Descriptive Analytics

Chetan Chitre

Course Outline

Sl No	Topics	Sub-topics	Number of Sessions	Software	Faculty
1	Introduction to Data Science and Business Analytics		2		Prof Mahima
2	Basic Statistical Measures	Visualizing and describing categorical and numerical data, Measures of central tendency and dispersion, Measures for population and sample data, Measures for data with more than one variable	2	Excel Toolpack	Prof Chetan
3	Probability and Random variables	Basic probability, conditional probability, Bayes' Theorem	2	Excel Toolpack	Prof Chetan
4	Discrete & Continuous Distributions	Binomial, Poisson,Normal distributions, Sampling distributions	5	Excel Toolpack	Prof Chetan
5	Hypothesis Testing	Confidence Intervals and Hypothesis testing with known and unknown population characteristics, Mean, Proportion and Variance	5	Excel Toolpack	Prof Chetan
6	Analysis of Variance	T-tests, 2 sample t-tests, One-way and Two-way ANOVA	6	Excel Toolpack	Prof Chetan

Course Outline

- **Recommended Reading:** Levine, D. M., Stephan, D. F., Szabat, K. A. (2017). Statistics for Managers Using Microsoft Excel. Pearson Education India.
- **Evaluation:** Group Assignment (20) and End Term Exam (30)

Business Decision

A well-known coffee brand from south India wants to enter the North Indian market. The CEO believes in data driven decision making. How will you go about it?

Terminology

- Population
- Sample
- Parameter
- Statistic
- Descriptive statistic
- Inferential statistic
- Levels of data:
  - Nominal
  - Ordinal
  - Interval
  - Ratio

Identify the Variable Levels

Because of increased competition for patients among providers and the need to determine how providers can better serve their clientele, hospital administrators sometimes administer a quality satisfaction survey to their patients after the patient is released. The following types of questions are sometimes asked on such a survey. These questions will result in what level of data measurement?

1. How long ago were you released from the hospital?
2. Which type of unit were you in for most of your stay?

   _____Coronary care

   _____Intensive care

   _____Maternity care

   _____Medical unit

   _____Pediatric/children's unit

   _____Surgical unit

3. In choosing a hospital, how important was the hospital's location?

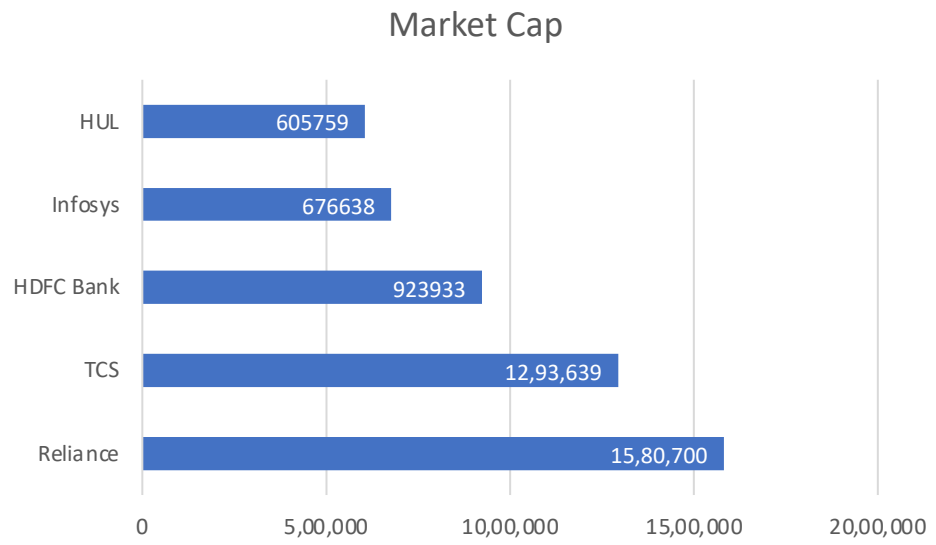
(circle one)			
Very Important	Somewhat Important	Not Very Important	Not at All Important

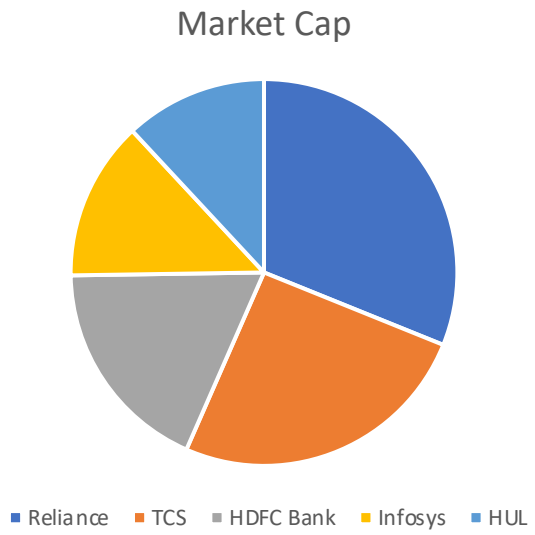
4. What was your body temperature when you were admitted to the hospital?
5. Rate the skill of your doctor:

(1) Excellent (2) Very Good (3) Good (4) Fair (5) Poor

Data Visualization

Company Name	Market Cap (in Rs. crore)
Reliance	15,80,700
TCS	12,93,639
HDFC Bank	923933
Infosys	676638
HUL	605759





Questions

Classify each of the following as nominal, ordinal, interval, or ratio data.

- The time required to produce each tire on an assembly line
- The number of quarts of milk a family drinks in a month
- The ranking of four machines in your plant after they have been designated as excellent, good, satisfactory, or poor
- The telephone area code of clients in the United States
- The age of each of your employees
- The dollar sales at the local pizza shop each month
- An employee's identification number
- The response time of an emergency unit

Questions

The Wiro Manufacturing Company makes electric wiring, which it sells to contractors in the construction industry. Approximately 900 electric contractors purchase wire from Wiro annually. Wiro's director of marketing wants to determine electric contractors' satisfaction with Wiro's wire. He developed a questionnaire that yields a satisfaction score between 10 and 50 for participant responses. A random sample of 35 of the 900 contractors is asked to complete a satisfaction survey. The satisfaction scores for the 35 participants are averaged to produce a mean satisfaction score.

- What is the population for this study?
- What is the sample for this study?
- What is the statistic for this study?
- What would be a parameter for this study?

Describing Data

- Frequency
- Frequency polygons
- Ogive

Describing Data

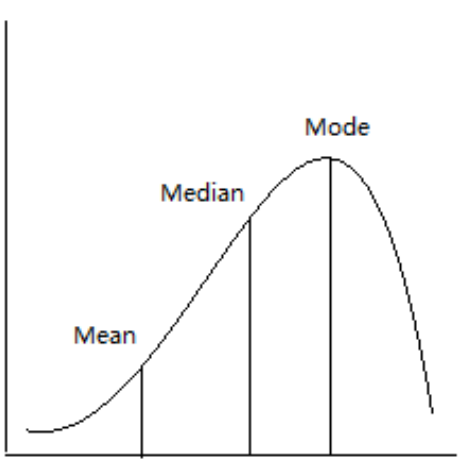
- Data with 2 variables

Locating Data

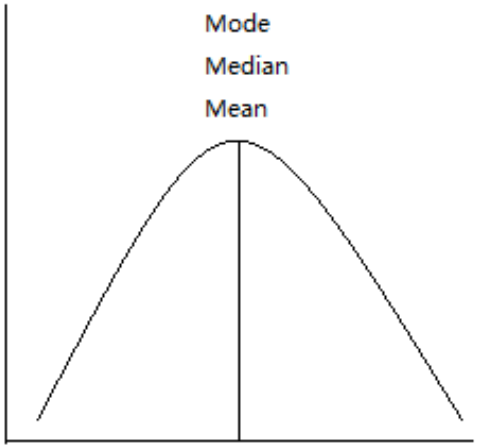
- Measures of Central Tendency
  - Mode:
    - Most frequently occurring number
    - Data could be bimodal or multimodal
    - Used when categories are important
  - Median:
    - Arrange data in ascending order
    - For odd N → entry no. (N+1)/2
    - For even N → Average of two central entries i.e. Average of N/2 and (N/2) + 1
    - Distribution of data does not affect Median
  - Mean
    - $(x_1 + x_2 + x_3 + ... + x_N)/N$
    - Affected by distribution of data and extreme values

Locating Data

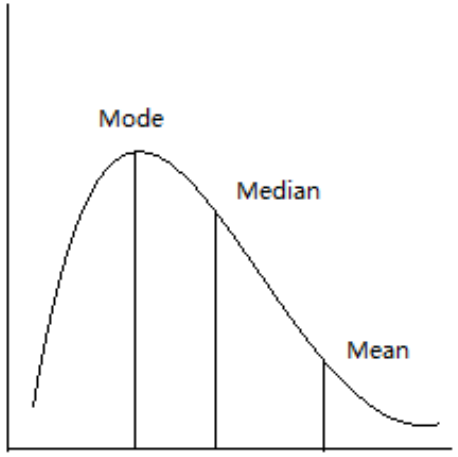
Impact of Distribution on Measures of Central Tendency



Negatively Skewed
Left Skew



Normal
No Skew



Positively Skewed
Right Skew

Check the location of Mean, Median and Mode in Employment data

Locating Data

- Percentiles: n^{th} percentile is the value such that at least n% of data are below that value
  - Arrange data in ascending order
  - Percentile location
    - $j = (P/100)N$
    - j = Percentile location
    - P = Percentile of interest
    - N – Number of observations
  - If j is a whole number, the P^{th} percentile is the average of j^{th} and $(j+1)^{th}$ entry
  - If j is not a whole number P^{th} percentile is located at the whole number part of (j+1) eg. If j = 2.3 → (j+1) = 3.3 → P^{th} percentile is the value located at the 3^{rd} entry
- Quartiles are specific cases of percentiles
- Find the 25^{th}, 50^{th} and 75^{th} quartile for examination score data

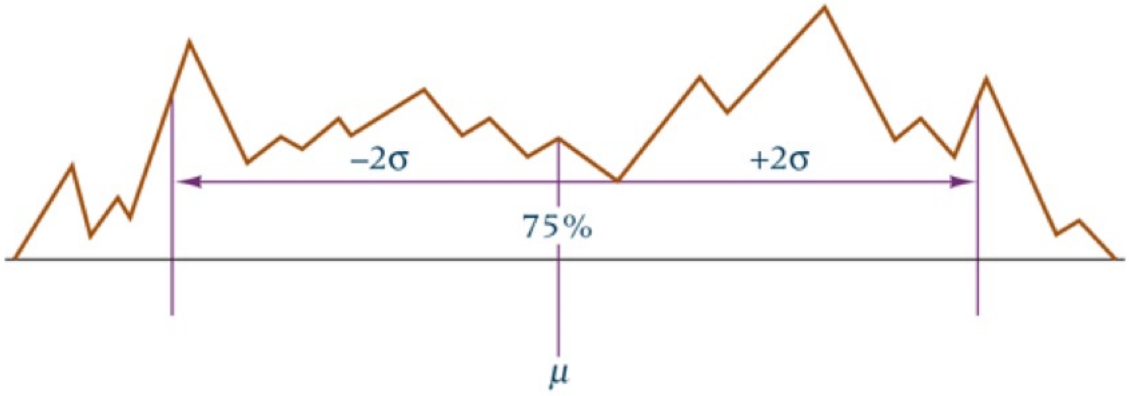
Dispersion of Data

- Range = Highest – Lowest
- Interquartile range: Range of the middle 50% of data ($Q_3 - Q_1$)
- Population Variance: $\sigma^2 = \frac{\sum(x_i - \mu)^2}{N}$
- Population Standard Deviation: $\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$
- Importance of Standard Deviation:
  - It is generally noticed that data are normally distributed
  - If that is the case:

Distance from Mean	Values within the distance
$\mu \pm 1\sigma$	68%
$\mu \pm 2\sigma$	95%
$\mu \pm 3\sigma$	99.7%

Dispersion of Data

- **Chebyshev's Theorem:** Regardless of the distribution → $1 - (1/k^2)$ percent values will fall within +/- k standard deviations from the mean (provided k > 1)



Dispersion of Data

- Sample variance: $s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$
- Sample Standard Deviation: $s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$

Question

A company produces a lightweight valve that is specified to weigh 1365 grams. Unfortunately, because of imperfections in the manufacturing process not all of the valves produced weigh exactly 1365 grams. In fact, the weights of the valves produced are normally distributed with a mean weight of 1365 grams and a standard deviation of 294 grams.

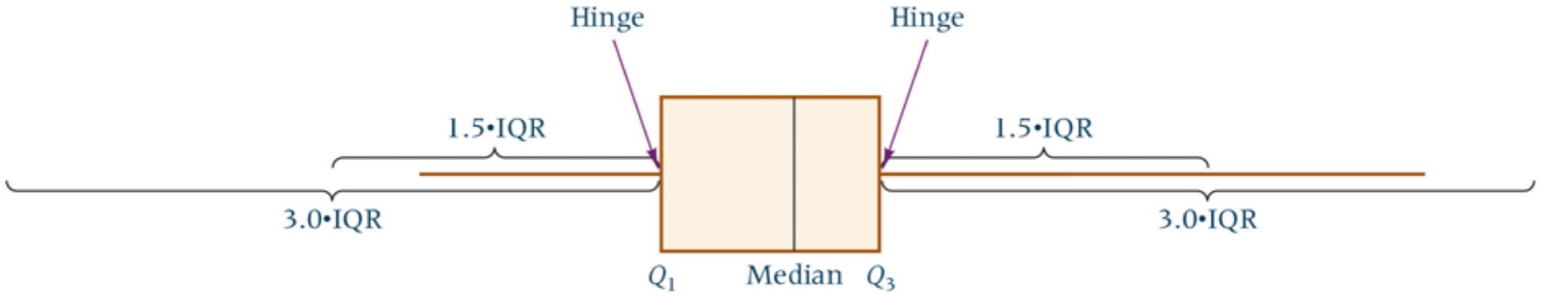
a) Within what range of weights would approximately 95% of the valve weights fall?

b) Approximately 16% of the weights would be more than what value?

c) Approximately 0.15% of the weights would be less than what value?

Dispersion of Data

- Box and Whisker Plot (5-number summary)
  - Median
  - Lower Quartile (Q1)
  - Upper Quartile (Q3)
  - Inner fence = +/- 1.5 IQR (mild outliers)
  - Outer fence = +/- 3IQR (extreme outliers)



Dispersion of Data

Box and Whisker Plot

- If the median is located on the right side of the box, then the middle 50% are skewed to the left.
- If the median is located on the left side of the box, then the middle 50% are skewed to the right.
- If the longest whisker is to the right of the box, then the outer data are skewed to the right and vice versa.