MEASURES OF CENTRAL TENDENCY OR AVERAGES

- The tendency to concentrate at certain values , usually somewhere in the centre of the distribution. Measures of this tendency are called averages.
- An average of a statistical series is the value of the variable which is representative of the entire distribution. The following are the five measures of central tendency that are in common use:
- Arithmetic mean
- Geometric mean
- Harmonic mean
- Median
- Mode

Arithmetic Mean: Arithmetic mean of a set of observations is their sum divided by the number of observations.

e.g. The arithmetic mean $\bar{x}$ of n observations $x_1, x_2, x_3, \ldots\ldots, x_n$ is given by

$$\bar{x} = \frac{1}{n}\left[x_1 + x_2 + x_3 + \ldots\ldots + x_n\right] = \frac{1}{n}\sum_{i=1}^{n} x_i$$

In case of frequency distribution

$$\begin{array}{lccccc} x: & x_1 & x_2 & x_3 & \cdots\cdots & x_n \\ f: & f_1 & f_2 & f_3 & \cdots\cdots & f_n \end{array}$$

$$\bar{x} = \frac{f_1x_1 + f_2x_2 + f_3x_3 + \cdots\cdots + f_nx_n}{f_1 + f_2 + f_3 + \cdots\cdots + f_n} = \frac{\sum_{i=1}^{n} f_ix_i}{\sum_{i=1}^{n} f_i}$$

Note: In case of grouped or continuous frequency distribution. x is taken as the mid. value of the corresponding class.

- If values of x and f are large, the prefer following formulas for arithmetic mean.

$let\ d_i = \frac{x_i - A}{h}$ where A is any arbitrary point and h is the common magnitude of class interval

$$\bar{x} = A + \frac{h}{N}\sum_{i=1}^{n} f_i d_i \;;\; N = \sum_{i=1}^{n} f_i$$

Geometric Mean

Geometric mean of a set of n observations is the nth root of their product.

Thus the geometric mean G of n observations $x_1, x_2, x_3, \ldots\ldots x_n$ is

$$G = (x_1 . x_2 . x_3 . \ldots\ldots x_n)^{1/n}$$

The computation is facilitated by the use of logarithms. Taking logarithms of both sides.

we get

$$\log G = \frac{1}{n}\left[\log x_1 + \log x_2 + \log x_3 \ldots\ldots\log x_n\right] = \frac{1}{n}\sum_{i=1}^{n}\log x_i$$

$$\text{or } G = \text{antilog}\left(\frac{1}{n}\sum_{i=1}^{n}\log x_i\right)$$

In case of frequency distribution (x_i, f_i); $i = 1, 2, 3, \ldots\ldots, n$

Geometric mean G is given by

$$G = \left(x_1^{f_1}, x_2^{f_2}, x_3^{f_3}, \ldots\ldots\ldots x_n^{f_n}\right)^{1/N} \quad ; \; N = \sum_{i=1}^{n} f_i$$

or G =antilog $\left(\frac{1}{N}\sum_{i=1}^{n} f_i \log x_i\right)$

Harmonic Mean

- Harmonic mean of number of observations, none of which is zero, is the reciprocal of the arithmetic mean of the: reciprocals of the given values.

Thus, harmonic mean H, of n observations $x_1, x_2, x_3, \ldots\ldots\ldots x_n$ is

$$H = \frac{1}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}}$$

In case of frequency distribution (x_i, f_i); $i = 1, 2, 3, \ldots\ldots n$

Harmonic mean H is given by

$$H = \frac{1}{\frac{1}{N}\sum_{i=1}^{n}\left(\frac{f_i}{x_i}\right)}$$

<u>Mode</u>

- Let us consider the following statements:
- The average height of an Indian (male) is 5'-6".
- The average size of tile shoes sold in a shop is 7.
- An average student in a hostel spends Rs.l50 p.m.

In all the above cases, the average referred to is mode. Mode is the value which occurs most frequently in a set of observations and around which the other items of the set cluster densely. In other words, mode is the value of the variable which is predominant in the series. Thus in the case of discrete frequency distribution mode is the value of corresponding to maximum frequency.

- But in anyone (or-more) of the following cases : (i) if the maximum frequency is repeated, (ii) if the maximum frequency occurs in the very beginning or at the end of the distribution, and ' (iii) if there are irregularities in the distribution, the value of mode is determined by the method of grouping.

Mode for continuous frequency distribution:

$$\text{Mode} = l + \frac{h(f_1 - f_0)}{(f_1 - f_0) - (f_2 - f_1)} = l + \frac{h(f_1 - f_0)}{2f_1 - f_0 - f_2}$$

Where l is the lower limit, h is the magnitude and f_1 f the frequency of the the modal class and f_0 and f_2 are frequencies of the class preceding and succeeding the modal class respectively.

Modal class: class corresponds to maximum frequency of the data.

Median

Median of a distribution is the value of the variable which divides it into two equal parts. It is the value which exceeds and is exceeded by the same number of observations, i.e., it is the value such that the number of observations above it is equal to the number of observations below it The median is thus a positional average.

In case of ungrouped data, if the number of observations is odd then median is the middle value after the values have been arranged in ascending or descending order of magnitude and In case of even number of observations, there are two middle terms and median is obtained by taking the arithmetic mean of the middle terms. For example, the median of the values 25, 20,15,35,18, i.e. of 15, 18, 20, 25 , 35 is 20

and the median of 8, 20, 50, 25, 15, 30, i.e. of 8, 15, 20, 25, 30, 50' is (20 +25)/2 = 22·5.

In case of discrete frequency distribution median is obtained by considering the cumulative frequencies. The steps for calculating median are :

(i) Find N/2, where $N = \sum_{i=1}^{n} f_i$

(ii) See the cumulative frequency (c.f.) just greater than N/2

(iii) The corresponding value of x is median.

Median for continuous frequency distribution:

- In the case of continuous frequency distribution, the class corresponding to the c.f. just greater than N/2 is called the median class and the value of median is obtained by the following formula:

$$\text{Median} = l + \frac{h}{f}\left(\frac{N}{2} - C\right)$$

where l is the lower limit of the median class, f is the frequency of the median class, h is the magnitude of the median class, 'c' is the c.f. of the class preceding the median class, and $N = \sum_{i=1}^{n} f_i$

Measures of Dispersion, Skewness and Kurtosis

Measures of central tendency are inadequate to give us a complete idea of the distribution. They must be supported and supplemented by some other measures, One such measure is Dispersion. Literal meaning of dispersion is scatteredness.

The following are the measures of dispersion

- Range,
- Quartile deviation
- Mean deviation
- Standard deviation.

Range

If A and B are the greatest and smallest observations respectively in a distribution, then it's range is given by :

Range = A – B

Range is the simplest but not a reliable measure of dispersion.

Quartile Deviation

$Q = \frac{1}{2}(Q_3 - Q_1)$. where Q_1 and Q_3 are the 1st and 3rd quartiles of the distribution respectively

Mean Deviation

Mean Deviation from average A $= \frac{1}{N}\sum_{i=1}^{n} f_i |x_i - A|$; $\sum_{i=1}^{n} f_i = N$

Standard deviation

Standard deviation is the positive square root of the A.M. of the squares of the deviations of the given values from their arithmetic mean

It is usually denoted by sigma 'σ'.

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{n} f_i\left(x_i - \bar{x}\right)^2} \; ; \quad \text{where } \bar{x} = \text{Arithmetic Mean and } \sum_{i=1}^{n} f_i = N$$

The square of standard deviation is called the variance and is given by

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{n} f_i\left(x_i - \bar{x}\right)^2$$

$$= \frac{1}{N}\sum_{i=1}^{n} f_i x_i^2 - \left(\bar{x}\right)^2$$

Note: Variance and standard deviation are independent of change of origin but not of scale.

$$\sigma_x^2 = \frac{1}{N}\sum_{i=1}^{k} f_i\left(x_i - \bar{x}\right)^2 = \frac{1}{N}\sum_{i=1}^{k} f_i\left(d_i - \bar{d}\right)^2 = \sigma_d^2 \; ; d_i = x_i - A$$

$$\text{If } d_i = \frac{x_i - A}{h} \quad \text{then} \quad \sigma_x^2 = \frac{1}{N}\sum_{i=1}^{k} f_i\left(x_i - \bar{x}\right)^2 = \frac{h^2}{N}\sum_{i=1}^{k} f_i\left(d_i - \bar{d}\right)^2 = h^2\sigma_d^2$$

Coefficient of variation:

$$\text{Coefficient of variation, } C.V. = 100 \times \frac{\sigma}{\bar{x}}$$

Note. 1. C.V. is the total variation in the mean.
2. for computing the variability of two series calculate C.V. for each series the series having greater C.V. is said to more variable than the other and series having lesser C.V. is more consistent than the other.

Moments

The r^{th} moment of a variable x about any point $x = A$, usually denoted by $\mu_r^{'}$ is given by

$$\mu_r' = \frac{1}{N}\sum_{i=1}^{n} f_i (x_i - A)^r$$

$$= \frac{1}{N}\sum_{i=1}^{n} f_i (d_i)^r \qquad ; d_i = x_i - A,\ \mathrm{N} = \sum_{i=1}^{n} f_i$$

The r^{th} moment of a variable x about mean $\bar{x}$, denoted by μ_r is given by

$$\mu_r = \frac{1}{N}\sum_{i=1}^{n} f_i \left(x_i - \bar{x}\right)^r$$

$$= \frac{1}{N}\sum_{i=1}^{n} f_i (z_i)^r \qquad ; \mathrm{z}_i = x_i - \bar{x},\ \mathrm{N} = \sum_{i=1}^{n} f_i$$

Relation between moments about mean in terms of moment about any point:

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

Pearson's β and γ coefficients

Karl Pearson defined the following four coefficients based upon the first four moments about mean:

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} \; ; \; \gamma_1 = +\sqrt{\beta_1}$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} \; ; \; \gamma_2 = \beta_2 - 3$$

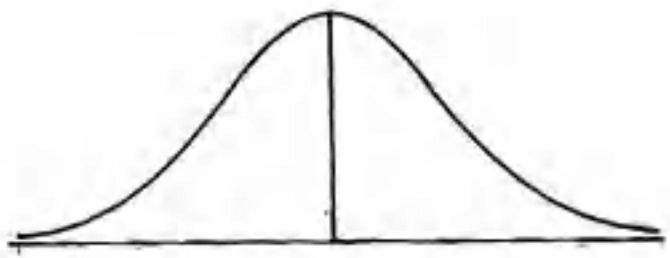
Skewness

Skewness means lack of symmetry. We study skewness to have an idea about the shape of the curve which we draw with the help of the given data. A curve is said to be skewed if:

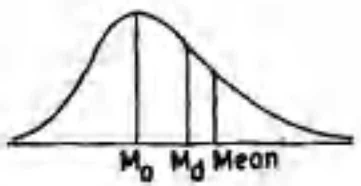
(i) Mean, median and mode fall at different points.

(ii) Quartiles are not equidistant from median and

(iii) The curve drawn with given data is not symmetrical but stretched more to one side than to the other.



$\bar{x}$ (*Mean*) = M_0 = M_d
(Symmetrical Distribution)



(Positively Skewed Distribution) (Negatively Skewed Distribution)

Measurement of Skewness:

(1) Karl Pearson's coefficient of skewness:

$$S_k = \frac{(M - M_0)}{\sigma}$$

$$= \frac{3(M - M_d)}{\sigma}$$ where σ is S.D. , M is mean, M_0 is mode and M_d is median.

Skewness is positive if $M > M_0$ or $M > M_d$ and is negative if $M < M_0$ or $M < M_d$.

(2) Bowley's coefficient of Skewness (Skewness based on Quartiles)

$$S_k = \frac{(Q_3 - M_d) - (M_d - Q_1)}{(Q_3 - M_d) + (M_d - Q_1)}$$

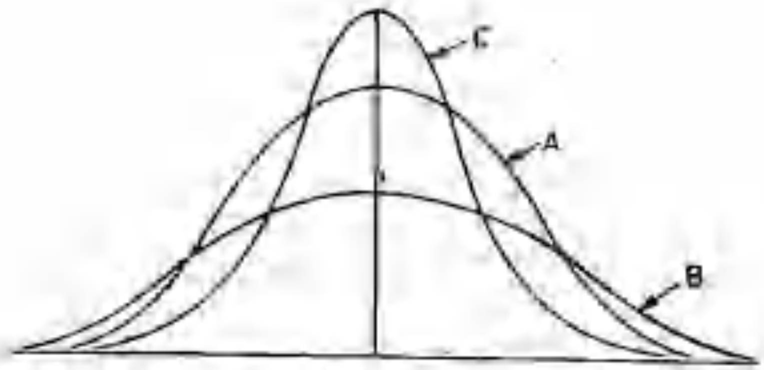
(3) Coefficient of Skewness based upon Moments:

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

Kurtosis

Convexity of the frequency curve or kurtosis enables us to have an idea about the flateness or peakedness of the frequency curve.

It is measured by the coefficients β_2 and γ_2 .



Curve of the type 'A' which is neither flat nor peaked is called the *normal curve* or *mesokurtic curve* and for such a curve $\beta_2 = 3$, i.e., $\gamma_2 = 0$. Curve of the type 'B' which is flatter than the normal curve is known as *platykurtic* and for such a curve $\beta_2 < 3$, i.e., $\gamma_2 < 0$. Curve of the type 'C' which is more peaked than the normal curve is called *leptokurtic* and for such a curve $\beta_2 > 3$, i.e., $\gamma_2 > 0$.

PROBABILITY AND PROBABILITY DISTRIBUTIONS

Random Experiment or Trail

- **Trial and Event**: Consider an experiment which, though repeated under essentially identical conditions, does not give unique results but may result in any one of the several possible outcomes .The experiment is known as a trial and the outcomes are known as events or casts

- For example: (i) Throwing of a die is a trial and getting 1 (or 2 or 3, ... or 6) is an event
- (ii) Tossing of a coin is a trial and getting head (H) or tail (T) is an event

- **Sample Space**: the set of all possible outcomes is called the sample space for particular experiment and is denoted by S.
- **Exhaustive Events**: The total number of possible outcomes in any trial is known as exhaustive events or exhaustive cases.

- For example: In tossing of a coin there are two exhaustive cases, viz., head and. Tail
- **Mutually exclusive events**: Events are said to be mutually exclusive or incompatible jf the happening of anyone of them precludes the happening of all the others (i.e., if no two Or more of them can happen simultaneously in the same. trial.

- For example: (i) In throwing a die all the 6 faces' numbered 1 to 6 are mutually exclusive since if anyone of these faces comes, ,the possibility of other in the same trial, is ruled out.
- **Independent events**: Several events are said to be independent if the happening (or non-happening) of an event is not affected by the supplementary knowledge concerning the occurrence of any number of the remaining events.

- For example: In tossing an unbiased coin the event of getting a head in the first toss is independent of getting a head in the second, third and subsequent throws.

Mathematical or Classical definition of probability:

If a trial results in n exhaustive, mutually exclusive and equally likely cases and m of them are favourable to the happening of an event E. Then the probability 'p' of happening of E is given by

$$p = P(E) = \frac{\text{Favourable number of cases}}{\text{Exhaustive number of cases}} = \frac{m}{n}$$

and $P(\bar{E}) = 1 - P(E) = 1 - \frac{m}{n}$

Statistical definition of Probability:

If in n trails, an event E happens m times, the Probability of happening of E is

$$P(E) = \lim_{n \to \infty} \frac{m}{n}$$

Note: (1) For any two events A and B, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

(2) P(S)=1, $0 \le P(E) \le 1$, $P(\phi) = 0$

Conditional Probability:

The probability of happening of an event E_1 when another event E_2 is known to have already happened is called conditional probability and defined as:

$$P(E_1/E_2) = \frac{P(E_1 \cap E_2)}{P(E_2)}$$

For mutually independent events $P(E_1 \cap E_2) = P(E_1).P(E_2)$

In general

$P(E_1 \cap E_2 \cap E_3 \cap \ldots \cap E_n) = P(E_1).P(E_2).P(E_3) \ldots P(E_n)$ for independent events.

Bayes Theorem. *If $E_1, E_2, \ldots, E_n$ are mutually disjoint events with $P(E_i) \neq 0$, $(i = 1, 2, \ldots, n)$ then for any orbitrary event A which is a subset of $\bigcup_{i=1}^{n} E_i$ such that $P(A) > 0$, we have*

$$P(E_i \mid A) = \frac{P(E_i)\, P(A \mid E_i)}{\sum_{i=1}^{n} P(E_i)\, P(A \mid E_i)}, \quad i = 1, 2, \ldots, n.$$

- **Random variables**
- Random variable is a function say X that assigns a unique real value to each outcomes of the random experiment.
- Thus a rule that assigns a real number to each outcome is called random variable.
- Eg. Let us tossing of two coins
- Then S= {HH,HT,TH,TT}
- Let us define Random variable X = No. of heads
- =\{0,1,2\}

- Random variables are of two types:
- Discrete Random variable
- Continuous Random variable

- Discrete probability distributions:
- **Binomial distribution**:

A random variable X is said to follow binomial distribution if it assumes only non-negative values and its probability mass function is given by

$$P(X=x)=p(x)=\begin{cases}\binom{n}{x}p^x q^{n-x}; x=0,1,2,\dots,n;\ q=1-p\\ 0,\ otherwise\end{cases}$$

The two independent constants n and p in the distribution are known as the *parameters* of the distribution. 'n' is also, sometimes, known as the degree of the binomial distribution.

Binomial distribution is a discrete distribution as X can take only the integral values, *viz.*, 0, 1, 2,..., n. Any variable which follows binomial distribution is known as *binomial variate.*

We shall use the notation $X \sim B(n,p)$ to denote that the random variable X, follows binomial distribution with parameters n and p.

Moments. The first four moments about origin of binomial distribution are obtained as follows :

$$\mu_1' = E(X) = \sum_{x=0}^{n} x \binom{n}{x} p^x q^{n-x} = np \sum_{x=1}^{n} \binom{n-1}{x-1} p^{x-1} q^{n-x}$$

$$= np(q+p)^{n-1} = np \qquad (\because q + p = 1)$$

Thus the mean of the binomial distribution is np.

$$\mu_2' = E(X^2) = \sum_{x=0}^{n} x^2 \binom{n}{x} p^x q^{n-x}$$

$$= \sum_{x=0}^{n} \{x(x-1) + x\} \frac{n(n-1)}{x(x-1)} \cdot \binom{n-2}{x-2} p^x q^{n-x}$$

$$= n(n-1)p^2 \left[\sum_{x=2}^{n} \binom{n-2}{x-2} p^{x-2} q^{n-x} \right] + np$$

$$= n(n-1)p^2 (q+p)^{n-2} + np = n(n-1)p^2 + np$$

- **Poisson Distribution**

Poisson Distribution (as a limiting case of Binomial Distribution). Poisson distribution was discovered by the French mathematician and physicist Simeon Denis Poisson (1781—1840) who published it in 1837. Poisson distribution is a limiting case of the binomial distribution under the following conditions:

(*i*) n, the number of trials is indefinitely large, *i.e.*, $n \to \infty$.

(*ii*) p, the constant probability of success for each trial is indefinitely small, *i.e.*, $p \to 0$.

(*iii*) $np = \lambda$, (say), is finite. Thus $p = \lambda/n$, $q = 1 - \lambda/n$, where λ is a positive real number.

$$\lim_{n \to \infty} b(x; n, p) = \frac{\lambda^x}{e^x \cdot x!} \cdot \frac{e^{-\lambda} \cdot 1}{e^{-x} \cdot 1} = \frac{e^{-\lambda} \cdot \lambda^x}{x!}; x = 0, 1, 2, ..., \infty;$$

which is the required probability function of the Poisson distribution. 'λ' is known as the parameter of Poisson distribution.

Moments of the Poisson Distribution

$$\mu_1' = E(X) = \sum_{x=0}^{\infty} x\, p(x, \lambda)$$

$$= \sum_{x=0}^{\infty} x \cdot \frac{e^{-\lambda}\lambda^x}{x!} = \lambda e^{-\lambda}\left[\sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!}\right]$$

$$= \lambda e^{-\lambda}\left(1 + \lambda + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + \dots\right) = \lambda e^{-\lambda} \cdot e^{\lambda} = \lambda$$

Hence the mean of the Poisson distribution is λ.

- **Normal Distribution**

Normal Distribution: The normal distribution was first discovered in 1733 by English mathematician De Moivre, who obtained this continuous distribution as a limiting case of the binomial distribution and applied it to problems arising in the game of chance. It was also known to Laplace, no later than 1774 but through a historical error it was credited to Gauss, who first made reference to it in the beginning of 19th century (1809), as the distribution of errors in Astronomy. Gauss used the normal curve to describe the theory of accidental errors of measurements involved in the calculation of orbits of heavenly bodies. Throughout the eighteenth and nineteenth centuries, various efforts were made to establish the normal model as the underlying law ruling all continuous random variables. Thus, the name "normal". These efforts, however, failed because of false premises. The normal model has, nevertheless, become the most important probability model in statistical analysis.

Definition. *A random variable X is said to ahve a normal distribution with parameters μ (called "mean") and σ^2 (called "variance") if its density function is given by the probability law :*

$$f(x;\mu,\sigma)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left\{\frac{x-\mu}{\sigma}\right\}^2\right]$$

or $$f(x;\mu,\sigma)=\frac{1}{\sigma\sqrt{2\pi}}e^{-(x-\mu)^2/2\sigma^2}$$

$$-\infty<x<\infty,\ -\infty<\mu<\infty,\ \sigma>0$$

Remarks. 1. A random variable X with mean μ and variance σ^2 and following the normal law is expressed by $X \sim N(\mu, \sigma^2)$

2. If $X \sim N(\mu, \sigma^2)$, then $Z = \dfrac{X-\mu}{\sigma}$, is a standard normal variate with

$$E(Z) = 0 \text{ and } \operatorname{Var}(Z) = 1$$

and we write $Z \sim N(0,1)$.

3. The p.d.f. of standard normal variate Z is given by

$$\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \quad -\infty < z < \infty$$

and the corresponding distribution function, denoted by $\Phi(z)$ is given by

$$\Phi(z) = P(Z \leq z) = \int_{-\infty}^{z} \varphi(u)\, du$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-u^2/2}\, du$$

Chief Characteristics of the Normal Distribution and Normal Probability Curve. The normal probability curve with mean μ and standard deviation σ is given by the equation

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}, \quad -\infty < x < \infty$$

and has the following properties :

(*i*) The curve is bell shaped and symmetrical about the line $x = \mu$.

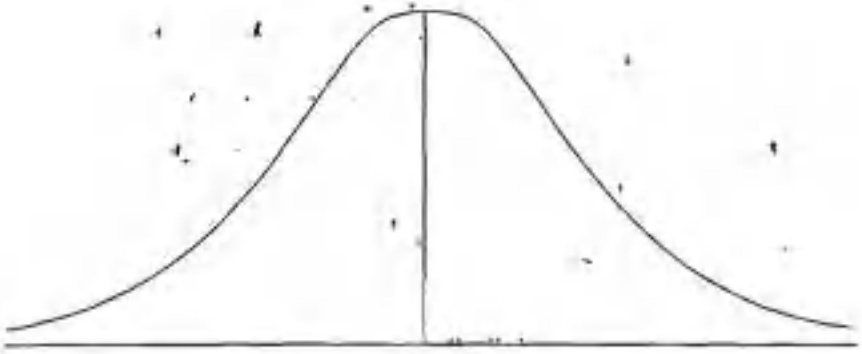
(*ii*) Mean, median and mode of the distribution coincide.

(*iii*) As x increases numerically, $f(x)$ decreases rapidly, the maximum probability occurring at the point $x = \mu$, and given by $[p(x)]_{max} = \dfrac{1}{\sigma\sqrt{2\pi}}$.

(*iv*) $\beta_1 = 0$ and $\beta_2 = 3$.

(*v*) $\mu_{2r+1} = 0, (r = 0, 1, 2, \ldots)$,
and $\mu_{2r} = 1.3.5 \ldots (2r-1)\sigma^{2r}, (r = 0, 1, 2, \ldots)$.

(*vi*) Since $f(x)$ being the probability, can never be negative, no portion of the curve lies below the x-axis.



(Normal Probability Cuve)

Curve Fitting and Principle of Least Squares

Curve Fitting. Let (x_i, y_i) ; $i = 1, 2, \ldots, n$ be a given set of n pairs of values, X being independent variable and Y the dependent variable. The general problem in curve fitting is to find, if possible, an analytic expression of the form $y = f(x)$, for the functional relationship suggested by the given data.

Fitting of curves to a set of numerical data is of considerable importance-theoretical as well as practical. Theoretically it is useful in the study of correlation and regression, *e.g.*, lines of regression can be regarded as fitting of linear curves to the given bivariate distribution

Fitting of a straight line. Let us consider the fitting of a straight line

$$Y = a + bX$$

$$\sum_{i=1}^{n} y_i = na + b \sum_{i=1}^{n} x_i \quad \text{and} \quad \sum_{i=1}^{n} x_i y_i = a \sum_{i=1}^{n} x_i + b \sum_{i=1}^{n} x_i^2$$

) are known as the *normal equations* for estimating a and b.

All the quantities $\sum_{i=1}^{n} x_i$, $\sum_{i=1}^{n} x_i^2$, $\sum_{i=1}^{n} y_i$ and $\sum_{i=1}^{n} x_i y_i$, can be obtained from the given set of points (x_i, y_i); $i = 1, 2, \ldots, n$

Fitting of second degree parabola. Let

$$Y = a + bX + cX^2$$

be the second degree parabola of best fit to set of n points (x_i, y_i); $i = 1, 2, ..., n$. Using the principle of least squares, we have to determine the constants a, b and c so that

$$E = \sum_{i=1}^{n} (y_i - a - bx_i - cx_i^2)^2$$

is minimum.

$$\sum y_i = na + b \sum x_i + c \sum x_i^2$$
$$\sum x_i y_i = a \sum x_i + b \sum x_i^2 + c \sum x_i^3$$
$$\sum x_i^2 y_i = a \sum x_i^2 + b \sum x_i^3 + c \sum x_i^4,$$

summation taken over i from 1 to n

Fitting of a Power Curve $Y = aX^b$

to a set of n points.

Taking logarithm of both sides, we get

$$\log Y = \log a + b \log X$$

$$\Rightarrow \quad U = A + bV$$

where $U = \log Y$, $A = \log a$ and $V = \log X$.

This is a linear equation in U and V.

Normal equations for estimating A and b are

$$\Sigma U = nA + b\Sigma V \quad \text{and} \quad \Sigma UV = A\Sigma V + b\Sigma V^2$$

These equations can be solved for A and b and consequently for a

$a = \text{antilog}\,(A)$

Correlation and Regression

Bivariate Distribution, Correlation. So far we have confined ourselves to univariate distributions, *i.e.*, the distributions involving only one variable. We may, however, come across certain series where each term of the series may assume the values of two or more variables. For example, if we measure the heights and weights of a certain group of persons, we shall get what is known as *Bivariate distribution*—one variable relating to height and other variable relating to weight.

In a bivariate distribution we may be interested to find out if there is any correlation or covariation between the two variables under study. If the change in one variable affects a change in the other variable, the variables are said to be correlated. If the two variables deviate in the same direction, *i.e.*, if the increase (or decrease) in one results in a corresponding increase (or decrease) in the other, correlation is said to be *direct* or *positive*. But if they constantly deviate in the opposite directions, *i.e.*, if increase (or decrease) in one results in corresponding decrease (or increase) in the other, correlation is said to be *diverse* or *negative*. For example, the correlation between (*i*) the heights and weights of a group of persons, (*ii*) the income and expenditure is positive and the correlation between (*i*) price and demand of a commodity, (*ii*) the volume and pressure of a perfect gas, is negative. Correlation is said to be *perfect* if the deviation in one variable is followed by a corresponding and proportional deviation in the other.

Scatter Diagram. It is the simplest way of the diagrammatic representation of bivariate data. Thus for the bivariate distribution (x_i, y_i); $i = 1, 2, \ldots, n$, if the values of the variables X and Y be plotted along the x-axis and y-axis respectively in the xy plane, the diagram of dots so obtained is known as *scatter diagram*. From the scatter diagram, we can form a fairly good, though vague, idea whether the variables are correlated or not, *e.g.*, if the points are very dense, *i.e.*, very close to each other, we should expect a fairly good amount of correlation between the variables and if the points are widely scattered, a poor correlation is expected. This method, however, is not suitable if the number of observations is fairly large.

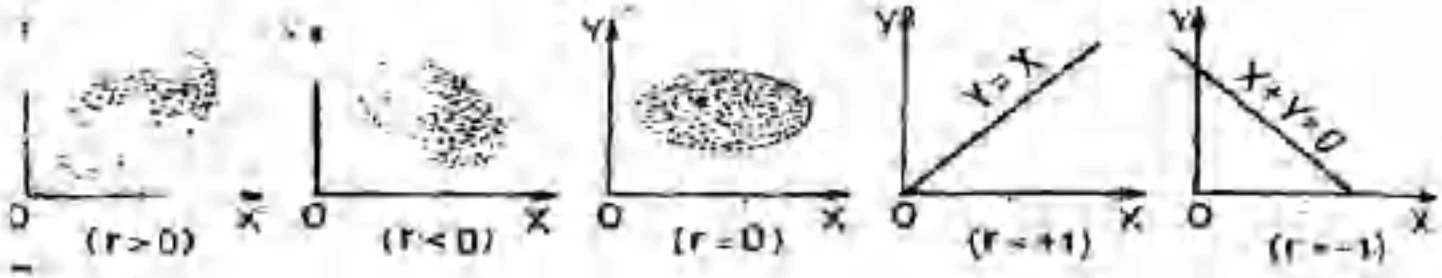
Karl Pearson Coefficient of Correlation.

Correlation coefficient between two random variables X and Y, usually denoted by $r(X, Y)$ or simply r_{XY}, is a numerical measure of *linear relationship* between them and is defined as

$$r(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

$$r(X, Y) = \frac{\mathrm{Cov}\,(X, Y)}{\sigma_X\,\sigma_Y} = \frac{\frac{1}{n}\Sigma\,(x_i - \bar{x})\,(y_i - \bar{y})}{\left[\frac{1}{n}\Sigma\,(x_i - \bar{x})^2 \,.\, \frac{1}{n}\Sigma\,(y_i - \bar{y})^2\right]^{1/2}},$$

Remarks 1. Following are the figures of the standard data for $r > 0$, < 0, $= 0$, and $r = \pm 1$.



2. It may be noted that $r(X, Y)$ provides a measure of *linear relationship* between X and Y. For nonlinear relationship, however, it is not very suitable.

3. Sometimes, we write : $\text{Cov}(X, Y) = \sigma_{XY}$

4. Karl Pearson's correlation coefficient is also called *product-moment correlation coefficient*, since

$\text{Cov}(X, Y) = E\{[X - E(X)]\,[Y - E(Y)]\} = \mu_{11}$.

Correlation coefficient is independent of change of origin and scale.

Let $U = \dfrac{X-a}{h}$, $V = \dfrac{Y-b}{k}$, so that $X = a + hU$ and $Y = b + kV$,

where a, b, h, k are constants; $h > 0$, $k > 0$.

$$\therefore r(X, Y) = r(U, V)$$

Rank Correlation. Let us suppose that a group of n individuals is arranged in order of merit or proficiency in possession of two characteristics A and B. These ranks in the two characteristics will, in general, be different. For example, if we consider the relation between intelligence and beauty, it is not necessary that a beautiful individual is intelligent also. Let (x_i, y_i); $i = 1, 2, \ldots, n$ be the ranks of the ith individual in two characteristics A and B respectively. Pearsonian coefficient of correlation between the ranks x_i's and y_i's is called the rank correlation coefficient between A and B for that group of individuals.

$$\rho = 1 - \frac{\sum_{i=1}^{n} d_i^2}{2n\sigma_X^2} = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)}$$

which is the *Spearman's formula for the rank correlation coefficient.*

Regression. *The term "regression"* literally means *"stepping back towards the average"*. It was first used by a British biometrician Sir Francis Galton (1822—1911), in connection with the inheritance of stature. Galton found that the offsprings of abnormally tall or short parents tend to "regress" or "step back" to the average population height. But the term "regression" as now used in Statistics is only a convenient term without having any reference to biometry.

Definition. *Regression analysis is a mathematical measure of the average relationship between two or more variables in terms of the original units of the data.*

Lines of Regression. If the variables in a bivariate distribution are related, we will find that the points in the scatter diagram will cluster round some curve called the "*curve of regression*". If the curve is a straight line, it is called the line of regression and there is said to be *linear regression* between the variables, otherwise regression is said to be *curvilinear*.

The line of regression is the line which gives the best estimate to the value of one variable for any specific value of the other variable. Thus the line of regression is the line of "*best fit*" and is obtained by the *principles of least squares*.

Let us suppose that in the bivariate distribution (x_i, y_i); $i = 1, 2, \ldots, n$; Y is dependent variable and X is independent variable. Let the line of regression of Y on X be $Y = a + bX$.

According to the principle of least squares, the normal equations for estimating a and b are

$$\sum_{i=1}^{n} y_i = na + b \sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} x_i y_i = a \sum_{i=1}^{n} x_i + b \sum_{i=1}^{n} x_i^2$$

$$Y - \bar{y} = r\frac{\sigma_Y}{\sigma_X}(X - \bar{x})$$

Starting with the equation $X = A + BY$

equation of the line of regression of X on Y becomes

$$X - \bar{x} = \frac{\mu_{11}}{\sigma_Y^2}(Y - \bar{y})$$

$$\Rightarrow \qquad X - \bar{x} = r\frac{\sigma_X}{\sigma_Y}(Y - \bar{y})$$

SAMPLING AND SAMPLING DISTRIBUTIONS

If the population is infinite, complete enumeration is not possible. Also if the units are destroyed in the course of inspection (*e.g.*, inspection of crackers, explosive materials, etc.), 100% inspection, though possible, is not at all desirable. But even if the population is finite or the inspection is not destructive, 100% inspection is not taken recourse to because of multiplicity of causes, *viz.*, administrative and financial implications, time factor, etc., and we take the help of *sampling*.

A finite subset of statistical individuals in a population is called a *sample* and *the* number of individuals in a sample is called the sample size.

For the purpose of determining population characteristics, instead of enumerating the entire population, the individuals in the sample only are observed. Then the sample characteristics are utilised to approximately determine or estimate the population. For example, on examining the sample of a particular stuff we arrive at a decision of purchasing or rejecting that stuff. The error involved in such approximation is known as *sampling error* and is inherent and unavoidable in any and every sampling scheme. But sampling results in considerable gains, especially in time and cost not only in respect of making observations of characteristics but also in the subsequent handling of the data.

Sampling is quite often used in our day-to-day practical life. For example, in a shop we assess the quality of sugar, wheat or any other commodity by taking a handful of it from the bag and then decide to purchase it or not. A housewife normally tests the cooked products to find if they are properly cooked and contain the proper quantity of salt.

Types of Sampling. Some of the commonly known and frequently used types of sampling are :

(*i*) *Purposive sampling.* (*ii*) *Random sampling.* (*iii*) *Stratified sampling.* (*iv*) *Systematic Sampling.*

Standard Error. The standard deviation of the sampling distribution of a statistic is known as its *Standard Error*, abbreviated as S.E. The standard errors of some of the well known statistics, *for large samples*, are given below, where n is the sample size, σ^2 the population variance, and P the population proportion, and $Q = 1 - P$. n_1 and n_2 represent the sizes of two independent random samples respectively drawn from the given population(s).

S.No.	*Statistic*	*Standard Error*
1.	Sample mean : $\bar{x}$	$\sigma/\sqrt{n}$
2.	Observed sample proportion 'p'	$\sqrt{PQ/n}$
3.	Sample s.d. : s	$\sqrt{\sigma^2/2n}$
4.	Sample variance : s^2	$\sigma^2\sqrt{2/n}$
5.	Sample quartiles	$1.36263\ \sigma/\sqrt{n}$
6.	Sample median	$1.25331\ \sigma/\sqrt{n}$

Tests of Significance. A very important aspect of the sampling theory is the study of the *tests of significance*, which enable us to decide on the basis of the sample results, if

(*i*) the deviation between the observed sample statistic and the hypothetical parameter value, or

(*ii*) the deviation between two independent sample statistics;

is significant or might be attributed to chance or the fluctuations of sampling.

Null Hypothesis. The technique of randomisation used for the selection of sample units makes the test of significance valid for us. For applying the test of significance we first set up a hypothesis—a definite statement about the population parameter. Such a hypothesis, which is usually a hypothesis of no difference, is called *null hypothesis* and is usually denoted by H_0. According to *Prof. R.A. Fisher, null hypothesis is the hypothesis which is tested for possible rejection under the assumption that it is true.*

For example, in case of a single statistic, H_0 will be that the sample statistic does not differ significantly from the hypothetical parameter value and in the case of two statistics, H_0 will be that the sample statistics do not differ significantly.

Alternative Hypothesis. Any hypothesis which is complementary to the null hypothesis is called an alternative hypothesis, usually denoted by H_1. For example, if we want to test the null hypothesis that the population has a specified mean μ_0, (say), *i.e.*, $H_0 : \mu = \mu_0$, then the alternative hypothesis could be

(*i*) $H_1 : \mu \neq \mu_0$ (*i.e.*, $\mu > \mu_0$ or $\mu < \mu_0$)

(*ii*) $H_1 : \mu > \mu_0$

(*iii*) $H_1 : \mu < \mu_0$

Errors in Sampling. The main objective in sampling theory is to draw valid inferences about the population parameters on the basis of the sample results. In practice we decide to accept or reject the lot after examining a sample from it. As such we are liable to commit the following two types of errors :

Type I Error : *Reject H_0 when it is true.*

Type II Error : *Accept H_0 when it is wrong, i.e., accept H_0 when H_1 is true.*

If we write,

$$\left.\begin{aligned} P(\text{Reject } H_0 \text{ when it is true}) &= P(\text{Reject } H_0 \mid H_0) = \alpha \\ \text{and } P(\text{Accept } H_0 \text{ when it is wrong}) &= P(\text{Accept } H_0 \mid H_1) = \beta \end{aligned}\right\}$$

then α and β are called the *sizes of type I error and type II error*, respectively.

In practice, type I error amounts to rejecting a lot when it is good and type II error may be regarded as accepting the lot when it is bad.

$$\left.\begin{aligned} \text{Thus} \quad P(\text{Reject a lot when it is good}) &= \alpha \\ \text{and} \quad P(\text{Accept a lot when it is bad}) &= \beta \end{aligned}\right\}$$

where α and β are referred to as *Producer's risk* and *Consumer's risk*, respectively.

Critical Region and Level of Significance. A region (corresponding to a statistic t) in the sample space S which amounts to rejection of H_0 is termed as *critical region* or *region of rejection*. If ω is the critical region and if $t = t(x_1, x_2, \ldots, x_n)$ is the value of the statistic based on a random sample of size n, then

$$P(t \in \omega \mid H_0) = \alpha, \quad P(t \in \bar{\omega} \mid H_1) = \beta$$

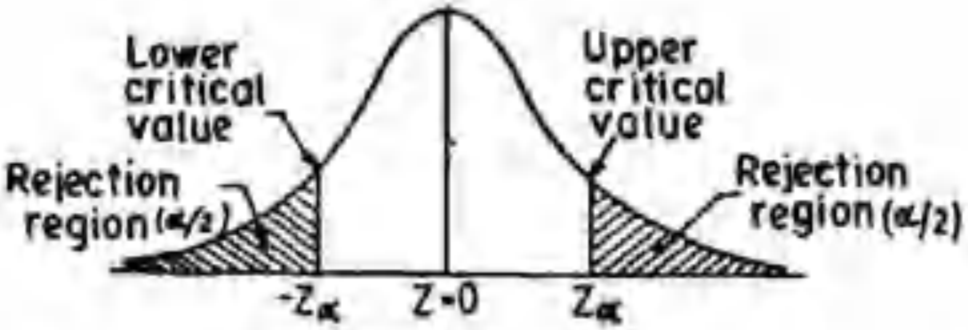
where $\bar{\omega}$, the complementary set of ω, is called the *acceptance region*.

Critical Values or Significant Values. The value of test statistic which separates the critical (or rejection) region and the acceptance region is called the *critical value* or *significant* value. It depends upon :

(*i*) The level of significance used, and

(*ii*) The alternative hypothesis, whether it is two-tailed or single-tailed.

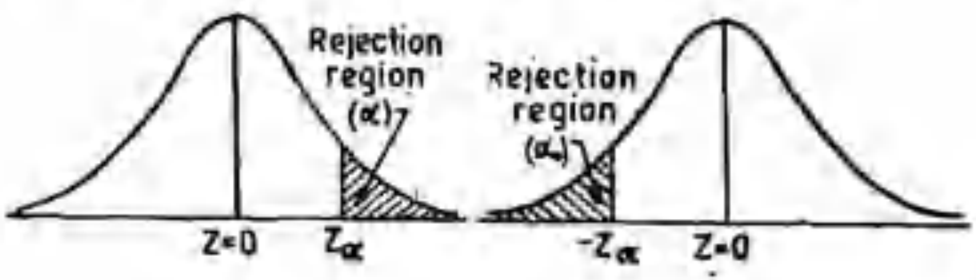
TWO-TAILED TEST
(*Level of Significance 'α'*)



In case of single-tail alternative, the critical value z_α is determined so that total area to the right of it (for right-tailed test) is α and for left-tailed test the total area to the left of $-z_\alpha$ is α (See diagrams below), *i.e.*, For Right-tail Test : $P(Z > z_\alpha) = \alpha$

For Left-tail Test : $P(Z < -z_\alpha) = \alpha$

RIGHT-TAILED TEST	LEFT-TAILED TEST
(Level of Signifiance 'α')	*(Level of Significance 'α')*



CRITICAL VALUES (z_α) OF Z

Critical Values (z_α)	*Level of significance* (α) 1%	5%	10%
Two-tailed test	$\lvert Z_\alpha \rvert = 2.58$	$\lvert Z_\alpha \rvert = 1.96$	$\lvert Z_\alpha \rvert = 1.645$
Right-tailed test	$Z_\alpha = 2.33$	$Z_\alpha = 1.645$	$Z_\alpha = 1.28$
Left-tailed test	$Z_\alpha = -2.33$	$Z_\alpha = -1.645$	$Z_\alpha = -1.28$

Procedure for Testing of Hypothesis. We now summarise below the various steps in testing of a statistical hypothesis in a systematic manner.

1. *Null Hypothesis.* Set up the Null Hypothesis H_0

2. *Alternative Hypothesis.* Set up the Alternative Hypothesis H_1. This will enable us to decide whether we have to use a single-tailed (right or left) test or two-tailed test.

3. *Level of Significance.* Choose the appropriate level of significance (α) depending on the reliability of the estimates and permissible risk. This is to be decided before sample is drawn, *i.e.*, α is fixed in advance.

4. *Test Statistic (or Test Criterion).* Compute the test statistic

$$Z = \frac{t - E(t)}{S.E.(t)}$$

under the null hypothesis.

5. Conclusion. We compare z the computed value of Z in step 4 with the significant value (tabulated value) z_{α}, at the given level of significance 'α'.

If $|Z| < z_{\alpha}$, *i.e.*, if the calculated value of Z (in modulus value) is less than z_{α} we say it is not significant. By this we mean that the difference $t - E(t)$ is just due to fluctuations of sampling and the sample data do not provide us sufficient evidence against the null hypothesis which may therefore, be accepted.

If $|Z| > z_{\alpha}$, *i.e.*, if the computed value of test statistic is greater than the critical or significant value, then we say that it is significant and the null hypothesis is rejected at level of significance α *i.e.*, with confidence coefficient $(1 - \alpha)$.

Test of Significance for Large Samples.

We have seen that for large values of n, the number of trials, almost all the distributions, *e.g.*, binomial, Poisson, negative binomial, etc., are very closely approximated by normal distribution. Thus in this case we apply the *normal test*, which is based upon the following fundamental property (*area property*) of the normal probability curve.

$$\text{If } X \sim N(\mu, \sigma^2), \text{ then } Z = \frac{X-\mu}{\sigma} = \frac{X - E(X)}{\sqrt{V(X)}} \sim N(0, 1)$$

Thus from the normal probability tables, we have

$$P(-3 \le Z \le 3) = 0.9973, \text{ i.e., } P(|Z| \le 3) = 0.9973$$

$$\Rightarrow \quad P(|Z| > 3) = 1 - P(|Z| \le 3) = 0.0027$$

i.e., in all probability we should expect a standard normal variate to lie between ± 3.

Also from the normal probability tables, we get

$$P(-1.96 \le Z \le 1.96) = 0.95 \text{ i.e., } P(|Z| \le 1.96) = 0.95$$

$$\Rightarrow \quad P(|Z| > 1.96) = 1 - 0.95 = 0.05$$

and $\quad P(|Z| \le 2.58) = 0.99$

$$\Rightarrow \quad P(|Z| > 2.58) = 0.01$$

Thus the significant values of Z at 5% and 1% level of significance for a two tailed test are 1.96 and 2.58 respectively.

Test for Single Proportion. If X is the number of successes in n independent trials with constant probability P of success for each trial

$$E(X) = nP \quad \text{and} \quad V(X) = nPQ,$$

where $Q = 1 - P$, is the probability of failure.

It has been proved that for large n, the binomial distribution tends to normal distribution. Hence for large n, $X \sim N(nP, nPQ)$ *i.e.*,

$$Z = \frac{X - E(X)}{\sqrt{V(X)}} = \frac{X - nP}{\sqrt{nPQ}} \sim N(0, 1)$$

and we can apply the normal test.

Test of Significance for Difference of Proportions. Suppose we want to compare two distinct populations with respect to the prevalence of a certain attribute, say A, among their members. Let X_1, X_2 be the number of persons possessing the given attribute A in random samples of sizes n_1 and n_2 from the two populations respectively. Then sample proportions are given by under $H_0 : P_1 = P_2$, the test statistic for the difference of proportions becomes

$$Z = \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0, 1)$$

Test of Significance for Single Mean. We have proved that if x_i, $(i = 1, 2, \ldots, n)$ is a random sample of size n from a normal population with mean μ and variance σ^2, then the sample mean is distributed normally with mean μ and variance σ^2/n, *i.e.*, $\bar{x} \sim N(\mu, \sigma^2/n)$.

Under the *null hypothesis*, H_0 that the sample has been drawn from a population with mean μ and variance σ^2, *i.e.*, there is no significant difference between the sample mean ($\bar{x}$) and population mean (μ), the test statistic (for large samples), is :

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

Confidence limits for μ. 95% confidence interval for μ is given by :

$$|Z| \le 1{\cdot}96, \text{ i.e., } \left|\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}\right| \le 1{\cdot}96$$

$$\Rightarrow \quad \bar{x} - 1{\cdot}96\sigma/\sqrt{n} \le \mu \le \bar{x} + 1{\cdot}96\sigma/\sqrt{n}$$

and $\bar{x} \pm 1{\cdot}96\sigma/\sqrt{n}$ are known as 95% confidence limits for μ. Similarly, 99% confidence limits for μ are $\bar{x} \pm 2{\cdot}58\sigma/\sqrt{n}$ and 98% confidence limits for μ are $\bar{x} \pm 2{\cdot}33\sigma/\sqrt{n}$.

Test of Significance for Difference of Means. Let $\bar{x}_1$ be the mean of a random sample of size n_1 from a population with mean μ_1 and variance σ_1^2 and let $\bar{x}_2$ be the mean of an independent random sample of size n_2 from another population with mean μ_2 and variance σ_2^2. Then, since sample sizes are large,

$$\bar{x}_1 \sim N(\mu_1, \sigma_1^2/n_1) \text{ and } \bar{x}_2 \sim N(\mu_2, \sigma_2^2/n_2)$$

Thus under $H_0 : \mu_1 = \mu_2$, the test statistic becomes (for large samples),

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{(\sigma_1^2/n_1) + (\sigma_2^2/n_2)}} \sim N(0, 1)$$

Student's 't'. *Definition.* Let x_i, $(i = 1, 2, \ldots, n)$ be a random sample of size n from a normal population with mean μ and variance σ^2. Then Student's t is defined by the statistic

$$t = \frac{\bar{x} - \mu}{S/\sqrt{n}}$$

where $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$, is the sample mean and

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

is an unbiased estimate of the population variance σ^2, and it follows Student's t-distribution with $\nu = (n-1)$ *d.f.* with probability density function,

$$f(t) = \frac{1}{\sqrt{\nu}\, B\left(\frac{1}{2}, \frac{\nu}{2}\right)} \cdot \frac{1}{\left[1 + \frac{t^2}{\nu}\right]^{(\nu+1)/2}}; -\infty < t < \infty$$

Applications of t-distribution. The t-distribution has a wide number of applications in Statistics, some of which are enumerated below.

(*i*) To test if the sample mean ($\bar{x}$) differs significantly from the hypothetical value μ of the population mean.

(*ii*) To test the significance of the difference between two sample means.

(*iii*) To test the significance of an observed sample correlation co-efficient and sample regression coefficient.

(*iv*) To test the significance of observed partial and multiple correlation coefficients.

t-Test for Single Mean. Suppose we want to test :

(*i*) if a random sample x_i ($i = 1, 2, \ldots, n$) of size n has been drawn from a normal population with a specified mean, say μ_0, or

(*ii*) if the sample mean differs significantly from the hypothetical value μ_0 of the population mean.

Under the null hypothesis H_0 :

(i) *The sample has been drawn from the population with mean μ or (ii) there is no significant difference between the sample mean $\bar{x}$ and the population mean μ,*

the statistic
$$t = \frac{\bar{x} - \mu_0}{S/\sqrt{n}}$$

where
$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \quad \text{and} \quad S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2,$$

follows Student's t-distribution with $(n-1)$ d.f.

We now compare the calculated value of t with the tabulated value at certain level of significance. If calculated $|t|$ > tabulated t, null hypothesis is rejected and if calculated $|t|$ < tabulated t, H_0 may be accepted at the level of significance adopted.

t-Test for Difference of Means. Suppose we want to test if two independent samples x_i ($i = 1, 2, \ldots, n_1$) and y_j ($j = 1, 2, \ldots, n_2$) of sizes n_1 and n_2 have been drawn from two normal populations with means μ_X and μ_Y respectively.

Under the null hypothesis (H_0) *that the samples have been drawn from the normal populations with means* μ_X *and* μ_Y *and under the assumption that the population variance are equal, i.e.,* $\sigma_X^2 = \sigma_Y^2 = \sigma^2$ (say), the statistic

$$t = \frac{(\bar{x} - \bar{y}) - (\mu_X - \mu_Y)}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where $\bar{x} = \frac{1}{n_1}\sum_{i=1}^{n_1} x_i, \quad \bar{y} = \frac{1}{n_2}\sum_{j=1}^{n_2} y_j$

and $S^2 = \frac{1}{n_1 + n_2 - 2}\left[\sum_i (x_i - \bar{x})^2 + \sum_j (y_j - \bar{y})^2\right]$

is an unbiased estimate of the common population variance σ^2, follows Student's t-distribution with $(n_1 + n_2 - 2)$ d.f.

F-test for Equality of Population Variances. Suppose we want to test (*i*) whether two independent samples x_i, $(i = 1, 2, \ldots, n_1)$ and y_j, $(j = 1, 2, \ldots, n_2)$ have been drawn from the normal populations with the same variance σ^2, (say), or (*ii*) whether the two independent estimates of the population variance are homogeneous or not.

Under the null hypothesis (H_0) that (*i*) $\sigma_X^2 = \sigma_Y^2 = \sigma^2$, *i.e., the population variances are equal or (ii) Two independent estimates of the population variance are homogeneous*, the statistic F is given by

$$F = \frac{S_X^2}{S_Y^2}$$

where $S_X^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (x_i - \bar{x})^2$ and $S_Y^2 = \frac{1}{n_2 - 1} \sum_{j=1}^{n_2} (y_j - \bar{y})^2$

are unbiased estimates of the common population variance σ^2 obtained from two independent samples and it follows *Snedecor's F-distribution* with (ν_1, ν_2) d.f. (where $\nu_1 = n_1 - 1$ and $\nu_2 = n_2 - 1$).

Chi-square Distribution

Chi-Square Variate *(Pronounced as Ki – Sky without S)*. The square of a standard normal variate is known as a chi-square variate with 1 degree of freedom (d.f.)

Thus if $X \sim N(\mu, \sigma^2)$, then $Z = \dfrac{X - \mu}{\sigma} \sim N(0, 1)$

and $$Z^2 = \left(\frac{X - \mu}{\sigma}\right)^2,$$ is a chi-square variate with 1 d.f.

In general, if X_i, $(i = 1, 2, \ldots, n)$ are n independent normal variates with mean μ_i and variance σ_i^2, $(i = 1, 2, \ldots, n)$, then

$$\chi^2 = \sum_{i=1}^{n} \left(\frac{X_i - \mu_i}{\sigma_i}\right)^2,$$ is a chi-square variate with n d.f.

If O_i and E_i ($i = 1, 2, \ldots, k$), be a set of observed and expected frequencies, then

$$\chi^2 = \sum_{i=1}^{k} \left[\frac{(O_i - E_i)^2}{E_i}\right], \quad \left(\sum_{i=1}^{k} O_i = \sum_{i=1}^{k} E_i\right)$$

follows chi-square distribution with $(k-1)$ d.f

Chi-square Test of Goodness of Fit. A very powerful test for testing the significance of the discrepancy between theory and experiment was given by Prof. Karl Pearson in 1900 and is known as *"Chi-square test of goodness of fit"*. It enables us to find if the deviation of the experiment from theory is just by chance or is it really due to the inadequacy of the theory to fit the observed data.

If O_i, $(i = 1, 2, \ldots, n)$ is a set of observed (experimental) frequencies and E_i $(i = 1, 2, \ldots, n)$ is the corresponding set of expected (theoretical or hypothetical) frequencies, then Karl Pearson's chi-square, given by

$$\chi^2 = \sum_{i=1}^{n} \left[\frac{(O_i - E_i)^2}{E_i} \right], \qquad \left(\sum_{i=1}^{n} O_i = \sum_{i=1}^{n} E_i \right)$$

follows chi-square distribution with $(n - 1)$ d.f.

Independence of Attributes. Let us consider two attributes A and B. A divided into r classes $A_1, A_2, \ldots, A_r$ and B divided into s classes $B_1, B_2, \ldots, B_s$. Such a classification in which attributes are divided into more than two classes is known as *manifold classification*. The various cell frequencies can be expressed in the following table known as $r \times s$ *manifold contingency table* where (A_i) is the number of persons possessing the attribute A_i, $(i = 1, 2, \ldots, r)$, (B_j) is the number of persons possessing the attribute B_j $(j = 1, 2, \ldots, s)$ and (A_iB_j) is the number of persons possessing both the attributes A_i and B_j, $(i = 1, 2, \ldots, r;\ j = 1, 2, \ldots, s)$. Also

$$\sum_{i=1}^{r} (A_i) = \sum_{j=1}^{s} (B_j) = N, \text{ is the total frequency.}$$

The exact test for the independence of attributes is very complicated but a fair degree of approximation is given, for large samples, (large N), by the χ^2-test of goodness of fit, *viz.*,

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \left[\frac{\{(A_iB_j) - (A_iB_j)_0\}^2}{(A_iB_j)_0} \right],$$

which is distributed as a χ^2-variate with $(r-1)(s-1)$ d.f.

SIGNIFICANT VALUES $\chi^2(\alpha)$ OF CHI-SQUARE DISTRIBUTION (RIGHT TAIL AREAS FOR GIVEN PROBABILITY α, where

$$P = P_\nu(\chi^2 > \chi^2(\alpha)) = \alpha$$

AND ν IS DEGREES OF FREEDOM (d.f.)

Degree of freedom (ν)	Probability (level of significance) P = .99	0.95	0.50	0.10	0.05	0.02	0.01
1	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
2	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
3	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
4	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
5	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
6	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
7	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
8	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
9	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
10	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
11	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
12	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
13	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
14	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
15	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
16	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
17	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
18	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
19	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
20	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
21	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
22	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
23	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
24	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
25	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
26	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
27	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
28	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
29	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
30	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Note. For degrees of freedom (ν) greater than 30, the quantity $\sqrt{2\chi^2} - \sqrt{2\nu - 1}$ may be used as a normal variate with unit variance.

SIGNIFICANT VALUES $t_v(\alpha)$ OF t-DISTRIBUTION
(TWO TAIL AREAS)

$$P\left[\,|t| \geq t_v(\alpha)\right] = \alpha$$

d.f. (v)	Probability (Level of Significance)					
	0.50	0.10	0.05	0.02	0.01	0.001
1	1.00	6.31	12.71	31.82	63.66	636.62
2	0.82	2.92	4.30	6.97	9.92	31.60
3	0.77	2.35	3.18	4.54	5.84	12.94
4	0.74	2.13	2.78	3.75	4.60	8.61
5	0.73	2.02	2.57	3.37	4.03	6.86
6	0.72	1.94	2.45	3.14	3.71	5.96
7	0.71	1.90	2.37	3.00	3.50	5.41
8	0.71	1.86	2.31	2.90	3.36	5.04
9	0.70	1.83	2.26	2.82	3.25	4.78
10	0.70	1.81	2.23	2.76	3.17	4.59
11	0.70	1.80	2.20	2.72	3.11	4.44
12	0.70	1.78	2.18	2.68	3.06	4.32
13	0.69	1.77	2.16	2.65	3.01	4.22
14	0.69	1.76	2.14	2.62	2.98	4.14
15	0.69	1.75	2.13	2.60	2.95	4.07
16	0.69	1.75	2.12	2.58	2.92	4.02
17	0.69	1.74	2.11	2.57	2.90	3.97
18	0.69	1.73	2.10	2.55	2.88	3.92
19	0.69	1.73	2.09	2.54	2.86	3.88
20	0.69	1.72	2.09	2.53	2.85	3.85
21	0.69	1.72	2.08	2.52	2.83	3.82
22	0.69	1.72	2.07	2.51	2.82	3.79
23	0.69	1.71	2.07	2.50	2.81	3.77
24	0.68	1.71	2.06	2.49	2.80	3.75
25	0.68	1.71	2.06	2.49	2.79	3.73
26	0.68	1.71	2.06	2.48	2.78	3.71
27	0.68	1.70	2.05	2.47	2.77	3.69
28	0.68	1.70	2.05	2.47	2.76	3.67
29	0.68	1.70	2.05	2.46	2.76	3.66
30	0.68	1.70	2.04	2.46	2.75	3.65
∞	0.67	1.64	1.96	2.33	2.58	3.29