Data Mining for Business Intelligence (IBM 312)

Sumit Kumar Yadav

Department of Management Studies

January 5, 2023

Introduction to the Course

What can you expect to learn?

- Basics of Probability and Statistics
- Distributions
- Linear / Multiple Linear Regression
- Logistic Regression, SVM, ANN, Association Rules, Clustering
- Text Analytics

Course Logistics

- Book - No single text book for the course, various references will be provided
- TA - Shivani Jaiswal - shivani_j@ms.iitr.ac.in
- Attendance - Institute Rules will be enforced
- Evaluation
  - MTE - 25%
  - Assignments/in-class quiz - 10%
  - Group Assignment - 15%
  - Course Participation - 10%
  - ETE - 40%
- All course material will be posted on MS Teams platform

Learning Outcomes Session - 1

- Descriptive Statistics
- Concept of Probability
- Random Variable, Discrete and Continuous
- Expected Value, Variance, Correlation

Definitions of Statistics

- Art of learning from data – Sheldon M. Ross, Introduction to Probability and Statistics for Engineers and Statistics
- Statistics is a branch of mathematics working with data collection, organization, analysis, interpretation and presentation. - Wikipedia
- Statistics may be regarded as (i) the study of populations, (ii) as the study of variation, (iii) as the study of methods of the reduction of data. – Fisher, 1925

Descriptive Statistics

- Describing the data, Summarize the data, etc
- Using numbers, pictures etc.
- https://www.ted.com/talks/hans_rosling_the_best_stats_you_ve_ever_seen
- https://www.ted.com/talks/hans_rosling_asia_s_rise_how_and_when

Pooled Testing Example - Warmup Problem

A LAB doing Covid testing gets 1000 samples to test everyday. However, due to the positivity rate drop in cases of Covid samples, the LAB is contemplating if it is better to mix the samples to get the result in lesser number of tests. Assuming 3% positivity rate, what is the number of samples that should be pooled together?

Summary Statistics

$x_1, x_2, \ldots, x_n$

$$\mu = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

$$\frac{\sum_{i=1}^{n} (x_i - \mu)^2}{n} = \sigma^2$$

- Measures of Central Tendency(mean, median, mode)
- Measures of Dispersion (Range, Variance)
- Chebyschev Inequality

10	8	5	0
10	9	10	10
10	10	10	10
10	11	15	20
10	12	20	

(A)
$\mu = 100$
$\sigma^2 = 10000$

(B)
$\mu = 100$
$\sigma^2 = 100$

Summary Statistics(multiple data-sets)

Co-variance and Correlation

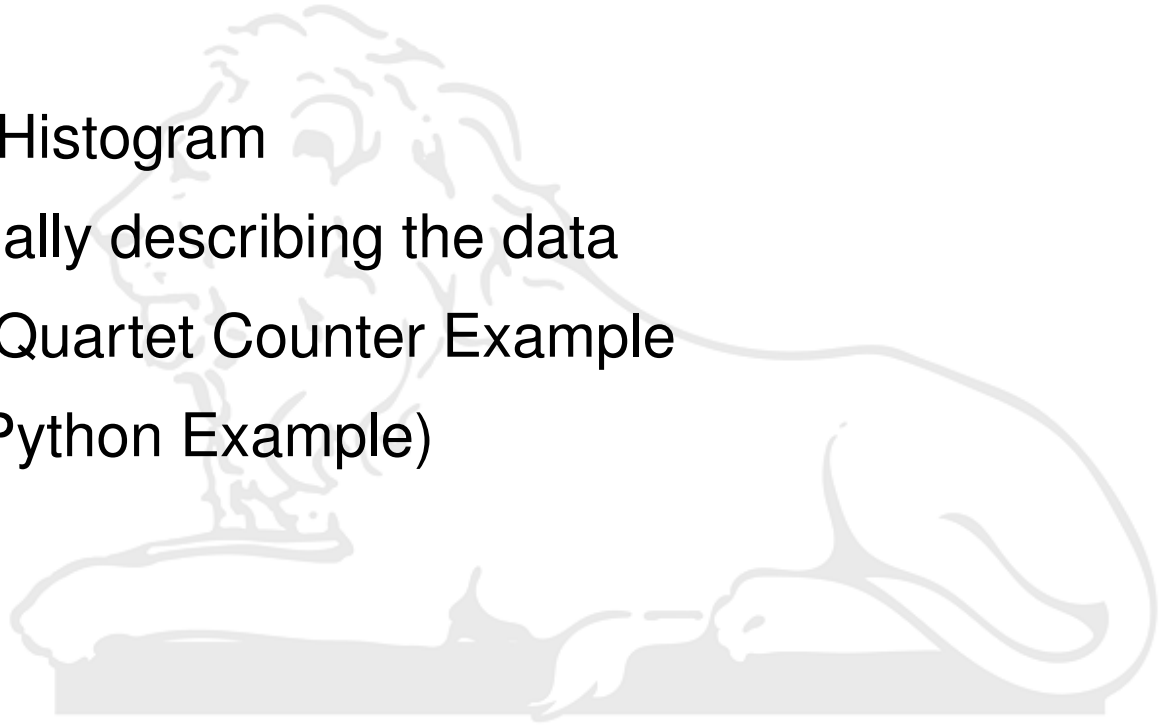
Visually describing the data

Scatter Plot, Histogram

Need for visually describing the data

Anscombe's Quartet Counter Example

Box-Plot (in Python Example)



A Few More terminologies

- Cross-sectional Data
- Time Series Data
- Panel Data

- Qualitative Data
  1. Nominal
  2. Ordinal
- Quantitative Data
  1. Interval
  2. Ratio

Probability Concepts

- What is probability??
- Concept of Experiment, Sample Space, Events
- A number associated with each Sample Point $P(E_i)$
- Less than 1
- Sum of all probabilities = 1
- $P(A_1 \cup A_2) <= P(A_1) + P(A_2)$
- Intersection of Events, Independent Events (Card Example)

Probability Concepts

- What is probability??
- Concept of Experiment, Sample Space, Events
- A number associated with each Sample Point $P(E_i)$
- Less than 1
- Sum of all probabilities = 1
- $P(A_1 \cup A_2) <= P(A_1) + P(A_2)$
- Intersection of Events, Independent Events (Card Example)

GodBole's Problem