Introduction to statistics

There are three kind of lies-"Lies, Damned lies and Statistics"

Statistics

- The word Statistics conveys a variety of meaning to people in different walks of life. Initially it was termed as "The Science of Statecraft" or "The Science of King".
- The word statistics comes from the Italian word **Statista** meaning **STATEMENT**
- The German word **Statistik** meaning **POLITICAL STATE.**
- **The processing of statistical information has a history that extends back to the beginning of humanity. As early as 3800 B.C., there were records of population in Babylonia and China.**

History

- The biblical census was conducted by David in 1017 B.C. and by Moses in 1491 B.C.
- Methods of census taking were also documented in Indian history, which dates back to the time of the northern Hindustani ruler Asoka (270–230 B.C.).
- When the food supply was in danger, the Athenians and other ancient Greeks would count the adult male citizens during times of war as well as the general populace.
- For military and administrative purposes, adult males and their belongings were registered by the Romans.
- Every male in the Roman Empire was required to come back to his birth Two thousand years ago, each male in the Roman Empire had to return to the city of his birth to be counted and taxed. Thus, the Bible gives an account of the return of Joseph and Mary to Bethlehem for such purpose, (Bible, Luke 2: 4-5)

Intro Contd…

- Statistics: Statistics is the branch of scientific knowledge refers to the body of techniques and methodology developed for the collection, classification, organization, presentation and analysis of statistical data and for the use of such data in decision-making in the face of uncertainty in any field of enquiry.

Characteristics

- Aggregate of Facts
- Affected to a substantial extent by a variety of reasons
- Numerical expression
- Enumerated and Estimated as per reasonable standard of accuracy
- Data collection is carried out in a systematic manner
- Data must be placed in relation to one another

Function

- Reduces complexities.
- Expresses facts in numbers.
- Presentation of data in condensed form.
- Increases the individual knowledge and experience.
- Different phenomena are compared.
- Helpful in the formulation of policies.
- Helpful in prediction and forecasting.

Limitation

- Statistics does not deal with individuals.
- Only deals with quantitative data.
- It may mislead to wrong conclusion in the absence of control.
- statistical laws are truly on averages.
- It does not reveal the entire story.
- Data should be uniform and homogeneous.
- Statistics is labeled to be misused

Uses of Statistics

- Quality Control and Reliability Testing
- Signal Processing
- Data Analysis and Interpretation
- Control Systems
- Telecommunications
- Reliability Engineering
- experimental Design
- Machine Learning and AI
- Power Systems Analysis
- Electromagnetic Compatibility (EMC)
- Optimization Problems
- Market Analysis and Forecasting

Important definitions

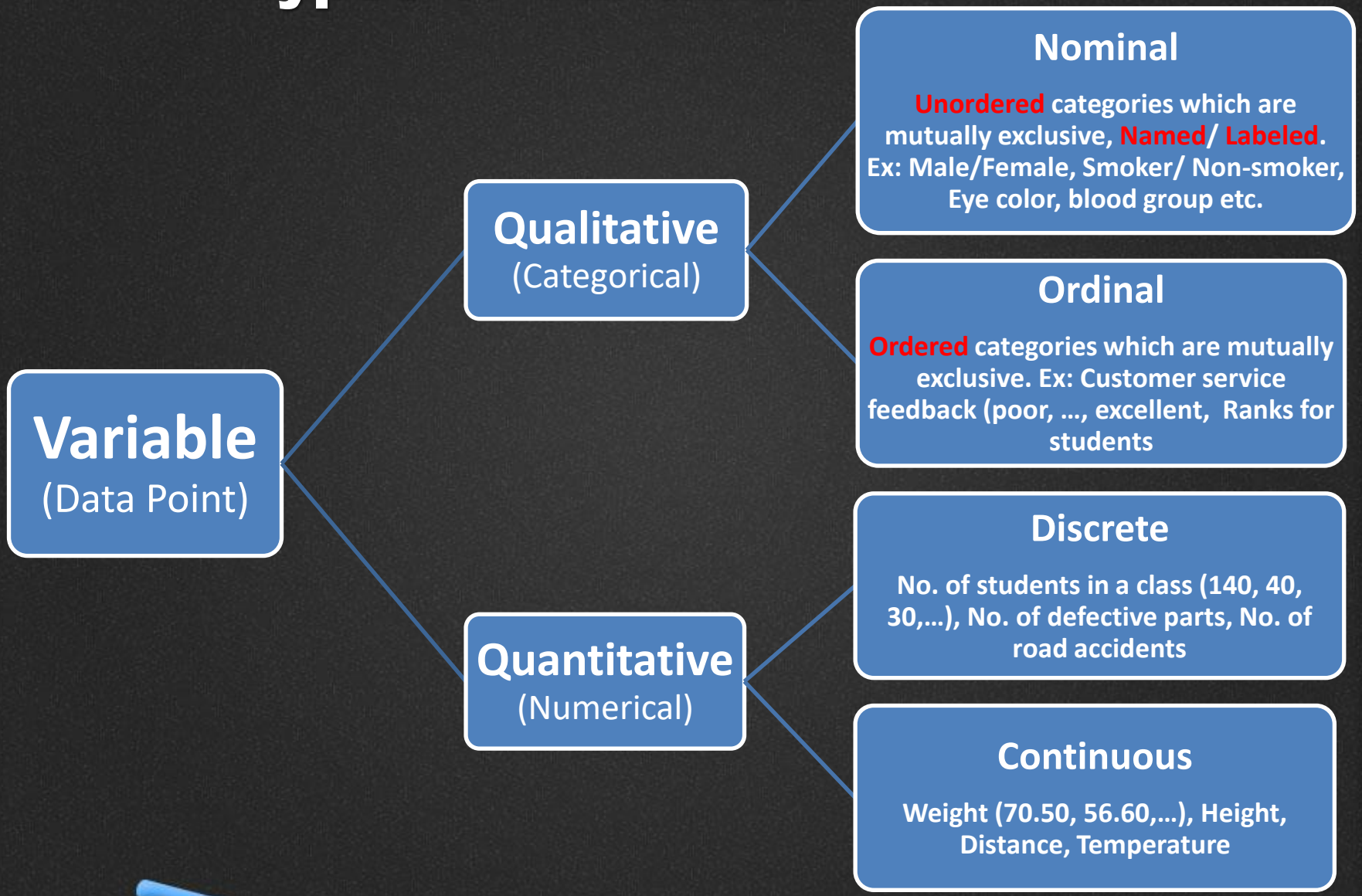
- Population: Population is the totality or collection of all objects or individuals on which observations are taken on the basis of some characteristics of the objects in any field of inquiry.

  i) Finite Population

  ii) Infinite Population

- Sample: A sample is a sub-set or part of the population selected to represent the population
- Experimental Unit: Each individual of a population is called an experimental unit. Observations are collected on experimental units.
- Statistic: Any numerical value which we calculate from the sample is called a statistic. Or, A descriptive measure of a sample is called statistic.
- Parameter: A descriptive measure of a population is called parameter
- Census: A process of gathering data from the whole population for a given measurement of interest.

Types of Variables

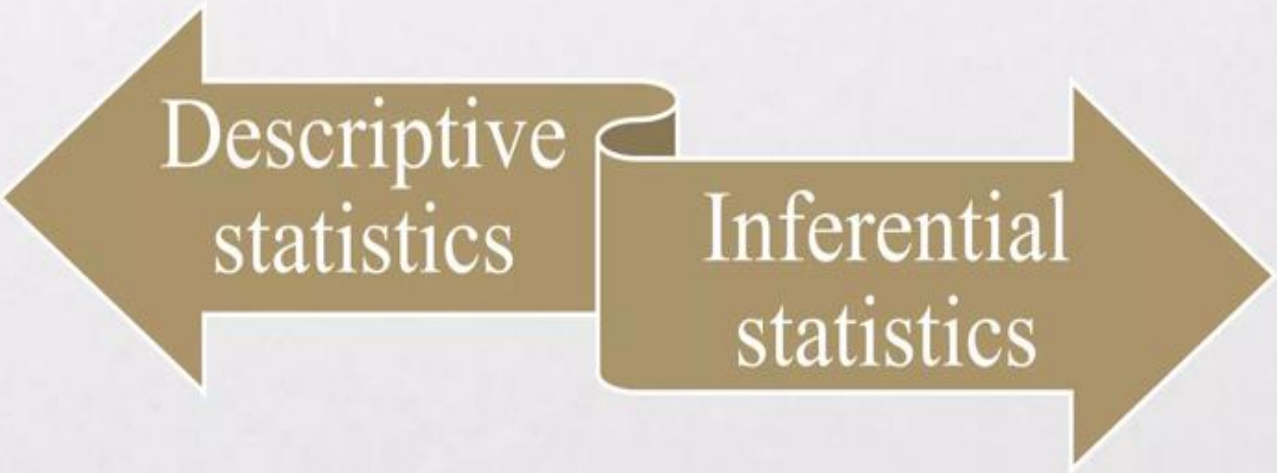


Types of Data

- According to origin (population data, sample data)
- According to variable (qualitative data, quantitative data, discrete data, continuous data)
- According to scale of measurement (nominal data, ordinal data, interval data, ratio data)
- According to time (time series data, cross section data, panel data),
- According to sources (primary data, secondary data),
- According to subject

Types of Statistics

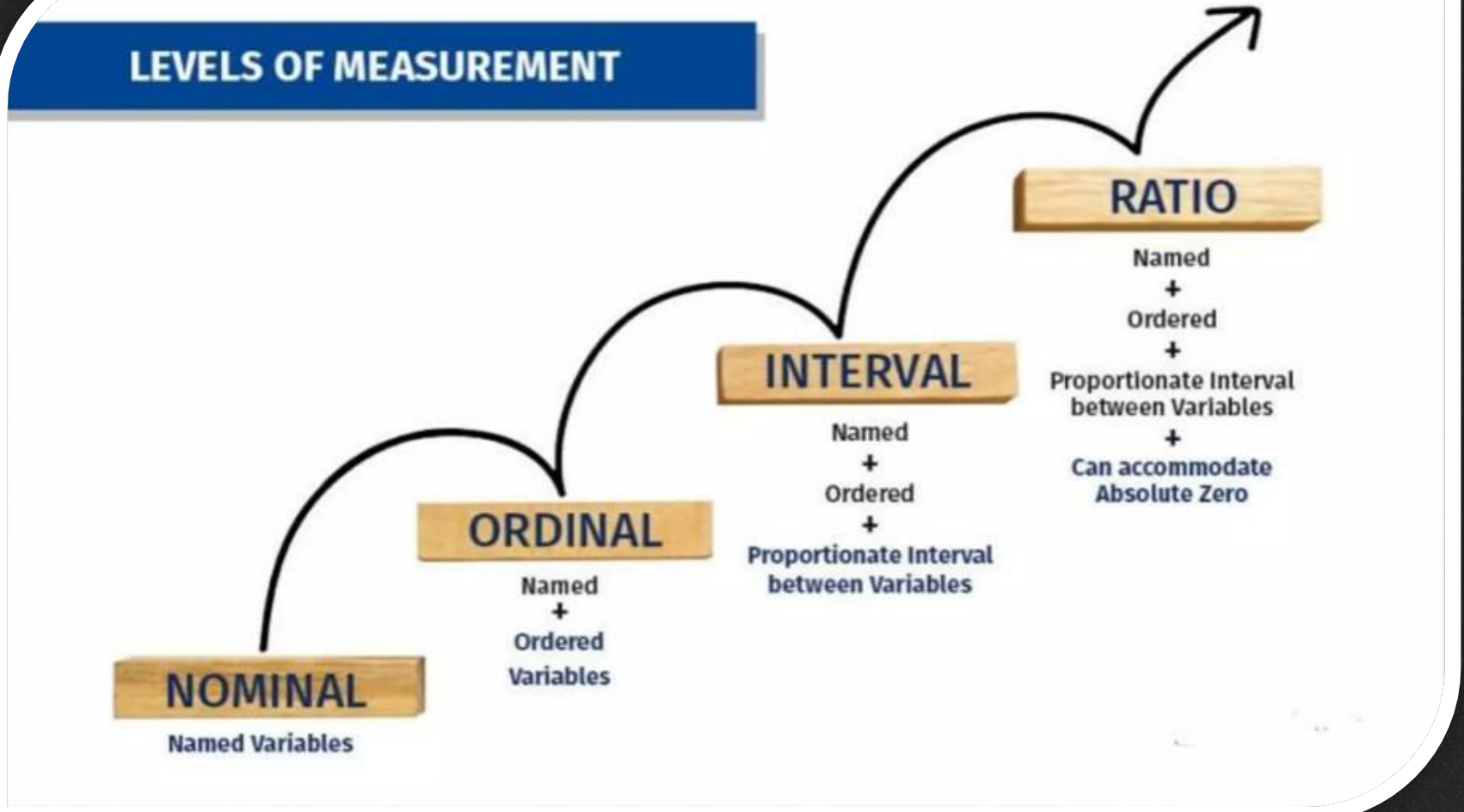
Statistics can be divided into two branches: Descriptive statistics and Inferential statistics.



Types of Statistics

- **Descriptive Statistics**: It comprises those methods concerned with collection and describing a set of data so as to yield meaningful information.
- If the data state is the entire population call mom we did only to draw conclusion based on the descriptive statistics.
- If it is not possible to get population data due to time, cost or other considerations, we have to take sample from the population.
- **Inferential statistics**: It consist of procedures used to make inferences about population characteristics from information contained in a sample drawn from the population.

Scales of measurement



Characteristics of different level of measurement

Scales	Characteristics	Examples
• Nominal	• Categories are homogeneous, mutually exclusive, end no assumptions about ordered relationships between categories made	• Sex of subject • eye color • Religion • political affiliation • place of residence • room number
• Ordinal	• All the above plus the categories can be rank ordered	• Examination grade • health status • level of education • rank in job
• Interval	• All the above plus exact differences between categories are specified and an arbitrary zero point is assumed	• Temperature, • IQ test score, • calendar time
• Ratio	• All the above with the exception that a true zero point is assumed	• Height • weight • fat consumed • wage

Collection of Data

- Data, whether qualitative or quantitative originate from two major sources:
  - i) Primary sources
  - ii) Secondary sources
- Data originated from primary sources are called primary data and those from secondary sources are secondary data.

Primary Data

- Focus group (8-12 people)
- Survey (Questionnaire, most common method in survey)(most common method in social sciences, management etc.)
- Interview (face-to-face)
- Observation(observing someone with or without letting him know)
- Experimentation
- Online survey or Internet survey

Secondary Data

- Internal Sources (Profit and loss statements, Balance sheets, Sales figures, Inventory records)
- External Sources (Govt. publications, Non-Govt. publications, Reviewing relevant literature)

Presentation of Data

- Once classification has been done, the classified data can be condensed and summarized in two basic forms:
  - Tabular form(known as Frequency distribution), and
  - Graphical form.

Frequency Distribution

- The most convenient method of organizing data is to construct a frequency distribution
- A frequency distribution is a tabular summary of data where observations are divided into different non overlapping classes or categories and frequency of each class or category is arranged accordingly.
- According to nature of variable, there may be three types of frequency distribution, namely:
  - Frequency distribution for qualitative or categorical variable
  - Frequency distribution for discrete variable and
  - frequency distribution for continuous variables.

Frequency Distribution

Leisure activity	Frequency	
	Count	Percentage
Watch TV	84	42
Read	26	13
Listen to music	46	23
Watch a video	24	12
Phone friends	8	4
Other	12	6
Total	200	100

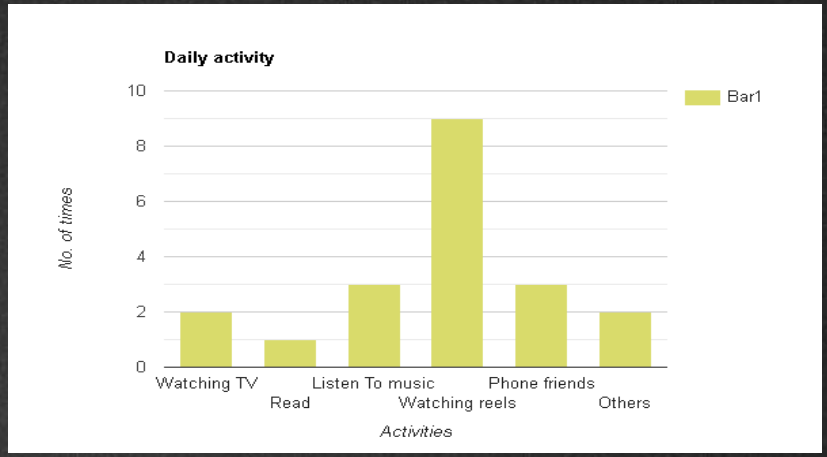
Frequency Distribution
(Categorical Data)

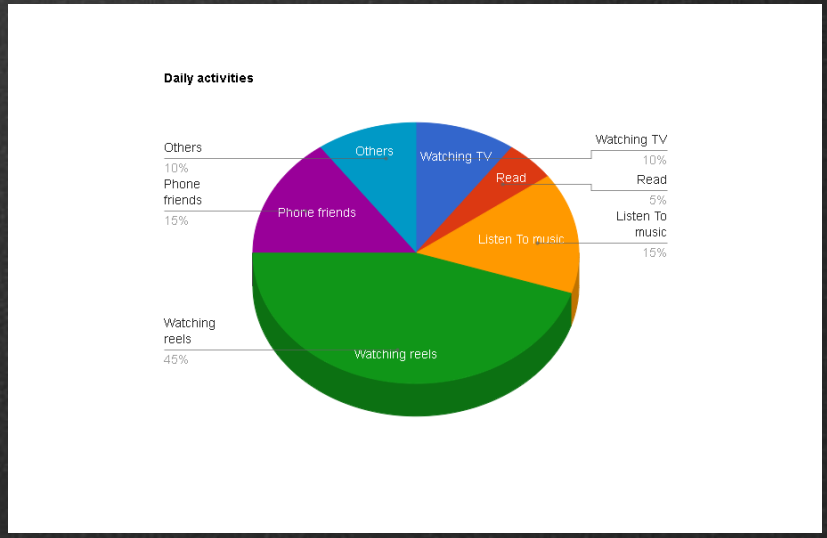
Consider an example of your weekly activities:

Weekly Activity	Frequency				
	Tally	Count	Relative	Percentage	Angles in degree
Watch TV	||	2	0.1	10	36
Read	|	1	0.05	5	18
Listen to music	|||	3	0.15	15	54
Watching reels	𝍸 𝍸	9	0.45	45	162
Phone friends	|||	3	0.15	15	54
Others	||	2	0.1	10	36
Total		20	1.00	100	360

Diagrams for Categorical data

- Bar Diagram
- Stacked Bar diagram
- Cluster or multiple
- Bar diagram
- Pie Diagram
- Pareto Diagram
- Pictogram





Example

- The management of an electrical components factory wants to know the monthly production pattern of different units. In this connection, a survey was conducted on 50 units of the factory. The following data give the number of units produced per month by 50 different machines:

140, 144, 187, 87, 40, 122, 203, 148, 150, 165, 133, 195, 151, 71, 94, 87, 42, 30, 62, 103, 204, 162, 149, 79, 113, 69, 121, 93, 143, 110, 175, 161, 157, 155, 108, 164, 128, 114, 178, 130, 156, 167, 124, 164, 146, 116, 149, 104, 141.

- Construct a frequency table using all the three methods

(i)Inclusive method (ii) Exclusive method (iii)Class boundary method (iv) relative frequency (v) Percent frequency (vi) Cumulative frequency (vii) Relative cumulative frequency (viii) Percent cumulative frequency

- Draw Histogram, Frequency Polygon, Frequency curve, Cumulative frequency Polygon or Ogive Polygon, Ogive curve

Frequency distribution (grouped data)

- Following steps need to keep in mind in constructing frequency distribution table
  - Array the given raw data in ascending order.
  - Find the largest and smallest values. Compute the $Range = Maximum - Minimum$
  - Compute for the tentative number of classes (K). The ideal number of classes is between 5 and 20 or should be $\sqrt{n}$ you may use the **Sturge's Method**:

$$K = 1 + 3.322 \log_{10} n$$

Where: K = tentative number of classes

n = total number of observations

log = common logarithm (base 10)

Frequency distribution cont.

- Compute for the Class Interval (I) by dividing the range by the tentative number of classes (K). Use class interval rounded to the nearest whole number.
- Sort the arrayed data into appropriate classes using convenient and easy to read class limits. Start the first class with a lower limit either equal to or a little bit less than the lowest observed value.
- Set up the class boundaries if necessary.
- Determine the class mark or midpoint.
- Count or tally the number of observations into the appropriate class intervals.
- If necessary, find the relative frequencies and/or relative cumulative frequencies.

Frequency Distribution Cont.

Class interval(Exclusive method)	Class interval(Inclusive method)	Class interval (Class boundary method)	Tally Marks	Frequency
30-55	30-54	29.5-54.5	𝍸	3
55-80	55-79	54.5-79.5	𝍸𝍸	4
80-105	80-104	79.5-104.5	𝍸𝍸𝍸	6
105-130	105-129	104.5-129.5	𝍸𝍸𝍸𝍸	10
130-155	130-154	129.5-154.5	𝍸𝍸𝍸𝍸𝍸	12
155-180	155-179	154.5-179.5	𝍸𝍸𝍸𝍸	10
180-205	180-204	179.5-204.5	𝍸𝍸	5
Total				50

Frequency distribution cont.

Class Interval	Mid value	Frequency	Relative Frequency	Percent Frequency	Cumulative Frequency	Cumulative frequency	Relative cumulative frequency	Percent cumulative frequency
30-55	42.5	3	0.06	6	3	50	0.06	6
55-80	67.5	4	0.08	8	7	47	0.14	14
80-105	92.5	6	0.12	12	13	43	0.26	26
105-130	117.5	10	0.20	20	23	37	0.46	46
130-155	142.5	12	0.24	24	35	25	0.70	70
155-180	167.5	10	0.20	20	45	15	0.90	90
180-205	192.5	5	0.10	10	50	5	1.00	100
Total		50	1.00	100				

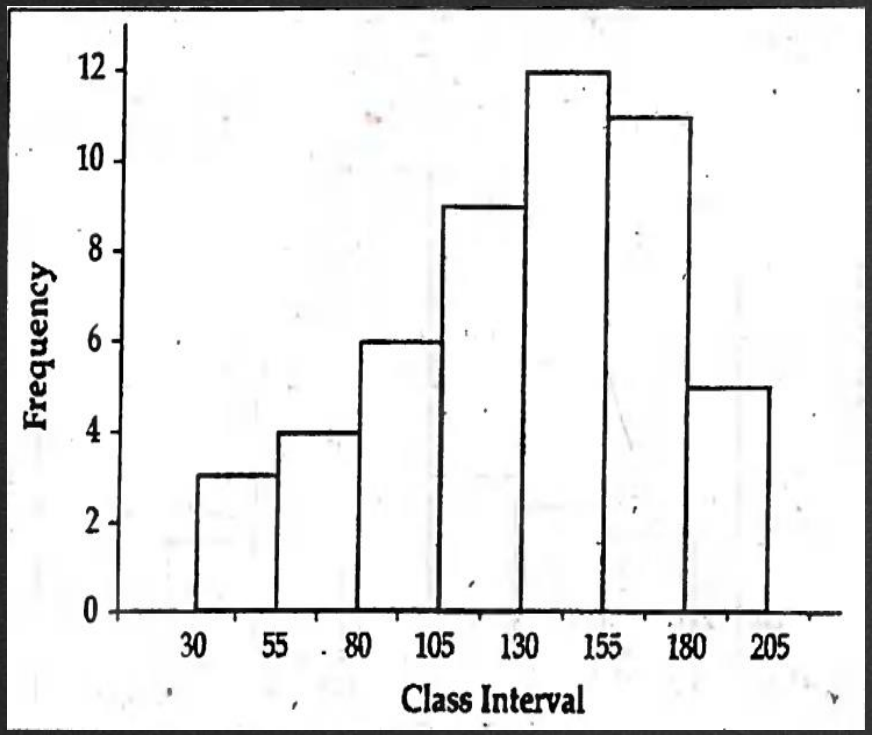
Graphical Representation

1. Dot Plot
2. Histogram
3. Frequency Polygon
4. Frequency Curve
5. Ogive Polygon
6. Ogive Curve
7. Line graph of time series data
8. Scatter Diagram

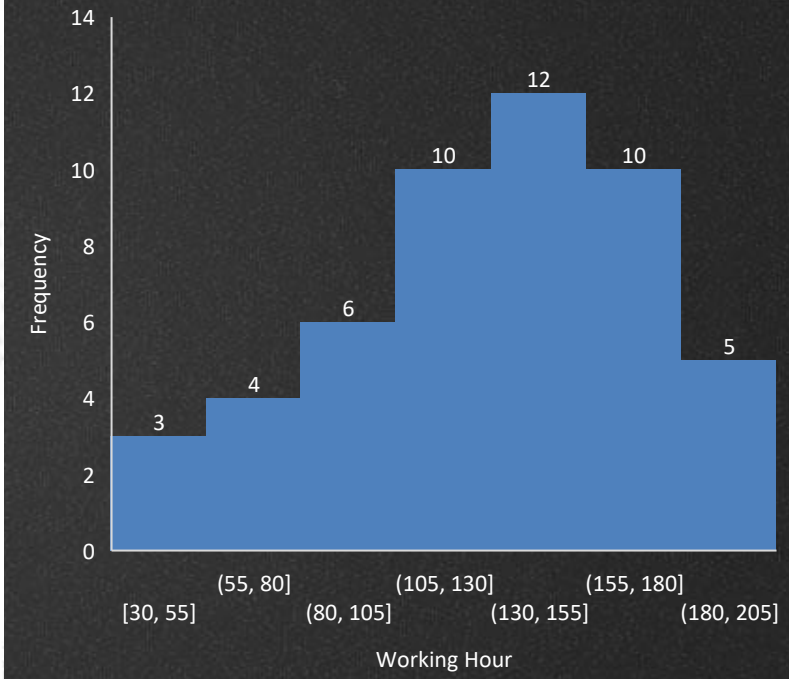
Graphical Representation

- Histogram:
  - Discrete frequency distribution
  - Continuous frequency distribution with equal class interval
  - Continuous frequency distribution with unequal class interval
- Difference between Bar diagram and Histogram

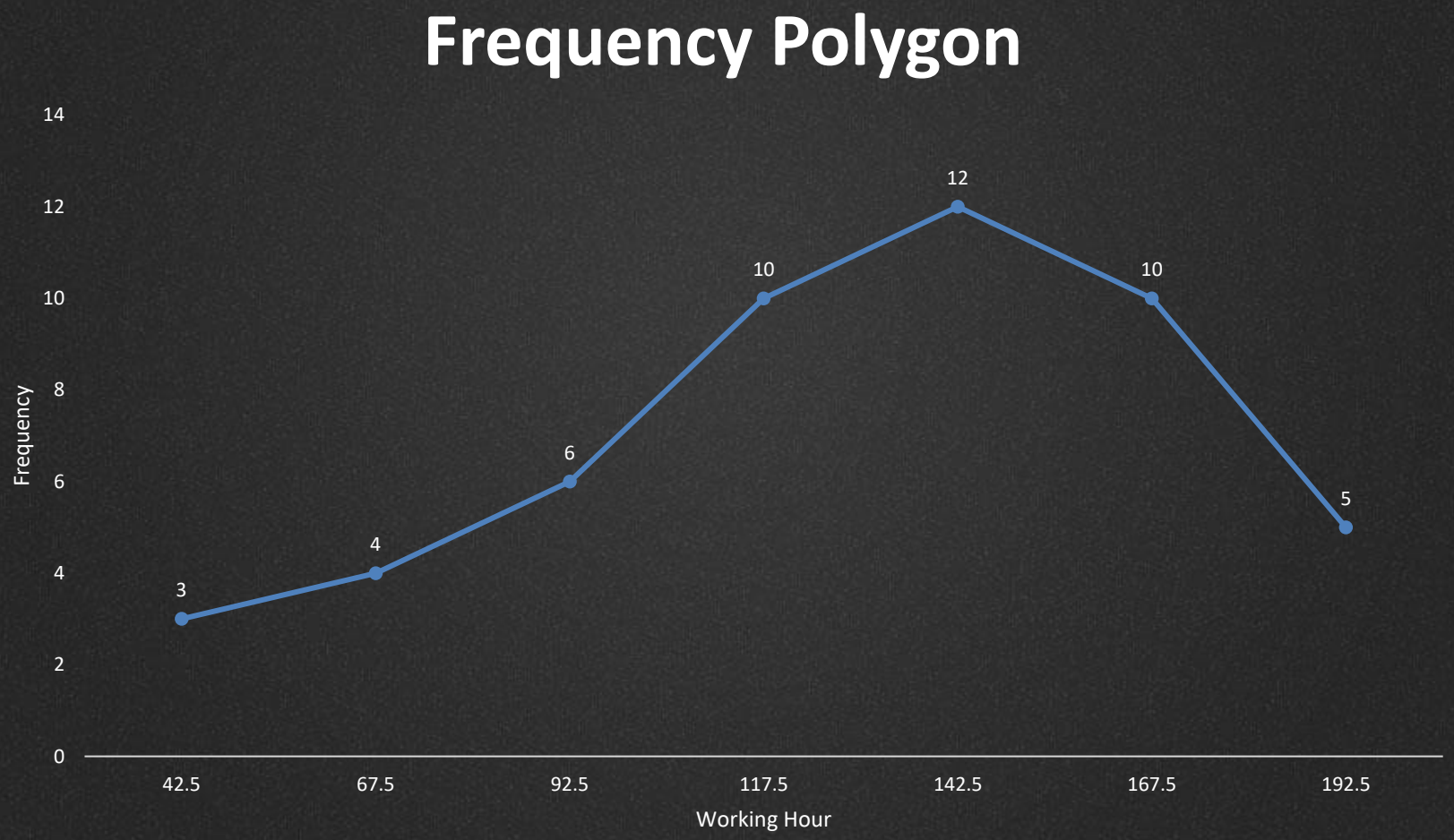
Graphical Representation



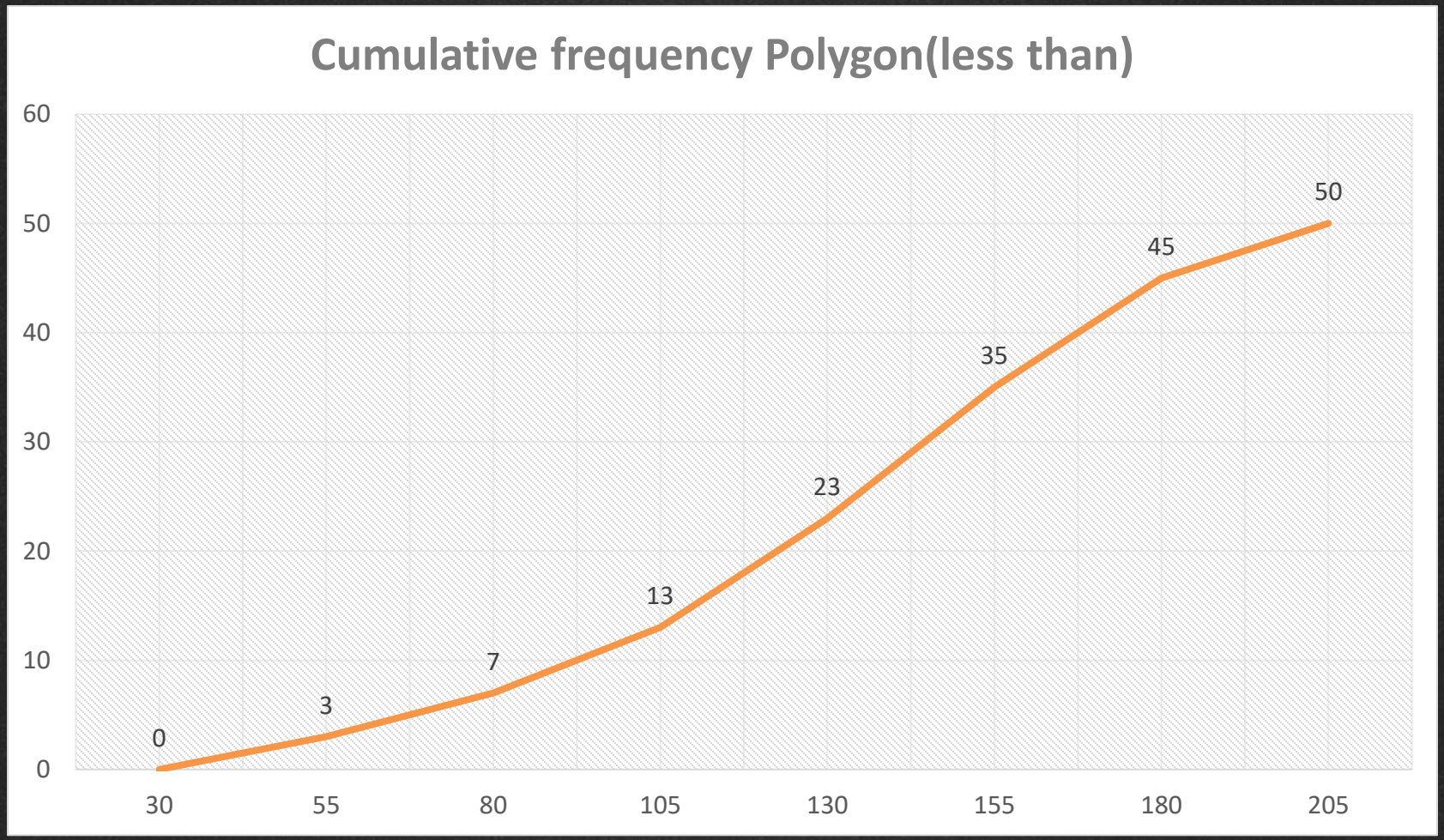
Histogram



Graphical Representation



Graphical Representation



Graphical Representation