Introduction to Data Types

Dr. Aanchal Anant Awasthi, Ph.D.
https://www.youtube.com/c/sscrindia

Content

- What is Data?
- Types of Data
- Measurement Levels in Statistics
- Do’s and Don’ts of analyzing Nominal Data
- Do’s and Don’ts of analyzing Ordinal Data
- Do’s and Don’ts of analyzing Interval/Ratio Data
- Use of descriptive statistics with different levels of measurements
- Quick Summary

What is Data?

- Data refers to a collection of facts, information, or statistics that are typically organized for analysis, interpretation, and presentation.
- It can be in various forms, such as numbers, text, images, audio, or video.
- Data is the raw material from which insights and knowledge are derived.
- Data is generated and collected from various sources, including sensors, instruments, human input, or automated systems.
- It plays a crucial role in fields such as science, business, technology, healthcare, and many others.
- By analyzing data, patterns, trends, and correlations can be discovered, leading to informed decision-making and the development of new insights and knowledge.
- Data can be categorized into various types based on different criteria.

Why Data Types?

- It's important to note that while data itself is valuable, the true value lies in the analysis and interpretation of data to extract meaningful information and derive actionable insights.

- Knowledge about types of data enables us to apply correct statistics

Structured or Unstructured Data

- Structured data is organized in a predefined format, such as spreadsheets or databases, where each data element has a specific meaning and purpose.

- Unstructured data, on the other hand, lacks a specific organization and can include things like emails, social media posts, or multimedia content.

Primary vs Secondary Data

- **Primary data** refers to information collected directly from the original source or firsthand by the researcher for a specific research purpose.
- This data is newly gathered and has not been previously published or analyzed. It can be collected through various methods, such as surveys, interviews, observations, experiments, and focus groups.
- **Secondary data**, on the other hand, refers to information that has already been collected by someone else for a different purpose and is made available for public use or research.
- This data is not gathered firsthand by the researcher but is obtained from sources like government agencies, research institutions, academic publications, reports, and databases.

Qualitative Vs Quantitative Data

- **Qualitative data** deals with characteristics and descriptors that can't be easily measured, but can be observed subjectively
- Smell, Taste, Attractiveness, Colour

- **Quantitative data** deals with numbers and things you can measure objectively
- Quantitative data is numerical information (numbers)
- It can be discrete or continuous
- Height, width, length, temperature, humidity, price, area and volume

Continuous vs Discrete Data

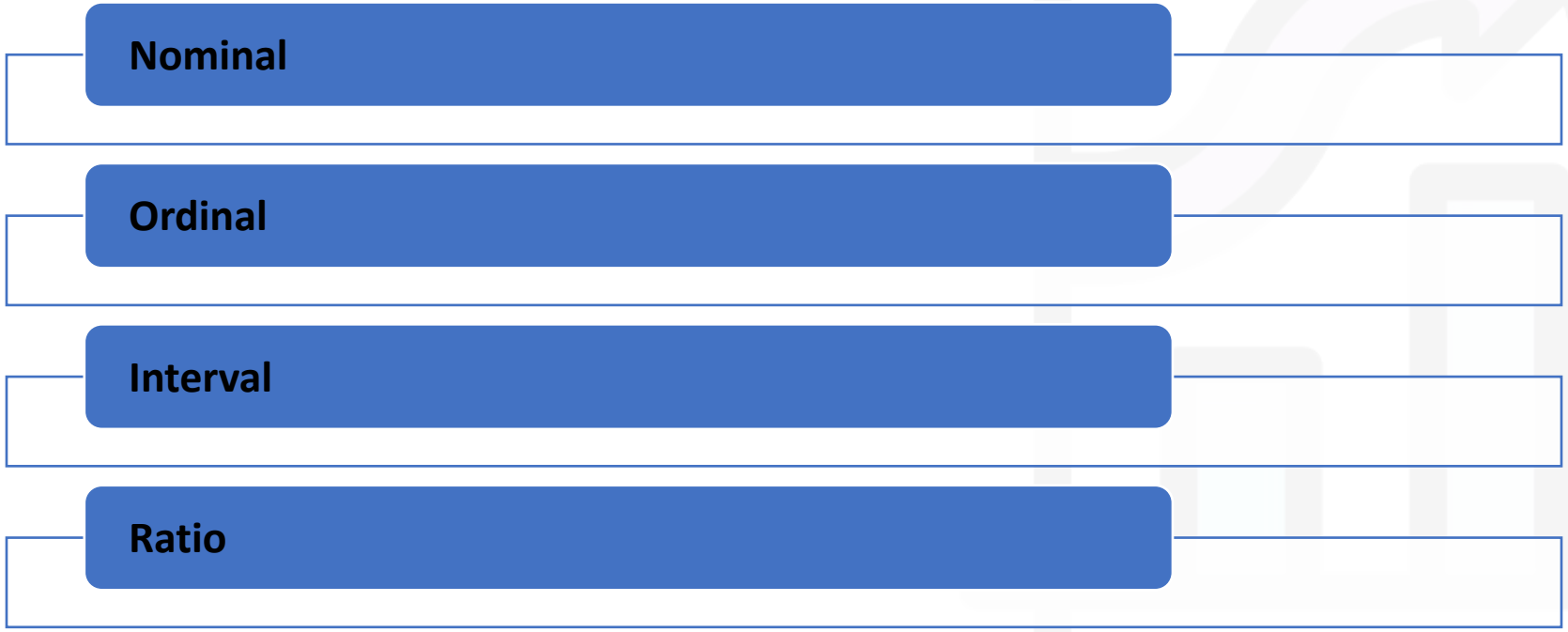
Continuous Data

- It could take any conceivable value within any observed range
- There is a possible value between any other two possible values
- Example- Height, Weight

Discrete Data

- It can take on only certain values
- Number of leaves on a plant, Number of tables

Levels of Measurement



Nominal Scale

- “Nominal” scales could simply be called “labels.”
- Objects fall into unordered categories
- Collect information through frequencies
- **Examples**
    - Hair colour: Brown, Red, Black, etc.
    - Race: Caucasian, African, American, Asian etc.
    - Smoking status: Smoker, Non-Smoker

Analysis of Nominal Data

- Frequency distribution: Mode
- Contingency table: Also known as a cross-tabulation or crosstab, this table displays the joint distribution of two or more nominal variables. It helps to examine the relationship and association between variables visually.
- Chi-squared test: This test is used to determine if there is a significant association between two nominal variables. It compares the observed frequencies with the expected frequencies based on independence assumptions.
- Association measures: Several measures can quantify the strength of association between nominal variables, such as phi coefficient, Cramér's V, or contingency coefficient. These measures indicate the degree of association or dependence between variables.
- Bar chart
- Multinomial logistic regression

What not to do when analyzing nominal data

- Avoid treating nominal data as numerical data: It is incorrect to apply mathematical operations or calculations such as addition, subtraction, or averaging to nominal data
- Avoid assuming equal intervals between categories: Nominal data does not have consistent or meaningful intervals between categories. Each category is distinct and lacks a quantitative relationship.
- Don't calculate measures of central tendency
- Avoid using complex statistical tests: Do not use tests like t-tests or ANOVA, which are suited for interval or ratio data. Instead, choose appropriate statistical tests designed for nominal data, such as the chi-squared test or Fisher's exact test.

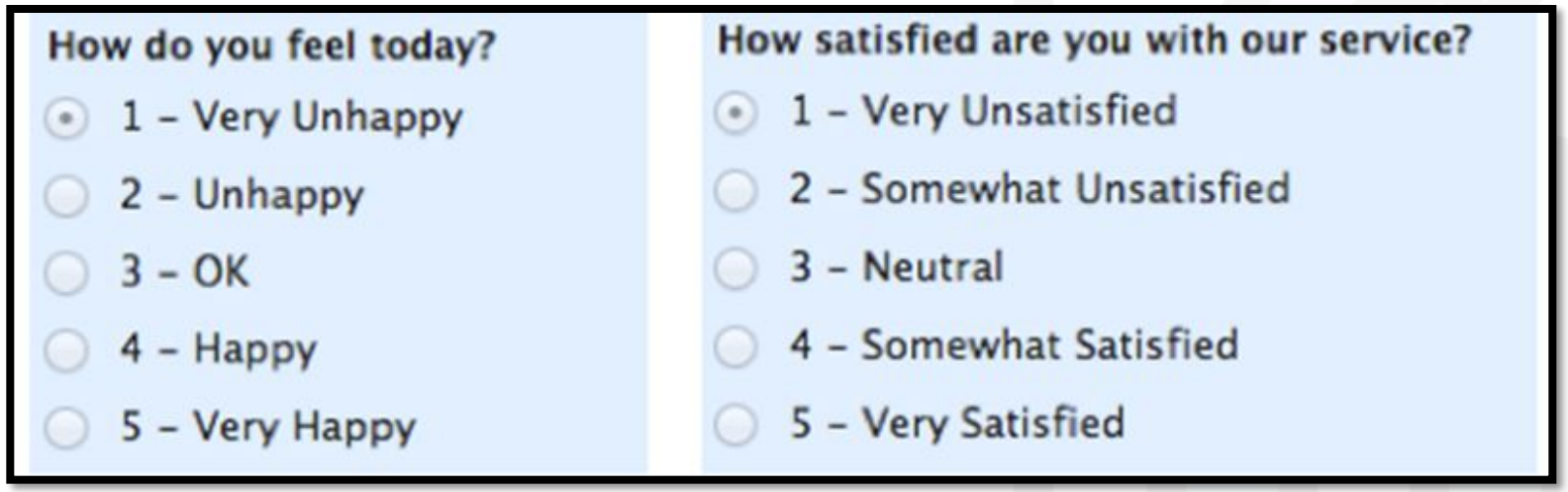
Ordinal Scale

- It is the order of the values is what's important and significant, but the differences between each one is not really known
- Ordinal scale dealing with **relative differences** rather than with quantitative differences
- Quantitative comparisons are impossible
- Examples:

Educational level, Satisfaction,
Happiness, Discomfort

Category	Level
Illetrate	0
Pre Primary	1
Primary	2
Junior High School	3
Highschool	4
Intermediate	5
Graduation	6
Post Graduation	7

Likert Scale



Ranking of cities as per Pollution status

City	Ranking
C	1
A	2
B	3

Analysis of Ordinal Data

- Descriptive statistics: frequencies, percentages, or proportions for each category or rank; Bar charts or pie charts can also be used to visualize the distribution of ordinal data.
- Rank-based tests: Since ordinal data can be ranked, nonparametric statistical tests that focus on rank comparisons are appropriate. These tests do not assume a specific distribution of the data. (Mann-Whitney U test/Wilcoxon rank-sum test, Kruskal-Wallis test)
- Ordinal regression
- Cumulative probability analysis: Cumulative probability analysis helps examine the likelihood of an ordinal outcome falling into different categories or ranks. This can be done using cumulative probability plots or cumulative odds ratios.

Analysis of Ordinal Data

- Nonparametric correlation: Spearman's rank correlation coefficient or Kendall's tau can be used. These measures evaluate monotonic relationships between variables.

- Ordinal factor analysis: If you have multiple ordinal variables and want to explore underlying factors or dimensions, ordinal factor analysis can be employed. This technique helps identify latent factors that explain the common variance among the variables.

What not to do when analyzing ordinal data

- Do not ignore the inherent ordering: Ordinal data has a meaningful order or ranking of categories. Ignoring this ordering or treating it as nominal data can result in a loss of valuable information.
- Avoid assuming equal intervals between categories: Such assumptions can lead to incorrect conclusions.
- Do not discard information: Ordinal data contains valuable information about the relative positions or ranks of categories. Avoid reducing the data to simple frequencies or proportions without considering the underlying ordering. Disregarding the ordinal nature of the data can result in a loss of important insights.
- Avoid using arithmetic operations: Ordinal data does not support arithmetic operations such as addition, subtraction, or multiplication. Performing mathematical calculations on ordinal data can lead to misleading results and invalid interpretations. Focus on analyzing the ranking and ordinal relationships instead.

What not to do when analyzing ordinal data

- Do not treat ordinal data as interval or ratio data: Ordinal data is a lower level of measurement compared to interval or ratio data. Avoid using statistical techniques or models designed specifically for interval or ratio data, as they may not be appropriate. Choose analysis methods that are specifically suited for ordinal data, such as rank-based tests or ordinal regression models.
- Avoid overlooking the context and meaning of categories: Ordinal data represents ordered categories with inherent meaning. Consider the context and significance of each category when analyzing the data. Understanding the meaning of the categories is crucial for accurate interpretation and drawing meaningful conclusions.
- Do not rely solely on summary statistics: While summary statistics like median or mode can provide a general overview of the data, they may not capture the full complexity of the ordinal relationships. Consider additional techniques like cumulative frequency plots, box plots, or rank correlations to gain a deeper understanding of the data distribution and relationships.

Interval Scale

- Possess a constant interval size

Examples: Temperature in centigrade

- Distance between 940˚C and 960˚C is the same as the distance between 1000˚C and 1020˚C

- Not a true zero (That means zero point is arbitrary)

    - But it can’t be said that a temperature of 20˚C is as twice as hot as a temperature of 10˚C

True zero point (Absence)

- Consider Temperature scale

F = 9/5C + 32

°C	5	10	15	0
°F	41	50	69	32

Circular Scales

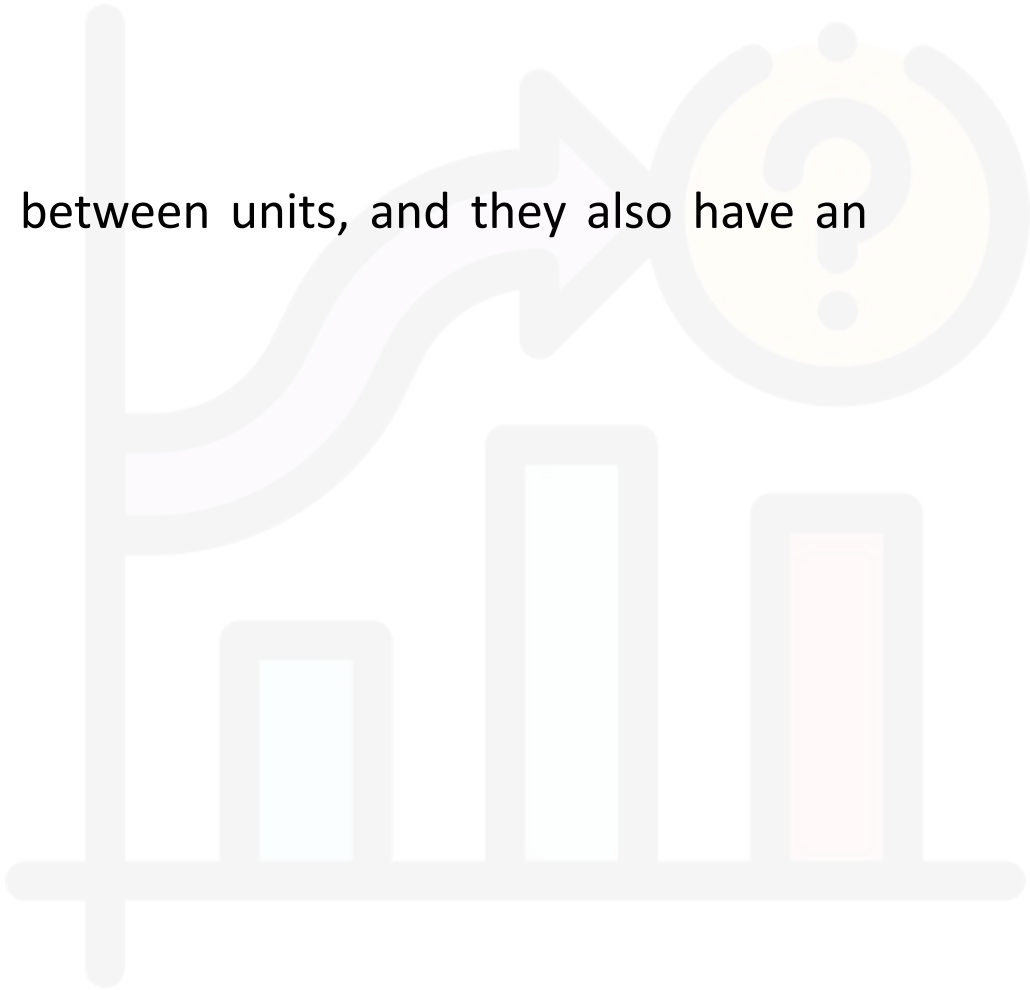
- The interval between 2:00 pm (1400hr) and 3:30 pm (1530hr) is the same as the interval between 08:00 am (0800hr) and 09:30am (0930hr).
- But one cannot speak of ratios of times of day because the zero point (midnight) on the scale is arbitrary.
- Circular biological data are usually like compass, as the designation of north as 0° is arbitrary.

Ratio Scale

They tell us about the order, they tell us the exact value between units, and they also have an absolute zero

Examples

- Weight
- Height
- Length of time (hr, days, year etc.)
- Volume



Question

- Can ordinal and discrete data types overlap?

Analysis of Interval/Ratio Data

- Descriptive statistics: Mean, median, and standard deviation are commonly used to describe the data's location, spread, and shape.
- Inferential statistics: Hypothesis testing, confidence intervals, and regression analysis can be used to make predictions, test relationships, or compare groups.
- Correlation analysis: The correlation coefficient, such as Pearson's correlation coefficient, is used to quantify the degree of association.
- Regression analysis: Linear regression and multiple regression are common techniques used for regression analysis. In case of ratio scale data, we can apply logistic regression as well.

Analysis of Interval/Ratio Data

- Analysis of variance (ANOVA): ANOVA is used to compare means across two or more groups or factors. One-way ANOVA is suitable for comparing means among multiple independent groups, while factorial ANOVA can handle multiple independent variables or factors.

- Time series analysis: Time series analysis is applied to interval data collected over time. It involves studying patterns, trends, and seasonality in the data, as well as making predictions or forecasting future values. Techniques such as autoregressive integrated moving average (ARIMA) models and exponential smoothing methods are commonly used in time series analysis.

- Cluster analysis: Cluster analysis is used to identify groups or clusters within the interval data based on similarities or dissimilarities between observations. It can help in data segmentation, customer segmentation, or pattern recognition.

What not to do when analyzing Interval/Ratio data

- Do not treat intervals as ratios: It is inappropriate to interpret ratios or proportions based on interval data. Avoid making statements like "twice as large" or "half as much" since the absence of a true zero renders such comparisons invalid.

- In case of ratio data, do not ignore the meaningful zero point: Ratio data has a true zero point, which indicates the absence of the attribute being measured. Ignoring or downplaying the significance of the zero point can lead to misinterpretations. Ensure that you consider the implications of the zero point in your analysis and conclusions

- Avoid using the mean with skewed distributions: The mean is a commonly used measure of central tendency, but it can be sensitive to outliers and skewed distributions. If the interval data is skewed, consider using the median as a more robust measure of central tendency, or transform the data to achieve a more symmetrical distribution.

What not to do when analyzing Interval/Ratio data

- Do not assume normality without justification: While many statistical techniques assume a normal distribution, it is important to assess the distribution of scaled data before applying such assumptions. Use visual tools like histograms or Q-Q plots, or conduct formal tests for normality like the Shapiro-Wilk test, to determine if normality assumptions are justified.

- Avoid extrapolating beyond the range of the data: Interval data is specific to the observed range of values in the dataset. Extrapolating beyond this range can lead to erroneous conclusions. Be cautious when making predictions or generalizations outside the range of the available data.

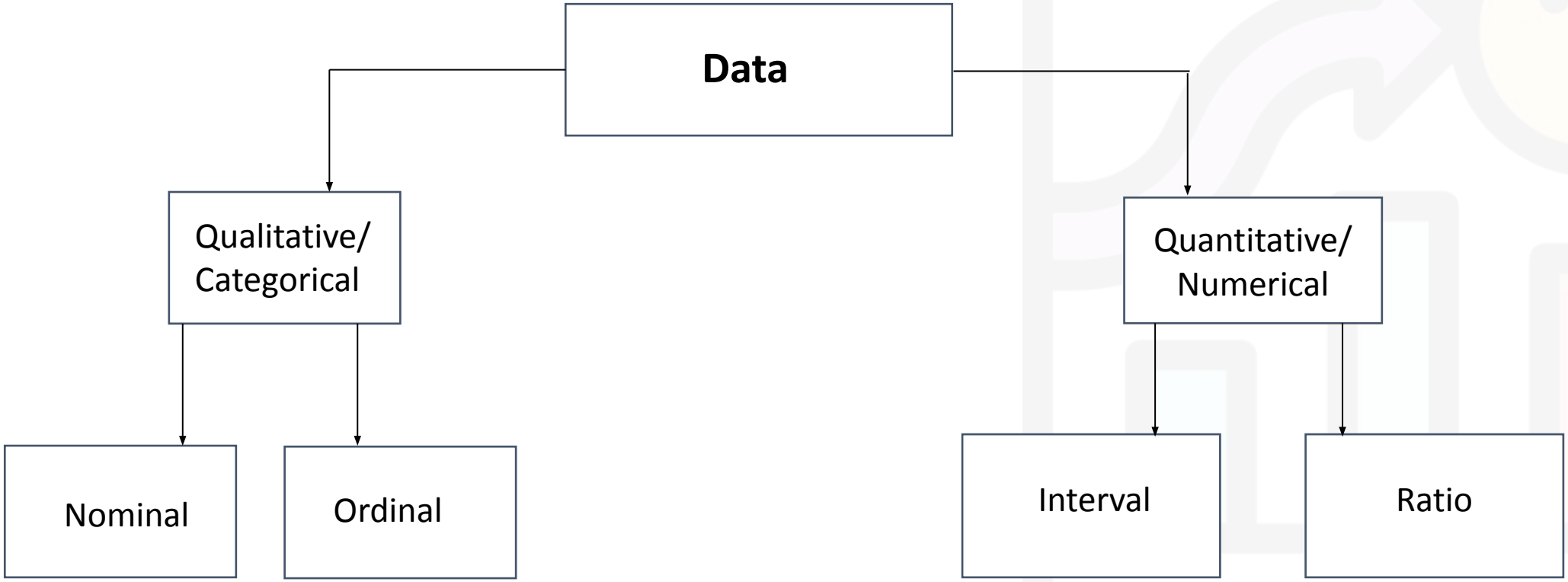
- Do not ignore missing data or outliers: Address missing data and outliers appropriately. Missing data can bias your analysis, so consider imputation techniques or evaluate the impact of missingness on the results. Outliers can distort your analysis, so assess their validity and consider their impact on the analysis outcomes.

What not to do when analyzing Interval/Ratio data

- Avoid inappropriate statistical tests: Select statistical tests that are appropriate for scaled data. Ensure that the assumptions of the chosen tests are met. For example, t-tests assume independence and normality, while ANOVA assumes equal variances. Violating these assumptions can lead to unreliable results.

- Do not neglect graphical representation: Visualizing interval data can provide valuable insights. Use appropriate graphical tools, such as histograms, box plots, or scatter plots, to explore the distribution, identify patterns, and detect outliers or influential observations.

Summary: Data Types



Summary: Level of Measurements

Scale	True Zero	Equal Intervals	Order	Category	Example
Nominal	No	No	No	Yes	Marital Status, Sex, Gender, Ethnicity
Ordinal	No	No	Yes	Yes	Student Letter Grade, NFL Team Rankings
Interval	No	Yes	Yes	Yes	Temperature in Fahrenheit, SAT Scores, IQ, Year
Ratio	Yes	Yes	Yes	Yes	Age, Height, Weight

Summary: Analysis as per type of data

	Nominal	Ordinal	Interval	Ratio
The “order” of value is known		Y	Y	Y
Counts/Frequency	Y	Y	Y	Y
Mode	Y	Y	Y	Y
Median		Y	Y	Y
Mean			Y	Y
Can Quantify the difference between each value			Y	Y
Can add or subtracts values			Y	Y
Multiplication and division				Y
Has ”true zero”				Y

The choice of analysis technique depends on the research question, nature of the data, and the objectives of the study.

Bibliography/Further Readings

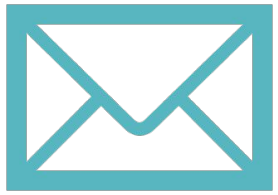
- Jerrold H. Zar. Biostatistical Analysis, Fourth Edition, Pearson Education India, 1999.

- David M. Levine, David F. Stephan, Kathryn A. Szabat. Statistics for Managers Using Microsoft Excel, 8th Edition. Pearson Publication, 2017

- I. Levin Richard, H. Siddiqui Masood, S. Rubin David, Rastogi Sanjay. Statistics for Management, Eighth Edition. Pearson Publication, 217

- https://www.kdnuggets.com/2019/10/data-sources-101.html

Lets Connect!



draanchalawasthi@gmail.com

https://www.youtube.com/c/sscrindia

+91 750.625.0403

Which among the following is example of Ratio scale

- Height in cm
- Gender
- Name
- Ranks