Introduction to Statistics

Module 12 Topic 0

Definition

Statistics is the science of collecting, organizing, summarising, analysing, and making inference from data

Descriptive stat. Includes collecting, organizing, summarising, analysing, and presenting data

Inferential stat. Includes Making inferences, hypothesis testing Determining relationship, and making prediction

Dr.H.Qotba

Introduction

- The word Statistics derived from Latin “status” meaning a political state. It was simply the collection of data, by kings, on different aspects useful to the state (1)
- Statistics is concerned with scientific methods for collecting, organising, summarising, presenting and analysing data as well as deriving valid conclusions and making reasonable decisions on the basis of this analysis.(2)

Introduction (contd)

- **Definition by Horace Secrist:** Statistics may be defined as the aggregate of facts affected to a marked extent by multiplicity of causes, numerically expressed, enumerated or estimated according to a reasonable standard of accuracy, collected in a systematic manner, for a predetermined purpose and placed in relation to each other(2)

Functions and scope of Statistics

- Condensation, comparison, forecasting, estimation and hypothesis testing(2)
- Scope in various areas(2)
  - Industry
  - Commerce
  - Economics
  - Education
  - Agriculture
  - Planning
  - Medicine

Limitations of Statistics

- Statistics is not suitable to the study of qualitative phenomenon
- Statistics does not study individuals
- Statistical laws are not exact
- Statistics is only, one of the methods of studying a problem (2)

Misuse of Statistics[3]

- False percentages
  - Adding percentages – taxi fares should be increased by 35 % because petrol went up by 10%,spares 15%, price index 10%
  - Decreasing percentages- strawberry yield reduced by 100%over last year *(impossible!!)*
  - Huge percentages - A earns 500% more than B *(actually 6 times !!!)*
  - Unsupported percentages – Body mass pwd showed weight gain in 80% people *(Was tried in 5,of which 4 were on regular exercise regimen!!!)*

Misuse of Statistics (contd)

- Fictitious precision
  - 13,12,0015 people in India speak Tamil *(how did they get the last 15???)*
  - Results up to many decimal places
  - Ask 100 patients how many days was their last bout of cold.
  - Patients recall is generally imprecise
  - Total adds to 689.3 days

  So average is 6.893!!!
- False Comparisons
- It is safer to be in the army since the death rate is 7/1000 vs. 17/1000 in civilians
- The army has able bodied youth vs. civilians of all ages!!

Misuse of Statistics (contd)

- Misleading presentation
  - Ina company annual report the same revenue number can be expressed as 1% return on sales,15% return on Investment, 1000% profit over 5 years back or 40% loss over last year
  - Misleading presentation using different scales
- Incomplete data
  - 3 independent reports showed that the risk of being killed was 5.5 times greater for persons small cars

(The same reports also mentioned that small cars got into lesser number of accidents!!!)

Deliverables

- You will understand statistics so
  - Planning a research project is possible
  - Inferences can be drawn from preliminary data
  - The statistical report can be understood
- You will not be mislead by play of numbers
- **Ralph Waldo Emerson**
  - “If you learn only methods, you'll be tied to your methods, but if you learn principles you can devise your own methods."

Samples and measurement

- The external validity of the results of statistical analyses, while key to the value of those results, is importantly a function of measurement and sampling
- A perfectly chosen and executed analysis will be at best misleading if it is conducted of the wrong data or data collected using an inaccurate measurement technique, or at the wrong time, and so on
- Garbage in garbage out
- Ref 4. .. Barkan H, Annals of Cardiac Anaesthesia (2015)18:74

Quantitative Techniques

- Powerful tool for helping making decisions regarding business or Industry
- Used for studying large number of items
  - Statistical Analysis
  - Programming Techniques (Operation Research)
- Statistical Analysis
  - Study of large number of items for inferential purposes
    - Census analysis
    - Sample analysis

What is census?

Collecting data of the entire population for the purpose of study

- Very accurate
- Time consuming and costly

Sampling

- Picking up of representative subset from the total population that is precise, accurate & reliable
- Benefits of Sampling
  - High feasibility
  - Lower cost than census
  - Shorter time required
  - Destructive analysis

Sampling (contd)

- A method to study a part of the whole or aggregate in order to draw inferences about the whole or aggregate
- Deliberate Sampling
  - Choosing samples with a deliberate bias
- Random Sampling
  - Unbiased selection of samples

Random Sampling

- Is one where EACH item in the population has the same chance of selection
- For a small sample use Random Numbers table or generate random numbers using simple computer program or even through a scientific calculator
- Random assignment can be done even using a flip of the coin. Heads – group A and Tails Group B

Deliberate Sampling

- Also known as purposive, non-random or judgment sampling
- Needed in special cases
- Makes studies easier and purposeful
- Cannot be used in all situations

Selective Sampling

- This is Non Probability sampling method, convenient and easy
- **Systematic Sampling:** selecting nth object

  e.g. : mailing list, production run (time, no.)
- **Stratified Sampling:** Population is divided (stratified) into groups (strata) based on a characteristic say smokers & non smokers and then a sample is formed (using Random Numbers) from the strata to achieve a certain composition

Selective Sampling (contd)

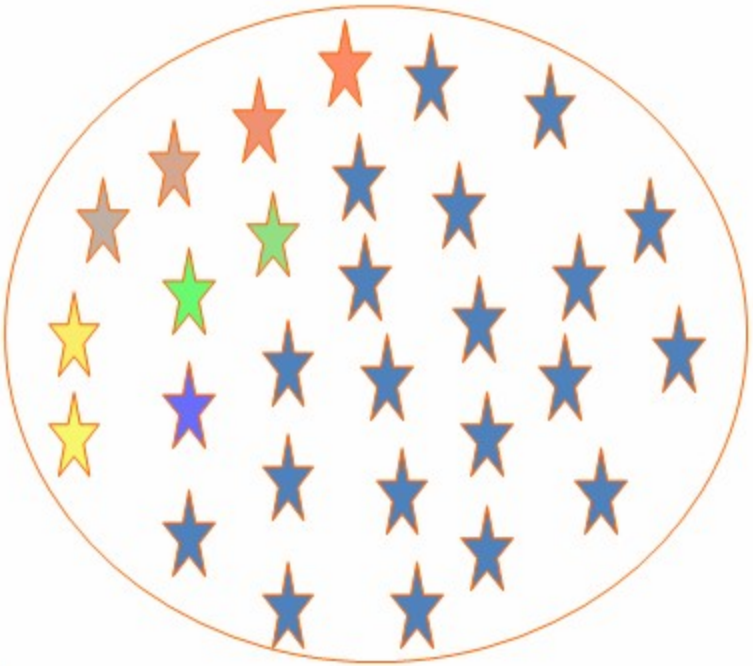
- Cluster Sampling : (Or) Multistage sampling is employed when there are many “primary units” that are Clustered together in “Secondary” larger units

e.g. : tablets packed in a bottle OR medicinal powder packed in large drums

Sampling Essentials

- Sample should be a true representative
- It should not be biased unless deliberate
- Possible to calculate error of sampling
- Results should be applicable to the universe.

Sampling



1/30

1/29

1/28

1/27

Sample with replacement

- Keep chances of each item equal
- True randomization
- Replacement not required when population is so large that drawing of samples does not affect chance

Measurement scales

- Nominal
- Ordinal
- Interval

Nominal scales (4)

- Use distinct and mutually-exclusive numbers to name each category of observation.
- Nominal scalings only classify observations.
- The numbers assigned in a nominal scale carry no further information about magnitude.
- Clinical examples of nominal scalings include
  - any notation that a disease is (simply) present or absent
  - demographic measures (such as gender and ethnic group),
  - and disease classification systems such as the International Classification of Disease (ICD)-10 and the Diagnostic and Statistical Manual of Mental Disorders-5

Ordinal Scales

- Ordinal scalings place observations in order- from least to most- but are not able to specify or compare the differences between pairs of measurements.
- Clinical measurements are ordinally scaled: e.g.
  - tumor grade,
  - pain scales,
  - Likert attitude scales

Interval scales

- Interval scales place observations in order and specify both the magnitude of individual measurements and the distance between pairs of measurements
- Interval scalings permit all the basic arithmetic operations and the calculations e.g.
  - anthropometric measurements of height and width,
  - blood pressure,
  - duration of time intervals

Discreet and continuous scales

- Two other interval scale options .The first is whether the source measurements are:
  - Discreet (e.g. number of children in the household) or
  - Continuous (e.g. blood pressure)
- This distinction bears on the source measurement and may influence how collected data are displayed graphically, but has no influence on the choice or calculation of statistical analyses

Ratio scale

- Another difference among interval scales is whether or not the scale has a true “0” point
- Those with a true “0” points are sometimes called ratio scales because the presence of a true “0” point makes division and hence the calculation of ratios possible
- The Kelvin scale has a true zero point at absolute zero and hence is a ratio scale. The Fahrenheit scales has a zero point that’s mathematically arbitrary and hence are interval scales

Ratio scale (contd)

- This difference bears on which conclusions regarding these measurements are meaningful. e.g. it is meaningful to say that the temperature of 30°K is half a temperature of 60°K while it is not valid to make the same statement regarding 30°F vs 60°F
- This difference has no bearing on the choice of statistical procedures to analyze these data

Frequency distribution

- Frequency distribution is a series when a number of observations with similar or closely related values are put in separate bunches or groups, each group being in order of magnitude in a series
- It is simply a table in which the data are grouped into classes and the number of cases which fall in each class are recorded

Frequency distribution (contd)

- It shows the frequency of occurrence of different values of a single Phenomenon
- Why frequency distribution
  - To facilitate the analysis of data.
  - To estimate frequencies of the unknown population distribution from the distribution of sample data and
  - To facilitate the computation of various statistical measures(2)

Discrete frequency distribition

- The frequency refers to discrete value
- Here the data are presented in a way that exact measurement of units are clearly indicated
- There are definite difference between the variables of different groups of items
- Each class is distinct and separate from the other class
- Non-continuity from one class to another class exist
- Examples Data as such facts like
  - The number of rooms in a house,
  - The number of companies registered in the country over years
  - The number of children per family in a locality

Freuqency tabulation

- In a survey of 40 families in a village, the number of children per family was recorded and the following data obtained
- 1 0 3 2 1 5 6 2 2 1 0 3 4 2 1 6 3 2 1 5 3 3 2 4 2 2 3 0 2 1 4 5 3 3 4 4 1 2 4 5

Freuqency tabulation (contd)

# of children	Tally marks	Frequency
0	III	3
1	𝍸 II	7
2	𝍸 𝍸	10
3	𝍸 III	8
4	𝍸 I	6
5	IIII	4
6	II	2
Total		40

Continuous Frequency Distribution

- A Class Interval is a subdivision of the total range of values which a Continuous variable may take
- The Class frequency is the number of observations of the variable which fall in a given interval
- The frequency distribution of a (continuous) variable is the set of class intervals for the variable, together with the associated class frequencies

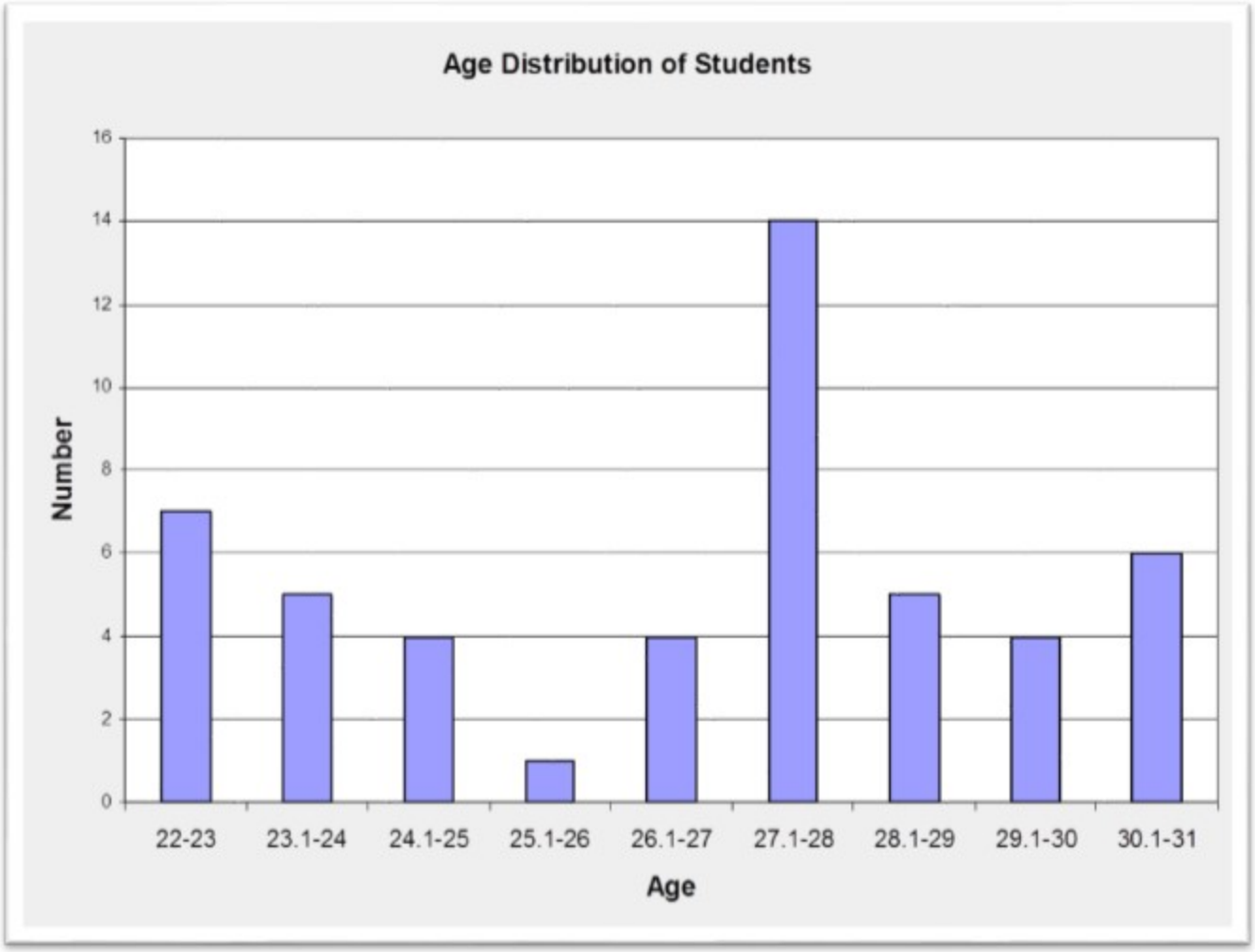
Frequency distribution

- 22, 22.25, 22.3, 22.6, 22.8, 23, 23, 23.2, 23.25, 23.5, 23.6, 23.75, 24, 24.2, 24.4, 24.6, 25, 25.1, 26.6, 26.6, 26.7, 27, 27.1, 27.15, 27.18, 27.2, 27. 3, 27.4, 27. 5, 27.5, 27.6, 27.7, 27.75, 27.8, 27.9, 27.95 28.1, 28.4, 28.5, 28.6, 28.7, 29.1, 29.2, 29.5, 29.8, 30.1, 30.4, 30.6, 30.7, 30.8, 31,

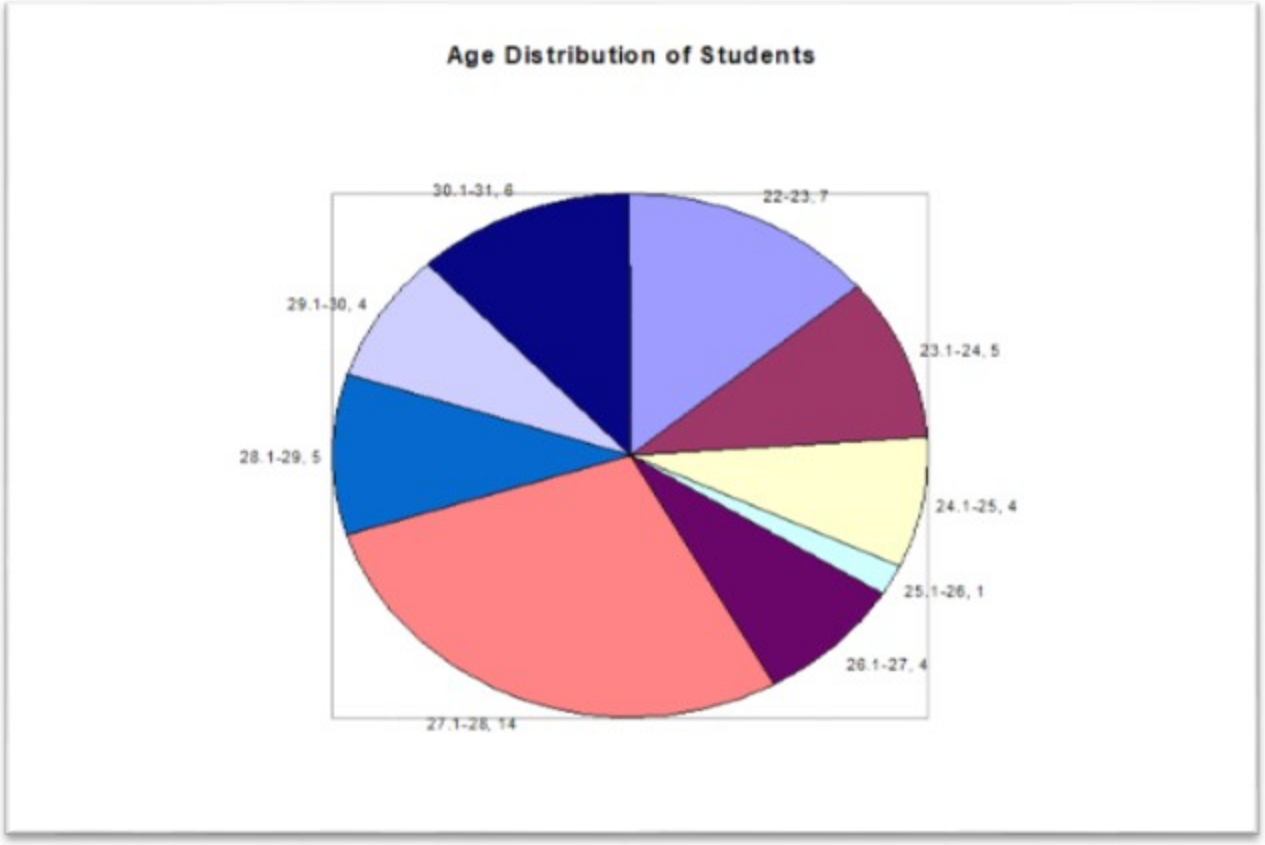
Frequency distribution

- 22-23 7
- 23.1-24 5
- 24.1-25 4
- 25.1-26 1
- 26.1-27 4
- 27.1-28 14
- 28.1-29 5
- 29.1-30 4
- 30.1-31 6

Graphic Displays



Graphic Displays



References

- Aggarwal YP, Statistical Methods, concepts, application, computation. Sterling publishers(1986)
- Varalakshmi V et al. Statistics – Higher Secondary, First Year. Tamil Nadu Textbook Corporation 2005
- Langley R, Practical Stats. The Chauser press 1968
- Barkan H, Annals of Cardiac Anaesthesia (2015)18:74
- Dunn OJ, Clark VA, Basic stats, a primer for biomed sciences. 4th Ed, 2009, John Weiley and sons inc.
- Harris M, Taylor G. Medical Statistics made easy. 1st Ed, Martin Dunitz, 2003
- Krousel wood M, clinicials guide to stats part 1 The Ochsner Journal (2006) 6:2,68
- KadamP ,Sample size calculation. International J of Ayurveda (2010)1:55-57
- Fox N, Hunn A, Mathews N, Sampling and sample size calculation. The NIHR RDS EM / YH(2009)