Lecture 1

Data Visualisation & Centrality Measures

Anees Parwez
Souhardya Ray

Table of contents

Introduction

Learning Goals:

- Develop mathematical background to statistical methodologies.
- Methodologies combined with industry knowledge can solve real life datascience challenges

- Machine Learning has two components:
  1. Algorithm implementation part can be done in R/Python
  2. Construction of these algorithm based on assumptions on data.
- Develop broader understanding of ML algorithms. Eg, logistic and linear regression are special cases of GLM.
- Ask statistical enquires to analysis done.

Statistical enquiries

Correlation and Causation

- Correlation means X and Y change together, it just may be a coincidence. Causality means X makes Y happen.
- In a study done in 2009, it had been speculated that homicide rates were higher when ice cream sales were on the rise.

Selection Bias

- Distorting in a measure of association due to a sample selection that does not accurately reflect the target population.
- A case-control study of smoking and chronic lung disease: the association of exposure(smoking) with lung disease will weaken if controls (non-smokers) are selected from a hospital population than if selected from the community.

Data Visualization

Presentation of Data

- The first step in a statistical analysis is visualization of data.
- It helps to visually summarize the information contained in data which is useful to identify and share real-time trends, outliers, and new insights.
- The goal of data visualization is to explore, analyse and present.
- Different data types require different visualisation tools.

Qualitative data

- This kind of data is divided into categories based on non-numeric characteristics.
- It can be of two types:
  1. Ordinal: Meaning it follows an order or sequence.
  2. Categorical: Meaning it follows no fixed order.
- Qualitative data can be easily represented through tables.

Education Level	Frequency
Primary	50
Secondary	20
Higher Secondary	20
Graduate	10

Table: Table for Ordinal

Blood Group	Frequency
A	82
B	178
AB	20
O	90

Table: Table for Nominal

Qualitative Data Pictorial Representation

- Pictorially qualitative data can be represented in bar or pie charts.

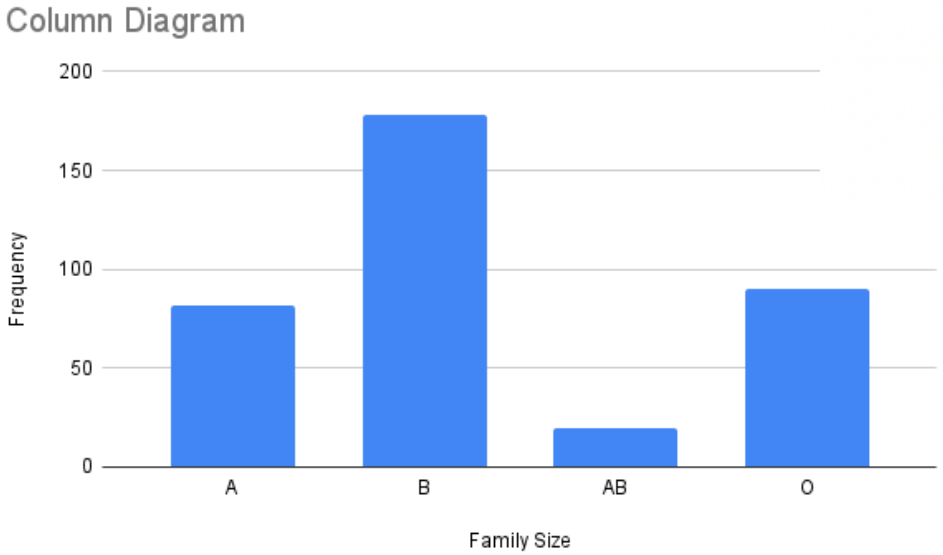


Figure: Column chart

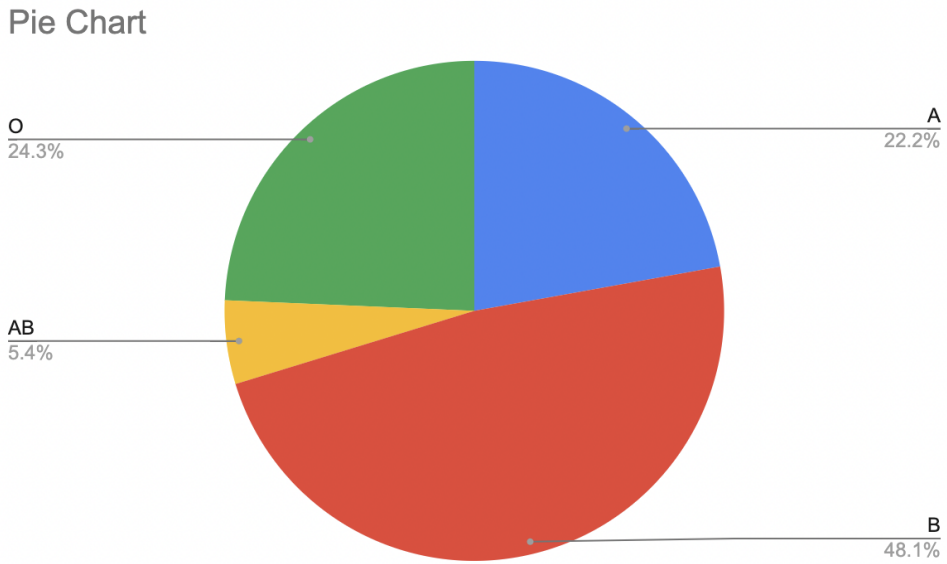


Figure: Pie chart

Bivariate Qualitative data representation

- Many a times data comes in pairs. For eg, number of cell phone users in $11-th$ and $12-th$ grades.

Cell Phone?	11th Grade	12th Grade	Total
Yes	59	50	109
No	6	3	9
Total	65	53	118

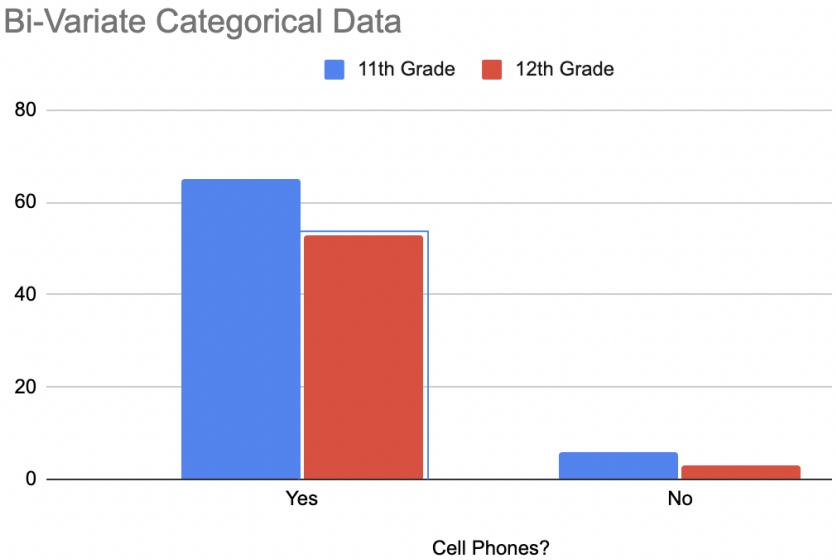


Figure: Bivariate Bar Chart

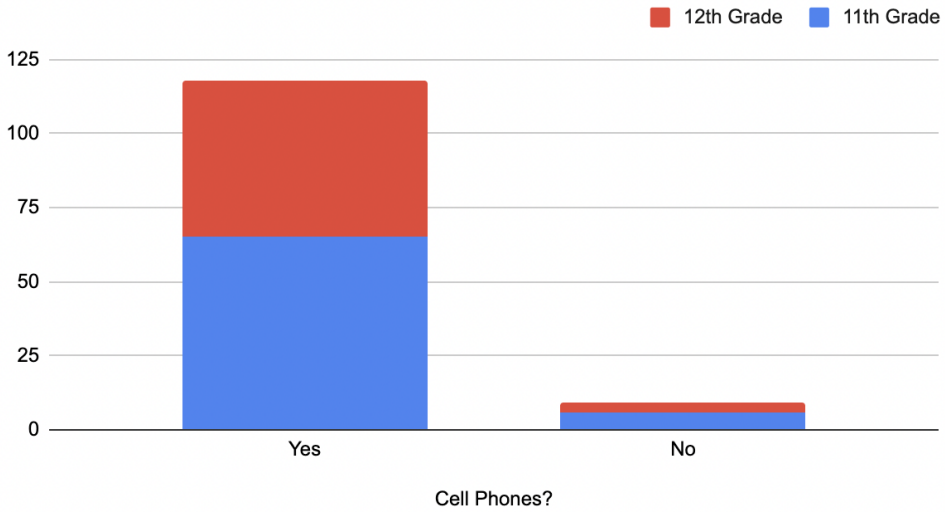


Figure: Stacked Bar Chart

Quantitative Data

- Data that can be quantified and measured. This category of data can be further subdivided into:
  - Discrete: Data that consists of whole numbers (0, 1, 2, 3, . . .). For example, e, the number of children in a family.
  - Continuous: Data that can take any value within an interval. For example, people's height (between 60 -70 inches)

- Grouped discrete quantitative data can be represented by bar chart or stick charts. Alternatively, it can also be represented as step diagram.

- Grouped continuous (quantitative) data can be represented by:
  - Histogram
  - Line chart
  - Frequency polygon
  - Scatter plot for bivariate data

Discrete Quantitative Data Representation

Family Size (x)	Frequency ($f(x)$)	Cumulative Frequency ($f(X \leq x)$)
2	2	2
3	3	5
4	7	12
5	5	17

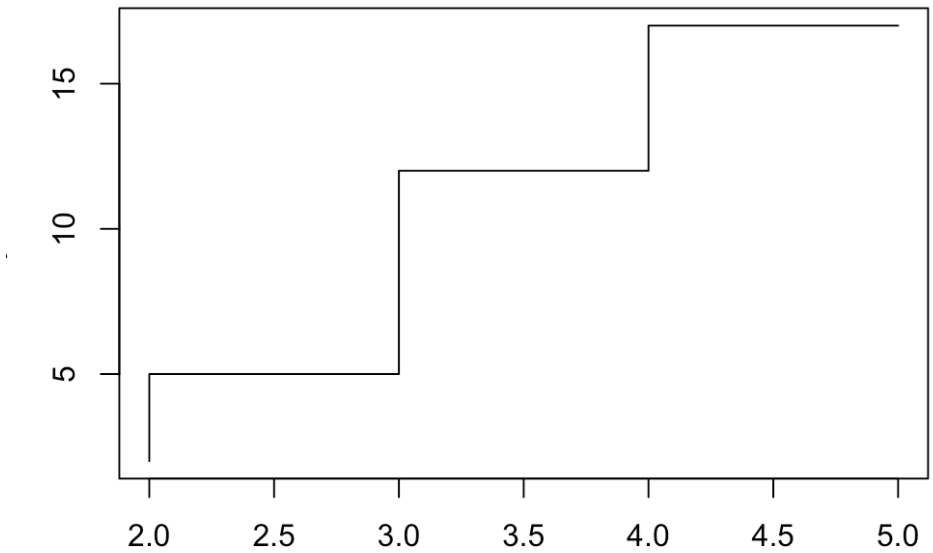
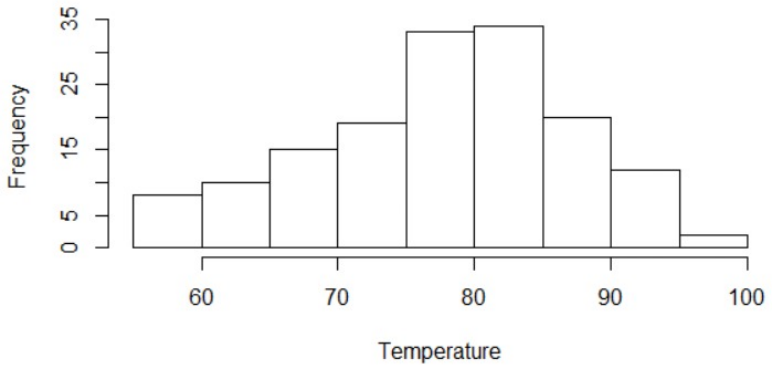
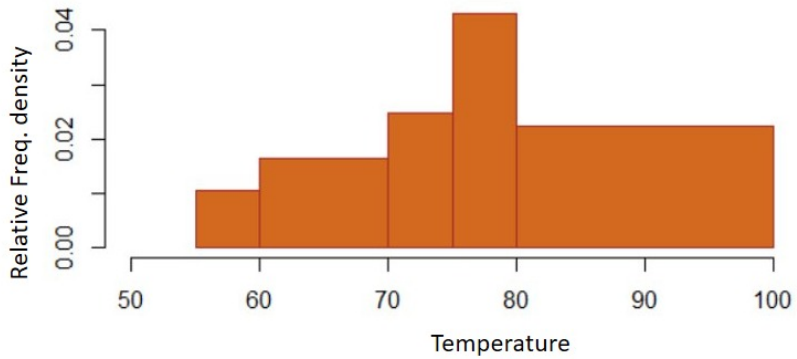


Figure: Step Diagram

Cont. Quantitative Data Representation

- Histogram is used to represent grouped continuous variables.
- Finer classing ensure more accurate empirical distribution of data.
- Height of each rectangle (on each class) of histogram represents the relative frequency density of that class.
- Relative Frequency of a class = $\frac{\text{Frequency of a Class}}{\text{Total Frequency}}$
- Relative Frequency Density = $\frac{\text{Relative Frequency}}{\text{Class Width}}$
- Area of each rectangle (on each class) represents the relative frequency of that class. Sum of area of all rectangles equals 1.





Central Tendency in Data

Central Tendency

- Many of our notations stem from capturing 'average' of given data.
- A measure of central tendency is a single value that attempts to describe data by identifying the central position.
- The mean, median and mode are all valid measures of central tendency, but under different conditions, some measures are more appropriate.
- Besides the centrality measure of data, it is necessary to capture how scattered the given values are about its 'average'. This is dispersion which we will cover later.

Mean

- Definition: If $x_1, x_2, ...x_n$ are the data points, then mean of the considered variable x,

$$\bar{x} = \sum_{i=1}^{n} x_i$$

- Translation & scale variance: If all the data points are shifted by b units and scaled by a units i.e.

$$x_i' = a + b \cdot x_i \quad \forall i \in 1 \ldots n \implies \bar{x'} = a + b\bar{x}$$

- Weighted mean: If there are two groups of variable x, one containing n_1 values with mean $\bar{x_1}$ and the other containing n_2 values with mean $\bar{x_2}$, then the mean of the combined data is given by,

$$\bar{x} = \frac{n_1.\bar{x}_1 + n_2.\bar{x}_2}{n_1 + n_2}$$

- Question: Compare $\bar{x}$ with $\min(\bar{x}_1, \bar{x}_2)$ and $max(\bar{x}_1, \bar{x}_2)$

Mean associated measures

- Geometric Mean: The geometric mean is defined as the nth root of the product of n numbers, i.e., for a set of numbers $x_1, x_2, \ldots, x_n$, the geometric mean is defined as

$$\left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}} = \sqrt[n]{x_1 x_2 \cdots x_n}$$

- Harmonic Mean: The harmonic mean H of the positive real numbers $x_1, x_2, \ldots, x_n$ is defined to be

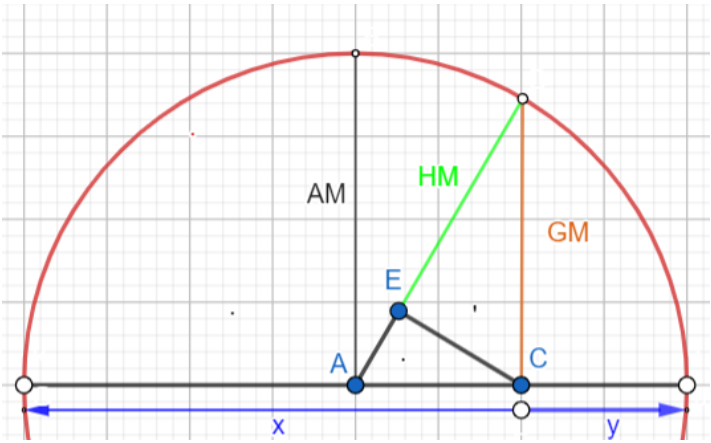
$$H = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} = \left(\frac{\sum_{i=1}^{n} x_i^{-1}}{n}\right)^{-1}$$

- Question: A car travels a distance d at a speed x and the same distance again at speed y. Calculate the average speed.

Mean Inequalities

- AFAP: Batting average of two baseball players for 2 years. Should the combined batting average of David necessarily be greater than Derek?

Batter \ Year	1995		1996	
Derek Jeter	12/48	.250	183/582	.314
David Justice	104/411	**.253**	45/140	**.321**



- AM-GM-HM inequality: For a set of positive real numbers $x_1, \ldots, x_n$,

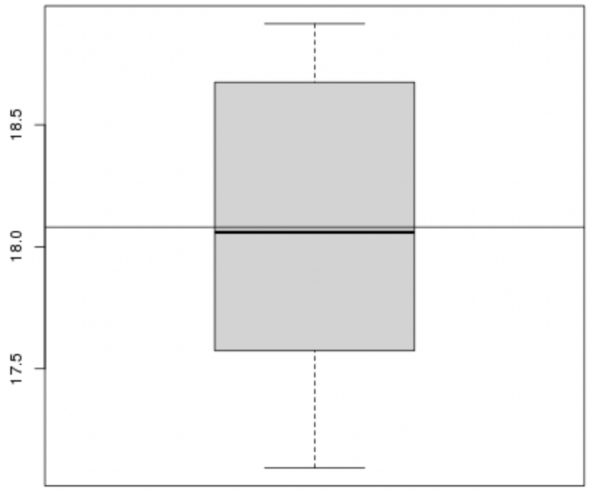
$$\frac{x_1 + \ldots + x_n}{n} \geq (x_1 \ldots x_n)^{1/n} \geq \frac{n}{x_1^{-1} + \ldots + x_n^{-1}}$$

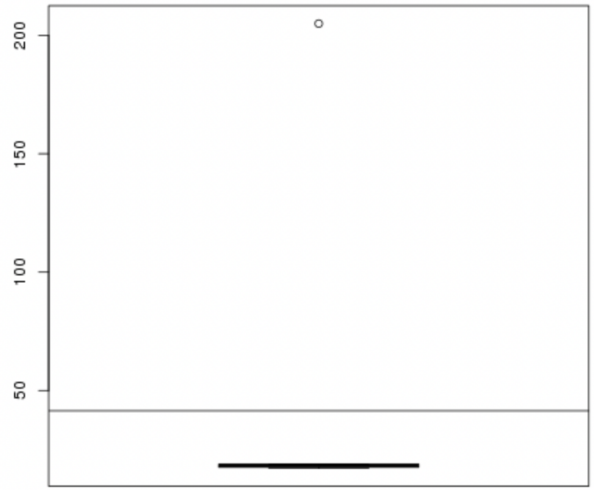
- Practice: If a, b, c are positive reals such that: $a + b > c$, $b + c > a$, $c + a > b$ and $a + b + c = 2$ then show that:

$$1 \geq ab + bc + ac - abc \leq 28/27$$

Median

- Consider 8 job market candidates from a department. Students 1 to 7 secured a mean CTC of 18.5 LPA (minimum 17.76, maximum 19.53) but Student 8, got an offer of 2.05 Crore.
- When the $8-th$ student is also considered, the mean rises to 42.01 from 18.72.
- But this new mean doesn't represent the actual central tendency of the data because 7 out of 8 candidates got far lower than mean salary in reality. This is evident from boxplot below:





Median

- Definition: The median of a variable is defined as the middlemost value when its values are arranged in ascending or descending order of magnitude.
- *n* odd case:If the total number of given values n, is an odd number, then there exists only one middlemost value, the $\frac{n+1}{2} - th$ value in the ordered arrangement of numbers.
- *n* even case:If n is even, median may not be unique. Any value between the $\frac{n}{2} - th$ and $(\frac{n}{2} + 1) - th$ number in the ordered arrangement, can be taken as median. As a standard however, the arithmetic mean of $\frac{n}{2}$ and $(\frac{n}{2} + 1) - th$ values represents the median of the values.
- Question: If $\{x_i\}_{i=1}^{n}$ and $\{y_i\}_{i=1}^{m}$ have median $\tilde{x}$ and $\tilde{y}$ respectively, show that median of combined dataset $\tilde{M}$ lies between $\tilde{x}$ and $\tilde{y}$

Median for frequency distribution

- For a frequency distribution of continous variable, the median may be supposed to be the value for which cumulative frequency is $n/2$.
- The median class can be found by cumulative frequency table.
- Let x_l and x_u denote the lower and upper class boundaries of median class. The corresponding C.F. are denoted by n_l and n_u.
- If C.F. is assumed to be linear between x_l and x_u, then the median M which is the value with C.F. $n/2$ satisfies:

$$\frac{M_i - x_i}{x_u - x_l} = \frac{n/2 - n_l}{n_u - n_l} \implies M_i = x_i + \frac{n/2 - n_l}{n_u - n_l} \times (x_u - x_l)$$

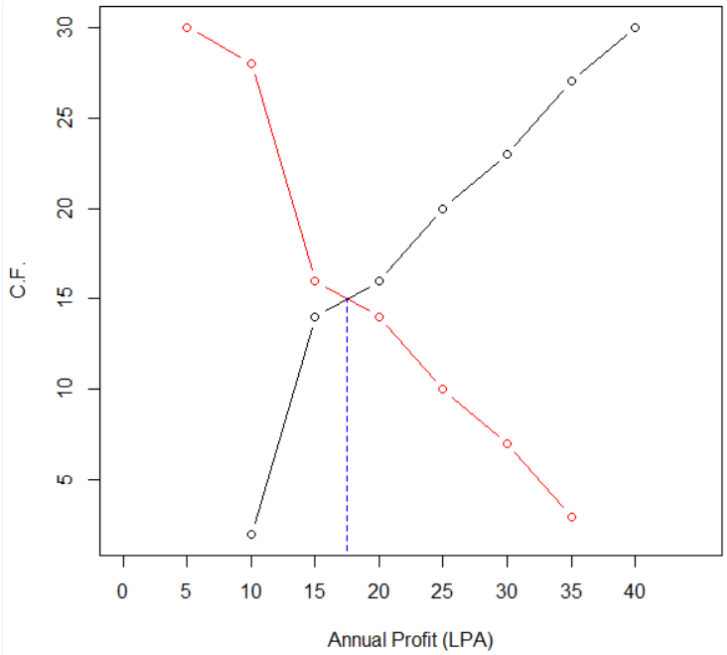
- The same value may also be obtained geometrically, from the ogive of frequency distribution.
- Question: The median is the point where the less than type C.F. and more than type C.F. intersect. Justify.

Finding Median: An example

- Example: Consider annual profits earned by 30 shops of a shopping complex.

Annual Profit	Frequency	C.F. (less than)	C.F. (More than)
5 to 10	2	2	30
10 to 15	12	14	28
15 to 20	2	16	16
20 to 25	4	20	14
25 to 30	3	23	10
30 to 35	4	27	7
35 to 40	3	30	3

Table: Annual Profits (in LPA)

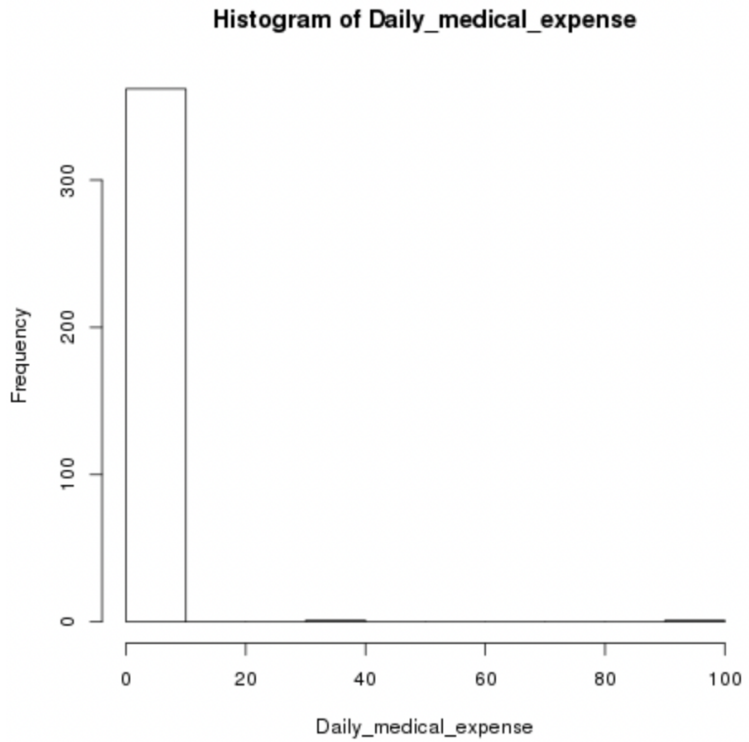


- Here $n/2 = 15$ which corresponds to the class $15 - to - 20$. Correspondingly, $x_l = 15$, $x_u = 20$, $n_l = 14$ and $n_u = 16$. Hence, the median is

$$M_i = 15 + \frac{15 - 14}{16 - 14} \times (20 - 15) = 17.5$$

Mode

- Tell me a situation where mean or median is not a good central tendency measure?
- Definition: The mode of a variable is that value of the variable which has the highest frequency or frequency density, according as the variable is discrete or continuous.
- Example. Daily Medical expense of a 18 year old boy.

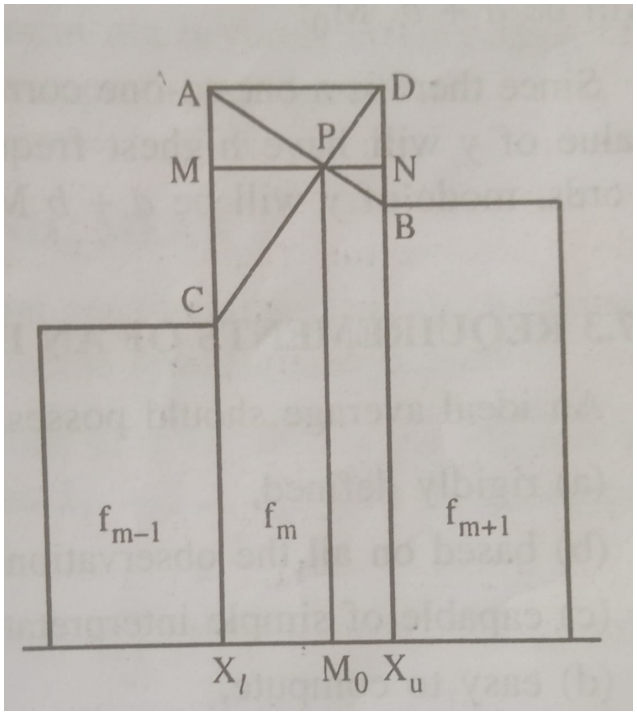


Mode for cont. frequency table

- For a continuous frequency distribution the model class is the class with highest frequency density.
- Suppose the modal class has width c and adjacent classes have class limits $(x_l - c, x_l)$ and $(x_u, x_u + c)$.
- Suppose x_l and x_u be lower and upper boundaries of modal class and f_{m-1}, f_m, f_{m+1} the frequencies as in figure.

- It can be assumed that frequencies increase linearly to mode before again decreasing.
- By similarity of ΔPAC and ΔPBD, $PM/PN = AC/BD$ hence,

$$\frac{M_o - x_l}{x_u - M_o} = \frac{f_m - f_{m-1}}{f_m - f_{m+1}}$$

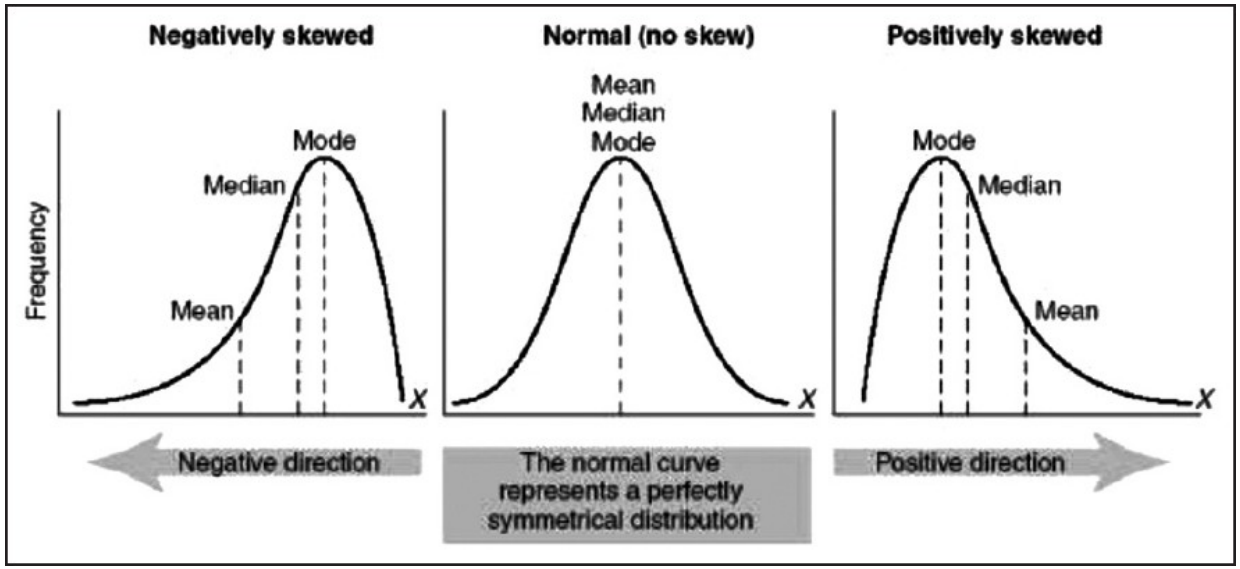


Mean Median and Mode comparison

- Suppose the variable are given as frequency table of which one or both the terminal classes are open. Here, computing mean is impossible (why?), but we can use median and mode.
- Empirically, the following relationship usually holds:

$$\bar{x} - Mo = 3(\bar{x} - Mi)$$

- Question: Using this relation, can you justify the following figure for skewed data.



Other measures

- Winsorised mean: To eliminate the effect of extreme values, we proceed as follows: replace each value lower than $1-st$ quartile by value of $1-st$ quartile and each value higher than $3-rd$ quartile, by the value of $3-rd$ quartile. Compute the mean of this modified data.

- Circular mean: The latitudes and longitudes of a place are recorded as (x_i, y_i). The minus(or plus) sign indicates West(or East) for longitudes and North (or South) for latitudes.

lat	1.14	1.96	-1.25	0.67	-2.36	-2.24	-1.43	2.73
long	-179.52	177.04	177.10	-177.59	-180.00	-178.82	178.12	177.70

Suggest a measure of centrality for such a data.

- Question Record hour data of road accidents during a day (0 to 23) hours. Suggest a measure of centrality.