Feb – Jun 2023

IT258 - Data Science

Exploratory Data Analysis

Exploratory Data Analysis

- Techniques for visualizing and summarizing data
  - What can the data tell us? (in contrast to "confirmatory" data analysis)
- Introduced many basic techniques:
  - Mean, median, mode …
  - Standard deviation
  - 5-number summary
  - box plots
  - stem and leaf diagrams,…

Levels of Data Analysis

- **Descriptive analysis**
- **Inferential analysis**

Levels of Data Analysis

- **Descriptive analysis**
  - helps **describe**, **show** or **summarize** data in a meaningful way for discovering new patterns.
  - Also,
    - Does not allow conclusions beyond the data analyzed.
    - Does not provide any conclusions w.r.t any hypotheses made.

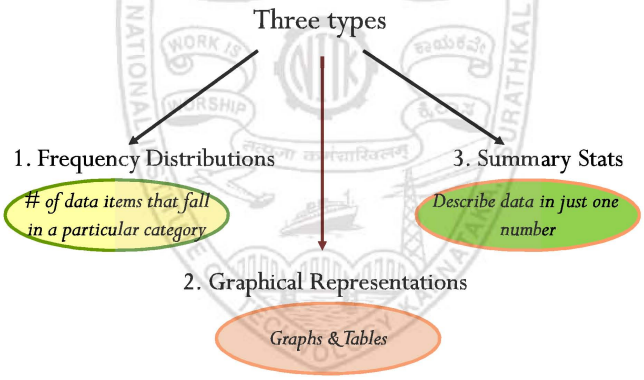
Levels of Data Analysis

- **Inferential analysis**
  - Techniques that can be **applied to well-defined samples** to make **generalizations about the collection** from which the samples were drawn.
  - Uses questions, models and hypotheses for investigation.
    - conclusions from inferential analysis extend beyond the immediate data.

Descriptive & Inferential Analysis

Descriptive	*Inferential*
• Organize • Summarize • Simplify • Present	• Generalize from samples to population • Hypothesis testing • Relationships among variables
↓	↓
Describe data	Make predictions

Descriptive Analysis



Descriptive Analysis

1. Frequency Distributions

Categorize on the basis of more that one variable at same time

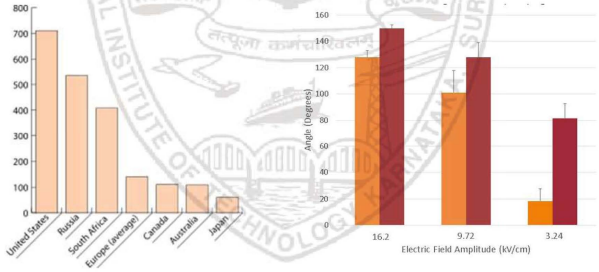
CROSS-TABULATION

			total
Technical Staff	14	11	25
Admin Staff	19	6	25
Total	33	17	50

Descriptive Analysis

2. Graphical Representations

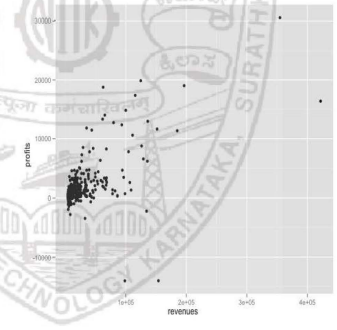
Bar graph (ratio data - quantitative)



Descriptive Analysis

2. Graphical Representations

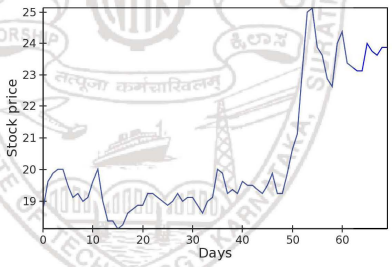
Scatter plot



Descriptive Analysis

2. Graphical Representations

Line plot

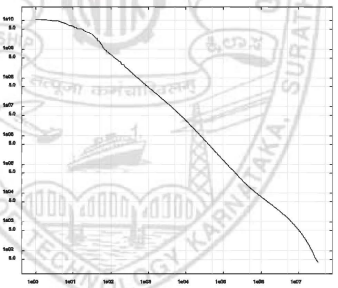


Descriptive Analysis

2. Graphical Representations

Log-log plot: Very useful for power law data

Frequency of words in tweets



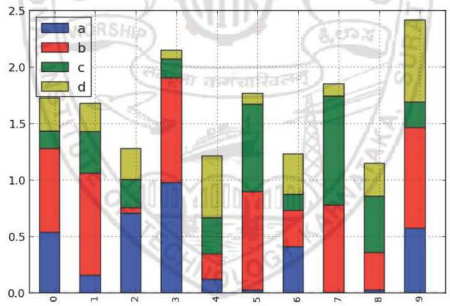
slope ~ -1

Rank of words in tweets, most frequent to least: I, the, you,…

Descriptive Analysis

2. Graphical Representations

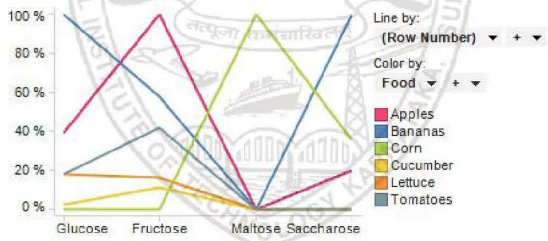
Stacked plot: stack variable is discrete:



Descriptive Analysis

2. Graphical Representations

Parallel coordinate plot: one discrete variable, an arbitrary number of other variables:



Descriptive Analysis

3. Summary Statistics:

describe data in just 1-2 numbers

Measures of central tendency	Measures of variability
• typical average score •Mean, Median, Mode ….	• typical average variation •Range •Standard Deviation •distribution …

Descriptive Data Analysis – Summary Statistics

- Three main categories of measures –
  - Central tendency
  - Variation or Dispersion
  - Shape

Descriptive Analysis

Central Tendency Measures

Descriptive Data Analysis

Central Tendency Measures

- describes a typical or **central** value in a given data distribution.
- Metrics
  - Arithmetic Mean, Median, Mode
  - Grouped Mean, Median, Mode
  - Geometric Mean, Harmonic Mean, Quadratic Mean
  - Quartiles
  - Midrange, Midhinge
  - Truncated Mean or trimmed mean
  - (Tukey's) Trimean
  - Interquartile Mean (IQM)
  - ….

Descriptive Data Analysis
Central Tendency Measures

Given - 13, 18, 13, 14, 13, 16, 14, 21, 13

- Arithmetic Mean

  (13 + 18 + 13 + 14 + 13 + 16 + 14 + 21 + 13) / 9 = **15**

- Median

  13, 13, 13, 13, **14**, 14, 16, 18, 21

- Mode

  13 is the mode

Descriptive Data Analysis
Central Tendency Measures

- How to find Mean, Median and Mode for Grouped Data?
  - For e.g. age of 21 volunteers involved in charity work at an orphanage

Age group	Frequency
51 - 55	2
56 - 60	7
61 - 65	8
66 - 70	4

Descriptive Data Analysis
Central Tendency Measures

- How to find Mean, Median and Mode for Grouped Data?
  - Only **estimates** can be obtained.
    - Estimate the values using the midpoints of the class intervals.

Descriptive Data Analysis
Central Tendency Measures

- Grouped Mean $= \frac{\sum fY}{n}$

  where:
  f → class frequency
  Y → class mean
  n → total number of values

- Grouped Median $= L + \frac{(n/2) - B}{G} \times w$

  where:
  L → the lower class boundary of the group containing the median
  n → the total number of values
  B → the cumulative frequency of the groups before the median group
  G → the frequency of the median group
  w → the group width

Descriptive Data Analysis
Central Tendency Measures

- Grouped Mode $= L + \frac{f_m - f_{m-1}}{(f_m - f_{m-1}) + (f_m - f_{m+1})} \times w$

where:
L → the lower class boundary of the group containing the median
f_{m-1} → the frequency of the group before the modal group
f_m → the frequency of the modal group
f_{m+1} → the frequency of the group after the modal group
w → the group width

Class Exercise

Given: Age of 21 volunteers involved in charity work at an orphanage. Find the grouped mean, median and mode.

Age group	Frequency
51 - 55	2
56 - 60	7
61 - 65	8
66 - 70	4

Homework

- The total number of cars sold by salesmen of a car showroom during different trimesters over two years were counted. The results are recorded in groups as follows. Estimate the median car sales of the showroom.

Range	Frequency
1 – 3	4
4 - 6	5
7 – 9	9
10 – 12	23
13 – 15	21
16 – 18	13
19 – 21	7
22 - 24	11

Descriptive Data Analysis
Central Tendency Measures

- Geometric Mean = n^{th} root of $(x_1 * x_2 * \ldots * x_n)$
  - Provides a way of finding a value in **between widely different values**.
    - *useful when we want to compare things with very different properties.*
    - *usually used for growth rates, like population growth or interest rates.*

Descriptive Data Analysis
Central Tendency Measures

- Harmonic Mean = $n.(x_1 * x_2 * \ldots * x_n)/(x_1 + x_2 + \ldots + x_n)$
  - best used for **fractions such as rates or multiples**.
  - is appropriate if the data values are ratios of two variables with different measures, called **rates**.

Descriptive Data Analysis
Central Tendency Measures

- Quadratic Mean (root mean squar $x_{\text{rms}} = \sqrt{\frac{1}{n}\left(x_1^2 + x_2^2 + \cdots + x_n^2\right)}$.
  - Used when the squares of the numbers add up to produce the net effect you are interested in.
  - gives a greater weight to larger items in a set
    - E.g. electrical current squared is proportional to power, so if you're interested in total power (rather than current), this can be used.

Descriptive Data Analysis
Central Tendency Measures

- Rule of thumb –
  - **If values have the same units**
    - Use the arithmetic mean.
  - **If values have differing units**
    - Use the geometric mean
  - **If values are rates**
    - Use the harmonic mean.
  - **If effect of square on values to be observed**
    - Use quadratic mean

Descriptive Data Analysis
Central Tendency Measures

- **Quartile** – Any value in a data distribution that divides the distribution into four parts of equal frequency.
  - The median of the **lower** half of data - **lower or first quartile Q_1**
  - The median of the **upper** half of data - **upper or third quartile Q_3**

Descriptive Data Analysis
Central Tendency Measures

- **Quartile** – Any value in a data distribution that divides the distribution into four parts of equal frequency.
  - The median of the **lower** half of data - **lower or first quartile Q_1**
  - The median of the **upper** half of data - **upper or third quartile Q_3**

13, 13, 13, 13, 14, 14, 16, 18, 21

$Q_1 = (13+13)/2 = 13$ $\quad$ $Q_3 = (16+18)/2 = 17$

Descriptive Data Analysis
Central Tendency Measures

Given - 13, 18, 13, 14, 13, 16, 14, 21, 13

- **Midrange** – the arithmetic mean of the maximum and minimum values in a dataset

$$M = \frac{\max x + \min x}{2} = (21 + 13)/2 = 17$$

Descriptive Data Analysis
Central Tendency Measures

Given - 13, 18, 13, 14, 13, 16, 14, 21, 13

- **Midhinge** - the average of the first and third quartiles of the given data.

$$\text{Midhinge} = \frac{Q_1 + Q_3}{2} = (13 + 17)/2 = 15$$

Descriptive Data Analysis
Central Tendency Measures

- **Truncated Mean (or trimmed mean)**
  - Mean computed after trimming any outliers at the two extremities of a data sample.
    - Typically required for extremely skewed distributions.
  - expressed in percentages.
    - indicates the percentage of data to remove.
    - E.g., 5% trimmed mean

Descriptive Data Analysis
Central Tendency Measures

Given - 13, 18, 10, 14, 13, 16, 14, 21, 13, 29

- 10% Trimmed Mean

-- 10, 13, 13, 13, 14 ,14, 16, 18, 21, 29

-- **10**, 13, 13, 13, 14 ,14, 16, 18, 21, **29**

10% Trimmed Mean = Arithmetic Mean(13, 13, 13, 14, 14, 16, 18, 21)
= 15.35

Descriptive Data Analysis
Central Tendency Measures

- **Trimean**
  - weighted average of the median and upper & lower quartiles.

$$\text{Trimean} = (Q_1 + 2Q_2 + Q_3)/4$$

  - where,
    - Q_1 *and* Q_3 *– lower and upper quartiles (or hinges)*
    - Q_2 *is the median (i.e. 50th quartile).*
  - Adv: reduced sensitivity to outliers in the data.

Descriptive Data Analysis
Central Tendency Measures

Given - 13, 18, 13, 14, 13, 16, 14, 21, 13

- Trimean -

  13, 13, 13, 13, 14, 14, 16, 18, 21

  Lower Quartile $Q_1 = (13+13)/2 = 13$

  Median $Q_2 = 14$

  Upper Quartile $Q_3 = (16+18)/2 = 17$

  Trimean $= (Q_1+2Q_2+Q_3)/4 = 14.5$

Descriptive Analysis

Dispersion Measures

Descriptive Data Analysis
Dispersion Measures

- Also called *Spread* or *Scatter* or *Variability*.
  - Used to capture the way the sample data is stretched/compressed, away/ towards the central value.
  - Reveals all the peculiarities and characteristics of the data sample.

* Opposite of central tendency.

Descriptive Data Analysis
Dispersion Measures

- Mathematical measures –
  - Range
  - Standard deviation
  - Mean absolute deviation (MAD)
  - Median absolute deviation
  - Interquartile range (IQR)
  - Variance
  - Trimmed Variance
  - Mean Absolute Difference/Gini mean absolute difference …
  - ….

Descriptive Data Analysis
Dispersion Measures

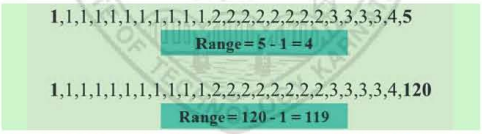
- **Range**
  - Depends only on extreme values and provides no information about how the remaining data is distributed.

Range = Largest (L) – Smallest (S)

Descriptive Data Analysis
Dispersion Measures

- **Range - Issues**
  - Ignores the way data is distributed

  - Sensitive to outliers



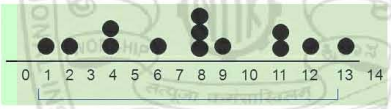
Descriptive Data Analysis
Dispersion Measures

- **Coefficient of Range**

$$\text{Coefficient of Range} = \frac{L - S}{L + S}$$

Descriptive Data Analysis
Dispersion Measures

- **Co-efficient of Range**



$$\text{Range} = \text{Largest (L)} - \text{Smallest (S)} = 12$$

$$\text{Coefficient of Range} = \frac{L - S}{L + S} = \frac{12}{14} = 0.857$$

Descriptive Data Analysis
Dispersion Measures

- **Range and Co-efficient of Range (example)**

Set A (Math marks, out of 25) : 10, 15, 18, 20, 20

Set B (English marks, out of 100) : 30, 35, 40, 45, 50

	Range	Coefficient of Range
Set A	20 -10 = 10	(20-10)/(20+10) = 0.33
Set B	50 – 30 = 20	(50-30)/(50+30) = 0.25

Descriptive Data Analysis
Dispersion Measures

- **Range and Co-efficient of Range (example)**

Set A (Math marks, out of 100) : 40, 60, 76, 80, 80

Set B (English marks, out of 100) : 30, 35, 40, 45, 50

	Range	Coefficient of Range
Set A	80 - 40 = 40	(80-40)/(80+40) = 0.33
Set B	50 – 30 = 20	(50-30)/(50+30) = 0.25

Descriptive Data Analysis
Dispersion Measures

- **Standard Deviation (σ)**
  - quantifies the amount of dispersion of a set of data values.

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

  - Where, x1,x2…xn are observed values,
  - $\bar{x}$ is the mean of the set
  - N is the number of observations.

Descriptive Data Analysis
Dispersion Measures

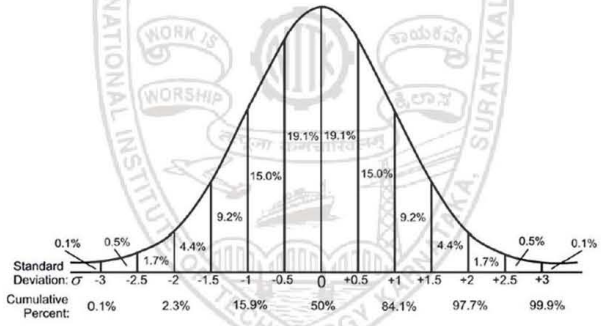
- **Standard Deviation (σ)**
  - quantifies the amount of dispersion of a set of data values.

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

  - Where, x1,x2…xn are observed values,
  - $\bar{x}$ is the mean of the set
  - N is the number of observations.

* Low σ = data points are closer to the mean (i.e. expected value)
High σ = data points are spread out over a wider range of values.

Descriptive Data Analysis
Dispersion Measures



Descriptive Data Analysis
Dispersion Measures

Standard Deviations Away From Mean	Abnormality	Probability of Occurance
beyond -3 sd	extremely subnormal	0.15%
-3 to -2 sd	greatly subnormal	2.35%
-2 to -1 sd	subnormal	13.5%
-1 to +1 sd	normal	68.0%
+1 to +2 sd	above normal	13.5%
+2 to +3 sd	greatly above normal	2.35%
beyond +3 sd	extremely above normal	0.15%

Descriptive Data Analysis
Dispersion Measures

- **Mean Absolute Deviation (MAD)**
  - the average distance between each data point in the given data and the mean.

$$\text{MAD} = \frac{\sum |x_i - \bar{x}|}{n}$$

Descriptive Data Analysis
Dispersion Measures

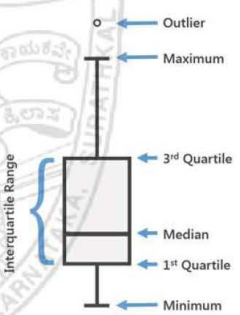
- Class exercise
  - Six of your Facebook photos got 10, 15, 15, 17, 18, 21 "likes". Find the mean dispersion of this dataset.

Descriptive Data Analysis
Dispersion Measures

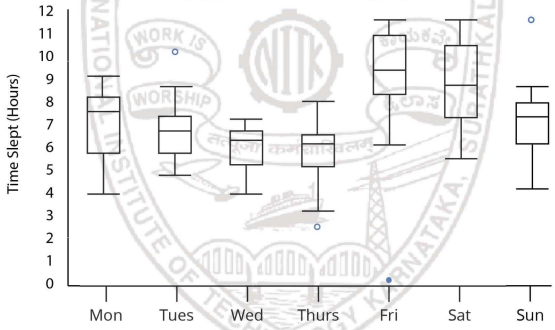
- **Inter-quartile range (IQR)**
  - difference between the upper and lower quartiles
  - a good indicator of the spread in the center region of the data
  - Often reported using the "five-number summary" called **Box Plot.**
    - **Minimum**
    - **lower quartile**
    - **Median**
    - **upper quartile**
    - **maximum.**

Box Plot

- *Data* - represented with a box
  - Box ends – 1st and 3rd quartiles
  - i.e., box height = IQR
- *Median* - marked by a line within box
- *Whiskers* - two lines outside the box extended to Minimum and Maximum
- *Outliers* - points beyond a specified outlier threshold, plotted individually



Box Plot

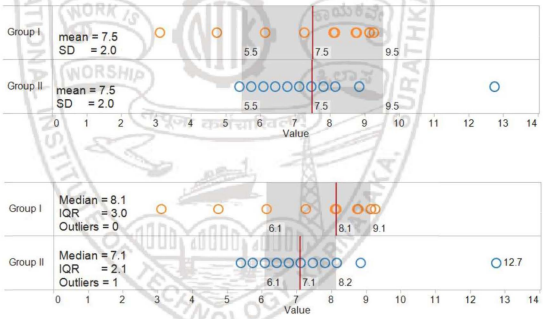


Descriptive Data Analysis
Dispersion Measures

- **Inter-quartile range (IQR)**
  - Also helps in -
    - **Outlier Identification.**
      - *Outliers - values more than one-and-a-half times the IQR distance below the first quartile or above the third quartile.*
    - **Skewness.**
      - *Comparing the median to the quartile values shows whether data is skewed.*

Descriptive Data Analysis
Dispersion Measures

- **Std. Dev v/s IQR**



Descriptive Data Analysis
Dispersion Measures

- **Mean Absolute Difference**
  - average absolute difference of two independent values drawn from a probability distribution.
  - Also called the **Gini Mean Difference** (GMD)

$$= E[|X - Y|] = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}|x_i - y_j|$$

Descriptive Data Analysis
Dispersion Measures

- **Variance**
  - Measures how far individuals in a group are spread out.
    - Conversely, SD measures how many observations differ from its mean.
  - Given by the square of Standard Deviation.

$$S^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$$

x_i = the value of the one observation
$\bar{x}$ = the mean value of all observations
n = the number of observations

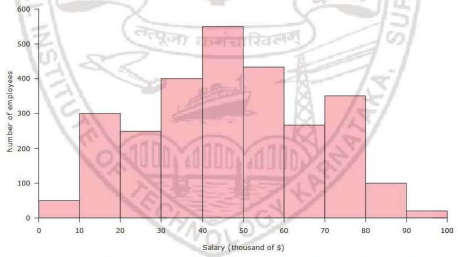
Descriptive Data Analysis
Dispersion Measures

- **Trimmed Variance**
  - Similar to variance, but less affected by outliers since the largest x% and the smallest x% of the sample are eliminated.
    - Typical range for x% is 5% to 25%.

Descriptive Analysis

Shape related Measures

- describe the distribution (or pattern) of the data within a dataset.
  - Typically plotted as histograms.



- Shape may be symmetrical or asymmetrical.
  - Symmetrical distributions:
    - the two sides of the distribution are a mirror image of each other.
  - Asymmetrical or Skewed distributions:
    - the two sides will not be mirror images of each other.
    - i.e. Data exhibits skewness.

- Mathematical measures –
  - Skewness
  - Kurtosis

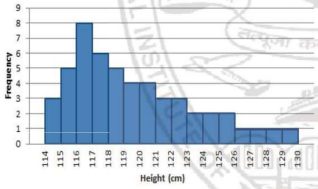
- **Skewness**
  - measure of asymmetry of the probability distribution of a real-valued random variable about its mean.
  - Types:
    - Symmetric
    - Positively skewed
    - Negatively skewed

- **Positively skewed**:
  - the tail on the **right side is longer than the left side** of the histogram.
    - i.e. **Smaller values tend to cluster toward the left side of the x-axis** with increasingly fewer values at the right side of the x-axis.
- **Negatively skewed**:
  - the tail on **the left side is longer than the right side**.
    - i.e. **larger values tend to cluster toward the right side of the x-axis**, with increasingly less values on the left side of the x-axis

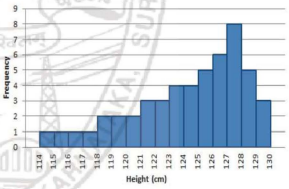
Descriptive Data Analysis

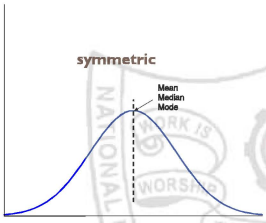
Shape of Data

Positively skewed

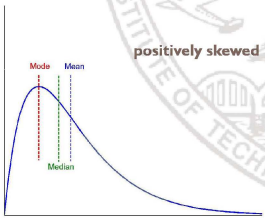


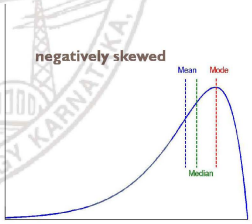
Negatively skewed





Median, mean and mode of symmetric, positively and negatively skewed data





Skewness

The **moment coefficient of skewness** of a data set

$$g_1 = m_3 / m_2^{3/2}$$

where,

$m_3 = \sum(x-\bar{x})^3 / n$

$m_2 = \sum(x-\bar{x})^2 / n$

$\bar{x}$ is the mean

n is the sample size.

m_3 -- **third moment** of the data set.

m_2 -- **variance**, the square of the standard deviation.

Skewness

The **moment coefficient of skewness** of a data set

$$g_1 = m_3 / m_2^{3/2}$$

where,

$m_3 = \sum(x-\bar{x})^3 / n$

$m_2 = \sum(x-\bar{x})^2 / n$

$\bar{x}$ is the mean

n is the sample size.

m_3 -- **third moment** of the data set.

m_2 -- **variance**, the square of the standard deviation.

If only a sample of the data is given, then

Sample skewness $G_1 = \dfrac{\sqrt{n(n-1)}}{n-2} g_1$

Skewness

- **How to interpret values ??**
  - Rule of thumb (put forth by Bulmer et al, 1979):
    - If skewness is less than −1 or greater than +1
      - *the distribution is highly skewed.*
    - If skewness is between −1 and −½ or between +½ and +1
      - *the distribution is moderately skewed.*
    - If skewness is between −½ and +½
      - *the distribution is approximately symmetric.*

Class Exercise

Grouped data w.r.t the marks of 100 randomly selected students in a college is given. Determine the extent to which this data is skewed

Marks (out of 100)	Number of students
59.5–62.5	5
62.5–65.5	18
65.5–68.5	42
68.5–71.5	27
71.5–74.5	8

Class Exercise

- **How to interpret this??**

Descriptive Data Analysis
Shape of Data

- **Kurtosis**
  - is a measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution.
  - is a measure of relative peakedness of a distribution. It is a shape parameter that characterizes the degree of peakedness.

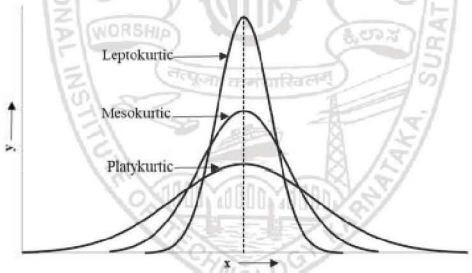
* Datasets with high kurtosis → heavy tails, or outliers.
Datasets with low kurtosis → light tails, or lack of outliers.

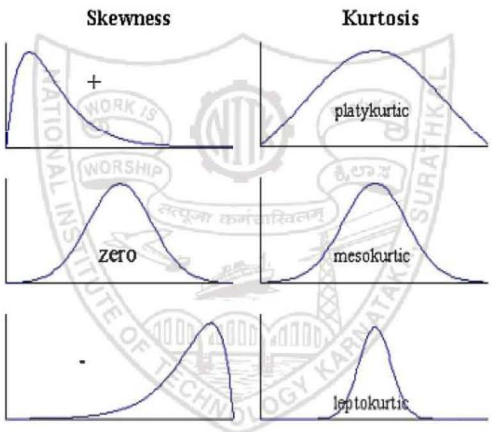
Descriptive Data Analysis
Shape of Data

- Kurtosis categorized based on difference w.r.t Normal distribution (kurtosis=3; excess kurtosis =0)
- Types
  - **Platykurtic:**
    - kurtosis <3 (excess kurtosis <0)
    - lighter tails that are shorter and contain fewer outliers.
  - **Mesokurtic:**
    - kurtosis ≈3 (excess ≈ 0)
    - tails are shorter and thinner, and often its central peak is lower and broader.
  - **Leptokurtic:**
    - kurtosis >3 (excess kurtosis >0)
    - have heavy tails that are longer and contain more outliers

Descriptive Data Analysis
Shape of Data

- **Kurtosis**





Descriptive Data Analysis
Shape of Data

The **moment coefficient of kurtosis** of a data set

$$a_4 = m_4 / m_2^2$$

Excess kurtosis:

$$g_2 = a_4 - 3$$

Where,

$m_4 = \sum(x-\bar{x})^4 / n$
$m_2 = \sum(x-\bar{x})^2 / n$
$\bar{x}$ is the mean
n is the sample size.

m_4 -- **fourth moment** of the data set.
m_2 -- **variance**, the square of the standard deviation.

Descriptive Data Analysis
Shape of Data

The **moment coefficient of kurtosis** of a data set

$$a_4 = m_4 / m_2^2$$

Excess kurtosis:

$$g_2 = a_4 - 3$$

Where,

$m_4 = \sum(x-\bar{x})^4 / n$
$m_2 = \sum(x-\bar{x})^2 / n$
$\bar{x}$ is the mean
n is the sample size.

m_4 -- **fourth moment** of the data set.
m_2 -- **variance**, the square of the standard deviation.

Sample excess kurtosis: $G_2 = \frac{n-1}{(n-2)(n-3)}[(n+1)g_2 + 6]$

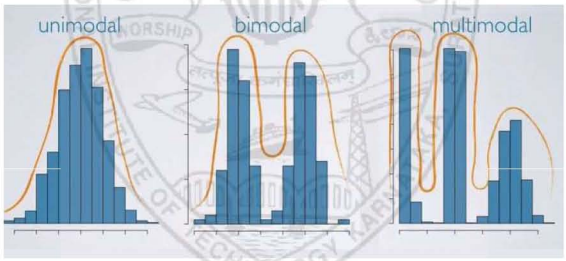
Homework

Grouped data w.r.t the marks of 100 randomly selected students in a college is given. Determine the extent to which this data exhibits heavy tailed behavior. Also, determine the type of kurtosis exhibited by this sample based on the computed values.

Marks (out of 100)	Number of students
59.5-62.5	5
62.5-65.5	18
65.5-68.5	42
68.5-71.5	27
71.5-74.5	8

Descriptive Data Analysis
Shape of Data

Unimodal , Bimodal, Multi-modal Data



Further reading

- Abt, K. (1987). Descriptive data analysis: a concept between confirmatory and exploratory data analysis. *Methods of information in medicine*, *26*(02), 77-88.
- Wilcox, R.R. and Keselman, H.J., 2003. Modern robust data analysis methods. *Psychological methods*, *8*(3), p.254.