Statistics

Frequency distribution

It is a table that shows classes or interval of data with a count of number of enteries in each class.

The frequency f of a class is the number of data entries in the class.

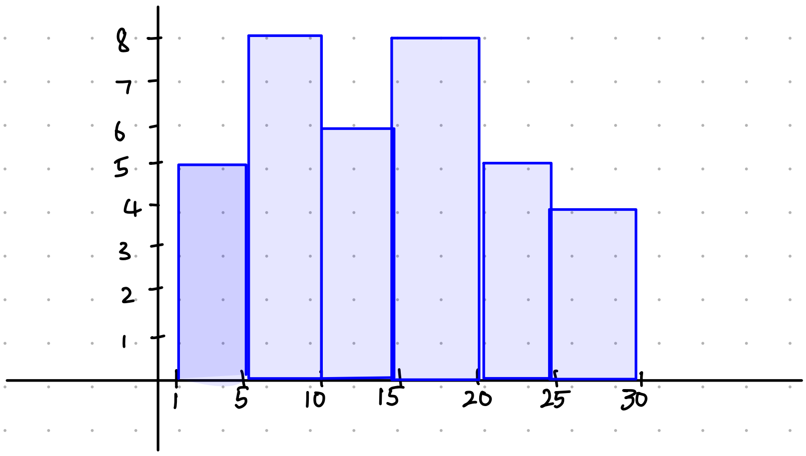
For instance if $X = \{x_1, x_2, \dots x_n\}$ are data set (Radom variable) and $F = \{f_1, f_2, \dots f_n)$ are corresponding freq., then Freq. distribution

X	x_1	x_2	x_3	$\dots$	x_n
F	f_1	f_2	f_3	$\dots$	f_n

Graph of frequency distribution

Ex:

Class	1–5	6–10	11–15	16–20	21–25	26–30
Freq, f	5	8	6	8	5	4



Fundamental task in many statistical analyses is to characterize the location and variability (or spread) of a data set.

1) Measure of Central tendency

It is a value that represents a typical, or central entry of data set.

Commonly used measures are

1) Mean
2) Median
3) Mode

The mean of a data set is sum of the enteries divided by no. of entries,

$$\bar{x} = \frac{\sum x_i}{N}$$

For the freq distribution

X	x_1	x_2	x_3	$\cdots$	x_n
F	f_1	f_2	f_3	$\cdots$	f_n

$$\bar{x} = \frac{\sum f_i x_i}{N}, \quad \text{where} \quad N = \sum f_i$$

The median of a data set is the value that lies in the middle of the data when the data set is ordered

Ex: For the data set

388 397 397 427 432 782 872

median is 427

The mode of a data set is the entry with highest frequency

For the above Ex mode = 397

2) Dispersion

It is a value that represents spread of data, it shows how squeezed or scattered the data are.

Commonly used measures are

1) **Variance**: It measures the degree of deviation of the data values from the mean of the distribution.

$$Var = \frac{\sum (x_i - \bar{x})^2}{N}$$

For grouped data:

$$Var = \frac{\sum f_i (x_i - \bar{x})^2}{N}, \quad N = \sum f_i$$

2) **Standard deviation**

$$\sigma = \sqrt{Var}$$

The Shapes of distribution

a) Skewness (the lack of symmetry)

b) Kurtosis (enable us to have an idea about flatness and peakedness of the curve)

Skewness

A distribution, or data set, is symmetric if it looks the same to the left and right of the center point.

Symmetric distribution

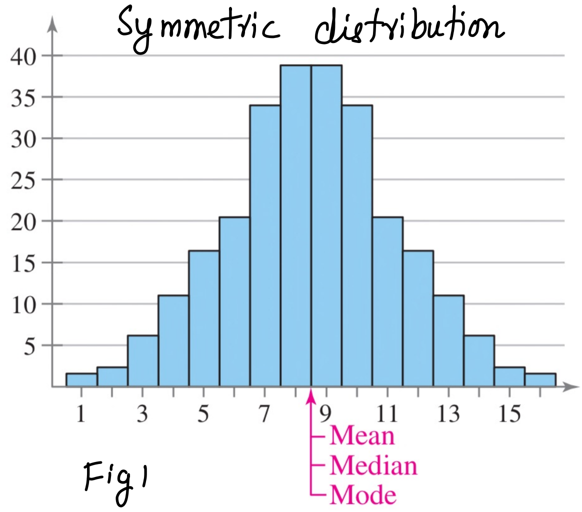
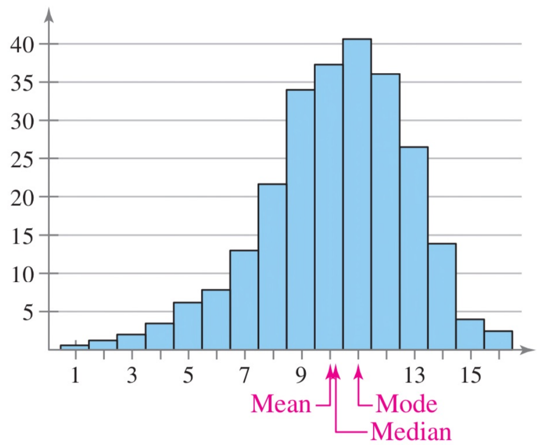


Fig 1

mean = median = mode

No skewness

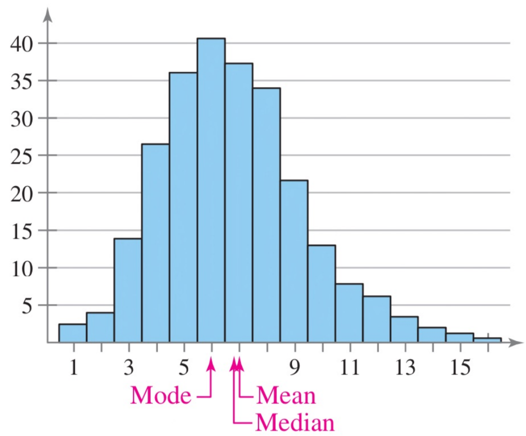
Skewness measures the degree and direction of departure from symmetry of a distribution.



skewed left (negatively skewed) distribution

Here tail is extended to the left

mean < median < mode.



skewed right (positively skewed) distribution

Tail is extended to the right

mean > median > mode

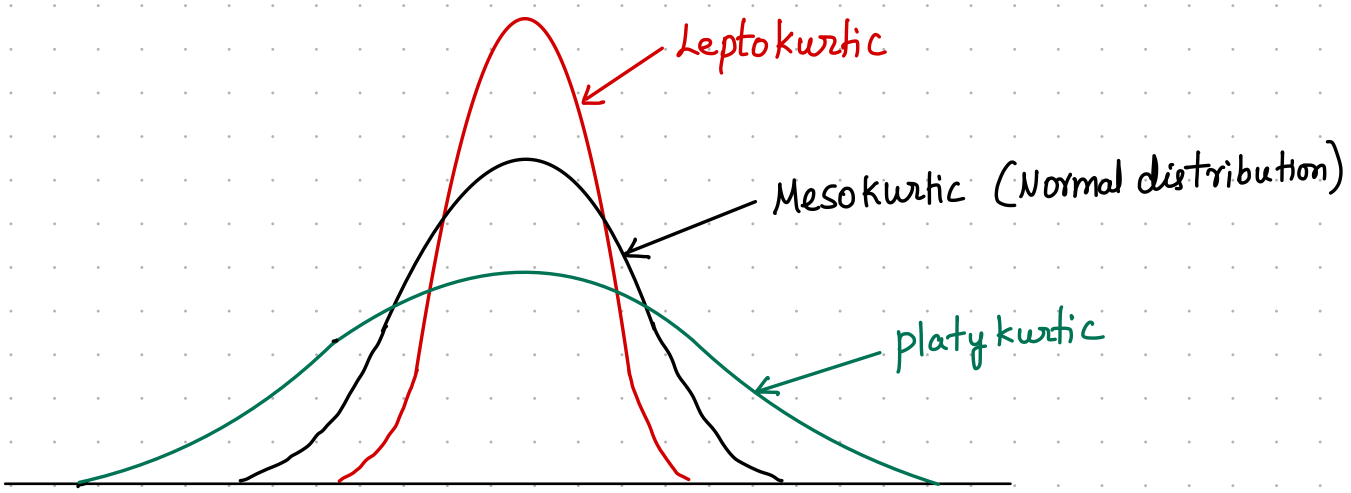
Kurtosis

It measures the thickness of the tail ends of a distribution in relation to the tails of a normal distribution.

Dataset with high kurtosis tend to have a distinct peak near the mean, decline rapidly, and have heavy tails.
(Leptokurtic distribution)

Data set with low kurtosis tend to have a flat top near the mean and have light tails
(platykurtic distribution)

Normal distribution is Mesokurtic distribution



Moments

It is a statistical measure used to describe and analyse the characteristic of a frequency distribution namely central tendency, dispersion, skewness and kurtosis.

Moments about mean (Central moment)

The r^{th} moment about mean (r^{th} central moment)

$$\mu_r = \frac{\sum f_i (x_i - \bar{x})^r}{N}, \quad r = 0, 1, 2, \ldots$$

where $N = \sum f_i$

In particular

$$\mu_0 = 1$$

$$\mu_1 = \frac{\sum f_i (x_i - \bar{x})}{N} = 0$$

$$\mu_2 = \frac{\sum f_i (x_i - \bar{x})^2}{N} = \text{Variance}$$

$$\mu_3 = \frac{\sum f_i (x_i - \bar{x})^3}{N} \quad \text{and} \quad \mu_4 = \frac{\sum f_i (x_i - \bar{x})^4}{N}$$

Moments about origin

The r^{th} moment about origin

$$\mu_r' = \frac{\sum f_i x_i^r}{N}$$

$r = 0, 1, 2, \ldots$

In particular

$$\mu_0' = 1$$

$$\mu_1' = \frac{\sum f_i x_i}{N} = \bar{x} \quad (\text{mean})$$

$$\mu_2' = \frac{\sum f_i x_i^2}{N}$$

$$\vdots$$

Moments about any point (Raw moments)

Let 'a' be arbitrary number. Then

$$\mu_r' = \frac{\sum f_i (x_i - a)^r}{N}, \qquad r = 0, 1, 2, \ldots$$

Relation between μ_r and μ_r'

$$\mu_2 = \mu_2' - (\mu_1')^2$$

$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2(\mu_1')^3$$

$$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'(\mu_1')^2 - 3(\mu_1')^4$$

In general

$$\mu_r = \mu_r' - {}^rC_1\, \mu_{r-1}'\, \mu_1' + {}^rC_2\, \mu_{r-2}'(\mu_1')^2 - {}^rC_3\, \mu_{r-3}'(\mu_1')^3 + \cdots + (-1)^r (\mu_1')^r$$

Conversely

$$\mu_2' = \mu_2 + (\mu_1')^2$$

$$\mu_3' = \mu_3 + 3\mu_2\mu_1' + (\mu_1')^3$$

$$\mu_4' = \mu_4 + 4\mu_3\mu_1' + 6\mu_2(\mu_1')^2 + (\mu_1')^4$$

In general

$$\mu_r' = \mu_r + {}^rC_1\, \mu_{r-1}\, \mu_1' + {}^rC_2\, \mu_{r-2}(\mu_1')^2 + {}^rC_3\, \mu_{r-3}(\mu_1')^3 + \cdots + (\mu_1')^r$$

Measure of skewness based on moments

*) If $\mu_3 = 0$, no skewness

*) If $\mu_3 > 0$, positive skewness

*) If $\mu_3 < 0$, negative skewness

Coefficient of skewness:

$$S_k = \frac{\sqrt{\beta_1}\,(\beta_2 + 3)}{2(5\beta_2 - 6\beta_1 - 9)}$$

Beta coefficient: $\beta_1 = \dfrac{\mu_3^2}{\mu_2^3}$

gamma coefficient $\gamma_1 = \pm\sqrt{\beta_1} = \dfrac{\mu_3}{\mu_2^{3/2}}$

sign of γ_1 depends on μ_3.

Measure of kurtosis based on moments

$$\beta_2 = \frac{\mu_4}{\mu_2^2}$$

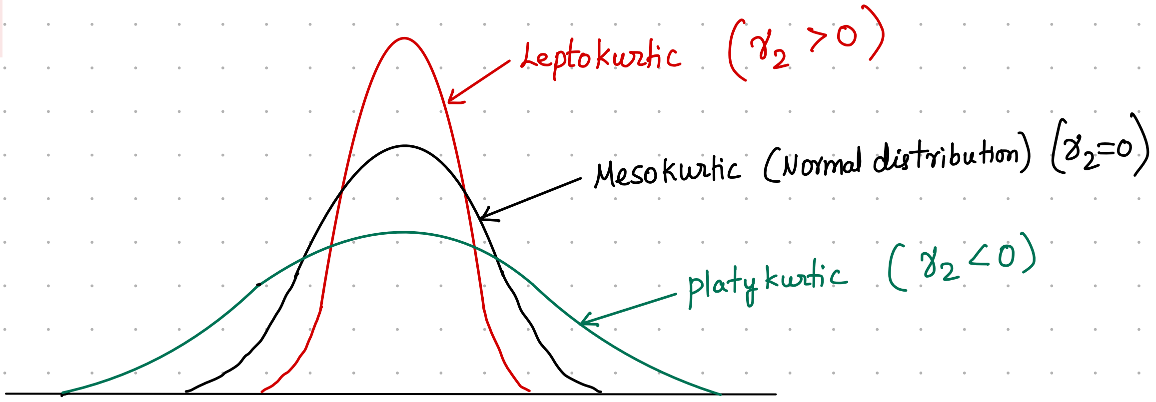
*) For mesokurtic (normal distribution), $\beta_2 = 3$

*) For leptokurtic, $\beta_2 > 3$

*) For platykurtic, $\beta_2 < 3$

The measure of kurtosis is also represented by gamma as

$$\gamma_2 = \beta_2 - 3$$



Ex1: The first four moments about the value 28.5 of a distribution are

0.294, 7.144, 42.409, and 454.98

Calculate moments about mean. Also find β_1 and β_2.

Ex2: Calculate μ_1, μ_2, μ_3 and μ_4 for the following frequency distribution. Also find β_1 and β_2

Marks	0–10	10–20	20–30	30–40	40–50	50–60
No. of students	1	6	10	15	11	7

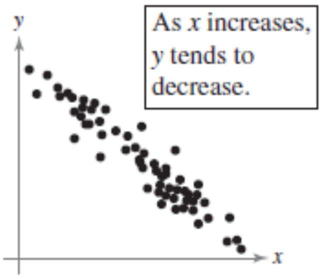
Ex3: From the foll. freq. distribution compute 1st four central moments

X	5	10	15	20	25	30	35
f	4	10	20	36	16	12	2

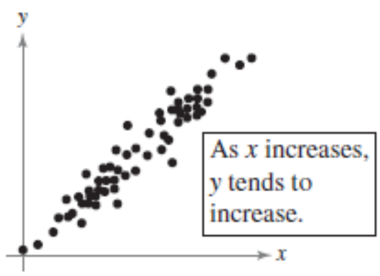
Correlation and Regression

Definition: A correlation is a relationship between two quantitative variables.

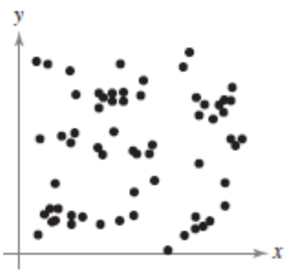
- The data can be represented by ordered pairs (x, y).
- The graph of ordered pair is called a scatter plot.



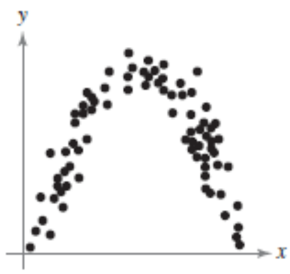
Negative Linear Correlation



Positive Linear Correlation



No Correlation



Nonlinear Correlation

Correlation coefficient

Definition: The numerical measure of linear correlation is called the correlation coefficient r. A formula for r is

$$r = \frac{\sum\left((x-\bar{x})(y-\bar{y})\right)}{\sqrt{\sum(x-\bar{x})^2}\sqrt{\sum(y-\bar{y})^2}}.$$

Or

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{n\sum x^2 - (\sum x)^2}\sqrt{n\sum y^2 - (\sum y)^2}}.$$

Note:

- Correlation coefficient r always lies between -1 and 1, that is $-1 \leq r \leq 1$.
- Correlation coefficient r is positive, if it lies between 0 and 1 and negative if $-1 \leq r < 0$.

- If x and y have a strong positive linear correlation, r is close to 1.
- If x and y have a strong negative linear correlation, r is close to -1.
- If there is no linear correlation or a weak linear correlation, r is close to 0.

Regression lines

Regression line, is a line of best fit for linearly correlated pair of data (x, y)

- We use the method of least squares to find regression lines.

The equation of regression line (y on x) for an independent variable x and a dependent variable y is

$$y = mx + b \ldots \ldots . . (1)$$

We find method of least squares to find $m$and b. Normal equations for (1) are

$$\sum y = m\sum x + bn \ldots \ldots \ldots \ldots . (2)$$

$$\sum xy = m\sum x^2 + b\sum x \ldots \ldots . . (3)$$

Divide (2) by n,

$$\frac{\sum y}{n} = m\frac{\sum x}{n} + b,$$

implies that

$$\overline{y} = m\overline{x} + b \ldots \ldots (4)$$

(4) implies that regression lines pass through the point $(\overline{x}, \overline{y})$. In view of equations (1) and (4), we see that

$$y - \overline{y} = m(x - \overline{x}) \ldots \ldots (5)$$

Normal equation for (5) is

$$\sum(x - \overline{x})(y - \overline{y}) = m\sum(x - \overline{x})^2 \ldots \ldots . (6)$$

From (4) and (6), we have

$$slope: \ m = \frac{\sum\big((x - \overline{x})(y - \overline{y})\big)}{\sum(x - \overline{x})^2} = \frac{n\sum xy - \sum x\sum y}{n\sum x^2 - (\sum x)^2} \text{ and } \ intercept: \ b = \overline{y} - m\overline{x}$$

Hence regression line of y on x is:

$$y - \overline{y} = \left(\frac{\sum\big((x - \overline{x})(y - \overline{y})\big)}{\sum(x - \overline{x})^2}\right)(x - \overline{x})$$

- If y is independent variable and x is dependent, then regression line of x on y is

$$x - \overline{x} = \left(\frac{\sum\big((x - \overline{x})(y - \overline{y})\big)}{\sum(y - \overline{y})^2}\right)(y - \overline{y})$$

Note that slopes of two regression lines are called coefficient of regressions and their product is r^2.

Example: An economist wants to determine whether there is a linear relationship between a country's gross domestic product (GDP) and carbon dioxide (CO_2) emissions. The data are shown in the table:

GDP(trillions of $), x	CO_2 emission(millions of metric tons), y
1.6	428.2
3.6	828.8
4.9	1214.2
1.1	444.6
0.9	264.0
2.9	415.3
2.7	571.8
2.3	454.9
1.6	358.7
1.5	573.5