Machine Learning

Introduction Basic Statistics

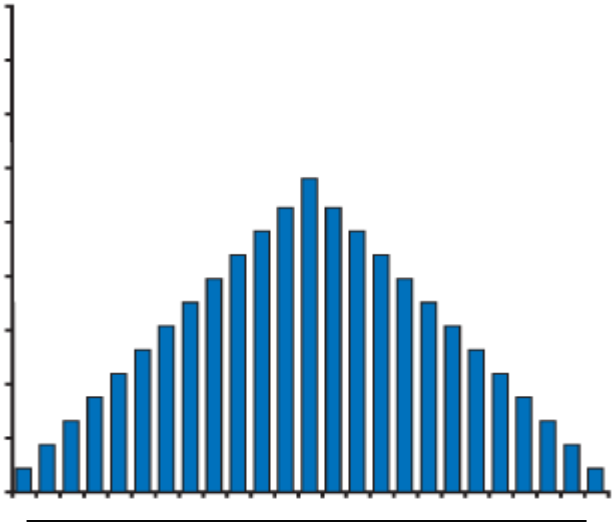
Machine Learning

- Discussion – How does Machine learning perform a prediction

Process

- A process may be well defined and well understood in terms of inputs, steps performed, output produced; or a process may be very abstract with many unknowns / ambiguities

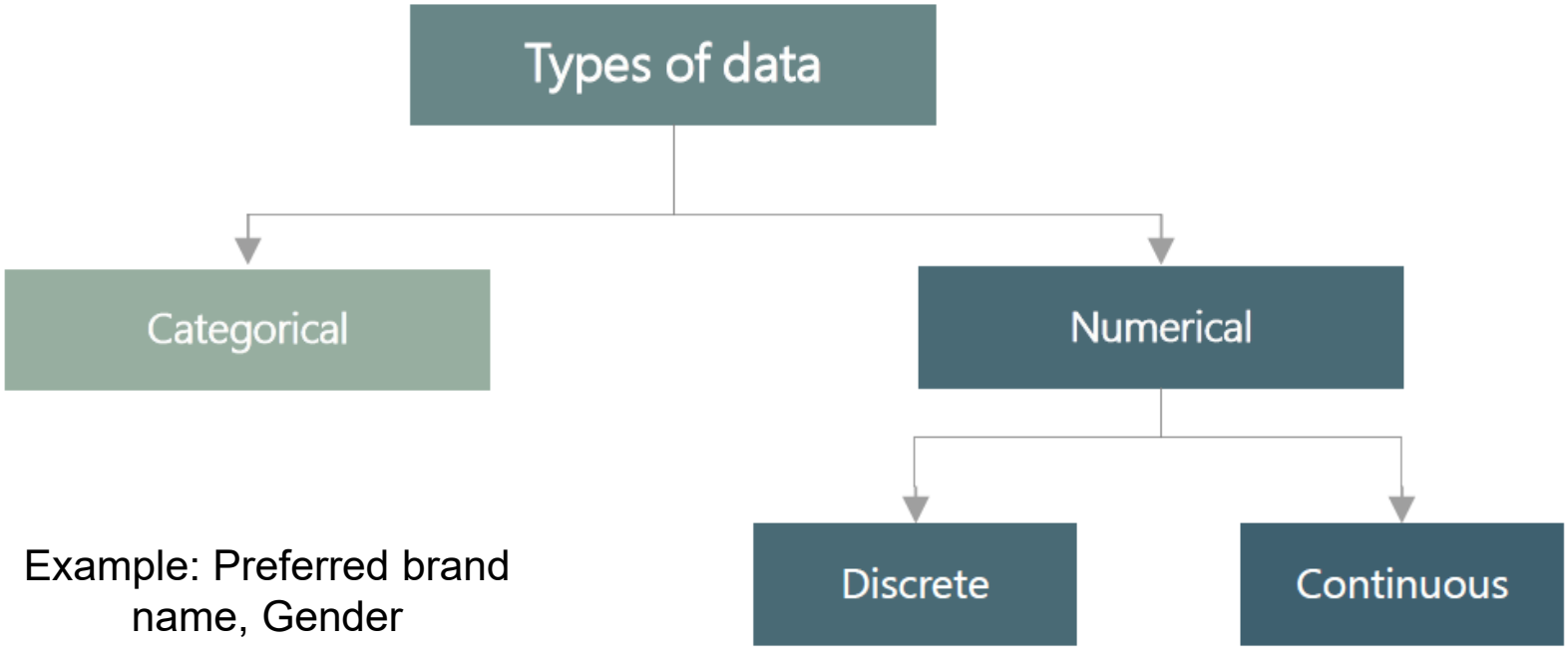
- A stable process is more predictable
- Analyzing the output produced by a process and its inputs, provides insight about the process



Population and Sample

- The collection of all data points is the "**population"** or the "**universe"** data for a process
- A subset of points drawn from a population is called "**sample"**
- Measurement of a characteristic of population is called "**parameter**"
- Measurement of a characteristic of sample is called "**statistic**"

Types of Data

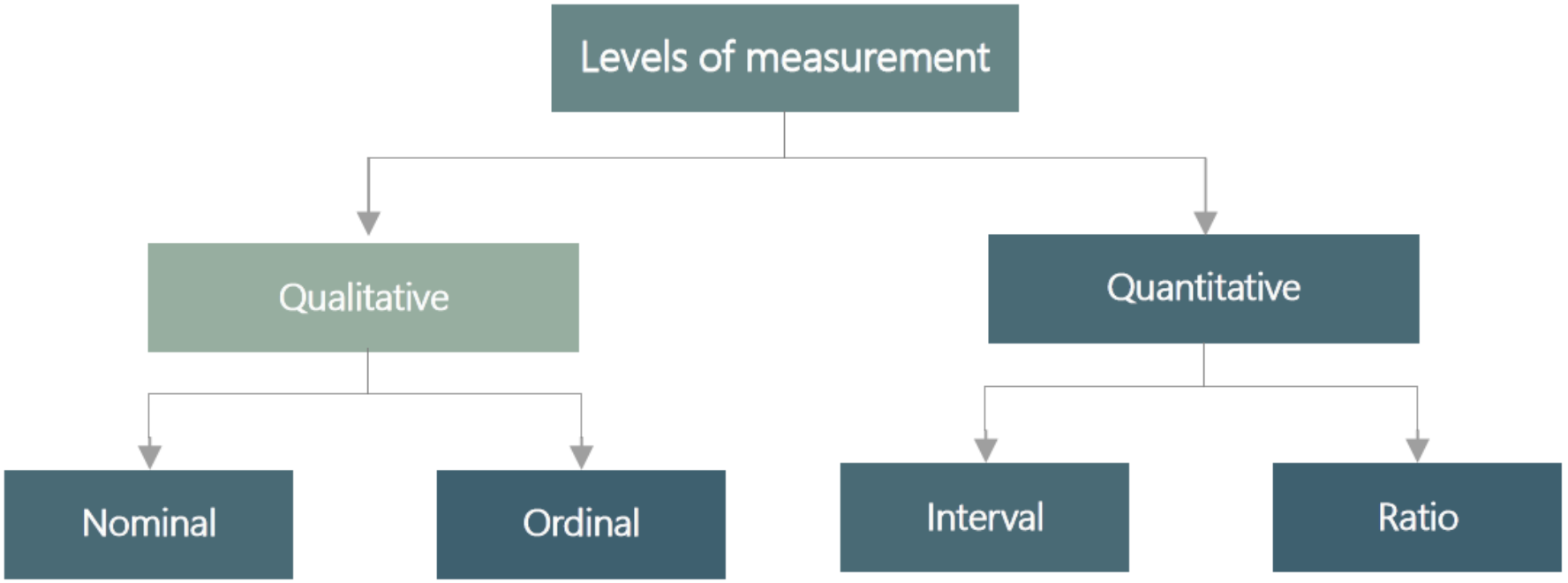


Example: Preferred brand name, Gender

Example: Number of items sold

Example: Weight of a product

Measurement Scale



Nominal does not have order (e.g. gender).

Ordinal has a meaningful order (e.g. appraisal rating)

Interval example: Temperature in Celsius.

Ratio example: Cost of an item

Descriptive Statistics

- Central Tendency
  - Mean: Arithmetic mean of numbers. Add the observations and divide by count of the observations. Mean is affected by extreme values
  - Median: When observations are sorted in ascending order, the middle observation is median. If we have n observations, the $(n+1)/2^{th}$ observation is median. The median can be an observation or between two observations
  - Mode: Mode is the most frequently occurring data point in a data set

Descriptive Statistics

ABC Pizzeria	6.5	6.6	6.7	6.8	7.1	7.3	7.4	7.7	7.7	7.7
XYZ Pizza To Go	4.2	5.4	5.8	6.2	6.7	7.7	7.7	8.5	9.3	10.0

	ABCPizzeria	XYZ Pizza To Go
Mean	7.15	7.15
Median	7.20	7.20
Mode	7.7	7.7

- The central tendency alone does not provide enough information. We need to understand the spread of data

Descriptive Statistics

- Range: It is the difference between the maximum and minimum values in a data set. Affected by extreme values

- Inter Quartile Range (IQR) – IQR is the distance between the first and the third quartile.
  - First quartile (Q1) has 25% observation lower than it.
  - Third quartile (Q3) has 75% observation lower than it
  - Median is also called second quartile (Q2)

- Variance is measured as the average of sum of squared difference between each data point (represented by xi) and the mean represented by

$$\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N} \qquad \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}$$

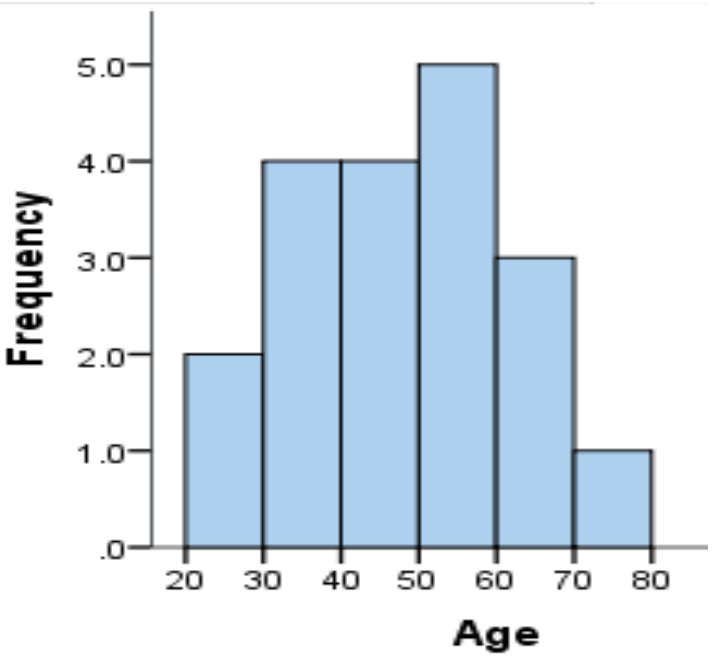
Descriptive Statistics

- Standard deviation is one of the most popular measure of spread. It is the square root of the variance.

$$\sqrt{\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}} \qquad \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}}$$

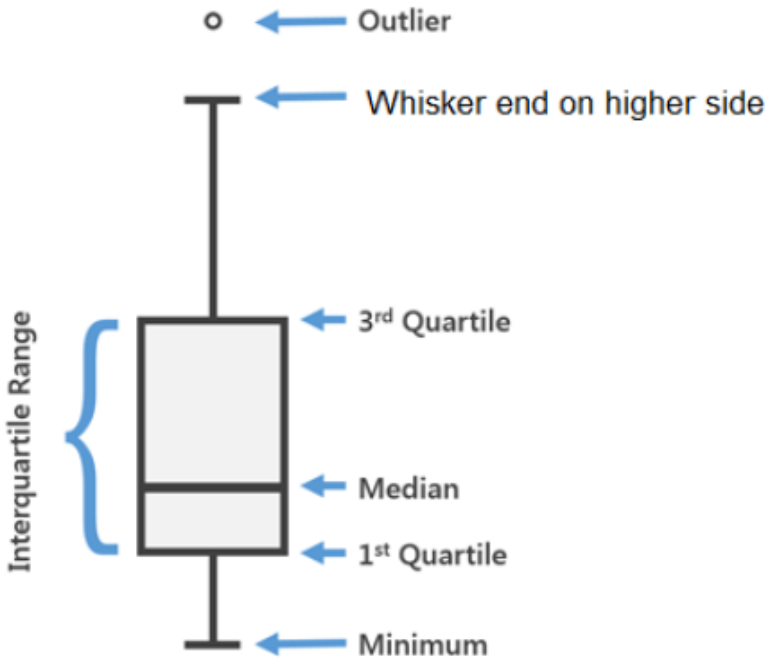
Descriptive Statistics

- Histogram: A histogram is a visual representation of the underlying frequency distribution of a data attribute.
  - Height of bars represents the frequency of occurrence
  - Width of the bars is called class intervals



Descriptive Statistics

- Boxplot: A boxplot is a standardized way of displaying the distribution of data based on a five-number summary (“minimum”, first quartile (Q1), median, third quartile (Q3), and “maximum”).
  - The box is drawn from Q1 to Q3
  - Whiskers extend maximum of (1.5 * IQR) beyond Q1 and Q3
  - Any points beyond whisker, called outliers, are also plotted



Covariance

- Covariance measures the joint variability between two numerical variables (X and Y).

- Covariance is calculated as

$$Cov(x,y) = \frac{\sum (x_i - \bar{x}) * (y_i - \bar{y})}{(n - 1)}$$

Coefficient of Correlation

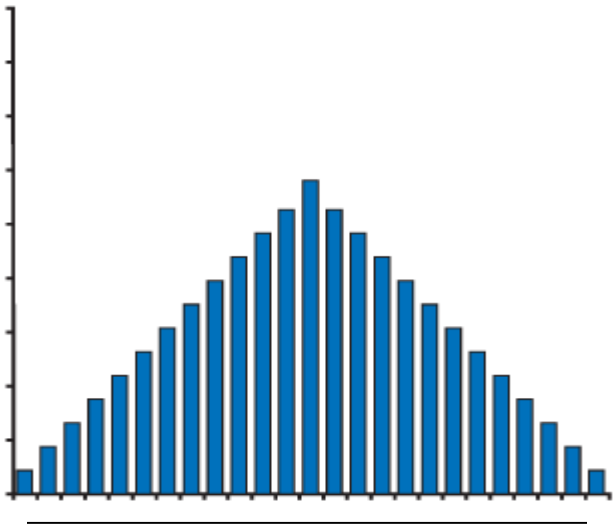
- Coefficient of correlation measures the strength of a linear relationship between two variables (X and Y).
- It is denoted by “r”. It's value can range between -1 to +1
- Value closer to +1 indicates strong positive relationship while a value closer to -1 indicates strong negative relationship

$$r = r_{xy} = \frac{\mathrm{Cov}(x, y)}{S_x \times S_y}$$

Normal Curve

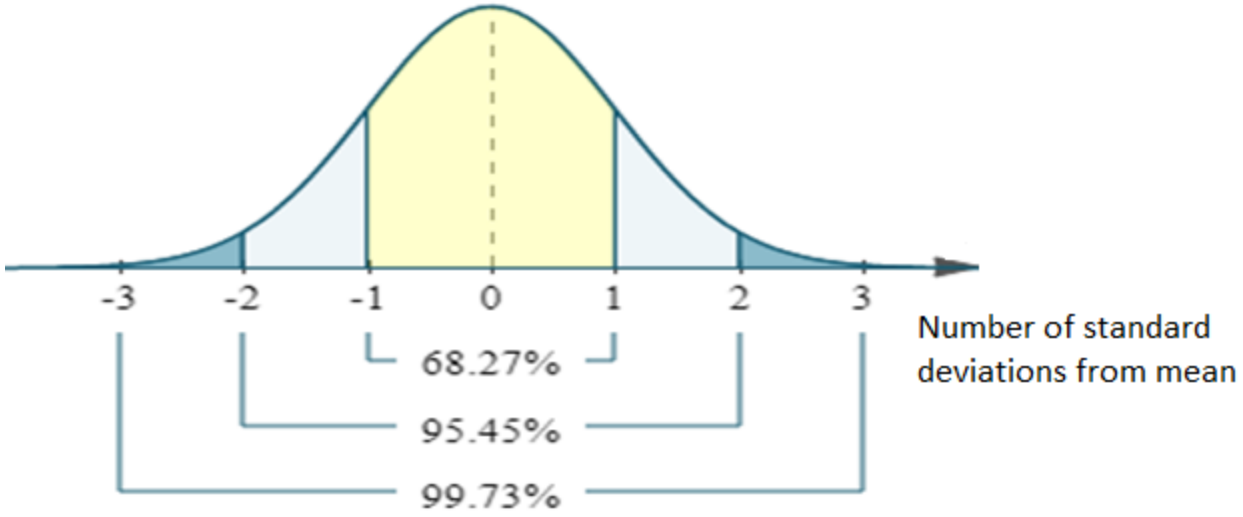
- Outputs of a process vary due to various factors that come into play during the process

- If it is possible to make probabilistic estimate of the value of the output, the process is said to be predictable



Normal Curve

- Normal distribution is a probability distribution
- A normal distribution is defined using parameters Mean and Standard Deviation
- Total area under the curve is 100%
- Area under the curve between particular values indicate probability of getting a value in that range



Normal Curve

- Normal curve is represented by the following equation

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\left[\frac{(x-\mu)^2}{2\sigma^2}\right]}$$

- Following transformation is used to convert normal distribution into Standard Normal distribution (Mean = 0 and Standard Deviation = 1)

$$z = \frac{x - \mu}{\sigma}$$

- This transformation converts a point into its Z-score

Hypothesis & Hypothesis testing

- A hypothesis is an educated guess or proposition that attempts to explain a set of facts or natural phenomenon.
- Hypothesis could be formulated based on initial analysis of available data, domain knowledge, prior experience etc.
- The goal of a hypothesis is to explain an observation and set the direction for further research
- Some sample hypothesis –
  - There is no impact of color on resale value of different cars
  - There is no impact of process improvement

Null and Alternate Hypothesis

- Null Hypothesis –
  - It is a “status quo”. Claims no significant change, no difference. E.g. When we are attempting to improve a process, we compare the metrics before and after the process change is implemented. Null Hypothesis will be – there is no change in process

- Alternate Hypothesis –
  - Alternate Hypothesis claims difference or change.

  - Alternate hypothesis stands proven when Null hypothesis is disproved In other words, if there is sufficient evidence to reject Null Hypothesis then alternate hypothesis is accepted

Hypothesis Testing

- One of the use of hypothesis testing in machine learning is to check whether there is a relationship between two attributes, for example – horse power of car engine and mileage of car.
- Null Hypothesis says “No relationship” while alternate hypothesis “There is a relationship”
- The data that shows the apparent relationship would have certain characteristics such as central values, spread, shape of the curve, tails etc.
- Statistical techniques are employed that assess the probability of getting such data (the observed values) if there was no such relationship between the attributes
- If the probability (indicated by p-value) is less than .05 (5%), then we reject null hypothesis, i.e. accept alternate hypothesis, i.e. it is considered as evidence of relationship between attributes.
-