Chapter 5

Probability and Statistics

CHAPTER HIGHLIGHTS

- *Probability*
- *Some special continuous distributions*
- *Statistics*
- *Hypothesis testing*

Probability

The word PROBABILITY is used, in a general sense, to indicate a vague possibility that something might happen. It is also used synonymously with chance.

Random Experiment

If the result of an experiment conducted any number of times under essentially identical conditions, is not certain but is any one of the several possible outcomes, the experiment is called a trial or a random experiment. Each of the out comes is known as an event.

Examples:

1. Drawing 3 cards from a well shuffled pack is a random experiment while getting an Ace or a King are events.
2. Throwing a fair die is a random experiment while getting the score as '2' or an odd number' are events.

Mutually Exclusive Events If the happening of any one of the events in a trial excludes or prevents the happening of all others, then such events are said to be mutually exclusive.

Example: The events of getting a head and that of getting a tail when a fair coin is tossed are mutually exclusive.

Equally Likely Events Two events are said to be equally likely when chance of occurrence of one event is equal to that of the other.

Example: When a die is thrown, any number from 1 to 6 may be got. In this trial, getting any one of these events are equally likely.

Independent Events Two events E_1 and E_2 are said to be independent, if the occurrence of the event E_2 is not affected by the occurrence or non-occurrence of the event E_1.

Example: Two drawings of one ball each time are made from a bag containing balls.

Here, we have two events drawing a ball first time (E_1) and drawing a ball second time (E_2). If the ball of the first draw is replaced in the bag before the second draw is made, then the outcome of E_2 does not depend on the outcome of E_1. In this case E_1 and E_2 are Independent events.

If the ball of the first draw is not replaced in the bag before the second draw is made, then the outcome of E_2 depends on the outcome of E_1. In this case, events E_1 and E_2 are Dependent events.

Compound Events When two or more events are in relation with each other, they are known as compound events.

Example: When a die is thrown two times, the event of getting 3 in the first throw and 5 in the second throw is a compound event.

Definition of Probability

If an event E can happen in m ways and fail in k ways out of a total of n ways and each of them is equally likely, then the probability of happening E is $\frac{m}{(m+k)} = \frac{m}{n}$ where $n = (m+k)$.

In other words, if a random experiment is conducted n times and m of them are favourable to event E, then the

probability of happening of E is $P(E) = \frac{m}{n}$. Since the event does not occur $(n - m)$ times, the probability of non-occurrence of E is $P(\bar{E})$.

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

Therefore, $P(E) + P(\bar{E}) = 1$.

NOTES

1. Probability $[P(E)]$ of the happening of an event E is known as the probability of success and the probability $[P(\bar{E})]$ of the non-happening of the event is the probability of failure.
2. If $P(E) = 1$, the event is called a certain event and if $P(E) = 0$ the event is called an impossible event.
3. Instead of saying that the chance of happening of an event is $\frac{m}{n}$, we can also say that the odds in favour of the event are m to $(n - m)$ or the odds against the event are $(n - m)$ to m.

Addition Theorem of Probability

If A and B are two events, then

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

This result follows from the corresponding result in set theory. If $n(X)$ represents the number of elements in set X, $n(X \cup Y) = n(X) + n(Y) - n(X \cap Y)$.

Example: If a die is rolled, what is the probability that the number that comes up is either even or prime?

A = The event of getting an even number = $\{2, 4, 6\}$

B = The event of getting a prime = $\{2, 3, 5\}$

$A \cup B = \{2, 3, 4, 5, 6\}$

$A \cap B = \{2\}$

$P(A) = \frac{3}{6}$, $P(B) = \frac{3}{6}$, $P(A \cup B) = \frac{5}{6}$ and $P(A \cap B) = \frac{1}{6}$. We can verify that $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

SOLVED EXAMPLES

Example 1

When a cubical dice is rolled, find the probability of getting an even integer.

Solution

When a dice is rolled, the number of possible out comes is 6. The number of favourable outcomes of getting an even integer is 3.

The required probability $= \frac{3}{6} = \frac{1}{2}$.

Example 2

If a card is drawn from a pack of cards, find the probability of getting a queen.

Solution

When a card is drawn, the number of possible outcomes is 52. The number of favourable outcomes of getting a queen card is 4.

The required probability $= \frac{4}{52} = \frac{1}{13}$.

Example 3

A bag contains 5 green balls and 4 red balls. If 3 balls are picked from it at random, then find the odds against the three balls being red.

Solution

The total number of balls in the bag = 9. Three balls can be selected from 9 balls in ${}^9C_3$ ways.

Three red balls can be selected from 4 red balls in ${}^4C_3$ ways.

Probability of picking three red balls

$$= \frac{{}^4C_3}{{}^9C_3} = \frac{4}{84} = \frac{1}{21};\ P(\bar{E}) = \frac{20}{21}$$

Odds against the three balls being red are $= P(\bar{E}) : P(E) = \frac{20}{21} : \frac{1}{21} = 20 : 1$.

Example 4

When two dice are rolled together, find the probability of getting at least one 4.

Solution

Let E be the event that at least one dice shows 4. $\bar{E}$ be the event that no dice shows 4. The number of favourable outcomes of $\bar{E}$ is $5 \times 5 = 25 \cdot P(\bar{E}) = \frac{25}{36}$

$$\therefore\ P(E) = 1 - P(\bar{E}) = 1 - \frac{25}{36} = \frac{11}{36}.$$

Example 5

When two dice are rolled together find the probability that total score on the two dice will be 8 or 9.

Solution

When two dice are rolled, the total number of outcomes $= 6 \times 6 = 36$.

Favourable outcomes for getting the sum 8 or 9 are {(2, 6), (6, 2), (3, 5), (5, 3), (4, 4), (3, 6), (6, 3), (4, 5), (5, 4)}, i.e., the total number of favourable outcomes = 9.

The required probability $= \frac{9}{36} = \frac{1}{4}$.

Example 6

If two cards are drawn simultaneously from a pack of cards, what is the probability that both will be jacks or both are queens?

Solution

Here two events are mutually exclusive, $P(J \cup Q) = P(J) + P(Q)$. Probability of drawing two jacks is $P(J) = \frac{^4C_2}{^{52}C_2}$

Probability of drawing two queens is $P(Q) = \frac{^4C_2}{^{52}C_2}$

$$P(J \cup Q) = P(J) + P(Q)$$

$$= \frac{^4C_2}{^{52}C_2} + \frac{^4C_2}{^{52}C_2} = 2 \cdot \frac{^4C_2}{^{52}C_2} = \frac{2}{221}.$$

Example 7

When two cards are drawn from a pack of cards, find the probability that the two cards will be kings or blacks.

Solution

The probability of drawing two kings $= \frac{^4C_2}{^{52}C_2}$

The probability of drawing two black cards is $= \frac{^{26}C_2}{^{52}C_2}$

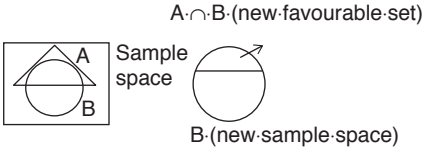
The probability of drawing two black kings is $\frac{^2C_2}{^{52}C_2}$

$\therefore$ The required probability

$$= \frac{^4C_2}{^{52}C_2} + \frac{^{26}C_2}{^{52}C_2} - \frac{^2C_2}{^{52}C_2} = \frac{55}{221}.$$

Conditional Probability

Let S be a finite sample space of a random experiment and A, B are events, such that $P(A) > 0$, $P(B) > 0$. If it is known that the event B has occurred, in light of this we wish to compute the probability of A, we mean conditional probability of A given B. The occurrence of event B would reduce the sample space to B, and the favourable cases would now be $A \cap B$.



Notation The conditional probability of A given B is denoted by $P\left(\frac{A}{B}\right)$.

$$\therefore \quad P\left(\frac{A}{B}\right) = \frac{n(A \cap B)}{n(B)} = \frac{\frac{n(A \cap B)}{n(S)}}{\frac{n(B)}{n(S)}} = \frac{P(A \cap B)}{P(B)}.$$

NOTES

1. This definition is also valid for infinite sample spaces.
2. The conditional probability of B given A is denoted by $P\left(\frac{B}{A}\right)$ and $P\left(\frac{B}{A}\right) = \frac{P(A \cap B)}{P(A)}$.

Multiplication Theorem

Let A and B be two events of certain random experiment such that A occurs only when B has already occurred. Then, for the conditional event $\frac{A}{B}$, the total possible outcomes are the outcomes favourable to the event B and its favourable outcomes are the outcomes favourable to both A and B.

So, $P\left(\frac{A}{B}\right) = \frac{n(A \cap B)}{n(B)}$

$$= \frac{n(A \cap B)}{n(S)} \times \frac{n(S)}{n(B)} = P(A \cap B) \times \frac{1}{P(B)}$$

That is, $P\left(\frac{A}{B}\right) \cdot P(B) = P(A \cap B)$

This is called the multiplication theorem on probability.

Example 8

A letter is selected at random from the set of English alphabet and it is found to be a vowel. What is the probability that it is '*e*'?

Solution

Let A be the event that the letter selected is '*e*' and B be the event that the letter is a vowel. Then, $A \cap B = \{e\}$ and $B = \{a, e, i\ o, u\}$

So, $P\left(\frac{A}{B}\right) = \frac{P(A \cap B)}{P(B)} = \frac{\left(\frac{1}{26}\right)}{\left(\frac{5}{26}\right)} = \frac{1}{5}$.

Independent Events In a random experiment, if A, B are events such that $P(A) > 0$, $P(B) > 0$ and if $P\left(\frac{A}{B}\right) = P(A)$ or $P\left(\frac{B}{A}\right) = P(B)$ (conditional probability equals to unconditional probability) then we say A, B are independent events. If A, B are independent, $P(A \cap B) = P(A)\ P(B)$.

Example 9

Two coins are tossed one after the other and let A be the event of getting tail on second coin and B be the event of getting head on first coin, then find $P\left(\frac{A}{B}\right)$.

Solution

Sample space = {HH, HT, TH, TT}, A = {HT, TT} and B = {HH, HT}, $(A \cap B)$ = {HT}

$$\therefore\ P(A) = \frac{2}{4} = \frac{1}{2} \quad \text{and} \quad P\left(\frac{A}{B}\right) = \frac{P(A \cap B)}{P(B)} = \frac{\frac{1}{4}}{\frac{1}{2}} = \frac{1}{2}$$

Thus $P\left(\frac{A}{B}\right) = P(A)$

$\therefore$ Logically too we understand that occurrence or non-occurrence of tail in 2nd coin.

Baye's Rule

Suppose $A_1, A_2, \ldots, A_n$ are n mutually exclusive and exhaustive events such that $P(A_i) \neq 0$. Then for $i = 1, 2, 3, \ldots, n$,

$$P\left(\frac{A_i}{A}\right) = \frac{P(A_i) \cdot P\left(\frac{A}{A_i}\right)}{\sum_{k=1}^{n} P(A_K) P\left(\frac{A}{A_K}\right)}$$

Where A is an arbitrary event of S.

Example 10

Akshay speaks the truth in 45% of the cases, In a rainy season, on each day there is a 75% chance of raining. On a certain day in the rainy season, Akshay tells his mother that it is raining outside. What is the probability that it is actually raining?

Solution

Let E denote the event that it is raining and A denote the event that Akshay tells his mother that it is raining outside.

Then, $P(E) = \frac{3}{4}$, $P(\bar{E}) = \frac{1}{4}$

$$P\left(\frac{A}{E}\right) = \frac{45}{100} = \frac{9}{20} \text{ and } P\left(\frac{A}{\bar{E}}\right) = \frac{11}{20}$$

By Baye's Rule, we have

$$P\left(\frac{E}{A}\right) = \frac{P(E)P\left(\frac{A}{E}\right)}{P(E)P\left(\frac{A}{E}\right) + P(\bar{E})P\left(\frac{A}{\bar{E}}\right)}$$

$$= \frac{\frac{3}{4} \times \frac{9}{20}}{\frac{3}{4} \times \frac{9}{20} + \frac{1}{4} \times \frac{11}{20}} = \frac{27}{38}.$$

Advanced Probability

Random Variable

A random variable is a real valued function defined over the sample space (discrete or continuous).

A **discrete random variable** takes the values that are finite or countable. For example when we consider the experiment of tossing of 3 coins, the number of heads can be appreciated as a discrete random variable (X). X would take 0, 1, 2 and 3 as possible values.

A continuous random variable takes values in the form of intervals. Also, in the case of a **continuous random variable** $P(X = c) = 0$, where c is a specified point. Heights and weights of people, area of land held by individuals, etc., are examples of continuous random variables.

Probability Mass Function (PMF)

If X is a discrete random variable, which can take the values $x_1, x_2, \ldots$ and $f(x)$ denote the probability that X takes the value x_i, then $p(x)$ is called the Probability Man Function (pmf) of X.

$p(x_i) = P(x = x_i)$. The values that X can take and the corresponding probabilities determine the probability distribution of X. We also have

1. $p(x) \geq 0$;
2. $\sum p(x) = 1$.

Probability Density Function (PDF)

If X is a continuous random variable then a function $f(x)$, $x \in I$ (interval) is called a Probability Density Function. The probability statements are made as $P(x \in I) = \int_I f(x)dx$.

We also have,

1. $f(x) \geq 0$
2. $\int_{-\infty}^{\infty} f(x)dx = 1$

The probability $P(X \leq x)$ is called the cumulative distribution function (CDF) of X and is denoted by $F(X)$. It is a point function. It is defined for discrete and continuous random variables.

The following are the properties of probability distribution function $F(x)$,

1. $F(x) \geq 0$
2. $F(x)$ is non-decreasing i.e., for $x > y$, $F(x) \geq F(y)$
3. $F(x)$ is right continuous
4. $F(-\infty) = 0$ and $F(+\infty) = 1$

Also,

5. $P(a < x \leq b) = F(b) - F(a)$.

For a continuous random variable:

6. $Pr\{x < X \leq x + dx\} = F(x + dx) - F(x) = f(x)\,dx$; where dx is very small
7. $f(x) = \frac{d}{dx}[F(x)]$ where;
   (a) $f(x) \geq 0\ \forall\ x \in R$.
   (b) $\int_R f(x)dx = 1$.

Mathematical Expectation [E(X)]

Mathematical Expectation is the weighted mean of values of a variable.

If X is a random variable which can assume any one of the values $x_1, x_2, \ldots, x_n$ with the respective probabilities $p_1, p_2, \ldots, p_n$, then the mathematical expectation of X is given by $E(X) = p_1x_1 + p_2x_2 + \cdots + p_n x_n$

For a continuous random variable,

$$E(X) = \int_{-\infty}^{+\infty} xf(x)dx \text{ where } f(x) \text{ is the PDF of } X.$$

Some Special Discrete Distributions

Discrete Uniform Distribution

A discrete random variable defined for values of x from 1 to n is said to have a uniform distribution if its probability mass function is given by

$$f(x) = \begin{cases} \frac{1}{n}; & \text{for } x = 1,\ 2,\ 3, \ldots,\ n \\ 0, & \text{otherwise} \end{cases}$$

- The cumulative distribution function $F(x)$ of the discrete uniform random variable x is given by

$$F(x) = \begin{cases} 0, & \text{for } x < 1 \\ \frac{x}{n}; & \text{for } 1 \le x \le n \\ 1; & \text{for } x > 1 \end{cases}$$

- Mean of $X = \mu = \frac{n+1}{2}$
- Variance of $X = \sigma^2 = \frac{n^2 - 1}{12}$

Binomial Distribution

An experiment which is made of n independent trials, each of which resulting in either 'success' with probability 'p' or 'failure' with probability 'q' ($q = 1 - p$), then the probability distribution for the random variable X when represents the number of success is called a binomial distribution. The probability mass function,

$$p(x) = b(x; n, p) = {}^nC_x\, p^x\, q^{n-x};\ x = 0, 1, 2, \ldots, n$$

Example: Hitting a target in 5 trials. Here the random variable (X) represents the number of trials made for hitting the target, i.e., $x = 0$ or 1 or 2 or 3 or 4 or 5.

We have a set of 5 trials $n = 5$

Each trial may hit the target termed to be success (p) or not termed to be failure (q), which are independent.

$\therefore$ This is an example for Binomial distribution.

Properties of Binomial Distribution

1. $E(X) = np$ (mean)
2. $V(X) = E(X^2) - (E(X))^2 = npq$; (variance)
   (mean > variance)
3. SD $(X) = \sqrt{npq}$
4. Mode of a binomial distribution lies between $(n + 1)$ $p - 1 \le x \le (n + 1)p$
5. If $X_1 \sim b(n_1, p)$ and $X_2 \sim b\ (n_2, p)$ and if X_1 and X_2 are independent, then . $X_1 + X_2 \sim b\ (n_1 + n_2, p)$ where (n, p) is the pmf of binomial distribution.

Poisson Distribution A random variable X is said to follow a **Poisson distribution** with parameter λ, $\lambda > 0$, if it assumes only non-negative values and its probability mass function is given by

$$p(x) = p(x;\ \lambda) = \begin{cases} \frac{e^{-\lambda}\lambda^x}{x!} & : x = 0,\ 1,\ 2, \ldots \\ 0 & \lambda > 0 \\ & \text{otherwise} \end{cases}$$

In a binomial distribution if n is large compared to p, then np approaches a fixed constant say λ. Such a distribution is called poisson distribution (limiting case of binomial distribution)

Properties of Poisson Distribution

1. $E(X) = \sum_x x \cdot \frac{e^{-\lambda}\lambda^x}{x!} = \lambda$
2. $V(X) = E(X^2) - (E(X))^2 = \lambda$

   SD $(X) = \sqrt{\lambda}$

   $\therefore$ Mean = λ = Variance
3. Mode of a Poisson distribution lies between $\lambda - 1$ and λ
4. If $X_1 \sim P(\lambda_1)$ and $X_2 \sim P(\lambda_2)$, and X_1, X_2 independent then $X_1 + X_2 \sim P(\lambda_1 + \lambda_2)$.

Some Special Continuous Distributions

Continuous Uniform Distribution or Rectangular Distribution

A continuous random variable x defined on $[a, b]$ is said to have a uniform distribution, if its probability density function is given by

$$F(x) = \begin{cases} \frac{1}{b-a}; & \text{for } x \in [a,\ b] \\ 0; & \text{otherwise} \end{cases}$$

- The cumulative distribution function of the continuous uniform random variable X is given by

$$F(x) = \begin{cases} 0; & \text{if } x \le a \\ \dfrac{x-a}{b-a}; & \text{if } a < x < b \\ 1; & \text{if } x \ge b \end{cases}$$

- Mean of $X = \mu = \dfrac{a+b}{2}$
- Variance of $X = \sigma^2 = \dfrac{(b-a)^2}{12}$.

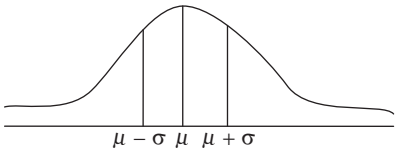
Normal Distribution

A continuous random variable X is said to have **a normal distribution** with parameters μ and σ^2 if its density function is given by the probability density function,

$$f(x) = \begin{cases} \dfrac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} & -\infty < x < \infty \\ & -\infty < \mu < \infty \\ & \sigma > 0 \\ 0 & \text{otherwise} \end{cases}$$

It is denoted as $X \sim N(\mu, \sigma^2)$.

The graphical representation of normal distribution is as given below.



Properties of Normal Distribution

1. The function is symmetrical about the value μ.
2. It has a maximum at $x = \mu$
3. The area under the curve within the interval $(\mu \pm \sigma)$ is 68%.

   That is, $P(\mu - \sigma \le X \le \mu + \sigma) = 0.68$.
4. A fairly large number of samples taken from a 'Normal' population will have average, median and mode nearly the same, and within the limits of average $\pm 2 \times$ SD, there will be 95% of the values.
5. $E(X) = \int_{-\infty}^{+\infty} x \cdot f(x)dx = \mu$.
6. $V(X) = \sigma^2$; S.D $(X) = \sigma$
7. For a normal distribution,

   Mean = Median = Mode
8. All odd order moments about mean vanish for a normal distribution.

   That is, $\mu_{2n+1} = 0 \forall = n = 0, 1, 2, \ldots$
9. If $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$, X_1, X_2 independent, then,

   $X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$

   Also, $X_1 - X_2 \sim N(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2)$
10. If $\mu = 0$ and $\sigma^2 = 1$, we call it as standard normal distribution. The standardization can be obtained by the transformation,

$$z = \frac{x-\mu}{\sigma}. \quad \text{Also, } \frac{X-\mu}{\sigma} \sim N(0, 1).$$

Exponential Distribution

A continuous random variable X is said to have an exponential distribution if its probability density function $f(x)$ is given by,

$$f(x) = \begin{cases} \lambda e^{-\lambda x}; & \text{for } x > 0 \\ 0; & \text{otherwise} \end{cases}$$

Here λ is the parameter of the exponential distribution and $\lambda > 0$.

The cumulative distribution function $F(x)$ of an exponential distribution with λ as parameter is

$$F(x) = \begin{cases} 1 - e^{-\lambda x}; & \text{if } x > 0 \\ 0, & \text{otherwise} \end{cases}$$

Mean $= \mu = \dfrac{1}{\lambda}$,

Variance $= \sigma^2 = \dfrac{1}{\lambda^2}$.

Example 11

An unbiased die is thrown at random. What is the expectation of the number on it?

Solution

Let X denotes the number on the die, which can take the values 1, 2, 3, 4, 5 or 6;

Probability of each will be equal to $\frac{1}{6}$

X	1	2	3	4	5	6
$P(X = x)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

$$E(X) = \sum_x xP(X = x)$$

$$= 1 \times \frac{1}{6} + 2 \times \frac{1}{6} + 3 \times \frac{1}{6} + 4 \times \frac{1}{6} + 5 \times \frac{1}{6} + 6 \times \frac{1}{6}$$

$$= \frac{1}{6}(1 + 2 + 3 + 4 + 5 + 6) = \frac{6 \times 7}{6 \times 2} = \frac{7}{2}$$

$$= 3.5.$$

Example 12

In a city 5 accidents take place in a span of 25 days. Assuming that the number of accidents follows the Poisson distribution, what is the probability that there will be 3 or more accidents in a day? (Given $e^{-0.2} = 0.8187$)

Solution

Average number of accidents per day $= \frac{5}{25} = 0.2$; $\therefore \lambda = 0.2$.

Probability (3 or more accidents per day)

$= 1 - P$ (2 or less accidents)

$= 1 - [P(X = 0) + P(X = 1) + P(X = 2)]$

$= 1 - [e^{-0.2} + 0.2e^{-0.2} + 0.02e^{-0.2}]$

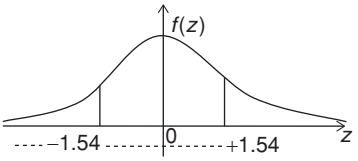
$= 1 - e^{-0.2}[1.22] = 1 - 0.99814 = 0.001186.$

Example 13

What is the area under the normal curve to the left of $Z = -1.54$ (given area between 0 and $-1.54 = 0.4382$)?

Solution

Required area $= 0.5 - 0.4382 = 0.0618$



Example 14

A family consists of five children. If the random variable (X) represents the number of boys in that family then,

1. Find the expected value $E(X)$ of X.
2. Find the variance of X.

Solution

This situation can be modelled as binomial distribution.

$$X \sim b\left(5, \frac{1}{2}\right);\ E(X) = np = 5 \times \frac{1}{2} = 2.5$$

$$V(X) = npq = 5 \times \frac{1}{2} \times \frac{1}{2} = 1.25$$

Example 15

Ram and Shyam play a game in which their chances of winning are in the ratio 2 : 3. Find Shyam's chance of winning at least 3 games out of five games played.

Solution

$$P(\text{Shyam wins}) = \frac{3}{5};$$

$$P(\text{Shyam loses}) = \frac{2}{5}$$

Let X denote the number of games won by Shyam.

P(Shyam wins at least 3 games) $= P(X \geq 3)$

$$= \sum_{x=3}^{5} {}^5C_x \left(\frac{3}{5}\right)^x \left(\frac{2}{5}\right)^{5-x} = \sum_{x=3}^{5} {}^5C_x \frac{3^x 2^{5-x}}{5^5}$$

$$= \frac{3^3}{5^5}[{}^5C_3 2^2 + {}^5C_4 \times 3 \times 2 + 1 \times 3^2 \times 1]$$

$$= \frac{27 \times 79}{3125} = 0.68.$$

Example 16

The PDF of a random variable X is

$$f(x) = \begin{cases} \left(\frac{1}{10}\right) e^{\left(-\frac{x}{10}\right)};\ x > 0 \\ 0 \quad \text{otherwise} \end{cases}$$

What is $P(X \leq 10)$? (given $e^{-1} = 0.3679$)

Solution

$$P(X \leq 10) = \int_0^{10} f(x)\,dx = \int_0^{10} \frac{1}{10} e^{-\frac{x}{10}} dx$$

$$= \frac{1}{10}\left(\frac{e^{-\frac{x}{10}}}{-\frac{1}{10}}\right)_0^{10} = 1 - e^{-1} = 0.6321.$$

Joint Distribution of Random Variables Joint Probability Mass Function

Let X and Y be two discrete random variables on the same sample space S with the range space of X as $R_x = \{x_1, x_2, \ldots, x_m\}$ and the range space of y as, $R_y = \{y_1, y_2, \ldots, y_n\}$ and $P_x(x)$ and $P_y(y)$ as the probability mass functions of x and y. Then the joint probability mass function $P_{xy}(x, y)$ of the two dimensional random variable (x, y) on the range space $R_x \times R_y$ is defined as,

$$P_{XY}(x_i, y_j) = \begin{cases} P(X = x_i,\ Y = y_j), \text{ for } (x_i, y_j) \in R_X \times R_Y \\ 0, \text{ otherwise} \end{cases}$$

This joint probability mass function can be represented in the form of a table as follows:

X \ Y	y_1	y_2	y_3	...	y_n	$\sum_{j=1}^{n} P_{xy}(x_i, y_j)$
x_1	$P_{xy}(x_1, y_1)$	$P_{xy}(x_1, y_2)$	$P_{xy}(x_1, y_3)$	...	$P_{xy}(x_1, y_n)$	$P_x(x_1)$
x_2	$P_{xy}(x_2, y_1)$	$P_{xy}(x_2, y_2)$	$P_{xy}(x_3, y_3)$	...	$P_{xy}(x_3, y_n)$	$P_x(x_2)$
x_3	$P_{xy}(x_3, y_1)$	$P_{xy}(x_3, y_2)$	$P_{xy}(x_3, y_3)$	...	$P_{xy}(x_3, y_n)$	$P_x(x_3)$
⋮	⋮	⋮	⋮		⋮	⋮
x_m	$P_{xy}(x_m, y_1)$	$P_{xy}(x_m, y_2)$	$P_{xy}(x_m, y_3)$	...	$P_{xy}(x_m, y_n)$	$P_x(x_m)$
$\sum_{i=1}^{m} P_{xy}(x_i, y_j)$	$P_y(y_1)$	$P_y(y_2)$	$P_y(y_3)$	...	$P_y(y_n)$	

From the above table, it can be easily observed that the marginal probability mass functions of X and Y namely $P_x(x)$ and $P_y(y)$ respectively can be obtained from the joint probability mass function $P_{xy}(x, y)$ as

$$P_x(x_i) = \sum_{j=1}^{n} P_{xy}(x_i, y_j), \text{ for } i = 1, 2, \ldots, m$$

And

$$P_y(y_j) = \sum_{j=1}^{m} P_{xy}(x_i, y_j) \text{ for } j = 1, 2, 3, \ldots, n$$

- $P_{xy}(x_i, y_j) \geq 0 \ \forall \ i, j$
- $\sum_{i=1}^{m} \sum_{j=1}^{n} P_{xy}(x_i, y_j) = 1$
- The cumulative joint distribution function of the two dimensional random variable (X, Y) is given by $F_{xy}(x, y) = P(X \leq x, Y \leq y)$.

Joint Probability Density Function

Let X and Y are two continuous random variables on the same sample space S with $f_x(x)$ and $f_y(y)$ as the probability density functions respectively. Then a function $f_{xy}(x, y)$ is called the joint probability density function of the two dimensional random variable (X, Y) if the probability that the point (x, y) will lie in the infinitesimal rectangular region of area $dx\,dy$ is $f_{xy}(x, y)\,dx\,dy$,

That is,

$$P\left(x - \frac{1}{2}dx \leq X \leq x + \frac{1}{2}dx,\ y - \frac{1}{2}dy \leq Y \leq y + \frac{1}{2}dy\right)$$

$$= f_{XY}(x, y)\,dx\,dy$$

- $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XY}(x, y)dxdy = 1$
- The marginal probability density functions $f_X(x)$ and $f_Y(y)$ of the two continuous random variables X and Y are given by,

$$f_x(x) = \int_{-\infty}^{\infty} f_{XY}(x, y)dy \text{ and } f_y(y) = \int_{-\infty}^{\infty} f_{XY}(x, y)dx$$

- The cumulative joint distribution function $F_{XY}(x, y)$ of the two-dimensional random variable (X, Y) (where X and Y are any two continuous random variables defined on the same sample space) is given by,

$$F_{XY}(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{XY}(x, y)dxdy.$$

Conditional Probability Functions of Random Variables

Let X and Y be two discrete (continuous) random variables defined on the same sample space with joint probability mass (density) function $f_{XY}(x, y)$, then

1. The conditional probability mass (density) function $f_{\frac{X}{Y}}\left(\frac{x}{y}\right)$ of X, given $Y = y$ is defined as

   $F_{\frac{X}{Y}}\left(\frac{x}{y}\right) = \frac{f_{XY}(x, y)}{f_Y(y)}$, where $f_Y(y) \neq 0$ and

2. The conditional probability mass (density) function $f_{\frac{Y}{X}}\left(\frac{y}{x}\right)$ of Y, given $X = x$ is defined as $f_{\frac{Y}{X}}\left(\frac{y}{x}\right) = \frac{f_{XY}(x, y)}{f_X(x)}$ where $f_X(x) \neq 0$.

Independent Random Variables

Two discrete (continuous) random variables X and Y defined on the same sample space with joint probability mass (density) function $P_{XY}(x, y)$ are said to be independent, if and only if,

$$P_{XY}(x, y) = P_X(x)\,P_Y(y)$$

Where $P_X(x)$ and $P_Y(y)$ are the marginal probability mass (density) functions of the random variables X and Y respectively.

NOTE

If the random variables X and Y are independent then

$$P_{xy}(a \leq X \leq b, c \leq Y \leq d) = P_x(a \leq X \leq b)\,P_y(c \leq Y \leq d)$$

Statistics

Statistics is basically the study of numeric data. It includes methods of collection, classification, presentation, analysis and inference of data. Data as such is qualitative or quantitative in nature. If one speaks of honesty, beauty, colour, etc., the data is qualitative while height, weight, distance, marks, etc are quantitative.

The present course aims to systematically study statistics of quantitative data. The quantitative data can be divided into three categories

1. Individual series
2. Discrete series and
3. Continuous series

Individual Series

Examples:

1. Heights of 8 students

   5.0, 4.9, 4.5, 5.1, 5.3, 4.8, 5.1, 5.3 (in feet)

2. The weight of 10 students

   46, 48, 52, 53.4, 47, 56.8, 52, 59, 55, 52 (in kgs)

Discrete Series

Example:

x : Number of children in a family

f : Number of families

Total number of families = 50

x	0	1	2	3	4
f	8	10	19	8	5

Continuous Series

Example: Total number of students = 50

Class Interval (CI)	Frequency (f)
0–10	8
10–20	12
20–30	13
30–40	10
40–50	7

In order to analyze and get insight into the data some mathematical constants are devised. These constants concisely describe any given series of data. Basically we deal with two of these constants,

1. Averages or measures of central tendencies
2. Measures of spread or dispersion

Measures of Central Tendencies These tell us about how the data is clustered or concentrated. They give the central idea about the data. The measures are

1. Arithmetic mean or mean
2. Geometric mean
3. Harmonic mean
4. Median
5. Mode

The first three are mathematical averages and the last two are averages of position.

Measures of Dispersion It is possible that two sets of data may have the same central value, yet they may differ in spread. So there is a need to study about the spread of the data.

The measures we deal with are,

1. Range
2. Quartile deviation or semi inter-quartile range
3. Mean deviation
4. Standard deviation (including variance)

The formulae for each of the above mentioned measures is listed for each of the series in what follows.

Measures of Central Tendencies

Arithmetic Mean (AM or $\bar{x}$)

1. Individual series:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\Sigma x_i}{n}$$

2. Discrete series:

$$\bar{x} = \frac{f_1x_1 + f_2x_2 + \cdots + f_nx_n}{f_1 + f_2 + \cdots + f_n} = \frac{\Sigma f_i x_i}{\Sigma f_i}$$

where $x_1, x_2, \ldots, x_n$ are n distinct values with frequencies $f_1, f_2, f_3, \ldots, f_n$ respectively.

3. Continuous series:

$$\bar{x} = \frac{f_1m_1 + f_2m_2 + \cdots + f_nm_n}{f_1 + f_2 + \cdots + f_n} = \frac{\Sigma f_i m_i}{\Sigma f_i}$$

where $f_1, f_2, f_3, \ldots, f_n$ are the frequencies of the classes whose mid-values are $m_1, m_2, \ldots, m_n$ respectively.

Some Important Results Based on AM

1. The algebraic sum of deviations taken about mean is zero.
2. Its value is based on all items.
3. Mean of first n natural numbers is $\frac{(n+1)}{2}$.
4. Arithmetic mean of two numbers a and b is $\frac{(a+b)}{2}$.
5. If b is AM of a and c then a, b, c are in arithmetic progression.

Combined Mean If x_1 and x_2 are the arithmetic means of two series with n_1 and n_2 observations respectively, the combined mean, $\bar{x}_c = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$.

Median

If for a value the total frequency above (or below) it is half of the overall total frequency the value is termed as median. Median is the middle-most item.

Individual Series If $x_1, x_2, \ldots, x_n$ are arranged in ascending order of magnitude then the median is the size of $\left(\frac{n+1}{2}\right)$th item.

Some Results Based on Median

1. Median does not take into consideration all the items.
2. The sum of absolute deviations taken about median is least.
3. Median is the abscissa of the point of intersection of the cumulative frequency curves.
4. Median is the best suited measure for open end classes.

Mode The most frequently found item is called mode. Being so, it is easy and straight forward to find for individual and discrete series.

Empirical Formula

- For moderately symmetrical distribution,
- Mode = 3 median – 2 mean
- For a symmetric distribution, Mode = Mean = Median. This formula is to be applied in the absence of sufficient data. Given any two, of the mean, median or mode the third can be found.

Measures of Dispersion

Range

The range of a distribution is the difference between the greatest and the least values observed.

Some Important Results Based on Range

1. Range is a crude measure of dispersion as it is based only on the value of extreme observations.
2. It is also very easy to calculate.
3. It does not depend on the frequency of items.

Quartile Deviation (QD)

$$\text{QD} = \frac{Q_3 - Q_1}{2}$$

Individual Series The numbers are first arranged in ascending or descending order, then we find the quartiles Q_1 and Q_3 as

$Q_1 \rightarrow$ size of $[(n + 1)/4]$th item

$Q_3 \rightarrow$ size of $[3(n + 1)/4]$th item

The first quartile (or the lower quartile) Q_1 is that value of the variable, which is such that one-quarter of the observations lies below it. The third quartile Q_3 is that value of the variable, which is such that three-quarters of the observations lie below it.

Mean Deviation (MD)

It is defined as the arithmetic mean of the deviation from origin, which may be either mean or median or mode.

Individual Series

$$\text{MD} = \frac{|x_1 - A| + |x_2 - A| + \cdots + |x_n - A|}{n}$$

where $x_1, x_2, \ldots, x_n$ are the n observations and A is the mean or median or mode.

Some Results Based on MD

1. Mean deviation depends on all items.
2. By default, mean deviation is to be computed about mean.
3. Mean deviation about the median is the least.
4. Mean deviation of two numbers a and b is $\frac{|a-b|}{2}$.

Standard Deviation (SD)

Standard deviation is referred to as root mean squared deviation about the mean.

Individual Series

$$SD(\sigma) = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n}}$$

where $x_1, x_2, \ldots, x_n$ are n observations with mean as $\bar{x}$.

Alternatively $\sigma = \sqrt{\frac{\Sigma x_i^2}{n} - \left(\frac{\Sigma x_i}{n}\right)^2}$ is a useful formula for computational purpose.

Some Results Based on SD

1. The square of standard deviation is termed as variance.
2. SD is the least mean square deviation.
3. If each item is increased by a fixed constant the SD does not alter or SD is independent of change of origin.
4. Standard deviation depends on each and every data item.
5. For a discrete series in the form $a, a+d, a+2d, \ldots$ (AP), the standard deviation is given by $\text{SD} = d\sqrt{\frac{n^2-1}{12}}$, where n is number of terms in the series.

Co-efficient of Variation (CV)

Co-efficient of variation (CV) is defined as, $\text{CV} = \frac{\text{SD}}{\text{AM}} \times 100$.

This is a relative measure, which helps in measuring the consistency. Smaller the co-efficient of variation, greater is the consistency.

Example 17

For the individual series, compute the mean, median and mode 8, 11, 14, 17, 20, 23, 26, 29.

Solution

$$\text{Mean} = \bar{x} = \frac{\Sigma x_i}{n} = \frac{8+11+\cdots+29}{8} = 18.5$$

Median: As the numbers are in ascending order and the number 17 and 20 being middle terms.

$$\text{Median} = \frac{17+20}{2} = \frac{37}{2} = 18.5$$

Mode: As no term can be regarded as 'most often found', mode is not-defined. However using empirical formula,

Mode = 3 median – 2 mean

= 3(18.5) – 2(18.5) = 18.5.

Example 18

The arithmetic mean of 8, 14, x, 20 and 24 is 16; then find x.

Solution

$$\bar{x} = \frac{8+14+x+20+24}{5} = 16$$

$$\Rightarrow \quad \bar{x} = 80 - 66 = 14.$$

Example 19

Calculate standard deviation of first five prime numbers.

Solution

Given set of observations {2, 3, 5, 7, 11}

$$\frac{\Sigma x^2}{n} = \frac{208}{5}$$

$$\frac{\Sigma x}{n} = \frac{28}{5}$$

$$\therefore \text{ SD} = \sqrt{\frac{\Sigma x^2}{n} - \left(\frac{\Sigma x}{n}\right)^2}$$

$$= \sqrt{\frac{208}{5} - \left(\frac{28}{5}\right)^2} = 3.2.$$

Example 19

In a series of observations, co-efficient of variation is 25 and mean is 50. Find the variance.

Solution

Co-efficient of variation: $\text{CV} = \frac{\text{SD}}{\bar{x}} \times 100$

$$\Rightarrow \quad \text{SD} = \frac{\text{CV}}{100} \cdot \bar{x}$$

$$= 50 \times \frac{25}{100} = 12.5$$

Variance = $(12.5)^2 = 156.25$.

Hypothesis Testing

Introduction

In probability theory, we set up mathematical models of processes and systems that are affected by 'chance'. In statistics, we check these models against the reality, to determine whether they are faithful and accurate enough for practical purposes. The process of checking models is called statistical inference.

Methods of statistical inference are based on drawing samples (or sampling). One of the most important methods of statistical inference is 'Hypothesis Testing'.

Some Basic Definitions

Population

Population is the set of individuals or objects, animate or inanimate, actual or hypothetical under study.

Size of the Population The number of individuals or objects or observations in the population.

- The size of the population is denoted by N.
- N can be finite or infinite (i.e., population can be finite or infinite)

Sample: Any subset of the population is called as sample.

- The size (i.e., the number of elements) of the sample is denoted by n.
- n is always finite.

Examples:

1. All the GATE applicants—Population
   GATE applicants form a city—Sample
2. Cars manufactured by Tata Motors—Population
   Nano cars manufactured by Tata Motors—Sample
3. All possible outcomes of 10 roles of a die—Population
   12 possible outcomes of 10 roles of a die—Sample
4. Number of units of electricity consumed by the residents of a colony in a city—Population
   Number of units of electricity consumed by the residents of 5 houses of that colony—Sample
5. Diameters of screws produced by a company—Population
   Diameters of screws produced on one machine of that company—Sample

Sampling

The process of drawing samples from the population is called sampling.

Random Sampling A sampling in which each member of the population has the same chance of being included in the sample is called random sampling.

Simple Sampling A random sampling in which the chance of being included in the sample for different members of the population is independent of whether included or not in the previous trails is called simple sampling.

Large and Small Samples If the size of the sample is greater than or equal to 30 (i.e., $n \geq 30$), then the sample is called a large sample. Otherwise it is called a small sample.

Parameter A statistical measure or constant of the population is called a parameter.

Examples:

1. Population mean (denoted by μ)
2. Population standard deviation (denoted by σ)

Statistic A statistical measure or constant of the sample drawn from the population is called a statistic. (statistic–singular, statistics–plural)

Examples:

1. Sample mean (denoted by $\bar{x}$)
2. Sample standard deviation (denoted by s)

NOTE

In general, the population parameters are not known and their estimates given by the corresponding sample statistics are used.

Sampling Distribution Consider samples of size n drawn from a given population. Compute some statistic S, say mean ($\bar{x}$) or variance (s^2) for each of the samples. The values of the statistics can be given in the form of a frequency table. The frequency table so formed is known as a sampling distribution of the statistic.

Example: Consider the set of numbers {1, 2, 3, 4, 5, 6} as population.

Consider the following 15 samples each of size 3 drawn from the above population.
(1, 2, 3), (3, 5, 5), (2, 4, 6), (5, 5, 5), (1, 2, 6)
(1, 3, 5), (6, 6, 6), (4, 4, 5), (2, 3, 4), (1, 1, 4)
(2, 5, 5), (2, 2, 5), (3, 4, 6), (2, 4, 5), (4, 5, 6)

Then the sampling distribution of means for these samples is

Sample Mean ($\bar{x}$)	2	3	3.67	4	4.33	5	6
Frequency	2	4	1	2	3	2	1

Standard Error The standard deviation of the sampling distribution of a statistic is called the standard error (SE) of that statistic.

- The standard deviation of the sampling distribution of means is called the standard error of means where as the standard deviation of the sampling distribution of variances is called the standard error of variances.

Precision: The reciprocal of the standard error is called precision.

NOTE

If the sample size n is large, (i.e., $n \geq 30$), then the sampling distribution of a statistic is approximately normal. (Irrespective of the population distribution being normal or not)

Testing of Hypothesis

We have some information about a characteristic of the population which may or may not be true. This information is called statistical hypothesis or briefly hypothesis. We wish to know, whether this information can be accepted or to be rejected. We choose a random sample and obtain information about this characteristic. Based on this information, a process that decides whether the hypothesis to be accepted or rejected is called testing of hypothesis. i.e., In brief, the test of hypothesis or the test of significance is a procedure to determine whether observed samples differ significantly from expected results.

Null Hypothesis and Alternative Hypothesis

Null Hypothesis A statistical hypothesis which is to be actually tested for acceptance or rejection is called a null hypothesis.

(According to RA Fisher, Null hypothesis is the hypothesis which is tested for possible rejection under the assumption that it is true)

- Null hypothesis is denoted by H_0

Alternative Hypothesis Any hypothesis other than the null hypothesis is called an alternative hypothesis.

- Alternative hypothesis is denoted by H_1
- Let θ be a population parameter and θ_0 be the specified value of θ. Then we define null and alternative hypotheses as follows.

Null hypothesis $H_0 : \theta = \theta_0$
Alternative Hypothesis $H_1 : \theta \neq \theta_0$ (two tailed alternative)
(OR) $H_1 : \theta > \theta_0$ (right tailed alternative)
(OR) $H_1 : \theta < \theta_0$ (left tailed alternative)

Type I and Type II Errors

Type I Error Rejecting the null hypothesis (H_0), when it should be accepted is called type I error.

Type II Error Accepting the null hypothesis (H_0) when it should be rejected is called type II error.

	Accept H_0	**Reject H_0**
H_0 is true	Correct decision	Type I error
H_0 is false	Type II error	Correct decision

Level of Significance

The probability level, below which, we reject the null hypothesis is called the level of significance.

(OR)

The probability of committing type I error is known as the level of significance.

- The level of significance is denoted by 'α'.
- It is customary to fix α, before sample information is collected.
- In most of the cases, we choose α as 0.05 or 0.01.
- $\alpha = 0.05$ is used for moderate precision and $\alpha = 0.01$ is used for high precision.
- Level of significance can also be expressed as percentage. $\alpha = 5\%$ means there are 5 chances in 100 that the null hypothesis H_0 is rejected when it is true or one is 95% confident that a right decision is made.
- The probability of committing type II error is denoted by β. $\therefore \beta = P$ (accept H_0 when H_0 is false)
- Level of significance (α) is also known as the size of the test.
- $1 - \beta$ is known as the power of the test.

Critical Region and Critical Value

Consider the area under the probability curve of the sampling distribution of the test statistic which follows some known distribution. The area under the probability curve is divided into two regions, namely the region of rejection where null hypothesis is rejected and the region of acceptance where null hypothesis is accepted.

Critical Region (or) the Region of Rejection (or) the Significant Region

The region under the probability curve of the sampling distribution of the test statistic, where the null hypothesis (H_0) is rejected is called the critical region.

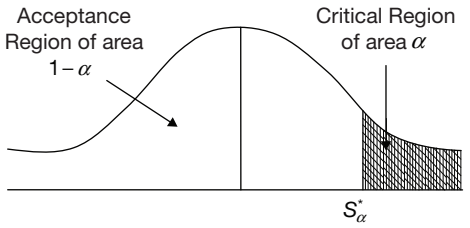
- The area of the critical region is equal to the level of significance α.

Critical Value (OR) Significant Value

The value of the test statistic (for given level of significance α) which separates the area under the probability curve into critical and non-critical regions.

One Tailed and Two Tailed Tests

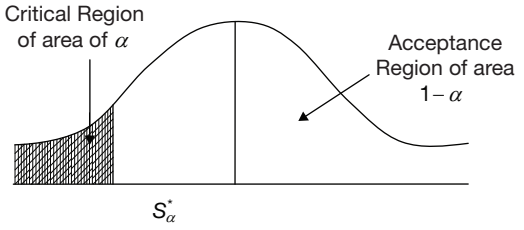
1. **Right one-tailed test:** If the alternative hypothesis H_1 is of greater than type (For example, $H_1 : \mu > \mu_0$ or $H_1 : \sigma_1^2 > \sigma_2^2$) then the entire critical region of area α lies on the right side tail of the probability curve of the test statistic S^* as shown in the figure. In this case, the test of hypothesis is known as right one-tailed test.



Right one-tailed test

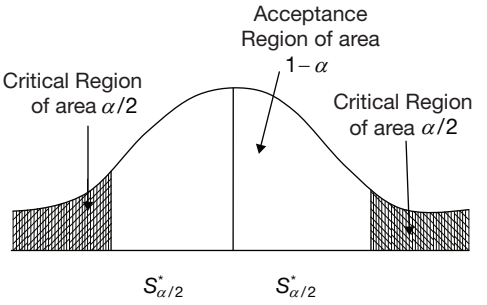
2. **Left one-tailed test:** If the alternative hypothesis H_1 is of less than type (For example, $H_1 : \mu_1 - \mu_2 < 0$ or $H_1 : \sigma^2 < \sigma_1^2$) then the entire critical region of area α lies on the left side tail of the probability curve of the test statistic S^* as shown in the figure.

   In this case, the test of hypothesis is known as left one-tailed test.



Left one-tailed test

3. **Two tailed test:** If the alternative hypothesis H_1 is of not equal to type (For example, $H_1 : \mu \neq \mu_0$ or $H_1 : \sigma_1^2 \neq \sigma_2^2$), then the critical region lies on both sides (right and left tails) of the probability curve of the test statistic S^* such that the critical region of area $\frac{\alpha}{2}$ lies on the right tail and the critical region of area $\frac{\alpha}{2}$ lies on the left tail as shown in the figure.



Two-tailed test

In this case, the test of hypothesis is known as two-tailed test.

Procedure for Test of Hypothesis

Step 1: Formulate null hypothesis H_0

Step 2: Formulate alternative hypothesis H_1

Step 3: Choose the level of significance α

Step 4: Identify the critical region based on the critical value S_α^* and the alternative hypothesis.

Step 5: Compute the test statistic S^* using the sample data (Formulae for finding the values of the test statistics under different tests of hypothesis were given while describing those tests).

Step 6: If the value of S^* comes under the critical region, then reject the null hypothesis H_0 and if the value of S^* comes under the non-critical (acceptance) region, then accept the null hypothesis H_0.

Central Limit Theorem If $\bar{x}$ is the mean of a sample of size n drawn from a population with mean μ and finite variance σ^2, then the limiting distribution of $Z = \dfrac{\bar{x} - \mu}{\dfrac{\sigma}{\sqrt{n}}}$ as $n \to \infty$ is the standard normal distribution (i.e., mean zero and standard deviation 1)

NOTES

1. If the sample size is large, then whether the population is normally distributed or not, the sampling distribution of means always follow a normal distribution.
2. If the sample size is small and the population from which the samples are drawn follows a normal distribution, then the sampling distribution of means also follows a normal distribution.

Tests of Hypothesis for Large Samples

If the sample size is large, then the standard error (SE) forms the basis for the testing of hypothesis. Also, we know that if sample size is large, then the sampling distribution of any statistic S is normal. So, in large sampling, we can relate the value of the test statistic S^* with the standard normal random variable Z as:

$$Z = \frac{S^* - E(S^*)}{SE(S^*)}$$

Where

S^* = Value of the test statistic

$E(S^*)$ = Expected Value (value of the corresponding population parameter)

$SE(S^*)$ = Standard error of the test statistic.

Following table gives the information about the standard errors and test statistics for various cases in testing of hypothesis for large samples.

Test of Hypothesis (significance)		Standard Error = SE = σ^*	Test Statistic = Z	Expansions for Notations
Difference between	Sample mean ($\bar{x}$) and population mean (μ)	$\frac{\sigma}{\sqrt{n}}$	$\frac{\bar{x} - \mu}{\sigma^*}$	μ = Population mean $\bar{x}$ = Sample mean
	Means of two samples ($\bar{x}_1$ and $\bar{x}_2$)	$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$	$\frac{\bar{x}_1 - \bar{x}_2}{\sigma^*}$	σ = Population standard deviation
	Sample standard deviation (s) and population standard deviation (σ)	$\frac{\sigma}{\sqrt{2n}}$	$\frac{s - \sigma}{\sigma^*}$	n = Sample size s = Sample standard deviation
	Sample standard deviations (s_1 and s_2)	$\sqrt{\frac{\sigma_1^2}{2n_1} + \frac{\sigma_2^2}{2n_2}}$	$\frac{s_1 - s_2}{\sigma^*}$	P = Population proportion $Q = 1 - P$
	Sample proportion (p) and population proportion (P)	$\sqrt{\frac{PQ}{n}}$	$\frac{p - P}{\sigma^*}$	p = Sample proportion
	Two sample proportions (P_1, P_2)	$\sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}$	$\frac{P_1 - P_2}{\sigma^*}$	

Example 21

The dean of an engineering college claims that the average attendance of students in the final semester of B.Tech is 72.5% with a standard deviation of 5.7%. To test this claim, the attendance of a random sample of 49 students of final semester of B.Tech were examined, which showed the average as 74.4%. Can the claim be accepted or not at 1% level of significance?

Solution

Here population mean = $\mu_0 = 72.5$

Simple mean = $\bar{x} = 74.4$

Population standard deviation = σ

$= 5.7$

Level of significance $= \alpha = \frac{1}{100} = 0.01$

Sample size = $n = 49$

Null hypothesis H_0: $\mu = \mu_0 = 72.5$

Alternative hypothesis H_1: $\mu \neq \mu_0$ $(= 72.5)$

Level of significance, $\alpha = 0.01$

Test statistic,

$$Z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} = \frac{74.4 - 72.5}{\frac{5.7}{\sqrt{49}}}$$

$$= \frac{1.9}{5.7} \times 7 = 2.333$$

Critical region: As the alternative hypothesis is of $\neq$ type, the test should be a two tailed test, where the critical region lies on both sides of the curve as shown in the figure.

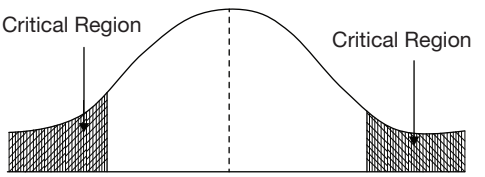
$\alpha = 0.01$

$\Rightarrow \frac{\alpha}{2} = 0.005$

$\therefore P(Z \leq -Z_{\alpha/2}) = 0.05$

$\Rightarrow -Z_{\alpha/2} = -2.575$

$\Rightarrow Z_{\alpha/2} = 2.575$



Decision: It can be easily observed that the value of the test statistic lies between $-Z_{\alpha/2}$ and $Z_{\alpha/2}$ i.e., The test statistic is not in the critical region.

Hence we accept the null hypothesis H_0

$\therefore$ The claim of the dean can be accepted.

Example 22

Can it be concluded that the average life span of an electric bulb is more than 200 hours, if a random sample of 100 electric bulbs has an average life span of 202 hours with a standard deviation of 8 hours with level of significance 0.05

Solution

Here population mean $\mu_0 = 200$

Sample mean = $\bar{x} = 202$

Sample standard deviation = $s = 8$

Sample size = $n = 100$

Null hypothesis H_0: $\mu = \mu_0 = 200$

Alternative hypothesis H_1: $\mu > \mu_0$ (= 202)

Level of significance, $\alpha = 0.05$

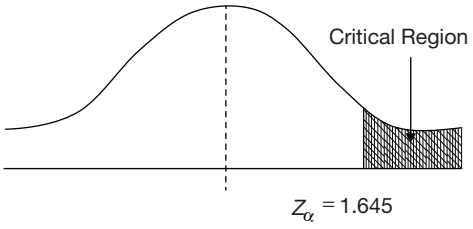
Test statistic, $Z = \dfrac{\bar{x} - \mu_0}{\dfrac{\sigma}{\sqrt{n}}} = \dfrac{\bar{x} - \mu_0}{\dfrac{s}{\sqrt{n}}}$

($\because$ As the population standard deviation σ is not given and the sample size is large, we can consider the sample standard deviation s as the population standard deviation)

$$\therefore Z = \frac{202 - 200}{\dfrac{8}{\sqrt{100}}}$$

$$Z = 2.5$$

Critical region: As the alternative hypothesis H_1 is of $>$ type, the test should be a right one tailed test, where the critical region lies on the right tail of the standard normal curve as shown in the figure.



Here $\alpha = 0.05$

$\therefore P(Z \geq Z_\alpha) = 0.05$

$\Rightarrow P(Z \leq -Z_\alpha) = 0.05$

$\Rightarrow -Z_\alpha = -1.645$

$\Rightarrow Z_\alpha = 1.645$

$\therefore$ The critical region is to the right of Z_α

(i.e., to the right of $Z = 1.645$) under the standard normal curve.

Decision: As the value of the test statistic $Z = 2.5$ is greater than that of $Z_\alpha = 1.645$, the test statistic lies in the critical region.

$\therefore$ Reject the null hypothesis $H_0 : \mu = 200$

Hence accept the alternative hypothesis $H_1 : \mu > 200$

$\therefore$ We can conclude that the average life span of an electric bulb is more than 200 hours.

NOTE

In the process of testing of hypothesis for large samples, if the population standard deviation σ is not given, then the sample standard deviation s can be assumed as the population standard deviation.

Example 23

In a city, a random sample of 36 men has an average life span of 71 years with a standard deviation of 9 years, while a random sample of 49 women has an average life span of 76 years with a standard deviation of 14 years. Does this substantiate the claim that the life span of men is less than that of women in that city with 1% level of significance?

Solution

Let $\bar{x}_1$ = Average life span of men = 71 years

And $\bar{x}_2$ = Average life span of women

= 76 years

$s_1 = 9$ and $s_2 = 14$

$n_1 = 36$ and $n_2 = 49$

Level of significance = $\alpha = 0.01$

Null hypothesis H_0: $\mu_1 = \mu_2$ (i.e., The average life span of men and women in the city is same)

Alternative hypothesis H_1: $\mu_1 < \mu_2$ (i.e., The average life span of men is less than that of women in the city)

Level of significance: $\alpha = 0.01$

Test statistic,

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$$

i.e., $$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$$

($\because$ The standard deviations s_1 and s_2 are given and the standard deviations of populations are unknown)

$$= \frac{(71 - 76)}{\sqrt{\dfrac{9^2}{36} + \dfrac{14^2}{49}}} = \frac{-5}{\sqrt{\left(\dfrac{9}{4} + 4\right)}}$$

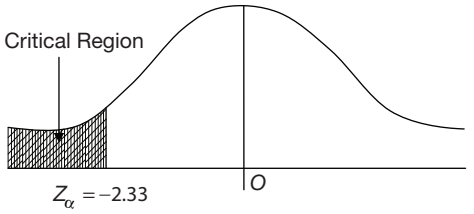
$$Z = -2$$

Critical region: As the alternative hypothesis H_1 is of $<$ type, the critical region should be in the left tail of the standard normal curve as shown in the figure.

Here $\alpha = 0.01$

$\therefore P(Z \le Z_\alpha) = \alpha = 0.01$

$\Rightarrow Z_\alpha = -2.33$



$\therefore$ The critical region is to the left of $Z_\alpha = -2.33$ under the standard normal curve.

Decision: As the value of the test statistic $Z = -2$ is greater than that of $Z_\alpha = -2.33$, the test statistic does not lie in the critical region.

$\therefore$ Accept the null hypothesis H_0

Hence reject the alternative hypothesis H_1

$\therefore$ The given information does not substantiate the claim that the life span of men is less than that of women in that city.

Example 24

The manufacturer of electronic weighing machines finds that in a random sample of 120 machines, 15 machines are defective. Find the standard error of the proportion of defective machines in the sample.

Solution

Total number of machines = Sample size = $n = 120$

$\therefore$ The number of defective machines = $15 = x$ (say)

$\therefore$ Proportion of defective machines

$$= p = \frac{x}{n} = \frac{15}{120} = \frac{1}{8}$$

$$\therefore q = 1 - p = 1 - \frac{1}{8} = \frac{7}{8}$$

As the population proportion P (and hence $Q = 1 - P$) is unknown, we take p and q as P and Q respectively.

$\therefore$ Standard error of he proportion of defective machines

$$= SE = \sqrt{\frac{PQ}{n}}$$

$$= \sqrt{\frac{pq}{n}} = \sqrt{\frac{\frac{1}{8} \times \frac{7}{8}}{120}}$$

$\therefore SE = 0.0302$

Example 25

A home appliances company claims that the life of its geysers has a standard deviation of 16 hours. The life of a sample of 98 geysers of that company was found to have a standard deviation of 18 hours. Find the test statistic Z that is used in the process of testing whether the claim of the company be accepted or not.

Solution

Population standard deviation = $\sigma = 16$

Sample standard deviation = $s = 18$

Sample size = $n = 98$

As the situation is a testing of hypothesis of difference between population and sample standard deviations, we have

$$\text{Test statistic } = Z = \frac{(s - \sigma)}{\frac{\sigma}{\sqrt{2n}}}$$

$$= \frac{(18 - 16)}{\frac{16}{\sqrt{2 \times 98}}} = \frac{2 \times 14}{16}$$

$\therefore Z = 1.75.$

Tests of Hypothesis for Small Samples

In case of large samples, we often made use of the fact that the sampling distribution of many statistics are approximately normal and values of sample statistics are considered best estimates of the parameters of a population. However, in case of small samples, the sampling distributions of many statistics are not normal and the approximations of population parameters by the corresponding sample statistics are not valid. So, we shall discuss different tests of hypothesis which are applicable to small sampling. Note that these tests of hypothesis for small samples can also be applied to the cases of large samples. First we will discuss three important distributions that are used in testing of hypothesis for small samples namely, t-distribution, F-distribution and χ^2-distribution. These distributions require the knowledge of the concept of 'Degrees of freedom'.

Degrees of Freedom

The number of degrees of freedom is defined as the number of values in a set, which may be assigned arbitrarily.

For example, if $x_1 + x_2 + x_3 + x_4 + x_5 = 18$, then assign any values for four of the five variables arbitrarily (say, x_1, x_2, x_3 and x_4 were given arbitrary values). Then the value of the fifth variable (x_5) has to be taken based on the values of x_1, x_2, x_3 and x_4. So, in this case, the degrees of freedom is 4.

With reference to statistics, if n is the number of observations in the small sample and k is the number of constraints

on them (or k values are already available), then the number of degrees of freedom can be obtained by $n - k$.

The number of degrees of freedom is denoted by ν.

Example: If $x_1, x_2, \ldots x_n$ are the observations given, then the number of degrees of freedom for the mean $\bar{x}$ is n ($\because$ we use all values to find $\bar{x}$)

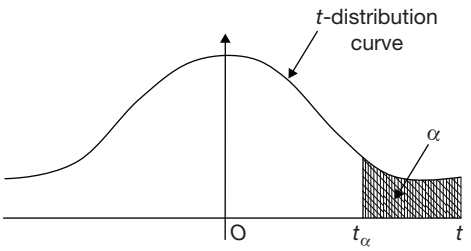
The number of degrees of freedom for the variance is $n - 1$ ($\because$ The variance depends on the mean)

Student's t-Distribution (or) t-Distribution

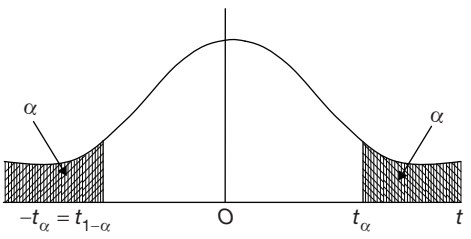
A random variable t is said to follow the t-distribution with $\nu = n - 1$ degrees of freedom, n being the sample size, if its probability density function is given by

$$f(t) = \frac{1}{\sqrt{\upsilon\beta\left(\frac{1}{2}, \frac{\upsilon}{2}\right)}}\left(1 + \frac{t^2}{\upsilon}\right)^{-(\upsilon+1)/2} ; -\infty < t < \infty$$

The t-distribution curve is as shown in the figure, which is symmetric about the mean 0 and bell shaped. The total area under the t-distribution curve is unity.



- The t-distribution curve is similar to normal curve.
- The variance of t-distribution is greater than 1 and depends on the degrees of freedom ν.
- As the sample size n (i.e., the degrees of freedom $n - 1$) becomes large, the variance of corresponding t-distribution approaches 1 and hence for large samples, t-distribution can be approximated by the standard normal distribution.
- Critical values of t-distribution (see the t-distribution tables) are denoted by t_α, which is such that the area under the curve to the right of t_α equals to α.
- As the t-distribution is symmetric, it follows that $t_{1-\alpha} = -t_\alpha$.

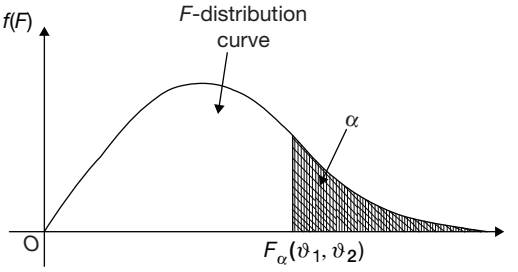


F-Distribution

A random variable F is said to follow the F-distribution with (ν_1, ν_2) degrees of freedom, if its probability density function $f(F)$ is given by

$$f(F) = \frac{\left(\frac{\nu_1}{\nu_2}\right)^{\frac{\nu_1}{2}} F^{\left(\frac{\nu_1}{2}\right)-1}}{\beta\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right)\left[1 + \left(\frac{\nu_1}{\nu_2}\right)F\right]^{(\nu_1+\nu_2)/2}}, F > 0$$

The graph of F-distribution is given below.



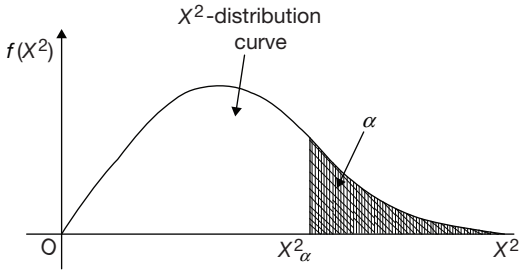
- The F-distribution curve entirely lies in the first quadrant.
- F-distribution is not symmetric.
- $F_\alpha(\nu_1, \nu_2)$ is the value of F with ν_1 and ν_2 degrees of freedom such that the area under the F-distribution curve to the right of $F_\alpha(\nu_1, \nu_2)$ is α.
- The value of $F_\alpha(\nu_1, \nu_2)$ not only depends on the values of the degrees of freedom ν_1 and ν_2, but also the order in which they were taken.
- $F_{1-\alpha}(\nu_1, \nu_2) = \dfrac{1}{F_\alpha(\nu_2, \nu_1)}$
- F-distribution is also known as variance ratio distribution.
- The values of F_a for $\alpha = 0.05$ and $\alpha = 0.01$ for various combinations of the degrees of freedom ν_1 and ν_2 were presented in the tables.
- For large values of ϑ_1 and ϑ_2, F-distribution can be approximated by a normal distribution $N\left[1, 2\left(\frac{1}{\partial_1} + \frac{1}{\partial_2}\right)\right]$ with mean 1 and variance $2\left(\frac{1}{\partial_1} + \frac{1}{\partial_2}\right)$.
- F-distribution and t-distribution can be related as follows. If a statistic t follows t-distribution with ν degrees of freedom, then t^2 follows F-distribution with degrees of freedom $\nu_1 = 1$ and $\nu_2 = \nu$.

Chi-square Distribution

If a random variable X follows chi-square distribution (denoted as χ^2-distribution or $\chi^2(\nu)$), then the probability density function of X is given by

$$f(\chi^2) = \frac{1}{2\frac{\nu}{2}\Gamma\left(\frac{\nu}{2}\right)} e^{-\chi^2/2}(\chi^2)^{\left(\frac{\nu 1}{2}\right)^{-1}}, 0 \le \chi^2 < \infty$$

where ν is the degrees of freedom
The graph of chi-square distribution is given below.



- The chi-square distribution curve entirely lies in the first quadrant.
- Chi-square distribution is not symmetric.
- χ^2-distribution depends only on ϑ, the degrees of freedom.
- If χ_1^2 and χ_2^2 are two independent distributions with ν_1 and ν_2 degrees of freedom respectively, then $\chi_1^2 + \chi_2^2$ will follow chi-square distribution with $(\nu_1 + \nu_2)$ degrees of freedom.
- χ_α^2 represents the value of χ^2 such that the area under the chi-square curve to the right of χ_α^2 is α
- The value of χ_α^2 for various combinations of α and ϑ were presented in the table.
- As the number of degrees of freedom $\vartheta \to \infty$, the χ^2-distribution tends to the normal distribution.

Identifying the Values of t_α, F_α and χ_α^2 from the Tables

1. The value of t_α for a given degrees of freedom ϑ is the value in the t-table at the intersection point of the column headed by α and the row headed by the degrees of freedom ϑ.

Example 26

Find the values of

(i) $t_{0.1}$ with degrees of freedom $\vartheta = 12$

(ii) $t_{0.05}$ with degrees of freedom $\vartheta = 17$

(iii) $t_{0.98}$ with degrees of freedom $\vartheta = 23$

Solution

(i) $t_{0.1}$ with degrees of freedom $\vartheta = 12$ = The value in the t-table at the intersection point of the column headed by $\alpha = 0.1$ and the row headed by $\vartheta = 12$ = 1.356

(ii) $t_{0.05}$ with degrees of freedom $\vartheta = 17$ = The value in the t-table at the intersection point of the column headed by $\alpha = 0.05$ and the row headed by $\vartheta = 17$ = 1.740.

(iii) $t_{0.98}$ with degrees of freedom $\vartheta = 23$. In the given t-table, there is no column corresponding to $\alpha = 0.98$

But we know that

$t_{1-\alpha} = -t_\alpha$

i.e., $t_\alpha = -t_{1-\alpha}$

$\therefore t_{0.98} = -t_{1-0.98}$

$\Rightarrow t_{0.98} = -t_{0.02}$ (1)

Now $t_{0.02}$ with degrees of freedom $\vartheta = 23$

= 2.177

$\therefore$ From (1), $t_{0.98}$ with degrees of freedom $\vartheta = 23$

= –2.177.

2. Two tables were given for F-distribution, one each for the values of $\alpha = 0.05$ and $\alpha = 0.01$ for various combinations of degrees of freedom ϑ_1 and ϑ_2.

- The value of F_α for a given pair of values of degrees of freedom ϑ_1 and ϑ_2 is the value in the respective F_α-table at the intersection point of column headed by ϑ_1 and the row headed by ϑ_2.

3. The value of χ_α^2 for a given degrees of freedom ϑ is the value in the χ^2-table at the intersection point of the column headed by α and the row headed by the degrees of freedom ϑ.

Example 27

Find the values of

(i) $F_{0.05}(6, 13)$;

(ii) $F_{0.01}(12, 17)$ and

(iii) $F_{0.95}(15, 24)$

Solution

(i) $F_{0.05}(6, 13)$ = The value in the F-table corresponding to $\alpha = 0.05$ at the intersection point of the column headed by $\vartheta_1 = 6$ and the row headed by $\vartheta_2 = 13$ = 2.92

(ii) $F_{0.01}(12, 17)$ = The value in the F-table corresponding to $\alpha = 0.01$ at the intersection point of the column headed by $\vartheta_1 = 12$ and the row headed by $\vartheta_2 = 17$ = 3.46

(iii) $F_{0.95}(15, 24)$

e know that $F_\alpha(\vartheta_1, \vartheta_2)$

$$= \frac{1}{F_{1-\alpha}(\vartheta_2, \vartheta_1)}$$

$\therefore F_{0.95}(15, 24)$

$$= \frac{1}{F_{1-0.95}(24,15)}$$

$$= \frac{1}{F_{0.05}(24,15)}$$

$$= \frac{1}{2.29} = 0.4367.$$

Example 28

Find the values of

(i) $\chi^2_{0.05}$ with degrees of freedom $\vartheta = 16$

(ii) $\chi^2_{0.01}$ with degrees of freedom $\vartheta = 21$

(iii) $\chi^2_{0.10}$ with degrees of freedom $\vartheta = 4$

Solution

(i) $\chi^2_{0.05}$ with degrees of freedom $\vartheta = 16$

= The value in the χ^2 – table at the intersection point of the column headed by $\alpha = 0.05$ and the row headed by $\vartheta = 16$

$= 26.296$

(ii) $\chi^2_{0.01}$ with degrees of freedom $\vartheta = 21$

= The value in the χ^2 – table at the intersection point of the column headed by $\alpha = 0.01$ and the row headed by $\vartheta = 21$

$= 38.932$

Similarly, we have

(iii) $\chi^2_{0.10}$ with degrees of freedom $\vartheta = 4$

$= 7.779.$

Test of Hypothesis	Test Statistic	Distribution with Degrees of Freedom	Expansions for Notations
Difference between means of population and sample (σ unknown)	$t = \dfrac{\bar{x} - \mu_0}{\dfrac{s}{\sqrt{n-1}}}$	t-distribution with $v = n - 1$	$\bar{x}$ = Sample mean
Difference between means of two samples (σ unknown) (a) If $n_1 \neq n_2$	$t = \dfrac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\left(\dfrac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}\right)\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$	t-distribution with $v = n_1 + n_2 - 2$	μ_0 = Population mean
(b) If $n_1 = n_2 = n$	$t = \dfrac{(\bar{x_1} - \bar{x_2})}{\sqrt{\dfrac{(s_1^2 + s_2^2)}{(n-1)}}}$	t-distribution with $v = 2n - 2$	s = Sample standard deviation n = Sample size ϑ = Degrees of freedom
(c) $n_1 = n_2 = n$ and the two samples are not independent i.e., they are related in some way (This implies that the pairs of observations (x_i, y_i) belong to same sample unit	$t = \dfrac{\bar{d}}{\dfrac{s}{\sqrt{n-1}}}$ where $\bar{d} = \bar{x_i} - \bar{y_i}$ $d_i = x_i - y_i$ and $s_2 = \dfrac{1}{n}\sum_i (d_i - \bar{d})^2$	t-distribution with $v = n - 1$	
Equality of the population variances **Note:** Take the larger of the estimates of variances of the samples as $\hat{\sigma}_1^2$ and the corresponding degrees of freedom as ϑ_1	$F = \dfrac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2}$ where $\hat{\sigma}_1^2 = \left(\dfrac{n_1}{n_1 - 1}\right)s_1^2$ and $\hat{\sigma}_2^2 = \left(\dfrac{n_2}{n_2 - 1}\right)s_2^2$	F-distribution with $v_1 = n_1 - 1$ $v_2 = n_2 - 1$	
Population variance	$\chi^2 = \dfrac{ns^2}{\sigma^2}$ 0 where $s^2 = \sum_{i=1}^{n} \dfrac{(x_i - \bar{x})^2}{n}$	χ^2-distribution with $v = n - 1$	

Example 29

The highest temperature in the month of June at a certain place is normally distributed with mean 40°C. The highest temperatures in June during the last five years are 43°C, 37°C, 35°C, 39°C and 38°C. From this data, can we conclude that the average highest temperature in June during the last five years is less than the normal highest temperature? (Test at 0.05 level of significance)

Solution

Population mean $= \mu_0 = 40$

Sample mean $= \bar{x} = \dfrac{43 + 37 + 35 + 39 + 38}{5}$

$\therefore \bar{x} = \dfrac{192}{5} = 38.4$

Sample size $v\ n = 5$

$\therefore$ Number of degrees of freedom $= \vartheta = n - 1 = 5 - 1 = 4$

$$\text{Sample variance} = s^2 = \left(\frac{1}{n}\sum_i x_i^2\right) - \overline{x^2}$$

$$= \frac{43^2 + 37^2 + 35^2 + 39^2 + 38^2}{5} - (38.4)^2$$

$= 1{,}481.6 - 1{,}474.56$

$\therefore s^2 = 7.04$

$\Rightarrow s = 2.6533$

Null hypothesis H_0: $\mu = \mu_0 = 40$

Alternative hypothesis: H_1: $\mu < \mu_0$ (=40)

Level of significance: $\alpha = 0.05$

Test statistic:

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n-1}}}$$

$$= \frac{38.4 - 40}{\frac{2.6533}{\sqrt{5-1}}} = \frac{-1.6}{\frac{2.6533}{2}}$$

$\therefore t = -1.2060$

Critical region: As the alternative hypothesis is < type, the critical region is in the left tail of the t-distribution curve as shown in the figure.

As $\alpha = 0.05$ with $\vartheta = 4$, we have

t_α with $\vartheta = 4$

= The area under the t-distribution curve to the right of $\alpha = 0.05$ with $\vartheta = 4$

$= 2.132$

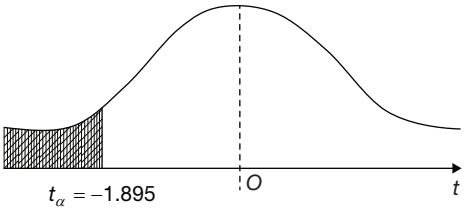
$\therefore$ The critical region is the region (area = 0.05) in the left tail of the t-distribution curve.

$= -t_\alpha$ with $\vartheta = 4$

($\because$ t-distribution is symmetric)

$\therefore$ The critical value is

$\therefore -t_\alpha = -2.132$



Decision: As the test statistic $t = -1.2060$ is greater than the critical value -2.132, it lies in the acceptance region.

$\therefore$ Accept the null hypothesis.

Hence there is no significant difference between the normal temperature and the average temperature of the last five years in the month of June.

Example 30

In a CBSE school marks scored in Mathematics by 10 students of section-A of X standard has a mean of 68 and a variance of 109 where as that of 8 students of Section-B has a mean of 57 with a variance of 128. Test of 2% level of significance whether there is a significant difference between the means of marks scored by the students of sections-A and B or not. Assume that the marks of the students of sections-A and B follow normal distribution with same variance.

Solution

Let μ_1 and μ_2 be the means of the marks of students of the sections A and B respectively.

Mean of first sample $= \bar{x}_1 = 68$

Mean of second sample $= \bar{x}_2 = 57$

Variance of first sample $= s_1^2 = 109$

Variance of second sample

$= s_2^2 = 128$

n_1=10 and $n_2 = 8$

Null hypothesis: H_0: $\mu_1 = \mu_2$

Alternative hypothesis: H_1: $\mu_1 \neq \mu_2$

Level of significance: $\alpha = 0.02$

Test statistic:

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\left(\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

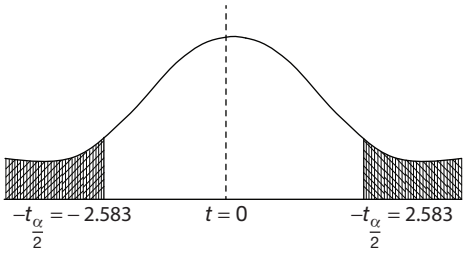
$$= \frac{(68 - 57)}{\sqrt{\left(\frac{10 \times 109 + 8 \times 128}{10 + 8 - 2}\right)\left(\frac{1}{10} + \frac{1}{8}\right)}}$$

$$= \frac{11}{5.4523}$$

$= 2.0175$

$\therefore$ The test statistic is $t = 2.0175$

Critical region: As the alternative hypothesis is of the $\neq$ type, the critical region lies on both sides of the t-distribution curve as shown in the figure.



Here $\alpha = 0.02$ and υ = degrees of freedom $= n_1 + n_2 - 2 = 10 + 8 - 2$

$\upsilon = 16$

$\frac{t_\alpha}{2}$ with $\upsilon = 16$

$= t_{0.01}$ at $\upsilon = 16$

$= 2.583$

$\therefore$ The critical region is to the left of $-\frac{t_\alpha}{2} = -2.583$ and to the right of $\frac{t_\alpha}{2} = 2.583$.

Decision: As the test statistic $t = 2.0175$ lies between $-\frac{t_\alpha}{2}$ and $\frac{t_\alpha}{2}$, we accept the null hypothesis.

$\therefore$ There is no significant difference between the means of marks scored by the students of sections A and B.

Example 31

The variances of two samples of sizes 9 and 13 are 15.778 and 19.175 respectively. Test whether the two samples be regarded as drawn from normal populations with the same variance at 5% level of significance.

Solution

Always the sample having higher variance will be taken as the first sample in this test of hypothesis.

$\therefore$ Variance of first sample $= s_1^2$

$= 19.175$

Variance of second sample $= s_2^2$

$= 15.778$

Sample size of first sample $= n_1 = 13$

Sample size of second sample $= n_2 = 9$

$$\hat{\sigma}_1^2 = \left(\frac{n_1}{n_1 - 1}\right) s_1^2 = \frac{13}{(13-1)} \times 19.175$$

$$\hat{\sigma}_1^2 = 20.7729$$

$$\hat{\sigma}_2^2 = \left(\frac{n_2}{n_2 - 1}\right) s_2^2 = \frac{9}{(9-1)} \times 15.778$$

$$\hat{\sigma}_2^2 = 17.7520$$

Null hypothesis: H_0: $\hat{\sigma}_1^2 = \hat{\sigma}_2^2$

Alternative hypothesis: H_1: $\hat{\sigma}_1^2 \neq \hat{\sigma}_2^2$

Level of significance: $\propto = 0.05$

Test statistic:

The test statistic is

$$F = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_{21}^2}$$

$$= \frac{20.7729}{17.7502}$$

$F = 1.1703$

Critical region:

Here $\upsilon_1 = n_1 - 1 = 13 - 1 = 12$

And $\upsilon_2 = n_2 - 1 = 9 - 1 = 8$

$\therefore$ The critical value is

$F_\alpha(\upsilon_1, \upsilon_2) = F_{0.05}(12, 8) = 3.28$

And the critical region is to the right of 3.28

As $F = 1.1703 < F_\alpha(\upsilon_1, \upsilon_2) = 3.28$,

we conclude that the two random samples are drawn from two normal populations with the same variance.

Non-Parametric Tests

Goodness of Fit Test

To determine, if a population follows a specified theoretical distribution such as binomial, Poisson or normal distribution, χ^2 test can be used. χ^2 test, which is based on how good a fit is there between the observed frequencies (O_i from the sample) and the expected frequencies (E_i from the theoretical distribution) is known as 'goodness of fit test'.

Let a distribution be given. Let O_i and E_i ($I = 1, 2, 3 \ldots n$) be the observed and expected frequencies of the ith class (or cell) such that $\sum_{i=1}^{n} O_i = \sum_{i=1}^{n} E_i = N =$ Total Frequency.

Test statistic (OR) Statistic for test of 'goodness of fit'

$$= \chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

where n is the number of class intervals or cells, in the given frequency distribution and χ^2 is a random variable which is very closely approximated with χ^2-distribution with degrees of freedom ν.

Degrees of Freedom for Goodness of Fit Test Based on the theoretical distribution given, the degrees of freedom are as given below.

1. For uniform distribution, $\nu = n - 1$
2. For binomial and Poison distribution, $\nu = n - 2$
3. For normal distribution, $\nu = n - 3$

NOTES

Given data should satisfy the following conditions.

1. Sample size (OR) the number of sample observations, N should be more than 50($N \geq 50$)
2. If individual frequencies (O_i and/or E_i) is/are small say less than 10, then combine neighbouring frequencies in such a way that, they will be ≥ 10.
3. The number of class or cells n should be neither too small nor too large. Generally, $4 \leq n \leq 16$.

Example 32

Fitting a Poisson distribution to the following data:

x_i	0	1	2	3	4
Observed Frequencies (O_i)	30	62	46	10	2

The following respective expected frequencies are obtained.

Expected Frequencies (E_i): 42 54 34 15 5

Test the goodness of fit of a Poisson distribution to the above data with 1% level of significance.

Solution

We have

Observed Frequencies (O_i)	30	62	46	10	2
Expected Frequencies (E_i)	42	54	34	15	5

Grouping the classes so that each class frequency is ≥10,

We have

O_i	30	62	46	12
E_i	42	54	34	20
$O_i - E_i$	–12	8	12	–8

Null hypothesis: H_0: Good fit exists between the theoretical (Poisson) distribution and given data (observed frequencies)

Alternative hypothesis: H_1: No good fit exists between the theoretical (Poisson) distribution and given data (observed frequencies).

Level of Significance: $-\alpha = 0.01$

Test Statistic:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

$$= \frac{(-12)^2}{42} + \frac{8^2}{54} + \frac{12^2}{34} + \frac{(-8)^2}{20}$$

$$= 3.4286 + 1.1852 + 4.2353 + 3.2$$

$\therefore \chi^2 = 12.0491$

Critical region: As we are testing the fitting of Poisson distribution,

Degrees of Freedom = $\upsilon = n - 2$

$= 4 - 2 = 2$

$\therefore \chi^2 \alpha = \chi^2_{0.01}$ With $\upsilon = 2$

$= 9.210$

$\therefore$ The critical region is to the right of 9.210 under χ^2 distribution curve.

Decision: As the value of the test statistic

$\chi^2 = 12.0491 > \chi^2_\alpha (= 9.210)$

We reject the null hypothesis.

$\therefore$ Accept the alternative hypothesis.

Hence no good fit exists between the theoretical (Poisson) distribution and given data (observed frequencies).

Analysis of rXc Contingency Tables

Consider two attributes A and B of the given population. Let each of these attributes are classified into different classes (categories), say the attribute A is divided into r classes $A_1, A_2, \ldots, A_r$ and the attribute B is divided into c classes $B_1, B_2, \ldots, B_c$. Let a table (matrix) be formed with the classed of attribute A as heading rows and the classes of attribute B as heading columns as shown below. In the table, the values $O_{i*} (i = 1, 2, \ldots, r$ and $j = 1, 2, \ldots,$ (C) are known as observed frequencies which denote the number of items belonging to both A_i and B_j; O_{i*} denote the number of items belonging to the class A_i and O_{*j} denote the number of items belonging to the class B_j. This table is known as *rxc* contingency table.

***rXc* Contingency Table**

A \ B	B_1	B_2	...	B_j	...	B_c	Row Total
A_1	O_{11}	O_{12}	...	O_{ij}	...	O_{1c}	O_{1*}
A_2	O_{21}	O_{22}	...	O_{2j}	...	O_{2c}	O_{2*}
⋮	⋮	⋮		⋮		⋮	⋮
A_i	O_{i1}	O_{i2}	...	O_{ij}		O_{ic}	O_{i*}
⋮	⋮	⋮		⋮		⋮	⋮
A_r	O_{r1}	O_{r2}	...	O_{rj}	...	O_{rc}	O_{r*}
Column Total	O_{*1}	O_{*2}	...	O_{*j}	...	O_{*c}	$\sum_{i=i}^{r} O_{i*} = \sum_{j=1}^{c} O_{*j} = N$

In general, these tables arise in two kinds of problems.

1. Test for Independence.
2. Test for Homogeneity.

Various requirements for testing of hypothesis in these two types of problems were described in the following table.

	Test for Independence	Test for Homogeneity
(A) Description	To test whether the given two attributes of the population are independent or not.	To test whether different classes of the attributes are homogeneous or not
(B) Expected frequency (E_{ij})	$\frac{\text{(Total observed frequency in } i\text{th row)} \times \text{(Total observed frequency in } j\text{th column)}}{\text{Total frequency}}$ i.e., $\frac{(O_{i\cdot})(O_{\cdot j})}{N}$	
(C) Test Statistic	$\chi^2 = \sum_{i-1}^{r}\sum_{j=1}^{c}\left[\frac{(O_{ij}-E_{ij})^2}{E_{ij}}\right]$	
(D) Degrees of Freedom (υ)	$v = (r-1)\times(c-1)$	
(E) Decision	1. If $\chi^2 < \chi^2_\alpha$ with v degrees of freedom, then accept the null hypothesis H_0. 2. If $\chi^2 > \chi^2_\alpha$ with v degrees of freedom, then accept the alternative hypothesis H_1.	

Example 33

Test the hypothesis with 1% level of significance that the heart problem is independent of drinking (alcoholic drinks) habits from the following experimental data on 200 persons.

	Non Drinkers	Moderate Drinkers	Heavy Drinkers
Heart problem	25	40	35
No Heart problem	50	30	20

Solution

Given contingency table is

	Non Drinkers	Moderate Drinkers	Heavy Drinkers	Row Total
Heart Problem	25	40	35	100
No Heart problem	50	30	20	100
Column Total	75	70	55	

Total number of persons $= N = 200$

The expected frequency E_{ij} is given by

$$E_{ij} = \frac{(i\text{th row total})\times(j\text{th column total})}{\text{Total number of persons } (N)}$$

$$E_{11} = \frac{100\times 75}{200} = 37.5,$$

$$E_{12} = \frac{100\times 70}{200} = 35,$$

$$E_{13} = \frac{100\times 55}{200} = 27.5,$$

$$E_{21} = \frac{100\times 75}{200} = 37.5,$$

$$E_{22} = \frac{100\times 70}{200} = 35,$$

$$E_{23} = \frac{100\times 55}{200} = 27.5$$

Null hypothesis: H_0: Heart problem and the drinking habits are independent.

Alternative hypothesis: H_1: Heart problem and the drinking habits are not independent.

Level of significance: $\alpha = 0.01$

Test statistic:

$$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\left[\frac{(O_{ij}-E_{ij})^2}{E_{ij}}\right]$$

$$= \sum_{i=1}^{2}\sum_{j=1}^{3}\left[\frac{(O_{ij}-E_{ij})^2}{E_{ij}}\right]$$

$$= \frac{(25-37.5)^2}{37.5} + \frac{(40-35)^2}{35} + \frac{(35-27.5)^2}{27.5}$$

$$= \frac{(50-37.5)^2}{37.5} + \frac{(30-35)^2}{35} + \frac{(20-27.5)^2}{27.5}$$

$= 4.1667 + 0.7143 + 2.0454 + 4.1667 + 0.7143 + 2.0454$

$\therefore \chi^2 = 13.8528$

Critical region:

Here $\alpha = 0.01$

r = Number of rows in the table = 2

c = Number of columns in the table = 3

$\therefore$ Degrees of freedom $= v = (r-1)(c-1)$

$= (2-1)(3-1) = 2$

$v = 2$

$\therefore \chi^2_\alpha$ with υ degrees of freedom.

$= \chi^2_{0.01}$ with 2 degrees of freedom

$= 9.210.$

Decision: As the test statistic $= \chi^2 = 13.8528 > \chi^2_\alpha$ $(= 9.210)$, reject the null hypothesis.

i.e., accept the alternative hypothesis.

$\therefore$ Heart problem and the drinking habits are not independent.

Example 34

Following table shows the opinions of 300 persons about 'Love Marriages'.

	Married Persons	Unmarried Persons
Good	40	50
Not good	20	60
Undecided	60	70

Test whether the opinions of married and unmarried persons are homogeneous (same) with respect to 'Love Marriages' at 0.01 level of significance.

Solution

Given contingency table is

	Married Persons	Unmarried Persons	Row Total
Good	40	50	90
Not Good	20	60	80
Undecided	60	70	130
Column total	120	180	

The expected frequencies are given by,

$$E_{11} = \frac{90 \times 120}{300} = 36,$$

$$E_{12} = \frac{90 \times 180}{300} = 54,$$

$$E_{21} = \frac{80 \times 120}{300} = 32,$$

$$E_{22} = \frac{80 \times 180}{300} = 48,$$

$$E_{31} = \frac{30 \times 120}{300} = 52,$$

$$E_{32} = \frac{130 \times 180}{300} = 78$$

Null hypothesis: H_0: The opinion of married and unmarried persons about 'Love Marriages' is homogeneous (same).

Alternative hypothesis: H_1: The opinion of married and unmarried persons about 'Love Marriages' is not homogeneous.

Level of significance: $\alpha = 0.01$

Test statistic:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \left[\frac{(O_{ij} - E_{ij})^2}{E_{ij}} \right]$$

$$\sum_{i=1}^{3} \sum_{j=1}^{2} \left[\frac{(O_{ij} - E_{ij})^2}{E_{ij}} \right]$$

$$= \frac{(40-36)^2}{36} + \frac{(50-54)^2}{54} + \frac{(20-32)^2}{32} + \frac{(60-48)^2}{48}$$

$$+ \frac{(60-52)^2}{52} + \frac{(70-78)^2}{78}$$

$\therefore \chi^2 = 10.292$

Critical region:

Here $\alpha = 0.01$

r = Number of rows in the table = 3

c = Number of columns in the table = 2

$\therefore$ Degrees of freedom = $\upsilon = (r-1)(c-1)$

$= (3-1)(2-1) = 2$

$\upsilon = 2$

$\therefore \chi^2_\alpha$ with υ degrees of freedom.

$= \chi^2_{0.01}$ with 2 degrees of freedom

$= 9.210$

Decision:

As the test statistic = χ^2

$= 10.292 > \chi^2_\alpha$ $(= 9.210)$, reject the null hypothesis.

i.e., Accept the alternative hypothesis.

$\therefore$ The opinion of married and unmarried persons about 'Love Marriages' is not homogeneous (not same).

Exercises

1. If eight unbiased coins are tossed together, then the probability that the number of heads exceeds the number of tails is

(A) $\frac{31}{128}$ (B) $\frac{1}{2}$

(C) $\frac{93}{256}$ (D) $\frac{57}{256}$

2. If A and B are two mutually exclusive and exhaustive events and the probability that the non-occurrence of A is $\frac{3}{4}$, then the probability of occurrence of B is

(A) $\frac{1}{4}$ (B) $\frac{1}{2}$

(C) $\frac{3}{4}$ (D) $\frac{1}{16}$

3. A bag contains five red balls, three black balls and a white ball. If three balls are drawn from the bag, the probability that the three balls are of different colours is

(A) $\frac{23}{28}$ (B) $\frac{5}{28}$

(C) $\frac{3}{28}$ (D) None of these

4. From a box containing 18 bulbs, of which exactly $\frac{1}{3}$ rd are defective, five bulbs are chosen at random to fit into the five bulb holders in a room. The probability that the room gets lighted is

(A) $1-\frac{{}^{6}C_5}{{}^{18}C_5}$ (B) $\frac{{}^{6}C_5}{{}^{18}C_5}$

(C) $\frac{{}^{12}C_5}{{}^{18}C_5}$ (D) $1-\frac{{}^{12}C_5}{{}^{18}C_5}$

5. On a biased dice, any even number appears four times as frequently as any odd number. If the dice is rolled thrice what is the probability that the sum of the scores on them is more than 16?

(A) $\frac{26}{375}$ (B) $\frac{112}{375}$

(C) $\frac{26}{3375}$ (D) $\frac{112}{3375}$

6. A five digit number is formed using the digits 0, 1, 2, 3, 4 and 5 at random but without repetition. The probability that the number so formed is divisible by 5 is

(A) $\frac{1}{5}$ (B) $\frac{2}{5}$

(C) $\frac{4}{25}$ (D) $\frac{9}{25}$

7. If six people sit around a circular table, the probability that two specified persons always sit side by side is

(A) $\frac{14}{15}$ (B) $\frac{11}{15}$

(C) $\frac{2}{5}$ (D) $\frac{4}{15}$

8. Eight letters are to be placed in eight addressed envelopes. If the letters are placed at random into the envelopes, the probability that exactly one letter is placed in a wrong addressed envelopes is

(A) $\frac{1}{6}$ (B) $\frac{1}{8!}$

(C) $\frac{1}{7!}$ (D) None of these

9. A puzzle in logic was given to three students A, B and C whose chances of solving it are $\frac{1}{2}, \frac{3}{4}$ and $\frac{1}{4}$ respectively. The probability that the problem being solved is

(A) $\frac{29}{32}$ (B) $\frac{31}{32}$

(C) $\frac{1}{8}$ (D) $\frac{7}{8}$

10. If A and B are two events of an experiment such that $P(A \cup B) = \frac{3}{4}$, $P(A) = \frac{7}{20}$, then find $P(B)$ given that

(i) A and B are mutually exclusive

(A) $\frac{1}{4}$ (B) $\frac{1}{5}$

(C) $\frac{3}{5}$ (D) $\frac{2}{5}$

(ii) A and B are equally likely

(A) $\frac{7}{20}$ (B) $\frac{3}{4}$

(C) $\frac{2}{5}$ (D) $\frac{13}{20}$

(iii) A and B are independent events

(A) $\frac{7}{13}$ (B) $\frac{8}{13}$

(C) $\frac{6}{13}$ (D) $\frac{2}{5}$

11. The probability that a square selected at random from a 8×8 chessboard is of size 3×3 is

(A) $\frac{8}{51}$ (B) $\frac{14}{17}$

(C) $\frac{3}{17}$ (D) $\frac{25}{204}$

12. A dice has two of its sides painted pink, two blue and two green. If the dice is rolled twice the probability that same colour appears both the times is

(A) $\frac{1}{3}$ (B) $\frac{2}{3}$

(C) $\frac{7}{9}$ (D) $\frac{8}{9}$

13. X and Y are independent events. The probability that both X and Y occur is $\frac{1}{8}$ and the probability that neither of these occur is $\frac{3}{8}$. The probability of occurrence of X can be

(A) $\frac{2}{3}$ (B) $\frac{1}{4}$

(C) $\frac{1}{3}$ (D) $\frac{3}{4}$

14. A bag contains 12 cards. 5 of these cards have the letter 'M' printed on them, 3 cards have the letter 'A' printed

on them and the remaining cards have the letter 'N' printed on them. If three cards are picked up one after the other at random, and placed on a table in that order, then what is the probability that the word formed will be 'MAN'?

(A) $\frac{5}{44}$ (B) $\frac{1}{22}$

(C) $\frac{3}{22}$ (D) $\frac{3}{44}$

15. A and B pick a card at random from a well shuffled pack of cards, one after the other replacing it every time till one of them gets a spade. The person who picks a spade is declared the winner. If A begins the game, then the probability that B wins the game is

(A) $\frac{5}{9}$ (B) $\frac{4}{9}$

(C) $\frac{3}{7}$ (D) $\frac{4}{7}$

16. A number is randomly chosen from the numbers 10 to 99. It is observed that the sum of the digits of the number is ten. Find the probability that it is divisible by five.

(A) $\frac{1}{9}$ (B) $\frac{1}{3}$

(C) $\frac{1}{2}$ (D) $\frac{2}{9}$

17. An unbiased coin is tossed a person gets ₹30 if the coin shows head, and he loses ₹15 if the coin shows tail. If three coins are tossed, the probability that the person gets ₹45 is

(A) $\frac{3}{8}$ (B) $\frac{1}{2}$

(C) $\frac{1}{10}$ (D) $\frac{1}{25}$

18. What is the probability of getting at least 6 heads when a coin is tossed 7 times if it is known that there are at least 5 heads?

(A) $\frac{5}{29}$ (B) $\frac{8}{29}$

(C) $\frac{9}{29}$ (D) None of these

19. If $P(A) = \frac{3}{5}$, $P(B^C) = \frac{6}{7}$ and $P(A \cap B) = \frac{1}{4}$, then find $P\left(\frac{A^c}{B^c}\right)$.

(A) $\frac{17}{60}$ (B) $\frac{71}{120}$

(C) $\frac{19}{60}$ (D) $\frac{29}{60}$

20. If two events A and B are such that $P(\bar{A}) = 0.4$, $P(B) = 0.7$ and $P(A \cap B) = 0.2$, then $P\left(\frac{B}{A \cup \bar{B}}\right)$ is

(A) $\frac{3}{5}$ (B) $\frac{2}{5}$

(C) $\frac{1}{4}$ (D) $\frac{4}{5}$

21. A cinema historian noted that for a brief period, all movies released were either directed by Nolan or starred Bale. Also no movie directed by Nolan starred Bale. The probability that a movie was directed by Nolan is 0.5, and the probability that a movie starred Bale is 0.5. The probability that a movie is a hit if directed by Nolan is 0.6, while the probability that a movie is a hit given that Bale acted in it is 0.4. Given that a movie is a hit, find the probability that it is directed by Nolan.

(A) 0.4 (B) 0.5

(C) 0.6 (D) 0.7

22. Probability mass function of a variate x is as follows:

x	0	1	2	3	4
$P(X = x)$	k	$2k$	$3k$	$4k$	$5k$

then $P(x \geq 3) =$

(A) $\frac{1}{3}$ (B) $\frac{4}{15}$

(C) $\frac{3}{5}$ (D) $\frac{5}{7}$

23. The expected number of trials required to open a door using a bunch of n keys of which only one is the correct key is

(A) $\frac{n}{2}$ (B) $\frac{n-1}{2}$

(C) $\frac{n+1}{2}$ (D) n

Direction for questions 24 and 25:

A variate x has the probability distribution as

x	4	8	12
$P(X = x)$	$\frac{1}{3}$	$\frac{3}{5}$	$\frac{1}{15}$

24. Values of $E(x)$ and $E(x^2)$ respectively are

(A) $\frac{104}{15}, \frac{160}{3}$ (B) $\frac{102}{15}, \frac{150}{3}$

(C) $\frac{21}{3}, \frac{160}{5}$ (D) $\frac{104}{15}, \frac{151}{3}$

25. The value of $E[(3x + 2)^2]$ is _____.

(A) 675.2 (B) 560.2

(C) 134.56 (D) 567.2

26. In the random experiment of drawing a card from 15 cards numbered 1 to 15, if x is the random variable defined by the number appeared on the card, then the expectation of x is

(A) 8 (B) 7
(C) 6 (D) 5

27. For a binominal distribution, mean is 6 and variance is 2. The number of Bernouli trials is

(A) 8 (B) 9
(C) 10 (D) 11

28. If $X(n, p)$ follows a binominal distribution with $n = 6$ such that $9P[X = 4] = P[X = 2]$, then $p =$

(A) $\frac{1}{3}$ (B) $\frac{1}{2}$
(C) 1 (D) $\frac{1}{4}$

29. The variance of a Poisson variate is given to be 1. Then, $P(X = 3)$ is

(A) $\frac{1}{e}$ (B) $\frac{1}{2e}$
(C) $\frac{1}{3e}$ (D) $\frac{1}{6e}$

30. A random variable X follows a Poisson distribution such that $P[X = 1] = P[X = 2]$. Its mean and variance are, respectively,

(A) 1, 1 (B) 2, 2
(C) $\sqrt{3}, 2$ (D) $\sqrt{2}, \sqrt{2}$

31. The probability that a person hits a target is 0.003. What is the probability of hitting the target with 2 or more bullets if the number of shots is 2000?

(A) $1 - e^{-6}$ (B) $1 - e^{6}$
(C) $1 - 7e^{6}$ (D) $1 - 7e^{-6}$

32. The expected value of a random variable with uniform distribution over the interval (2, 5) is

(A) 2 (B) $2\frac{1}{2}$
(C) $3\frac{1}{2}$ (D) $4\frac{1}{2}$

33. If X is a continuous random variable with PDF $f(x) = \frac{1}{4}$ if $-2 \le x \le 2$ and $f(x) = 0$ elsewhere, the mean of X is _____.

(A) 1 (B) 1.5
(C) 2 (D) 0

34. If X is a uniformly distributed random variable in [1, 4] then $P\left(x > \frac{3}{2}\right)$ is

(A) $\frac{1}{6}$ (B) $\frac{1}{2}$
(C) $\frac{5}{6}$ (D) $\frac{1}{4}$

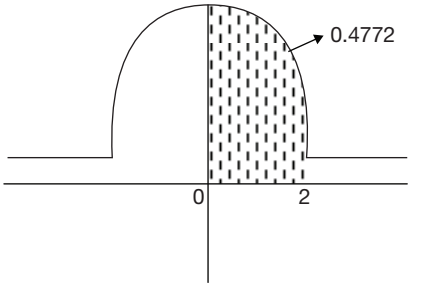
35. If X is a uniformly distributed random variable in [2, 5] then $E(X^2)$ is

(A) 2 (B) 8
(C) 13 (D) 15

36. If the life time of bulbs (in months) is exponential with mean 5 months, then the probability that the bulb lasts for atleast 7 months is

(A) 0.2466 (B) 0.7534
(C) 0.4932 (D) 0.5068

37. x is a normal variate with mean 35 and variance 25 probability of $31 \le x < 45$ is $(-0.8 \le z < 0 = 0.2881)$



(A) 0.6735 (B) 0.7563
(C) 0.7653 (D) 0.5736

38. Let X_1 and Y_1 be two discrete random variables with joint probability mass function as given below

X_1 \ Y_1	2	3	$P(X_1 = x_i)$
1	$\frac{1}{15}$	$\frac{2}{15}$	$\frac{1}{5}$
4	$\frac{4}{15}$	$\frac{8}{15}$	$\frac{4}{5}$
$P(Y_1 = y_i)$	$\frac{1}{3}$	$\frac{2}{3}$	

Let X_2 and Y_2 be two discrete random variables with joint probability mass function given as follows:

X_2 \ Y_2	0	4	7	$P(X_2 = x_i)$
−1	$\frac{1}{7}$	$\frac{3}{14}$	$\frac{1}{14}$	$\frac{3}{7}$
3	$\frac{4}{21}$	$\frac{2}{7}$	$\frac{2}{21}$	$\frac{4}{7}$
$P(Y_2 = y_i)$	$\frac{1}{3}$	$\frac{1}{2}$	$\frac{1}{6}$	

Which of the following statements is TRUE about the random variables X_1, X_2, Y_1 and Y_2?

(A) Only X_1 and Y_1 are independent.
(B) Only X_2 and Y_2 are independent.
(C) X_1 and Y_1 are independent as well as X_2 and Y_2 are independent.
(D) Neither X_1 and Y_1 are independent nor X_2 and Y_2 are independent.

39. If X and Y are two independent random variables with expectations 3 and 4 respectively. Then the expectation of XY is
(A) 1 (B) 7
(C) 12 (D) 16

40. If X and Y are two independent random variables that are uniformly distributed over the same interval [2, 5] then $P\left(X \le \frac{11}{4}, Y \ge \frac{11}{3}\right)$ is
(A) $\frac{1}{9}$ (B) $\frac{2}{9}$
(C) $\frac{1}{3}$ (D) $\frac{4}{7}$

41. The mean of cubes of first 10 natural numbers is
(A) 305 (B) 300
(C) 302.5 (D) 310

42. The mean of 25 observations was found to be 38. It was later discovered that 23 and 38 were misread as 25 and 36, then the mean is
(A) 32 (B) 36
(C) 38 (D) None of these

43. If 3, 2 and 9 occur with frequencies 2, 5 and 3 respectively, then their arithmetic mean is
(A) 4.3 (B) 5
(C) 6 (D) 4.8

44. The median of first ten prime numbers is
(A) 11 (B) 13
(C) 12 (D) 10

45. If the mean of a set of 12 observations is 10 and another set of 8 observations is 12, then the mean of combined set is
(A) 12.6 (B) 10.8
(C) 12.8 (D) 10.6

46. The mode of a distribution of 13 and its mean is 4 then its median is
(A) 7 (B) 9
(C) 8 (D) 11

47. Consider the non-decreasing series of the numbers, 1, 8, 8, 13, 14, 14, x, y, 18, 20, 31, 34, 38 and 40. If the median of the series is 15, then the mode of the series is
(A) 14 (B) 16
(C) 18 (D) Cannot be determined

48. The standard deviation of 5, 5, 5, 5, 5, 5, 5, 13 is
(A) $2\sqrt{2}$ (B) $\sqrt{6}$
(C) 5 (D) $\sqrt{7}$

49. If the standard deviation of 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 is M, then the standard deviation of 101, 102, 103, 104, … and 111 is
(A) M (B) $100 + M$
(C) $100 - M$ (D) $M - 100$

50. If the standard deviation of 10, 20, 30, 40 and 50 is S, then the standard deviation of 20, 30, 40, 50 and 60 is
(A) S (B) $S + 10$
(C) $S - 10$ (D) $10S$

51. The arithmetic mean of five observations is 6.4 and the variance is 8.24. If three of the observations are 3, 4, 8, then find the other two observations.
(A) 6, 11 (B) 10, 7
(C) 8, 9 (D) 5, 12

52. The director of a sports academy claims that the average height of sports persons in their academy is more than 170 cms. A random sample of 40 sports persons of that academy has an average height of 174 cm with a standard deviation of 15 cm. Then the claim of the director can be accepted with _____.
(A) both 1% as well as 5% levels of significance
(B) 5% level of significance but not with 1% level of significance
(C) neither 5% nor 1% levels of significance
(D) no level of significance

53. In a survey conducted on the increase in pay packages of male and female managers across IT industry by taking a random sample of 32 male managers and another random sample of 36 female managers, the following information was derived.

	Sample Size	Average Increases in Pay Package/ Annum	Standard Deviation of Increase in Pay Package/Annum
Male Managers	32	20%	4%
Female Managers	36	17%	3%

The standard error of the difference between the average increase in pay packages is _____
(A) $\frac{\sqrt{3}}{2}$ (B) $\frac{3}{\sqrt{2}}$
(C) $\frac{3}{2}$ (D) $\frac{3}{4}$

54. Match the following:

	List I		List II
P.	Standard error	**1.**	Level of significance
Q.	Type I error	**2.**	Standard deviation of the sampling distribution of a statistic
R.	Type II error	**3.**	Accepting the null hypothesis when it should be rejected
S.	Size of a test	**4.**	Rejecting the null hypothesis when it should be accepted

Codes:
(A) P – 1, Q – 2, R – 3, S – 4
(B) P – 2, Q – 4, R – 3, S – 1
(C) P – 2, Q – 3, R – 4, S – 1
(D) P – 1, Q – 3, R – 2, S – 4

55. If the population distribution is not normal, then for which of the following sample sizes, the sampling distribution of a statistic will always be normal?
(A) 4 (B) 8
(C) 16 (D) 32

56. One can reduce both type I and type II errors by
(A) reducing the sample size.
(B) increasing the sample size.
(C) changing the null hypothesis.
(D) changing the alternative hypothesis.

57. Assume that the marks obtained in GATE by the students of Civil Engineering follow normal distribution. The mean and standard deviation of marks of two groups with 10 students in each group are as given below.

	Mean	Standard Deviation
Group 1	57	3.36
Group 2	53	5.44

Then the test statistic that is to be used to test whether the difference in means of marks is significant or not is _____.
(A) 1.8768 (B) 2.3541
(C) 3.6172 (D) 4.5376

58. To test whether there is any significant difference in the marks scored by 13 students in a test before and after a yoga course using the t-distribution, the degrees of freedom to be taken is _____.
(A) 12 (B) 13
(C) 24 (D) 26

59. If $F_{0.01}$ with the degrees of freedom $\upsilon_1 = 8$ and $\upsilon_2 = 24$, is 3.36, then the value of $F_{0.99}$ with the degrees of freedom $\upsilon_1 = 24$ and $\upsilon_2 = 8$ is _____.
(A) 0.64 (B) 6.33
(C) 4.95 (D) 0.2976

60. Fitting a normal distribution to the following data:

Class	5–9	10–14	15–19	20–24	25–29	30–34	35–39
Observed frequencies (O_i)	1	10	37	36	13	2	1

The respective expected frequencies when the data is fitted to normal distribution are:

Expected Frequencies (E_i): 2, 12, 32, 36, 15, 3 and 0

The critical value to test the goodness of fit of normal distribution to the above data with 5% level of significance is _____.
(A) 14.067 (B) 9.488
(C) 3.841 (D) 2.706

61. For the contingency table given below, the test stastic (chi-square value) is _____.

A \ B		
	10	20
	30	40

(A) 0.7936 (B) 7.8361
(C) 4.8312 (D) 3.8142

Previous Years' Questions

1. If the standard deviation of the spot speed of vehicles in a highway is 8.8 km/h and the mean speed of the vehicles is 33 km/h, the coefficient of variation in speed is **[GATE, 2007]**
(A) 0.1517 (B) 0.1867
(C) 0.2666 (D) 0.3646

2. A person on a trip has a choice between private car and public transport. The probability of using a private car is 0.45. while using the public transport, further choices available are bus and metro, out of which the probability of commuting by a bus is 0.55. In such a situation, the probability (rounded up to two decimals) of using a car, bus and metro, respectively would be **[GATE, 2008]**
(A) 0.45, 0.30 and 0.25
(B) 0.45, 0.25 and 0.30
(C) 0.45, 0.55, and 0.00
(D) 0.45, 0.35 and 0.20

3. If probability density function of a random variable x is $F(x) = x^2$ for $-1 \le x \le 1$ and $= 0$ for other value of x then, the percentage probability $P\left(-\frac{1}{3} \le x \le \frac{1}{3}\right)$ is **[GATE, 2008]**
(A) 0.247 (B) 2.47
(C) 24.7 (D) 247

4. Two coins are simultaneously tossed. The probability of two heads simultaneously appearing is **[GATE, 2010]**
(A) $\frac{1}{8}$ (B) $\frac{1}{6}$
(C) $\frac{1}{4}$ (D) $\frac{1}{2}$

5. There are two containers, with one containing 4 Red and 3 Green balls and the other containing 3 Blue and 4 Green balls. One ball is drawn at random from

each container. The probability that one of the ball is Red and the other is Blue will be **[GATE, 2011]**

(A) $\frac{1}{7}$ (B) $\frac{9}{49}$

(C) $\frac{12}{49}$ (D) $\frac{3}{7}$

6. In an experiment, positive and negative values are equally likely to occur. The probability of obtaining at most one negative value in five trials is **[GATE, 2012]**

(A) $\frac{1}{32}$ (B) $\frac{2}{32}$

(C) $\frac{3}{32}$ (D) $\frac{6}{32}$

7. The annual precipitation data of a city is normally distributed with mean and standard deviation as 1000 mm and 200 mm, respectively. The probability that the annual precipitation will be more than 1200 mm is **[GATE, 2012]**

(A) <50% (B) 50%

(C) 75% (D) 100%

8. Find the value of λ such that the function $f(x)$ is a valid probability density function

$F(x) = \lambda(x-1)(2-x)$ for $1 \le x \le 2$

$= 0$ otherwise **[GATE, 2013]**

9. A fair (unbiased) coin was tossed four times in succession and resulted in the following outcomes; (i) Head, (ii) Head (iii) Head (iv) Head. The probability of obtaining a 'Tail' when the coin is tossed again is **[GATE, 2014]**

(A) 0 (B) $\frac{1}{2}$

(C) $\frac{4}{5}$ (D) $\frac{1}{5}$

10. The probability density function of evaporation E on any day during a year in a watershed is given by

$$F(E) = \begin{cases} \frac{1}{5} & 0 \le E \le 5 \text{ mm/day} \\ 0, & \text{otherwis} \end{cases}$$

The probability the E lies in between 2 and 4 mm/day in a day in the watershed is (in decimal). **[GATE, 2014]**

11. A traffic office imposes on an average 5 number of penalties daily on traffic violators. Assume that the number of penalties on different days is independent and follows a Poisson distribution. The probability that there will be less than 4 penalties in a day is _____. **[GATE, 2014]**

12. If $\{x\}$ is a continuous, real valued random variable defined over the interval $(-\infty, +\infty)$ and its occurrence is defined by the density function given as $f(x) = \frac{1}{\sqrt{2\pi} * b} e^{\frac{1}{2}\left(\frac{x-a}{b}\right)^2}$ where 'a' and 'b' are the statistical attributes of the random variable $\{x\}$. The value of the integral $\int_{-\infty}^{a} \frac{1}{\sqrt{2\pi} * b} e^{\frac{1}{2}\left(\frac{x-a}{b}\right)^2} dx$ is **[GATE, 2014]**

(A) 1 (B) 0.5

(C) π (D) $\frac{\pi}{2}$

13. Consider the following probability mass function (pmf) of a random variable X:

$$p(x,q) = \begin{cases} q & \text{if } X = 0 \\ 1-q & \text{if } X = 1 \\ 0 & \text{othewise} \end{cases}$$

If $q = 0.4$, the variance of X is _____. **[GATE, 2015]**

14. The probability density function of a random variable, x is

$f(x) = \frac{x}{4}(4 - x^2)$ for $0 \le x \le 2$

$= 0$ otherwise

The mean, μ_x of the random variable is _____. **[GATE, 2015]**

15. X and Y are two random independent events. It is known that $P(X) = 0.40$ and $P(X \cup Y^C) = 0.7$. Which one of the following is the value of $P(X \cup Y)$? **[GATE, 2016]**

(A) 0.7 (B) 0.5

(C) 0.4 (D) 0.3

16. Probability density function of a random variable X is given below

$$f(x) = \begin{cases} 0.25 \text{ if } 1 \le x \le 5 \\ 0 \quad \text{otherwise} \end{cases}$$

$P(x \le 4)$ is **[GATE, 2016]**

(A) $\frac{3}{4}$ (B) $\frac{1}{2}$

(C) $\frac{1}{4}$ (D) $\frac{1}{8}$

17. If $f(x)$ and $g(x)$ are two probability density functions,

$$f(x)=\begin{cases}\frac{x}{a}+1 & : -a \le x < 0\\ -\frac{x}{a}+1 & : 0 \le x \le a\\ 0 & : \text{otherwise}\end{cases}$$

$$g(x)=\begin{cases}-\frac{x}{a} & : -a \le x < 0\\ \frac{x}{a} & : 0 \le x \le a\\ 0 & : \text{otherwise}\end{cases}$$

Which one of the following statement is true? **[GATE, 2016]**

(A) Mean of $f(x)$ and $g(x)$ are same; Variance of $f(x)$ and $g(x)$ are same.

(B) Mean of $f(x)$ and $g(x)$ are same; Variance of $f(x)$ and $g(x)$ are different.

(C) Mean of $f(x)$ and $g(x)$ are different; Variance of $f(x)$ and $g(x)$ are same.

(D) Mean of $f(x)$ and $g(x)$ are different; Variance of $f(x)$ and $g(x)$ are different.

18. The spot speeds (expressed in km/h) observed at a road section are 66, 62, 45, 79, 32, 51, 56, 60, 53 and 49. The median speed (expressed in km/h) is _____. (Note: answer with one decimal accuracy) **[GATE, 2016]**

19. Type II error in hypothesis testing is **[GATE, 2016]**

(A) acceptance of the null hypothesis when it is false and should be rejected.

(B) rejection of the null hypothesis when it is true and should be accepted.

(C) rejection of the null hypothesis when it is false and should be rejected.

(D) acceptance of the null hypothesis when it is true and should be accepted.

Answer Keys

Exercises

1. C	**2.** C	**3.** B	**4.** A	**5.** D	**6.** D	**7.** C	**8.** D	**9.** A	
10. (i) D	(ii) A	(iii) B	**11.** C	**12.** A	**13.** B	**14.** B	**15.** C	**16.** A	**17.** A
18. B	**19.** B	**20.** B	**21.** C	**22.** C	**23.** C	**24.** A	**25.** D	**26.** A	**27.** B
28. D	**29.** D	**30.** B	**31.** D	**32.** C	**33.** D	**34.** C	**35.** C	**36.** A	**37.** C
38. C	**39.** C	**40.** A	**41.** C	**42.** C	**43.** A	**44.** C	**45.** B	**46.** A	**47.** D
48. D	**49.** A	**50.** A	**51.** A	**52.** B	**53.** A	**54.** B	**55.** D	**56.** B	**57.** A
58. A	**59.** D	**60.** C	**61.** A						

Previous Years' Questions

1. C	**2.** A	**3.** B	**4.** C	**5.** C	**6.** D	**7.** A	**8.** 6	**9.** B	**10.** 0.4
11. 0.26 to 0.27		**12.** B	**13.** 0.23 to 0.25		**14.** 1.06 to 1.07		**15.** A	**16.** A	**17.** B
18. 54.5	**19.** A								