MEDICAL UNIVERSITY – PLEVEN
FACULTY OF PUBLIC HEALTH

DEPARTMENT OF PUBLIC HEALTH SCIENCES
CENTRE FOR DISTANT LEARNING

LECTURE No4

DESCRIPTIVE STATISTICS FOR QUANTITATIVE DATA. MEASURES OF SPREAD

Assoc. Prof. Gena Grancharova, MD, PhD

Plan of the lecture

Part 1. MEASURES OF SPREAD

Part 2. THE CONCEPT OF NORMS AND NORMAL GROUPS' LIMITS

Part 1

MEASURES OF SPREAD (DISPERSION, VARIABILITY)

Why we need measures of dispersion?

Measures of central tendency do not give a total picture of a distribution.

1. Two sets of data with identical means could be very different from one another.

2. Two distributions with the same means could be very different in shape - they could be skewed in opposite directions.

3. Even when two sets of data have equal means, medians, modes, and the same form of distribution, they could be different from one another.

Consider the following two sets of data:

Set №1: 18, 21, 23, 23, 25, 27, 27, 28, 30, 33

Set №2: 23, 23, 24, 25, 26, 26, 27, 27, 27, 27

The means for the two samples = 25.5
The medians = 26
The modes = 27

But the two sets are very different:
the range for set №1 is 33 – 18 = 15
the range for set №2 is 27 – 23 = 4

So, the knowledge of summary measures, describing the central tendency in a sample or in a population is not enough without a measure of the extent of variability or spread of the measurements around these summary indices.

This means that no description of any health data by summary measures is complete without the measures of variability.

The most common measures of variability or spread include the following:

- the range
- the standard deviation
- the variance
- the inter- and semiquartile range
- the coefficient of variation.

Range

The range is simply the difference between the highest and the lowest values in an array of the variable in a given empiric distribution.

The range can be easily computed but a single outlier may have a large impact on the range.

Another disadvantage is that it ignores completely the variations in scores between the highest and the lowest values, as it takes into account just the two extreme values.

For these reasons, the range is used only as a gross descriptive index and is typically reported in conjunction with other measures of variability.

Standard deviation and variance

The standard deviation (denoted by SD or s for a sample and σ for a population) is the most commonly reported measure of variability, especially with interval or ratio data.

Standard deviation describes the degree of variation among the individual observations in the sample around the mean, and like the mean it considers every score in a given distribution.

For this reason, means and standard deviations are generally reported together in the text or in tables.

The calculation of SD includes the following steps:

1. Firstly, we calculate how much each individual varies from the mean by subtracting the mean from each individual value ($x - \bar{x}$).

2. Secondly, we add the individual variations together.

$$\sum(x - \bar{x}).$$

To calculate the average deviation, the sum should be divided by the number of scores -

$$\frac{\sum(x - \bar{x})}{n}$$

Unfortunately, summing the differences of deviations will always lead us to zero.

The reason is that those individuals who have values larger than the mean will simply cancel out those that have values below the mean.

3. Third, to overcome this problem, we can square each difference and calculate the sum of squared deviations around the mean.

$$\sum(x - \bar{x})^2$$

4. Fourth, the sum of squares has to be related to the number of results under study.
So, to allow fair comparisons between studies of different sizes, we should take the study size into account by calculating an "average" variation, called variance – s^2.

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

s^2 = variance of a sample
x = value of a single unit
$\bar{x}$ = arithmetic mean of a sample
n = number of units in a sample studied
$n-1$ = degrees of freedom (df)

Calculation of variance

5. Fifth, the variance measures variation in squared units which is not convenient. To solve this problem, we take the square root of the variance and we finally come to the most meaningful and most widely used measure of variability - the standard deviation – s or SD.

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

s = standard deviation of a sample

x = value of a single unit

$\bar{x}$ = arithmetic mean of a sample

n = number of units in a sample studied

n-1 = degrees of freedom (df)

Calculation of standard deviation

In summary, the standard deviation is a useful index of variability and also can be used to interpret the score of the performance of an individual in relation to others in the sample. It is a stable estimate and is also used in more advanced statistical procedures.

The standard deviation is the preferred measure of distribution's variability but it is appropriate only for variables measured on interval or ratio scale.

Interquartile range

The interquartile range (IQR) is the difference between the third (Q_3) and the first (Q_1) quartiles in a dataset (where quartiles are the values that divide the data into four equal sized parts).

Characteristics of IQR

1. The advantage of the IQR over the range is that it is quite robust to outliers.

2. The IQR is commonly quoted in conjunction with the sample median.

The semiquartile range

The semiquartile range (SQR), used as a term in many statistical texts instead of IQR, is half of the distance between Q_1 and Q_3.

Because these two measures of variability are based on middle cases rather than on extreme scores, they are considerably more stable than the range.

Coefficient of variation

The standard deviation *s* and the variance s^2 have the same measurement units as the mean and because of this they are not appropriate for comparing the relative variability of different distributions where the variables are measured in different units (height in cm, weight in kg, blood pressure in mm mercury, etc.).

This problem can be overcome by calculating another measure of variation called the coefficient of variation (denoted by C_V), also known as relative variability.

It expresses the sample standard deviation as a proportion or percentage of the mean value and can be calculated very easily by the following formula:

$$C_V = \frac{s}{\overline{X}} \times 100$$

The main advantage of the coefficient of variation is its independence of any unit of measurement, and thus, it is useful for comparison of variability in two or more distributions having variables expressed in different units.

For example, if we measure height and weight in a sample, it is not possible to say which variable varies greatly because these two variables have different measurement units. Using the coefficient of variation we can transform the standard variations in comparable units, expressed in percent.

Interpretation of C_V:

1. When the value of C_V is less than 10%, it means that the degree of variation is low and the sample is quite homogeneous.

2. In a situation when $10\% < C_V < 30\%$ - the variation is moderate.

3. When $C_V > 30\%$ the variation is considerable, and this is a clear evidence of heterogeneity of the sample or population under study.

SUMMARY

Measures of central tendency and variability are the two essential measures of location for describing and representing frequency distributions.

The mode and the median are used as measures of central tendency for discrete data, and the mean for continuous data.

The range, the variance, the standard deviation, the inter- and semiquartile range, and the coefficient of variation are measures of variability.

The mean and the standard deviation are the most appropriate for interval or ratio data when the distribution is normal or nearly normal.

The median and the inter- or semi-quartile range are used when the data was measured on an ordinal scale, or when interval or ratio data has a highly skewed distribution.

Practical assignment

Determine the measures of central tendency and spread

Data set 1

X	x - $\bar{x}$	(x– $\bar{x}$)2	
18			
21			
23			
23			
25			
27			
27			
28			
30			
33			
∑ x	∑ (x- $\bar{x}$)	∑ (x– $\bar{x}$)2	

Data set 2

X	x - $\bar{x}$	(x– $\bar{x}$)2	
23			
23			
24			
25			
26			
26			
27			
27			
27			
27			
∑ x	∑(x - $\bar{x}$)	∑(x - $\bar{x}$)2	

Part 2.

THE CONCEPT OF NORMS AND NORMAL GROUPS' LIMITS

Basic principle of a normal distribution

In a normal or nearly normal distribution there are fixed percentages of cases that fall within certain distances from the mean.

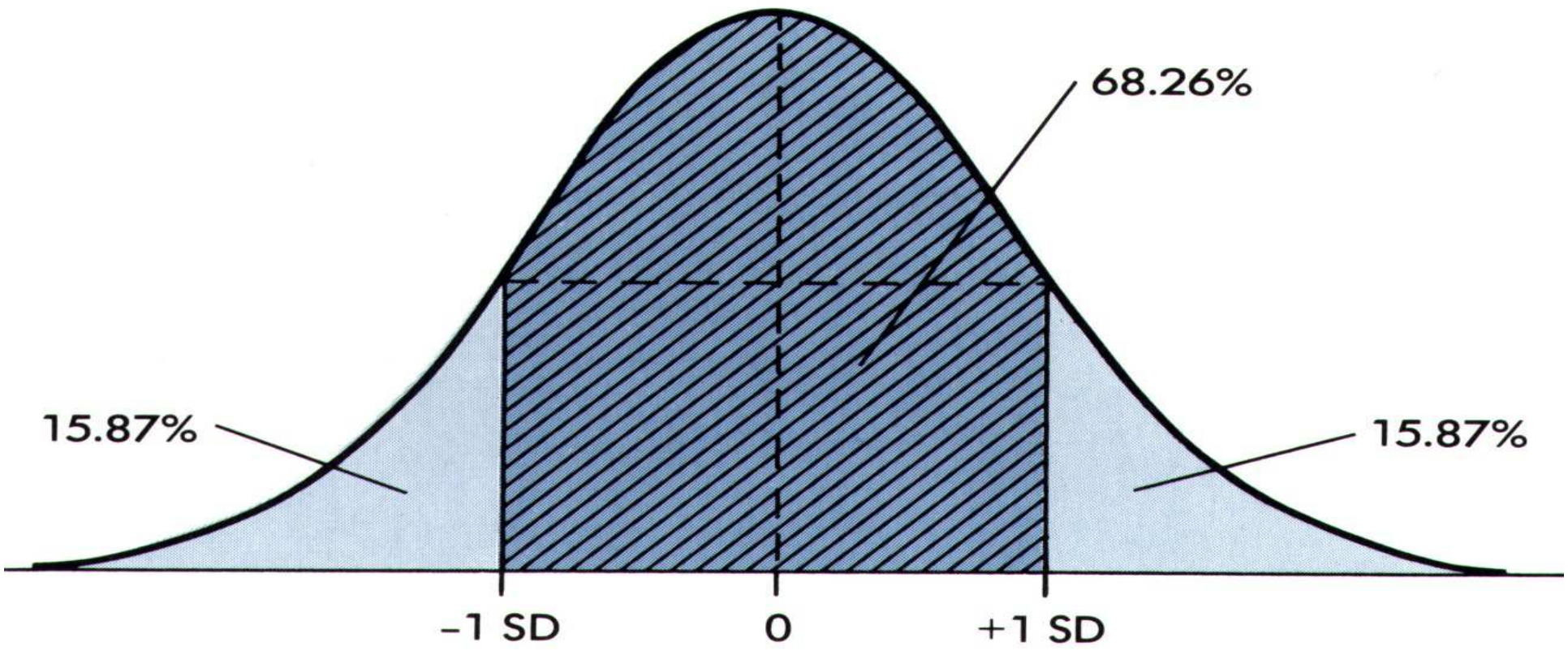
NORMAL DISTRIBUTION

Number of SD from the mean	Results lying inside this (%)	Results lying outside this (%)
0.5	38.3	61.7
1	68.26	31.74
1.64	90	10
1.96	95	5
2.58	99	1
3.00	99,7	0.3
3.29	99.9	0.1

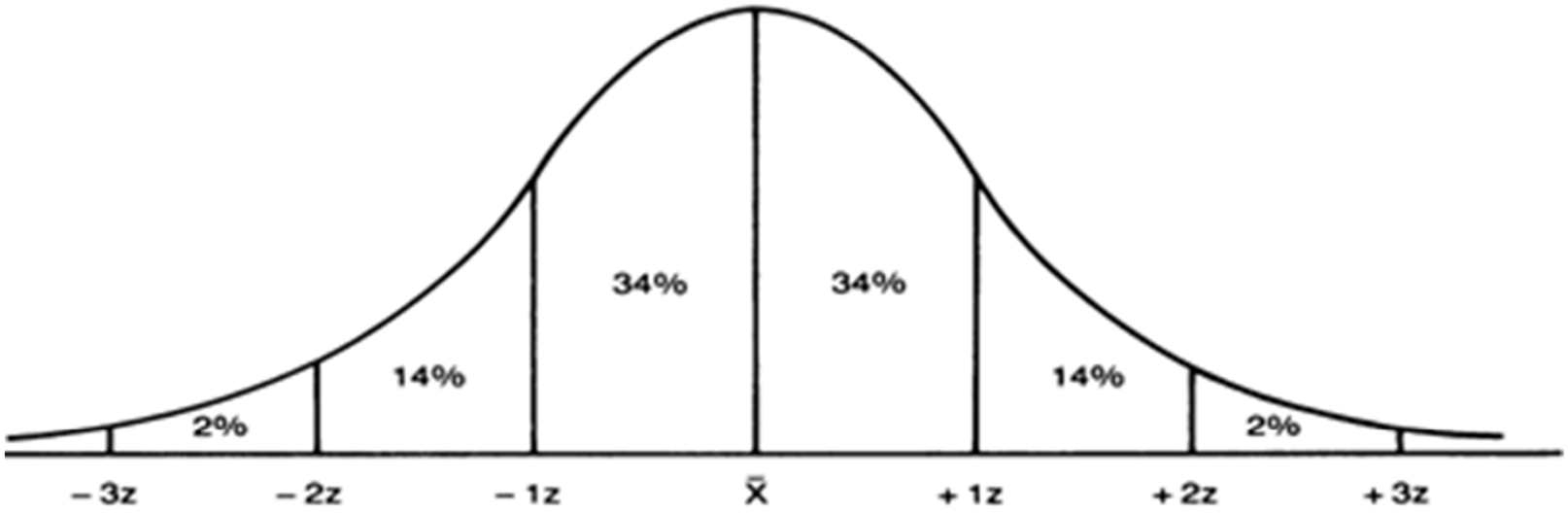
Using this principle, we can easily create "normal limits" and interpret individual scores for clinical and laboratory tests.

Using the basic principle of normal distribution, we can determine the limits of different groups of normality.

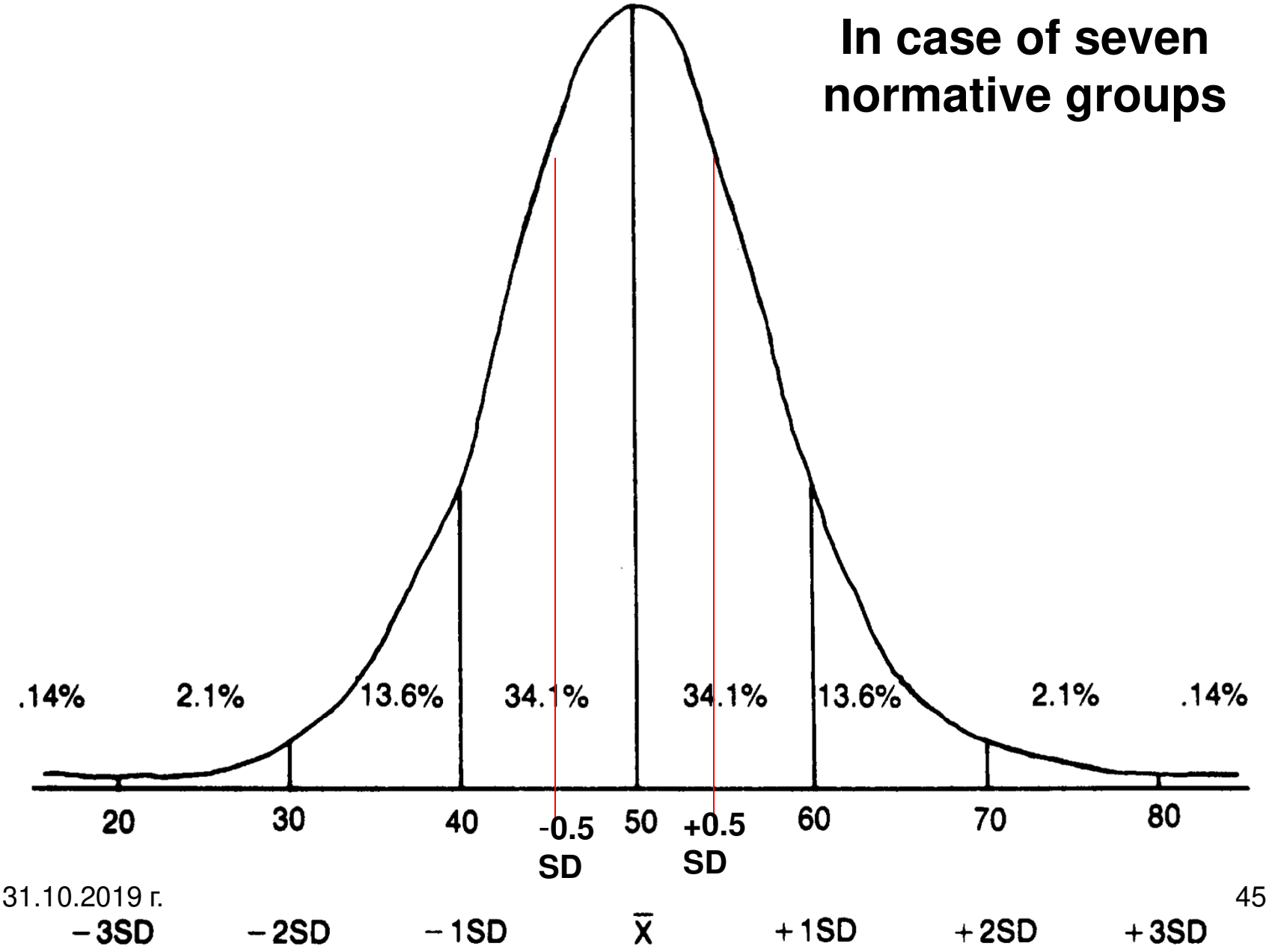
In case of three groups of normality:
normal, above the norm and below the norm
Mean $\pm$ 1 s (SD)



In case of five normative groups



In case of seven normative groups



Practical assignment

Using the basic principles of the normal distribution, determine the limits of 7 normative groups for a variable "weight" for males based on a sample of 1000 males aged 60-69 years with

a mean =80 kg and a standard deviation=8 kg

The Concept of “Norms” or “Normal Limits”

Groups	***Limits***		***% of cases***
Strongly below the norm	Below $\bar{x} - 2s$		2,3
Average below the norm	From $\bar{x} - 1s$ to $\bar{x} - 2s$		13,6
Slightly below the norm	From $\bar{x} - 0{,}5s$ to $\bar{x} - 1s$		15,0
Normal	From $\bar{x} - 0{,}5s$ to $\bar{x} + 0{,}5s$		38,2
Slightly above the norm	From $\bar{x} + 0{,}5s$ to $\bar{x} + 1s$		15,0
Average above the norm	From $\bar{x} + 1s$ to $\bar{x} + 2s$		13,6
Strongly above the norm	Over $\bar{x} + 2s$		2,3

In other words,

- **38,2% percent of all cases fall within 0.5 SD of the mean;**

68,2% percent of all cases fall within 1 SD of the mean;

- 95% of the scores fall within 2 SDs from the mean (exactly 1.96 SD);

- only a handful of cases – about 2% at each extreme.

What can be done when the shape of a distribution is skewed to the left or right?

In such situations it is recommended to use percentiles to determine the limits of different groups of normality.

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).

When the investigator prefers to use seven reference groups the limits of “normal” values are determined by P_{25} and P_{75} whereas P_{50} corresponds to the mean.

Groups of normality using percentiles

Groups	Limits	% of cases falling in a group
Strongly below the norm	Below P_3	3
Moderately below the norm	P_3 to P_{10}	7
Slightly below the norm	P_{10} to P_{25}	15
Normal	P_{25} to P_{75}	50
Slightly above the norm	P_{75} to P_{90}	15
Moderately above the morm	P_{90} to P_{97}	7
Strongly above the norm	Above P_{97}	3

1. If a distribution is negatively skewed, then:

A. the median is greater than the mean
B. the mode is greater than the median
C. the mean is greater than the median
D. both A and B are true
E. none of the above are true

2. In a normal distribution, the mean, the median and the mode:

A. always have the same value
B. the mean has the higher value
C. the mean has the lower value
D. have no particular relationship
E. cannot take the same value

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

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

4. A group of females aged 30-39 years has a mean weight of 60 kg and a standard deviation of s = 5 kg. What are the limits of a "norm' in case of seven normative groups?

A. 65 ÷ 70 kg
B. 57.5 ÷ 62.5 kg
C. 50 ÷ 70 kg

5. Select the statement which you believe to be true. ***The standard deviation:***

A. Is a measure of central tendency (of location).

B. Is a measure of spread which is equal to the range.

C. Is unaffected by outliers.

D. Is an inappropriate measure of spread for skewed data.

6. The square root of the variance is called the standard deviation.

A. True **B.** False

7. Standard deviation indicates the extent to which scores are distributed about the mean.

A. True **B.** False

8. When a distribution consists of very different scores, standard deviation will be relatively large.

A. True **B.** False

9. *Select the statement which you believe to be true.* ***The standard deviation of a set of observations:***

A. Is a measure of central tendency (of location).

B. Is the square root of the variance.

C. Is a measure of spread which is equal to the range.

D. Is unaffected by outliers.

10. *Select the statement which you believe to be true.* ***The standard deviation of a set of observations:***

A. Is a measure of central tendency (of location).

B. Has the same units of measurement as the raw data.

C. Is a measure of spread which is equal to the range.

D. Is unaffected by outliers.

11. The smaller the variance the less spread of the data around the mean.

A. True B. False

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

A, True B. False

13. The mean of the sample means is

A**.** A biased estimator of the population

B. An unbiased estimator of the population mean

C. Neither biased nor unbiased

14. Select all of the following statements which you believe to be true. The arithmetic mean of a set of values:

A. Is a particular type of average.

B. Is a useful summary measure of location if the data are skewed to the right.

C. Is always greater than the median.

15. Select all of the following statements 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.

16. Select all of the following statements which you believe to be true. The median:

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

B. Is a useful summary measure when the data are skewed to the right.

C. Is greater than the arithmetic mean when the data are skewed to the right.

17***. Select all of the following statements which you believe to be true.***

A. The first percentile has 99% of the observations in the ordered set below it.

B. The first decile is equal to the 90th percentile and has 10% of the observations in the ordered set below it.

C. The median is equal to the 50th percentile.

18. Select all of the following statements which you believe to be true.

A. The median is equal to the 50^{th} percentile.

B. The first percentile has 99% of the observations in the ordered set below it.

C. The first decile is equal to the 90^{th} percentile and has 10% of the observations in the ordered set below it.

19. Select all of the following statements which you believe to be true.

A. The range is the difference between the 1st and 99th percentiles.

B. The interquartile range lies between the 1st and 3rd quartiles.

C. Quartiles divide the data set into 100 equally sized groups

20. The more dispersed, or spread out, a set of scores is:

A. The greater the difference between the mean and the median
B. The greater the value of the mode
C. The greater the standard deviation
D. The smaller the interquartile range

21. Select all of the following statements which you believe to be true. The standard deviation of a set of observations:

A. Is unaffected by outliers.
B. Has the same units of measurement as the mean.
C. Is a measure of spread which is equal to the range.

22. The range is calculated by adding the lowest score to the highest score in a distribution.

A. True B. False

23. Standard deviation indicates the extent to which scores are distributed around the mean.

A. True B. False

24. The mean must have a value equal to one of the scores in the distribution.

A. True B. False

25. When a distribution consists of very similar scores, standard deviation will be relatively low.

A. True B. False

26. The range is the simplest indicator of variability.

A. True B. False

27. 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 50th percentile rank have different values.

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

28. The mean height of a student group is 167 cm. Assuming that height is normally distributed this enables us to deduce that:

A. Approximately half of all students are taller than 167 cm

B. Approximately half of all students are shorter than 167 cm

C. Both statements are true

29. The interquartile range of the following set of data
3 3 4 5 6 7 8 9 9 10
is equal to:
A. 5.0*
B. 4.5
C. 6.0
D. 9.0

30. In a normal distribution the percentage of cases falling between the mean ± 1 SD is:
A. 16.8%
B. 33.6%
C. 34.1%
D. 68.3%*

Answers:

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