manipalglobal
Academy of Data Science

Statistical Techniques for Data Science

Introduction to Statistics

Collection, Categorization and Presentation of Data

Dr. Subhabaha Pal
Manipal Global Academy of Data Science

Objective

manipalglobal
Academy of Data Science

After attending this session, you will be able to –

- **Delineate different data collection techniques**
- **Perform meaningful classification of large mass of data and interpret the same**
- **Present a set of observations using a frequency table and histogram**

Data Collection

manipalglobal
Academy of Data Science

- **The first activity that is needed to be performed before undertaking any statistical analysis project is collecting relevant data/information**
- **Information can be gathered from varied range of sources**
- **Data can be sub-grouped into 2 types based on data collection source – Primary Data and Secondary Data**

Areas of Difference	Primary Data	Secondary Data
Meaning	Data collected by researcher himself	Data collected by other person
Originality	Original or unique information	Not original or unique information
Adjustment	Does not need adjustment, is focused	Needs adjustment to suit actual aim
Sources	Surveys, observations, experiments	Internal records, Govt. published data etc.

Data Collection Techniques

manipalglobal
Academy of Data Science

- **Various techniques can be used for collecting data**

Techniques	Key Facts	Example
Interviews	Interviews can be conducted in person or over the telephone. Interviews can be done formally (structured), semi-structured, or informally. Questions should be focused, clear, and encourage open-ended responses. Mostly the data gathered is qualitative.	Interviewing the farmers directly to know about the crops produced in the current season in a specific area
Questionnaires and Surveys	Responses can be analyzed with quantitative methods by assigning numerical values to Likert-type scales. Results are generally easier to analyse.	Results of a satisfaction survey or opinion survey
Observations	Allows for study of dynamics of a situation with frequency count of the target behaviour.	Going to Real Estate Project site to check how far the work is completed and at what pace the work is proceeding
Focus Groups	A facilitated group interview with individuals that have something in common. Gathers information about combined perspectives and opinions.	A group of software engineers with 1 year of work experience are invited in a group meeting to understand their future career needs and how they want further
Ethnographies, Oral History and Case Studies	Most holistic approach to evaluation involving studying a single phenomenon and examining people in their natural settings and using combination of techniques such as observation, interviews and surveys	Split-Cable Single Source Experiment
Documents and Records	Consists of examining existing data in the form of databases, meeting minutes, reports, attendance logs, financial records, newsletters, etc.	Going through the attendance records of the students in a school of a locality to understand the attendance rate of students during a particular time-period

Data Variables

manipalglobal
Academy of Data Science

- Constant – A characteristic which remains same everywhere is termed as constant. Example – Value of π (Pi)
- Local and Global Constant
- Variable – A characteristic which takes on different values in different person, places and things. Example – Diastolic Blood Pressure, heart rate and height of males.
- Quantitative Variable – One that can be measured and expressed numerically. Example – Heart rate, Blood Pressure
- Qualitative Variable – The characteristics that can not be measured quantitatively but can be categorized. Measurement convey information regarding the attribute. Example – Gender of a patient.
- Random Variable – Values obtained as a result of chance event/factor, so that can not be exactly predicted in advance. Example – Weights of a group of randomly selected infants.
- Discrete Random Variable – Characterized by gaps or interrupts in the values that it can assume. It assumes values with definite jump. It can not take all possible values within a range. Example – Number of daily admission to a general hospital.
- Continuous Random Variable – It can take all possible values positive, negative, integral and fractional values within a specified relevant interval. Example – Height of a person.

More about Data Variables

manipalglobal
Academy of Data Science

- **Qualitative Data can be classified as –**
  - **Nominal Data – Data has name only. Example – Gender, Colour of a car.**
  - **Ordinal Data – Data has order but it is not numeric. Example – Very Good, Good, Neutral, Bad, Very Bad.**

Data Categorization and Classification

manipalglobal
Academy of Data Science

- In order to get insight from the collected data, it is first needed to be organized
- Basic insight about the data can be obtained through listing of values in an ordered array – Helps in quick determination of smallest and largest values

Example – Heights of 10 person

160 cm, 155 cm, 170 cm, 175 cm, 168 cm, 190 cm, 174 cm, 180 cm, 177 cm, 182 cm, 183 cm.

Ordered Array – 155 cm, 168 cm, 170 cm, 174 cm, 175 cm, 177 cm, 180 cm, 182 cm, 183 cm, 190 cm.

- We can tell from the above data that 155 cm is the minimum height and 190 cm is the maximum height.
- Data Classification – It is the grouping of related facts/data into different classes according to certain characteristics. Helps in condensing mass of data such that similarities and dissimilarities can be readily distinguished. It facilitates comparison.
- Basic Data classifications are as follows – 1. Geographical, 2. Chronological or Temporal, 3. Qualitative and 4. Quantitative.

Data Classification

manipalglobal
Academy of Data Science

- **Geographical Classification – Classification based on Geographical Location. Example – Classify Students of an Institute based on the states they belong to.**

Student Name	State
Amit	Uttar Pradesh
Samit	Himachal Pradesh
Arnab	West Bengal
Purushottam	Karnataka
Haricharan	Karnataka
Amrita	Uttar Pradesh
Suresh	Himachal Pradesh

State	Student Name
Uttar Pradesh	Amit
	Amrita
Himachal Pradesh	Samit
	Suresh
Karnataka	Purushottam
	Haricharan
Arnab	West Bengal

Data Classification

manipalglobal
Academy of Data Science

- **Chronological or Temporal Classification – Classification is based on Time. Example – Babies born in a hospital in current year and last year.**

Baby Name	Birth Date
Sreyash	21.01.2016
Dolly	20.02.2017
Sarika	15.04.2016
Hari	18.06.2017
Asmit	14.05.2016
Amrita	16.07.2016
Sarbesh	10.07.2016

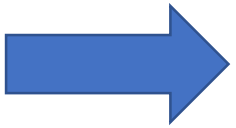
Birth Year	Student Name
2017	Dolly
	Hari
2016	Sreyash
	Sarika
	Asmit
	Amrita
	Sarbesh

Data Classification

manipalglobal
Academy of Data Science

- **Qualitative Classification – Classification based on some attributes. Example – Classify few individuals based on area, gender and literacy.**

Name	Area	Gender	Education
Amit	Urban	Male	Literate
Samit	Rural	Male	Literate
Arnab	Urban	Male	Literate
Purushottam	Urban	Male	Illiterate
Haricharan	Rural	Male	Illiterate
Amrita	Urban	Female	Literate
Suresh	Rural	Male	Illiterate
Sreyash	Urban	Male	Literate
Dolly	Urban	Female	Illiterate
Sarika	Rural	Female	Illiterate
Hari	Rural	Male	Illiterate
Asmit	Urban	Male	Literate
Amrita	Rural	Female	Illiterate
Sarbesh	Urban	Male	Literate
Pampa	Rural	Female	Literate



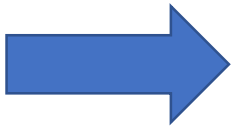
Urban				Rural			
Male		Female		Male		Female	
Literate	Illiterate	Literate	Illiterate	Literate	Illiterate	Literate	Illiterate
Amit	Purushottam	Amrita	Dolly	Samit	Haricharan	Pampa	Sarika
Arnab					Suresh		Amrita
Sreyash					Hari		
Asmit							
Sarbesh							

Data Classification

manipalglobal
Academy of Data Science

- **Quantitative Classification – On the basis of Quantitative Class Intervals. Example – Classify 10 individuals based on their Annual Income.**

Name	Annual Income (INR)
Amit	150000
Samit	250000
Arnab	200000
Purushottam	300000
Haricharan	350000
Amrita	700000
Suresh	550000
Sreyash	400000
Dolly	800000
Sarika	600000



Income Range (INR)	Names
< 240000	Amit, Arnab
240000 - 500000	Samit, Purushottam, Haricharan, Sreyash
> 500000	Suresh, Dolly, Sarika

Presentation of Data

manipalglobal
Academy of Data Science

- A Frequency Distribution is a grouping of data into mutually exclusive categories showing the number of observations in each class
- Constructing a frequency distribution involves –
  - Determining the question to be addressed
  - Collecting raw data
  - Organizing data (frequency distribution)
  - Presenting data (Histogram)
- Example – As Marketing Manager of your company, you are looking for prospective clients for your product which is electricity-driven car. You want to target a particular section of IT employees in certain location of Bengaluru. From your past experience, you found that people who use to travel between 3 KM to 6 KM everyday for office were more interested to buy such car. As reaching each and every employee in the IT park may incur huge cost, you decided to do a pilot survey to get some idea about the prospective market of your product in the IT park. You engaged an executive who were supposed to ask every employee coming to office at morning about how much they need to travel to come to office.

Presentation of Data

manipalglobal
Academy of Data Science

- **The executive could collect data from 100 employees (Distance Travel in KM)**

0.3	13.9	7.1	0.5	1.8	8.1	18.3	6.2	15.9	17.4
16.2	8.6	9.2	20.9	15.1	18.0	2.0	6.5	6.1	19.2
3.5	14.1	17.6	20.9	20.9	9.3	17.8	4.7	12.5	11.6
8.7	18.2	14.8	18.2	0.3	13.3	9.1	17.0	6.0	10.4
10.2	10.5	11.7	15.8	1.3	20.4	20.5	1.1	9.7	2.8
5.9	14.8	5.5	16.7	9.4	16.4	15.5	16.1	14.6	8.4
7.5	8.0	4.5	18.6	1.9	14.2	17.8	9.3	9.8	8.1
15.4	14.8	4.7	12.5	3.5	15.8	9.6	2.8	5.5	3.8
19.4	0.3	8.7	10.1	9.8	6.1	1.3	7.5	14.8	0.7
3.1	13.5	3.4	3.1	9.4	2.2	20.1	16.6	10.0	0.9

- **How the above data can be interpreted and presented for further analysis?**

Presentation of Data

manipalglobal
Academy of Data Science

- **First you need to create a Frequency Distribution of the same which will give some insight on the data**
- **In order create a Frequency distribution, you first need to import the data in R through following code –**

```
freq_data <- read.csv("freq_dist_data.csv") #Importing the Data
```

- **Find the range of the data (minimum value and maximum value)**

```
> range(freq_data$Distance)
[1] 0.3 20.9
```

- **As lowest is 0.3 and highest is 20.9, you can take lowest class limit of the first interval as 0 and highest class limit of the last interval as 21**
- **If you break the total range into 7 class intervals, you will get the class 3KM – 6KM also for which you have special interest**

Presentation of Data

manipalglobal
Academy of Data Science

- **Creating frequency distribution with 7 intervals in the range 0 to 21**

```
bins = seq(0, 21, by = 3)
> intervals = cut(freq_data$Distance, bins)
> transform(table(intervals))
intervals Freq
1 (0,3] 15
2 (3,6] 13
3 (6,9] 14
4 (9,12] 17
5 (12,15] 12
6 (15,18] 17
7 (18,21] 12
```

Intervals	Frequency
0-3	15
3-6	13
6-9	14
9-12	17
12-15	12
15-18	17
18-21	12

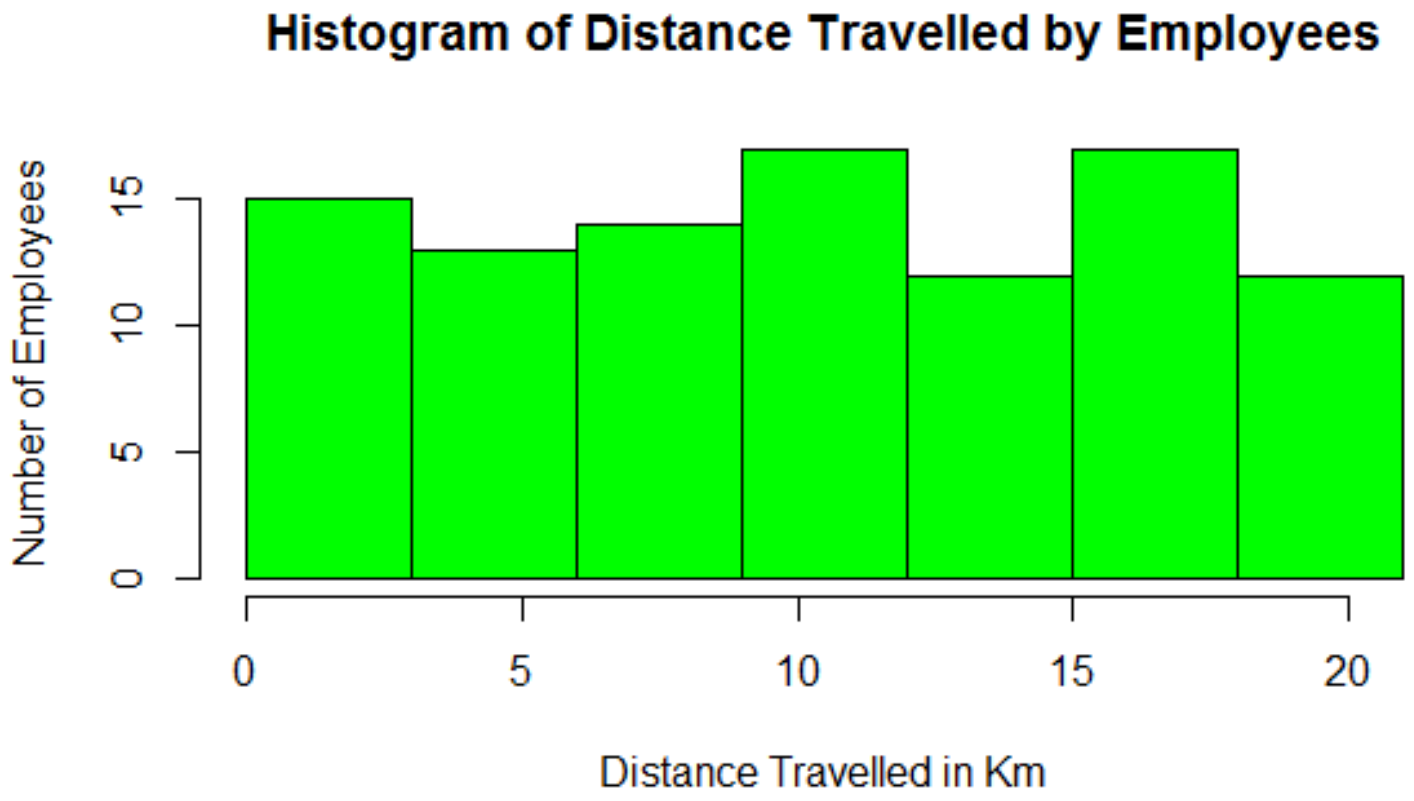
- **From the frequency table, you can observe that 13 employees out of total sample of 100 employees travel 3KM – 6KM everyday to come to office**

Presentation of Data

manipalglobal
Academy of Data Science

- **For better visibility of the result, you can present it through a histogram**

```
hist(freq_data$Distance, breaks = bins, main = "Histogram of Distance Travelled by Employees",
col = "green", xlab = "Distance Travelled in Km", ylab = "Number of Employees")
```



Presentation of Data

manipalglobal
Academy of Data Science

- **You can create relative frequency and cumulative frequency from the data**

```
> transform(table(intervals), Relative = prop.table(Freq), Cumulative = cumsum(Freq))
intervals Freq Relative Cumulative
1 (0,3] 15 0.15 15
2 (3,6] 13 0.13 28
3 (6,9] 14 0.14 42
4 (9,12] 17 0.17 59
5 (12,15] 12 0.12 71
6 (15,18] 17 0.17 88
7 (18,21] 12 0.12 100
```

manipalglobal
Academy of Data Science