Introduction to Data and Statistics

Prof. Vishal Singh Patyal
Associate Professor
Decision Sciences
Indian Institute of Management Visakhapatnam

1

Outline

- Introduction to Data and its types
- Descriptive Statistics vs Inferential Statistics
- Caselet Discussion

2

What is Statistics?

The science of collecting, analysis, interpreting data and presentation of numerical data

"Statistics is a way to get information from data"

Data: Facts, especially numerical facts, collected together for reference or information.

Information: Knowledge communicated concerning some particular fact.

*Statistics is a **tool** for creating **new understanding** from a set of numbers.*

3

Population

- A population consists of all the items or objects about which you want to draw a conclusion.
- The objects can be **people, animals, plants**, etc.
- Population size is usually very large (human beings) but can be very small also (panda).
- A study that involves a population is called census.
- The population size denoted by N

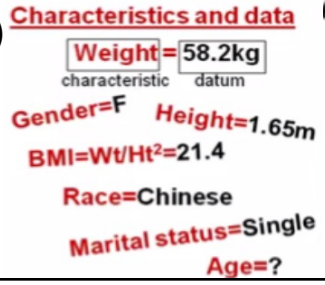
4

Object Characteristics & Data

We are usually interested in certain characteristics of the object

Data on these *characteristics can be obtained by*

- ***Measuring*** **(weight)**
- **Counting (moles)**
- **Asking (marital status)**
- **Observing (eye colour)**
- **Computing (BMI)**



5

Sample

- A ***sample*** is a subset of objects drawn from population of interest
- The sample size is denoted by n.
- n is usually smaller than population size N
- Important to determine sample size n before drawing sample from population
- ***E.g. a sample of 765 voters exit polled on election day***

6

Descriptive Statistics

- Descriptive Statistics is that branch of Statistics that **summarizes, presents and analyzes and reach conclusion about same group**.
- Descriptive statistic includes methods of
  - organizing
  - summarizing
  - analyzing
  - presenting data in an informative way

7

Descriptive Statistics

- Descriptive statistics in healthcare involve summarizing key aspects **of medical data to understand trends and patterns**.
  - Examples include c**alculating average patient age, disease prevalence rates**, and hospital performance metrics like **readmission rates**.
- It helps healthcare providers make informed decisions, **assess treatment effectiveness**, and **improve patient care**
  - by offering a clear snapshot of data **such as patient demographics, clinical trial outcomes, and health service performance.**

8

Inferential Statistics

- Another facet of statistics is inferential statistics-also called statistical inference and inductive statistics.
- Statistical inference is that branch of **Statistics that deals with drawing valid inferences about the population parameters on the basis of sample data**.

9

Inferential Statistics

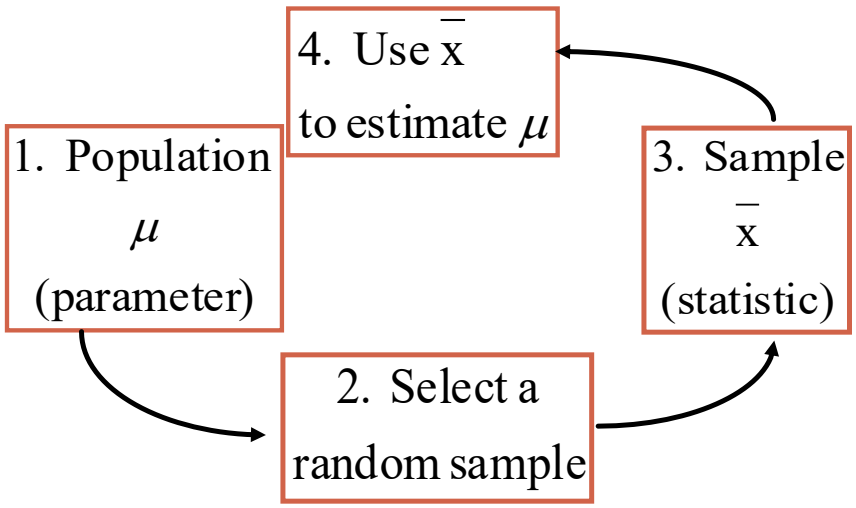
- **Estimation**
  - ex. Estimate the population mean weight using the sample mean weight
- **Hypothesis testing**
  - **Test hypotheses, and make informed decisions about medical treatments and interventions**
  - ex. Test the claim that the population mean weight is 100 kg

- ✓ **Predict and forecast values of population parameters**
- ✓ **Test hypotheses about values of population parameters**
- ✓ **Make decisions**

Drawing conclusions and/or making decisions concerning a **population** based on **sample** results.

10

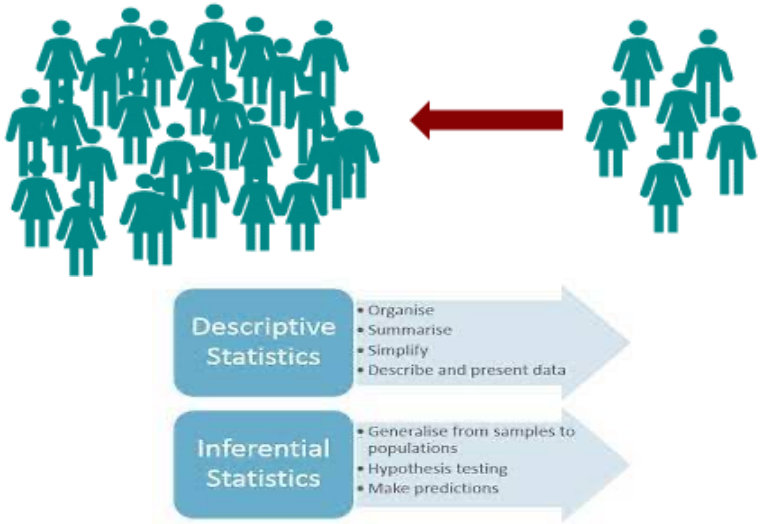
Process of Inferential Statistics



11

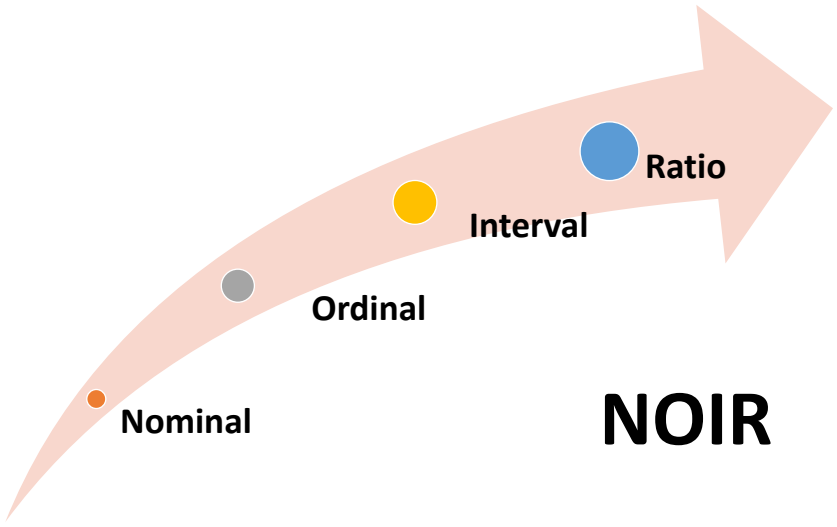
Inferential statistics:

Testing of statements about the population on the basis of sample characteristics.



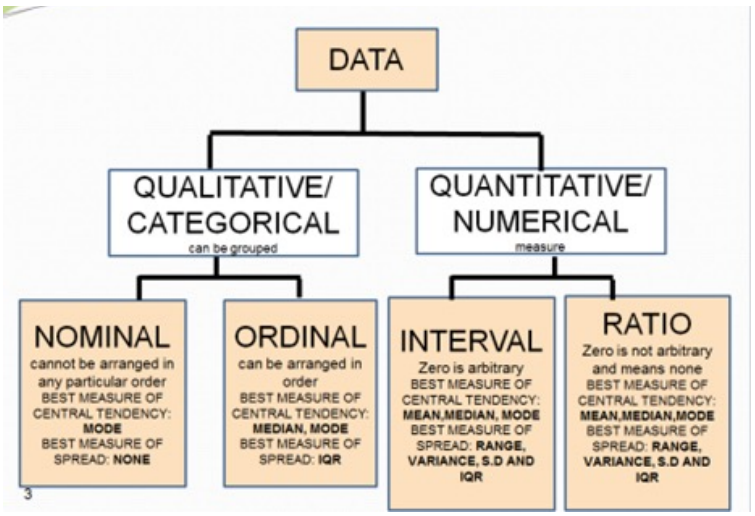
12

Level of Measurement



13

Data types

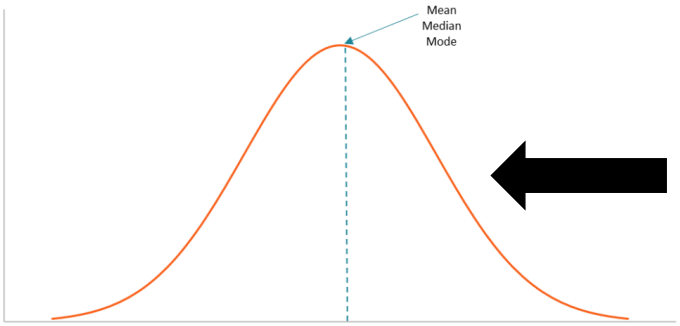


14

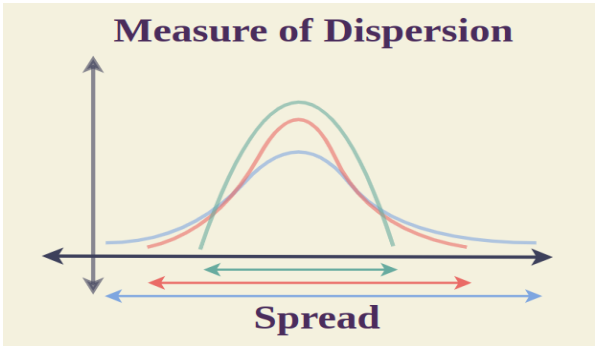
Numerical Measures (DS) [Summary]

- **Measures of Central Tendency [MoCT]**
  - ✓Median
  - ✓Mode
  - ✓Mean
  - ✓Quartiles
- **Measures of Dispersion[MoD]**
  - ✓Range
  - ✓Interquartile range
  - ✓Variance
  - ✓Standard Deviation
  - ✓Coefficient of Variation
  - ✓Quartile Deviation
- **Measures of Shape [MoS]**
  - ✓ Skewness
  - ✓ Kurtosis
- **Measures of Association [MoA]**
  - ✓ Correlation

16



MOCT

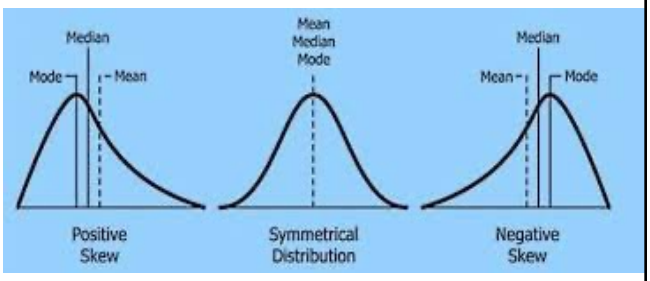


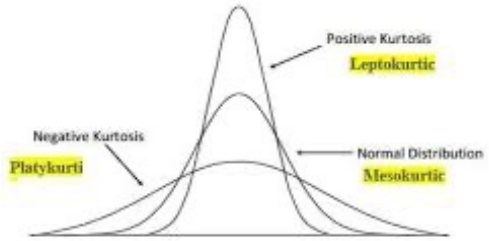
MOD

17

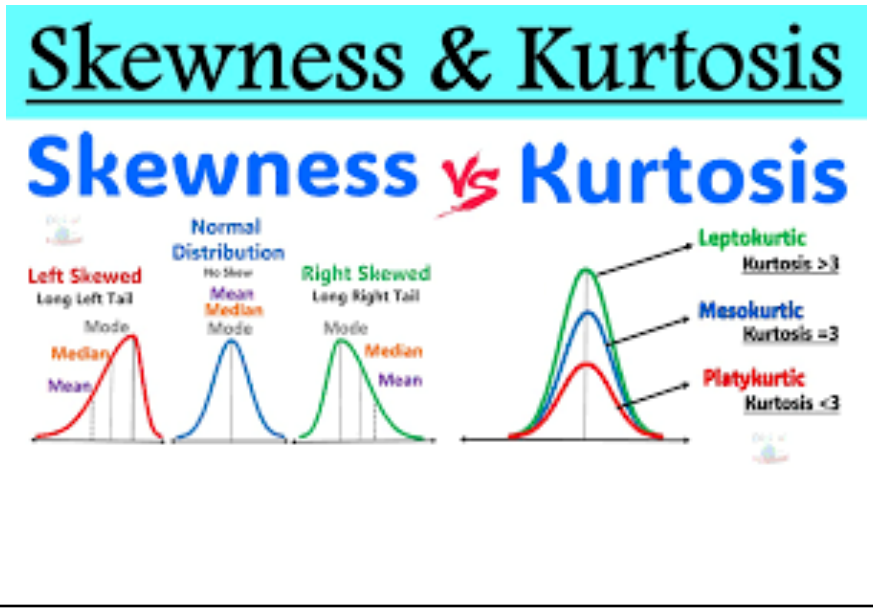
MoS

- **Skewness**
  - Absence of symmetry
  - Extreme values in one side of a distribution
- **Kurtosis**
  - Peakedness of a distribution

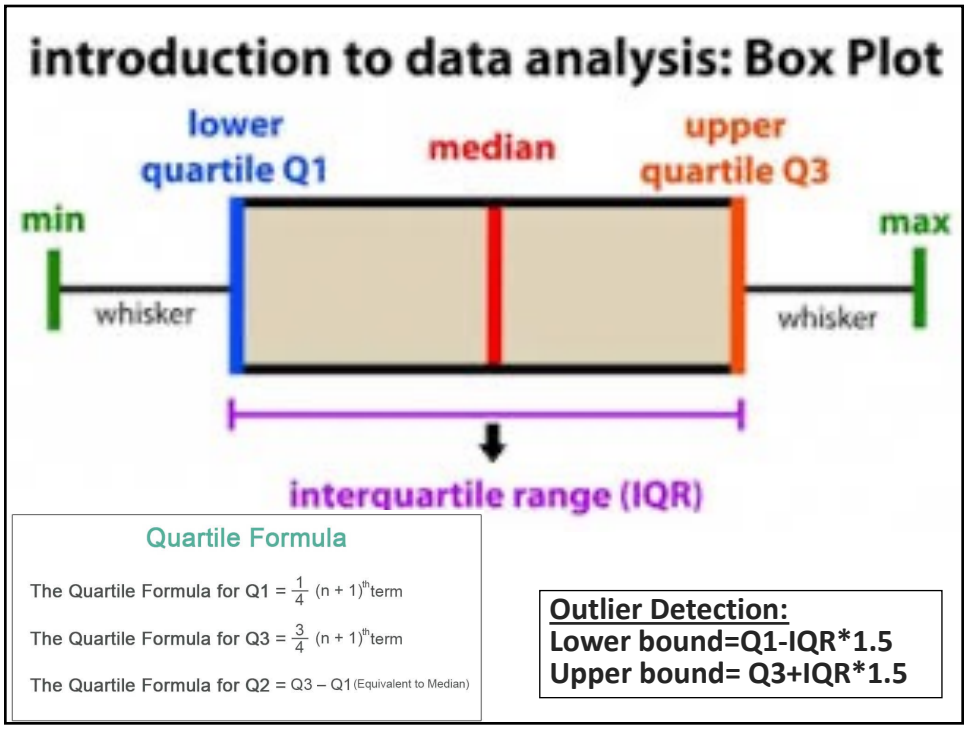




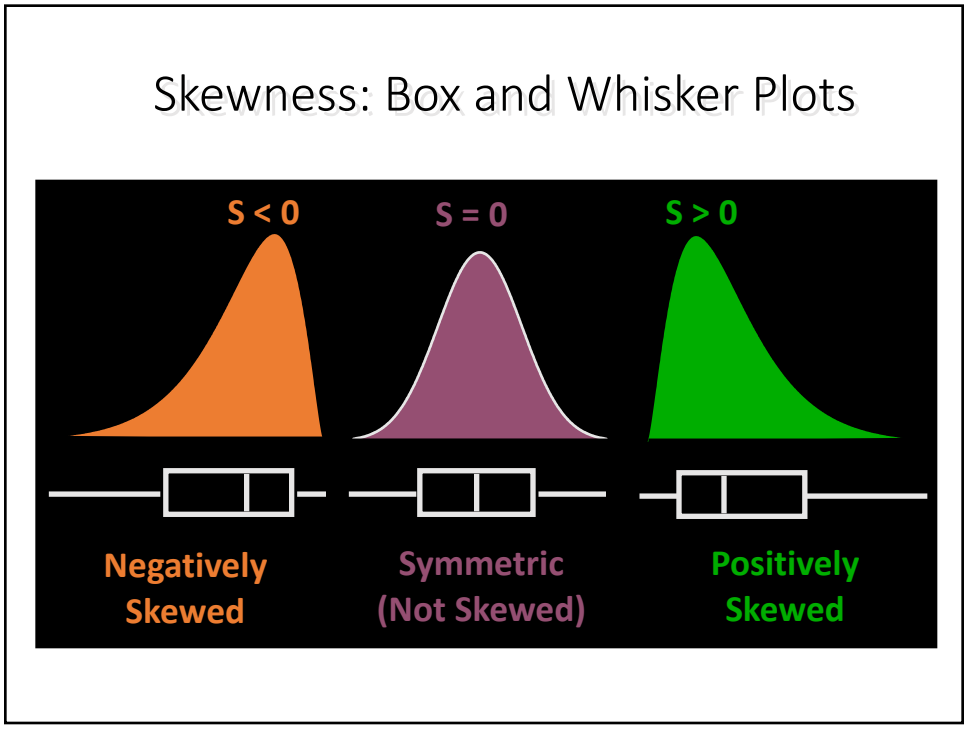
18



19

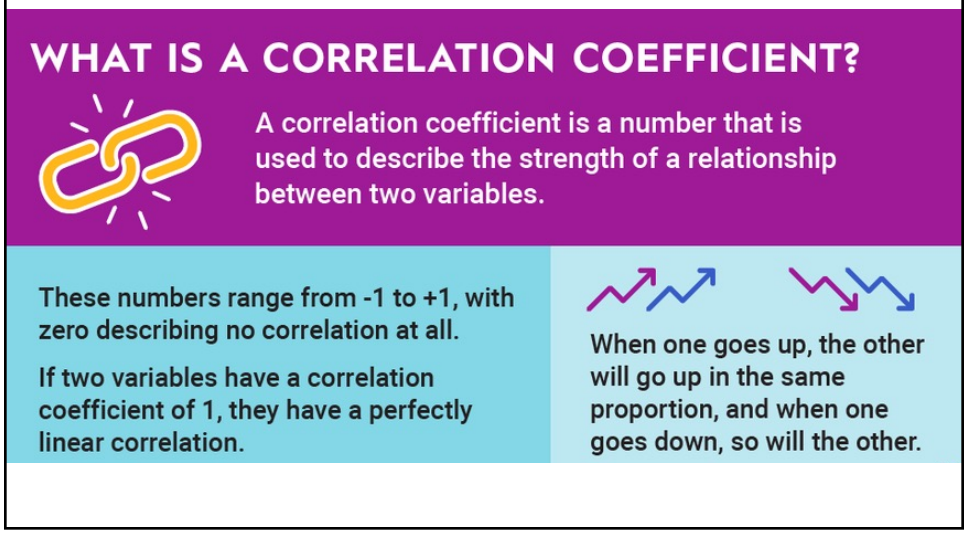

introduction to data analysis: Box Plot
lower quartile Q1
median
upper quartile Q3
min
max
whisker
whisker
interquartile range (IQR)
Quartile Formula
The Quartile Formula for Q1 = 1/4 (n + 1)th term
The Quartile Formula for Q3 = 3/4 (n + 1)th term
The Quartile Formula for Q2 = Q3 – Q1 (Equivalent to Median)
Outlier Detection:
Lower bound=Q1-IQR*1.5
Upper bound= Q3+IQR*1.5

20

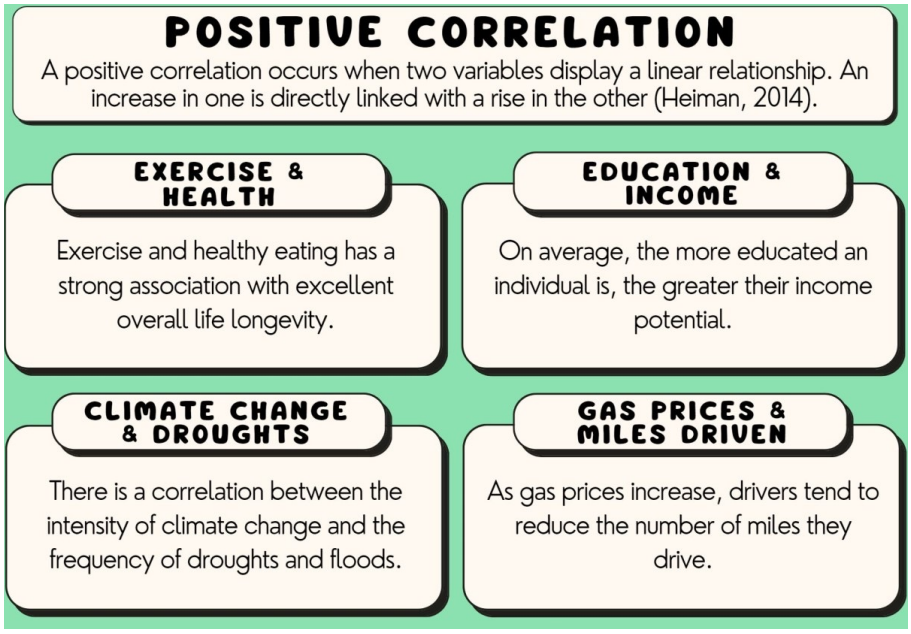

Skewness: Box and Whisker Plots
S < 0
S = 0
S > 0
Negatively Skewed
Symmetric (Not Skewed)
Positively Skewed

21

MoA: Correlation



22



23

NEGATIVE CORRELATION

A negative correlation is a relationship between two variables in which one variable decreases as the other increases.

SUGAR & ORAL HYGENE

Excessive sugar intake has been linked to decreased oral hygiene, as it not only stimulates bacterial growth but also weakens tooth enamel.

BOOZE & COORDINATION

When alcohol consumption rises, mental acuity and physical coordination decrease sharply.

INTEREST RATES & HOUSE PRICES

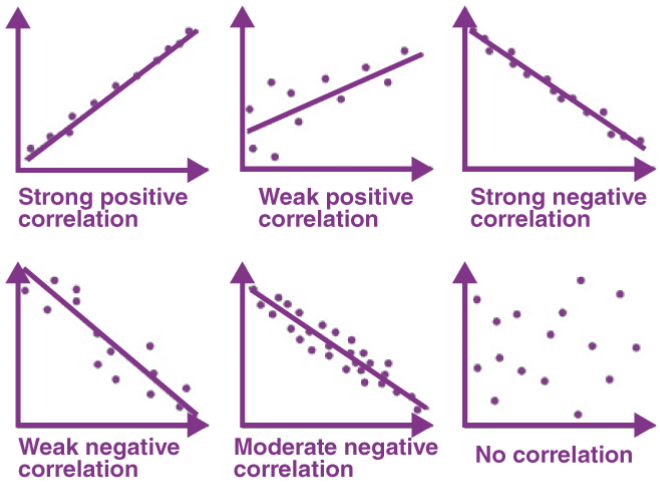
As interest rates go up, house prices go down. This is because high interest rates make it harder to get a mortgage.

WINTER & ICE CREAM

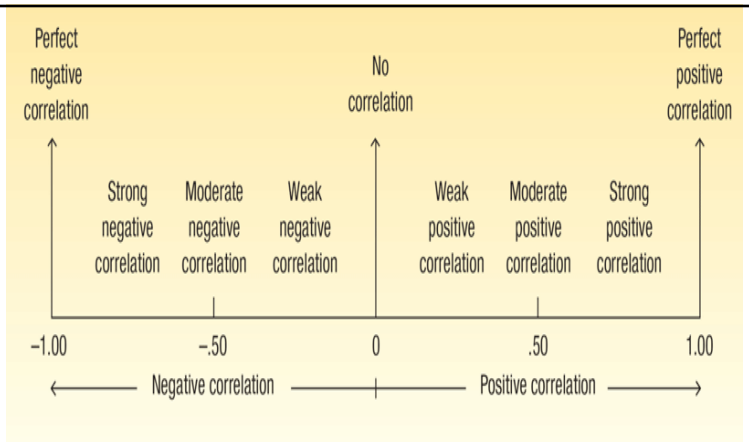
As the temperature increases, consumption of icecream goes down.

24

MoA: Scatter Plot



25



Positive	Negative	Interpretation
+1.00	-1.00	Perfect
+0.80 to +0.99	-0.80 to -0.99	Very Strong
+0.60 to +0.79	-0.60 to -0.79	Strong
+0.40 to +0.59	-0.40 to -0.59	Moderate
+0.20 to +0.39	-0.20 to -0.39	Weak
+0.01 to +0.19	-0.01 to -0.19	Very Weak

Source: Adopted from Salkind (2009)

26

Level of Measurement : Characteristics

	Nominal	Ordinal	Interval	Ratio
Labeled	☑ Yes	☑ Yes	☑ Yes	☑ Yes
Ordered	☒ No	☑ Yes	☑ Yes	☑ Yes
Known difference	☒ No	☒ No	☑ Yes	☑ Yes
Zero is arbitrary	N/A	☑ Yes	☑ Yes	☒ No
Zero means none	N/A	☒ No	☒ No	☑ Yes

27

Level of Measurement: Statistical Tests

	Nominal	Ordinal	Interval	Ratio
Mode	☑ Yes	☑ Yes	☑ Yes	☑ Yes
Median	☒ No	☑ Yes	☑ Yes	☑ Yes
Mean	☒ No	☒ No	☑ Yes	☑ Yes
Frequency distribution	☑ Yes	☑ Yes	☑ Yes	☑ Yes
Range	☒ No	☑ Yes	☑ Yes	☑ Yes
Add & Subtract	☒ No	☒ No	☑ Yes	☑ Yes
Multiply & Divide	☒ No	☒ No	☒ No	☑ Yes
Standard deviation	☒ No	☒ No	☑ Yes	☑ Yes

28

Analytics

- **Analytics** is the scientific process of **transforming data into insight** for making better decisions.
- Data-driven or fact-based **decision making**
- Improving our ability to more accurately **forecast for planning, quantify risk, and yielding better alternatives** through analysis
  - Simple reports to the most advanced optimization techniques

29

Descriptive analytics

- Descriptive analytics: It includes analytical techniques that describe w**hat has happened in the past**
- Examples
  - Data queries
  - Reports
  - Descriptive statistics
  - Data visualization
  - **Data dash boards**
  - Basic what-if spreadsheet models

30

Predictive analytics

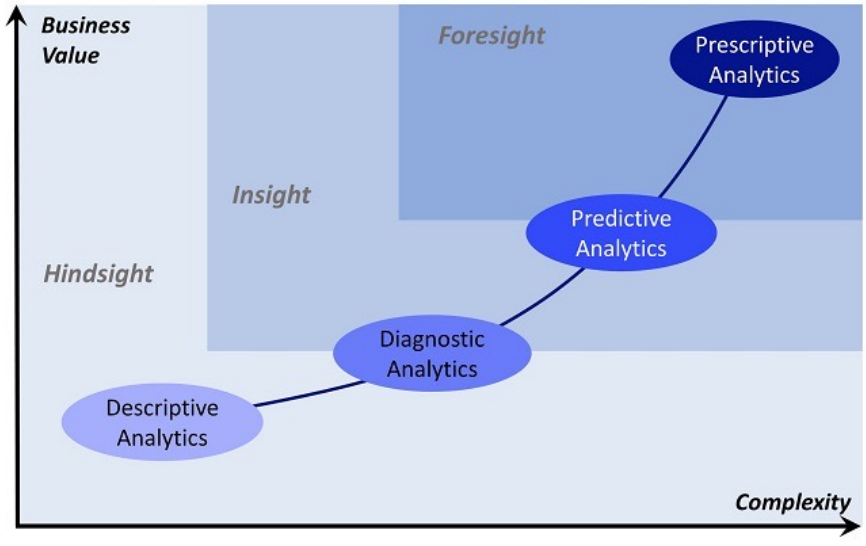
- Predictive analytics It includes analytical techniques that use **models constructed from past data to predict the future** or to assess the **impact of one variable on another**
- **Example**
  - **Product Sales** (used to construct a mathematical model that predicts future sales)
    - Factors:growth trajectory and seasonality of the product's sales (based on past growth and seasonal patterns)
  - **Point-of sale** (data from retail outlets may be used by a packaged food manufacturer to help estimate the lift in unit sales associated with coupons or sales events)
  - **Survey data** and **past purchase behavior** (used to help predict the market share of a new product)
- Methods
  - Linear regression, time series analysis, and forecasting models
  - Simulation

31

Prescriptive analytics

- Prescriptive Model: prescriptive models yield a **best course of action** to take.
  - the output of a prescriptive model is a best decision.
- Optimization Models
  - generate solutions that maximize or minimize some objective subject to a set of constraints
- Example
  - Airline industry's revenue management model
  - past purchasing data as **inputs** into a model
  - recommends the pricing strategy across all flights
  - In turn maximize revenue for the company

32



38

Caselet Discussion

40

Thank You

41