Kenya Medical Training College -Port Reitz Campus Department of Clinical Medicine Year Two Semester One Descriptive Statistics 22nd October 2020

Willis J. Opalla

Descriptive Statistics

- Learning Objective

  To demostrate knowledge of descriptive statistics and be able to utilize it in addressing emerging needs in healthcare practice.

Learning Outcomes

- By the end of this session, you should be able to
  1. Define descriptive statistics.
  2. Explain the components of descriptive statistics.
  3. Define the measures of central tendency.
  4. Explain the component measures of central tendency.
  5. Define each measure of central tendency.
  6. Utilize each measure of central tendency to solve problems requiring knowledge of health statistics.

Types of Statistics

1. Descriptive Statistics
2. Inferential Statistics.

- **Statistics**
  - The branch of mathematics that transforms data into useful information for decision makers.

Descriptive Statistics

Collecting, summarizing, and describing data

Inferential Statistics

Drawing conclusions and/or making decisions concerning a population based only on sample data

Introduction to Descriptive Statistics

- Definition
  - Descriptive statistics are the type of statistics that analyse data through description or summarization in a meaningful way and show patterns of relationship among the data.
  - Descriptive statistics do not, allow making conclusions beyond the data analyzed or reach conclusions regarding any hypotheses.
  - They are simply a way to describe data.

Introduction to Descriptive Statistics

- Importance of Descriptive statistics
  - If data are simply presented in raw form it would be hard to understand what they are showing.
  - Descriptive statistics therefore enables presentation of the data in a more meaningful way, to allow simpler interpretation of the data.

Importance of Descriptive Statistics

- Example:
  - For the coursework scores of 100 of students, descriptive statistics will help determine:
  1. the overall performance of those students.
  2. the distribution or spread of the marks.
- There are two general types of statistic that are used to describe data:
  1. Measures of central tendency.
  2. Measures of dispersion (or spread or scatter.)

Descriptive Statistics

- In descriptive statistics a group of data is summarized using a combination of tabulated description (i.e., tables), graphical description (i.e., graphs and charts) and statistical commentary (i.e., a textual discussion of the results).

Importance of Descriptive Statistics

- Measures of central tendency
  - Are ways of describing the central position of a frequency distribution for a group of data.
  - e.g. simplify the distribution and pattern of marks scored by the 100 students from the lowest to the highest.
  - This central position can be described using the mode, median, and mean.

Importance of Descriptive Statistics

- Measures of dispersion (or spread or scatter)
  - Ways of summarizing a group of data by describing how spread out or scattered they are.
  - e.g, the mean score of the 100 students may be 65%.
  - However, not all students will have scored 65%.
  - Their scores will be spread out with some lower and others higher.
- Measures to describe this spread, include the range, quartiles, absolute deviation, variance and standard deviation.

Measures of Central Tendency

- A measure of central tendency is a single value that attempts to describe a set of data by identifying the central position within that set of data.
- For this reason, the measures of central tendency are also termed *measures of central location* or *summary statistics.*

Measures of Central Tendency

- Measures of central tendency convey information regarding the average value of a set of data.
- The three most commonly used measures of central tendency are the ***mean***, the ***median*** and the ***mode.***
- In each of the measures of central tendency, there is a single value that is considered to be typical of the set of data as a whole.

Mean

- Arithmetic Mean
  - This is the most familiar measure of central tendency.
  - It measures the “average” and may be referred to simply as the mean.
  - It is obtained by adding all the values in a population or sample and diving by the number of values that are added.
- Others:
  - Geometric Mean, GM.
  - Harmonic Mean, HM.

Mean

- Example: Mean age of 189 subjects

30 34 35 37 37 38 38 38 38 39 39 40 40 42 42 43 43 43 43 43 43 44 44
44 44 44 44 44 45 45 45 46 46 46 46 46 46 47 47 47 47 47 47 48 48 48
48 48 48 48 49 49 49 49 49 49 49 50 50 50 50 50 50 50 50 51 51 51 51
52 52 52 52 52 52 53 53 53 53 53 53 53 53 53 53 53 53 53 53 53 53 53
54 54 54 54 54 54 54 54 54 54 54 55 55 55 56 56 56 56 56 56 57 57 57
57 57 57 57 58 58 59 59 59 59 59 59 60 60 60 60 61 61 61 61 61 61 61
61 61 61 61 62 62 62 62 62 62 62 63 63 64 64 64 64 64 64 65 65 66 66
66 66 66 66 67 68 68 68 69 69 69 70 71 71 71 71 71 71 71 72 73 75 76
77 78 78 78 82

- Mean age = $\frac{30 + 34 + 35 + \ldots + 78 + 82}{189} = 55.032$
- The three dots in the numerator represent the values not shown in order to save space.

Ages of 189 Subjects who Participated in a Study on Smoking Cessation

Subject No.	Age	Subject No.	Age	Subject No.	Age	Subject No.	Age
1	48	49	38	97	51	145	52
2	35	50	44	98	50	146	53
3	46	51	43	99	50	147	61
4	44	52	47	100	55	148	60
5	43	53	46	101	63	149	53
6	42	54	57	102	50	150	53
7	39	55	52	103	59	151	50
8	44	56	54	104	54	152	53
9	49	57	56	105	60	153	54
10	49	58	53	106	50	154	61
11	44	59	64	107	56	155	61
12	39	60	53	108	68	156	61
13	38	61	58	109	66	157	64
14	49	62	54	110	71	158	53
15	49	63	59	111	82	159	53
16	53	64	56	112	68	160	54
17	56	65	62	113	78	161	61
18	57	66	50	114	66	162	60
19	51	67	64	115	70	163	51
20	61	68	53	116	66	164	50
21	53	69	61	117	78	165	53
22	66	70	53	118	69	166	64
23	71	71	62	119	71	167	64
24	75	72	57	120	69	168	53
25	72	73	52	121	78	169	60
26	65	74	54	122	66	170	54
27	67	75	61	123	68	171	55
28	38	76	59	124	71	172	58

Ages of 189 Subjects who Participated in a Study on Smoking Cessation

Subject No.	Age	Subject No.	Age	Subject No.	Age	Subject No.	Age
29	37	77	57	125	69	173	62
30	46	78	52	126	77	174	62
31	44	79	54	127	76	175	54
32	44	80	53	128	71	176	53
33	48	81	62	129	43	177	61
34	49	82	52	130	47	178	54
35	30	83	62	131	48	179	51
36	45	84	57	132	37	180	62
37	47	85	59	133	40	181	57
38	45	86	59	134	42	182	50
39	48	87	56	135	38	183	64
40	47	88	57	136	49	184	63
41	47	89	53	137	43	185	65
42	44	90	59	138	46	186	71
43	48	91	61	139	34	187	71
44	43	92	55	140	46	188	73
45	45	93	61	141	46	189	66
46	40	94	56	142	48		
47	48	95	52	143	47		
48	49	96	54	144	43		

The General Formula for the Mean

- Let X be the random variable of interest which in the smoking cessation study is age.
- Specific values of a random variable are designated by the small letter *x*.
- To distinguish one value from another, a subscript is attached to the *x* so that x_1 is the first item in the data set, x_2 is the second, …, and x_{189} is the 189^{th}, etc.

The General Formula for the Mean

- From the table, $x_1 = 48$, $x_2 = 35$, …, $x_{189} = 66$

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

- where,

  μ is the population mean,

  ∑ is the summation sign,

  x_i is the ith item in the data set and

  N is the total number of items in the population.

- $\sum_{i=1}^{N}$ is an instruction to add all values of the variable from the first to the last.

The Genaral Formula for Sample Mean

- Since a sample is a subset of a population and provides the characteristics of the population to be researched on,
- Sample mean, $\bar{x}$ (read as *x bar*) is calculated using the formula:

$$\overline{X} = \frac{\sum_{i=1}^{n} Xi}{n}$$

- where, n is the total number of items in the sample.

Sample Mean

- Example
  - To compute the mean age of 10 subjects selected through simple random sampling from the population of 189 respondents in the smoking cessation study.
  - The ages of the 10 subjects in the sample are $x_1 = 43$, $x_2 = 66$, $x_3 = 61$, $x_4 = 64$, $x_5 = 65$, $x_6 = 38$, $x_7 = 59$, $x_8 = 57$, $x_9 = 57$, $x_{10} = 50$,

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{43 + 66 + \cdots + 50}{10} = 56$$

Properties of the Mean

1. Uniqueness. For a given set of data there is only one arithmetic mean.
2. Simplicity. The arithmetic mean is easily understood and easy to compute.
3. Since each and every value in a set of data enters into the computation of the mean, it is affected by each value. Extreme values therefore have an influence on the mean and in some cases, can distort it that it becomes undesirable as a measure of central tendency.

Mean of Grouped Data

- In the case of the grouped data , arithmetic mean is calculated assuming that each observation in a class interval is equal to the midpoint of the class interval.

$$m = \frac{\sum x}{n}$$

- Where m is the mean, $\sum x$ the sum of all items in the data set, n is sample size.

Mean of Grouped Data

- Cumulative frequency, from the frequency didtribution table can also be used to calculate the arithmetic mean: $m = \frac{\sum fx}{\sum f}$ or $m = \frac{\sum fx}{n}$
- Where

m is the mean,

f is frequency,

x is the matching score for that frequency,

$\sum fx$ the sum of the products of frequency and the score (or item) in the data set,.

Geometric Mean, GM.

- Definition
  - The average of a set of products, commonly calculated to determine the performance results of an investment or portfolio.
  - Defined as "the *nth* root product of *n* numbers".
- Must be used when working with %ges, which are derived from values, while the standard *arithmetic mean* is calculated from the values themselves.

Geometric Mean, GM.

- GM is more suitable as a measure of central tendency when values change exponentially.
- If there are only 2 observations, then the GM is the square root of the product of two observations.
- If they are 3 observations, then it is the cube root of the product of the three observations.

Geometric Mean, GM.

- If there are n observations, then GM will be the n^{th} root of the product of the n observations.

$$GM = \sqrt[n]{(x1)(x2)\ldots\ldots(xn)}$$

- The computation is not as menacing as the above equation!
- Simplified, $\text{Log}_{10}GM = \frac{\sum(\log x)}{n}$
- The logarithm of the geometric mean is the arithmetic mean of the logarithms of individual observations.

Geometric Mean, GM.

- The number of bacteria ($x10^3$) observed in an experiment at hourly intervals are as follows: 10, 25, 76, 148, 302

$$GM = \sqrt[5]{10 \times 25 \times 76 \times 148 \times 302}$$

$$= \sqrt[5]{33,968,960}$$

$$= 61.07 \times 10^3$$

Geometric Mean, GM, Cont…

- The same result can be obtained by taking the arithmetic mean of the log of these 5 values and then determining the antilog as follows:

Value	Log
10	1.0000
25	1.3979
76	1.8808
148	2.1703
302	2.4800

Total divide by 5 = 8.9290/5 = 1.7858

GM = antilog (1.7858) = 61.07

Harmonic Mean, HM.

- Definition:
  - A type of numerical **average** that is calculated by dividing the number of observations by the reciprocal of each number in the series.
  - Thus, the harmonic mean is the reciprocal of the *arithmetic mean* of the reciprocals of values in a data set.

Harmonic Mean, HM.

- Harmonic mean is used where reciprocals of the actual values seem more useful to determine the central tendency.
- e.g. it has been suggested that the sensitivity to detect clusters of observation is increased by measuring reciprocal of distance rather than using distance directly.

$$HM = \frac{1}{\left(\frac{1}{n}\right)\Sigma\left(\frac{1}{x}\right)} \text{ or } \frac{n}{\Sigma\left(\frac{1}{x}\right)}$$

Harmonic Mean, HM.

- Taking the reciprocal of both sides,

$$1/HM = \frac{\Sigma\left(\frac{1}{x}\right)}{n}$$

- The reciprocals of the harmonic mean is the mean of the reciprocals of the individual observations.

Harmonic Mean, HM.

- Example:
  - The distance (Km) of 17 chronic bronchitis patients from a factory are: 0.8, 1.2, 3.2, 1.6, 1.1, 2.7, 2.1, 1.3, 0.9, 1.3, 1.5, 1.1, 0.9, 1.8, 2.2, 2.4
  - Thus, $\frac{1}{HM} = \frac{\text{Sum } (1/x)}{n}$

    $\text{Sum} = \frac{12.98}{17} = 0.7635$

    HM= 1.31 Km.
  - If using the distance directly, then

    $HM = \frac{26.8}{17} = 1.58$ Km.

Median

- The median is the middle observation when all observations in a set of data are ranked.
- It divides the set of data into two equal partsh such that the number of values equal to or greater than te median is equal to the number of values equal to or less than the median.
- If odd number of values: the median will be the middle value after arranging all the values in order of magnitude.
- For even number of values, then the mean of the two middle values after arranging all the values in the order of their magnitude. Median = (n + 1)/2th item.

Calculation of the Median

- Median = $\frac{(n + 1)}{2}$th item.
- Example, to determine the median of the data for the smoking cessation study, the data is arranged in ascending order from the least to the largest.
- Then:

$$\text{The middle value is the } (n+1)/2 = (189+1)/2 = 190/2 = 95\text{th one.}$$

- Counting from the first which is smallest age up to the 95^{th} value, the 95^{th} value is found to be 54 years.
- The median age of the 189 respondents is 54 years.

Calculation of the Median Cont…

- Worked Example 2
  - To determine the median age of the sample of 10 representing the larger population of 189 subjects:
  - The 10 ages are arranged in order of magnitude from smallest to largest, i.e. 38, 43, 50, 57, 57, 59, 61, 64, 65, 66.
  - Since this data set has an even number of ages, i.e. 10, there is no middle value. The two middle values, however, are 57 and 59.
  - The median age, then, is $\frac{(57+59)}{2}$ = 58 years.

Median for Grouped Data

- When data is grouped, the median can be determined using the formula:

$$\text{Median} = L + \frac{\left(\frac{n}{2} - F\right)C}{f}$$

- Where,

  L is the lower limit of the median class

  n is the total number of observations,

  F is the number of observations up to the median class,

  f is the frequency in the median class

  C class interval of the median class.

Properties of the Median

1. Uniqueness, there is only one median for a given set of data
2. Simplicity, it is easy to calculate.
3. It is not as drastically affected by extreme values as is the mean.
4. It is the more representative measure when the data set is skewed (i.e. has extreme values or outliers).

Mode

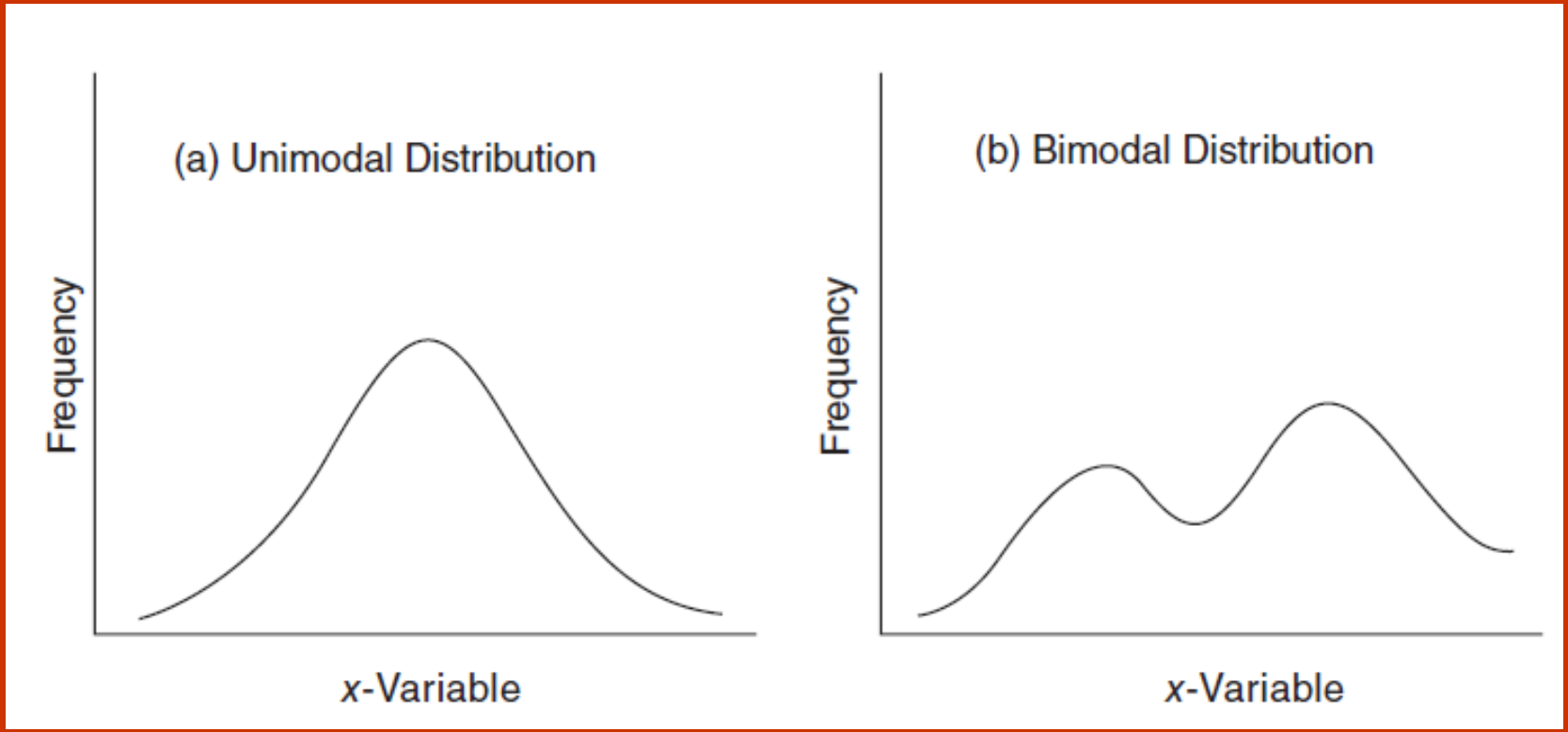
- Definition:
  - The mode is the value (or values) which appear(s) most frequently in a set of data.
  - If all the values are different there is no mode; on the other hand, a set of data may have more than one mode.
  - It is determined by counting the number of times each value appears in a data set.

Mode

- Example
  1. A data set with more than one mode:
  - The ages of 10 members of staff laboratory whose ages are 20, 21, 20, 20, 34, 22, 24, 27, 27 and 27 years.
  - There are two modes, 20 and 27 in this data set, i.e. it is a bimodal data set.
  2. The sample consisting of 10, 21, 33, 53, and 54 has no mode since all the values are different.

Modal Distributions

- Unimodal or Bimodal or Multimodal distribution



Bimodal Distribution

- Although the mode can sometimes be a good measure of central tendency, at least in the case of the symmetric bimodal distribution, the natural center is in the “ middle ” between the two modes at where there is a trough.
- That middle of the valley between the peaks is where the median and mean are located.

Mode

- The mode may be used for describing qualitative data.
- e.g., suppose patients in a mental health clinic in one given year had one of the diagnoses: mental retardation, organic brain syndrome, psychosis, neurosis and personality disorder.
- The diagnosis occurring most frequently in the group of patients would be called the modal diagnosis.

Mode

- For a moderately asymmetric distribution, the mode can be calculated using the following empirical relationship.

  Mode = 3 median – 2 mean, or using formula:

$$\text{Mode} = L_M + \frac{d1C}{d1 + d2}$$

- Where,

  L_M = Lower limit of modal class

  d_1 = frequency in modal class *minus* frequency in the preceding class.

  d_2 = frequency in modal class minus frequency in the succeeding class

  C = class interval of the modal class.

Properties of the Mode

1. Sometimes, it is not unique, i.e. there may be a multimodal distribution.
2. It may be used for describing qualitative data.

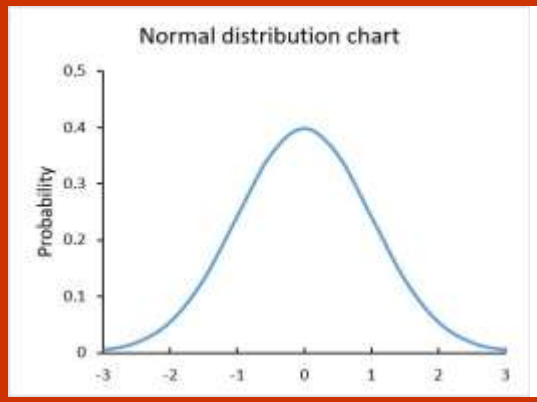
The Bell-Shaped Curve

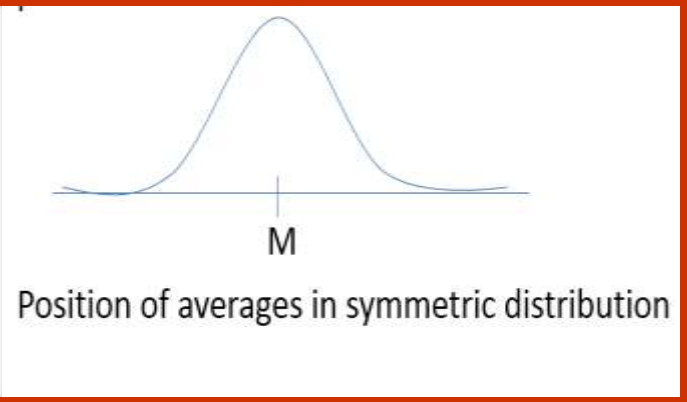
- An attractive property of a data distribution occurs when the mean, median, and mode are all equal.
- The well-known “bell-shaped curve” is a graphical representation of a distribution for which the mean, median, and mode are all equal.
- Much statistical inference is based on this distribution, the most common of which is the normal distribution.

The Normal Distribution Curve

- For symmetric unimodal distribution:

  Mean = Median = Mode
- Bell-shaped.





- Symmetric distribution, i.e. the left half of the frequency polygon is a mirror image of its right half.

Skewness

- Definition
  - Skewness is the extent of asymmetry of the distribution of data.
- Data distribution may be classified on the basis of whether they are symmetric or asymmetric.
- If the graph (histogram or frequency polygon) of a distribution is asymmetric, the distribution is said to be skewed.

Skewness

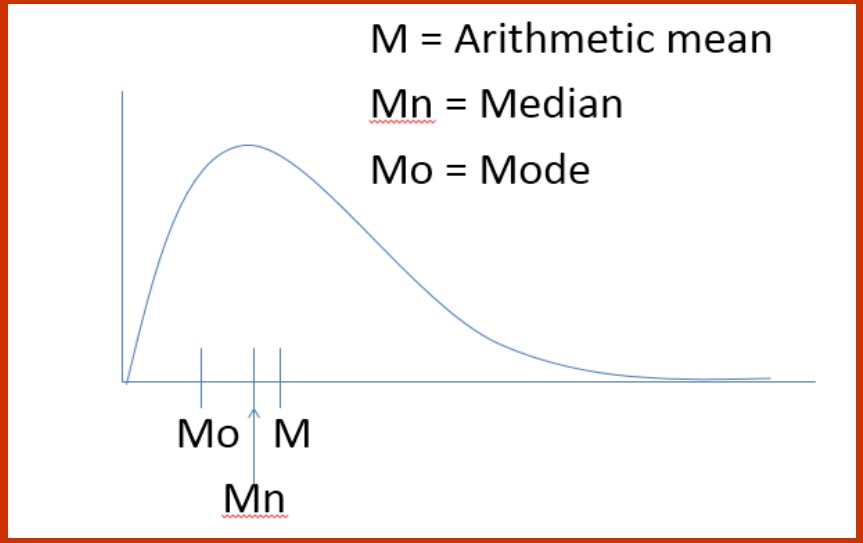
- A distribution is symmetric if the left and right halves of its graph (histogram or frequency polygon) will be a mirror images of each other.

- The distribution is asymmetric when the left half and right half of the graph of a distribution are not mirror images of each other.

Skewness

- If a distribution is not symmetric because its graph extends further to the right than to the left, i.e. if it has a long tail to the right, then that distribution is skewed to the right or is positively skewed.
- For unimodal distributions that are right skewed:

  mean > median > mode.

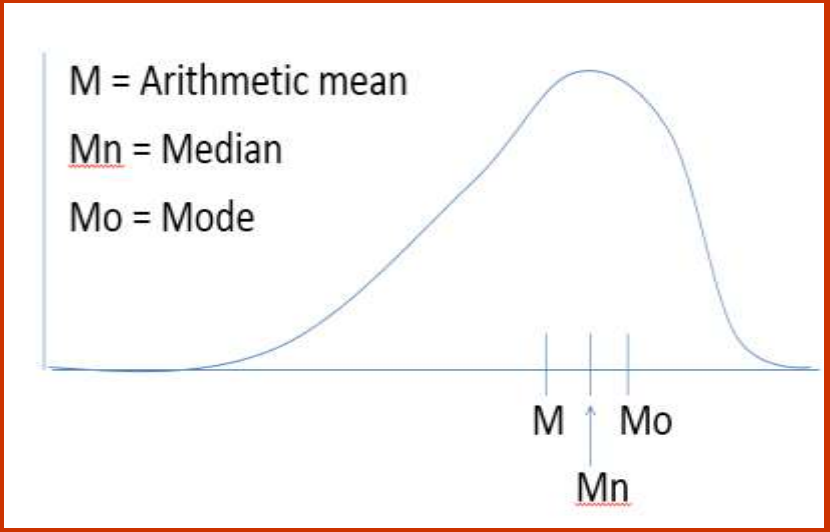


- A distribution is positively skewed, if its mean > its mode.

Skewness

- If a distribution is not symmetric because its graph extends further to the left than to the right,
- i.e. if it has a long tail to the left, then that distribution is said to be skewed to the left or is negatively skewed.
- For unimodal distributions that are left skewed:

  mean < median < mode.



- A distribution is negatively skewed, if its mean < its mode.

Summary

- Descriptive statistics comprise measures of central tendency and measures of dispersion.
- Measures of central tendency are values that describe a set of data by identifying the central position within that set of data.
- They comprise, the mean, median and mode.
- A distribution is positively skewed, if its mean > its mode.
- A distribution is negatively skewed, if its mean < its mode.

References

- Chernick, M. R. (2011) *The Essentials of Biostatistics for Physicians, Nurses and Clinicians*, First Edition, John Wiley & Sons, Hoboken.
- Holmes, L, Jr. (2018) *Applied Biostatistical Principles and Concepts: Clinician's Guide to Data Analysis and Presentation*, Routledge, Taylor & Francis group, New York.