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

Willis J. Opalla

Statistical Data

- Learning Objective

  To demonstrate understanding of measures of dispersion and apply it in health statistics..

Learning Outcomes

- By the end of this session, you should be able to
  1. Define the measures of dispersion.
  2. Explain measures of dispersion.
  3. Analyze data using the range.
  4. Explain the concepts of quartiles and percentiles.
  5. Analyze data using variance, standard deviation and coefficient of variation.
  6. Explain kurtosis in relation to normal distribution.

Basic Terminologies

- **Variable:** The measured characteristics of the research problem that is under observation.
- **Population:** The largest collection of items or entities with common observable characteristics, that are of research interest at a particular time.
- **Sample:** A subset of a population that represents and provides the characteristics of the population to be researched on.

Basic Terminologies Cont…

- **Parameter**: A measurable characteristic that assumes different values in a population.

- **Statistics**: A measurable characteristic that assumes different values in a sample.

Some Commonly Used Notations

Quantity	Parameter	Statistic
Mean	μ	$\bar{x}$
Variance	σ^2	s^2
Standard deviation	σ	s
Proportion	Π	p

Measures of Dispersion

- Definition
  - A measure of dispersion is a statistical description of the degree or amount of variability present in a set of data.
  - If all the values are the same then there is no dispersion; if they are all not the same, dispersion is present in the data.
  - Other terms used synonymously with dispersion include *variation*, *spread* and *scatter.*

Measures of Dispersion

- Examples
  1. Range
  2. Interquartile range
  3. Variance
  4. Standard deviation
  5. Coefficient of Variation

Range

- The range is the difference between the largest and smallest value (observation) in a data set.
- Range (R) = $x_{max} - x_{min}$
- x_{max} is the largest observation,
- x_{min} is the smallest observation.
- Range is easy to calculate and understand.
- Is based on only two observations and tends to increase with sample size.
- Difficult for mathematical manipulation.

Range

- It is a poor measure of dispersion because it takes into account only two values.
- Since the range, expressed as a single measure, imparts minimal information about data, is of limited use, it is preferable to express the range as a pair.

Percentiles and Quartiles

- These descriptive measures are called location parameters or measures of location because they can be designated certain positions on the horizontal axis when the distribution of a variable is graphed.
- Given a set of *n* observations $x_1, x_2, \ldots\ldots x_n$, the P^{th} percentile *P* is the value of *X* such that *p* percent or less of the observations are less than *P* and (100-*p*) percent or less of the observations are greater than *P*.
- The 10th percentile is designated as P_{10} whereas the 70th is designated as P_{70}

Percentiles

- Percentiles:
  - Observations which divide a set of data that has been ranked into 100 equal parts.
  - e.g. 1st percentile: 1% of the data is less than or equal

    to this value.

    10th percentile:10% of the data is less than or equal to this value.

    50th percentile: 50% of the data is less than or equal to this value.

Percentiles and Quartiles

- The 25th percentile is often referred to as the *first quartile* and denoted Q_1.
- The 50th percentile (the median) is referred to as the second or *middle quartile* and written Q_2.
- The 75th percentile is referred to as the *third quartile*, Q_3.
- Interquartile range (IR) is the difference between the third and first quartiles: that is,
  - **IR =** $\boldsymbol{Q_3 - Q_1}$
- IR gives an idea on where the data is located and is a measure of location.

Interquartile Range (IR)

- Quartiles:
  - In the data set with observations ranked into 100 equal parts.
  - The 100 equal parts can be divided into
  1. The lower quartile or 1st quartile or Q1 or 25th percentile.
  2. Q3 or upper quartile or 3rd quartile or 75th percentile.
- Interquartile Range , IR = Q3 – Q1

Interquartile Range

- Determination of IR
  - Rank of Q1 = $\frac{n}{4}$, Rank of Q3 = $\frac{3n}{4}$
  - IR encloses the central 50% of observations.
  - It is not based on all observations.
  - Can be used to select cut-off points during development of clinical standards.
- The semi-interquartile range can be calculated from a set of observations, by SIR = $\frac{Q3 - Q1}{2}$.

Variance

- A measure of dispersion.
- Measures variability of values or observations in a data set.
- The variance is the average deviation of each number from its mean.
- For the observations 1, 2, 3, the mean is 2, hence Variance , $\sigma^2 = \dfrac{(1-2)^2 + (2-2)^2 + (3-2)^2}{3}$

$$= 0.667$$

Variance in a Population

- Using summation notation, the variance in a population:

$$\sigma^2 = \frac{\sum(x-\mu)^2}{N}$$

where μ is the mean and N is the number of observations.

Variance in a Sample

- In a sample, the variance is:

$$s^2 = \frac{\sum(x\text{-M})^2}{N}$$

where M is the mean of the sample and s^2 is a biased estimate of σ^2.

- The more common formula for variance in a sample is $s^2 = \frac{\sum(x\text{-M})^2}{N-1}$ or $s^2 = \frac{\Sigma(X-\overline{X})^2}{n-1}$
- Subtracting 1 from n makes sample variance an unbiased estimate of population variance, σ^2.

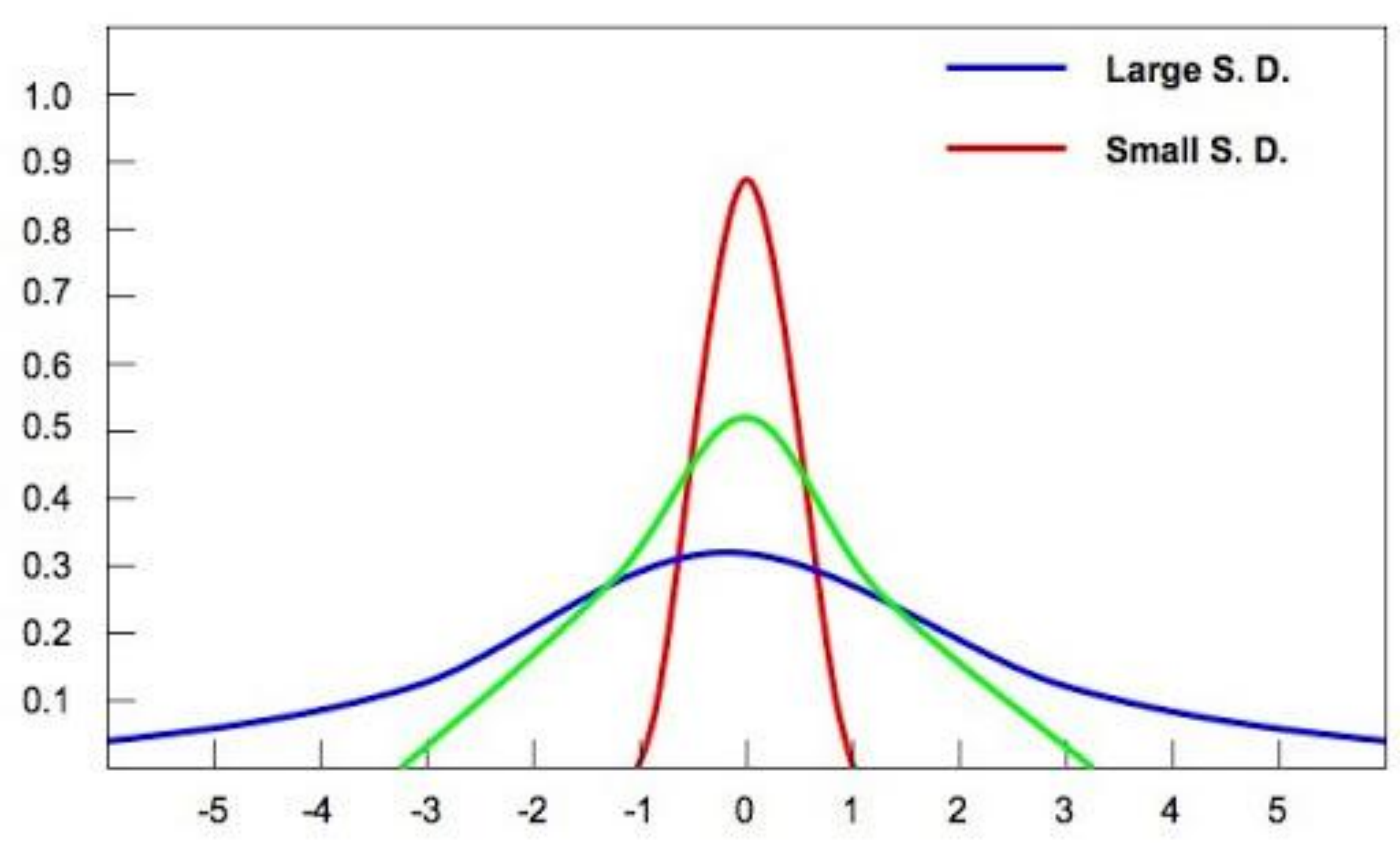
Variance in a Sample

- The formula $s^2 = \dfrac{\sum(x\text{-}\textsc{m})^2}{N - 1}$ or $s^2 = \dfrac{\Sigma(X - \overline{X})^2}{n - 1}$
  - Gives an unbiased estimate of σ^2.
  - Since samples are used to estimate populations, s^2 is the most commonly used unit for variance.
- Calculation of variance is the first step in calculation of standard deviation, *s*.
- Calculation of variance is important in many statistical analyses.

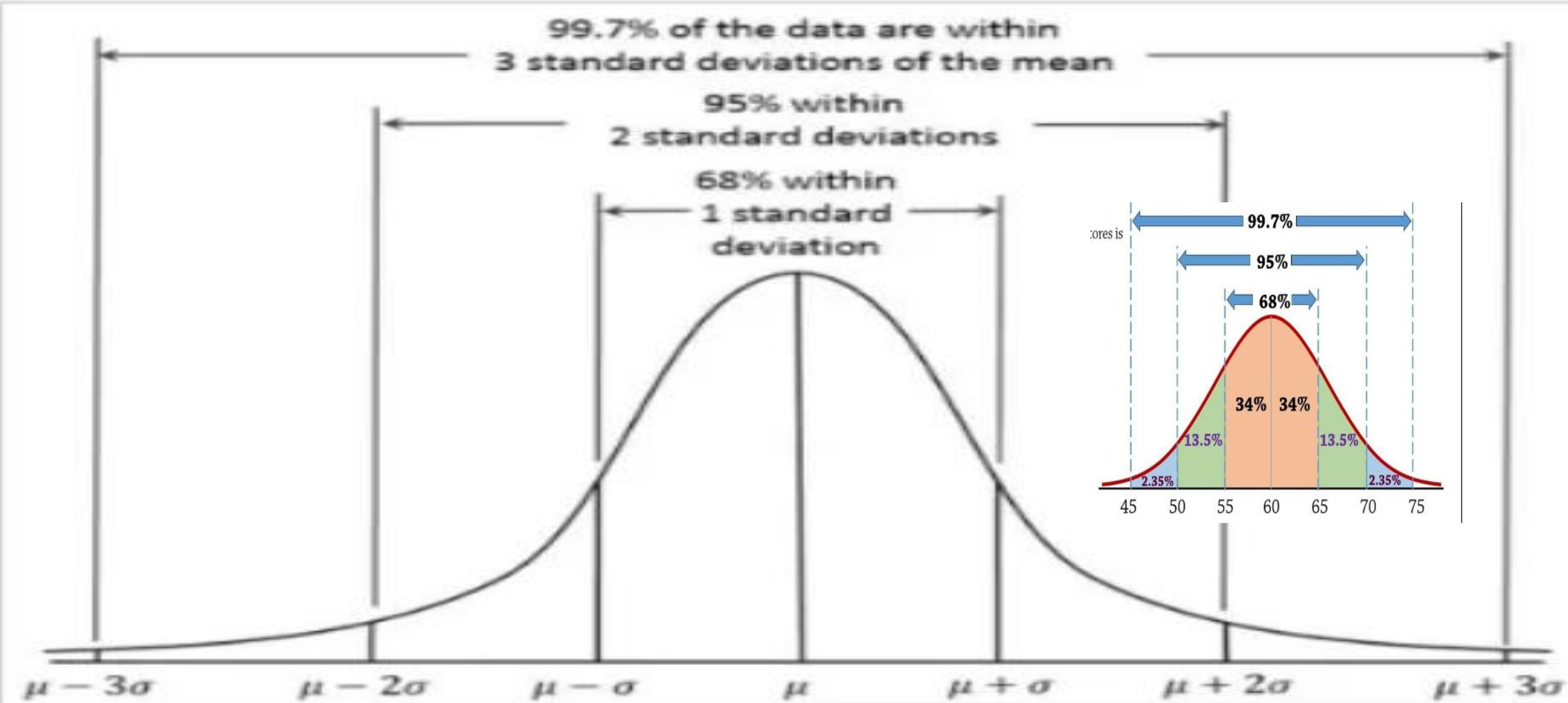
Standard Deviation

- Definition:
  - Standard deviation is a measure of how spread out (or scattered) the observations in a set of data are from the mean.
  - A small standard deviation indicates that the values tend to be close to the mean of the set, while a large standard deviation indicates that the values are spread out over a wider range.
- It is the most commonly used measure of dispersion.

Small vs Large Standard Deviation



1, 2 and 3 Standard Deviations



For the normal distribution, the values less than one standard deviation away from the mean account for 68.27% of the set; while two standard deviations from the mean account for 95.45%; and three standard deviations account for 99.73%.

Calculation of Standard Deviation

- From a data set 1, 3, 4, 6, 9, 19.
  - Step 1: Calculate the mean. Mean = $\frac{(1+3+4+6+9+19)}{6}$.

    = 7
  - Step 2: Subtract the mean from every number in the data set to get a list of deviations. i.e. 1 – 7, 3 -7, 4 – 7, etc. This gives -6, -4, -3, -1, 2, 12.
  - Step 3: Square each number in the list of deviations, i.e. $-6^2 = 36$, $-4^2 = 16$, etc. Add up all the resulting squares to get their total sum. (36 + 16 + 9 + 1 + 4 + 144 = 210).

Calculation of Standard Deviation

- Step 3 cont…: Having obtained the sum of the squares. (36 + 16 + 9 + 1 + 4 + 144 = 210). Divide the sum by one less than the number of items in the data set, i.e.

$$\frac{210}{6-1} = \frac{210}{5} = 42$$

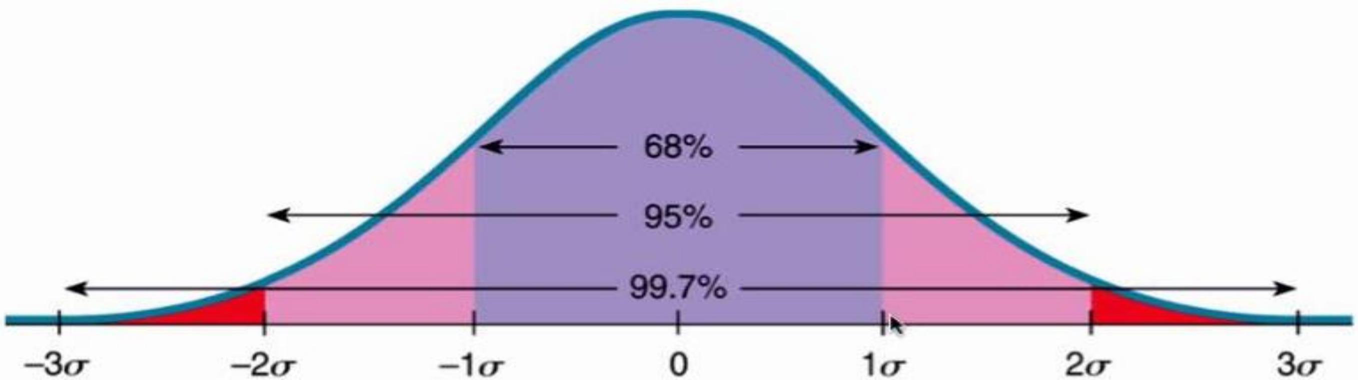
- Step 4: Calculate the square root of the resulting number, i.e. Standard deviation, $s = \sqrt{42} = 6.48$.

- The extent of spread or scatter or variability of each observation in the data set from the mean is 6.48.

The Empirical 68-95-99.7 Rule for Normal Distribution

The 68 – 95 – 99.7 Rule

- In a Normal model:
  - About 68% of the values fall within 1 standard deviation of the mean
  - About 95% of the values fall within 2 standard deviations of the mean
  - About 99.7% of the values fall within 3 standard deviations of the mean ☺



Standard Deviation

- The formula for calculation of standard deviation is

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

Where

s is the standard deviation,

x is each observation in the data set,

$\bar{x}$ is the mean,

n is the number of all observations in the data set (the sample size).

Standard Deviation and Variance

- S^2 is the symbol for variance, S: sample standard deviation (SD) and σ: population SD.
- Hence standard deviation = √variance, i.e. $S = \sqrt{S^2}$.
- If $S = \sqrt{S^2}$, but variance, $s^2 = \frac{\Sigma(X - \overline{X})^2}{n-1}$
- Then sample standard deviation, $s = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}$

Standard Deviation and Variance

- Consider the data set 8, 5, 4, 12, 15, 5, 7. To calculate the variance, and standard deviation, S

X	X – mean	$(x - \text{mean})^2$
8	0	0
5	-3	9
4	-4	16
12	4	16
15	7	49
5	3	9
7	1	1
n = 7 Mean = 8 ∑x = 56		$S^2 = 100/6 = 16.67$ S = $\sqrt{16.67}$ = 4.8

Worked Example

- Diastolic blood pressures from some 8 hypertensive patients: 106, 98, 96, 110, 102, 108, 100, 105 mmHg.
- Calculate a) The Range b) IR c) Variance d) Standard deviation.

a) Range, $R = x_{max} - x_{min}$

$$R = 110 - 96 = 14$$

b) IR = $Q_3 - Q_1$,

but Rank of Q1 = $\frac{n}{4} = \frac{8}{4}$, Rank of Q3 = $\frac{3n}{4} = \frac{24}{4}$

Worked Example

b) $IR = Q_3 - Q_1$,

but Rank of Q1 $= \frac{n}{4} = \frac{8}{4} = 2^{nd}$,

Rank of Q3 $= \frac{3n}{4} = \frac{24}{4} = 6^{th}$

Rearranging: 96, 98, 100, 102, 105, 106, 108, 110

Therefore $IR = Q_3 - Q_1$

$IR = 106 - 98$

$= 8$

Worked Example

c) Variance, $s^2 = \frac{\Sigma(X - \overline{X})^2}{n - 1}$ but Mean, $\bar{x}$ = 103mmHg

- $\sum(x - \overline{x})^2 = (96 - 103)^2 + (98 - 103)^2 + (100 - 103)^2 +$

  $(102 - 103)^2 + 105 - 103)^2 + (106 - 103)^2 +$

  $(108 - 103)^2 + (110 - 103)^2$

  $= 49 + 25 + 9 + 1 + 4 + 9 + 25 + 49$

  $= 171$

Therefore $S^2 = \frac{171}{n - 1} = \frac{171}{8 - 1} = \frac{171}{7}$,

$= 24.4$

Worked Example

d) Sample standard deviation = √Variance

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

Therefore $S = \sqrt{24.4}$

$= 4.939$

Coefficient of Variation (CV)

- Definition:
  - The ratio of the standard deviation to the mean.
- Also known as relative standard deviation.
- It is a measure of dispersion of a frequency distribution and is expressed as a percentage.
  - The higher the CV, the greater the level of dispersion around the mean.
  - The lower the value of the CV, the more precise the estimate.

Coefficient of Variation

- The standard deviation is useful as a measure of variation within a given set of data.
- To compare dispersion in two variables a measure of relative variation rather than absolute variation is required.
- The coefficient of variation measures relative variation.
- It is standard deviation expressed as a percentage of the mean., i.e. C.V. = $\frac{S}{\bar{X}} \times 100$

Coefficient of Variance, CV.

- Suppose two samples of a group of males yield the following results:

	Sample 1	Sample 2
Age	25 years	11 years
Mean weight	145 Kg	80 Kg
Standard deviation	10 Kg	10 Kg

- Of interest is: which is more variable, the weights of the 25 years olds or the weights of the 11 year olds?

Coefficient of Variance, CV

- C.V. for 25 year olds:

  C.V. = 10/145 (100) = 6.9%

- C.V. for 11 year olds

  C.V. = 10/80 (100) = 12.5%

- When compared, variation is seen as much higher in the sample of 11 years olds than in the sample of 25 year olds.

Coefficient of Variation

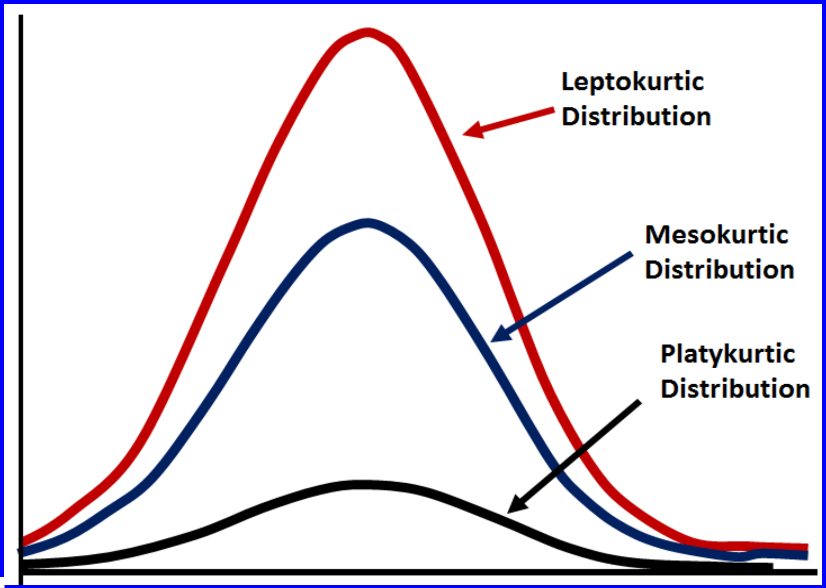
- The coefficient of variation is also useful in comparing the results obtained by different persons who are conducting investigations involving the same variable.
- Since the coefficient of variation is independent of the scale of measurement, it is a useful statistic for comparing the variability of two or more variables measured on different scales.

Kurtosis

- Kurtosis is the ‘humpedness’ or a measure of the ‘flat-toppedness’ of the distribution curve.
- It is a measure of the degree by which a distribution is “peaked” or flat in comparison with a normal distribution whose graph is bell-shaped.
- Knowledge of the shape of the distribution is crucial for statistical data analysis.

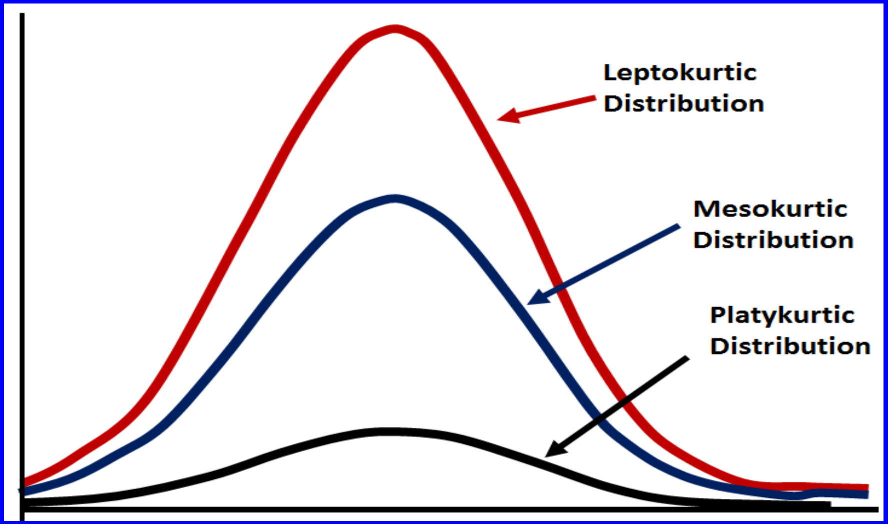
Platykurtic Distribution

- In comparison with a normal distribution, a platykurtic distribution may possess an excessive proportion of observations in its tails, so that its graph exhibits a flattened appearance, i.e. appears more flat than the normal distribution curve.
- Platykurtic distributions have more values in the distribution tails and fewer values close to the mean.



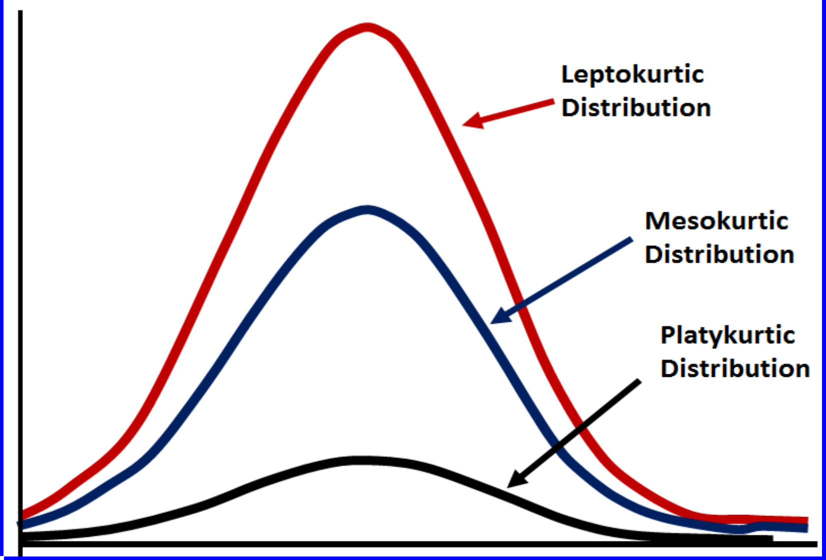
Mesokurtic Distribution

- A mesokurtic distribution is a statistical term used to describe the outlier characteristic of a probability distribution in which extreme events are close to zero.
- A mesokurtic distribution has a similar extreme value character as a normal distribution.
- Mesokurtic distributions are moderate in breadth and curves with a medium peaked height.



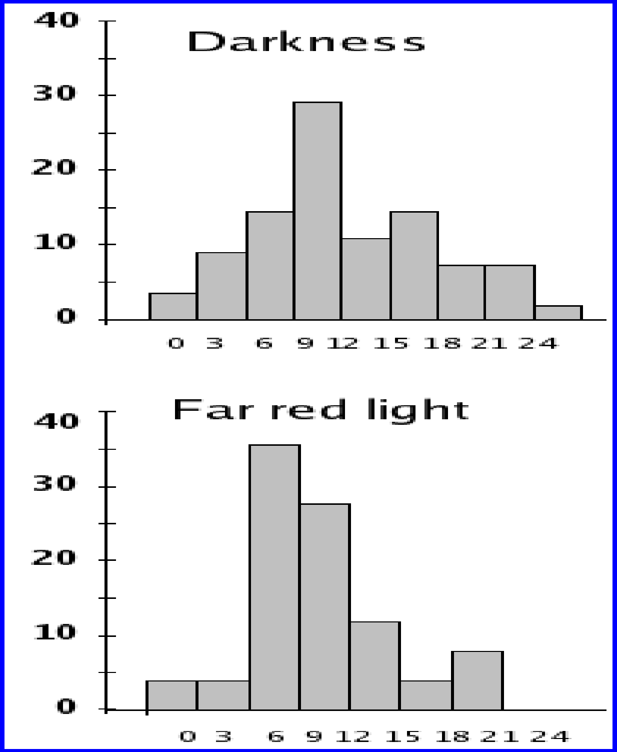
Leptokurtic Distribution

- In comparison with a normal distribution, a leptokurtic distribution may possess a smaller proportion of observations in its tails, so that its graph exhibits a more peaked appearance, i.e. has a sharper peak as compared with that of the normal distribution curve.
- Leptokurtic distributions have fewer values in the distribution tails and more values close to the mean.



Platykurtosis and Leptokurtosis

- The "Darkness" data is platykurtic, while "Far Red Light" shows leptokurtosis.

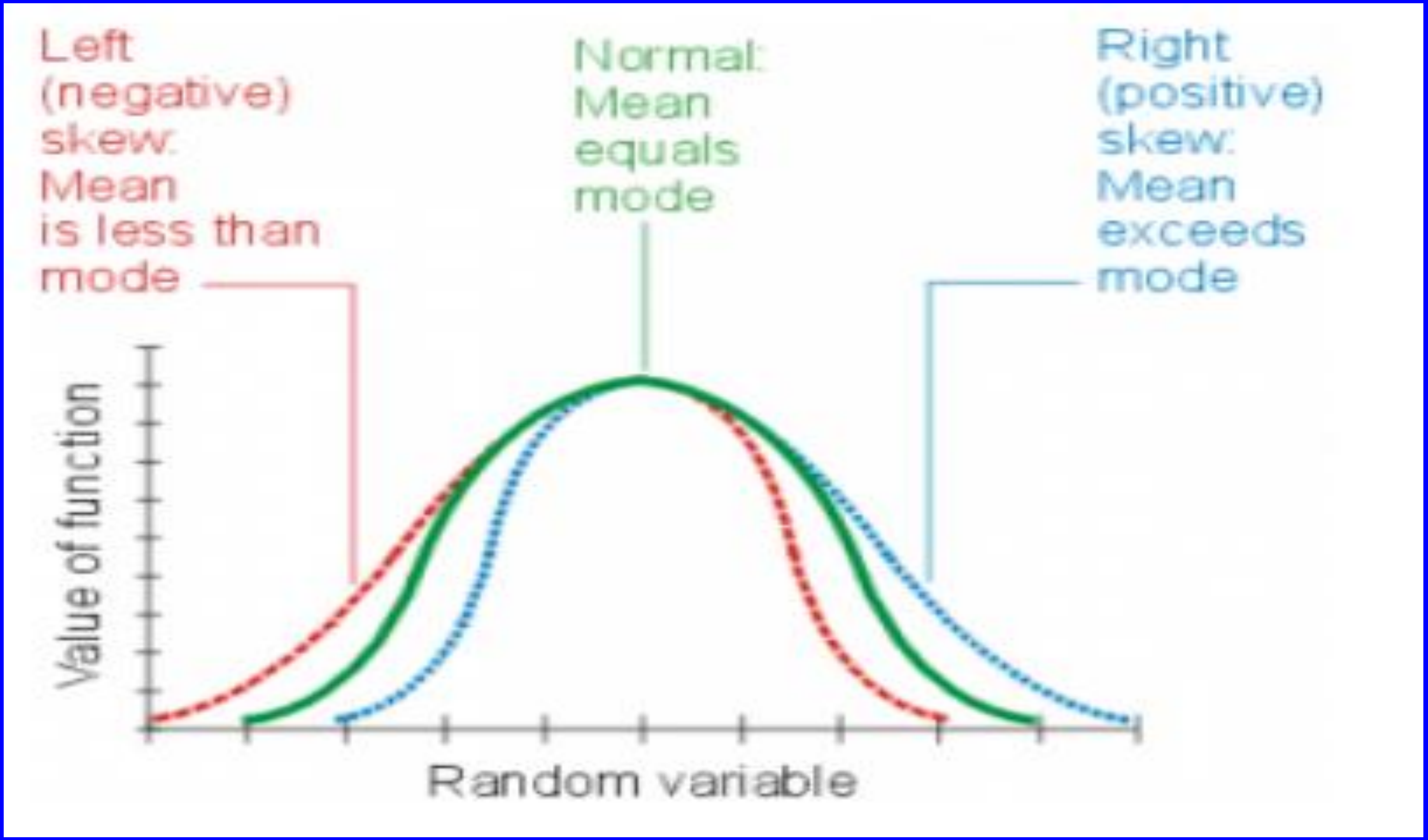


Skewness

- Definition
  - Skewness is the extent of asymmetry of the distribution of data in which the curve appears more drawn either to the left or to the right.
- Skewness can be quantified to define the extent to which a distribution differs from a normal distribution.

Skewness

- Definition



Skewness Calculation Formula

$$\tilde{\mu}_3 = \frac{\sum_i^N \left(X_i - \bar{X}\right)^3}{(N-1) * \sigma^3}$$

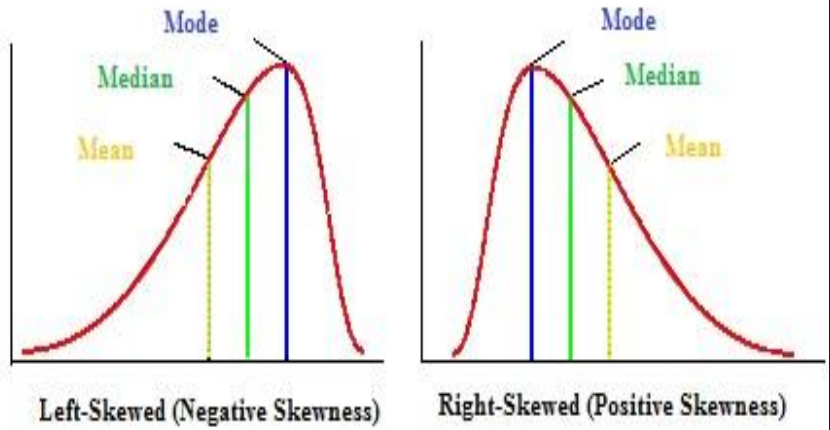
$\tilde{\mu}_3$ = skewness

N = number of variables in the distribution

X_i = random variable

$\bar{X}$ = mean of the distribution

σ = standard deviation



Assignment

- Length of illness (days) for 22 patients diagnosed with pneumonia: 6, 7, 8, 8, 10, 11, 11, 11, 8, 10, 10, 12, 12, 14, 14, 15, 15, 17, 18, 6, 5, 4.
- Calculate:

  a) The range, b) Interquartile range

  c) Semi-interquartile range, d) Variance,

  e) Coefficient of Variance

Summary

- Measures of dispersion provide information about the extent of variability among observations in a set of data and how spread each observation is relative to the mean.
- Measures of dispersion include range, quartiles, percentiles, variance, standard deviation and coefficient of variation.
- Observations in a data set may have a platykurtic, mesokurtic or leptokurtic distribution.

References

- Heumann, C., Schomaker, M. and Shalabh (2016) *Introduction to Statistics and Data Analysis With Exercises, Solutions and Applications in R*, Springer International Publishing Switzerland.
- Holmes, L, Jr. (2018) *Applied Biostatistical Principles and Concepts: Clinician's Guide to Data Analysis and Presentation*, Routledge, Taylor & Francis group, New York.