

ONLINE TRAINING WORK SHOP ON BIOSTATISTICS MEASURES OF DISPERSION & NORMAL DISTRIBUTION

Dr. P.Sasikala

Assistant professor

Dept of community medicine

Narayana medical college, Nellore.

INTRODUCTION

- The Measures of central tendency gives us a birds eye view of the entire data
- it serves to locate the centre of the distribution but they do not reveal how the items are spread out on either side of the central value.
- Dispersion is the scatteredness of the data series around the average.
- Dispersion is the extent to which values in a distribution differ from the average of the distribution.
- Greater the variation amongst different items of a series, the more will be the dispersion.

Other terms:

- measures of variability or
- scatter as against the average

OBJECTIVES OF MEASURING DISPERSION

- To determine the reliability of an average
- To compare the variability of two or more series
- For facilitating the use of other statistical measures
- Basis of Statistical Quality Control

Dispersion: example

- Number of minutes waited by 10 patients to see a consultant :

Doctor X		Doctor Y	
05	15	15	16
12	3	12	18
4	19	15	14
37	11	13	17
6	34	11	15
total=146		Total=146	

Mean waiting time for doctor X= $146/10=14.6$ min.

Mean waiting time for doctor Y= $146/10=14.6$ min.

What is the difference b/n two data series??

Doctor X: high variability

Doctor Y: less variability

How dispersions are measured for individual observations?

- Range
- Inter-quartile range
- Mean deviation / average deviation
- Standard deviation
- Coefficient of variation

Measures of variability of samples

- Standard error of mean
- Standard error of difference between two means
- Standard error of proportion
- Standard error of difference between two proportion
- Standard error of correlation coefficient
- Standard deviation of regression coefficient

Range

- ❑ Defines the normal limits of a biological characteristic.
- ❑ Observations falling within particular range are considered normal & those falling outside the normal range are considered abnormal.
- ❑ Simplest measure of dispersion
- ❑ Not a satisfactory measure as it is based on two extreme values.
- ❑ Normal range covers the observations falling in 95% confidence limits.
- ❑ Gives idea of variability quickly

- Calculated by finding the difference b/n the maximum & minimum measurements in the series.
- $R = L - S$
- $R = \text{Range}$, $L = \text{Largest Value}$, $S = \text{Smallest Value}$

□ What is the range of the following data?

□ blood serum cholesterol (mg/dl) levels of 10 persons are given below.

260, 240, 200, 240, 260, 150, 220, 190, 210, 240.

□ Largest = 260 smallest = 150

Range = $260 - 150 = 110$

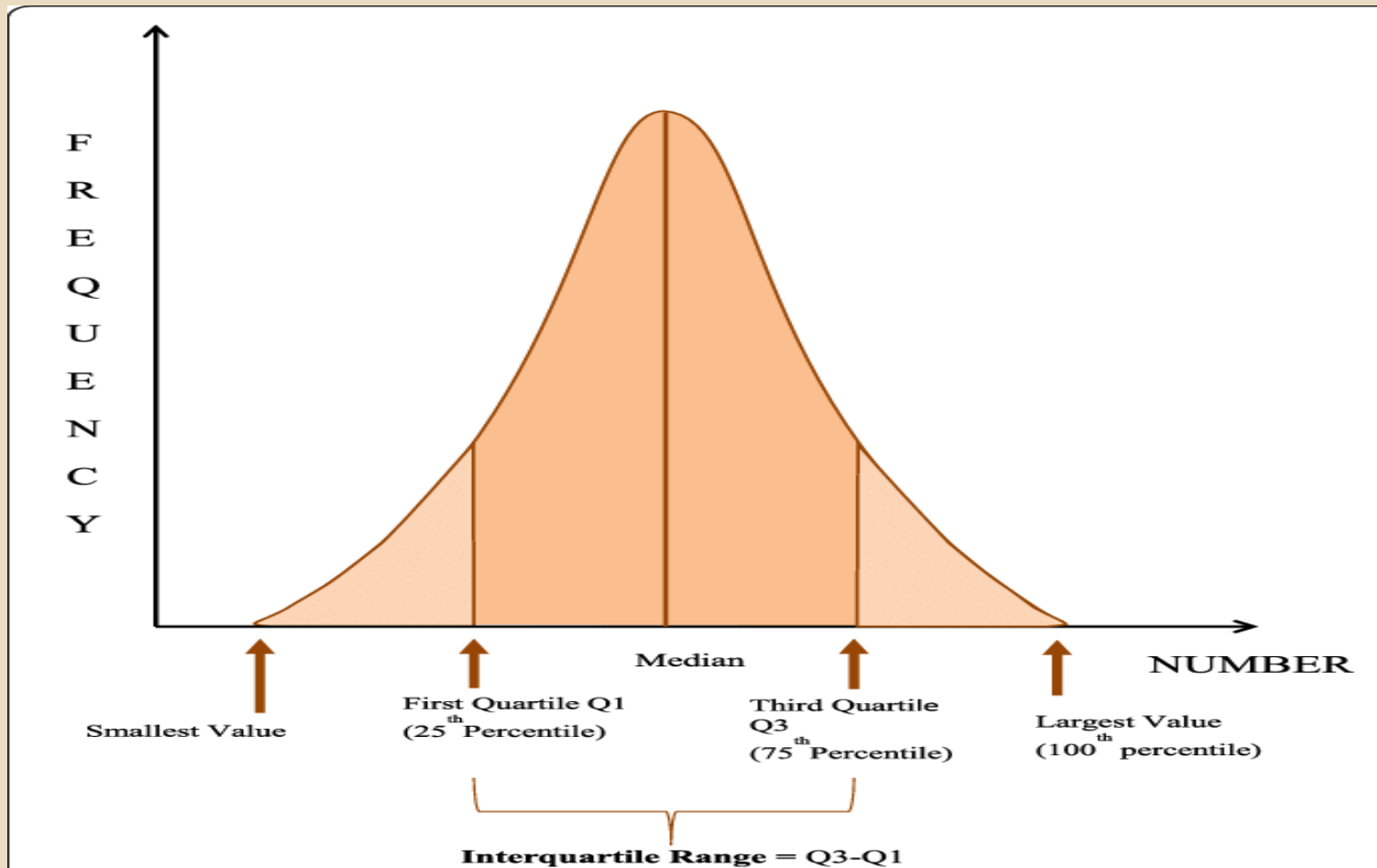
Inter-quartile range

- To calculate the inter-quartile range we must first find the quartiles.
- There are three quartiles, called Q1, Q2 & Q3 divides the data into 3 equal parts
- Q1 - lower quartile will have 25% observations falling on its left & 75% on its right.
- Q2/median: 50% observations on either side
- Q3- upper quartile, will have 75% of observations to left & 25% on right.

Inter-quartile range

- Interquartile Range is the difference between the upper quartile (Q3) and the lower quartile (Q1)
- Better than range as it involves middle half of the results.
- Inter-quartile range = $Q3 - Q1$

Quartiles :




Semi-Interquartile Range/quartile deviation

- The semi-inter quartile range (or SIR) is average of inter quartile range.
- $SIR(Q) = (Q3 - Q1) / 2$
- Slightly better than range.
- Takes into account only the middle half of the data b/n Q1 & Q3.
- If Q is greater → greater scatter in IQR
- If Q is smaller → less scatter, more in middle.

MEAN DEVIATION/AVERAGE DEVIATION

- Mean deviation is the arithmetic mean of the absolute differences of the values from their average .
- To find out MD of observations:
- calculate the mean (\bar{x})
- calculate the difference b/n observation (x) & mean ($x - \bar{x}$)
- add the differences by ignoring the sign, $\sum |x - \bar{x}|$
- divide by number of observations.


$$MD = \frac{\sum |X - \bar{X}|}{N}$$

Simple to understand & easy to calculate, not used in statistical conclusions.

- A surgeon of a medical college performed the following number of operations during the 6 days of a week . 10, 8, 11, 8, 9, 8. calculate mean deviation

Operations	mean	Observation-mean($X - \bar{X}$)
8	9	-1
8	9	-1
8	9	-1
9	9	0
10	9	1
11	9	2
Total = 54. mean = $54/6=9$		6

$$\frac{\sum |X_i - \bar{X}|}{n}$$

$$= 6/6$$

$$= 1$$

Standard Deviation(SD)

- Most important & widely used measure of dispersion
- First used by Karl Pearson in 1893
- Also called “Root- Means Square Deviations”
- It is defined as the square root of the arithmetic mean of the squares of the deviation of the values taken from the mean
- Denoted by σ (sigma) & initials S.D

- For small sample ($n < 30$)

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

- Large sample
 $n > 30$

$$S = \sqrt{\frac{\sum (X - \bar{X})^2}{N}}$$

*where S = the standard deviation of a sample,
Σ means "sum of,"
X = each value in the data set,
X̄ = mean of all values in the data set,
N = number of values in the data set.*

□ Computed by

(i) Take the deviation of each value from the

arithmetic mean, i.e., $(x - \bar{x})$

(ii) Square the each deviation, i.e., $(x - \bar{x})^2$

(iii) Sum the squared deviations. $\sum (x - \bar{x})^2$

(iv) Divide the above result by the number of observations (n) (or (n-1) for small sample).

(v) calculate the square root for the above result, which gives the standard deviation.

Operations	\bar{x}	$X - \bar{x}$	$(x - \bar{x})^2$
8	9	-1	1
8	9	-1	1
8	9	-1	1
9	9	0	0
10	9	1	1
11	9	2	4
Total		6	8

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

$$= \sqrt{8/5}$$

$$= \sqrt{1.6}$$

$$= 1.26$$

Uses of standard deviation:

- Summarizes the deviations of a large distribution from mean in one figure used as unit of variation.
- Indicates whether the variation of difference of an individual from the mean is by chance
- Helps in finding the standard error
- Helps in finding sample size to draw valid conclusions.

Coefficient of variation

- Method used to compare relative variability i.e., variation of same character in two or more different series of data.
 - E.g. variation of pulse rate in boys & girls
- or
- To compare variability of two characters in one individual
 - E.g., pulse rate & BP, height & weight

□ Coefficient of variation (CV)=

$$\frac{\text{SD}}{\text{mean}} \times 100$$

$$\text{CV} = 1.26/9 \times 100 = 14\%$$

- In two series of adults aged 21 years & children 3 months old following values were obtained for height. Find which series shows greater variation?

Persons	Mean height	SD	CV
Adults	160 cm	10cm	$(10/160)*100=6.25\%$
Children	60 cm	5 cm	$(5/60)*100=8.33\%$

- Thus, heights of children shows greater variation.

Normal distribution


- The normal distribution is a descriptive model that describes real world situations.
- It is defined as a continuous frequency distribution of infinite range
- This is the most important probability distribution in statistics and important tool in analysis of epidemiological data and management science.

Importance

- Many dependent variables are commonly assumed to be normally distributed in the population
- If a variable is approximately normally distributed we can make inferences about values of that variable

Normal distribution

- When large number of observations of any variable characteristic such as height, blood pressure are taken at random to make it a representative sample, & frequency distribution table is prepared by keeping group interval small, then it is seen that:
- Some observations are above the mean & some are below the mean.
- If they are arranged in order, deviating towards the extremes from the mean, on plus or minus side, maximum numbers will be seen in middle around mean & fewer at extremities, decreasing smoothly on both sides.

- 
- Normally all observations are symmetrically distributed on each side of mean.
 - A distribution of this nature or shape is called normal distribution or Gaussian distribution.

Characteristics of normal curve

- bell shaped curve
- It is symmetrical
- Mean, median, mode coincide
- Area under the curve is 1
- Normal distributions are denser in the center and less dense in the tails.
- 68.27% of the area of a normal distribution is within mean ± 1 SD limits
- Approximately 95% of the area of a normal distribution is with in mean ± 2 SD
- 99.7% of the area of normal distribution is with in mean ± 3 SD
- These limits on either side of mean are called confidence limits

- The normal distribution is asymptotic - the curve gets closer and closer to the x-axis but never actually touches it.
- The points at which the curvature changes are called inflection points.
- Normal distributions are defined by two parameters, the mean (μ) and the standard deviation (σ).

- Observations larger or smaller than mean ± 1 SD are fairly common, nearly one-third.
- Values that differ from mean by more than twice the SD are rare, being only 4.55%.
- Values higher or lower than 3 SD are very rare, being only 0.27%. The chance of being normal is 0.27 in 100. such values are abnormal & may be even pathological also.

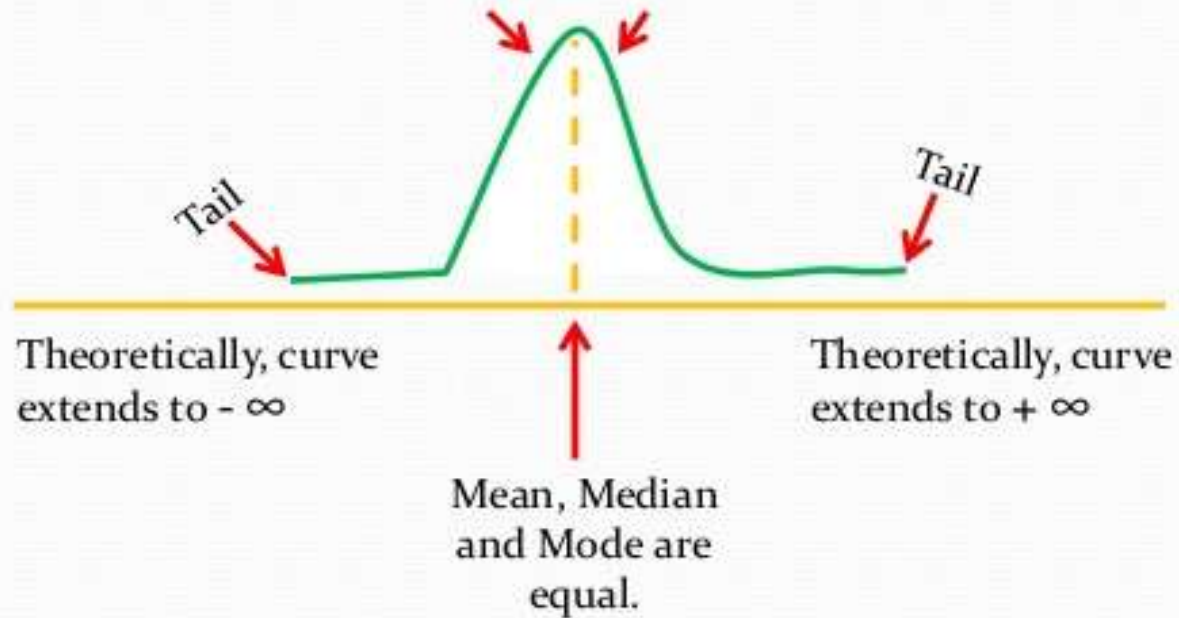
We say the data is "normally distributed":



The **Normal Distribution** has:

- mean = median = mode
- symmetry about the center
- 50% of values less than the mean and 50% greater than the mean

**Normal Curve is Symmetrical
Two halves identical**



Normal distribution of height of 1000 students.

<i>Height in cm</i>	<i>Frequency of each group</i>	<i>Frequency within height limits of</i>		
142.5	3			
145.0	8			
147.5	15			
150.0	45			
152.5	90			
155.0	155	Mean	Mean	Mean
157.5	194	± 1 SD	± 2 SD	± 3 SD
160.0M	195	680	950	995
162.5	136	68%	95%	99%
165.0	93			
167.5	42			
170.0	16			
172.5	6			
175.0 - 177.5	2			
Mean = 160 cm SD = 5 cm				

Relative or standard normal deviate or variate (Z)

- The Deviation of value (x) from the mean in a normal distribution curve is called Relative or standard normal deviate
- Denoted by “Z”
- Measured in terms of SD & indicates how much an observation is bigger or smaller than mean in unit of SD.

STANDARDIZED SCORE (Z-value)

Formula:

$$Z = \frac{x - \mu}{\delta}$$

z = Normal Value

X = value of any particular observation

μ = mean of the distribution

δ = standard deviation

Skewness & kurtosis

- Symmetric distribution of data means that the right and the left of the distribution are perfect mirror images of one another.
- Not every distribution of data is symmetric, some distributions are asymmetrical & skewed.

Skewness & kurtosis

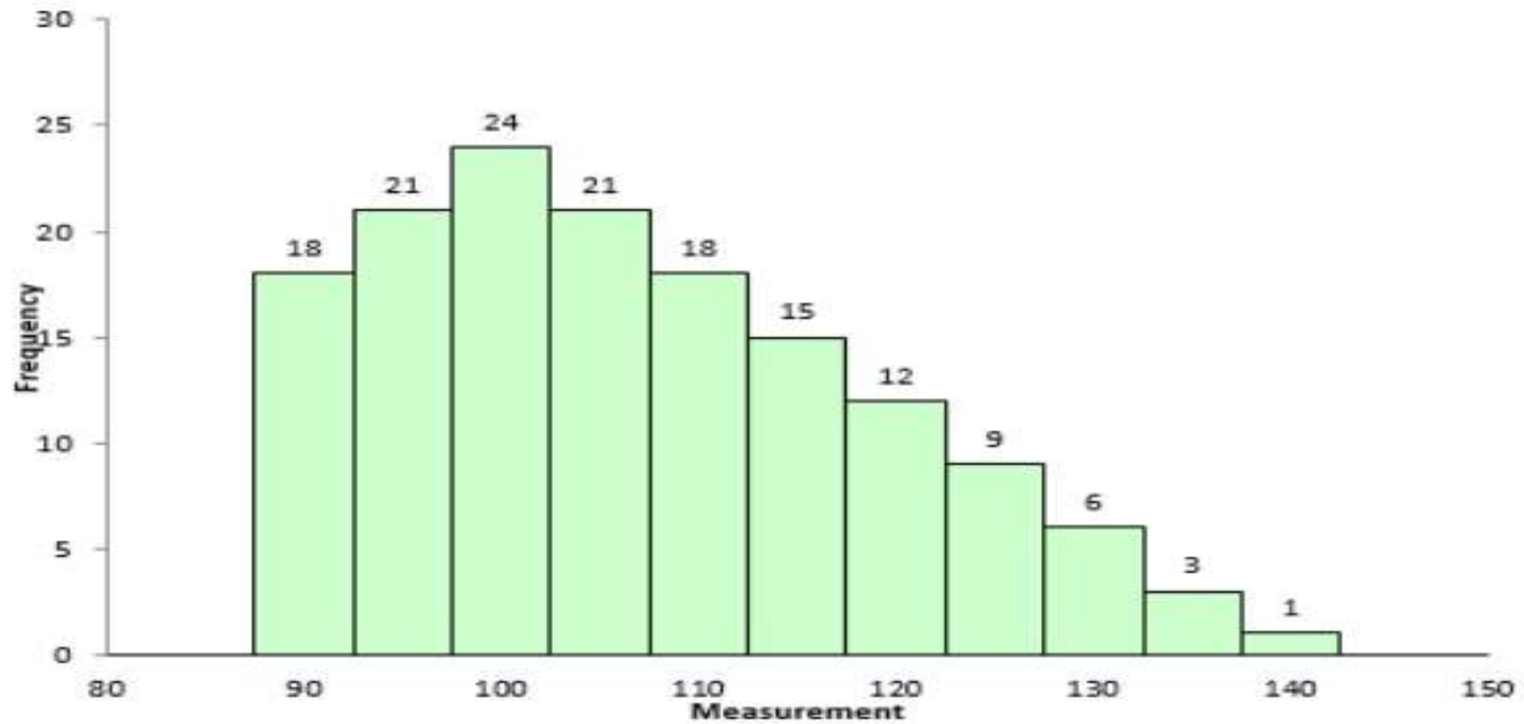
□ **SKEWNESS :**

- Skewness is usually described as a measure of a dataset's symmetry – or lack of symmetry.
- A perfectly symmetrical data set will have a skewness of 0.
- The normal distribution has a skewness of 0.

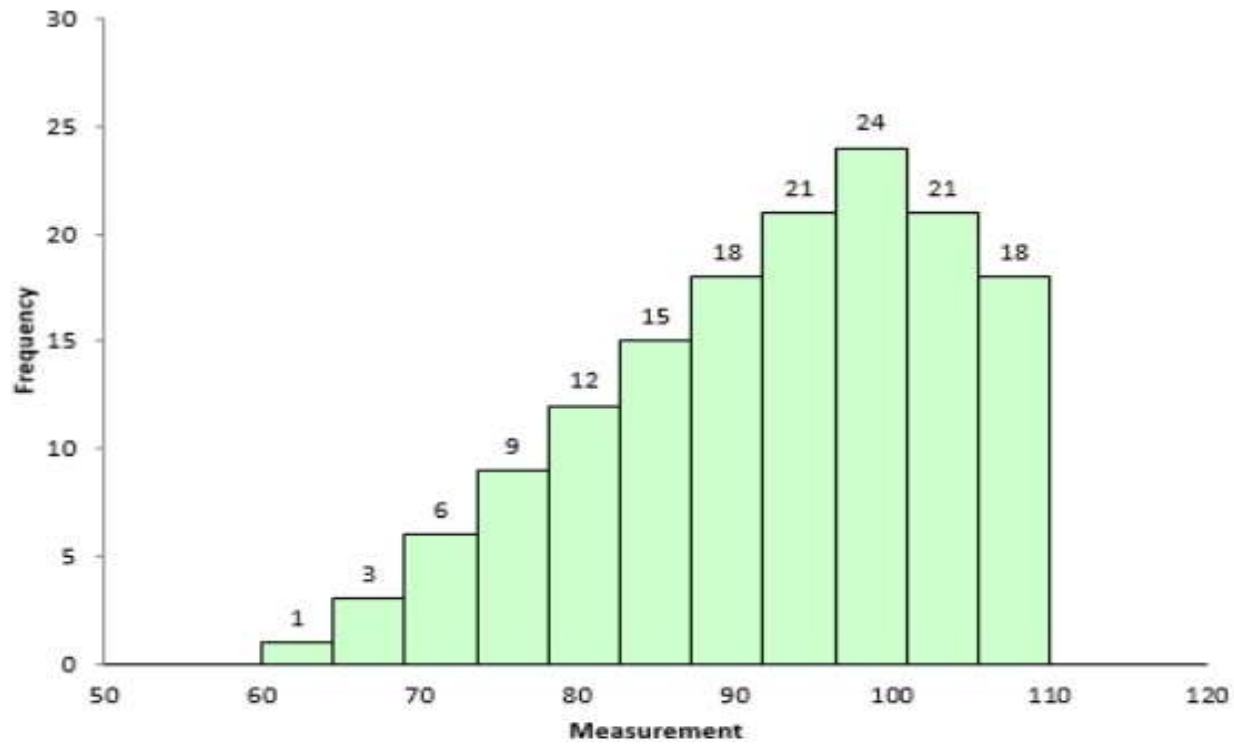
□ **KURTOSIS :**

- Kurtosis is the degree of peakedness of a distribution
- Is a measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution.

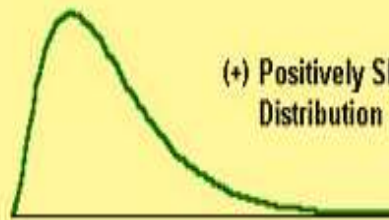
Skewed to right



Skewed to left

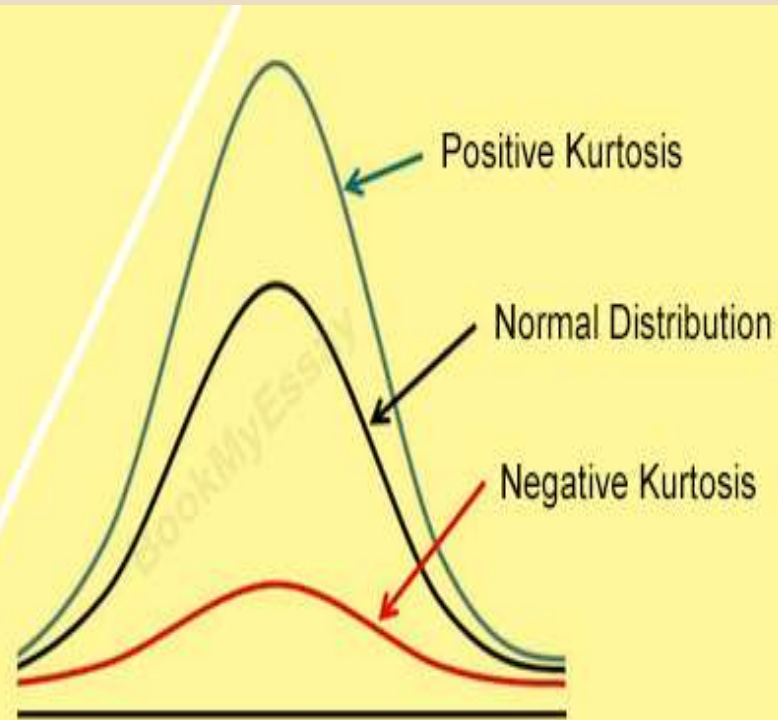


Skewness



(+) Positively Skewed Distribution

(-) Negatively Skewed Distribution



Kurtosis



Exercises by using MS-Excel

Exercise 1

Depression scores of 15 persons. calculate measures
of dispersion

51, 55, 52, 45, 49, 57, 51, 48, 49, 50, 48, 43, 52,
53, 47

Exercise- 2

Mid arm circumference (cm) of 10 male children aged 4 months is given below:

14, 11, 11, 10, 12, 13, 10, 14, 11, 11,

Calculate Q1, Q2, Q3

Exercise-3

- calculate measure of dispersion for the following data.
- 1) The diastolic blood pressure values (mm Hg) of 10 male adults are given below.
80, 60, 70, 80, 65, 74, 66, 80, 70, 55.

Exercise-4

- 1) blood serum cholesterol levels (mg/dl) of 10 subjects are given below:

260, 200, 240, 240, 260, 150, 220, 190, 210, 240.

- Calculate range, SD.

Exercise-5

- The weights (in Kgs) for aged two years of children:
- 9, 11, 10, 12, 10, 8, 10, 11, 10, 9, 12, 11, 8, 9, 12, 11, 10, 10, 12, 10
- Calculate range, SD, IQR

Exercise-6: construction of normal curve

1. Calculate mean, SD
2. Do Z distribution , function is
`=norm.dist(data point, $mean,$SD, false)`
3. Select data & Z distribution
4. Insert → scatter

- In a series of boys the mean SBP was 120 mm Hg & SD was 10 mm of Hg. In the same series mean height & SD were 160 & 5 cm respectively. Find which character shows greater variation.

- CV of BP = $10/120 * 100 = 8.3\%$
 - CV of height = $5/160 * 100 = 3.1\%$
- BP is more variable than height.

Exercise 7: Construct normal curve.

data

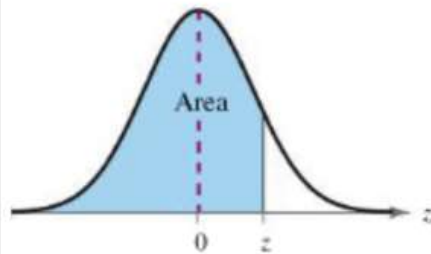
- 1, 3, 3, 5, 6, 9, 10, 10, 13, 19, 19, 20, 21, 22, 23, 25, 25, 27, 31, 32, 36, 37, 37, 40, 44, 45, 46, 48, 49, 49, 54, 55, 57, 57, 57, 61, 62, 62, 63, 63, 64, 66, 74, 75, 76, 77, 78, 88, 89, 90, 91, 91, 91, 91, 95, 95, 95, 98.

Z-calculation

- Average weight of baby at birth is 3.05 kg with SD of 0.39kg. If the birth weight are normally distributed would you regard:
 - a. Wt of 4 kg abnormal
 - b. Wt of 2.5 as normal
- Normal limits of weight: mean $\pm 1.96SD$
 - = $3.05 \pm 1.96 \times 0.39$ (0.7644)
 - = $3.04 + 0.7644 - 3.04 - 0.7644$
 - = $3.81 - 2.29$
- Wt of 4kg fall out side the normal \rightarrow abnormal
- Wt of 2.5kg lies within normal limits \rightarrow normal

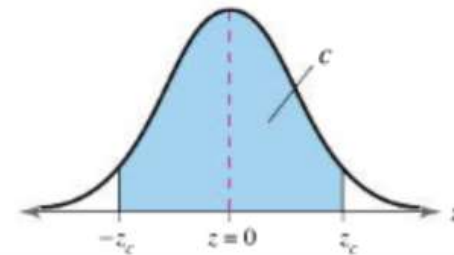


Standard Normal Distribution



Critical Values

Level of Confidence c	z_c
0.80	1.28
0.90	1.645
0.95	1.96
0.99	2.575



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389

□ Menstrual cycles (MC) in woman following normal distribution has a mean of 28 days & SD of 2 days. How frequently would you expect menstrual cycle of

a. More than 30 days

b. Less than 22 days

- $Z = \text{observation} - \text{mean} / \text{SD}$

$$= \frac{(x - \mu)}{\sigma}$$

- $Z = 22 - 28 / 2 = -3$ ($Z = -3$, Z distribution value = 0.0013) i.e., 0.13% will have MC of <22 days

- $Z = 30 - 28 / 2 = 1$ ($Z = 1$, Z distribution value = 0.1587) i.e., 15.85% will have MC of >30 days



Thank you....