**Week 4 Day 1: Afternoon Session**

**Data engineering pipeline**

**1. Data Collection**

**1.1 Population and Sample Statistic**

**Population:**

* A complete collection of the objects or measurements is called the population.
* Or else everything in the group we want to learn about will be termed as population.
* In statistics population is the entire set of items from which data is drawn in the statistical study. It can be a group of individuals or a set of items.

**Population is the entire group you want to draw conclusions about.**

The population is usually denoted with **N**

1. The number of citizens living in the State of Rajasthan represents a population of the state
2. All the chess players who have FIDE rating represents the population of the chess fraternity of the world
3. the number of planets in the entire universe represents the planet population of the entire universe
4. The types of candies and chocolates are made in India.

population mean is usually denoted by the Greek letter μ

μ (population mean) =

**For example**: let us assume that there are 5 employees in a company, so 5 people is a complete set hence it will represent the population of my company

If we want to find the average age of the company then we simply add their ages and divide it by N which is the number of population

ages = {23,45,12,34,22}

                                                                  μ =

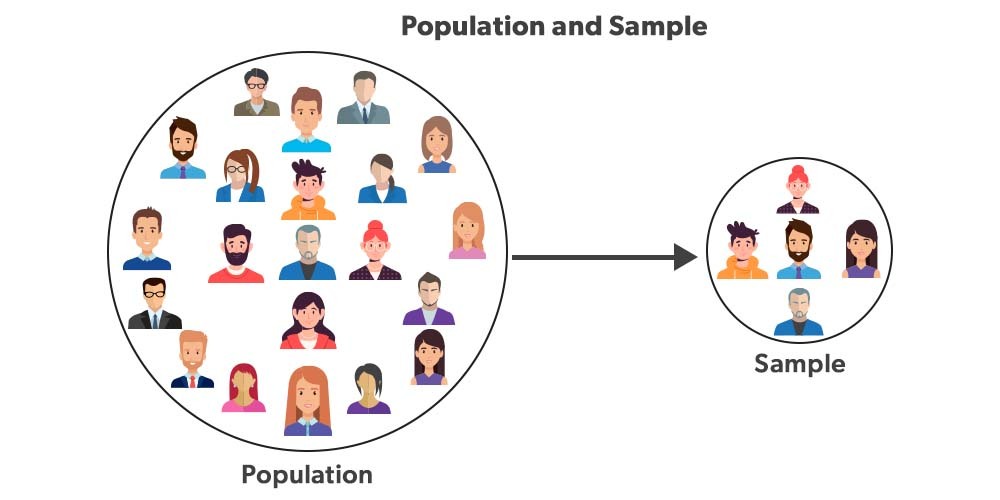
= (23 + 45 + 12 + 34 + 22) / 5

The average age of the company is 27.2 years

* Here the population was quite less so the calculation of the population means was an easy task.
* What if we want to calculate the average height of the Indians, it would be next to an impossible task.
* Even if it is not impossible it would be a difficult task.
* **Since in this case and many others it is impossible to observe the entire statistical population due to the time constraints, constraints on geographical accessibility and the constraints on the researcher resources a researcher would instead observe a sample of the same population in order to attempt to learn something about the population as a whole.**

**Sample:**

* A sample represents a group of the interest of the population which we will use to represent the data.
* The sample is an unbiased subset of the population in which we represent the whole data.
* A sample is a group of the elements actually participating in the survey or study.
* A sample is the representation of the manageable size.
* Samples are collected and stats are calculated from the sample so one can make interferences or extrapolations from the sample. This process of collecting info from the sample is called **sampling.**
* The sample is denoted by **n**



1. 500 people from a total population of the Rajasthan state will be considered as a sample
2. 150 total chess players from all total number of chess players will be considered as a sample

*Sample mean is denoted by x –*

x̄ *(****sample mean****) =*=

**for example:**

1. The total number of users of a is the population and, all student accounts of the website is a sample.
2. All the FIDE rated chess players in the population and players having a rating of more than 1700 is a sample.

***Random Sampling***

***Proportional sampling***

***Stratified Sampling***

**2. Types of Data**

### 2.1 Data Type

### 2.1.1 Type 1

### a) Cross-sectional Data

Cross-section data is collected in a single time period and is characterized by individual units - people, companies, countries, etc. Some examples include:

* Student grades at the end of the current semester;
* Household data of the previous year - expenditure on food, unemployment, income, etc.
* Car data - average speed, horsepower, colour, etc.
* **With cross-sectional data the ordering of the data does not matter**. In other words, we can order the data by ascending, descending or even randomized order and this will not affect out modelling results.
* Cross-sectional data or cross-section of a population is obtained by taking observations from multiple individuals at the same point in time.
* Cross-sectional data can comprise of observations taken at different points in time, however, in such cases time itself does not play any significant role in the analysis.
* Exam scores of high school students in a particular year is an example of cross-sectional data.
* Gross domestic product of countries in a given year is another example of cross-sectional data.
* However, we need to note that, in case of exam scores of students and GDP of countries, all the observations have been taken in a single year and this makes the two datasets cross-sectional.
* In essence, the cross-sectional data represents a snapshot at a given instance of time in both the cases.

### b) Time Series Data

Data collected at a number of specific points in time is called time series data. Such examples include stock prices, interest rates, exchange rates as well as product prices, GDP, etc.

* Time series data can be observed at many different frequencies (hourly, daily, weekly, monthly, quarterly, annually etc)
* **Unlike cross-sectional data, the ordering of the data is important in time-series data.**
* Each point represents the values at specific points in time.
* As such, time series data are typically presented in chronological order.
* Changing the order of the data ignores the time-dimensionality of the data.

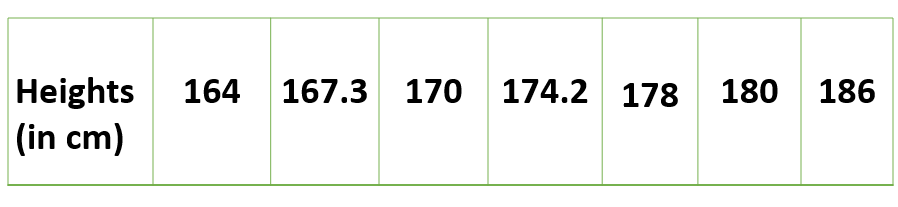
**Example:**

* If we consider only one country, and take a look at its military expenses and central government debt for a span of 10 years from 2001 to 2010, that would get two time series - one about the military expenditure and the other about debt of central government debt.
* Therefore, in essence, a time series is made up of quantitative observations on one or more measurable characteristics of an individual entity and taken at multiple points in time.
* Time series data is typically characterized by several interesting internal structures such as trend, seasonality, stationarity, autocorrelation, and so on.

**2.1.2 Type 2**

**a) Univariate data**

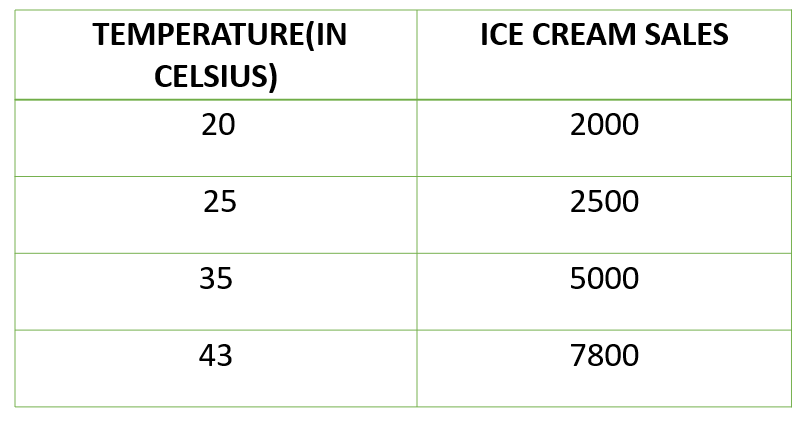
* This type of data consists of **only one variable**.
* The analysis of univariate data is thus the simplest form of analysis since the information deals with only one quantity that changes.
* It does not deal with causes or relationships and the main purpose of the analysis is to describe the data and find patterns that exist within it.
* The example of a univariate data can be height.



* Suppose that the heights of seven students of a class is recorded, there is only one variable that is height and it is not dealing with any cause or relationship.
* The description of patterns found in this type of data can be made by drawing conclusions using central tendency measures (mean, median and mode), dispersion or spread of data (range, minimum, maximum, quartiles, variance and standard deviation) and by using frequency distribution tables, histograms, pie charts, frequency polygon and bar charts.

**b) Bivariate data**

* This type of data involves **two different variables**.
* The analysis of this type of data deals with causes and relationships and the analysis is done to find out the relationship among the two variables.
* Example of bivariate data can be temperature and ice cream sales in summer season.



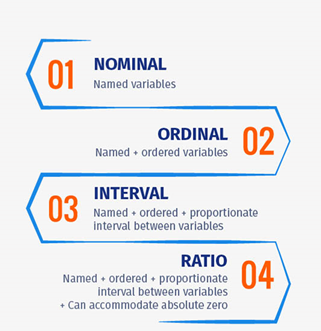
* Suppose the temperature and ice cream sales are the two variables of a bivariate data.
* Here, the relationship is visible from the table that temperature and sales are directly proportional to each other and thus related because as the temperature increases, the sales also increase.
* Thus, bivariate data analysis involves comparisons, relationships, causes and explanations.
* These variables are often plotted on X and Y axis on the graph for better understanding of data and one of these variables is independent while the other is dependent.

**c) Multivariate data**

* When the data involves **three or more variables**, it is categorized under multivariate.
* Example of this type of data is suppose an advertiser wants to compare the popularity of four advertisements on a website, then their click rates could be measured for both men and women and relationships between variables can then be examined.
* It is similar to bivariate but contains more than one dependent variable.
* The ways to perform analysis on this data depends on the goals to be achieved.
* Some of the techniques are regression analysis, path analysis, factor analysis and multivariate analysis of variance (MANOVA)

**2.2 Variable Types**

* A variable is a characteristic that can be measured and that can assume different values.
* A variable can occur in any form, such as trait, factor or a statement that will constantly be changing according to the changes in the applied environment.
* Height, age, income, province or country of birth, grades obtained at school and type of housing are all examples of variables.
* In the 1940s, Stanley Smith Stevens introduced four scales of measurement:
  + nominal
  + ordinal
  + interval
  + ratio
* These are still widely used today as a way to describe the characteristics of a variable.
* Knowing the scale of measurement for a variable is an important aspect in choosing the right statistical analysis.

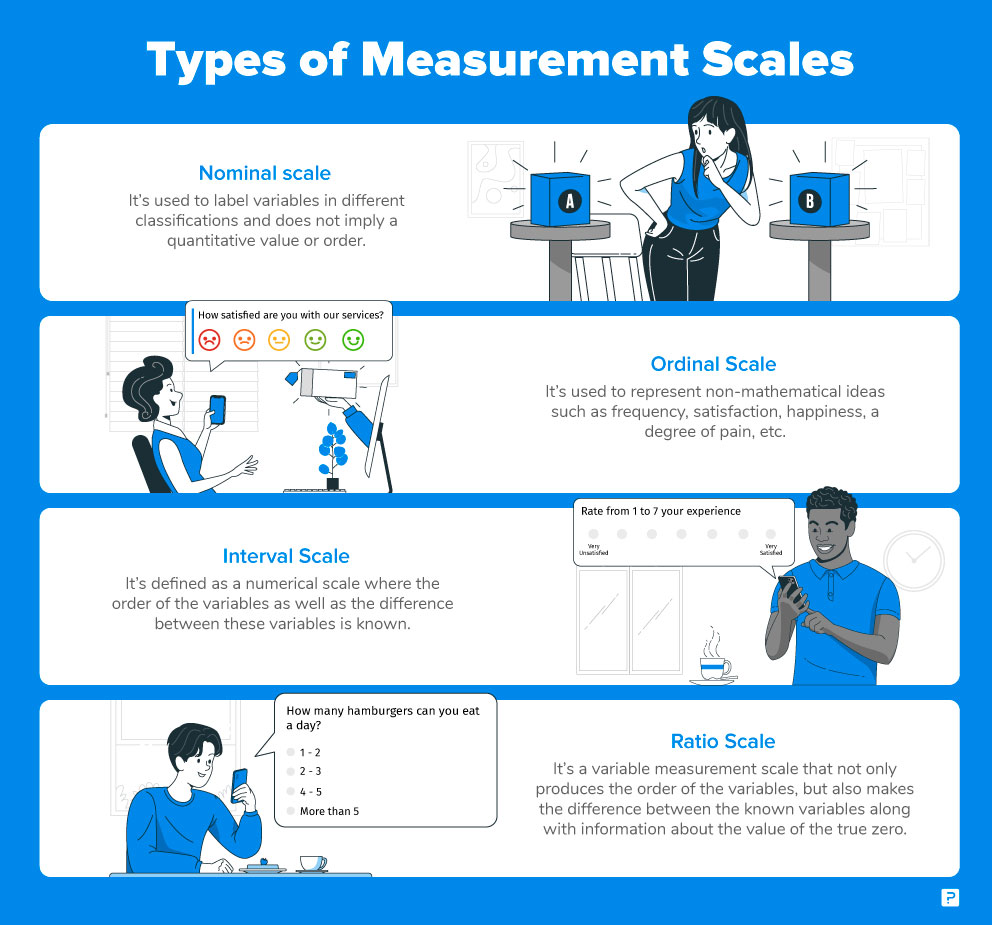


## Levels of Measurement in Statistics

* To perform statistical analysis of data, it is important to first understand variables and what should be measured using these variables.
* There are different levels of measurement in statistics and data measured using them can be broadly classified into qualitative and quantitative data.
* The level of measurement of a variable decides the statistical test type to be used. The mathematical nature of a variable or in other words, how a variable is measured is considered as the level of measurement.

## What are Nominal, Ordinal, Interval and Ratio Scales?

* Nominal, Ordinal, Interval, and Ratio are defined as the four fundamental levels of measurement scales that are used to capture data in the form of [surveys](https://www.questionpro.com/blog/surveys/) and [questionnaires](https://www.questionpro.com/blog/what-is-a-questionnaire/), each being a [multiple choice question](https://www.questionpro.com/article/multiple-choice-questions.html).
* Each scale is an incremental level of measurement, meaning, each scale fulfils the function of the previous scale



### a) Nominal /Categorical Scale: 1st Level of Measurement

* [Nominal Scale,](https://www.questionpro.com/blog/nominal-scale/) also called the categorical variable scale, is defined as a scale used for labelling variables into distinct classifications and doesn’t involve a quantitative value or order.
* This scale is the simplest of the four variable measurement scales.
* There are cases where this scale is used for the purpose of classification – the numbers associated with variables of this scale are only tags for categorization or division.
* Calculations done on these numbers will be futile as they have no quantitative significance.

For a question such as:

|  |  |  |
| --- | --- | --- |
| What is your Gender? | What is your Political preference? | Where do you live? |
| * M- Male * F- Female | * 1- Independent * 2- Democrat * 3- Republican | * 1- Suburbs * 2- City * 3- Town |

Nominal scale is often used in research surveys and questionnaires where only variable labels hold significance.

For instance, a customer survey asking “Which brand of smartphones do you prefer?”

Options :

“Apple”- 1

“Samsung”-2

“OnePlus”-3

* In this [survey question](https://www.questionpro.com/article/survey-question-answer-type.html), only the names of the brands are significant for the researcher conducting consumer research.
* There is no need for any specific order for these brands.
* However, while capturing nominal data, researchers conduct analysis based on the associated labels.
* In the above example, when a survey respondent selects Apple as their preferred brand, the data entered and associated will be “1”.
* This helped in quantifying and answering the final question – How many respondents selected Apple, how many selected Samsung, and how many went for OnePlus – and which one is the highest.
* This is the fundamental of quantitative research, and nominal scale is the most fundamental research scale.

### b) Ordinal Scale: 2nd Level of Measurement

* [Ordinal Scale](https://www.questionpro.com/blog/ordinal-scale/) is defined as a variable measurement scale used to simply depict the order of variables and not the difference between each of the variables.
* These scales are generally used to depict non-mathematical ideas such as frequency, satisfaction, happiness, a degree of pain, etc.
* It is quite straightforward to remember the implementation of this scale as ‘Ordinal’ sounds similar to ‘Order’, which is exactly the purpose of this scale.
* Ordinal Scale maintains descriptional qualities along with an intrinsic order but is void of an origin of scale and thus, the distance between variables can’t be calculated.
* Descriptional qualities indicate tagging properties similar to the nominal scale, in addition to which, the ordinal scale also has a relative position of variables. Origin of this scale is absent due to which there is no fixed start or “true zero”.

For example, a semantic differential scale question such as:

How satisfied are you with our services?

* Very Unsatisfied – 1
* Unsatisfied – 2
* Neutral – 3
* Satisfied – 4
* Very Satisfied – 5
* Here, the order of variables is of prime importance and so is the labelling. Very unsatisfied will always be worse than unsatisfied and satisfied will be worse than very satisfied.
* This is where ordinal scale is a step above nominal scale – the order is relevant to the results and so is their naming.

### c) Interval Scale: 3rd Level of Measurement

In **interval** measurement the distance between attributes does have meaning. For example, when we measure temperature (in Fahrenheit), the distance from 30-40 is same as distance from 70-80. The interval between values is interpretable. Because of this, it makes sense to compute an average of an interval variable, where it doesn’t make sense to do so for ordinal scales. An interval scale is one where there is order and the difference between two values is meaningful. You can categorize, rank, and infer equal intervals between neighboring data points, but there is no true zero point.

The difference between any two adjacent temperatures is the same: one degree. But, zero degrees is defined differently depending on the scale – it doesn’t mean an absolute absence of temperature.

The same is true for test scores and personality inventories. A zero on a test is arbitrary; it does not mean that the test-taker has an absolute lack of the trait being measured.

For instance, consider a Celsius/Fahrenheit temperature scale –

* 80 degrees is always higher than 50 degrees and the difference between these two temperatures is the same as the difference between 70 degrees and 40 degrees.
* Also, the value of 0 is arbitrary because negative values of temperature do exist – which makes the Celsius/Fahrenheit temperature scale a classic example of an interval scale.
* Interval scale is often chosen in research cases where the difference between variables is a mandate – which can’t be achieved using a nominal or ordinal scale. The Interval scale quantifies the difference between two variables whereas the other two scales are solely capable of associating qualitative values with variables.
* The mean and median values in an ordinal scale can be evaluated, unlike the previous two scales.
* In statistics, interval scale is frequently used as a numerical value can not only be assigned to variables but calculation on the basis of those values can also be carried out.

The following questions fall under the Interval Scale category:

* What is your family income?
* What is the temperature in your city?

### Ratio Scale: 4th Level of Measurement

[Ratio Scale](https://www.questionpro.com/blog/ratio-scale/) is defined as a variable measurement scale that not only produces the order of variables but also makes the difference between variables known along with information on the value of true zero. It is calculated by assuming that the variables have an option for zero, the difference between the two variables is the same and there is a specific order between the options.

With the option of true zero, varied inferential, and descriptive analysis techniques can be applied to the variables. In addition to the fact that the ratio scale does everything that a nominal, ordinal, and interval scale can do, it can also establish the value of absolute zero. The best examples of ratio scales are weight and height. In market research, a ratio scale is used to calculate market share, annual sales, the price of an upcoming product, the number of consumers, etc.

* Ratio scale provides the most detailed information as researchers and statisticians can calculate the central tendency using statistical techniques such as mean, median, mode, and methods such as geometric mean, the coefficient of variation, or harmonic mean can also be used on this scale.
* Ratio scale accommodates the characteristic of three other variable measurement scales, i.e. labeling the variables, the significance of the order of variables, and a calculable difference between variables (which are usually equidistant).
* Because of the existence of true zero value, the ratio scale doesn’t have negative values.
* To decide when to use a ratio scale, the researcher must observe whether the variables have all the characteristics of an interval scale along with the presence of the absolute zero value.
* Mean, mode and median can be calculated using the ratio scale.

#### Ratio Scale Examples

The following questions fall under the Ratio Scale category:

* What is your daughter’s current height?
  + Less than 5 feet.
  + 5 feet 1 inch – 5 feet 5 inches
  + 5 feet 6 inches- 6 feet
  + More than 6 feet
* What is your weight in kilograms?
  + Less than 50 kilograms
  + 51- 70 kilograms
  + 71- 90 kilograms
  + 91-110 kilograms
  + More than 110 kilograms

### Summary –  Levels of Measurement

The four data measurement scales – nominal, ordinal, interval, and ratio –  are quite often discussed in academic teaching. Below easy-to-remember chart might help you in your statistics test.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Offers:** | **Nominal** | **Ordinal** | **Interval** | **Ratio** |
| The sequence of variables is established | – | Yes | Yes | Yes |
| Median | – | Yes | Yes | Yes |
| Mean | – | – | Yes | Yes |
| Difference between variables can be evaluated | – | – | Yes | Yes |
| Addition and Subtraction of variables | – | – | Yes | Yes |
| Absolute zero | – | – | – | Yes |