examples

1. ***In case there are too many outliers in the data set, the most representative average value is***

A. Mean

B. Mode

C. Median

2. ***Given a set of nominally scaled scores, the most appropriate measure of central tendency is the:***

A. mean

B. mode

C. standard deviation

D. range

3. ***Which of the following statements is true?***

A. The mode is the most useful measure of central tendency.

B. The variance is the square root of the standard deviation.

C. The median and the 50^{th} percentile rank have different values.

D. The mean is more affected by extreme scores than the median.

4. ***Given the group of scores 1, 4, 4, 4, 7, it can be said of the mean, the median, and the mode that:***

A. the mean is larger than either the median or the mode

B. all are the same

C. the median is larger than either the mean or the mode

D. all are different

E. the mode is larger than either the median or the mode

5. ***Inferential statistics are used to describe specific characteristics of the data.***

A. True **B.** False

6. ***Select the statement which you believe to be true. The arithmetic mean of a set of values:***

A. Is a useful summary measure of central tendency (of location) if the data are symmetrical.

B. Is always greater than the median

C. Cannot be calculated if the data set contains both positive and negative values.

7. ***Central tendency describes the 'typical' value of a set of scores.***
A. True **B.** False

Example: *A nurse recorded the number of analgesic preparations taken by patients in a surgical ward. The resulting data were: 5, 2, 8, 2, 3, 2, 4, 12.*
Questions 8-11 refer to this data.

8. ***The mode for this distribution is:***
A. 2
B. 3
C. 8
D. there is no mode

9. ***The median is:***
A. 2.00
B. 3.50
C. 3.00
D. 3.25

10. ***The mean is:***
A. 3.52
B. 5.43
C. 4.75
D. 4.15

11. ***The range is:***
A. 9
B. 10
C. 12
D. 2

12. ***Descriptive statistics are used to describe specific characteristics of the data.***

A. True **B.** False

13. ***With nominal data, the mean should be used as a measure of central tendency.***

A. True **B.** False

14. ***The mode represents the most frequently occurring score in a distribution.***

A. True **B.** False

15. ***With ordinal data we can use both the mode and the mean as a measure of central tendency.***

A. True **B.** False

16. ***When the data are interval or ratio, we can use the mean as a measure of central tendency.***

A. True **B.** False

17. ***If a continuous distribution is highly skewed, the median might be the appropriate measure of central tendency.***

A. True **B.** False

18. ***When a frequency distribution is positively skewed, the mean is greater than the median or the mode.***

A. True **B.** False

19. ***Given a normal distribution, the three measures of central tendency are equivalent.***

A. True **B.** False

20. ***If we subtract the value of the mean from every score in a set of scores the sum of the remaining values will be:***

A. impossible to determine

B. equal to the mean

C. a measure of the dispersion around the mean

D. zero

21. ***Given a normally distributed continuous variable the best measure of central tendency is the:***

A. mode

B. median

C. mean

D. standard deviation

22. *Select the statement which you believe to be true.* ***The arithmetic mean of a set of values:***

A. Cannot be calculated if the data set contains both positive and negative values.

B. Is always greater than the median.

C. Coincides with the median if the distribution of the data is symmetrical.

23. *Select the statement which you believe to be true.* ***The median:***

A. Is a measure of the spread of the data.

B. Is greater than the arithmetic mean when the data are skewed to the left.

C. Can be distorted by outliers.

24. In case there are too many outliers in the data set, the most representative average value is

A. Mean
B. Mode
C. Median

25. One way to measure the spread is to calculate the difference between the third and first quartile. This measure is called

A. The interquartile range
B. The mid quartile
C. The differential quartile

26. Since mode is the most frequently occurring score, it can be determined directly from a frequency distribution or a histogram

A. True B. False

27. The 50th percentile score and the median will always be the same value.

A. True B. False

28. Twenty five percent (25%) of the scores fall between Q1 and the median.

A. True B. False

29. When the data are interval or ratio, we can use the mean as a measure of central tendency.

A. True B. False

30. With nominal data, the mean should be used as a measure of central tendency.

A. True B. False

Answers

1-C	11-B	21-C
2-B	12-A	22-C
3-D	13-B	23-B
4-B	14-A	24-C
5-B	15-B	25-A
6-A	16-A	26-A
7-A	17-A	27-A
8-A	18-A	28-A
9-B	19-A	29-A
10-C	20-D	30-B