INTRODUCTION

- by Aindrila Garai, IITKANPUR

1. What is Statistics?

A burning subject that accounts for collecting, presenting, analysing data & extracting certain information or interpreting regarding some phenomenon.

2. Classification of Data:

2.1 Based On Method Of Collection-

- **Primary Data:** This more reliable data is acquired directly from the field of study. eg. *Census*.
- **Secondary Data:** This is collected from indirect sources and needs cross validation. eg. *any data from newspaper or book*.

2.2 Based On Characteristics-

- **Variable or Quantitative Data:** It is presented in terms of numerical numbers and it has two types-

1. Discrete Data: It takes some isolated value like *Class size, number of countries etc.*

2. Continuous Data: It takes values within its range of variation such as *Weight, income etc.*

- **Atribute or Quanlitative/Categorical Data:** The character observed is not measurable in numerical terms and it is of three types-

1. Ordinal Data: Categories of data can be ordered in form of superiority or inferiority. eg. Grades, Finacial Status, Drug dose etc.

2. Non Ordinal Data: Categories cannot be ordered in heirarchy. eg. Martial Status, Blood Group etc.

3. Nominal Data: Those characteristics can be expressible as numerical figures but not fit for arithmatic operations. eg. *phone number, PIN code etc.*

2.3 Based On Time & Space-

These are non-frequency type data.

- **Cross Sectional Data/ Spatial Data:** This type of data varies across different geographical regions. eg. Annual temperature in 2023 in different states.
- **Time Series Data/ Historical Data/ Chronological Data:** Time series data varies across different time points. eg. Annual temperature in India from 2001-2023.

3. Scale:

There are different types of scaling for numerical data.

3.1 Interval Scale: Here, different levels are indexed by maintaining ordering of the numerical values. Linear transformation does not hamper the measurement and measure "zero" does not imply absence of the character. eg. Centigrade and Fahrenheit scale of temperature $\frac{C}{5} = \frac{F-32}{9}$.

3.2 Ratio Scale: Same as Interval Scale but "zero" implies absence of the character. eg. *Weight*.

3.3 Nominal Scale: When numbers or other symbols are used to identify the group to which various objects belong to these numbers or other symbols constitute a Nominal Scale. eg. *the numbers on automobile license plates*

3.4 Ordinal Scale: It may happen that objects in one category of scale are not just different from the objects in other categories of that scale but that they stand in same kind of relationto them. eg. *The no of class 8 students are greater than the no of class 7 students.*

4. Collection of Data:

In any statistical activity, the primary task is to collect data using some method.

4.1 Questionnaire: A questionnaire means a systematically arranged series of relevant questions.

4.2 Schedule: A list of items on which information will be collected.

4.3 Interview Method: An interviewer gathers the data directly from the field of enquiry.

4.4 Mailed Questionnaire Method: Questionnaires are sent to respondents for answering and post it back.

5. Scrutiny of Data:

Collection of data, it is important to cross validate each and every response in order to make a fair and proper analysis. If any data seems inconsistent, preprocessing or **data cleansing** should be conducted before final analysis. eg. *a govt employee claims that his monthly income is Rs.1500 - so, here may be a chance of a lack of zero/s from the right.*

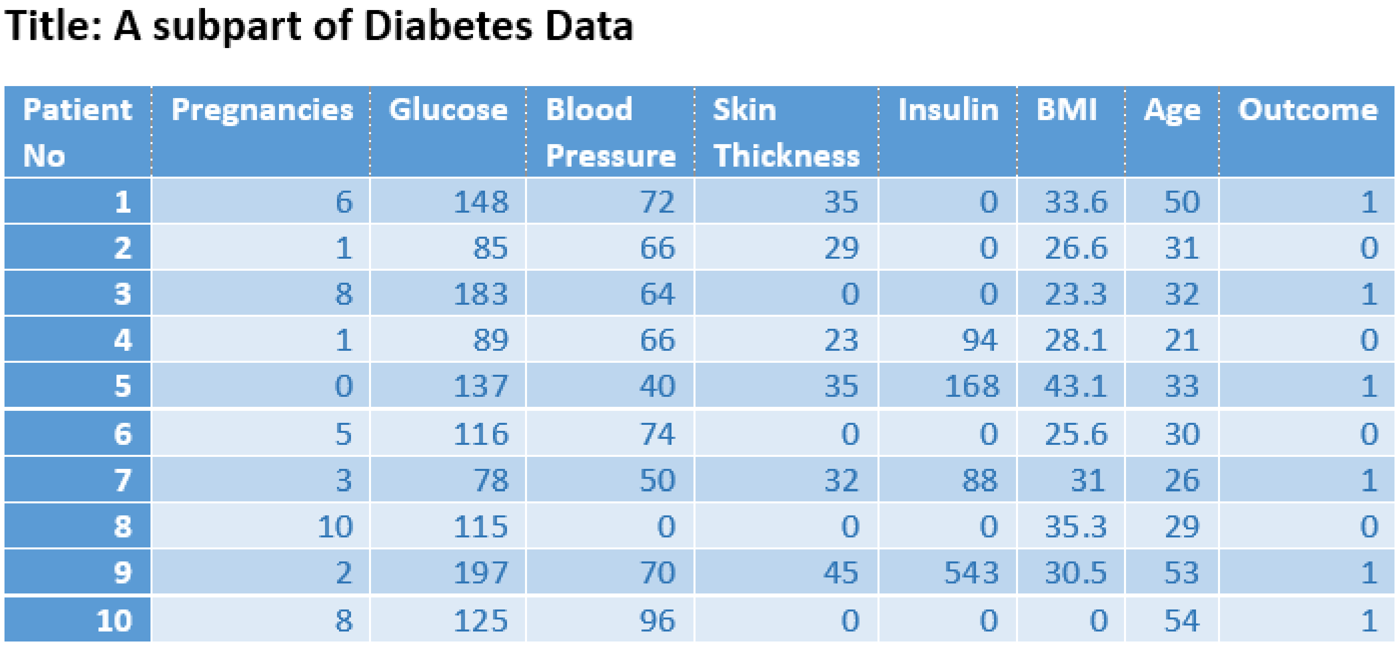
5. Representation of Data:

5.1 Textual Representation: Data is represented through plain text. It is too ineffective.

5.2 Tabular Representation: A well defined table should have following features-

- Title- A brief description of contents of the table.
- Stub- The extreme left part describing the rows.
- Caption- The upper part describing the columns.
- Body- Exhibiting all relevant information.
- Footnote- Source of the data.

Table No: 1



Title: A subpart of Diabetes Data

Patient No	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Age	Outcome
1	6	148	72	35	0	33.6	50	1
2	1	85	66	29	0	26.6	31	0
3	8	183	64	0	0	23.3	32	1
4	1	89	66	23	94	28.1	21	0
5	0	137	40	35	168	43.1	33	1
6	5	116	74	0	0	25.6	30	0
7	3	78	50	32	88	31	26	1
8	10	115	0	0	0	35.3	29	0
9	2	197	70	45	543	30.5	53	1
10	8	125	96	0	0	0	54	1

Source: Diabetes Data

5.3 Graphical Representation: We have discussed this part in Data Visualisation.

6. Frequency Distribution:

The features to construct a frequency table are-

i) **Range**(r) = difference between the maximum and minimum of dataset.

```
range <- max(data) - min(data) #range.default()
```

ii) **Class Limit** = Some intervals where the whole range of variables is divided into some groups.

```
class_limit <- function(data,class_width)
{
  upper_limit <- NULL
  lower_limit <- first_lowest_class_limit
  new_upper_limit <- 0
  new_lower_limit <- 0
  while(max(data)>new_upper_limit)
  {
    new_upper_limit <- max(lower_limit)+class_width
    upper_limit <- c(upper_limit, new_upper_limit)

    new_lower_limit <- ceiling(new_upper_limit)
    lower_limit <- c(lower_limit, new_lower_limit)

  }
  lower_limit <- lower_limit[-(which.max(lower_limit))]
  rtn <- data.frame(lower_limit, upper_limit )
  rtn
}
class_limit <- class_limit(data,0.9)
```

iii) **Class Width**(w) = Upper class limit — Lower class limit.

```
# one of the class width or number of classes should be given
class_width <- 0.9 #given
```

iii) **Number of classes**(c) = $\frac{r}{w}$, converted to nearest integer.

```
C <- round((range)/class_width) # check round function in numerical analysis
# you may try with fixed classes
```

iv) **Class boundaries** = For a continuous variate, class limits are smoothen to class boundaries in the following mathod.

```
class_boundary <- function(class_limit)
{
  lim <- class_limit
  gap <- (lim$lower_limit[2] - lim$upper_limit[1]) /2
  lower_boundary <- lim$lower_limit - gap
  upper_boundary <- lim$upper_limit + gap
  rtn <- data.frame(lower_boundary, upper_boundary )
  return(rtn)
}
class_boundary <- class_boundary(class_limit)
```

Class Mark = The mid value of the class limits or boundaries.

```
class_mark <- function(class_limit)
{
  len <- dim(class_limit)[1]
  classMark <- numeric(len)
  for(i in 1:len)
  {
    classMark[i] <- sum(class_limit[i,])/2 # try to use apply()
  }
  data.frame(classMark)
}
class_mark <- class_mark(class_limit)
```

v) **Frequency** = Number of data in each class

```
freq <- function(data,class_limit)
{
  lim <- class_limit
  len <- dim(lim)[1]
  fre <- numeric(len)
  for(i in 1:len)
  {
    vec <- numeric(length = length(data))
    for(j in 1:length(data))
    {
```

```
    vec[j] <- (data[j] >= lim$lower_limit[i] && data[j] <= lim$upper_limit[i])

    fre[i] <- sum(vec)
    }
  }
  return(data.frame(fre))
}
freq <- freq(data,class_limit)
```

Cumulative Frequency:

cumulative more than type of a particular class = total number of data points + frequency of the class — cumulative less than type of the class. (*Hint: sum- CF more + CF less*)

```
cumulative <- function(freq, type)
{
  len <- dim(freq)[1]
  vec <- numeric(len)
  if(type=="increasing")
  {
  for (i in 1:len)
  {
    vec[i] <- sum(freq[1:i,])
  }
  return(data.frame(vec))
  }
  if(type=="decreasing")
  {
   for (i in len:1)
  {
    vec[len-i+1] <- sum(freq[len-i+1:i,])
  }
  return(data.frame(vec))
  }
}
```

```
cum_in <- cumulative(freq,"increasing")
cum_de <- cumulative(freq,"decreasing")
```

vi) **Relative frequency** = $\frac{\text{Frequency of the class}}{\text{Total Frequency}}$

```
relativeFreq <- function(freq)
{
  len = dim(freq)[1]
  vec <- numeric(len)
  for(i in 1:len)
  {
    vec[i] <- freq[i,]/sum(freq[,1])
  }
  return(data.frame(vec))
}
rel_freq <- relativeFreq(freq)
sum(rel_freq) # sum of relative frequencies should be 1
```

```
     1
```

vii) **Frequency Density** = $\frac{\text{Frequency of the class}}{\text{Class Width}}$

```
Freqden <- function(freq)
{
  len = dim(freq)[1]
  vec <- numeric(len)
  for(i in 1:len)
  {
    vec[i] <- round(freq[i,]/class_width, 2)
  }
  return(data.frame(vec))
}
freqden <- Freqden(freq)
```

```
# Let's make a frequency table
data <- c(11.1, 13.0, 9.9, 13.3, 14.5, 11.0, 10.8,
10.0, 10.6, 11.4, 12.5, 12.6, 15.2, 13.7,
11.3, 14.2, 12.4, 12.4, 12.3, 13.9, 13.1,
13.7, 10.9, 13.4, 12.5, 14.5, 13.4, 15.9,
13.5, 13.6, 12.5, 10.2, 9.4, 12.0, 11.1,
14.3, 12.6, 13.6, 12.6, 11.4, 13.9, 13.9,
11.4, 11.9, 15.7, 12.6, 14.4, 11.6, 14.5,
15.3)

class_width <- 0.9 #given
first_lowest_class_limit <- 9 #(by default=floor(min(data)))

freq_table <- data.frame(c(class_limit, class_boundary, class_mark, freq, cum_in, cum_de, rel_freq, freqden))
freq_table
```

A data.frame: 7 × 10

lower_limit	upper_limit	lower_boundary	upper_boundary	classMark	fre	vec	vec.1	vec.2	vec.3
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
9	9.9	8.95	9.95	9.45	2	2	50	0.04	2.22
10	10.9	9.95	10.95	10.45	5	7	48	0.10	5.56
11	11.9	10.95	11.95	11.45	9	16	43	0.18	10.00
12	12.9	11.95	12.95	12.45	11	27	34	0.22	12.22
13	13.9	12.95	13.95	13.45	13	40	23	0.26	14.44
14	14.9	13.95	14.95	14.45	6	46	10	0.12	6.67
15	15.9	14.95	15.95	15.45	4	50	4	0.08	4.44