# SM727

## BUSINESS ANALYTICS & DECISION MAKING

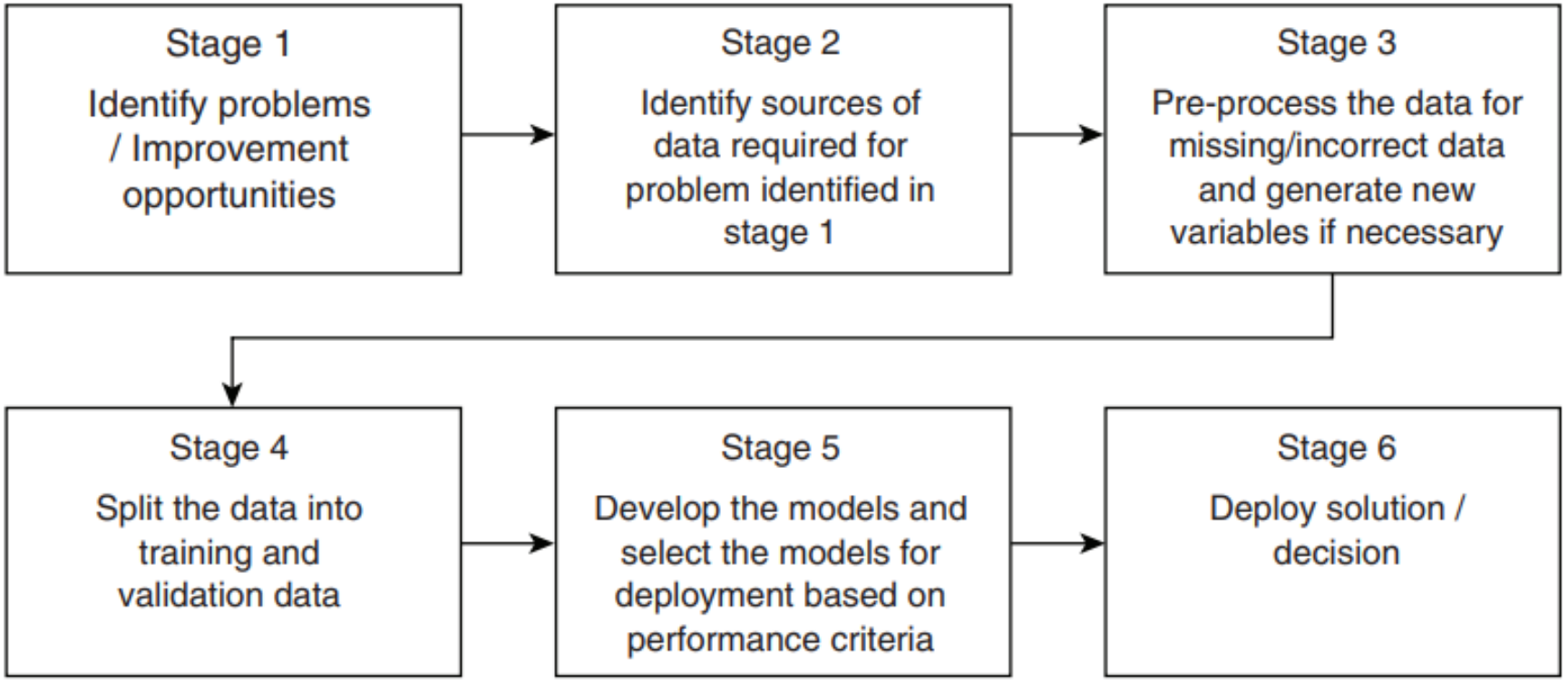
# Course Contents

| | | | |
|---|---|---|---|
| **1. Introduction to Business Analytics** | **5. Confidence Intervals** | **9. Correlation Analysis** | **13. Forecasting Techniques** |
| **2. Descriptive Analytics** | **6. Hypothesis Testing** | **10. Simple Linear Regression** | **14. Clustering** |
| **3. Introduction to Probability** | **7. Analysis of Variance** | **11. Logistic Regression** | **15. Prescriptive Analytics** |
| **4. Sampling and Estimation** | **8. Key Performance Indicator (KPI)** | **12. Decision Trees** | **16. Six Sigma** |

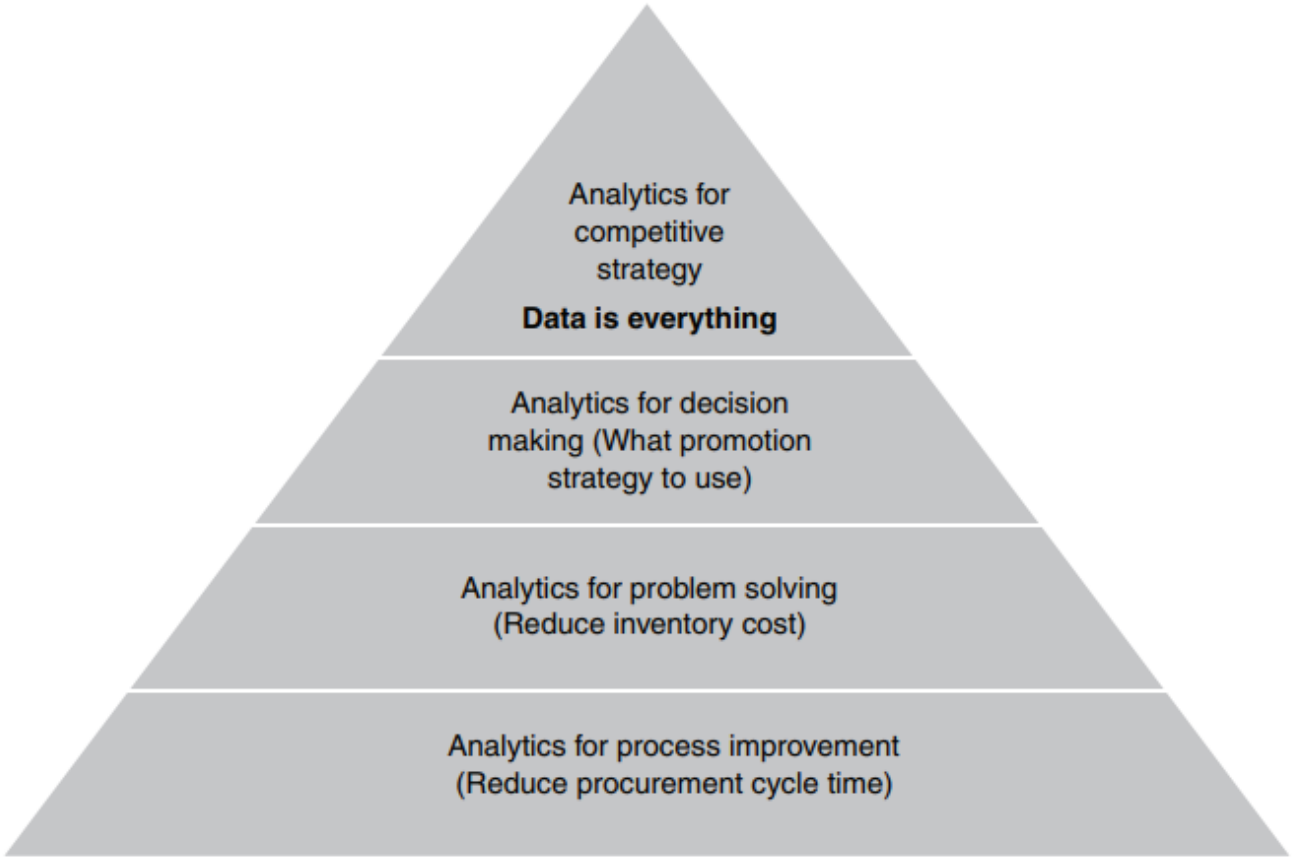
# Chapter 1

## Introduction to Business Analytics

# Data-driven Decision Making



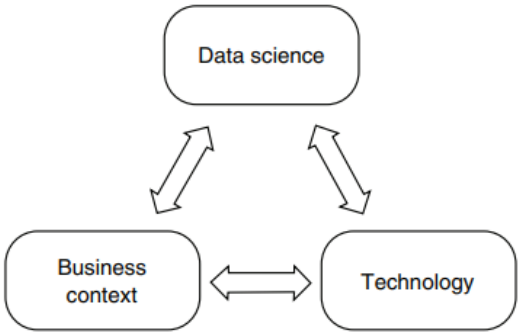
# Pyramid of Analytics



# BUSINESS ANALYTICS: THE SCIENCE OF DATA-DRIVEN DECISION MAKING

- Business analytics is a set of statistical and operations research techniques, artificial intelligence, information technology and management strategies used for framing a business problem, collecting data, and analysing the data to create value to organizations.

- Business Analytics can be broken into 3 components:

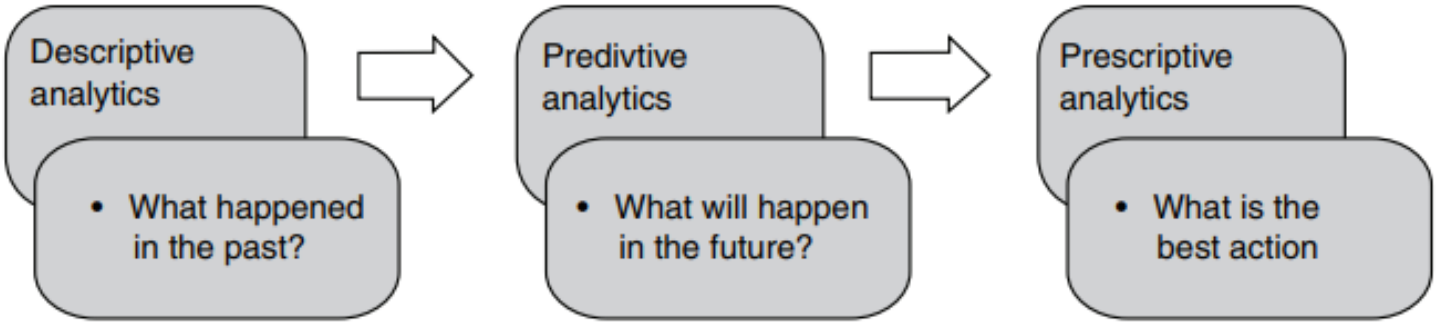


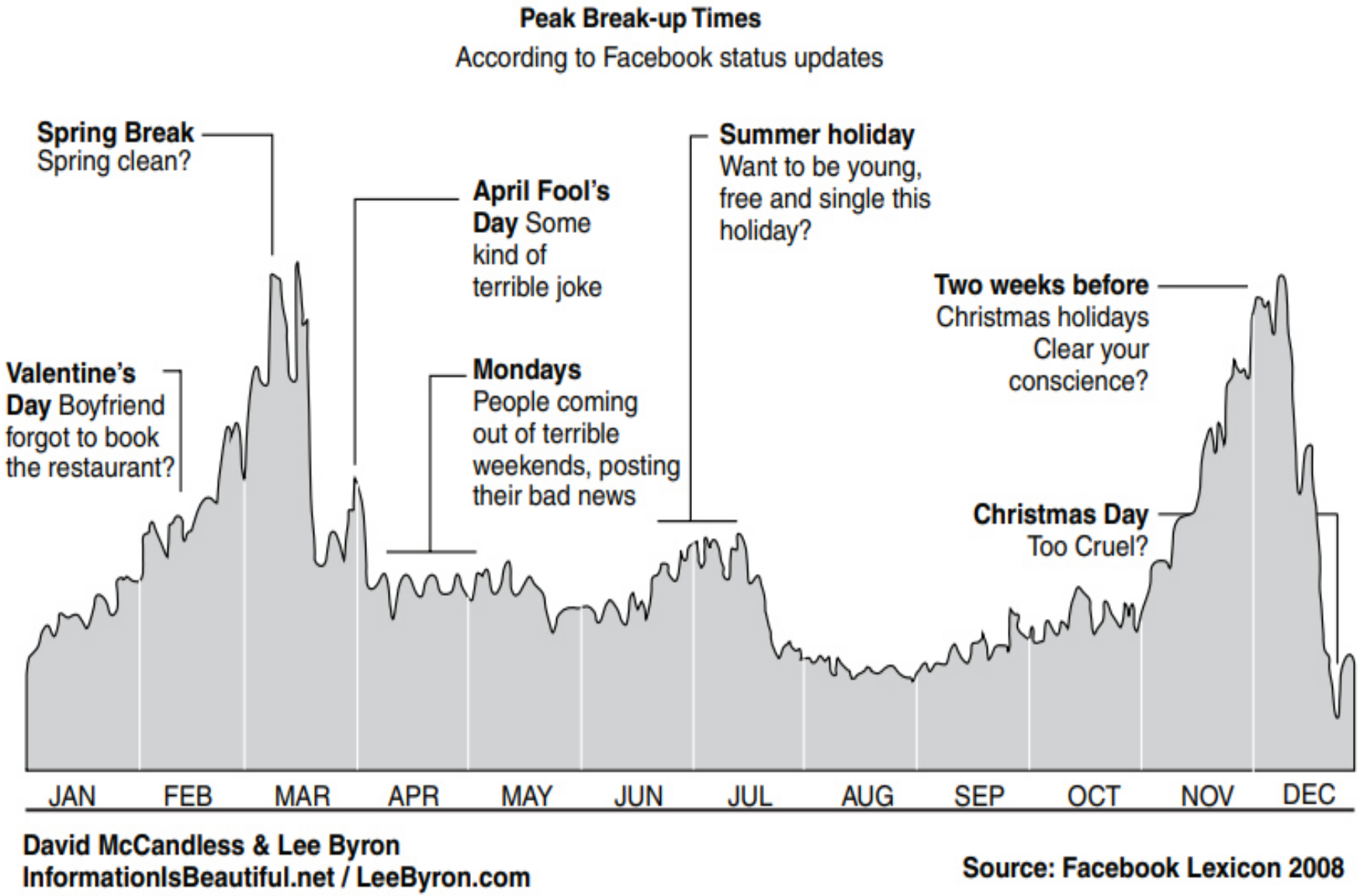
# Business context, Technology, Data science

- In analytics, knowledge of ***business context*** is important for the ability to ask the right questions to start the analytics project.
- Information ***Technology*** (IT) is used for data capture, data storage, data preparation, data analysis, and data share
- Objective of the ***data science*** component of analytics is to identify the most appropriate statistical model/machine learning algorithm that can be used.

# Types of Analytics

- Descriptive
- Predictive
- Prescriptive




Peak Break-up Times
According to Facebook status updates
Spring Break
Spring clean?
April Fool's Day Some kind of terrible joke
Summer holiday
Want to be young, free and single this holiday?
Two weeks before
Christmas holidays
Clear your conscience?
Valentine's Day Boyfriend forgot to book the restaurant?
Mondays
People coming out of terrible weekends, posting their bad news
Christmas Day
Too Cruel?
JAN
FEB
MAR
APR
MAY
JUN
JUL
AUG
SEP
OCT
NOV
DEC
David McCandless & Lee Byron
InformationIsBeautiful.net / LeeByron.com
Source: Facebook Lexicon 2008

# Descriptive Analytics

# Descriptive Analytics

# Descriptive Analytics

# Six Sigma

- https://www.youtube.com/watch?v=-K-QIwXoGHE

# Descriptive Analytics

- ***Descriptive analytics*** *is the starting point of analytics based solution to problems.*
- *It helps to understand the data and provide directions for predictive and prescriptive analytics.*
- *Business Intelligence (BI), which largely involves creating reports and business dashboard that lead to actionable insights, is essentially a descriptive analytics exercise.*
- *measures of central tendency, measures of variation and measures of shape, histogram, bar chart, pie-chart, box-plot, scatter plot and tree diagram.*

# Data types & Scales

- Structured and Unstructured Data

**TABLE 2.1** Structured data consisting of nominal and ratio scales

| No. | Gender | Age | Percentage SSC | Board SSC | Percentage HSC | Percentage Degree | Salary |
|---|---|---|---|---|---|---|---|
| 1 | M | 23 | 62 | Others | 88 | 52 | 270000 |
| 2 | M | 21 | 76.33 | ICSE | 75.33 | 75.48 | 220000 |
| 3 | M | 22 | 72 | Others | 78 | 66.63 | 240000 |
| 4 | M | 22 | 60 | CBSE | 63 | 58 | 250000 |
| 5 | M | 22 | 61 | CBSE | 55 | 54 | 180000 |
| 6 | M | 23 | 55 | ICSE | 64 | 50 | 300000 |
| 7 | F | 24 | 70 | Others | 54 | 65 | 240000 |
| 8 | M | 22 | 68 | ICSE | 77 | 72.5 | 235000 |
| 9 | M | 24 | 82.8 | CBSE | 70.6 | 69.3 | 425000 |
| 10 | F | 23 | 59 | CBSE | 74 | 59 | 240000 |

**TABLE 2.2** Unstructured data (sample clickstream data)

| |
|---|
| https://en.wikipedia.org/wiki/Clickstream |
| http://hortonworks.com/hadoop-tutorial/how-to-visualize-website-clickstream-data/ |
| http://searchcrm.techtarget.com/definition/clickstream-analysis |
| https://www.qubole.com/blog/big-data/clickstream-data-analysis/ |

# Data types & Scales

1. ***Cross-Sectional Data***: Data collected on many variables at the same time or duration of time.

2. ***Time Series Data***: Data collected for a single variable such as demand for smartphones collected over several time intervals (weekly, monthly, etc.)

3. ***Panel Data:*** Data collected on several variables (multiple dimensions) over several time intervals is called panel data (also known as longitudinal data).

# TYPES OF DATA MEASUREMENT SCALES

- ***Nominal scale*** refers to variables that are basically names (qualitative data) and also known as categorical variables.

(ex: gender, marital status etc.)

- ***Ordinal scale*** is a variable in which the value of the data is captured from an ordered set, which is recorded in the order of magnitude.

(ex: Likert scale)

- ***Interval Scale :*** the value is chosen from an interval set.

(ex: temperature, IQ)

- ***Ratio Scale :*** Any variable for which the ratios can be computed and are meaningful.

(ex: salary)

# Population & Sample

- ***Population*** is the set of all possible observations (often called cases, records, subjects or data points) for a given context of the problem. The size of the population can be very large in many cases.

- Population (also known as universal set) is the set of all possible data for a given context whereas ***sample*** is the subset taken from a population.

# MEASURES OF CENTRAL TENDENCY

- measures that are used for describing the data using a single value.

| MEAN | MEDIAN | MODE |
| --- | --- | --- |
| • arithmetical average value of the data<br>$\bar{X} = \sum_{i=1}^{n} \frac{X_i}{n}$ | • value that divides data into 2 equal parts<br>• (n + 1)/2 when n is odd and when n is even, average value of (n/2)th & (n + 2)/2th observation | • most frequently occurring value in the data set.<br>• valid for qualitative (nominal) data |

# PERCENTILE, DECILE, AND QUARTILE

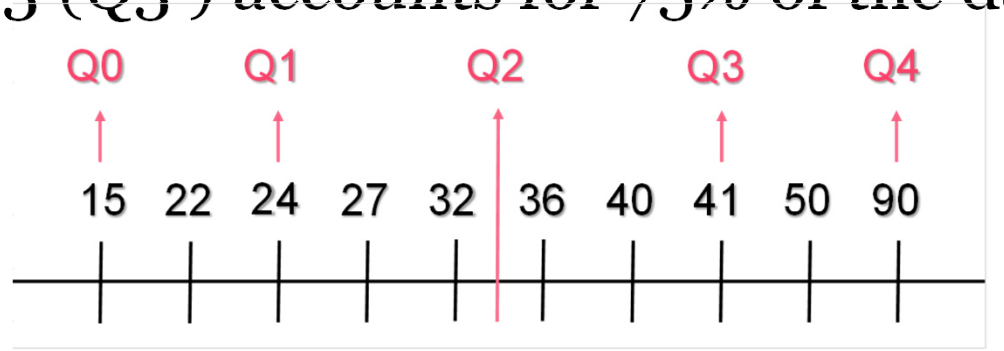
- used to identify the position of the observation in the data set.
- ***Percentile***, denoted as $P_x$ , is the value of the data at which x percentage of the data lie below that value.

$$\text{Position corresponding to } P_x \approx \frac{x(n+1)}{100}$$

where n is the number of observations in the data

- ***Decile*** corresponds to special values of percentile that divide the data into 10 equal parts. First decile contains first 10% of the data and second decile contains first 20% of the data and so on.

- ***Quartile*** divides the data into 4 equal parts. The first quartile (Q1 ) contains first 25% of the data, Q2 contains 50% of the data and is also the median. Quartile 3 (Q3 ) accounts for 75% of the data.



**EXAMPLE 2.1**

Time between failures (in hours) of a wire cutter used in a cookie manufacturing oven is given in Table 2.4. The function of the wire-cut is to cut the dough into cookies of desired size.

**TABLE 2.4** Time between failures of wire-cut (in hours)

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 2 | 22 | 32 | 39 | 46 | 56 | 76 | 79 | 88 | 93 |
| 3 | 24 | 33 | 44 | 46 | 66 | 77 | 79 | 89 | 99 |
| 5 | 24 | 34 | 45 | 47 | 67 | 77 | 86 | 89 | 99 |
| 9 | 26 | 37 | 45 | 55 | 67 | 78 | 86 | 89 | 99 |
| 21 | 31 | 39 | 46 | 56 | 75 | 78 | 87 | 90 | 102 |

# Answer the following:

- (a) Calculate the mean, median, and mode of time between failures of wire-cuts.
- (b) The company would like to know by what time 10% (ten percentile or P10) and 90% (ninety percentile or P90) of the wire-cuts will fail?
- (c) Calculate the values of P25 and P75.

# Solution

(a) Mean = 57.64, median = 56, and mode = 46, 89 and 99.

(b) The position of P10 = 10 × (51)/100 = 5.1. We can round off 5.1 to its nearest integer which is 5.

The corresponding value from table is 21 (10 percentage of observations in Table 2.4 have a value of less than or equal to 21). That is, by 21 hours, 10% of the wire-cuts will fail.

Instead of rounding the value, we can use the following approximation:
Position corresponding to P10 = 10 × (51)/100 = 5.1
Value at 5th position is 21.
Value at position 5.1 is approximated as
21 + 0.1 × (value at 6th position – value at 5th position) = 21 + 0.1(1) = 21.1

Position corresponding to P90 = 90 × 51/100 = 45.9
The value at position 45 is 90 and the value at position 45.9 is
90 + 0.9 (value at 46th position - value at 45th position) = 90 + 0.9 × (3) = 92.7
That is, 90% of the wire-cuts will fail by 92.7 hours.

# Solution

(c) Position corresponding to P25 (1st Quartile or Q1 ) = 25 × 51/100 = 12.75 Value at 12th position is 33, so P25 = 33 + 0.75 (value at 13th position – value at 12th position) = 33 + 0.75 (1) = 33.75

Position corresponding to P75 (3rd Quartile or Q3 ) = 75 × 51/100 = 38.25
Value at 38th position is 86, so P75 = 86 + 0.25 (value at 39th position – value at 38th position) = 86 + 0.25 (0) = 86

# MEASURES OF VARIATION

- to understand the variability in the data.

## RANGE

- Difference between maximum and minimum value of the data

## INTER-QUARTILE DISTANCE (IQD/IQR)

- distance between Quartile 1 (Q1 ) and Quartile 3 (Q3 )
- useful measure for identifying outliers in the data (Values below Q1 – 1.5 IQD & above Q3 + 1.5 IQD)

## VARIANCE

- variability in the data from the mean value

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

$$S^2 = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1}$$

## STANDARD DEVIATION

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

$$S = \sqrt{\sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1}}$$

# Possibility of measurements

| | Mean | Median | Mode | Standard Deviation | Ratio between two quantities |
|---|---|---|---|---|---|
| Nominal | ✘ | ✘ | ✓ | ✘ | ✘ |
| Ordinal | ✘ | ✘ | ✓ | ✘ | ✘ |
| Interval | ✓ | ✓ | ✓ | ✓ | ✘ |
| Ratio | ✓ | ✓ | ✓ | ✓ | ✓ |

# DATA VISUALIZATION

# Sampling & Estimation

- ***Sampling*** is the process of selecting subset of observations from a population to make inference about various population parameters such as mean, proportion, standard deviation, etc.

- It is an important step in inferential statistics since an incorrect sample may lead to wrong inference about the population.

- Sampling process itself has several steps and each step is important to ensure that the ideal sample is used for estimation of population parameters and for making inferences about the population.

- Under Big Data context, we may use almost the entire population; however, in most cases we will still be dependent on samples to make inference.

# Why and When to Sample?

- Sampling is necessary when it is difficult or expensive to collect data on the entire population.
- In many cases, all members of the population are not known for sampling purpose.
- Example1: A study on diabetic patients (64.5 million)
- Example2: Crash test of vehicles
- Incorrect sample may lead to incorrect inference about the population.

# Population Parameters and Sample Statistic

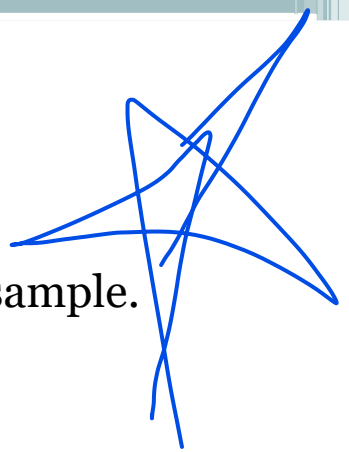
- Measures such as mean and standard deviation calculated using the entire population are called ***population parameters***. ($\mu$ , $\sigma$)

- Population parameters estimated from sample are called ***sample statistic*** or ***statistic***. ($\bar{X}$ , S)

# Sampling Process

1. Identification of target population that is important for a given problem under study
2. Decide the sampling frame
3. Determine the sample size
4. Sampling method (probabilistic / non-probabilistic)

# PROBABILISTIC SAMPLING



***1. Random Sampling*** (with/without replacement)

- ideal when the population is homogeneous.
- every subject in the population has equal probability of selection in the sample.

***2. Stratified Sampling***

- necessary when the population is heterogeneous
- population are divided into homogeneous groups called strata.

***3. Cluster Sampling***

- population is divided into mutually exclusive clusters.
- clusters are randomly selected and then all units within the selected clusters are included in the sample
- used when the clusters are large in number.

***4. Bootstrap Aggregating (Bagging)***

- sampling with replacement used in machine learning algorithms, especially the random forest algorithm

***5. Systematic Sampling***

- Sample elements are selected from the population at uniform intervals

# NON-PROBABILISTIC SAMPLING

***1. Convenience Sampling***

- not recommended since it is likely to result in biased estimates.

***2. Voluntary Sampling***

- data is collected from people who volunteer for such data collection.
- there could be bias

# PREDICTIVE ANALYTICS
# Linear Regression

# SIMPLE LINEAR REGRESSION

- Statistical technique for finding the existence of an association relationship between a dependent variable and an independent variable.

- Regression establishes existence of an association relationship between two variables, and not a causal relationship. The use of the term ‘dependent variable’ does not imply that the changes in the values of that variable are dependent on the changes in the values of the independent variable(s).

- Ex: India v/s England (test series) in 2014, studies & exam, price of apartment

- The dependent variable Yi is often known as response variable or outcome variable, and the independent variable Xi is also known as predictor variable or explanatory variable.

- Organizations use several key performance indicators (KPIs) such as cost of goods sold, customer lifetime value, growth rate, market share, productivity, profit, return on investment (ROI), etc. to measure their performance.

- KPIs, in turn, may be influenced by several factors.

- For example, the market share of a product sold by a company may be associated with factors such as:
    1. Price of the product
    2. Promotion expenses
    3. Competitors' price
    4. Competitor's promotion expenses
    5. New product introductions
    6. Macro-economic variables such as GDP, inflation, unemployment, and so on.

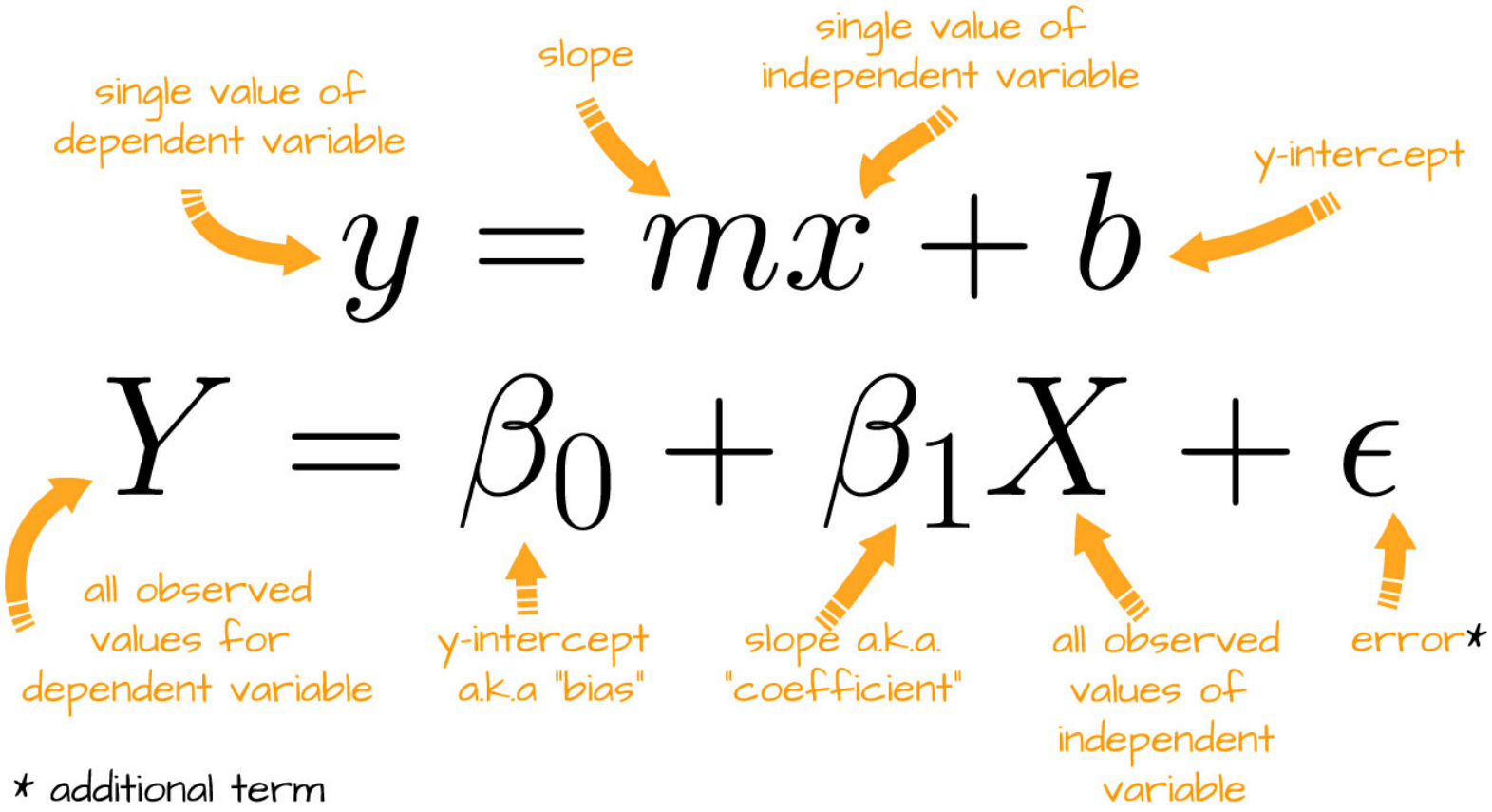
# Functional form of relationship

$$y=f(x)$$

$$Y = a + bX$$

(Sample)

# Functional form of relationship



Linear Regression: Single Variable

$$\hat{y} = \beta_0 + \beta_1 x + \epsilon$$

Predicted output — Coefficients — Input — Error

Linear Regression: Multiple Variables

$$\hat{y} = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \epsilon$$

# Testing your regression

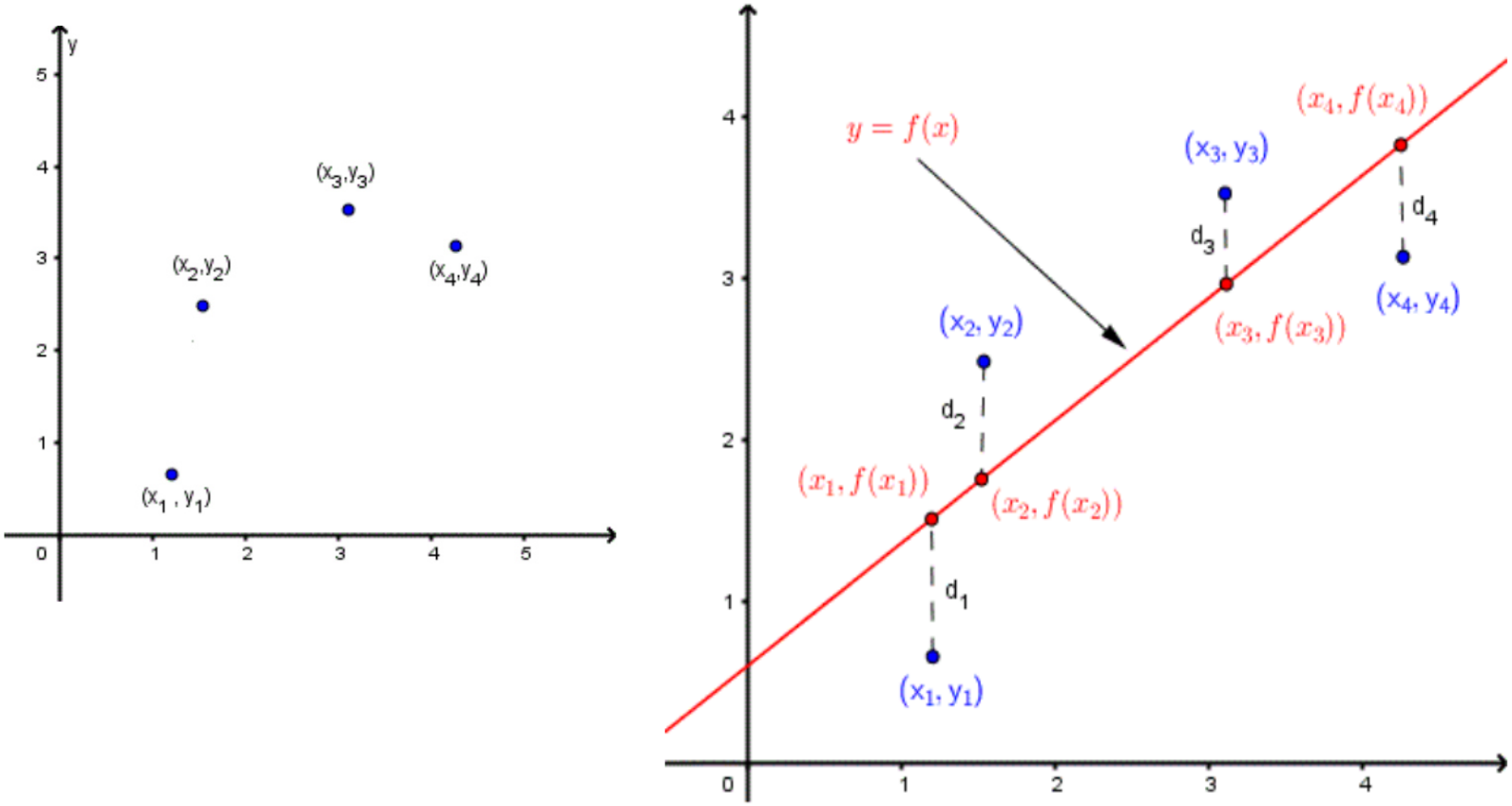
- does $y$ really depend on $x$?

$H_o: \beta = 0$

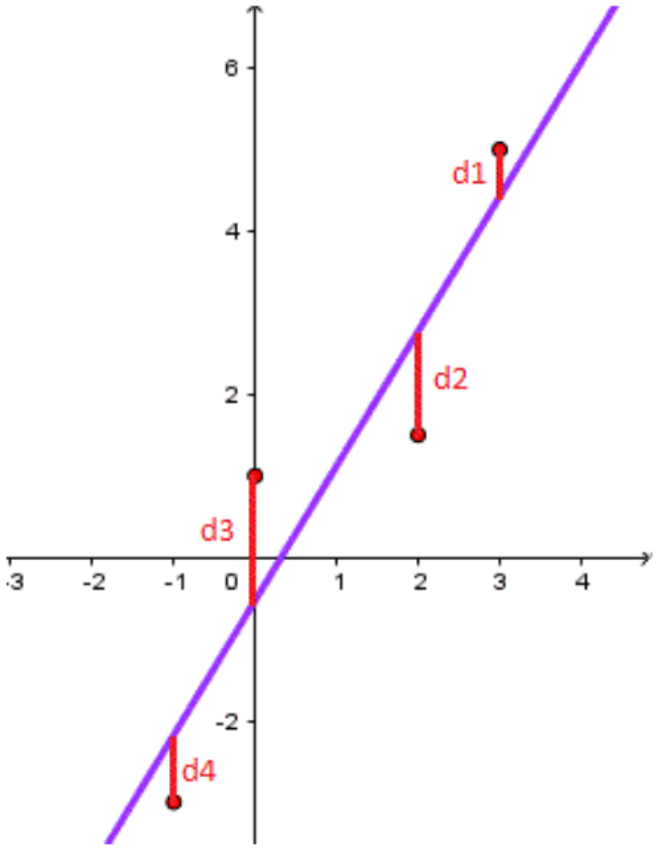
$H_a: \beta \neq 0$

- does this equation really help predict?

# Linear Least Squares Fitting



# Linear Least Squares Fitting




The least squares regression line is the line that minimizes the sum of the squares

**(d1 + d2 + d3 + d4)**

of the vertical deviation (between observed & predicted) from each data point to the line.

# Linear Least Squares Fitting

- The least square regression line for the set of n data points is given by the equation of a line in slope intercept form:

$$y = a\,x + b$$

Where **a** and **b** are given by:

$$a = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

$$b = \frac{1}{n}\left(\sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} x_i\right)$$

# Solve the following

**Problem 1**

- Consider the following set of points:

$$\{(-2 , -1) , (1 , 1) , (3 , 2)\}$$

a) Find the least square regression line for the given data points.

b) Plot the given points and the regression line in the same rectangular system of axes.

# Solve the following

**Problem 2**

The values of y and their corresponding values of y are shown in the table below

| x | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| y | 2 | 3 | 5 | 4 | 6 |

a) Find the least square regression line

$y = a x + b$.

b) Estimate the value of y when x = 10.

# Solve the following

Find linear regression equation for the following data:

| | | | | |
|---|---|---|---|---|
| x | 2 | 4 | 6 | 8 |
| y | 3 | 7 | 5 | 10 |

# FORECASTING TECHNIQUES

# Introduction to forecasting

- Every organization prepares long-range and short-range planning for the organization and forecasting demand for product and service is an important input for both.

- Boeing 747-400
- Amazon.com
- Walmart
- Manpower planning

# Demand Forecasting Objectives

- Financial planning
- Pricing Policy
- Manufacturing policy
- Sales
- Marketing planning
- Capacity planning
- Expansion
- Labor planning
- Capital Expenditure.

# TIME-SERIES DATA

- Time-series data is a data on a response variable observed at different time points t. Whenever we have a forecasting problem, we will be using a time-series data.

- univariate & multi-variate time series

# COMPONENTS OF TIME-SERIES DATA

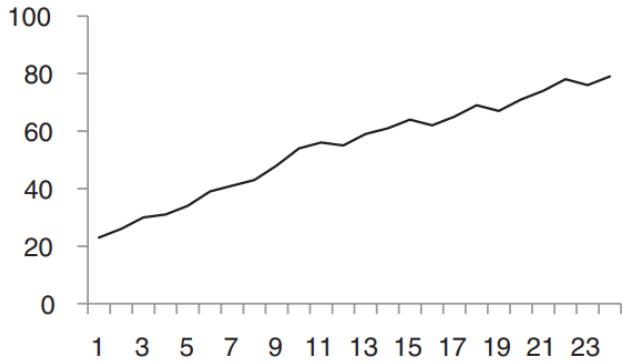
1. **Trend Component**: Trend is the consistent long-term upward or downward movement of the data over a period of time.

2. **Seasonal Component**: Repetitive upward or downward movement (or fluctuations) from the trend that occurs within a calendar year such as seasons, quarters, months, days of the week, etc.

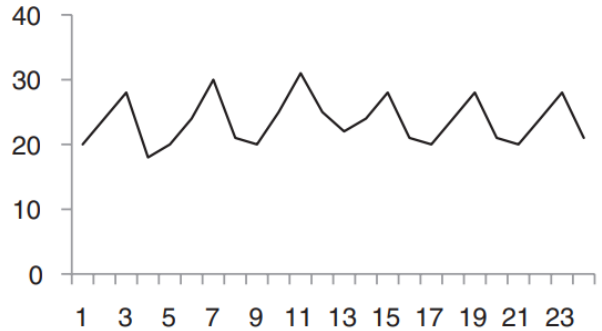
3. **Cyclical Component**: Fluctuation around the trend line that happens due to macro-economic changes such as recession, unemployment, etc. Cyclical fluctuations have repetitive patterns with a time between repetitions of more than a year.

4. **Irregular Component** : Random uncorrelated changes

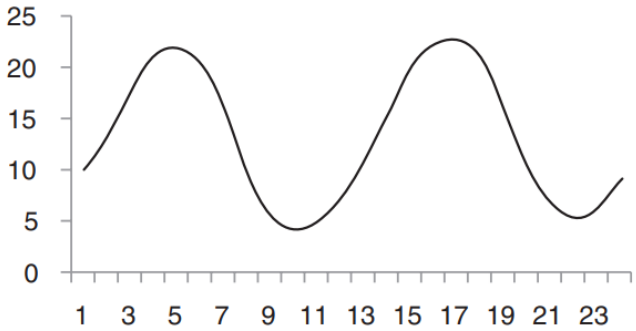
# COMPONENTS OF TIME-SERIES DATA



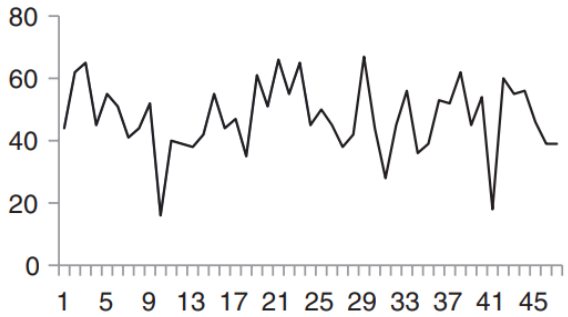

(a) Trend




(b) Seasonality (fixed periodicity)




(c) Cyclicality




(d) Irregular

# FORECASTING TECHNIQUES

1. Naïve Approach
2. Moving Average
3. Weighted Moving Average
4. Exponential Smoothing

# Forecasting Accuracy Measures

1. Mean absolute error
2. Mean absolute percentage error
3. Mean squared error
4. Root mean square error

# Mean Absolute Error (MAE)

$$MAE = \sum_{t=1}^{n} \frac{\left| Y_t - F_t \right|}{n}$$

$Y_t$ is the actual value of Y at time t
$F_t$ is the corresponding forecasted value
n is the number of observations

# Mean Absolute Percentage Error (MAPE)

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\frac{\left|Y_t - F_t\right|}{Y_t} \times 100\%$$

$Y_t$ is the actual value of Y at time t
$F_t$ is the corresponding forecasted value
n is the number of observations

# Mean Square Error (MSE)

$$MSE = \frac{1}{n}\sum_{t=1}^{n}\left(Y_t - F_t\right)^2$$

$Y_t$ is the actual value of Y at time t
$F_t$ is the corresponding forecasted value
n is the number of observations

# Root Mean Square Error (RMSE)

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(Y_t - F_t\right)^2}$$

$Y_t$ is the actual value of Y at time t
$F_t$ is the corresponding forecasted value
n is the number of observations

# MOVING AVERAGE METHOD

Of the past 'N' observations =

$$F_{t+1} = \frac{1}{N} \sum_{k=t+1-N}^{t} Y_k$$

**Weighted** moving average is given by

$$F_{t+1} = \sum_{k=t+1-N}^{t} W_k \times Y_k$$

# Solve

| year | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Sales | 5.2 | 4.9 | 5.5 | 4.9 | 5.2 | 5.7 | 5.4 | 5.8 | 5.9 | 6 | 5.2 | 4.8 |

Calculate 5 year Simple Moving Average forecast

**Also calculate the forecasting errors (MAE, MSE, RMSE and MAPE)**

| (1) year | (2) Sales | (3) 5 year moving average |
|---|---|---|
| 1 | 5.2 | |
| 2 | 4.9 | |
| 3 | 5.5 | |
| 4 | 4.9 | |
| 5 | 5.2 | |
| 6 | 5.7 | 5.14 |
| 7 | 5.4 | 5.24 |
| 8 | 5.8 | 5.34 |
| 9 | 5.9 | 5.4 |
| 10 | 6 | 5.6 |
| 11 | 5.2 | 5.76 |
| 12 | 4.8 | 5.66 |
| 13 | | 5.54 |

# PRESCRIPTIVE ANALYTICS

# Prescriptive Analytics

- ***Prescriptive analytics*** provides the optimal solution (or the best action) to a problem. Traditionally, Operations Research (OR) techniques are used for finding the optimal solution to a problem.

# Prescriptive Analytics Problems in Business

1. **Travelling salesman problem** which occurs in industries such as e-commerce (delivery of goods to customers), electronics parts manufacturing (movement of robot during manufacturing), and logistics service providers 

2. **Airlines:** Assigning an aircraft to a flight subject to several constraints (such as connectivity, minimum ground time, runway restrictions, etc.). Scheduling maintenance of aircraft for various maintenance checks in an optimal manner is a complex optimization.

3. **Water distribution system** in every city requires optimal design of water pipe network which is a complex prescriptive analytics problem.

4. **Finding optimal location** for utilities such as water tanks, fire stations, mobile towers, police stations are complex decision problems which are solved using prescriptive analytics techniques

5. **Retails stores**

6. **Markdown optimization**

7. **Scheduling** such as nurse scheduling and airline crew. Hospital with 250 beds, >1500 nurses of varying skills , 3 shifts. The hospital has to schedule nurses during each shift satisfying several constraint to optimize an objective function.

# Prescriptive Analytics Problems

**1. Product Mix Problem:** to decide the number of different products to be produced using common resources

**2. Blending Problem**: mixing various ingredients to optimize yield of various end products and objective (such as profit)

**3. Cutting Stock Problem**: objective is to minimize the loss.

**4. Transportation Problem**: objective is to minimize the cost of transportation of goods from multiple origins (production centers) to several destinations (consumption centers)

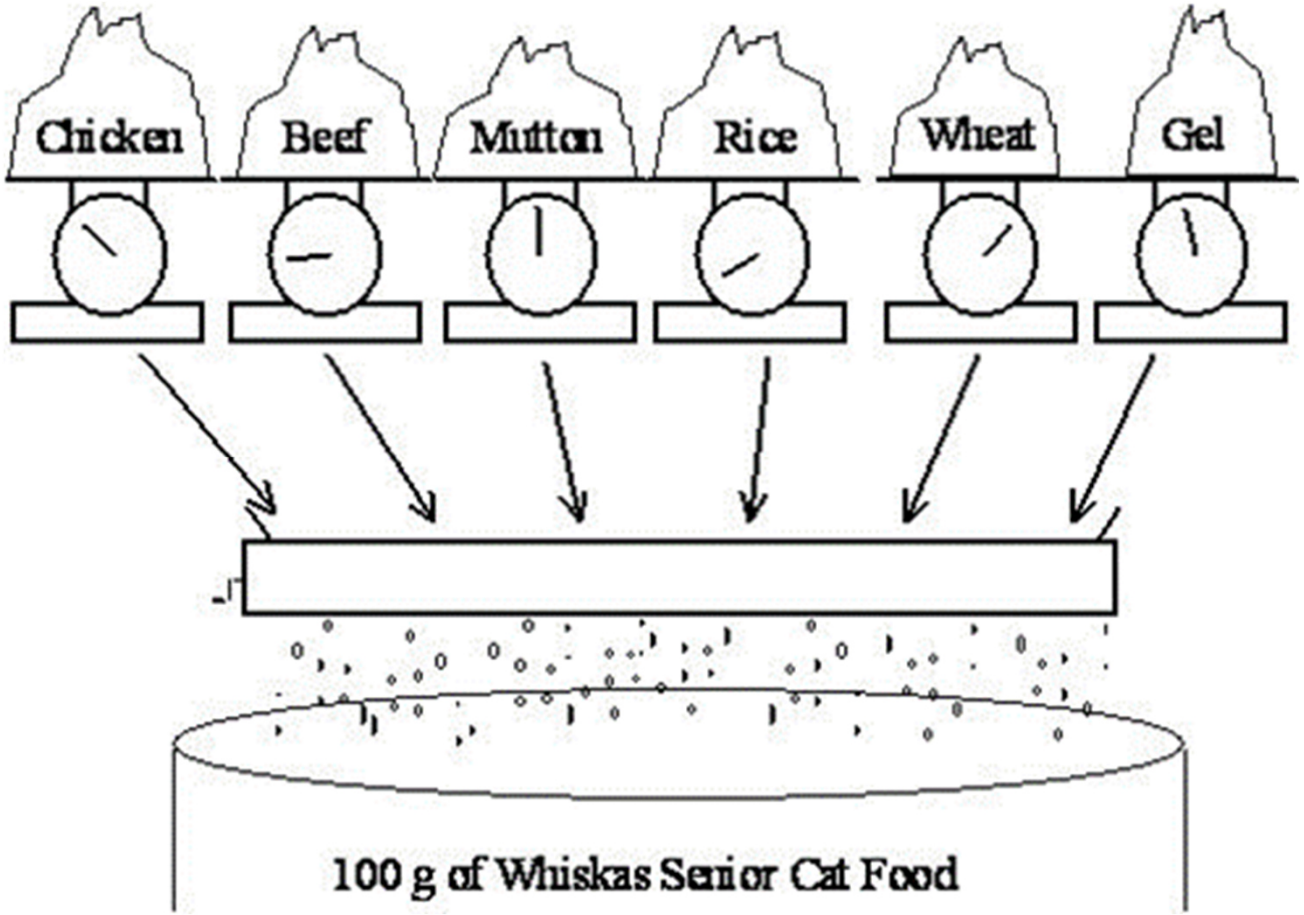
**5. Assignment Problem**: to assign task among agents that minimize the total assignment cost.

**6. Location Problem**: optimal location of facilities to optimize objectives such as median (total distance) or minimize the maximum distance

**7. Set Covering Problem**: objective is to identify a subset from a set of elements such that the subset will cover the entire problem under given conditions.

# Product Mix (Example)

# Blending problem (example)

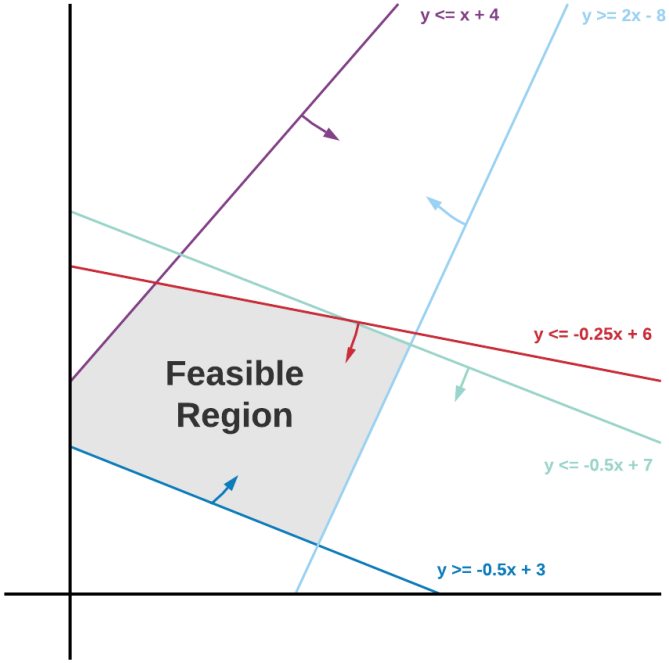


# Techniques

- Linear programming
- Integer programming
- Goal programming
- Non-linear programming
- Meta-heuristic

# Linear Programming

- used when the objective function and the constraints of the problem can be expressed as linear equation of decision variables



# LINEAR PROGRAMMING (LP) MODEL BUILDING

The following steps are used in formulating a problem as LPP:

1. **Identification of decision variables**
2. **Identify the objective function:** a linear function of decision variables. The goal is either to minimize or maximize the objective function value.
3. **Identify constraints**: Restrictions such as availability of resources that a linear programming problem should satisfy.
4. **Identify implicit constraints**: Conditions that the model has to satisfy
5. **Solve the problem**
6. **Perform sensitivity analysis:** The values of objective function coefficients and resource availability may change due to several factors such as market conditions. It is important to understand the impact of the changes in objective function coefficient and resource availability on the optimal solution; this is achieved through sensitivity analysis.

# Example

## Product Mix

Appukuttan Menon is the co-founder and CEO of Appukuttan Halva (AH) with headquarters in Kuttanad, Kerala. AH manufactures two types of halva: (a) Death by Halva (DH) and (b) Travancore Halva (TH). The main ingredients of halva are: (a) corn flour, (b) sugar, (c) fruit and nut, and (d) ghee. The quantity of each ingredient required for the two types of halva for every one kilogram is given in Table 15.1.

**TABLE 15.1** Ingredients required for 1 kg of halva

| Halva Type | Ingredients (in grams) Required for 1 kg of Halva | | | |
|---|---|---|---|---|
| | Corn Flour | Sugar | Fruit and Nut | Ghee |
| Death by Halva (DH) | 500 | 750 | 150 | 200 |
| Travancore Halva (TH) | 500 | 625 | 100 | 300 |

# Example

The profit from DH and TH per kilogram are INR 45 and 50, respectively. The maximum daily demand for DH and TH are 50 kg and 20 kg, respectively. Appukuttan Menon is a big fan of the Japanese lean management concept and used JIT procurement. All the ingredients necessary for the daily production are delivered on the day of production at 6.00 am and AH maintained no safety stock. The DH and TH are delivered to the customers (local retail stores in Kuttanad) from 12.00 Noon onwards every day. The suppliers of the ingredients are located in Coimbatore, which is about 300 km from Kuttanad.

Due to some supply chain disturbance, suppliers of AH have informed Appukuttan Menon that they will be unable to supply the raw material on 25$^{th}$ January 2017; however, the supply will be restored from 26$^{th}$ January onwards. To manage the supply of halva on 25$^{th}$ January 2017, Appukuttan Menon decided to procure the ingredients locally. His procurement manager George Varghese informed him that since AH uses specific brands of the ingredients, the availability of raw material is limited in the local market and is shown in Table 15.2.

# Example

**TABLE 15.2** Availability (in grams) of ingredients in local market

| Corn Flour | Sugar | Fruit and Nut | Ghee |
|---|---|---|---|
| 20,000 | 42,000 | 10,400 | 9,600 |

Use linear programming to find the optimal product mix for 25th January for AH that will maximize the profit for AH? Assume that there will be no change in the profit of DH and AH due to procurement of ingredients from the local market.

**Solution:**

**STEP 1** *Identification of the decision variables*

In this case, AH has to decide the quantity (in kilograms) of DH and TH to be produced. Let

$X_1$ = Quantity (in kilogram) of DH to be produced

$X_2$ = Quantity (in kilogram) of TH to be produced

**STEP 2** *Identification of the objective function*

The objective is to maximize the profit. The profit on DH per kg is 45 and the profit on TH per kg is 50. The objective function is

Maximize $45X_1 + 50X_2$

**STEP 3** *Identify the constraints*

In this example, the constraints are availability of various ingredients.

**Constraint for corn flour:** 20,000 grams of corn flour is available. Each kg of AH requires 500 grams of corn flour and each kg of TH 500 grams of corn flour. Thus, the corresponding constraint is

$$500\ X_1 + 500\ X_2 \leq 20{,}000$$

**Constraint for sugar:** 42,000 kg of sugar is available. Each kg of DH requires 750 grams of sugar and each kg of TH 625 grams of sugar. Thus the corresponding constraint is

$$750\ X_1 + 625\ X_2 \leq 42{,}000$$

**Constraint for fruit and nut:** 10,400 kg of fruit and nut is available. Each kg of DH requires 150 grams of fruit and nut and each kg of TH requires 100 grams of fruit and nut. Thus the corresponding constraint is

$$150X_1 + 100X_2 \leq 10{,}400$$

**Constraint for ghee:** 9,600 kg of ghee is available. Each kg of DH requires 200 grams of ghee and each kg of TH requires 300 grams of ghee. Thus the corresponding constraint is

$$200X_1 + 300X_2 \leq 9{,}600$$

**Maximum demand constraint:** The maximum daily demand for DH and TH are 50 kg and 20 kg, respectively, which can be written as $X_1 \leq 50$ and $X_2 \leq 20$.

**STEP 4** *Identify implicit constraints*

In this case the implicit constraints are the quantity of DH and TH which cannot be negative. Thus the values $X_1$ and $X_2$ are non-negative. That is, $X_1 \geq 0$ and $X_2 \geq 0$.

The complete LP formulation is

$$\text{Maximize } 45X_1 + 50X_2$$

subject to constraints

$$500X_1 + 500X_2 \leq 20{,}000$$

$$750X_1 + 625X_2 \leq 42{,}000$$

$$150X_1 + 100X_2 \leq 10{,}400$$

$$200X_1 + 300X_2 \leq 9{,}600$$

$$X_1 \leq 50$$

$$X_2 \leq 20$$

$$X_1 \geq 0 \text{ and } X_2 \geq 0$$

A company buying scrap metal has two types of scrap metal available to him. The first type of scrap metal has 30% of metal A, 20% of metal B and 50% of metal C by weight. The second scrap has 40% of metal A, 10% of metal B and 30% of metal C. The company requires at least 240 kg. of metal A, 100 kg. of metal B and 290 kg. of metal C. The price per kg. of the two scraps are Rs. 120 and Rs. 160 respectively. 'Determine the optimum quantities of the two scraps to be purchased so that the requirements of the three metals are satisfied at a minimum cost.

A company owns two flour mills, A and B, which have different production capacities, for high, medium and low grade flour. This company has entered a contract to supply flour to a firm every week with at least 12, 8 and 24 quintals of high, medium and. low grade respectively. It costs the company Rs. 1000 and Rs. 800 per day to run mill A and B respectively. On a day, mill A produces 6, 2 and 4 quintals of high, medium and low grade flour respectively. Mill B produces 2, 2 and 12 quintals of high, medium and low grade flour respectively. How many days per week should each mill be operated in order. to meet the contract order most economically.

A company buying scrap metal has two types of scrap available to them. The first type of scrap metal has 20% of metal A, 10% of impurity and 20% of metal by weight. The second type of scrap has 30% of metal A,10% of impurity and 15% of metal B by weight. The company requires at least 120 kg. of metal A, at most 40 kg. of impurity and at least 90 kg. of metal B. The price for the two scraps are Rs. 200 and Rs. 300 per kg. respectively. Determine the optimum quantities of the two scraps to be purchased so that the requirements of the two metals and the restriction on impurity are satisfied at minimum cost.

# Solve the following LPP using Graphical Method

1 The Consumer Product Corporation wishes to plan its advertising strategy. There are 'two magazines under consideration, magazine I and magazine II. Magazine I has a reach of 2000 potential customers per advertisement and magazine II has a reach of 3000 potential customers per advertisement, The cost per advertising is Rs. 6000 and Rs. 9000 in magazines I and II respectively and the firm has a monthly budget of Its, 1 lakh. There is an important requirement that the total reach for the income group under Its 20000 per annum should not exceed 3000 potential customers. The reach of magazine I and II for this income group is 300 and 150 potential customers respectively per advertisement. How many times the company should advertise in the two magazines to maximise the total reach?

2 A company owns two flour mills, A and B, which have different production capacities, for high, medium and low grade flour. This company has entered a contract to supply flour to a firm every week with at least 12, 8 and 24 quintals of high, medium and. low grade respectively. It costs the company Rs. 1000 and Rs. 800 per day to run mill A and B respectively. On a day, mill A produces 6, 2 and 4 quintals of high, medium and low grade flour respectively. Mill B produces 2, 2 and 12 quintals of high, medium and low grade flour respectively. How many days per week should each mill be operated in order. to meet the contract order most economically.

# CONFIDENCE INTERVALS

# Interval Estimate

- An interval estimate of a population parameter such as mean and standard deviation is an interval or range of values within which the true parameter value is likely to lie with certain probability.

- For example, confidence interval for population mean may be stated as

$$\mathbf{30 \leq \mu \leq 50}$$

(that is, the population mean lies between values 30 and 50)

# Confidence level

- Confidence level, usually written as

$$\mathbf{(1 - \alpha)100\%}$$

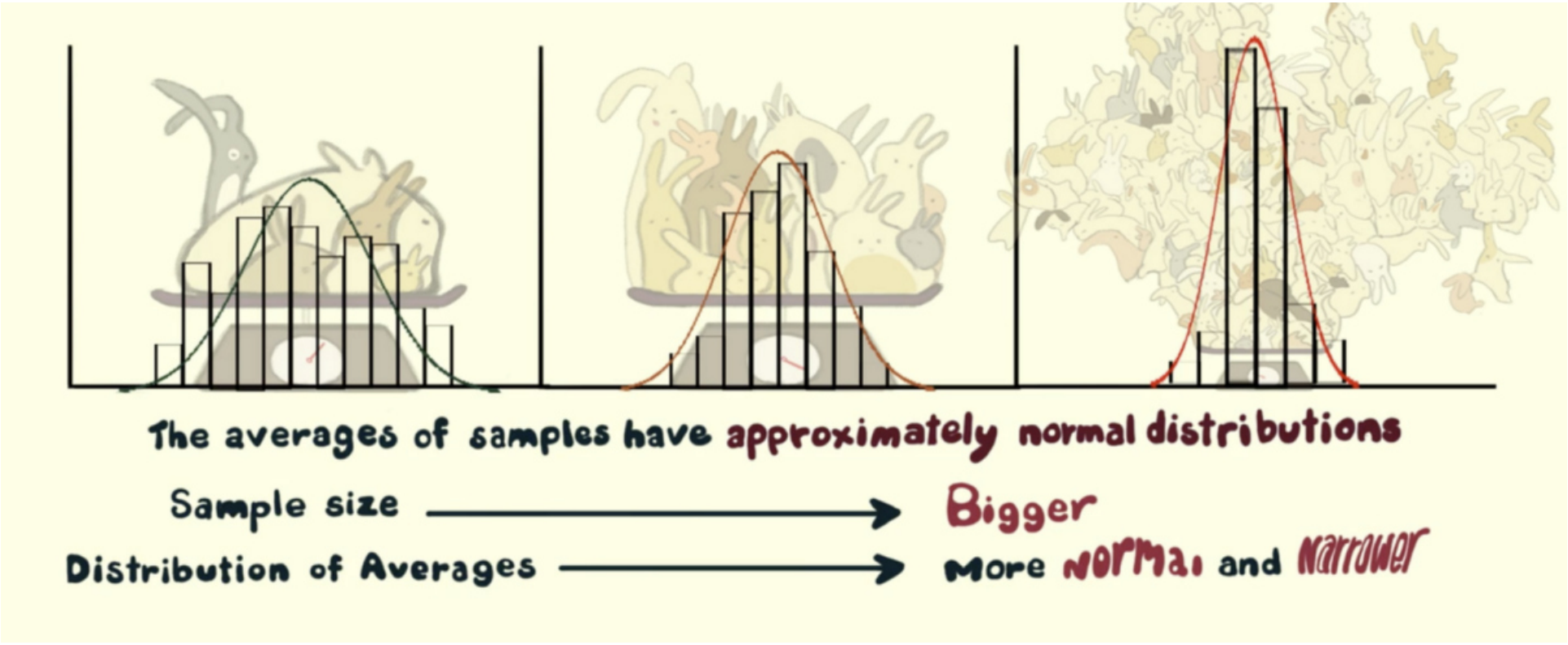
on the interval estimate of a population parameter is the ***probability*** that the interval estimate will contain the true population parameter.

- When (significance) $\boldsymbol{\alpha}$ = 0.05, 95% is the confidence level and 0.95 is the probability that the interval estimate will have the population parameter.

- The choice of $\boldsymbol{\alpha}$ depends on the context of the problem

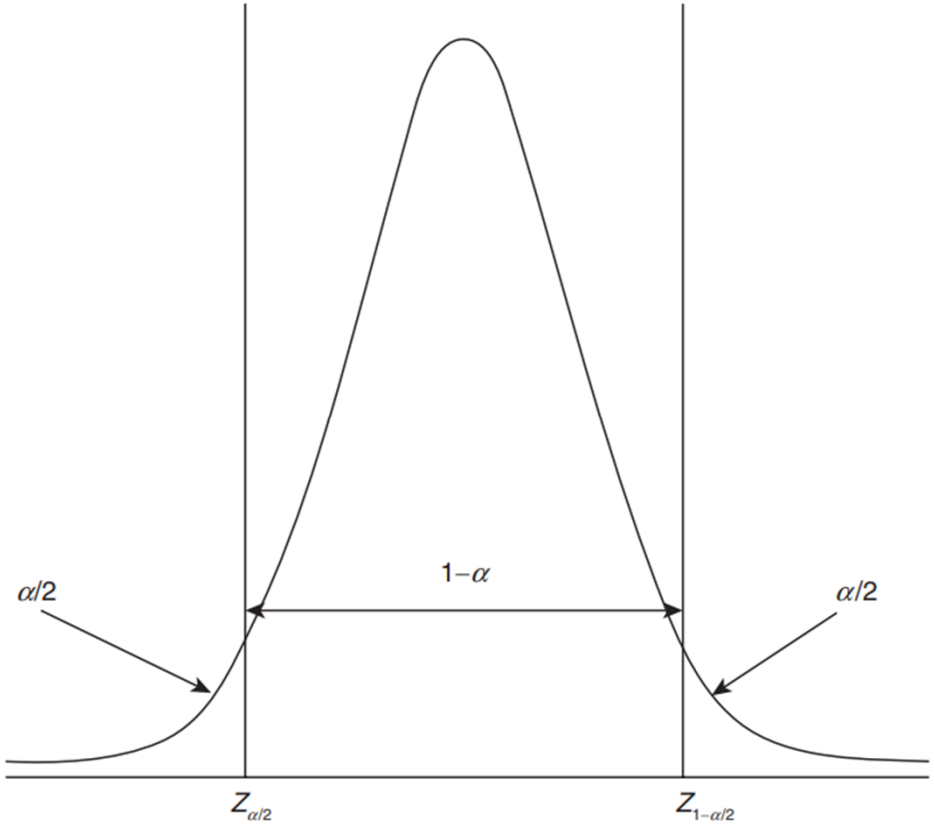
# Central Limit Theorem (CLT)

- For a large sample drawn from a population with mean **μ** and standard deviation **σ**, the sampling distribution of mean, X, follows an approximate normal distribution with mean **μ** and standard deviation (standard error) **σ** / $\sqrt{n}$ irrespective of the distribution of the population.



# Central Limit Theorem (CLT)

# CONFIDENCE INTERVAL FOR POPULATION MEAN



# SURVEY DATA ANALYTICS PROJECT

# Steps:

1. **Ask a question**
2. **Formulate your hypothesis**
3. **Identify variables.**
4. **Collect Survey Data**
5. **Analyze Survey Data**
6. **(What Do You Want To Know? What are the Variables used? Which Patterns Stand Out Is The Data Reliable? (statistical significance))**
7. **Statistical test (Variance, SD, Z-score, Confidence level, hypothesis, t-test)**
8. **Results & Implications**

# Hypothesis examples

| Research question | Hypothesis | Null hypothesis |
|---|---|---|
| What are the health benefits of eating an apple a day? | Increasing apple consumption in over-60s will result in decreasing frequency of doctor's visits. | Increasing apple consumption in over-60s will have no effect on frequency of doctor's visits. |
| Which airlines have the most delays? | Low-cost airlines are more likely to have delays than premium airlines. | Low-cost and premium airlines are equally likely to have delays. |
| Can flexible work arrangements improve job satisfaction? | Employees who have flexible working hours will report greater job satisfaction than employees who work fixed hours. | There is no relationship between working hour flexibility and job satisfaction. |
| How effective is high school sex education at reducing teen pregnancies? | Teenagers who received sex education lessons throughout high school will have lower rates of unplanned pregnancy than teenagers who did not receive any sex education. | High school sex education has no effect on teen pregnancy rates. |
| What effect does daily use of social media have on the attention span of under-16s? | There is a negative correlation between time spent on social media and attention span in under-16s. | There is no relationship between social media use and attention span in under-16s. |

# Key Survey Analysis Variables

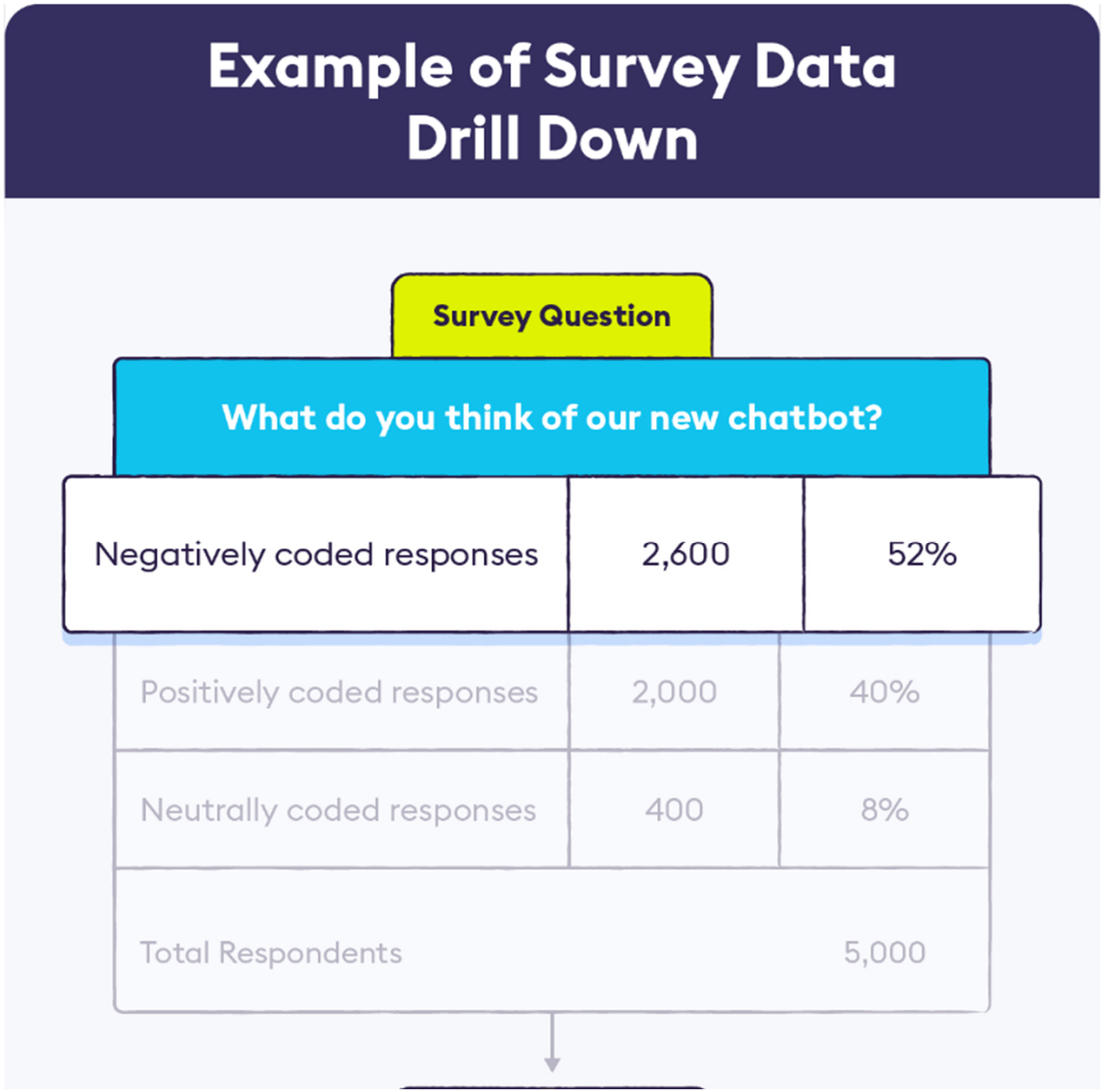
# How To Collect Survey Data

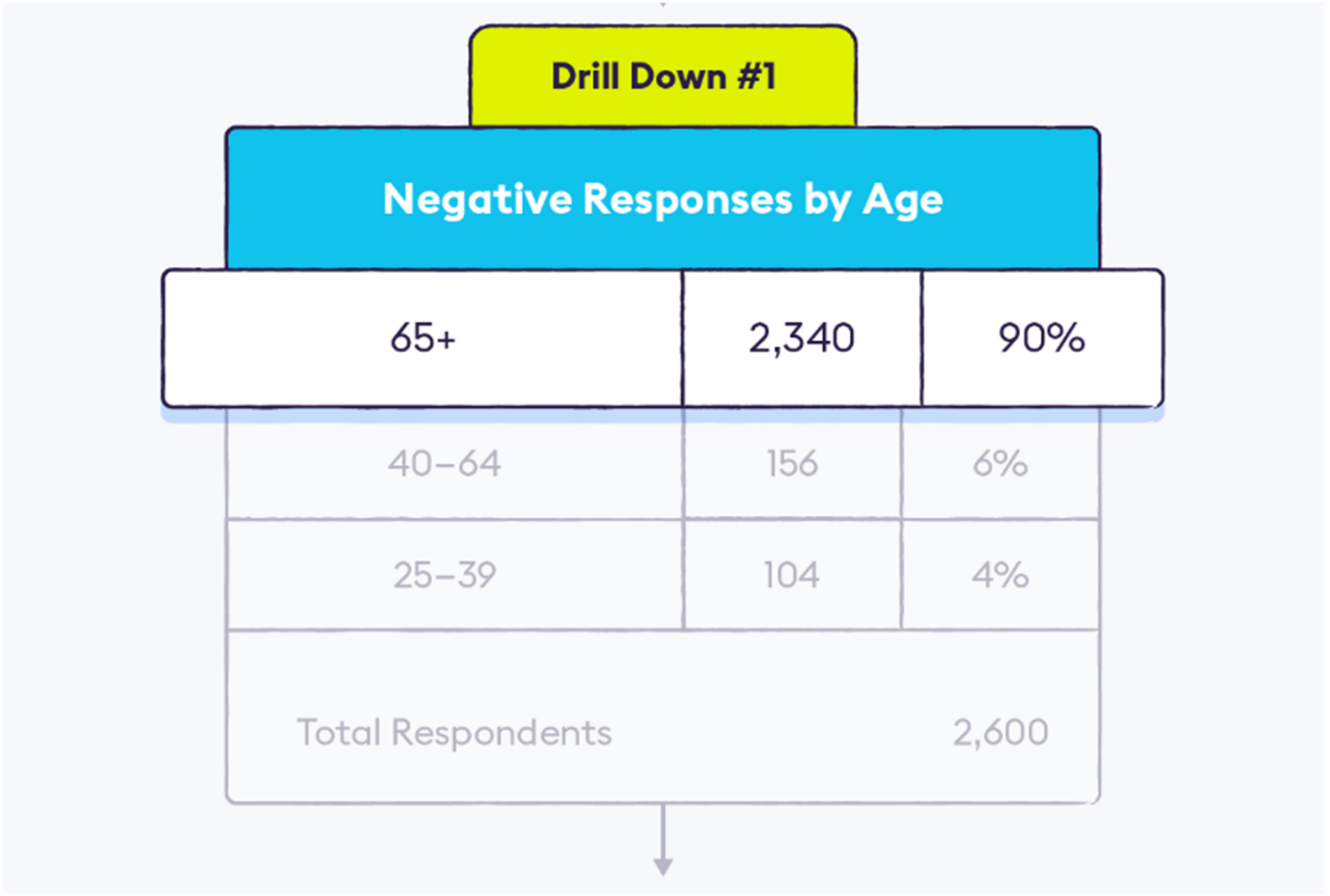
- **Consistent Metrics**
- **Different Descriptors**
- **Numerical Scales**
- **Freeform vs Multiple-Choice**
- **Open vs Closed Questions**
- **Other Common Pitfalls** (asking leading questions, surveys that are too long, over-surveying people, and starting with an already-biased audience)

# Coding Qualitative Data

| Customer Response | Positive vs. Negative | Coding (1–5) |
|---|---|---|
| Amazing customer service team, helped me change my order last minute and it still shipped on time. | Very Positive | 5/5 |
| Customer service is a bit slow, but we got there in the end. | Neutral | 3/5 |
| Don't try to contact customer service, they're nice but they don't know what they're doing. | Negative | 2/5 |

# How To Analyze Survey Data




Drill Down #1
Negative Responses by Age
65+
2,340
90%
40–64
156
6%
25–39
104
4%
Total Respondents
2,600

## Drill Down #2

### Negative Responses for 65+ by Income

| | | |
|---|---|---|
| $200K+ | 2,106 | 90% |
| $100K–199K | 93 | 4% |
| Less than 100K | 23 | 1% |
| Total Respondents | | 2,340 |

## Conclusion

Wealthy, older customers
**do not like your chatbot.**

**Note that you will need to prove this number is statistically significant**

# Is The Data Reliable?

- statistical significance
- demographic spread

## Population

The **total number of people**. For example, everyone in your customer base.

## Sample

The **part of the total population** that you survey. Used to make assumptions about the whole.

# HYPOTHESIS TESTING