Basic Biostatistics

Dr. Kiran Chaudhary

Dr. Mina Chandra

Overview

1. **Importance of Biostatistics**
2. **Biological Variations, Uncertainties and Sources of uncertainties**
3. **Terms- Population/Sample, Validity/ Reliability, Precision/Accuracy, Gaussian distribution, Confidence intervals.**
4. **Sampling Techniques**
5. **Summarisation of Medical Data-Averages, Tables, Graphs, Plots etc.**

Introduction

What is “Statistics”?

- **A science that deals with**
- **Collection,**
- **Organization/Classification,**
- **Tabulation,**
- **Analysis,**
- **Interpretation and**
- **Presentation of information, that can be presented numerically and/or graphically, to summarize results in a meaningful way, to help us answer a question of interest.**

Why Study Biostatistics?

- **Clinician-One drug superior to another with P<0.05**
- **To assess contribution of risk factors (hereditary, biological, environmental),in development of disease-e.g. Association between lifestyle and heart disease**
- **To decide between two or three alternative strategies treatment /diagnostic e.g. ELISA, NAT, Both**
- **Policy makers-Situational analysis for carrying out a control program e.g. AIDS control program, sex workers and IV drug users**
- **Administrators-e.g. Bed occupancy rates, projecting requirement, needs assessment**

Why Study Biostatistics

- **Biological variations and uncertainties bound to be present in Nature**
- **Biostatistics helps to control the impact of uncertainties, to measure its magnitude**
- **And to take valid decisions, that is least affected by such uncertainties.**
- **GOAL of Research-to minimize errors that threaten conclusions based on inferences/ to keep errors at an acceptable level**

Classification

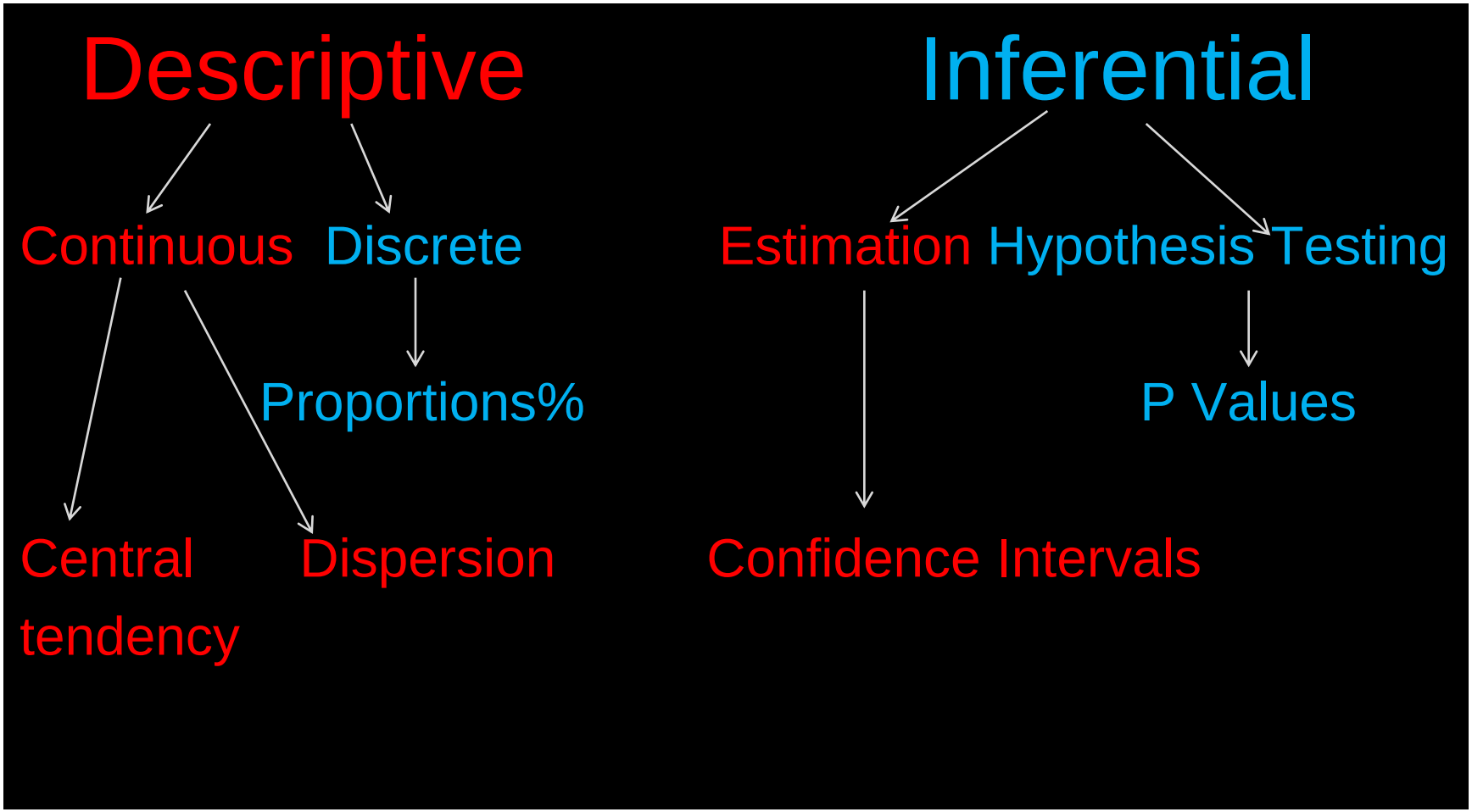
Types of Statistics

- **Descriptive Statistics- Description of data in a study**
- **Includes techniques for organizing, summarizing and presenting data using tables, graphs, plots etc.**
- **E.g. Mean, SD of height of children**
- **E.g. Proportion of overweight children is 20%**

Types of Statistics

- **Inferential Statistics- consists of statistical methods for making inferences about a population based on information obtained from a sample**
- **Estimation-Using results from a sample, to estimate levels in a larger population**
- **Testing Hypothesis-Comparing two groups and ascertaining whether there is true difference between the groups**

BIOSTATISTICS



Biological Variations &

UNCERTAINTIES

DEALING WITH IMPERFECT INFORMATION

Sources of Medical Uncertainties

1. **Genuine Variability-Biological, Environmental, Sampling Fluctuations, Chance variability**
2. **Variability due to un standardized methods-Observer variability, Instrument variability, Laboratory variability**
3. **Variability due to partial or erroneous information-Unavoidable incomplete information, Avoidable incomplete information, Partial compliance, Errors**

Sources-Genuine V

Biological Variability: from person to person, from time to time, within healthy groups and within sick groups e.g. BP, blood sugar levels

Environmental Variability: Nutrition, Smoking, Pollution e.g. water pollution, sanitation

Sampling Variability: Cannot study the whole population- sampling variability or sampling error e.g.-Two different samples show different results e.g.

Different mean HB.

Chance Variability: This is "Unknown" cannot be explained

V due to Un Standardized Methods

- **Observer V-Different opinions-e.g.,**

  X Ray findings in TB
- **Instrument V- e.g. B.P. on different instruments,**
- **Pain Intensity-on visual analogue scale and visual rating scale.**
- **Un standardized Questionnaire-inconsistent results**
- **Laboratory V- Improper Quality Control**

V due to Partial or Erroneous Information

- **Unavoidable Incomplete Info-e.g. comatose**
- **Avoidable Incomplete Info-e.g. particular test not done for economic reasons, unavailability of facility, suppression of facts in history taking**
- **Partial Compliance- leads to varied response of treatment e.g.- Effect of exercise, Drug**
- **Errors-Occur due to carelessness and lack of knowledge-wrong calibration, wrong reagents, wrong recording, misinterpretation e.g. Blood Group-Negative for positive**

Concept of Population and Sample

Population/Sample

- **Population-larger group of units that is the target of investigation-i.e., the group to which the findings are generalized**
- **E.g.-A drug for treatment of hyper emesis,**
- **Target population- All pregnant females with hyper emesis**
- **But cannot study the whole population- Why?? -Values for population are unknown**
- **Sample-i.e., a fraction or part of a population is actually studied.**

Population/Sample

- **SAMPLING VARIATIONS ARE BOUND TO OCCUR**
- **In view of this-**
- **Can samples be really used to estimate the characteristic of a population?**
- **Can legitimate inference be drawn about population from the sample results?**
- **If somebody else has done so can you believe these inferences?**
- **ANSWER IS--??**

Population/Sample

- **YES !!!!! PROVIDED-**
- **The sample is chosen with due precaution**
- **A. Random selection**
- **B. Inclusion of sufficient number of individuals so that the sample is able to represent the entire spectrum of features of the population-Representative Sample**
- **Then, sample mean,(median, proportion) is a "good estimate" of the target population.**

Parameter vs Statistic

- **DEF: A measurable characteristic of a population, such as a mean or standard deviation, is called a parameter, but a measurable characteristic of a sample is called a statistic.**
- **Mean of a population is denoted by the symbol μ**
- **Mean of a sample is denoted by symbol- $\bar{x}$**

	Population (Greek)	**Sample (Roman)**
	Parameter	Statistic
Mean	**μ**	**$\bar{x}$**
Standard Deviation	σ	**s**
Proportion	π	p

Increasing the Sample Size CLOSER TO 'TRUTH'

RQ: What is the Sero positivity for HIV among Blood donors in RMLH?

By this study sample we want to know the **TRUTH** among all Blood donors. By increasing Sample size we can get closer to the truth

Sample Size	No. Positive	%
10	x	0%
100	3	3%
1000	15	1.5%
10000	80	0.8%

Planning the Measurements

Precision and Accuracy

Precision

- **Repeatability/Reproducibility/Reliability /Consistency**
- **Precision-The degree to which the value of a measurement has nearly the same value when measured several times**
- **Precision has a very important influence on the 'Power' of the study**
- **The more precise the measurement the greater the statistical power at a given sample size to test mean values and to test hypothesis**

Precision is Affected by Random Error

Error: The difference between the measurement and the truth

- **Random error-A chance event**

Sources of error: observer, instrument, subject variability

Reliability

- **The extent to which repeated measurements of a stable phenomenon - by different people and instruments, at different times and places - get similar results**
- **Reproducibility and Reliability are similar**
- **Random error affects Reliability**

Reliability

- **Intra-Rater reliability**
  - **Comparison of measures by same person (test-retest; Repeatability)**
- **Inter-Rater Reliability**
  - **Comparison of measures by different people**

Accuracy

- **Accuracy-The degree to which the measurement actually represents what it was intended to represent**
- **How close is the measurement to the truth?**
- **Compare result to a 'GOLD STANDARD'**
- **Eg, serum sodium can be measured on an instrument recently calibrated against solutions made up with known concentrations of sodium**

Accuracy

- **Important influence on the Validity of the study-**
- **The degree to which the observed findings lead to correct Inference about phenomenon taking place in the study sample and in the universe**

Accuracy is Affected by Systematic Error

S.E.-The difference between the 'measurement' and the 'truth'

- **Systematic error/Bias**
  - **Design flaws**
  - **Protocol flaws (instrument or observer bias)**
    - **Use of a rectal thermometer for oral temperatures**
    - **Mis calibration of a scale**
  - **Subject limitations (subject bias)**
    - **Memory**
  - **Mistakes in Data entry and Analysis**

Sources of error**: observer, instrument, subject**

Precision/Accuracy

	PRECISION	ACCURACY
DEFINITION	The degree to which a variable has nearly same value when measured several times	The degree to which a variable actually represents what it is supposed to represent
Best way to assess	Comparison among repeated measures	Comparison with a reference standard
Value to study	Increase power to detect effects	Increase validity of conclusions
Threatened by	Random error – Chance Observer Subject Instrument	Systematic error- Bias Observer Subject Instrument

Strategies to Reduce Random Error-

Increasing Precision

- **Standardizing the measurement method**
- **Training and certifying the observer**
- **Refining the Instrument**
- **Automating the Instrument**
- **Repeating the measurement**

Strategies to Reduce Systematic Error-

- **Increasing Accuracy**
- **Standardizing the measurement method**
- **Training and certifying the observer**
- **Refining the instrument**
- **Automating the instrument**
- **Making unobtrusive measurements**
- **Calibrating the instrument**
- **Blinding**

Strategies to Reduce R Error Increasing Precision

	Source	Error	Prevention
Standardizing methods-opera-manual	Observer	Variation in BP due to variable rate of cuff deflation	Specify how cuff will be deflated 2mm/sec
	Subject	Variable length of quiet sitting before BP check	Specify
Training observer	Observer	Variable observer technique	Training
Refining the Instrument	Instrument+ observer	Rounding off of BP	Conceals the reading
Automating the instrument	Observer	Technique Variable	Automatic device
	Subject	Emotional Reaction	Automatic Device
Repeating the measurement	O+S+I	All sources of measurements	Use mean of two or more measurements

Strategies to Reduce Systematic Error-Increase Accuracy

	Source	Error	Prevention
Making unobtrusive measurements	Subject	Tendency to overestimate compliance with drug	Measure drug level in urine
Calibrating the Instrument	Instrument	High reading adjustment wrong	Periodic calibration
Blinding	Observer	Reads lower BP in t/t group	Double blind placebo Assignment concealment
	Subject	Tendency to over report side effects	Same

Internal /External Validity

Validity of the Study

- **DEF: Validity is the degree of correctness of the results.**
- **Did you measure what you were intended to measure? (Accuracy)**
  - **Internal validity**
  - **External validity**

Internal Validity

Internal validity: **DID the study actually measure what it set out to? Are the findings true for the study subjects?**

Depends on- Selection bias, information bias, measurement bias and confounding are threats to internal validity

- If the study has been undertaken using **wrong methods**
- E.g, use of outdated kits, untrained technician, hemolysed samples
- This is a problem of "**Internal validity**"
- The study has to be rejected, **Achieving internal validity takes a priority**

External Validity

External validity: To what extent are the findings generalizable to the population?

Depends on- Inclusion & Exclusion criteria

E.g,HIV Sero Positivity rates among those attending STD clinics?

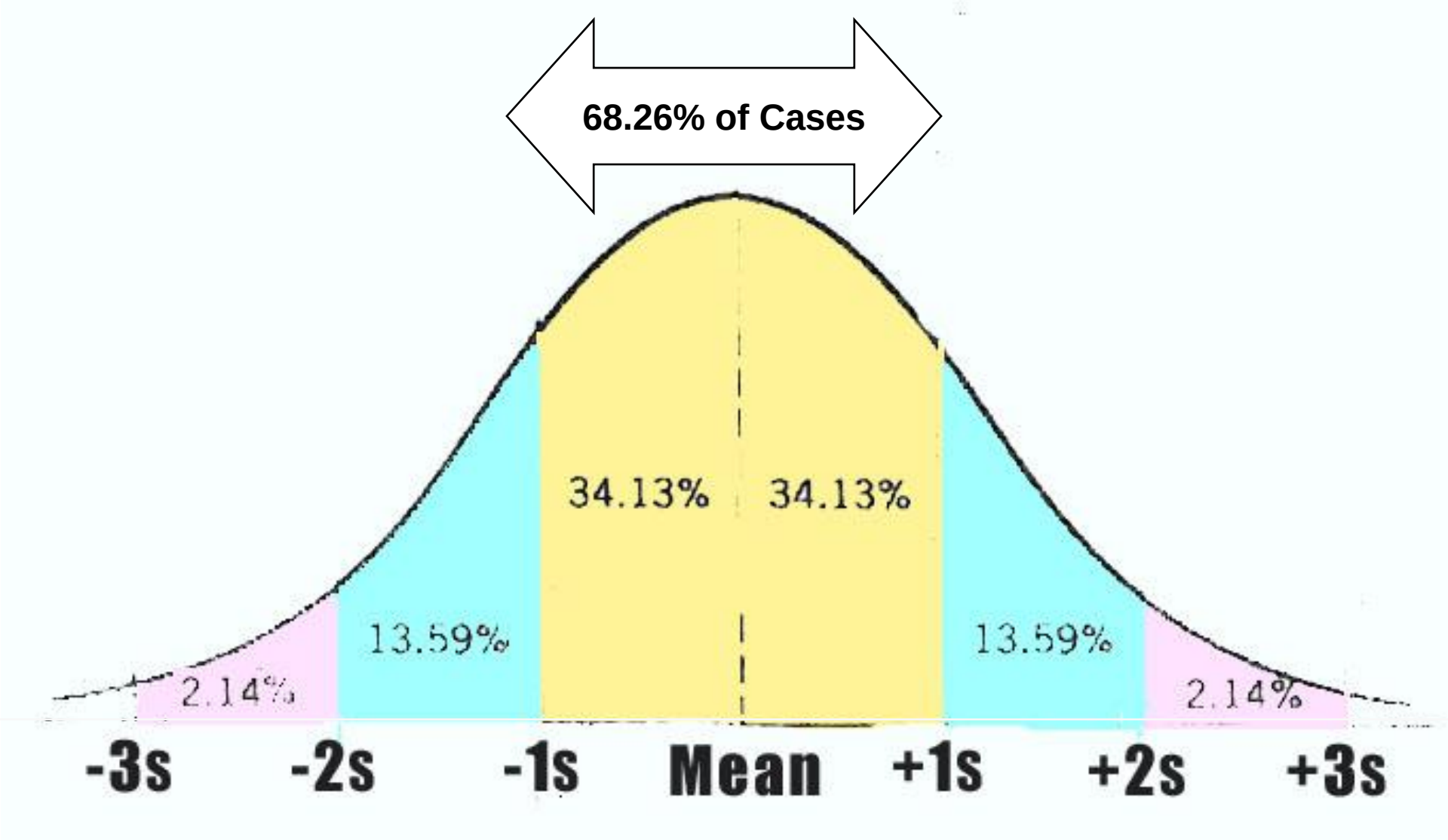
- **Sample is Non Representative- Systematically different from the General population**
- **"Loss of generalizability" if this sample is used to predict the prevalence rate in the population**
- **Problem of "Representativeness" arises**

Normal Distribution

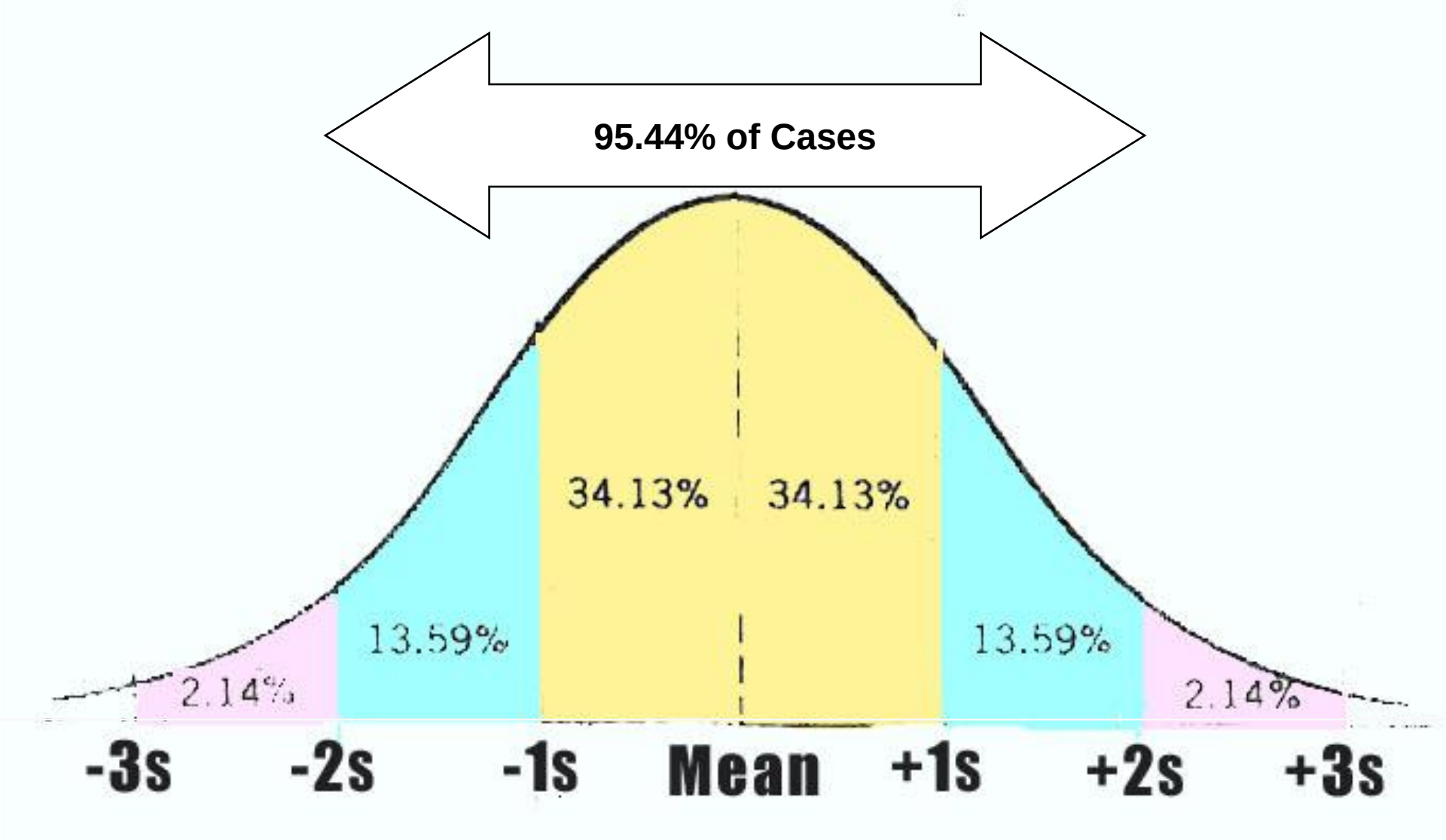
Properties Of Normal Curve

- **Normal curves are symmetrical, uni modal, have a bell-shaped form.**
- **Mean, median, and mode all have the same value, are in the centre**
- **How do we know if data is normally distributed ?**
  - **Visual appearance of histogram**
  - **Skewed: If SD > half of mean**

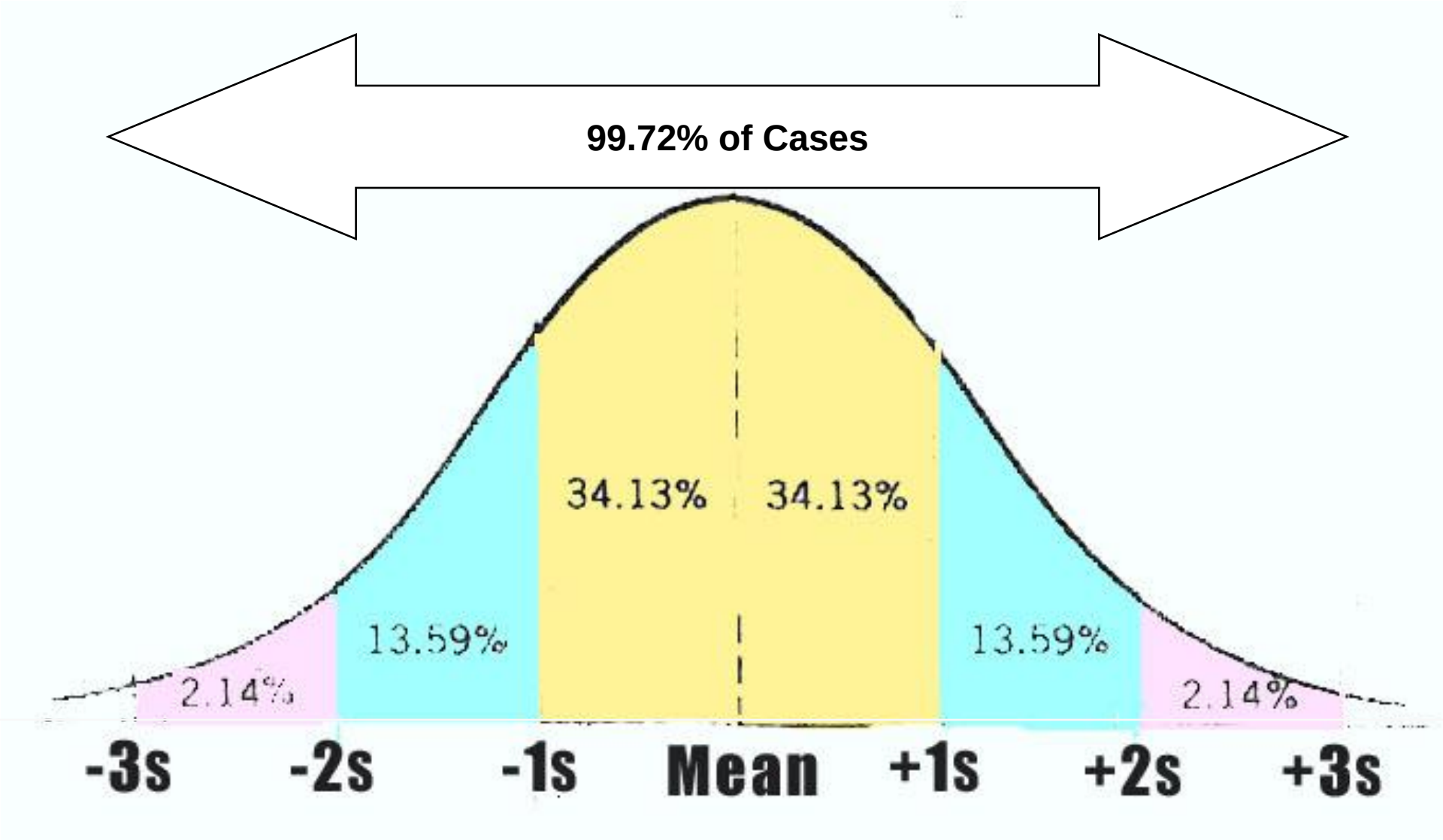
Percent of Values Within One Standard Deviations



Percent of Values Within Two Standard Deviations



Percent of Values Within Three Standard Deviations



Confidence Intervals

Confidence Interval

- **CI provides a Range that is likely to contain the population mean when so obtained for repeated samples.**
- **Conventionally 95% is fixed as a limit and is called Confidence level**
- **Eg. A range from 10mg/dl-16mg/dl has 95 % chances of containing the actual population mean for HB, such a range is called 95% CI.**
- **The limits 10 and 16 are called confidence limits.**
- **There is only 5 % chance that the actual population mean is outside the range.**

Confidence Interval

Example-:

In a study of a sample of 100 subjects it was found that the mean systolic blood pressure was 120 mm Hg with a standard deviation of 10 mm Hg. Find out 95% confidence limits for the population mean of systolic blood pressure.

SE = SD / (n) = 10/ (100) = 10/10 =1

LL :--- mean - 1.96*1 :--- 120 - 1.96 = 118.04

UL :--- mean +1.96*1 :--- 120 + 1.96 = 121.96

i.e. the population mean value of systolic blood pressure will lie between 118.04 and 121.96 and we can have a confidence of 95% for making this statement.

Confidence Interval

- **CI can be calculated for the following:**
- **Is our question that of estimation-estimation of mean, proportion**
- **Is our question that of Hypothesis testing-**
- **Difference between two means**
- **Difference between two proportion**
- **Relative Risk,Odds Ratio**
- **‘Smaller the CI- more Precise the measurements**
- **Wider CI- Increase the sample size’**

In Summary

1. **All biological phenomenon-differ in characteristics from each other**

"Variability"

2. **We cannot study the entire population so we have to study SAMPLES-**

"Representative"

Ensures external validity/ generalisability

#Role of Biostatistics

Summary contd…

3. Sample should be of **Adequate Size** to obviate the play of '**chance**'

4. Select correct tools and techniques for correct measurements-to measure what we really "**intend to measure**",

   To ensure "Internal validity"

Summary contd…

5. This data should be collected, presented and analyzed in a scientific way, so that meaningful inferences could be drawn
6. Analyse data appropriately using appropriate statistical tests

" the extent of chance, random or sampling variation" should be kept at a minimum acceptable level"

Summary

7. **Interpret results according to their clinical or public health significance and not simply go by statistical significance-**

   “Hurray the “p” value is <0.05,lets celebrate!!!!

8. **Draw correct inferences and Generate new research Hypothesis**

TAKE HOME MESSAGE

M, x̄ ,σ, s, π, p ~!!!!!