Statistics 1 Notes

by Gen-Z IITian

About & Quick Links

Gen-Z IITian by Sriram

Topic: Statistics 1 Detailed Notes

YouTube: Gen-Z IITian

Personal YouTube: SriRam_in

Instagram: @curious_sri

Community: IITM BS Unofficial Community

PYQs Resources: IIT Pathshala Study Materials

Best Lecture: Foundation Term-1 Course

Practice Stats 1 Quiz 1 PYQs: Foundation Term-1 Course

Reference Links

- **Practice PYQs for free:** Website
- **Explore Term 1 Course** Best and Affordable
- **Unofficial Community:** WhatsApp Link

Week 1: Table of Contents

Week 1 Stats 1 by Gen-Z IITian

1. Statistics

Definition: Statistics

Statistics is the art of learning from data. It is concerned with the collection of data, their subsequent description, and their analysis, which often leads to the drawing of conclusions.

Hinglish Explanation

Statistics ka matlab hai data se seekhna. Isme hum data (jaankari) ko collect karte hain, use describe karte hain, aur analyze karke usse kuch important conclusions nikalte hain. Simple bhasha mein, yeh data ko story mein badalne ka science hai.

1.1 Population and Sample

Population

The total collection of all the elements that we are interested in is called a **population**.

Sample

A subgroup of the population that will be studied in detail is called a **sample**.

Hinglish Explanation

Socho aapko ek bade patile mein bani biryani ka namak check karna hai.

- **Population:** Poore patile ki biryani.
- **Sample:** Ek chammach biryani jo aap taste karne ke liye nikalte ho.

Aap ek chammach (sample) taste karke poore patile (population) ka andaza laga lete ho.

Example: Population vs. Sample

Suppose a survey is conducted to know the prices of all houses in Tamil Nadu and 1000 houses were randomly selected from the urban areas of Tamil Nadu for this study. It is concluded that the price of a house per square foot is roughly 5680 Rs.

- **Sample:** The selected 1000 houses from the urban areas of Tamil Nadu.
- **Population:** All houses in Tamil Nadu.

1.2 Major branches of statistics

Descriptive Statistics

The part of statistics concerned with the description and summarization of data.

- It involves using numbers or graphs to summarize the main points of data.
- A descriptive study can be done on either a sample or a full population.

Inferential Statistics

The part of statistics concerned with drawing conclusions from the data. This often means using sample data to make guesses (inferences) about the entire population.

Hinglish Explanation

Descriptive (): Ye data ki kahani batata hai, "Kya hai?". Jaise, ek class ke students ke average marks 75 hain. Ye sirf uss class ke data ko describe kar raha hai.
Inferential (): Ye sample data se population ke baare mein anumaan lagata hai, "Kya ho sakta hai?". Jaise, 1000 logon ke survey se anumaan lagana ki poore India mein kaunsi party election jeetegi.

1.3 Purpose of statistical analysis

- If the purpose of the analysis is to examine and explore information about the **collected data only**, then the study is **descriptive**.

Descriptive Study

A class of 50 students gave an exam (of 100 marks) and the average marks of the class is calculated as 65. This is descriptive statistics because we are just summarizing the data for the whole class (which is our entire group of interest here).

- If the information is obtained from a **sample** and the purpose is to use that information to draw **conclusions/inferences about the population**, the study is **inferential**.

Inferential Study

A teacher wants to know the average marks of all students in the school. Since there are too many students, the teacher takes a sample of 100 students and finds their average is 60. Using this, the teacher concludes that the average marks of *all students in the school* is likely around 60. This is inferential statistics.

2. Data

Definition: Data

Data are the facts and figures collected, analyzed, and summarized for presentation and interpretation.

- **Purpose to collect data:** Generally, we collect data when we are interested in understanding the characteristics or attributes of some group of people, places, things, or events.

Why Collect Data?

1. To know about temperatures in a particular month in Chennai, India.
2. To know about the marks obtained by students in their Class X.

2.1 Unstructured and Structured Data

Unstructured Data

Unstructured data is a dataset that is not organized in a predefined manner. It is typically text-heavy, but may contain data such as dates, numbers, and facts as well. Unstructured data requires more work to process and understand.
Examples: YouTube comments, image files, social-media posts, lyrics of a song, etc.

Structured Data

Structured data is a standardized format for providing information about a dataset. It is clearly defined and searchable. Structured data is easy to analyze and understand because it is organized.

Hinglish Explanation

Structured Data: Ek saaf-suthri Excel sheet ki tarah, jisme har cheez apne row aur column mein aaram se set ho. Jaise, students ka roll number, naam, aur marks. Ise samajhna aur analyze karna bahut aasan hai.
Unstructured Data: Ek WhatsApp group chat ki tarah, jisme text, photos, voice notes, stickers sab mix-up hai. Isme se kaam ki information nikalna mushkil hota hai.

Example 1: Structured Student Dataset

The student dataset shown in Table 1 can be considered as structured data. This data is in a tabular form and provides clear information about the Gender, Date of Birth, Marks, and Board of the students. It is easy to analyze and understand, as we can quickly find information like "Anjali scored 484 marks in the State board" or "Pradeep is Male and his date of birth is 3rd June, 2002".

Table 1: Student dataset

Name	Gender	Date of Birth	Marks	Board
Anjali	F	17 Feb, 2003	484	State Board
Pradeep	M	3 June, 2002	514	ICSE
Divya	F	22 Mar, 2003	397	State Board
Sarita	F	19 May, 2002	533	ICSE
Harsha	M	4 March, 2002	436	CBSE
Bhavana	F	7 Apr, 2003	526	State Board
Rohit	M	4 March, 2002	378	CBSE
Vikash	M	11 Oct, 2001	526	CBSE

Example 2: Structured Fertilizers Dataset

The fertilizers dataset shown in Table 2 is also structured data. It's in a tabular form and clearly provides information about different fertilizers. The data is easy to understand; for instance, we can see that Potassium is an inorganic fertilizer used for pulse crops, with a recommended amount of 320 Kg for a 1.6-acre field.

Table 2: Fertilizers dataset

Fertilizer	Type	Area (acres)	Crop	Amount (Kg)
Nitrogen	Inorganic	1	Rice	200
Phosphorus	Inorganic	2	Wheat	400
Manure	Organic	1.5	Potato	300
Compost	Organic	1.3	Rice	260
Potassium	Inorganic	1.6	Pulse	320

2.1.1 Variables and Cases

Case (or Observation)

A **case** or **observation** is a unit for which data is collected. Cases should uniquely identify each row in a dataset.

Variable

A **variable** is a characteristic or attribute that varies across all units. Intuitively, a variable is something that "varies".

Hinglish Explanation

Simple language mein:

- **Case:** Hum kiske baare mein data collect kar rahe hain? Student dataset mein, har **student** ek case hai. Table ki har **row** ek case hai.
- **Variable:** Hum uss case ke baare mein kya-kya information collect kar rahe hain? Student dataset mein, **Name, Gender, Marks** yeh sab variables hain. Table ka har **column** ek variable hai.

Cases and Variables in Student Dataset

In the student dataset (Table 1):

- The **cases** are the students: "Anjali", "Pradeep", "Divya", etc. Data is collected for each student, and their names uniquely identify each row.
- The **variables** are "Name", "Gender", "Date of Birth", and "Board", as their values keep varying from one student to another.

Important Note on Tabular Data

If we want to organise data in a tabular form, then the following two points should be taken into consideration:

- **Rows represent cases:** For each case (e.g., each student), the same set of attributes is recorded.
- **Columns represent variables:** For each variable (e.g., Marks), the same type of value is recorded for every case.

2.2 Classification of Data

Data is broadly classified into two categories: categorical data and numerical data.

2.2.1 Categorical Data and Numerical Data

Categorical Data (Qualitative)

Categorical data, also called qualitative variables, identifies group membership. We cannot perform any meaningful mathematical operations (like addition or averaging) on it.

Categorical Variables

In the student dataset (Table 1):

- **Gender** is a categorical variable with two categories: F and M.
- **Board** is a categorical variable with three categories: State Board, ICSE, and CBSE.

Numerical Data (Quantitative)

Numerical data, also called quantitative variables, describes numerical properties. We can perform mathematical operations on this data.

Numerical Variables

In the student dataset (Table 1), **Marks** is a numerical variable. We can describe its numerical properties, for instance, Rohit's marks (378) are less than Pradeep's marks (514).

Measurement Units

The scale defines the meaning of numerical data, such as weights measured in **kilograms (kg)**, prices in **rupees ()**, or heights in **centimeters (cm)**. All data points for a numerical variable must share a common unit.

2.2.2 Time-Series and Cross-Sectional Data

Time-Series Data

If data for a single unit is recorded over a period of time, it is called time-series data. A graph of a time series showing values in chronological order is known as a Time-plot.

Time-Series Data

The data collected to observe the temperature in **Delhi** for seven different days is a time-series data. The unit (Delhi) is fixed, and the measurement is taken over a period of time (seven days).

Cross-Sectional Data

If data for multiple units is observed at the same point in time, it is called cross-sectional data.

Cross-Sectional Data

The data collected to observe the temperature of **Delhi, Chennai, Jaipur, and Bhopal** on a particular day is cross-sectional data. The time is fixed (a single day), and the measurement is taken across several units (cities).

2.2.3 Scales of Measurement

We have four scales of measurement: nominal, ordinal, interval, and ratio. Data collection requires using one of these scales.

2.2.3.1 Nominal scale of measurement

Nominal Scale

When the data for a variable consists of labels or names used to identify the characteristic of an observation, the scale of measurement is considered a nominal scale.

Nominal Variables

Name, Board, Gender, Blood group, etc.

Properties of Nominal Scale

- It represents categories or labels with no natural order.
- Nominal variables can be numerically coded (e.g., Male=1, Female=2), but the numbers are just labels and have no mathematical value.

2.2.3.2 Ordinal scale of measurement

Ordinal Scale

When data exhibits the properties of nominal data and the order or rank of the data is meaningful, the scale of measurement is an ordinal scale.

Ordinal Variable

A restaurant customer provides a service rating of **Excellent, Good, or Poor**. These are labels (like nominal data), but they also have a clear order or rank.

Properties of Ordinal Scale

- It represents categories with a meaningful order.
- While there is a rank (e.g., Excellent ¿ Good ¿ Poor), the difference between the ranks is not uniform or measurable. The gap between "Good" and "Excellent" isn't necessarily the same as the gap between "Poor" and "Good".

2.2.3.3 Interval scale of measurement

Interval Scale

If the data have all the properties of ordinal data and the interval between values is expressed in terms of a fixed unit of measure, the scale of measurement is an interval scale.

Interval Variable

Temperature in degrees Celsius or Fahrenheit is an interval variable. We know the difference between 20°C and 30°C is the same as the difference between 30°C and 40°C (both are 10°C).

Properties of Interval Scale

- It has order and a constant, meaningful difference between values.
- Ratios are meaningless because there is **no true zero**. 0°C does not mean "no temperature"; it's just another point on the scale (the freezing point of water). For this reason, we cannot say 40°C is "twice as hot" as 20°C.

2.2.3.4 Ratio scale of measurement

Ratio Scale

If the data have all the properties of interval data and the ratio of two values is meaningful, the scale of measurement is a ratio scale. This is possible because the ratio scale has a true, absolute zero.

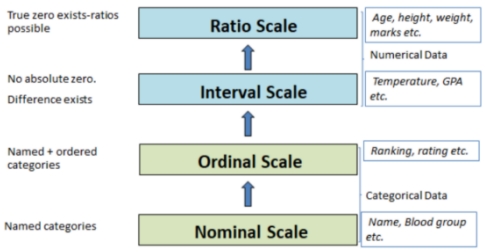

Figure 1: Enter Caption

Ratio Variables

Height (in cm), Weight (in kg), and Marks. All these have an absolute zero. A height of 0 cm means "no height". Because of this, ratios are meaningful: a person who is 180 cm tall is twice as tall as a person who is 90 cm tall.

Scales of Measurement: Quick Summary

- **Nominal (Naam):** Sirf naam ya label. Koi order nahi. *Ex: Jersey numbers (10, 7, 18).*
- **Ordinal (Order):** Naam bhi aur order bhi. Par difference ka pata nahi. *Ex: Race ranks (1st, 2nd, 3rd).*
- **Interval (Difference):** Order bhi aur equal difference bhi. Par "true zero" nahi hota. *Ex: Temperature in Celsius.*
- **Ratio (Sab Kuch):** Order, equal difference, aur "true zero" bhi. Isme ratio nikal sakte hain. *Ex: Marks in an exam (80 marks is double of 40 marks).*

A summary about all scales of measurement can be described as follows:

Unsolved Problems

Question 1

An analyst wants to conduct a survey for testing the maintenance of hospitals in a particular district in Bihar, for which he selects 25 hospitals randomly from that district. Identify the sample and population. [2 Marks]

Options

(a) The population is all the hospitals in Bihar and the sample is all the hospitals in the district.
(b) The population is all the hospitals in Bihar and the sample is 25 selected hospitals in Bihar.
(c) The population is all hospitals in the district of Bihar and the sample is 25 selected hospitals in the district.
(d) None of the above.

Answer

The correct option is **(c)**. The study is focused on a *particular district*, so that district's hospitals are the population. The 25 selected hospitals form the sample.

Tutor Calc Space

Question 2

In the 2011 Cricket ODI World Cup quarter-final match between India and Australia, a media organization estimated that Australia would beat India by 50 runs if Australia bats first, based on the information of matches played between the two teams previously. Which branch of statistics does the above analysis belong to?

Answer

Inferential Statistics. The organization is using past data (a sample) to make a prediction or inference about a future event.

Tutor Calc Space

Question 3

Values of temperature and humidity of a room are measured for 24 hours at a regular time interval of 30 minutes. Based on this information, choose the correct option:

Options

(a) It is a cross-sectional data.
(b) It is time-series data.

Answer

The correct option is **(b)**. The data is recorded for one unit (a room) over a period of time (24 hours).

Tutor Calc Space

Question 4

What kind of data is "Social media posts"?

Options

(a) Unstructured data
(b) Structured data

Answer

The correct option is **(a)**. Social media posts contain a mix of text, images, videos, and numbers with no predefined format.

Tutor Calc Space

Question 5

What kind of variable is the qualification of a candidate sitting for a job interview?

Options

(a) Numerical/ Quantitative
(b) Categorical/ Qualitative
(c) Numerical and discrete
(d) Numerical and continuous

Answer

The correct option is **(b)**. Qualifications (e.g., "B.Tech", "M.Sc.", "Ph.D.") are labels or categories.

Tutor Calc Space

Question 6

If addition and subtraction can be performed on a variable, then the scale(s) of measurement of the variable could be:

Options

(a) Ordinal
(b) Ratio
(c) Interval
(d) Nominal

Answer

The correct options are **(b) and (c)**. Both **Interval** and **Ratio** scales support addition and subtraction because the difference between values is meaningful and constant.

Tutor Calc Space

Question 7

Which of the following variable(s) have nominal scale of measurement?

Options

(a) Education qualification of a person.
(b) Hair color
(c) Brand name of mobile phone
(d) Number plate of cars

Answer

The correct options are **(b), (c), and (d)**. Hair color, brand names, and number plates are just labels with no inherent order. Education qualification (a) is Ordinal.

Tutor Calc Space