NEET SS OBG

BIOSTATISTICS

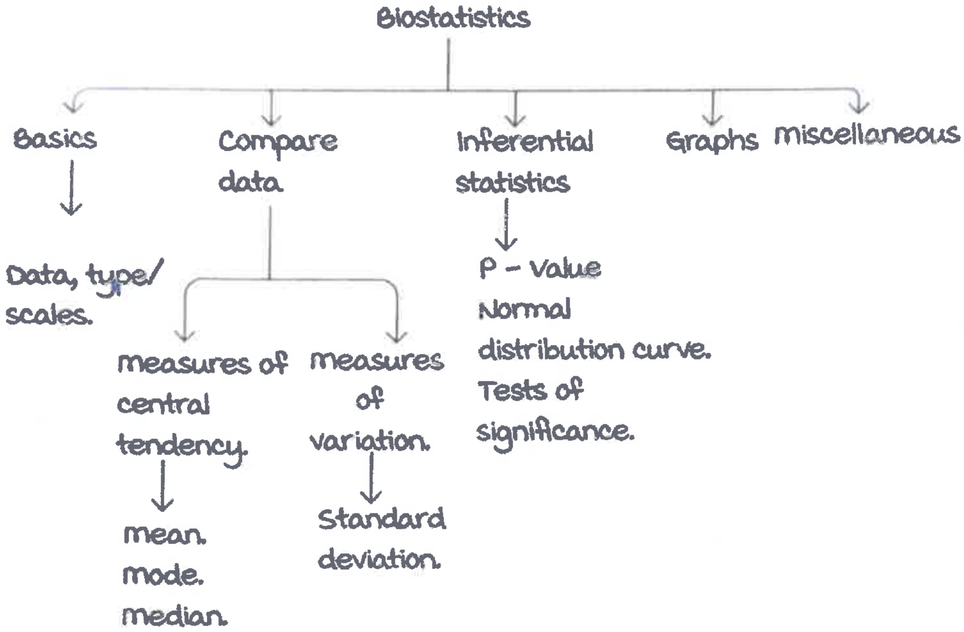
4

CONTENT

INTRODUCTION TO DATA IN BIOSTATISTICS

Uses :

- Define cut-offs.
- understand variation.
- To present data.
- To make inference (provide evidence).



Data

00:07:54

Quantitative	Qualitative
• Continuous. • measurable. • E.g. weight, height, AST, ALT levels. • mean of data can be calculated.	• Discrete. • Countable. • E.g. No. of people who are sick/ healthy, alive/dead. Gender. • Proportions/ percentages can be calculated.

Pulse rate is a data which is discrete and countable, however it is quantitative as we calculate its mean.

BP is quantitative data.

Scales of data

00:14:25

Nominal	Ordinal	Interval	Ratio
• Named data. • No sequence • E.g. Gender, religion, blood groups.	• Inherent order. • Has a sequence. • E.g. Stage, grade, the severity of the disease.	• Interval between two values is present. • No start point/no absolute zero. • E.g. °C, dB.	• Ratio can be calculated. • There is zero point/ absolute zero. • E.g. Na, K, FEV levels.

Interval type of data :

Example : 20 °C is not half as hot as 40 °C, but colder compared to 40 °C. Here the intensity of data is measured. Also, the temperature can go below 0 °C (in minus °C), which means there is no absolute zero.

Ratios :

Example : A weak fragile child weighs 20 kg when the ideal weight should have been 40 kg in the same age group. The ideal weight is 2 x child's age, which means the values can be expressed in multiples (double, triple) of each other
i.e calculation of ratios is possible.
Also, there is absolute zero/ no value below zero.

MEASURES OF CENTRAL TENDENCY AND VARIATION

Measures of Central tendency 00:01:59

mean :

1. Arithmetic mean :
   - Average = $\frac{\Sigma(\text{summation})}{n}$
2. Geometric mean :
   - Calculated in case of : Exponential data.
     Extreme values.
   - Example : Human development index
     (India = 0.647, ranked at 129 in 2019)
3. Harmonic mean :
   - Calculated in case of : Inverse data.
     Fractional values.

Advantages :
- Best measure of central tendency.
- Easiest to calculate.

Disadvantages :
- most affected by extreme values.

median :

Central value after arranging in ascending or descending order.

Advantages :
- Least affected by extreme values.

mode :

The most frequently occurring value.

mode = 3 median − 2 mean.

Advantages :
- The most robust measure of central tendency.
- The last to be affected by extreme values.

Data with extreme values : Preferred measure is median.
Preferred mean is geometric mean.

Measures of variation

1. Range :

   Range = maximum to minimum.

2. Standard deviation :

Gives the mean deviation of every value from the mean.

Formula : The root of the mean of squared deviation.

$$SD = \sqrt{\frac{\Sigma (x - \bar{x})^2}{n}}$$

In case of a small sample,

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

$n - 1$ is the correction for the small sample ($n < 30$).

3. Variance :

   Variance (V) = SD^2

$$V = \frac{\Sigma (x - \bar{x})^2}{n}$$

4. Coefficient of variation (CV) :

   Absolute variation between a different populations.

$$CV = \frac{S.D}{mean} \times 100$$

5. Standard error :

Gives the error in different studies in terms of standard deviation.

Alternatively, gives the variation between values when different researches are done.

a. Standard error for mean :

- For quantitative data.
- $SE_m = \frac{SD}{\sqrt{n}}$

b. Standard error for proportions :

- For qualitative data.
- $SE_p = \sqrt{\frac{PQ}{N}}$

P : Prevalence.

Q : 100 – prevalence.

n : Sample size.

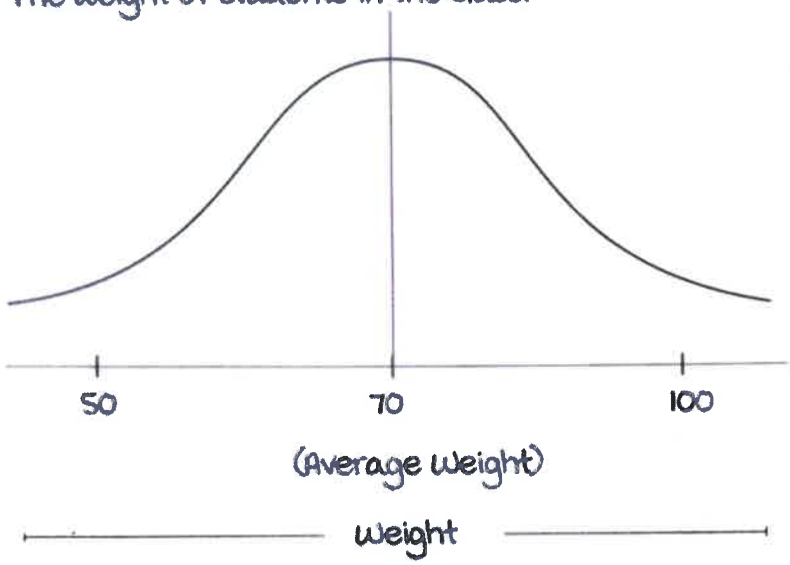
If p- value or Confidence interval is provided as input, Standard error has to be calculated and not the Standard deviation.

NORMAL DISTRIBUTION CURVE

Normal distribution curve 00:00:08

It represents the distribution of data in a bell-shaped curve, in a large sample.

Eg : The weight of students in the class.



Features of Normal distribution curve :

It is also known as the Gaussian distribution curve.

It is a bilaterally symmetrical bell-shaped curve.

The ends never touch the baseline.

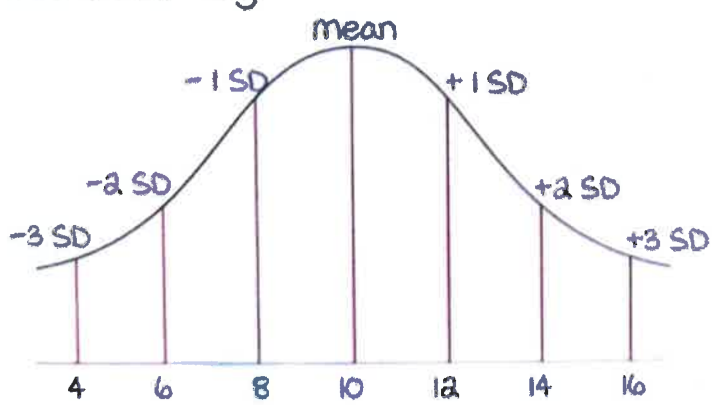
mean = median = mode → Coincide at 0 or the centrepoint.

SD = 1.

AUC = 1 (Area Under Curve), means the whole population is accounted for.

Eg : mean Hb ($\bar{x}$ Hb) at a place = 10 gm% ± 2 g%.

where 1 SD = 2 g%

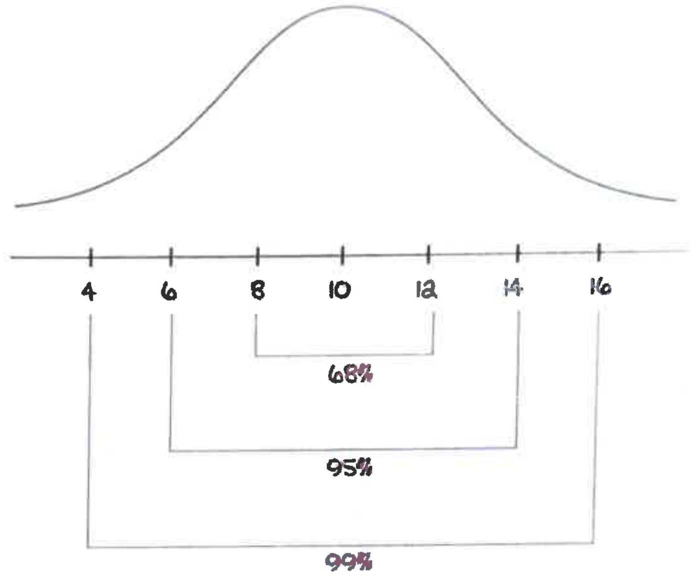


Assumptions in normal distribution curve :

First assumption

00:07:28

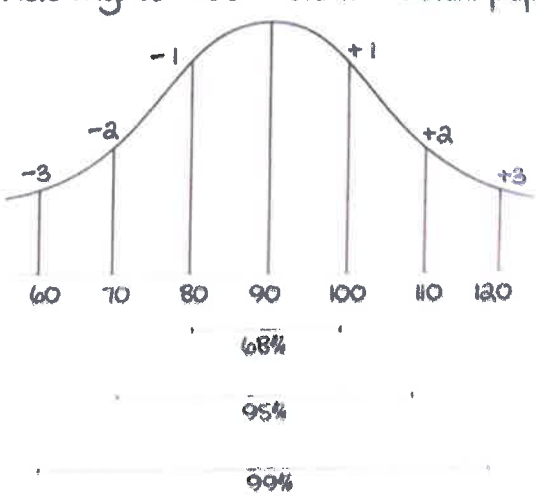
Between the – 1 SD and + 1 SD : 68% of the population lies.
Between the – 2 SD and + 2 SD : 95% of the population lies.
Between the – 3 SD and + 3 SD : 99% of the population lies.



Eg : mean blood glucose = 90 ± 10 SD.

How much of the population will be expected to fall between :

- 80 to 100 mg/dl = 68% population.
- 70 to 110 mg/dl = 95% population.
- 70 to 100 mg/dl = 68% + 13.5 % population [(95–68)/2=13.5]
- more than 70 mg/dl = 100 – 2.5 % = 97.5% population.
- Less than 100 mg/dl = 84% population (100–13.5+2+0.5).
- more than 100 mg/dl = 16 % population.
- Less than 60 mg/dl = 100 – 99 = 1/2 = 0.5% population.
- Less than 120 mg/dl = 100 – 0.5 % = 99.5% population.



Q. The mean blood glucose from 5929 ANC females in the state of maharashtra was found to be 130 ± 5 mg/dl. The cut off for diagnosing GDM was kept as higher than 140 mg/dl. How many pregnant females are expected to be GDM diagnosed?

A. < 50.
B. 50 to 100.
C. 100 to 200.
D. 200 to 500.

mean = 130, + 1 SD = 135, + 2 SD = 140, +3 SD = 145
− 1 SD = 125, − 2 SD = 120, −3 SD = 115.

To be GDM diagnosed, they must belong to above + 2 SD of population.
Above +2 SD = 100 − 95% (between +2 and −2 SD) − 2.5% (less than −2 SD) = 2.5%
2.5% of 5929 ~ 150 females, which falls under range of 100−200.

Second assumption : Zone of Normalcy 00:20:51

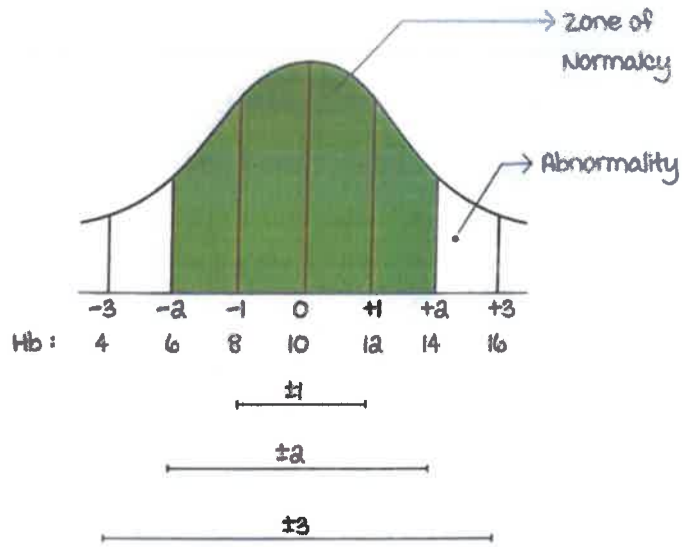
Zone of normalcy/normal zone :
Between the − 2 SD and + 2 SD = 95% of population.

Z score :
- It is also called standard deviate.
- It gives the location of the value in terms of the standard deviation (SD).
- The cut off for Z score : ± 2 SD/± 1.96 SD.
- If the Z score is > 2 : Abnormal Z score.

It is calculated by = $\frac{\text{Observed value} - \text{Expected value}}{\text{SD}}$

Eg : Observed value of Hb = 15 gm /dl.
Expected value (always the mean value) = 10
SD = 2
Z score = $\frac{15 - 10}{2} = 2.5$

Z score 2.5 : It lies 2.5 SD away from the mean.


Zone of Normalcy
Abnormality
-3 -2 -1 0 +1 +2 +3
Hb : 4 6 8 10 12 14 16
±1
±2
±3

TESTS OF SIGNIFICANCE

Statistical mathematical formula to derive a p-value.
Determines if P-value is significant or non significant.

Types of tests of significance

00:02:59

Types

1. Parametric
   Quantitative
   Normal distribution data

2. Non-parametric
   Qualitative
   Non-normal distribution data

Parametric test	Situation	Non-parametric test
Paired 't' test.	Single group	mc nermar's test.
Unpaired 't' test A/K/A Independent sample 't' test.	Two groups	Chi square test (χ^2).
Analysis of variance (ANOVA)	Three or more groups	Kruskal-wallis test. Chi square for trend.

Advance tests of significance

00:08:59

- Large sample (n > 30) = 'Z' test.
- Ordinal data : wilcoxan rank test (W/R)

W/R sign test
For grouped data

W/R sum test
For ungrouped data

- Normalcy of data : Kolmogorov smirnov test.
- Outliers : Dixon's Q test.
- Internal consistency of questionnaire : Cronbach's α score
- Compare a new test with a gold standard test : Bland altman analysis.

- Level of agreement : KAPPA test.

$$\text{Formula} = \frac{\text{Observed level of agreement} - \text{expected level of agreement}}{1 - \text{expected level of agreement}}$$

CONCEPT OF PROBABILITY VALUE

P value

00:01:05

P value :

Probability value (chance of events expressed in decimals).

Normal value ranges from 0 to 1.

0 : Lowest probability.

1 : maximum probability.

Standard errors (SE) :

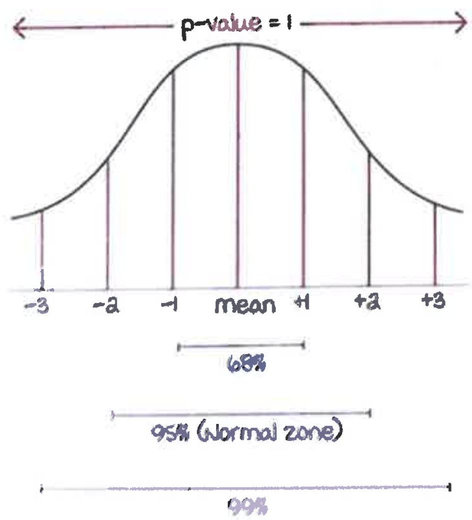
±1, ±2, ±3,...

Confidence limit/interval :

+1 to − 1 = 68% confidence interval.

+2 to − 2 = 95% confidence interval.

+ 3 to − 3 = 99% confidence interval.



In the normal distribution curve :

The highest probability is towards the centre : 1.

The lowest probability lies on either side of the curve.

At +2 to − 2 standard deviation the P value is : 0.05 − zone of normalcy.

P value – abnormal zone

00:06:20

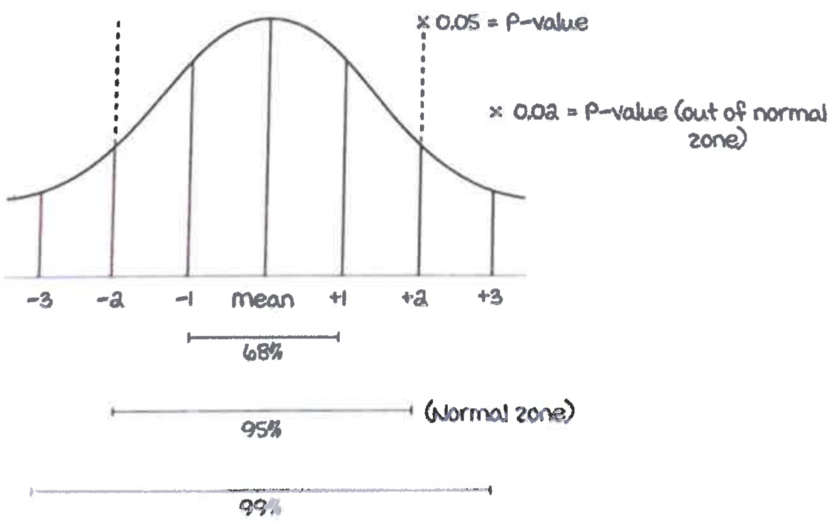
Example : Randomised clinical trial – two groups A and B

Group A	Group B
↓	↓
Drug is given (Remdesivir)	Placebo given (vitamin C/zinc)
↓	↓
check A'	Check B'

The collected data is incorporated in a machine : Gives P value.

If the P value is 0.02 : Abnormal/out of the normal zone.

P value > 0.05	P value < 0.05
Normal variant	Abnormal variant
Non-significant	significant
No effect found	Effect is found
Null hypothesis : Accepted	Null hypothesis : Rejected



P value – normal zone and changes

00:16:42

The normal zone for P value – 95% confidence interval.

If the normal zone moved from 95% to 68% :

Previously non-significant becomes significant.
Chances of finding an effect increases.
The chances of reject of null hypothesis increases.
The chances of alpha error increases.

If the normal zone moves from 95% to 99% :

Previously significant becomes non-significant.
The chances of finding an effect decreases.
The chances of accepting of null hypothesis increases.
The chances of beta errors increase.

Alpha error, type I & II error 00:23:08

Definition :

It is the probability of finding an effect (just by chance) which in reality does not exist.
It corresponds to the P value/confidence interval/limit.

Example : P value of 0.02 corresponds to α value 2%.
It means there is 2% chance of error in the study.
It also means there is 98% of confidence in the study.

68% corresponds to 32% alpha.
95% corresponds to 5% alpha.
99% corresponds to 1% alpha.

FPER : The chance of finding disease in a healthy patient.

Type I error :

Rejecting a null hypothesis, which in reality is true.

Type II error :

Accepting a null hypothesis, which is false in reality.

CORRELATION, REGRESSION AND SKEW

Correlation 00:00:13

Relation between 2 variables.

Scatter plots are used.

Types:

Linear

- Also known as Pearson-Karl correlation.
- Represented by : r
- Range : −1 to +1

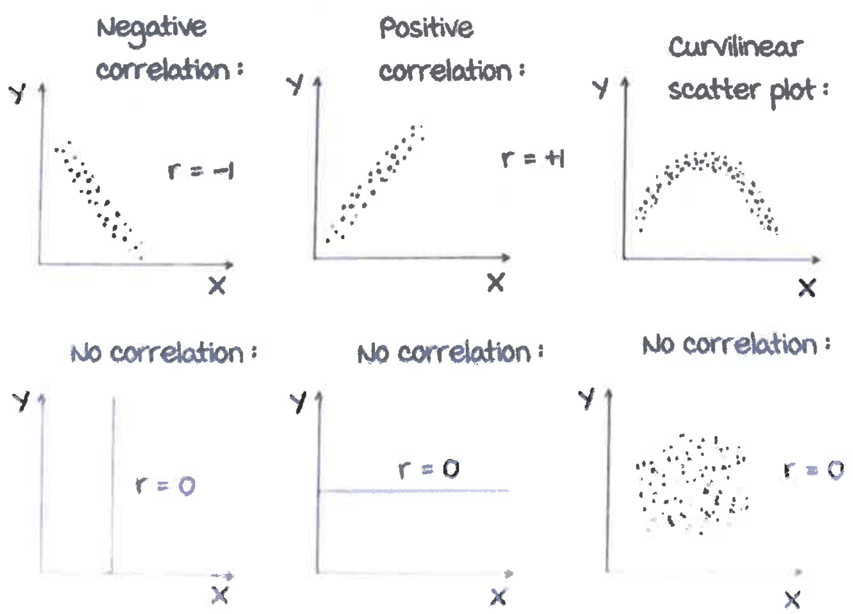
  −1 : Perfect negative correlation.

  +1 : Perfect positive correlation.

  r = 0 : No correlation.

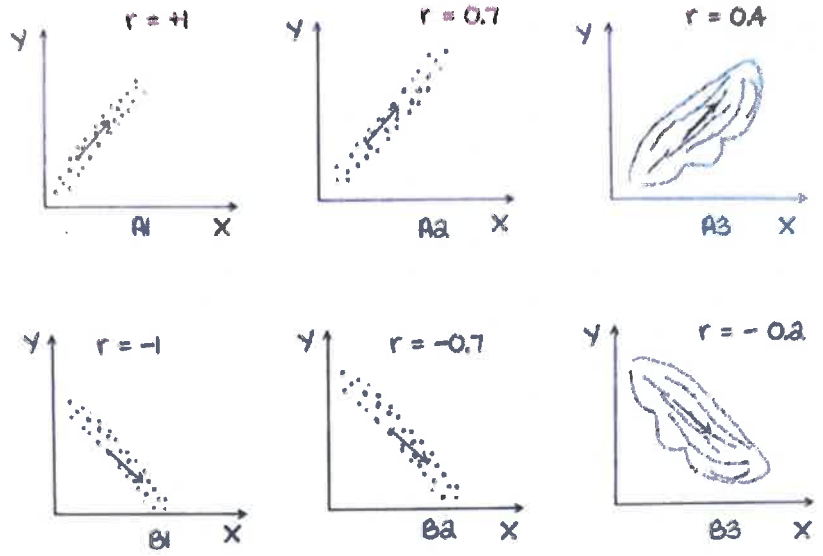
Curvilinear

- Also known as non-linear/ Spearman correlation.
- Represented by : ρ



Scatter plots

Coefficient of determination 00:10:04



+1 : Perfect positive correlation (1 unit change in X axis = 1 unit change in Y axis).

> 0.7 : Strong positive correlation.

0.5 – 0.7 : moderately positive correlation.

< 0.5 : Weak correlation.

< 0.3 : Very weak correlation.

Coefficient of determination (CD) :

The percentage change in one variable which is accounted for by a unit change in another variable.

CD = r^2 in %.

Regression 00:18:26

Primarily refers to prediction.

Types :

1. Linear : If variables as quantitiative.
2. Logistic : If variables are qualitative.
    1. Univariate linear regression :
       Eg : Predicting renal failure based on GFR.
    2. Univariate logistic regression :
       Eg : Predicting MI based on obesity levels.
    3. Multivariate linear regression :
       Eg : Predicting the renal status based on serum Na, urea, creatinine and GFR levels.