Introduction to Statistics

Open Data Lab

GCK

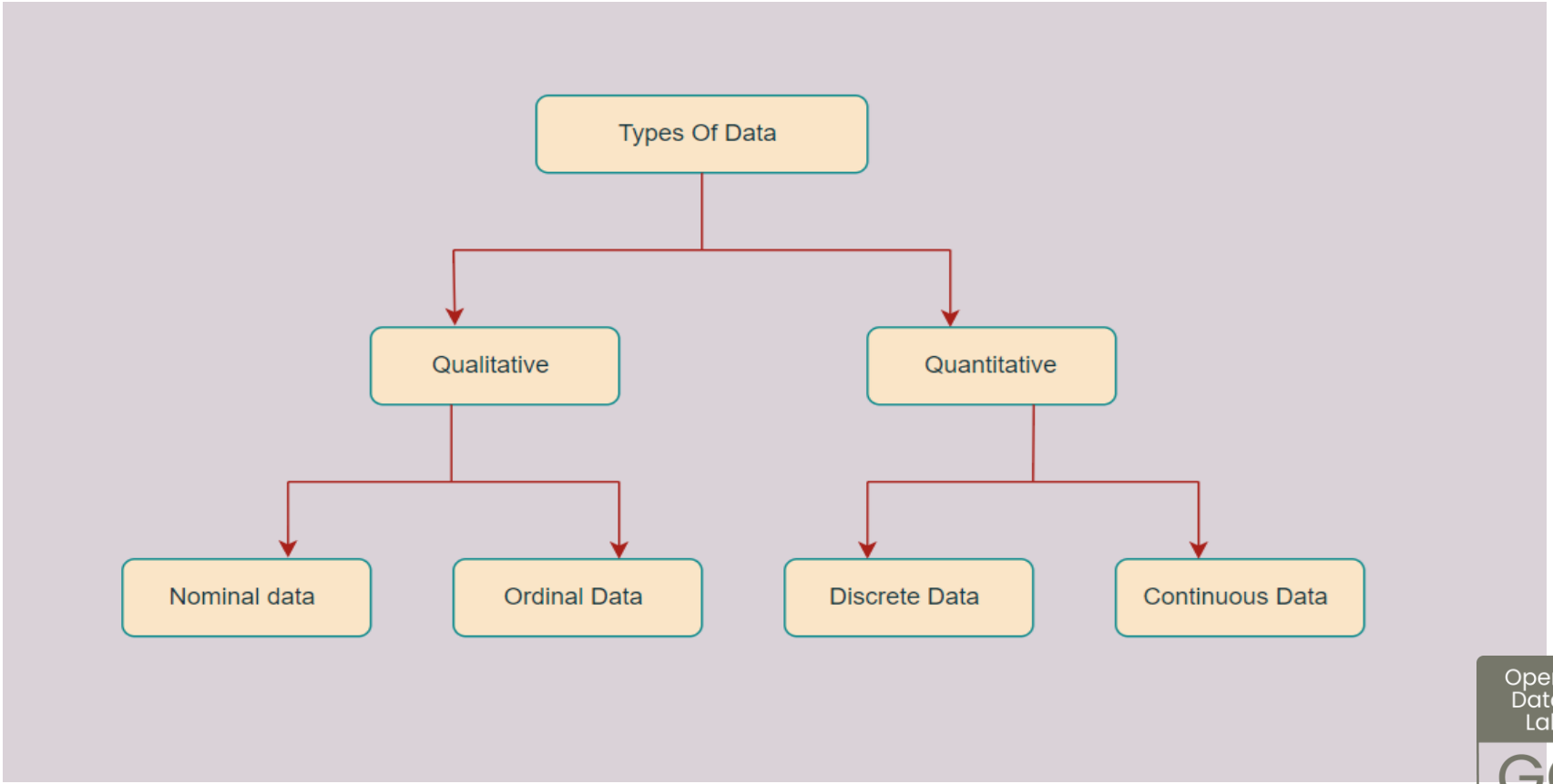
Why statistics in data analysis?

- To understand dataset better
- To perform better data manipulation

Some statistical concepts for data analysis

- Types of Data
- Population and Sample
- Descriptive Statistics
- Correlation and Regression

Types of Data

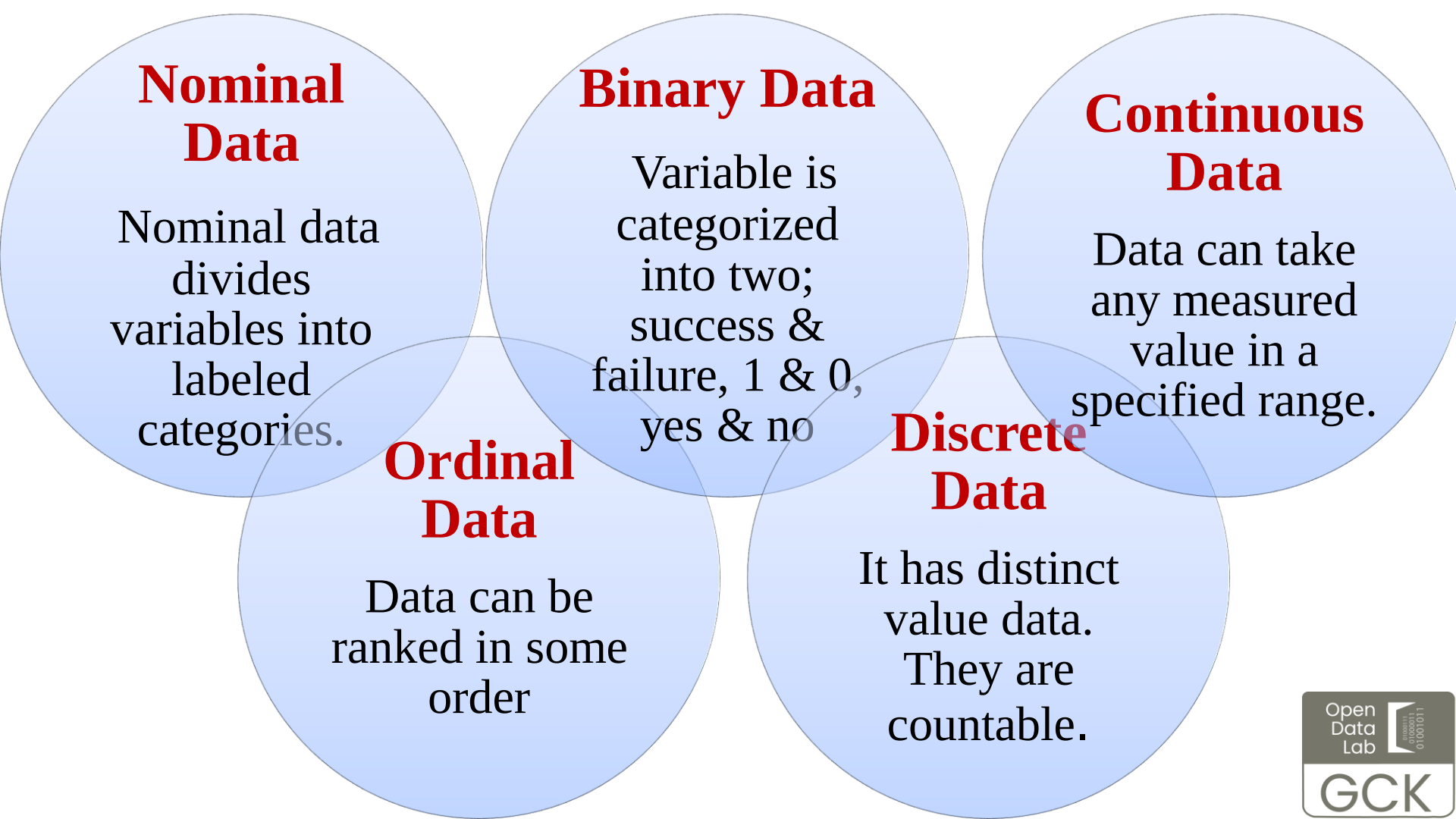


Qualitative Data

- Cannot be measured
- Categorical Data
- Eg: sex, marital status

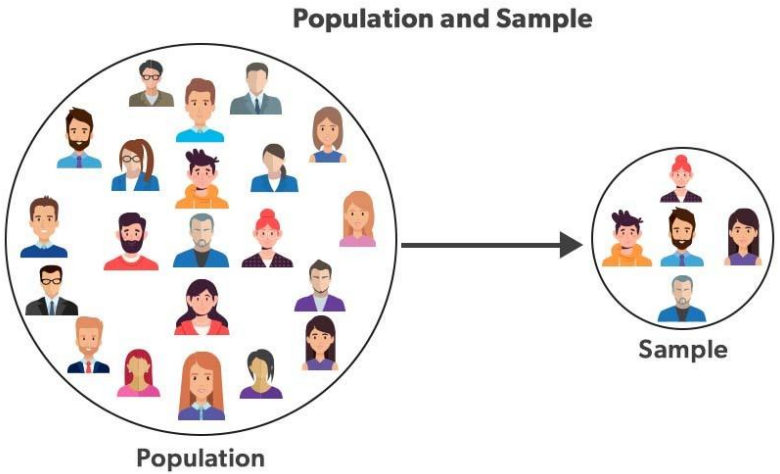
Quantitative Data

- Can be measured
- Numerical Data
- Eg: temperature, Age


Nominal Data
Nominal data divides variables into labeled categories.
Binary Data
Variable is categorized into two; success & failure, 1 & 0, yes & no
Continuous Data
Data can take any measured value in a specified range.
Ordinal Data
Data can be ranked in some order
Discrete Data
It has distinct value data. They are countable.

Population

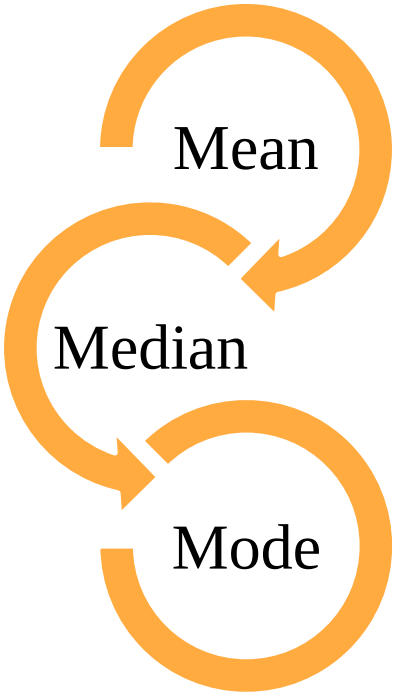
- All elements of a group
- Eg: All teachers in Kerala



Sample

- Subset of population
- True representation of population
- Eg: 1000 teachers in kerala.

Measures of Central Tendency



Mean

Average of the data

Ratio of sum of all observation to number of observation in the dataset.

Arithmetic Mean

$$\text{Mean}(X) = \sum_{i=1}^{n} \frac{x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

$$\text{Arithmetic Mean} = \frac{3+2+6+5}{4} = 4$$

Median

- Middle most vale when arranged in ascending or descending order
- When distribution has even number- median is average of two middle values
- When distribution has odd number- median is the middle most value

1, 3, 3, **6**, 7, 8, 9

Median = 6

1, 2, 3, **4**, **5**, 6, 8, 9

Median = (4 + 5) ÷ 2

= 4.5

Open Data Lab GCK

Mode

- Most frequently occurring value

Measures of dispersion

- Range
- Quartiles
- Variance
- Standard Deviation

Range
Difference between largest and smallest value in the dataset.

Quartile
Values that divide the dataset into four equal parts

Standard Deviation
measure of how far the data deviates from the mean of dataset

Variance
Average squared deviation from the mean.

Formula

{1, 3, 8, 3, 7, 11, 8, 3, 9, 10}

Data	Sample Mean x	Deviation $(x-\bar{x})$	Deviation ^2 $(x-\bar{x})^2$
1	6.3	-5.3	28.09
3	6.3	-3.3	10.89
8	6.3	1.7	2.89
3	6.3	-3.3	10.89
7	6.3	0.7	0.49
11	6.3	4.7	22.09
8	6.3	1.7	2.89
3	6.3	-3.3	10.89
9	6.3	2.7	7.29
10	6.3	3.7	13.69
			110.1

$= \sum (x-\bar{x})^2$

"n-1" = 9 (for denominator of sample st. deviation and variance)

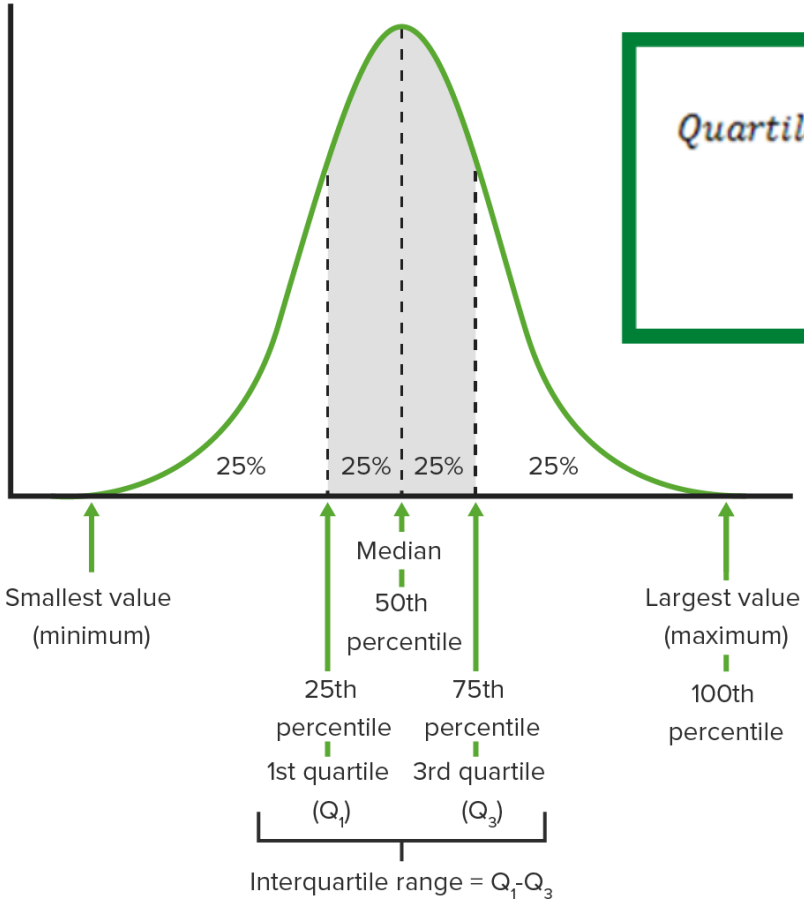
Standard Deviation Calculation

$$s = \sqrt{\frac{\sum (x-\bar{x})^2}{n-1}} = \sqrt{\frac{110.1}{9}} = 3.5$$

Variance Calculation (equals standard deviation squared)

$$s^2 = \frac{\sum (x-\bar{x})^2}{n-1} = 12.2\overline{3}$$

Quartile Deviation and Interquartile Range



$$Quartile\ Deviation = \frac{Third\ Quartile - First\ Quartile}{2}$$

$$Quartile\ Deviation\ (Q.D) = \frac{Q_3 - Q_1}{2}$$

REGRESSION

describes how an independent variable is numerically related to the dependent variable

CORRELATION

measures co-relationship or association of two variables

Correlation	Regression
Correlation is a statistical measure that determines the association between two variables.	Regression describes how to numerically relate an independent variable to the dependent variable.
To represent a linear relationship between two variables.	To fit the best line and to estimate one variable based on another.
No Difference.	Both variables are different.
To find a numerical value expressing the relationship between the variables.	To estimate values of random variables based on the values of fixed variables.

MSE

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

MSE = mean squared error

n = number of data points

Y_i = observed values

$\hat{Y}_i$ = predicted values