Descriptive Statistics and Exploratory Data Analysis

Dr. Upendra Pratap Singh

LNMIIT

August 27, 2023

Descriptive Statistics

Types of Data

1. Numerical data:
   - Continuous: real numbers
   - Discrete: discrete numbers

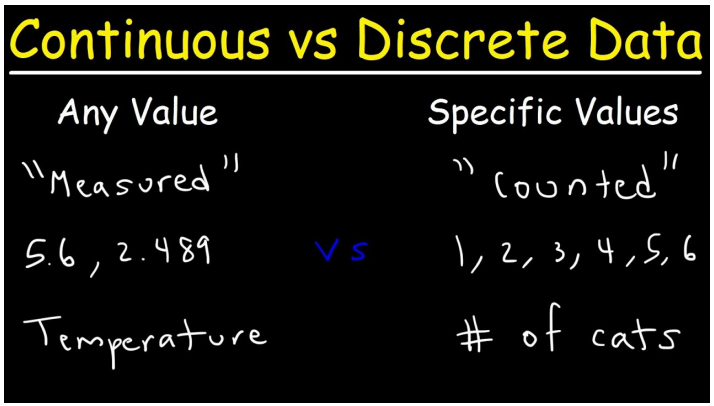

Figure: Discrete and continuous data

Descriptive Statistics

Types of Data

1. Categorical
   - Nominal: order immaterial
   - Ordinal: order remains significant

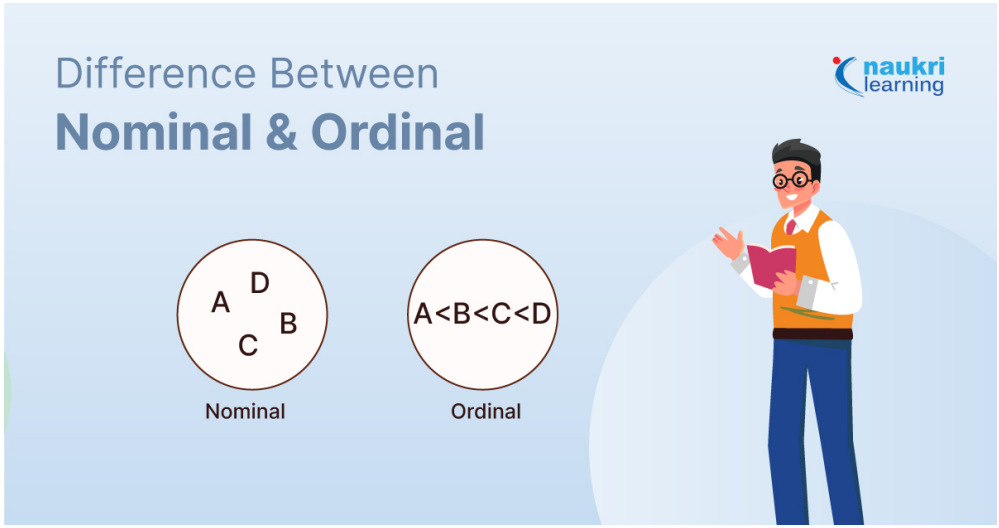

Figure: Nominal and Ordinal data

Descriptive Statistics

Types of Data

1. Time series data: data points collected at fixed intervals.
2. Text, image and audio data.
3. Relational data: data stored in tables with keys connecting them.
4. Geo-spatial data: related to geographic locations and spatial information.

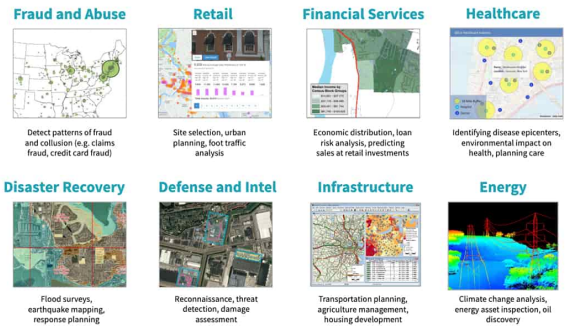

Figure: Geo-Spatial Data: Applications

Descriptive Statistics

Types of Data

1. Meta-data: provides information about other data, such as data dictionaries, data schemas, and data lineage.

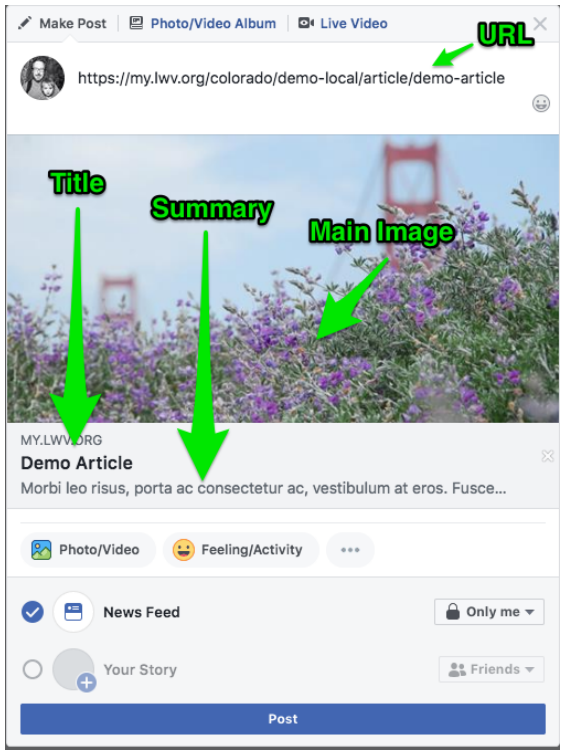

Figure: Meta Data: Application

Descriptive Statistics

Introduction

1. A set of techniques and summary measures used to *describe* and *summarize* the key characteristics of data.
2. Allows analysts and researchers to better understand the underlying data distribution, central tendencies, dispersion, and other such attributes.

Descriptive Statistics

Methods

1. **Graphical methods:** bar charts, histograms, pie-charts, scatter plots, and box-plots.
2. **Non-graphical methods:** measures of central tendency, measures of dispersion, percentiles and quartiles, correlation coefficients, and summary tables.

Descriptive Statistics

Graphical Methods - Bar Charts

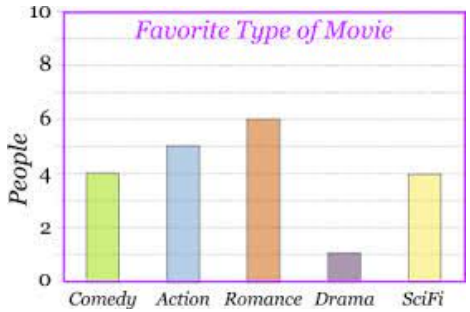

Figure: Bar Charts

1. *Categorical* Data Representation; bars can be vertical as well as horizontal.
2. Height or length measures the quantity.
3. Excellent for *comparison* and *ranking*.
4. **Limitation:** Not suited for *numerous* categories.

Descriptive Statistics

Graphical Methods - Histograms

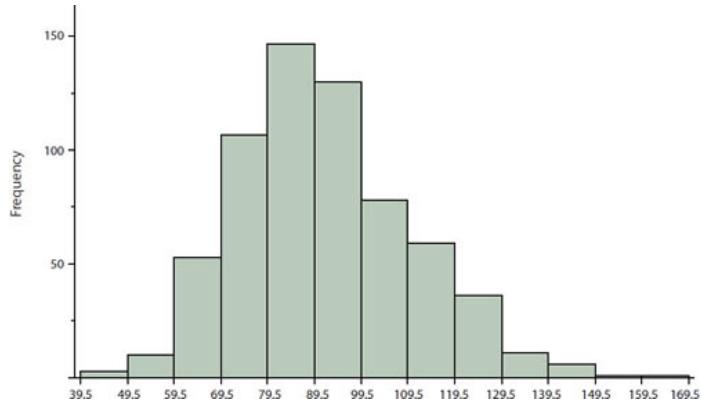

Figure: Histograms

1. Provide a visual representation of the distribution of *numerical* data.
2. Useful for understanding the *frequency/count* of data points within different ranges or bins.

Descriptive Statistics

Graphical Methods - Histograms

1. Bar height corresponds to the frequency/count of data points within the associated bin.
2. Interpretation remains *sensitive* to bar width.
3. Measures of central tendency and outliers are easily observed.

Descriptive Statistics

Graphical Methods - Bar Charts Vs Histograms

1. **Data types:** bar charts for categorical data and histograms for numerical data.
2. **Data representation:** separate bars vs bins.
3. **Bar spacing:** separate vs contiguous.
4. **Utility:** ranking and comparison vs density shape estimation.

Descriptive Statistics

Graphical Methods - Histograms

Limitations:

1. Bin size sensitivity.
2. Assumption of *data continuity*.
3. Interpreting data *between the bin boundaries* can be challenging.
4. Not suited for *complex patterns*.

Descriptive Statistics

Graphical Methods - Pie Charts

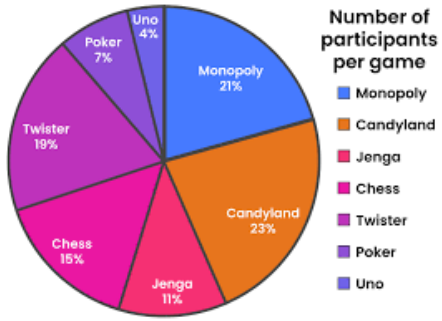

Figure: Pie-Charts

1. **Part-to-Whole Representation:** used to display the composition or distribution of a whole into its constituent parts.
2. Shape resembles a pie divided into slices; slice angle follows *proportional representation*.
3. Most suited for representing *limited* number of categories.

Descriptive Statistics

Graphical Methods - Pie Charts

1. *Invariant* to slice ordering.
2. Computation of slice angle:

$$\boxed{angle(data) = \frac{frequency(data)}{total\ data\ frequency} \times 360}$$

3. Limited interpretability with *numerous categories*.

Descriptive Statistics

Graphical Methods - Scatter Plots

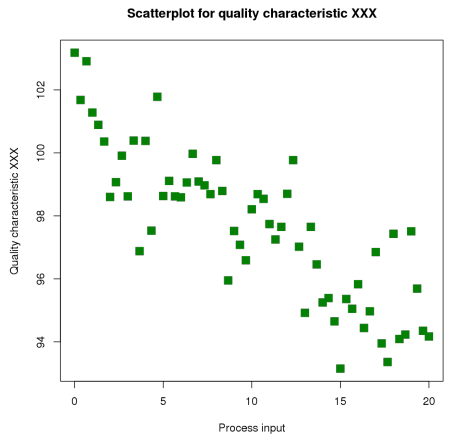

Figure: Scatter Plots

1. Depict the *relationship* between two numerical variables.
2. Each data point represents a single observation, with its coordinates (x, y).
3. *Outliers* and *correlation* information are easily observable.

Descriptive Statistics

Graphical Methods - Scatter Plots

1. Can easily *reveal* clusters/groups.
2. Primarily used to analyze relationships between variables in various fields, including economics, social sciences, natural sciences, and engineering.
3. **Correlation Vs Causation:** scatter plots do not imply causation
4. May not be useful for *large* datasets.

Descriptive Statistics

Graphical Methods - Box Plots

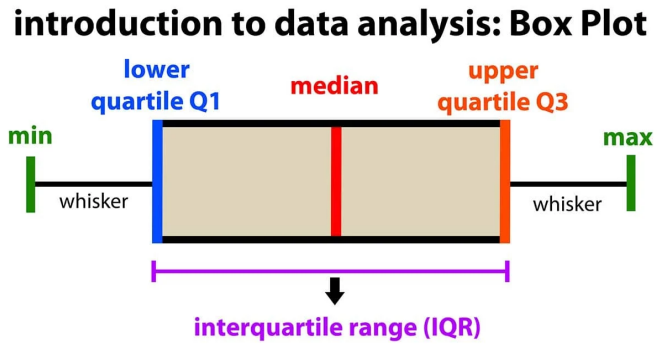

Figure: Box-Plot

1. AKA box-and-whisker plots, provide a concise summary of the distribution of numerical data.
2. Five-number summary: minimum, first quartile (Q1), median (Q2 or middle value), third quartile (Q3), and maximum.

Descriptive Statistics

Graphical Methods - Box Plots

1. Excellent for comparing distributions of multiple datasets.
2. *Assumption of Continuity:* Box plots assume a continuous distribution of data.
3. Concise representation of *5* numbers.
4. **Limitation:** they do not tell about the underlying functional form of the data.

Descriptive Statistics

Measures of Central Tendency

1. AKA measures of *central location*.
2. Provide information about the *central* or *average* value of a dataset.

Descriptive Statistics

Measures of Central Tendency

Mean

1. Calculated by adding up all the values in a dataset and then dividing by the total number of values

$$\bar{X} = \frac{1}{N} \sum_{i=1}^{i=N} X_i$$

Descriptive Statistics

Measures of Central Tendency

Mean

1. *Sensitivity* to *outliers* and *missing* values.
2. **Balancing property:** the sum of the deviations of each value from the mean is always zero.
3. **Non-integer Mean:** The mean computed may not necessarily be one of the actual values in the data.
4. May not provide a reasonable picture of the data is the distribution is skewed.

Descriptive Statistics

Measures of Central Tendency

Mode

1. Identifies the most *prevalent/frequent* value in the dataset.
2. Unlike the mean and median, which are numerical averages, the mode deals with the *occurrence* of specific values.
3. *Not always unique:* A dataset can have one mode, multiple modes, or no mode at all.
4. Applicable for both *numerical* and *categorical* data.

Descriptive Statistics

Measures of Central Tendency

Mode

1. *Insensitive* to outliers.
2. *Relationship with mean and median:* mode, mean, and median are usually close to each other for non-skewed dataset.
3. **Limitations:** it doesn't provide insight into the spread or variability of the data.

Descriptive Statistics

Measures of Central Tendency

Median

1. The median is the *middle* value in a dataset when it is ordered from least to greatest.
2. **Prerequisite:** requires the data to be *ordered*
3. Computation:
   - Arrange the data in ascending order and then find the middle value.
   - If the dataset has an odd number of values, the median is the middle value.
   - If the dataset has an even number of values, the median is the average of the two middle values.

Descriptive Statistics

Measures of Central Tendency

Median

1. *Less sensitive* to outliers compared to mean, hence, a better choice when the distribution is skewed or when outliers are present.
2. It provides a representative value that's not heavily influenced by extreme observations.
3. **Limitations:** Not directly applicable to *nominal categorical* data.

Descriptive Statistics

Measures of Dispersion

Range

1. *Difference* between the maximum and minimum values in a dataset.

$$\boxed{\textit{Range} = \textit{Maximum Value} - \textit{Minimum Value}}$$

2. *Simple* to compute.
3. Despite being a numerical value, it can be visualized using box plots.
4. It provides a basic measure of the variability or spread of the data points.

Descriptive Statistics

Measures of Dispersion

Range

1. **Not Robust:** The range is *sensitive* to *extreme* values or *outliers*.
2. **Limitation:** While the range provides insight into the extent of data dispersion, it doesn't provide information about the distribution's *shape* or *central tendency*.

Descriptive Statistics

Measures of Dispersion

Standard Deviation

1. Measures the average amount by which individual data points in a dataset deviate from the mean of the dataset.

$$\sigma = \sqrt{\frac{\sum (X_i - \mu)^2}{N}}$$

2. **Interpretation:** A larger standard deviation indicates greater variability or spread of the data points around the mean.

Descriptive Statistics

Measures of Dispersion

Standard Deviation

1. Standard deviation has the *same unit* as the original data.
2. *Sensitive* to outliers.
3. **Relation with Range:** While the range gives you the extent of the spread in the dataset, the standard deviation provides a more detailed and comprehensive measure of spread by considering the deviations of each data point from the mean.
4. **Best Practice:** use with mean.

Descriptive Statistics

Measures of Dispersion

Standard Deviation

1. **Bessel's Correction:** A mathematical adjustment applied when calculating the sample variance and sample standard deviation.
   - Sample Mean

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

   - Population Mean

$$\sigma = \sqrt{\frac{\sum (X_i - \mu)^2}{N}}$$

Descriptive Statistics

Measures of Dispersion

Variance

1. Measures the average of the squared differences between each data point and the mean of the dataset.

$$\boxed{Variance = \frac{\sum (X_i - \mu)^2}{N}}$$

2. Variance has *squared* units, which might not be directly interpretable in the same units as the original data.
3. Sensitive to outliers.
4. Variance is always a non-negative value.